Tracking Error and Depeg Mechanics on DeFi Synthetic Products

Internship Report

Alexandre Lemière

Host company:	Derivalink
Placement agency:	Acensi
Company mentor:	James Kouthon
Academic tutor:	Cyril Grunspan
School / program:	ESILV – École Supérieure d'Ingénieurs Léonard de Vinci
Internship period:	May – August 2026 (16 weeks)

August 27, 2026

Confidentiality notice. This report describes research conducted against a proprietary internal specification document owned by Derivalink. That document is never quoted or reproduced here; where its content is referenced, it is cited only by section number (e.g. §4.4), consistent with the confidentiality practice followed throughout this internship's own technical deliverables. All quantitative results presented were derived from public on-chain data, public exchange data, or code written during the internship.

Acknowledgments

I would like to thank James Kouthon, my mentor at Derivalink, for his guidance throughout this internship, for the structure of the weekly checkpoints that kept a sixteen-week solo research programme on track, and for the trust placed in an unusually open-ended research mandate. I would also like to thank Acensi for making this placement possible, and Cyril Grunspan for his supervision on the academic side. Finally, thank you to everyone at Derivalink who made a solo intern feel like part of a real research effort.

Executive Summary

This report covers a sixteen-week research internship at Derivalink, a company developing decentralized-finance (DeFi) synthetic-asset infrastructure, on the subject of tracking error and depeg mechanics across existing DeFi synthetic products. The internship's purpose was to build an empirical, data-grounded understanding of how and why synthetic assets — tokens designed to track the price of another asset without holding it directly — lose their peg, in order to inform the design of Derivalink's own second-generation protocol.

The work proceeded in four planned phases, of which the first three were substantially completed. Phase 1 built an on-chain data-engineering pipeline covering roughly twenty protocols, from long-lived survivors such as MakerDAO's DAI to infamous collapses such as Iron Finance's IRON/TITAN pair and Terra's UST. Phase 2 built a multi-level tracking-error decomposition and an implied-volatility smirk engine, clearing a hard go/no-go gate on free data after an initial paid-data assumption turned out to be wrong. Phase 3 applied econometrics, machine learning and graph theory — including a DebtRank systemic-risk model validated against three real historical collapses — to the resulting dataset.

The central, recurring finding is a disciplined one: real DeFi collapses are directionally predictable from public on-chain data with fairly simple, auditable models, but precisely how much and how fast is consistently harder to pin down than any single model admits. That includes negative results reported as plainly as positive ones — a reinforcement-learning alert system that never cleared its own pre-registered false-alarm bar, and five supervised model families that all lost to a trivial persistence baseline on every asset — and one retraction, an early Phase 2 result communicated to the mentor and later withdrawn under audit. An internship built to inform a real production design decision only has value if its failures are reported as honestly as its successes.

Derivalink's own V2 specification and Solidity prototype, and the final stress-testing and synthesis phase, were not reached within the sixteen-week window; this report states that plainly and closes with a handover.

Keywords: decentralized finance; synthetic assets; stablecoins; tracking error; depeg; on-chain data engineering; volatility modelling; cointegration; Hawkes processes; anomaly detection; systemic risk; DebtRank; implied-volatility skew.

Contents

Glossary

This report assumes no prior familiarity with decentralized finance. The terms below are used throughout and defined here once.

DeFi (decentralized finance) Financial applications built on public blockchains (mainly Ethereum and its ecosystem) that run via self-executing code ("smart contracts") rather than a bank or broker.

Synthetic asset A token engineered to track the price of another asset (a currency, a stock, a commodity, another cryptocurrency) without holding that asset directly. DAI tracks the US dollar; sTSLA (historically) tracked Tesla's share price.

Peg / depeg The target price a synthetic asset is meant to track is its *peg* (often \$1.00 for a dollar-stable token). A *depeg* is a meaningful, often sudden, departure from that target.

Tracking error (TE) The gap between a synthetic asset's actual market price and its intended peg, usually expressed in basis points (bp; $1\,\text{bp} = 0.01\%$).

Collateral The assets backing a synthetic asset's peg. Collateral can be *fully collateralized* (locked value $\geq$ synthetic supply, e.g. DAI), *partially collateralized* (a mix of hard collateral and an algorithmic mechanism, e.g. Iron Finance's IRON), or *algorithmic* (no locked collateral at all, e.g. Terra's UST).

CDP (collateralized debt position) A smart contract into which a user locks collateral to mint a synthetic asset against it, redeemable by repaying the debt.

Oracle A service that reports real-world price data on-chain, since a blockchain cannot query external data on its own. Most synthetic-asset protocols depend on one or more price oracles to know whether they are still at peg.

DEX (decentralized exchange) / liquidity pool An on-chain trading venue where assets can be swapped against a pool of locked liquidity, rather than matched against a traditional order book.

Depeg replay / historical replay This report's own term for reconstructing a real historical depeg event from on-chain data and comparing a model's prediction, made *as of* the eve of the event, against what actually happened.

Implied volatility (IV) smirk The pattern, observed across options at different strike prices, in which implied volatility is not flat but skewed — a signal option markets sometimes price in ahead of a large move in the underlying asset.

DebtRank A network-based systemic-risk metric (Battiston et al., 2012) that measures how much economic distress propagates through a network of financial dependencies when one node is shocked.

Granger causality / cointegration Standard econometric tests used throughout this report to establish whether one time series helps predict another (Granger causality) or whether two series that each wander individually move together in the long run (cointegration).

Hawkes process A statistical model of "self-exciting" event sequences, in which one event (e.g. a liquidation) temporarily raises the probability of further similar events — used here to study liquidation cascades.

General Introduction

Context

A *synthetic asset* is a token engineered to track the price of some reference — a dollar, an ounce of gold, a share of Tesla, an index — without its issuer holding that reference directly. The mechanism that holds the tracking together differs from protocol to protocol: over-collateralization, a partial collateral buffer topped up with an endogenous share token, a pure arbitrage loop with no collateral at all, a delta-neutral derivatives position, or an off-chain custodian. What these designs share is a single measurable failure mode. The token's market price drifts away from the reference it is supposed to track, and in the tail of that distribution the drift stops being noise and becomes a *depeg*: a self-reinforcing loss of the peg that, in several documented cases, destroyed the protocol outright within hours.

Those cases are not hypothetical, and they are not rare. Between 2020 and 2026 the sector produced a sequence of full-scale natural experiments: MakerDAO's DAI surviving the March 2020 liquidation cascade and the March 2023 Silicon Valley Bank contagion; Iron Finance's IRON destroyed in roughly twenty-four hours in June 2021; Terra's UST, at the time the third-largest stablecoin in existence, collapsing in May 2022; Lido's stETH decoupling under the Celsius and Three Arrows Capital unwind the following month. Every one of those events is recorded, permanently and publicly, in blockchain state. The raw material for a quantitative science of peg failure already exists; what has been missing is the work of turning it into one.

This internship was carried out at Derivalink, a company building infrastructure for on-chain synthetic assets, through the placement agency Acensi, as part of the engineering programme at ESILV (École Supérieure d'Ingénieurs Léonard de Vinci). It ran for sixteen weeks between May and August 2026, as a solo research mandate: one person, one internal research specification, structured mentor checkpoints on a Monday/Wednesday/Thursday rhythm, and no day-to-day pairing. A prior personal interest in market microstructure, developed independently by building market-making and taker trading bots on the prediction-market platform Polymarket, turned out to be directly relevant and is referenced at several points below where it shaped an engineering or modelling decision.

Problem Statement

Derivalink's motivation for commissioning this work was concrete. The company is designing a second-generation protocol, and the design decisions in front of it — how a collateral ratio should respond to volatility, whether one oracle or an aggregate of several is safer, what should trigger a minting freeze and how early — are all, underneath, empirical questions about how real synthetic assets have actually behaved under stress. They are usually answered by intuition, and by whichever two or three collapses the designer happens to remember. The premise of this internship is that they can be answered with evidence instead.

That premise turns into a single research problem:

> *Can five years of public, on-chain history across many heterogeneous synthetic-asset protocols be turned into quantitative, transferable knowledge about how and why pegs fail — knowledge precise enough to inform the design of a new protocol, rather than merely to narrate past ones?*

Answering that requires breaking it into questions that can each be tested against data, and that can each come back negative. Four were fixed at the start of the internship; the concluding chapter answers each one in turn.

	Research question	Addressed in
RQ1	**Measurement.** Can peg deviation be decomposed into interpretable components, so that an episode of tracking error is attributable to a specific part of the mechanism — collateral, trading venue, oracle — rather than reported as one undifferentiated number?	Chapter 5
RQ2	**Causality.** Within a protocol, which side of the mechanism leads the other — do shocks to the collateral drive deviation in the synthetic, or the reverse — and do the resulting liquidation cascades show measurable self-excitation?	Chapter 6
RQ3	**Prediction.** Is a depeg detectable *before* it happens, from public on-chain and options-market data, at a horizon and a false-alarm rate that would make an automated circuit breaker usable in production?	Chapters 5, 6
RQ4	**Systemic risk.** Does a protocol's structural position in the network of on-chain dependencies — what collateralizes it, which oracles it reads, which pools quote it — predict how much it suffers when a neighbour is shocked?	Chapter 6

Table 1: The four research questions this internship was organized around.

Two of these questions are answered affirmatively in this report, one is answered negatively with an unusual degree of consistency, and one turns out to be partially unanswerable from this data for a reason that is itself a finding. Which is which is stated in Chapter 10, and not softened.

Objectives

The internship's governing document is an internal research specification, referred to throughout this report only by section number (e.g. §4.4), since the document itself is proprietary to Derivalink and is not reproduced here. It laid out a sixteen-week, four-phase programme and seven target deliverables, summarized in Table 2.

Id	Deliverable	Status
D1	Unified database (Parquet + TimescaleDB) across all covered protocols	In progress
D2	Implied-volatility smirk engine (Deribit + Derive data)	Complete
D3	Historical depeg audit report (40–60 pages)	In progress
D4	Derivalink V2 smart contracts (Solidity, testnet)	Not started
D5	Final research report (60–100 pages)	Not started
D6	Bonus: at least one academic paper preprint	Not started
D7	Bonus: open-source tracking-error multi-asset dataset	Not started

Table 2: Original deliverable framework and status at the end of the internship.

Beyond producing those artefacts, the programme carried two hard go/no-go gates, both of which this internship reached, and both of which are decision points rather than milestones. The first, at the end of Phase 1, asked whether reconstructing five years of historical on-chain state was realistic on schedule across enough protocols to be useful, with an explicit fallback to a narrow three-asset scope if it was not. The second, at the end of Phase 2, asked whether the implied-volatility smirk signal showed genuine predictive value, with an explicit instruction to redesign the corresponding deliverable if it did not. Both outcomes are reported in the chapters covering their respective phases.

This report presents the work honestly against that framework. Three of the four planned phases were substantially completed; one deliverable was fully delivered; the rest were left at various stages of progress or not started, for reasons this report explains rather than glosses over.

Organisation of This Report

The report is organized in three parts.

Part I — State of the art establishes what is already known. Chapter 2 introduces synthetic assets and the peg problem for a reader with no prior exposure to decentralized finance: the taxonomy of peg mechanisms, the failure mode characteristic of each, a chronology of the depeg events this report analyses, and the formal definition of tracking error used throughout. Chapter 2.6 reviews the analytical toolkits the internship applies — volatility modelling, cointegration and causality testing, self-exciting point processes, anomaly detection and supervised early warning, reinforcement learning for alerting, network models of financial contagion, and options-implied signals — stating for each what the literature has established and what gap this work addresses.

Part II — Contribution presents the work itself. Chapter 3 sets out the research design: the technology stack, the notation and formal definitions used throughout, the experimental protocol, and the methodological commitments that govern every result that follows. Chapters 4 to 6 then follow the internship's own phase structure — data infrastructure, tracking-error decomposition and the smirk engine, and the four parallel research workstreams of Phase 3. Chapter 7 draws the cross-cutting findings together and states what they imply for Derivalink's own design decisions.

Part III — Assessment and conclusion steps back from the results. Chapter 8 accounts for the technical and professional skills the internship developed, Chapter 9 offers a critical personal assessment including what would be done differently, and Chapter 10 answers each of the four research questions of Table 1 in turn, then hands over what remains for Derivalink beyond this internship.

Part I — Context and State of the Art

Chapter 1

The Placement: Company, Team and Curriculum

The General Introduction stated the research problem. This chapter states the circumstances under which it was posed: which organisation asked the question, how a sixteen-week solo mandate was run and reviewed, and what five years of engineering school did — and did not — prepare the person who answered it to do.

It is deliberately short on company description and long on working arrangement. A research report can be read without knowing the host firm's headcount; it cannot be read fairly without knowing that there was one researcher, no peer on the same subject, three fixed checkpoints a week, and a governing document the reader is not permitted to see. Those conditions determined what could be attempted in sixteen weeks, and they are stated here once rather than offered later as excuses.

1.1 Acensi and the Derivalink project

The placement has two parties. **Derivalink** is the source of the subject: it builds infrastructure for on-chain synthetic assets, and is designing a second-generation protocol of its own, the subject of the internal specification's §10. **Acensi** is the firm through which the internship was carried out, and whose finance research team the work sat in. The company mentor, James Kouthon, set the subject, held the checkpoints and reviewed every deliverable; academic supervision was provided by Cyril Grunspan at ESILV.

Acensi is a French IT consultancy and engineering services firm (*ESN*) founded in 2003 and registered as an SAS at Courbevoie, in the La Défense business district west of Paris. It employs on the order of two thousand consultants — public sources quote figures between roughly 1,800 and 2,200, and the company's own profile gives the higher end — and operates from France alongside offices in North America, Europe, Africa and Asia. Its work spans consulting and project management, application development and integration, cloud, data science and cybersecurity, delivered mainly into banking and financial markets, telecommunications, energy, insurance and industry. The finance research team sits in that banking-and-markets practice: it is a delivery team for quantitative and data-intensive work commissioned by clients, not a product team, which is why a research mandate of this kind lands there rather than inside a protocol engineering group.

Derivalink itself is described here only through what this internship touched — the internal specification, the V2 design questions it poses, and the deliverables handed back — because the protocol it is building is unreleased and its own corporate details are not public. That is a real boundary on this chapter, and it is stated rather than filled with approximations.

The finance research team is not built around this subject alone. The same mentor has supervised other end-of-studies projects in it, including one on financial-portfolio optimisation combining deep learning with graph theory; that report was made available as a reference for what an end-of-studies deliverable is expected to contain, and the structure of the present report was checked against it. Working alongside a project that shares a mentor and a set of methods — graph construction, clustering, deep sequence models — but no subject matter fixed the standard of evidence expected without offering any shortcut on the content.

The placement also supplied infrastructure, and that mattered technically rather than merely administratively. Acensi provided a PREPROD virtual machine — two cores and 4 GiB of RAM — hosting the project's PostgreSQL/TimescaleDB database, the target of deliverable D1 in Table 2. Its first full load carried 396,453,410 rows across 25 tables, and had to be performed year by year with the query engine capped at one thread and one gibibyte, because a single-transaction load exhausted the machine outright (Section 3.3). That constraint produced the division of labour running through the whole project: every result in this report was computed on the author's own laptop against Parquet files, which are disposable, while the curated subset on the company's

virtual machine is the artefact that outlives the internship and that another analyst can query without any of this project's code.

One constraint of the placement is visible on every page. The governing document is proprietary to Derivalink: it is never quoted here, and is referenced only by section number. That is a real limitation on what this report can show — the reader sees the questions and the answers, but not the document that set them — and it is why Chapter 3 restates the research design in full rather than pointing at a specification the reader could otherwise consult.

1.2 The sector, commercially

Chapter 2 treats the design taxonomy properly; the point here is the commercial one. A synthetic-asset protocol is three things stacked: an issuance mechanism, a price feed, and a risk policy. The token is the visible layer and the least differentiated — any competent team can deploy a contract that mints a dollar-denominated claim. What is hard, and what a firm in this sector actually sells, is the set of parameters and defences deciding whether that claim still trades at a dollar on the day its collateral falls twenty percent in an hour.

Revenue comes from the same machinery. Mint and redemption fees, a stability fee on outstanding debt, and the spread earned by the protocol's own stabilisation facilities are each simultaneously an income line and a peg defence: Frax's asymmetric mint-versus-redeem schedule (Kazemian et al., 2020), Liquity's hard redemption right against the least-collateralised position (Liquity AG, 2021) and Curve's debt-ceiling-bounded PegKeepers (Curve Finance, 2023) are revenue mechanisms that exist because they are also stabilisers. That coupling is the structural peculiarity of the sector, and it is why a mispriced parameter is not a margin problem but an existential one.

The case for a second-generation protocol rests on the first generation now being a completed experiment with a published result: each design family has a documented characteristic failure, and no design in production dominates the others on every axis. Incumbency here is liquidity and integrations rather than code — a fork costs nothing, while the arbitrage capital that keeps a peg tight is scarce and slow to move (Makarov and Schoar, 2020). That leaves one differentiator plausibly available to a new entrant: risk instrumentation. Being able to state, with numbers drawn from real history rather than from a model of it, how a peg behaves under stress and at what point minting should stop is not something a competitor can fork. It is also exactly the deliverable this internship was commissioned to produce.

1.3 How the internship was organised

The mandate was a solo research programme: one person, sixteen weeks, four phases and the seven deliverables of Table 2, with no day-to-day pairing and no second engineer on the same subject. The internal specification was the governing document throughout. It defined the phases, the deliverables and — unusually for an internship brief — two hard go/no-go gates with defined consequences attached (Section 3.2). A brief specifying in advance what happens if the answer is no is a brief written by someone who expects to act on the result.

Progress was reviewed on a fixed Monday/Wednesday/Thursday checkpoint rhythm across the whole sixteen weeks. Three slots a week with a mentor whose own time was limited imposes a specific discipline: arriving with a written statement of what moved, what did not, and what decision is being asked for, rather than with a narrative of the week. The instrument that made that possible was a living roadmap updated after every completed task, with a dated changelog entry and an atomic, conventionally-formatted commit (Section 3.7). That record was built for the checkpoints and turned out to serve two further purposes: it is how this report's own numbers were verified before being written down, and it is what makes every claim in Part II traceable to a date and a commit.

Review was substantive rather than ceremonial, and one instance changed a deliverable outright. The first closing report on the machine-learning workstream stated its verdicts without the protocol,

the intermediate numbers or the mechanism behind them; the mentor judged it too thin, and it was rewritten as a full technical report — data, features, splits, folds, results, mechanism — running to 60 pages against the original nine. Nothing in the underlying result changed. What changed was that the reader could now check it, which was the whole point of commissioning the work. That review is the clearest single demonstration that the evidentiary standard set out in Chapter 3 came from the company as much as from the author.

Working alone on an open-ended mandate has one cost no amount of process removes: there is no peer to sanity-check an approach informally before a scheduled checkpoint, so a wrong turn can run for days before anyone else sees it. The compensating habit, and its consequences, are discussed in Chapter 9.

1.4 Where this internship sits in the ESILV curriculum

This internship closes the fifth and final year of ESILV's engineering programme, in the Fintech major. Four courses fed into it directly. *Blockchain Programming* supplied the mental model of a smart contract and the ability to read protocol source code, which is how most of the mechanism detail in Chapter 2 was established — documentation describes intent, the contract describes behaviour, and where they disagree the contract wins. *Rust Programming* contributed less through the language than through its discipline about ownership and allocation, which transferred directly to working under a hard memory ceiling. The *cryptography* courses explain why an on-chain record is auditable at all: signatures, hashing and commitment structures are the reason a five-year reconstruction of historical state can be trusted as evidence rather than treated as a vendor's claim. *Cryptocurrency Markets* supplied the market vocabulary — venues, order books, funding, liquidations — without which the corpus would have been unreadable. An Erasmus exchange at VUM in Varna, Bulgaria, in 2025 contributed something less tangible and used every day here: technical work conducted, documented and reported entirely in English.

The school's PI^2 industrial project ran in parallel on crypto-asset portfolio management — mean-variance allocation, GARCH volatility forecasting, on-chain indicators as features. The overlap is real but narrower than the shared vocabulary suggests. GARCH appears in both, but on different objects: the project forecasts the volatility of a return series, while this internship models the volatility of a *deviation* from a fixed reference (Bollerslev, 1986), where the quantity of interest is a tail the Gaussian-innovation default handles badly. On-chain indicators appear in both, but the project consumes them as features from a provider, whereas here the extraction layer that produces them had to be built first — and most of Phase 1 is that layer. Everything else in Part II lies outside the syllabus entirely: cointegration and error-correction modelling (Engle and Granger, 1987), self-exciting point processes (Hawkes, 1971), copulas, options inversion (Black, 1976), reinforcement learning, and a from-scratch implementation of a published contagion algorithm (Battiston et al., 2012).

Three gaps in the coursework had to be closed on the job, and naming them is more useful than listing what transferred. The first is data engineering at scale: no course covers rate-limited archival RPC extraction, resumable state surviving multi-day interruptions, partitioned columnar storage or memory-bounded query planning — and that work was the single largest consumer of hours in the internship. The second is experimental design for rare events. The degree teaches models; it does not teach that accuracy is uninformative when the positive class is a few percent of days (Saito and Rehmsmeier, 2015), nor how to build a walk-forward protocol whose purge and embargo make leakage impossible — and it is that protocol, not any of the models, that produced this report's most decisive result. The third is the domain itself: the Fintech major's finance content is oriented towards traditional markets, so collateralised debt positions, automated-market-maker invariants, oracle update models and liquidation engines had to be learned from the literature of Chapter 2 and from contract source.

1.5 Discovering how research work is actually done

A school project has a known finish line and a grade at the end of it. This did not. The output was an input to a real design decision — Section 7.2 states which parameters it now informs — taken by people who would act on the answer, which changes the failure mode completely. A school project fails by being incomplete. This kind of work fails by being confidently wrong, because a design decision taken on the strength of an artefact carries unearned confidence.

An open-ended mandate also does not tell you which model to fit. The deliverable is not a model but a defensible answer, which means the evaluation standard is itself part of the deliverable and has to be fixed before the numbers arrive. Fix it afterwards and it becomes, unavoidably, the standard that the result you happen to have already passes. Writing the label, the folds, the metric and the baseline down in advance felt like bureaucracy in week six; it is the reason the results in Chapter 6 can be believed. The third difference — reporting a negative result, and retracting a positive one, to a company that was hoping otherwise — is the substance of Chapter 9, and is left there rather than previewed here.

Chapter 2

Background: Peg Mechanisms and the Methods for Studying Them

This chapter establishes the ground the contribution stands on: what holds a synthetic asset's peg, how each design family characteristically breaks, the formal statement of tracking error this report uses throughout, and the analytical toolkits Part II applies together with what each one costs when moved out of the setting it was built for. Nothing here is a finding of this internship. It is deliberately compressed; the per-family chronologies, secondary specifications and derivations behind it live in the technical sub-reports listed in Appendix A.2.

2.1 Pegs and the arbitrage that holds them

A *synthetic asset* is a token engineered to track the price of some reference without its issuer necessarily holding that reference. The target price is its *peg*; a *depeg* is a meaningful departure from it; the running measure of the departure is *tracking error* (TE), quoted in basis points. The token is issued against *collateral*, which may cover the issued supply fully, partially or not at all, most often through a *collateralized debt position* (CDP). Because a blockchain cannot query the outside world, the price the protocol acts on comes from an *oracle*, and the token usually trades on a *decentralized exchange* (DEX) quoting mechanically from its own reserves.

None of these components is what holds a peg. What holds a peg is an arbitrage loop: some agent must be able to buy the token below target, convert it into something worth the target through the protocol's own rail, and keep the difference — and the symmetric operation must exist above target. The peg is the fixed point of that loop, and is only as tight as the loop is cheap, open and fast. Lyons and Viswanath-Natraj (2023) identify primary-market arbitrage as the channel through which deviations are retired, and the same framing underlies the treatment of stablecoins as privately issued liabilities whose depegs are runs rather than mispricings (Gorton and Zhang, 2023; Catalini et al., 2022). Three design choices, and only three, characterise a peg mechanism at the level this report needs: what backs the claim, who may redeem it and at what price, and which price the protocol reads as truth.

Five frictions determine how well the loop closes: available depth; the cycle's steps, fees and per-leg exposure; the asymmetry of slippage between its two directions, since a mechanism making restoration from below expensive and departure from above cheap produces a systematically one-sided deviation (Makarov and Schoar, 2020; Milionis et al., 2022); the fact that the price the protocol reads is not the price the market trades at, because an oracle updates discretely on a heartbeat or deviation trigger (Eskandari et al., 2021; Chainlink Labs, 2024); and the public transaction queue, where part of the value the loop should deliver is extracted before it arrives (Qin et al., 2022). The last two have no counterpart in the index-fund setting from which the vocabulary of tracking error is borrowed.

2.2 A taxonomy of peg mechanisms

Design is not incidental to stability (Gadzinski et al., 2023), and Catalini et al. (2022) give the taxonomy this report extends to eight families, because the corpus covers designs a three-way stablecoin classification has no slot for.

Fiat-reserve-backed (USDC, USDT): claims on off-chain cash and short-dated paper. There is no collateralization ratio, no liquidation engine and no on-chain reserve to observe, so most of the machinery applied elsewhere is replaced by a study of oracle behaviour, and the redemption gate sets the band width — Tether's $100,000 authorized-participant minimum permits a structurally wider deviation before redemption becomes economic. **Over-collateralized CDP** splits into three different peg machines: MakerDAO's DAI pins its upper side with a fee-bounded 1:1 USDC–DAI swap facility that transmits shocks in the other direction just as faithfully (MakerDAO, 2021);

Liquity's LUSD grants *any* holder the right to redeem \$1 of ETH at any time against the least-collateralized troves (Liquity AG, 2021), a hard mechanical floor against a merely behavioural ceiling, which is why LUSD spends 74.2% of its history above \$1; and Curve's crvUSD has no holder-side redemption at all, its peg an algorithmic balance struck by four PegKeeper agents (Curve Finance, 2023), all four pools trading below \$1 between 71% and 79% of the time.

Partial collateral mixes exogenous collateral with an endogenous share token in proportions set by a collateral ratio CR (Kazemian et al., 2020). Minting V dollars against c units of collateral priced at p_c burns share tokens priced at p_s for the uncollateralized remainder:

$$V = \frac{c\,p_c}{\mathrm{CR}}, \qquad s_{\mathrm{burn}} = \frac{c\,p_c\,(1-\mathrm{CR})}{\mathrm{CR}\,p_s}. \tag{2.1}$$

Redemption runs the identity backwards: a holder surrendering debt d, net of the redemption fee φ_r so that $d' = d(1-\varphi_r)$, receives collateral and *newly minted* share tokens:

$$c_{\mathrm{out}} = \frac{d'\,\mathrm{CR}}{p_c}, \qquad s_{\mathrm{out}} = \frac{d'\,(1-\mathrm{CR})}{p_s}. \tag{2.2}$$

Equation (2.2) is the entire family in one line, and it is literally the same equation in FRAX V1 and in Iron Finance, which reproduced the FRAX contract decomposition almost verbatim. Iron Finance introduced one change with consequences: minting uses the *target* ratio while redemption pays out at the *effective* ratio,

$$\mathrm{ECR}_t = \min\left(\frac{V_t^{\mathrm{collateral}}}{\mathrm{supply}_t},\ 1\right), \tag{2.3}$$

so under stress, when the two diverge, redeemers are paid the larger share-token fraction implied by the lower effective ratio.

Pure algorithmic seigniorage (Terra's UST, Mirror's mAssets) has no exogenous collateral at all: the "collateral" is the protocol's own seigniorage token (U.S. Securities and Exchange Commission, 2023), and both are dead (Briola et al., 2023; Liu et al., 2023). **Shared debt pool** (Synthetix) issues synths against one pooled obligation, so there is no per-synth collateral and no per-synth redemption (Synthetix, 2025); because a synth reads the underlying's own canonical feed, *oracle fidelity* and *venue peg* become separate questions that move independently. **Liquid staking** (Lido's stETH) is a rebasing claim whose fair value is 1.0 ETH by construction (Lido DAO, 2024), redeemed through a withdrawal queue that did not exist before May 2023. **Delta-neutral synthetic dollars** (Ethena's USDe, Resolv's USR) pair a spot leg with an offsetting short perpetual, making sustained negative funding the principal mechanism-side stress vector (Ethena Labs, 2024); USDe held within ±1% of par on 99.27% of observations across 920 days. **Physically backed** tokens (Backed's xStocks) hold the underlying equity one-for-one in custody (Backed Assets (JE) Limited, 2025; Solana Foundation, 2025), tracking a live price rather than a constant, so the failure surface shifts off-chain and the deviation accumulates while the reference market is shut. Two constructions in the corpus, GMX's GLP and Vertex's perpetuals, carry no peg at all and are retained precisely because stating where the machinery stops transferring matters: a drawdown in net asset value is not a tracking error.

2.3 Failure modes

Organising failures by family produces a list; organising them by mechanism produces an analysis, because the same five mechanisms recur.

(i) The reflexive share-token spiral. The asset settling the uncollateralized part of a redemption is issued by the protocol itself, so redeeming dilutes the very thing backing the next redemption. Clements (2021) argues this fragility is inherent rather than incidental, and Klages-Mundt and Minca (2022, 2021) model it as a deleveraging spiral. Iron Finance is the cleanest documented instance, reconstructed in full as a case study in Chapter 4. Table 2.1 is why this

internship also built the FRAX pipeline: the two ran substantially the same contract logic and their outcomes differ on five measurable variables at once. The decisive ones are that FRAX's collateral ratio ratcheted *up* under stress while IRON's fell, and that FRAX's mint fee exceeded its redemption fee while IRON's did the reverse, so the redeem-and-dump loop was rewarded rather than taxed. IRON held roughly 75% hard USDC collateral when it collapsed and died anyway, which supports a claim narrower than the one usually made: partial collateralization was *necessary but not sufficient* for survival. That claim must be kept away from mechanism fatalism — a contemporaneous twin on BNB Smart Chain with identical logic did not break its peg, so the mechanism alone was not sufficient for failure either.

	FRAX V1 (survivor)	**IRON / TITAN (casualty)**	**UST / LUNA (algorithmic)**
Mechanism	Fractional CR; redeeming mints FXS	FRAX fork; redeeming mints TITAN	Fully algorithmic; burning UST mints LUNA
Share-token dilution	$\times$1.008 over two years	$\times$334,400	To trillions of units
CR direction under stress	Up +8.0 pp (85 to 93%)	Down $-$25.7 pp (99.8 to 74.1%)	0%; nothing to ratchet
Fee asymmetry	Mint 0.95% $>$ redeem 0.45%	Mint $<$ redeem; loop rewarded	Not applicable
Redeem loop ignited?	Never	53.2% of bars, mean premium +134.4%	Not applicable
Peg outcome	Held \$0.99998	\$0.582 trough, settled near \$0.76	\$0.0059; no floor

Table 2.1: The survivor-versus-casualty comparison across the reflexive family, on five variables that differ simultaneously. This is the report's own reconstruction, not a result taken from the literature.

(ii) Collateral shock through a shared leg. A fully collateralized system can depeg without becoming insolvent. In March 2023 USDC's depeg propagated into DAI through the Peg Stability Module at a mean basis of -7.6 bp, and the recurring lesson concerns measurement rather than mechanism: DAI's oracle bottomed at 0.8888 while its traded price on the Curve 3pool bottomed at 0.8283, so the price of record understated the dislocation by more than six hundred basis points. The same gap appears one layer up, where USDC's oracle floor of 0.88 sat 697 bp above the Bitstamp wick of 0.81029. Any monitoring system watching only the oracle is, by construction, watching the smaller number.

(iii) Redemption-rail closure. Where the loop is periodically shut, the deviation widens on a schedule: USDC's absolute deviation averages 3.30 bp at weekends against 1.45 bp on weekdays, while the split within the banking day is flat — day-of-week, not hour-of-day, exactly the signature of a primary rail that closes when banks do.

(iv) Off-chain and operational failure. Two of the most recent events are not mechanism failures at all. USDC's March 2023 depeg originated in reserve concentration, \$3.3 billion of roughly \$40 billion at a single failed bank, invisible to every on-chain series until disclosure (Circle Internet Financial, 2023; Federal Deposit Insurance Corporation, 2023). Resolv's March 2026 depeg originated in a leaked credential: three unauthorized calls minted 80,019,895 USR against approximately 300,000 USDC deposited, against a protocol over-collateralized at 137% the day before (Resolv Labs, 2026). Its loss-absorbing tranche was bypassed entirely, because the loss vector was supply dilution rather than trading profit and loss, and the generalisable finding is a latency asymmetry: the pause landed two hours and fifty-five minutes after the first illicit mint, a delay attributable to a four-of-four multisig requirement the single-key attacker did not face.

(v) Protocol mortality. The failure mode the literature discusses least is the one this corpus documents best. Synthetix's Wezen release froze its low-volume L1 synths at a fixed redemption rate in September 2021 (Momtazi, 2021): the synthetic Tesla tracker had tracked its reference at an oracle root-mean-square error of 0.205% while it lived, but from the freeze onward the real share price kept moving while the token did not. Mirror's mAssets met the same end through their oracle, the terminal Band Protocol push landing at block 7,870,817, 2022-06-01 21:21:15 UTC — a timestamp reconstructed on-chain for this report, two months earlier than the secondary sources record — after which tracking error is not merely large but undefined. The peg does not degrade, it stops.

Finally, a failure mode of the *detectors* rather than the pegs. Several assets above are structurally biased away from par by design. When "abnormal" stops being rare, an anomaly detector's edge over a plain threshold rule shrinks or disappears — expected behaviour of the technique, and predicting it from the taxonomy before running the model is one of the practical payoffs of writing this section first.

2.4 What the dated record implies

The events analysed directly span Black Thursday (2020-03), Iron Finance (2021-06), the Wezen freeze (2021-09), Terra and the USDT contagion trough (2022-05), the Band terminal push (2022-06), the stETH discount under the Celsius and Three Arrows unwind (2022-06) (Celsius Network LLC, 2022; Nansen Research, 2022), FTX (2022-11), Silicon Valley Bank and DAI's pass-through (2023-03), the October 2025 flash crash, and the Resolv compromise (2026-03). Three observations follow from the record rather than any single event, and each governs how inference is done in Part II.

First, the 2022 cluster is not four independent events: Terra, the stETH discount, the USDT trough and FTX are four faces of one credit unwind, propagating through overlapping collateral and holders, so treating them as independent draws would badly overstate the effective sample size available for any test. Second, March 2023 is the only event where a *fiat* rail failed and propagated on-chain, and it did so along two channels at once: mechanically into DAI through the Peg Stability Module at −7.6 bp, and *behaviourally* into USDT as a flight to safety, with USDT carrying a +149 bp premium and the USDC–USDT basis reaching −414 bp. Same shock, opposite signs — and the mechanical channel is the one a graph model can see, the behavioural one is the one it cannot. Third, the two most recent events are not peg-mechanism failures at all: Ethena's October 2025 drawdown was a venue-specific oracle failure while the on-chain rail cleared at par throughout (LlamaRisk, 2025), and Resolv's was an operational-security incident. As designs mature, the residual failure surface migrates off the mechanism and onto the operations around it, which is exactly where the least on-chain telemetry exists.

2.5 Tracking error: the formal statement this report uses

The term is borrowed from index-fund management, where Roll (1992) defines tracking error as the standard deviation of the return differential between a replicating instrument and its benchmark. Written for a synthetic priced at P^s against a reference priced at P^u, sampled at a frequency with F periods per year:

$$\mathrm{TE}^{\mathrm{Roll}} \;=\; \mathrm{sd}(r^s_t - r^u_t)\,\sqrt{F}, \qquad r^s_t = \log\frac{P^s_t}{P^s_{t-1}}, \quad r^u_t = \log\frac{P^u_t}{P^u_{t-1}}. \tag{2.4}$$

That is a *dispersion* measure and it throws away the level. For a peg the level is the point, so this report also carries the pointwise log deviation $\ell_t = \log P^{\mathrm{obs}}_t - \log P^{\mathrm{ref}}_t$ and its root-mean-square, mean-absolute and *bias* summaries; where an asset has no constant peg the same object is defined against a net asset value instead, the closest on-chain analogue to the exchange-traded-fund premium literature (Petajisto, 2017).

The second layer decomposes the deviation rather than merely measuring it, which is what makes RQ1 answerable. Conditionally, tracking error is regressed on the drivers a designer can act on:

$$\mathrm{TE}(t) \;=\; \alpha + \beta_1\,\sigma_C(t) + \beta_2 \log \mathrm{Liq}(t) + \beta_3\,\mathrm{Funding}(t) + \beta_4\,\mathrm{Gas}(t) + \varepsilon(t), \tag{2.5}$$

and unconditionally, its variance is attributed across three components:

$$\mathrm{Var}(\mathrm{TE}) \;=\; \sigma_C^2 + \sigma_S^2 + \sigma_P^2 \;+\; 2(\mathrm{Cov}_{CS} + \mathrm{Cov}_{CP} + \mathrm{Cov}_{SP}), \tag{2.6}$$

where σ_C is the collateral-volatility contribution, σ_S the synthetic-side microstructure contribution and σ_P the peg and oracle-staleness contribution. The cross-covariances are reported explicitly rather than folded into a residual, so the genuinely joint part of the variance stays visible. Both specifications are estimated with heteroskedasticity- and autocorrelation-consistent standard errors (Newey and West, 1987), since peg residuals are heteroskedastic and serially correlated.

Three respects separate this object from the index-fund quantity it is named after, and each determined how Part II's numbers were computed. *There is not one benchmark but several*, and they disagree materially under exactly the conditions one cares about, so choosing a reference is a modelling decision that must be stated. *Bias is signal*: its sign says whether the asset trades at a premium or discount, and in this corpus that sign is diagnostic of the mechanism — LUSD's positive bias follows from a hard floor paired with a soft ceiling, crvUSD's negative bias appears independently in all four PegKeeper pools. And *imputation destroys tracking-error variance*: upsampling an hourly series to one-minute resolution inserts fifty-nine zero returns for every real one and biases the estimator toward zero by the upsampling ratio, so every estimator in this project refuses imputed rows rather than silently computing on them.

Three conventions follow. Deviation is quoted in basis points against the stated reference. A 25 bp absolute deviation is the gate for the early-warning work of Chapter 6, fixed before any model was fitted. And for assets with no constant peg, the stress regime is the 95th percentile of $|\mathrm{TE}|$ within the asset's own history rather than any absolute cutoff, because an absolute cutoff would classify a tokenized equity as permanently stressed and a tight stablecoin as never stressed.

2.6 Methods, and what each one costs when transplanted

Every tool below was built somewhere else, for something else. None was designed for an object like a peg — a price that is supposed *not* to move, quoted on venues that disagree, referenced by an oracle updating on its own schedule, and held in place by a mechanism a governance vote can switch off. Each paragraph therefore closes on what its method does *not* settle, and that is where the research questions come from. Two evaluation principles recur: every model is scored against a trivial baseline on identical days, and negative results are reported at the same length as positive ones.

Volatility. Engle (1982) formalised volatility clustering and Bollerslev (1986) generalised it to the GARCH$(1,1)$ specification that remains the default baseline:

$$\sigma_t^2 \;=\; \omega + \alpha\,\varepsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2, \qquad \omega > 0, \quad \alpha, \beta \geq 0, \quad \alpha + \beta < 1, \tag{2.7}$$

with $\alpha + \beta$ read as persistence; asymmetry is captured by GJR-GARCH (Glosten et al., 1993) and EGARCH (Nelson, 1991), whose persistence statistics are *not* comparable across families, a standard and easily made error. A parallel literature measures variance directly from intraday data (Andersen et al., 2003), its persistence captured by an additive regression on three horizons (Corsi, 2009):

$$\mathrm{RV}_t \;=\; c + \beta_d\,\mathrm{RV}_{t-1} + \beta_w\,\overline{\mathrm{RV}}_{t-5:t-1} + \beta_m\,\overline{\mathrm{RV}}_{t-22:t-1} + \varepsilon_t, \tag{2.8}$$

with two-scales and kernel estimators correcting the microstructure bias that accumulates as the sampling interval shrinks (Zhang et al., 2005; Barndorff-Nielsen et al., 2008). *Does not settle:* on a

tightly pegged token these estimators can return essentially zero on a calm day, and a synthetic has several price series rather than one — an oracle feed and the raw tape can select different families and disagree about whether a leverage effect exists at all.

Cointegration and causality. A peg is a claim about a long-run relationship: two prices may each wander and yet a linear combination of them be stationary (Engle and Granger, 1987). The Engle–Granger procedure regresses one on the other and tests the residuals for a unit root,

$$y_{1,t} = \mu + \gamma\, y_{2,t} + u_t, \qquad \Delta\hat{u}_t = \rho\, \hat{u}_{t-1} + \sum_{i=1}^{k} \phi_i\, \Delta\hat{u}_{t-i} + e_t, \quad H_0 :\ \rho = 0, \tag{2.9}$$

with Johansen (1991) testing the number of relations in larger systems and predictive precedence (Granger, 1969) testing direction. *Does not settle:* the order-of-integration assumption is load-bearing and routinely skipped — applied to series stationary by construction, Johansen returns a rank equal to the system dimension, which is mechanically true and economically empty, so the augmented Dickey–Fuller (Dickey and Fuller, 1979) and KPSS (Kwiatkowski et al., 1992) tests must be run as a pair. Cointegration also establishes that a relation exists, not what enforces it, and a failure to cointegrate can be the mechanism speaking rather than the test failing.

Tail dependence. Linear correlation is close to uninformative about the corners of a joint distribution (Embrechts et al., 2002); Sklar (1959) showed any joint distribution factorises into its marginals and a copula carrying the dependence. Clayton (Clayton, 1978) and Gumbel (Gumbel, 1960) are asymmetric in opposite directions and the choice between them is itself the finding, both calibrating in closed form by inverting Kendall's rank correlation (Joe, 1997). *Does not settle:* the tail coefficients are limits estimated from a handful of joint exceedances, a point estimate can rest on two or three co-exceedances, and a copula has no clock, so it cannot distinguish contagion from a shared shock — which is why dependence is refit on volatility-standardised residuals before it is believed (Patton, 2006).

Self-excitation. Liquidations arrive in bursts, and the natural model is a process whose own history raises its future arrival rate (Hawkes, 1971):

$$\lambda(t) \;=\; \mu + \sum_{t_i < t} \alpha\, e^{-\beta\,(t - t_i)}, \qquad n \;=\; \frac{\alpha}{\beta}, \tag{2.10}$$

where the *branching ratio* n is the expected number of further events each event triggers. If $n < 1$ every cascade is finite; if $n \geq 1$ it sustains itself — a formal, estimable version of "death spiral". Estimation is by maximum likelihood, the exponential kernel admitting a recursion linear in the number of events (Ozaki, 1979); Bacry et al. (2015) survey the finance applications. *Does not settle:* two incompatible parametrisations are in common use, so a reported α means different things in different papers; the exponential kernel is an assumption; and self-excitation is not identified against a time-varying exogenous background, so a high n alone cannot distinguish a cascade from a common shock.

Early warning. A supervised framing labels each observation by whether deviation exceeds a threshold within a forward horizon; the window must be strictly future, or measured performance is tautological. Three unsupervised detectors dominate and disagree usefully about what "normal" means — Isolation Forest (Liu et al., 2008), one-class SVM (Schölkopf et al., 2001) and an autoencoder scored by reconstruction error (Hinton and Salakhutdinov, 2006; Sakurada and Yairi, 2014) — and the supervised toolkit is boosted and bagged trees (Chen and Guestrin, 2016; Ke et al., 2017; Breiman, 2001) and gated recurrent networks (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), with attribution by Shapley values (Lundberg and Lee, 2017; Lundberg et al., 2020) and text as an extra channel (Devlin et al., 2019; Araci, 2019), all mature open implementations (Pedregosa et al., 2011; Paszke et al., 2019). The interesting question is not how to fit them but how to score them. To compare models independently of a threshold, the area under the precision–recall curve is estimated by average precision,

$$\text{PR-AUC} \;=\; \sum_{n} (R_n - R_{n-1})\, P_n. \tag{2.11}$$

Under severe imbalance this and the receiver-operating characteristic disagree sharply (Saito and Rehmsmeier, 2015): the false-positive rate divides false alarms by an enormous negative count, so a detector can raise ten false alarms per true one and still trace a flattering ROC curve, while precision divides by the alerts actually raised, which is what an operator experiences. A no-skill classifier's PR-AUC also equals the base rate, so PR-AUC is comparable only within a fixed base rate. *Does not settle:* its metrics score rows while an operator experiences episodes and a monthly budget; baselines are usually absent, and on an autocorrelated peg series the zero-parameter rule "score current stress by the current deviation" is a formidable and partly tautological competitor, correctable only by restricting evaluation to observations still inside the band. Every detector also assumes abnormal is rare — an assumption the structurally biased pegs above violate by design.

Reinforcement learning poses monitoring as sequential decision-making (Sutton and Barto, 2018), with value-based deep RL (Mnih et al., 2015) and policy-gradient methods (Schulman et al., 2017) the two standard families (Raffin et al., 2021; Farama Foundation, 2026). *Does not settle:* benchmark results come from simulators sampled without limit, whereas a historical replay offers one fixed trajectory and of order ten independent stress episodes per asset; and the reward constants silently encode a prevalence, so a policy trained at one base rate is mis-calibrated the moment it meets another.

Network contagion. DebtRank (Battiston et al., 2012) removes the all-or-nothing discontinuity of clearing-vector models (Eisenberg and Noe, 2001) by propagating a continuous distress variable, which is the relevant formulation for a peg that loses part of its backing while remaining nominally solvent. Propagation runs in synchronous rounds with each node carrying a distress level and a three-valued state, so a node propagates exactly once and the recursion terminates even on a cyclic graph; the score weights the distress acquired *beyond* the initial shock by node value v_i:

$$R \;=\; \frac{\sum_{i} \big(h_i(T) - h_i(0)\big)\, v_i}{\sum_{i} v_i}. \tag{2.12}$$

The same graph supports centrality (Freeman, 1978, 1977; Bonacich, 1972; Katz, 1953; Page et al., 1999) and community detection (Newman, 2006), optimised by Louvain (Blondel et al., 2008) or Leiden (Traag et al., 2019) with spectral clustering (von Luxburg, 2007) as a third opinion; path-based metrics need weights inverted to a distance, whereas influence metrics consume the raw weight. *Does not settle:* on-chain there is no balance sheet, so v_i must be proxied; edges weighted in updates and in dollars are not commensurable; the model is static and single-round while a real collapse unfolds over days; and centrality answers "how connected", not "how healthy".

Options-implied signals. Out-of-the-money puts trade at higher implied volatilities than equidistant calls, an asymmetry read since Bates (1991) and Rubinstein (1994) as the market pricing a fat left tail, time-varying and priced (Bakshi et al., 2003; Carr and Wu, 2007; Bollerslev et al., 2013; Xing et al., 2010), and studied for crypto on far less data (Hou et al., 2020). Crypto options are futures options quoted in coin terms with rates near zero, so the natural model is Black's formula (Black, 1976), the futures adaptation of Black and Scholes (1973):

$$c_{\text{coin}} = \mathcal{N}(d_1) - (K/F)\,\mathcal{N}(d_2), \qquad p_{\text{coin}} = (K/F)\,\mathcal{N}(-d_2) - \mathcal{N}(-d_1), \tag{2.13}$$

with $d_1 = \big(\ln(F/K) + \tfrac{1}{2}\sigma^2\tau\big)/(\sigma\sqrt{\tau})$ and $d_2 = d_1 - \sigma\sqrt{\tau}$; equities need the continuous-dividend extension instead (Merton, 1973). The market-standard summary of the asymmetry is the twenty-five-delta risk reversal,

$$\text{RR}_{25} \;=\; \text{IV}(\Delta_{\text{put}} = -0.25) - \text{IV}(\Delta_{\text{call}} = +0.25)\,, \tag{2.14}$$

whose value rests on being a property of the *shape* of the surface: a model-free implied-variance index (Britten-Jones and Neuberger, 2000), of which the exchange-published crypto index is the working example (Fernando, 2021), compresses the surface into one at-the-money scalar blind to the wing. *Does not settle:* the predictive evidence is cross-sectional whereas a peg application needs a time-series test, and the effective sample is the number of independent stress clusters rather than of days, forcing block resampling (Efron, 1979; Künsch, 1989), paired tests (McNemar, 1947) and false-discovery control (Benjamini and Hochberg, 1995). A steepening skew may also be compensation for bearing tail risk rather than information about it (Carr and Wu, 2009); a graded covariate and a thresholded alarm are separate claims that can come apart; and a historical per-strike surface is a commercial product, which has kept such work rare — though a free reconstruction path exists (Deribit, 2026).

Causal designs. Interrupted time series, difference-in-differences and regression discontinuity (Thistlethwaite and Campbell, 1960) identify a discrete treatment's effect when its date is known, and decentralized protocols suit that unusually well, since a parameter change or governance vote is a timestamped transaction (Momtazi, 2021; Solana Foundation, 2025), with robust standard errors in either direction (MacKinnon and White, 1985). *Does not settle:* the counterfactual. Governance acts on the whole protocol at once, and an emergency change during a crisis often has two or three observations per side, so distinguishing "no effect" from "no power" is the central reporting obligation of this design on protocol data.

2.7 The gap this internship addresses

Each toolkit above is individually mature and each has been applied to decentralized finance. What has not been done is to apply them together, across many heterogeneous protocols at once, on data reconstructed from the chain rather than taken from a vendor, under a single measurement convention and a single evaluation standard fixed before the results were seen. The literature is organised one protocol and one method at a time, which makes regularities that appear only across protocols structurally invisible, and means the field's survivors are studied far less than its casualties — FRAX V1 and Iron Finance ran substantially the same contract logic and only one has a literature. There is also no public multi-protocol tracking-error dataset on a single native-granularity convention, which is why the magnitudes in those studies are not comparable.

Four consequences map onto the research questions. RQ1 needs one definition applied identically before its components mean anything comparable. RQ2 is a single-protocol finding until tested the same way in several, which is the only way to tell a mechanism from a coincidence. RQ3 is worth nothing without a trivial baseline scored on identical days — exactly what the early-warning literature usually omits. And RQ4 requires a graph built from real edges rather than a stylised network, and a propagation model whose error against real collapses has been measured rather than assumed. Part II builds that common protocol, Chapter 3 states it explicitly, and Chapter 7 reports what it produced, including where the answer was negative.

Part II — Contribution

Chapter 3

Research Design, Notation and Experimental Protocol

Part I established what is already known. This chapter is the hinge: it fixes, once and for the whole report, the objects every later chapter measures and the rules under which every later claim was produced.

A separate chapter for this is not ceremony. The central claims of Part II are comparative — they hold a fiat-backed stablecoin, an over-collateralized debt position, an algorithmic seigniorage token and a delta-neutral synthetic dollar against each other — and a comparison across mechanisms this heterogeneous means nothing unless one measurement convention was applied identically to all of them. Several of those claims are also negative, and a negative result is only informative if the standard it failed was written down before it was measured. Both properties are protocol claims a reader cannot check unless the protocol is stated somewhere it can be audited. The granular inventories — per-protocol coverage, the folds one by one, hyperparameters, the pre-registration list — are in the technical sub-reports listed in Appendix A.2.

The stakes are practical. Derivalink's second-generation design depends directly on this research: how a collateral ratio should respond to volatility, whether a single or an aggregated oracle is safer, what circuit-breaker logic should freeze minting under stress. A design decision taken on the strength of a result that turns out to be an artefact is worse than one taken on intuition, because it carries unearned confidence. Everything that follows exists to make the difference between those two cases visible to someone who did not run the analysis.

3.1 Methodological commitments

Real data, not synthetic fixtures. Every quantitative result is computed from real on-chain, exchange or oracle data. Where a synthetic example appears — to unit-test an algorithm before trusting it on real data — it is labelled as such, never presented as a finding.

Single-protocol pilot before generalizing. Every new capability was built and validated end to end on one well-understood protocol first, then deliberately generalized once the pipeline itself was trusted. This is slower than building for generality from the outset, but it caught real bugs earlier.

Negative results are reported as plainly as positive ones. An internship whose purpose is to inform a real design decision has no use for a report that only shows what worked. This is not a stylistic choice but the single commitment that makes the rest of the report worth reading as evidence rather than as a highlight reel.

The three are not independent: the second is what makes the third credible. A capability validated on one protocol before being generalized has a known failure mode, so when it fails on the seventh asset the failure reads as information about that asset rather than as a suspicion about the code.

3.2 The gate structure

Two hard go/no-go decision points were built into the internal specification, both of which this internship reached. The **end of Phase 1** gate asked whether historical on-chain reconstruction was realistic on schedule across enough protocols to be useful — the difference between a handful of flagship protocols and a broader systematic scope. The **end of Phase 2** gate asked whether the implied-volatility smirk showed genuine predictive value ahead of a depeg, or whether deliverable D3 should be redesigned around a different signal. A gate is only a decision point if a negative answer has a defined consequence, and both did: an explicit fallback to a three-asset scope in the first case, an explicit instruction to redesign D3 in the second. Neither was a milestone that could be declared met by having done work.

3.3 Technology stack

The stack was chosen for a solo, sixteen-week programme handling large time-series volumes on consumer hardware, not for production scale. Python throughout, with Polars (Polars Development Team, 2026) and DuckDB (Raasveldt and Mühleisen, 2019) preferred to pandas specifically for their lower memory footprint; `statsmodels` (Seabold and Perktold, 2010) and `arch` (Sheppard, 2026) for econometrics; `scikit-learn` (Pedregosa et al., 2011), XGBoost (XGBoost Contributors, 2026), LightGBM (Microsoft Corporation, 2026) and PyTorch (Paszke et al., 2019) for machine learning, with the reinforcement-learning pilot on Stable-Baselines3 (Raffin et al., 2021) against a Gymnasium interface (Farama Foundation, 2026) and the sentiment feature scored through `transformers` (Wolf et al., 2020); NetworkX (Hagberg et al., 2008), `python-igraph` (Csárdi and Nepusz, 2006) and `leidenalg` (Traag, 2026) for graphs; `web3.py` (web3.py Contributors, 2026) for direct RPC access.

Storage is split, which is why the same numbers can be reproduced two ways. The research path reads Hive-partitioned Parquet on local disk and is what every result here was computed on; the live path is a curated subset in PostgreSQL/TimescaleDB hypertables (Timescale, Inc., 2026) for a monitoring service that outlives the internship (deliverable D1). That first full load carried 396,453,410 rows across 25 tables, roughly 44 GB after loading and 6.9 GB after compression, on a two-core virtual machine with 4 GiB of RAM.

Two conventions are load-bearing. Algorithm and implementation are cited separately, because a reproduction needs the version and a critique needs the paper. And every estimator with a documented small-sample failure mode is taken from a mature library rather than reimplemented — with one deliberate exception, the DebtRank propagation of Section 2.6, written from scratch against Battiston et al. (2012) because no maintained implementation of the continuous-distress variant exists, and therefore verified against three cases with hand-computable answers before being run on real data.

3.4 Notation and formal definitions

Section 2.5 already stated tracking error formally. This section supplies what the contribution chapters need on top: which series plays each role, the estimators actually computed, the early-warning label with its constants filled in, and the quantities the verdicts are rendered on. Table 3.1 collects the symbols; secondary conventions are stated in the sub-report corpus (Appendix A.2).

Every measurement compares two price series at a common timestamp: P_t^{obs}, what the synthetic actually traded at on a named venue, and P_t^{ref}, what it was supposed to be worth, inner-joined on UTC timestamps before any statistic is computed. The choice of P^{ref} is a modelling decision, not a data-availability accident, because the candidates disagree by hundreds of basis points under exactly the conditions of interest. Four kinds are used and every statement names which: a *constant parity* τ_{peg}, an *oracle* series, a *market* series, and an *attested or computed net asset value*. Where several exist the metric is computed against each, and the disagreement between the columns is itself a result. Against a constant parity,

$$d(t) \;\equiv\; \text{dev}_{\text{bp}}(t) \;=\; \frac{P_t^{\text{obs}} - \tau_{\text{peg}}}{\tau_{\text{peg}}} \times 10^4. \tag{3.1}$$

Two tracking-error estimators are implemented, and they are not the same object. The first is the log-return estimator of Equation (2.4), annualised by $\sqrt{F}$. The second measures the gap itself rather than the variance of its returns:

$$\text{TE}^{\text{rel}} = \text{sd}\left(\frac{P_t^s - P_t^u}{P_t^u}\right), \qquad \text{MAE}^{\text{rel}} = \mathbb{E}\left[\frac{|P_t^s - P_t^u|}{P_t^u}\right], \qquad \text{MDD}^{\text{peg}} = \max_t \frac{|P_t^s - P_t^u|}{P_t^u}. \tag{3.2}$$

The distinction is enforced in code rather than by convention: the log-return estimator refuses imputed rows and raises rather than computing on them, while Equation (3.2) tolerates them

because it describes a price gap rather than a return variance. The same discipline forbids comparing a natively one-minute tracking error against one upsampled from hourly data.

What the variance decomposition actually estimates, from three aligned series (synthetic returns a_t, collateral returns c_t, peg residuals p_t), is

$$\sigma_C = \mathrm{sd}(c), \quad \sigma_S = \mathrm{sd}(a), \quad \sigma_P = \mathrm{sd}(p), \qquad \mathrm{Var(TE)} = \mathrm{Var}(a + p), \tag{3.3}$$

together with the three pairwise covariances. The load-bearing convention is that σ_C does not enter the returned variance: Equation (2.6) is the conceptual identity and Equation (3.3) the computed quantity, and reading the first as though it were the second overstates how much of the collateral channel the decomposition mechanically absorbs. One defect is carried openly because it constrains a Phase 3 result: on sUSD the estimated σ_S and σ_P are exact duplicates, traced to the peg residual having been defined as the log price itself. A collinearity guard detects the degeneracy and falls back to a bivariate system rather than reporting a rank-deficient fit. The defect is flagged and worked around; it is not fixed, and the affected estimates are marked where they appear.

The supervised early-warning label is

$$y_\tau(t) = \mathbf{1}\left[\max_{s \in (t,\, t+H]} |d(s)| > \tau \right], \qquad H = 24 \text{ h}, \qquad \tau \in \{10,\, 25,\, 50,\, 100\} \text{ bp}, \tag{3.4}$$

with a secondary regression target on future *amplitude*. Labels are computed on hourly closes so a one-minute wick cannot flip a label; wick information enters on the feature side, where it is a signal rather than a target. The thresholds are exogenous, and this is the single most important guard in the workstream: the 25, 50 and 100 bp rungs reuse band conventions that already existed in this project's own USDC sub-report and in the Chainlink USDC/USD push oracle, which triggers an update at 25 bp. No threshold was chosen by looking at the data. A single rung is the decision gate — **25 bp, for all assets**, declared in advance — and three arguments converge on it: it is a band a protocol actor already reacts to; at 10 bp the base rate explodes to between 14% and 92% depending on the asset and the label stops describing stress at all; and at 100 bp no trainable fold contains a single test hour above the threshold, which fails not as a fitting problem but as an evaluation famine.

Hours are not the unit an operator experiences, so contiguous exceedance hours are merged into episodes, two blocks separated by less than 24 hours joining into one. The collapse this produces is the real sample-size statement of the whole early-warning workstream: for USDC, 2,220 hours labelled positive at 25 bp become **73 episodes** over the asset's entire lifetime, of which 33 fall inside the unsupervised scored window and **15** inside the supervised out-of-sample window. Fifteen independent observations, not thirty thousand rows, is the ceiling every model in Chapter 6 works against.

Row-level classification is scored threshold-free on PR-AUC (Equation (2.11)), for the reason Section 2.6 gives. The operating point is chosen to maximise F_1 on validation and then frozen; it is never chosen on the test period. Event-level evaluation adds three rules that turn a row score into something an operator could act on. An episode counts as caught if the score crosses the operating point anywhere from 48 hours before its start to its end. Coverage is distinguished from silence: an episode whose window contains no scored hour is *uncovered*, not missed, recall denominators count covered episodes only, and the uncovered count is published beside every recall figure. And false alerts are counted on calm time only and merged before counting:

$$\mathrm{FA}_{30} \;=\; \frac{\#\{\text{merged false-alert spans}\}}{\text{calm days}} \times 30, \qquad \text{calm} = \{t : t \notin [\, s_e - 72\text{h},\; e_e + 72\text{h}\,]\ \forall e\}, \tag{3.5}$$

where the buffer excludes hours adjacent to *any* episode, including an uncovered one, and consecutive alert hours collapse into a single span so serial correlation cannot manufacture precision. The pre-registered budget is $\mathrm{FA}_{30} \leq 2$.

Finally, the bar. Every supervised verdict is rendered against a zero-parameter persistence rule $s_{\text{pers}}(t) = |d(t)|$, and evaluated a second time on the *onset slice* $\{t : |d(t)| \le \tau\}$, which keeps only hours whose current deviation is still inside the band. That turns the task from nowcasting into forecasting, and is where the persistence advantage should mechanically shrink. Every supervised result in Chapter 6 is reported on both slices.

Symbol	Meaning	Fixed value or convention
P_t^{obs}, P_t^{ref}	Observed price; the reference it is measured against	Inner-joined on timestamp, UTC
τ_{peg}	Constant reference parity, where one exists	1.00; 1.0 ETH for stETH
$d(t)$	Deviation from parity in basis points, Eq. (3.1)	1 bp = 0.01%
F	Annualisation periods per year	525,600 / 105,120 / 8,760 / 365
$\sigma_C, \sigma_S, \sigma_P$	Collateral, synthetic-microstructure and peg-staleness channels, Eq. (3.3)	Population sd; covariances unbiased
H	Forward label horizon	24 h, all assets
τ	Exceedance threshold	$\{10, 25, 50, 100\}$ bp; **gate 25 bp**
$y_\tau(t)$	Forward-exceedance label, Eq. (3.4)	Null, never 0, in the last H rows
$e = (s_e, e_e)$	Stress episode	24 h merge gap
θ	Operating point on the score	Max-F_1 on validation, then frozen
FA_{30}	Merged false alerts per 30 calm days, Eq. (3.5)	Budget ≤ 2; 72 h calm buffer
g	Purge plus embargo between train and test	$24 + 24 = 48$ h
T_k	Start of fold k's test block	6 calendar months per block
$s_{\text{pers}}(t)$	Persistence baseline score	Zero fitted parameters

Table 3.1: Symbols fixed for the whole report.

3.5 Experimental protocol

The data inventory. The research corpus is roughly 77 GiB of ZSTD-compressed Parquet across approximately 1.24 million files, organised as 31 raw and 54 derived tables, drawn from a registry of 182 declared upstream sources: exchange tapes, several families of decentralized exchange, push and pull oracles, protocol subgraphs, direct archival event-log extraction for dead protocols, first-party issuer APIs and one news corpus. Every `source` column is a foreign key into that registry, enforced at ingestion, so no observation exists without a declared provenance. Coverage runs from 2019 to 2026 across sixteen protocols, the largest single table — Terra Classic — holding 208,794,291 observations; the per-protocol windows and counts are tabulated in the sub-report corpus (Appendix A.2).

Certifying a pipeline before analysing it. No analysis was permitted to run on a dataset that had not passed a written audit: seven families of check over every raw partition — partition coverage and density, field presence and type, native-layer invariants, value-domain sanity, a schema round-trip, null counts, primary-key uniqueness — then cross-table invariants specific to the protocol. The load-bearing one is supply reconciliation, for a specific reason: the token supply implied by cumulative mint and burn events is built from exactly the same event stream that feeds every downstream flow statistic, so an exact match against the contract's own archival total supply, read independently at sampled block heights, certifies that the reconstruction is neither dropping nor double-counting events. It matched to the cent at all six sampled blocks for both Iron Finance and both FRAX tokens, and to the byte for DAI; the idea generalises to non-mintable constructions, Lido's reconstructed share price matching the recorded rebase net asset value to 4.4×10^{-16}. DAI passed 118 of 118 checks, UST 154 of 154, Vertex 49 of 49, against acceptance criteria written before extraction started and recorded per protocol in the sub-report corpus (Appendix A.2). Nothing downstream of a failing audit was written.

Train, validation and test. The early-warning workstream covers six assets, admitted by one criterion stated in advance — a continuous market price series on disk *and* a measurable peg — which excludes UST and sUSD, GLP, and the equity and commodity synthetics whose tracking error is dominated by the closing gap of a market that is shut most of the time. Validation is a repeated walk-forward with an expanding origin: a training core, a validation tail and a six-month test block, separated by a fixed gap $g = 24 + 24 = 48$ h. The purge is exactly the label horizon, not a safety margin — a training row stamped 30 April at noon carries a label computed from 1 May prices, so removing exactly H hours removes exactly the rows whose labelling window overlaps the test block — and the embargo is the safety margin for the residual serial correlation of the rolling features. The origin expands rather than rolls because a two-year rolling window would eventually push Terra, FTX and Silicon Valley Bank out of training entirely. The six assets give **47 folds** , listed fold by fold in the sub-report corpus (Appendix A.2), with USDC's folds five through nine covering two full years without one positive hour at 25 bp between them. Designing this geometry correctly mattered more to the eventual results than the choice of model did.

Models and capacity. Nine models and four baselines were fitted across an unsupervised, a supervised tabular and a sequential tier, plus a deep Q-network for the reinforcement-learning pilot. Capacity is deliberately small and deliberately smallest at the top of the stack: with of the order of fifteen to seventy independent positive episodes at the gate, the recurrent models are given eight or sixteen hidden units — less capacity than the trees below them, not more, because the sequential tier carries the highest overfitting risk on a problem with five to ten major events. Unsupervised hyperparameters are fixed a priori at published defaults and never tuned, because without labels there is nothing honest to tune against; the supervised ones are tuned exactly once on a single calibration fold, then frozen and reused identically everywhere , the frozen values being recorded in the sub-report corpus (Appendix A.2). No transformer model was trained, and no conclusion in this report concerns one; that is written down so the absence is not read as an oversight.

Pre-registration. One decision rule governs the entire workstream, written before any training run: *a model is retained only if it beats every trivial baseline, on the metric and at the false-alert budget declared in advance. A model that improves on another model without clearing the baseline is not promising; it is a negative result, and it is reported as one.* The sub-report corpus (appendix A.2) inventories what that required fixing ahead of time and records three occasions on which the commitment survived an unfavourable result rather than being quietly relaxed. Traceability closes the loop: every computation stage writes a manifest recording the git fingerprint of the executed code, the random seed — 42 everywhere, with no multi-seed dispersion study, which is a stated limitation — the library versions and fingerprints of its inputs.

3.6 Data acquisition reality

Free-tier blockchain access has sharp limits, and a meaningful fraction of this internship's engineering effort went into working within them rather than around them with paid services. The caps are not uniform, which is what makes them expensive: on one chain the same nominal query returned ten thousand blocks of logs from one provider, one thousand from another and ten from a third; one provider silently omits an entire class of bridge logs, and another returns an empty result rather than an error on a pruned block range — the worst possible behaviour, because the pipeline reads it as "no events happened". Two standing rules came out of this: verify that a known-active historical window returns a non-zero result before trusting any endpoint at all, and page adaptively, halving the block range on a too-large response and growing it back after a success, over resume state written atomically so a crash restarts at the last completed block. In one case the constraint forced an entirely different data source, after the planned paid options vendor turned out to cost roughly one hundred times more than estimated — a pivot described in Chapter 5, and a good example of a constraint turning into a methodological finding rather than an obstacle.

3.7 Working discipline and AI-assisted engineering

The specification was tracked in a single living roadmap, updated after every completed task with a status change, a dated changelog entry and an atomic, conventionally-formatted commit. That produced an auditable record of the sixteen weeks: every decision, dead end and result is traceable to a commit and a date, which is also how this report's own numbers were checked before being written down. A quieter constraint shaped the pace: the development laptop has 11.4 GB of RAM, which is not much for Polars/DuckDB operations over tens of millions of rows. That forced a discipline — checking available memory before any large operation, scoping queries to exact date partitions, streaming engines by default — which in one case directly caused a real out-of-memory crash mid-pipeline, caught, diagnosed and fixed rather than worked around with a bigger machine.

Claude Code was used throughout as a day-to-day development tool, in a role closer to an always-available pair programmer than an autocomplete tool: writing and reviewing extraction pipelines, drafting statistical code against a specification, and helping maintain the roadmap and commit discipline. Every result in this report was directed, reviewed and is owned by the author; the tool did not choose what to study, decide when a result was strong enough to report, or make the judgment calls Chapter 9 requires. It is named here because it materially shaped how much technical ground one person could cover in sixteen weeks, and omitting it would give an inaccurate picture of how this internship was carried out.

3.8 What this chapter fixes

For RQ1 this chapter fixes one measurement convention applied identically across mechanisms: one deviation (Equation (3.1)), two clearly separated tracking-error estimators with an enforced rule about which may be computed on which data (Equation (3.2)), and an explicit statement of what the variance decomposition computes as against what it conceptually claims (Equation (3.3)). It also names the choice of reference price as a modelling decision, which is the precondition for a finding that recurs throughout Part II: the oracle and the traded price disagree by hundreds of basis points at exactly the moments that matter. For RQ3 it fixes the label with its thresholds taken from conventions that predate the data (Equation (3.4)), the 47-fold walk-forward with a purge sized exactly to the label horizon, an operational metric (Equation (3.5)) and a zero-parameter baseline scored on identical folds — without which the machine-learning results of Chapter 6 would be uninterpretable rather than negative. For RQ2 and RQ4 the contribution is upstream of the models: a Granger or cointegration verdict and a contagion replay both inherit whatever defects their input event streams carry, so the audit procedure is what licenses those chapters to make claims at all.

One thing the protocol cannot fix, and does not pretend to. It can prevent a leak, remove a tautology and stop a threshold from being mined, but it cannot manufacture events. Fifteen independent out-of-sample episodes for the pilot asset, thirteen stress clusters carrying a true positive in the options panel, three historical collapses available for replay — and, as Section 2.4 argues, the four 2022 events are four faces of one credit unwind rather than four independent draws. Every negative result in Part II should be read against that ceiling, and every positive one should be read as having had to clear it.

Chapter 4

Phase 1: Infrastructure and Data Pipelines

None of the series this report uses — a stablecoin's traded price at the second, the collateral ratio a contract held at a given block, the exact minute a redemption oracle returned zero — exists as a downloadable file. They exist as event logs and contract state scattered across six blockchains, several dead protocols and a dozen exchange APIs, and Phase 1 turned them into something a regression can consume. The interesting question is never whether a pipeline runs but whether what it produced is true, so this chapter gives equal room to the architecture, the audit gate every dataset had to clear, and three occasions on which the pipeline was silently wrong.

4.1 ETL architecture

Each protocol has its own extractor, written against its own event log or subgraph, but all write the same shape: Hive-partitioned Parquet, written atomically so a crash mid-write cannot leave a corrupt partition, against a shared schema registry that only evolves additively — a field can be added, never silently renamed or dropped. That uniformity let every later tool, from tracking-error computation to the graph pipeline of Chapter 6, be written once and pointed at any covered protocol.

Underneath sits one commitment, made in week one and never relaxed: the store has four layers with different rules, and only one is ground truth (Table 4.1). Raw time-series and raw event tables are append-only and *never* imputed — a gap is left as a gap, because the alternative is a row that looks like an observation and is not. Everything computed lives in a derived layer, regenerable from raw, carrying per-row audit columns marking whether a value was observed or manufactured.

Layer	Contents	Mutability	Imputation
Reference	Asset registry, oracle configurations, source registry	Slowly-changing-dimension history	Not applicable
Raw time-series	Native-cadence observations from one named source	Append-only	**Never**
Raw event	Block-level events: mints, burns, liquidations, admin actions	Append-only	**Never**
Derived	Anything computed from raw: uniform grids, tracking error, regimes, graph snapshots	Regenerable	Required, with audit columns

Table 4.1: The four-layer data model. The raw layer is ground truth; regridding and aggregation happen only in the derived layer, with provenance recorded per row.

That rule is arithmetic, not fastidiousness. Forward-fill a series observed every g_{native} seconds onto a grid of g_{target} seconds: each synthetic row repeats the previous price, so its log return is exactly zero and only one return in k is real. The upsampled sample variance is then

$$\widehat{\mathrm{Var}}\left(r^{\text{ffill}}\right) \;\approx\; \frac{1}{k}\,\widehat{\mathrm{Var}}\left(r^{\text{native}}\right), \qquad k \;=\; \frac{g_{\text{native}}}{g_{\text{target}}}, \tag{4.1}$$

so an hourly series forward-filled to one minute ($k = 60$) reports a tracking error biased toward zero by a factor near sixty, and does so silently — the output is a number, not an error. The same mechanism denatures the persistence parameters of any GARCH family fitted on top of it. The estimator helpers therefore do not default to the raw layer, they *raise* when handed an upsampled frame — the difference between a convention and a guarantee.

Two storage conventions carry results. Prices and token quantities are 128-bit decimals with eighteen fractional digits rather than floats, because reconstructing minted-minus-burned-equals-total-supply to the cent is impossible if precision was lost on the way in; and each file's footer carries the schema name and the git commit hash that wrote it, so any file on disk traces back to exact code. Every `source` column is a foreign key into a hand-maintained registry, validated at ingestion, which by the end held **182 registered sources** across twenty-two categories (Eskandari et al., 2021; Chainlink Labs, 2024). For any number in this report, the chain from value to file to source identifier to endpoint is mechanical. Analytical work is Polars and DuckDB throughout (Polars Development Team, 2026; Raasveldt and Mühleisen, 2019), extraction `web3.py` against archive endpoints (web3.py Contributors, 2026).

4.2 Protocol coverage

By the end of Phase 1, and still expanding through a parallel track that ran alongside Phases 2 and 3, the pipeline covered on the order of twenty protocol families, spanning every peg-mechanism category of Section 2.2: MakerDAO, Iron Finance, FRAX V1, Terra, Synthetix's four cohorts, USDC and USDT, Lido, Ethena, Resolv, GMX and Vertex, and Backed Finance (Appendix A.1).

Coverage is not uniform and the report never pretends it is. Some assets are reconstructed at block cadence over an entire life — Iron Finance's fifty-nine days, FRAX V1's two years, Terra's twenty months to the chain halt. Others are daily panels because that is all the chain emits: GLP publishes one net asset value per day, so 1,122 rows is complete coverage, not thin coverage. Others end at an event rather than at the limit of effort. The MakerDAO window closes on 2026-03-31 because several major exchanges then auto-converted customer DAI balances and suspended DAI trading; the Synthetix equity cohort closes on 2021-09-16 because the Wezen release froze those synths (Momtazi, 2021); the Mirror cohort closes on 2022-06-01 because the terminal Band Protocol push landed at block 7,870,817 and the contracts were bricked thereafter. Each boundary is a property of the protocol and is recorded as one.

On disk this is **roughly 77 GiB of compressed Parquet across about 1.24 million files** — thirty-one raw tables, fifty-four derived tables and three hundred and twenty-four per-asset intermediates — written by 127 Python modules in twelve extractor families. The shape is uneven and the unevenness is informative: the largest single table is Terra Classic's raw event log at 50.3 GiB in 268,530 files, a dead chain holding two-thirds of the corpus, because a Cosmos-SDK chain emits its state as verbose JSON attributes and nothing about UST is reconstructible any other way (U.S. Securities and Exchange Commission, 2023; Chainalysis, 2022). The last infrastructure task moved the curated subset off the laptop: a PostgreSQL/TimescaleDB instance (Timescale, Inc., 2026) on a two-vCPU, four-gibibyte virtual machine took its first full load of **25 tables totalling 396,453,410 rows, with no empty table**, compression across 2,147 chunks taking it from approximately 44 GB to 6.9 GB.

4.3 Working inside free-tier limits

Rate limits took a great deal of this phase, and they changed the architecture rather than merely slowing it down. Every endpoint used is free — a constraint at the start and a finding by the end, since the complete Vertex archive (96 million events, 9.18 GB) came off the public Arbitrum RPC at a total data cost of zero dollars.

What ranked the endpoints was not the per-call caps, which differ by three orders of magnitude between providers, but the failure modes, established by live probing rather than by reading documentation. One archival provider silently drops the Polygon state-sync logs through which bridged USDC arrives, so a supply reconstruction built on it would be quietly incomplete. Another returns an empty list rather than an error when asked for a pruned range, which is worse than a failure because it looks like a fact. Hence the rule the extraction layer now applies everywhere:

before trusting any endpoint, verify that a known-active historical window returns a non-zero number of logs.

The other half of the strategy is resumability. Each extractor persists its cursor only *after* the corresponding Parquet batch has been flushed, so an interrupted run restarts at the last block that genuinely reached disk rather than the last block requested. That made multi-day backfills survivable on domestic Wi-Fi: the six-year MakerDAO backfill ran end to end across four days, survived six in-flight crashes and produced 1,044 month-partitions. And after an early pipeline was declared ready on the strength of unit tests that passed against a fictional schema, a six-hour live smoke run against every extractor became mandatory before any full backfill.

4.4 Verification: the audit gate

No dataset was analysed before it passed a dedicated audit script, and each script returns a verdict rather than a report. Seventeen exist. Each walks every raw partition on disk and checks seven families of property — partition completeness, declared columns and types, the raw-layer audit triple on every row, value-domain sanity, sample round-tripping through a strict validation model, non-null discipline and primary-key uniqueness — then per-asset cross-table invariants.

The load-bearing check is supply reconciliation. Cumulative net issuance from the zero-address transfer stream must equal the contract's own archival total supply, read independently at K sampled block heights:

$$\sum_{t \le b_k} m_t \;-\; \sum_{t \le b_k} u_t \;\overset{!}{=}\; \texttt{totalSupply}(b_k), \qquad k = 1, \dots, K, \quad K \ge 5. \tag{4.2}$$

The left side is built from exactly the event stream that feeds every downstream flow statistic; the right side is an independent read of contract state, so an exact match certifies that the reconstruction is neither dropping events nor double-counting them. It held to the cent at all six sampled blocks for both Iron Finance tokens and both FRAX V1 tokens, and to the byte for MakerDAO. Without mint-and-burn semantics the idea changes form: Lido's reconstructed share price matches the recorded rebase net asset value to 4.4×10^{-16}, machine precision (Lido DAO, 2024); GMX's independently-sourced net asset value and its assets-over-supply identity differ by at most 3.8×10^{-10} across all 1,122 matched days; Resolv's three illicit March-2026 mint events sum to 80,019,895 USR, within 0.02% of the figure in the protocol's own post-mortem (Resolv Labs, 2026). MakerDAO passed 118 of 118 checks, Terra 154 of 154, Vertex 49 of 49, Ethena 43 of 43.

A second gate operates on the written deliverable rather than on the data: an eight-part checklist that re-derives every number appearing more than once, requires the standard-error method to be named, flags any test with five or fewer observations as underpowered, and demands that causal language be backed by a formal test or reframed. It is deliberately read-only — it finds and records, it does not edit — so a finding cannot be quietly absorbed by the pass that discovered it.

4.5 Three times the pipeline was silently wrong

The audits exist because of what follows. None of these produced an error, an exception or an empty file; each was found by a check looking for something else; each had already contaminated downstream numbers.

An interface declaration that discarded every liquidation. MakerDAO's liquidation events were extracted against a vendored contract interface declaring the vault identifier as a 32-byte value, while the deployed contract emits it as an address. The mismatch changes the event's topic hash, so the log filter matched nothing: the extractor ran to completion, wrote no rows and reported success. The audit's month-coverage check caught it. Correcting the interface and refilling recovered 5,539 rows, of which 4,775 fall inside the March 2020 liquidation cascade, the single most analytically important cluster in MakerDAO's history. The same day surfaced a coupled

defect: both liquidation decoders scaled collateral amounts by each token's own decimal precision, but the protocol's core accounting stores those amounts in a fixed eighteen-decimal representation regardless of the token, so only the ETH vault type was correct by coincidence. Every previously stored row for the USDC, WBTC, GUSD and USDT vault types was wrong by a factor of 10^{18-d} — for the two-decimal GUSD vault, by 10^{16}.

A pool ordering that substituted one asset for another. The Curve pool used to price Synthetix's sUSD holds four coins and the extractor assumed sUSD occupied index zero. It occupies index three. The filter therefore kept the swaps of the coin that *is* at index zero — DAI — and silently discarded every genuine sUSD trade. Across the 72,211 affected rows only 36.2% lie in a plausible range around par, 16.6% exceed 10^9 and 47.2% fall below 10^{-9}, the last two the arithmetic signature of eighteen-decimal scaling applied to a six-decimal token in each direction. The severity recorded at the time was blunt and correct: the bug blocked *every* numerical claim in the corresponding sub-report. The fix was one constant; the remediation was a full re-extraction, recovering 218,000 clean events — and it retracted an earlier finding. A "two-and-a-half-year coverage gap" had been raised as a critical data-availability problem on the strength of the broken extractor's partial output; after re-extraction the gap was an artefact, and real coverage runs continuously with no month below fifty observations. In both cases a contract layout was assumed rather than read from the chain.

An out-of-memory crash on a fifty-gigabyte table. The UST historical replay of Chapter 6 required scanning the Terra Classic event table, roughly 48 GB for the UST asset identifier alone because fourteen distinct event types share the same partitions. An unscoped first scan crashed the eleven-gigabyte development laptop. The fix was not a larger machine: it was scoping every scan to the exact year and month sub-partitions the analysis window actually touches, and running Polars' streaming engine by default. This is the constraint of Chapter 3 producing an actual failure and an actual architectural response rather than a stated intention.

4.6 Case study: Iron Finance, June 2021

IRON, launched on Polygon in May 2021, targeted a $1 peg backed by hard USDC collateral plus an endogenous share token, TITAN, in proportions set by a moving collateral ratio — a near-verbatim fork of FRAX's partial-collateral design, which is what makes the two comparable. Mint and redeem are Equations (2.1) and (2.2); the one substantive deviation, minting priced at the target collateral ratio while redemption pays out at the effective ratio of Equation (2.3), is stated in Section 2.2.

The dataset spans deployment to epilogue, 2021-05-17 to 2021-07-15, so this is a complete life, not a sample of one. A seven-extractor pipeline against archival Polygon RPC assembled eight data layers per token for fifteen distinct series and approximately **18.38 million rows with zero imputed observations**. Liquidity was concentrated enough that pricing the deepest venue is very nearly pricing the market: the deepest TITAN market was SushiSwap's TITAN/WMATIC pool at $1.08 billion, also the pool whose time-weighted average price the redemption contract consulted, making it the conduit of the entire failure. Two limitations were accepted and recorded rather than worked around: no off-chain context at all, and dollar references that inherit whatever staleness the Chainlink heartbeat imposed.

On 16–17 June 2021 IRON was destroyed in a reflexive bank run that wiped out roughly $2 billion of value in about twenty-four hours. TITAN peaked intraday at $64.53, then collapsed to $0.000008 by 01:39 UTC the following morning. IRON broke below $0.99 within an hour of that peak, reached an intraday low of $0.582, and settled near $0.76 once redemptions froze — close to its own effective collateral ratio, a pattern examined quantitatively by the replay of Section 6.4. The regime break is unusually clean in the tracking-error series itself: over the pre-crash period the deviation from par averaged +27.08 bp with a standard deviation of 57.71 bp — a functioning peg — while during the run the same quantity averaged −2,171.93 bp with a standard deviation of 610.29 bp, and the IRON float went from 176 million tokens to 1.36 million.

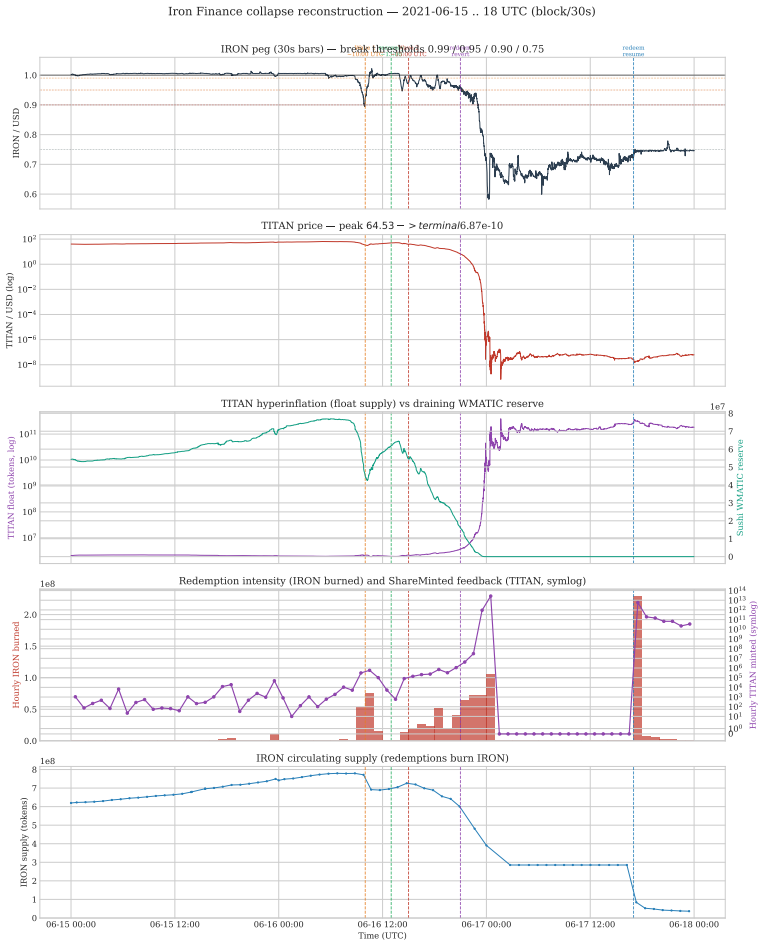

Figure 4.1: The Iron Finance collapse of 16–17 June 2021, reconstructed block by block from a continuous 30-second on-chain series. TITAN peaks at \$64.53, falls −51.7%, bounces +67.3%, then free-falls to approximately 6.9×10^{-10}. IRON holds through the morning, breaks \$0.99 at 05:46 UTC and reaches \$0.582 at 00:17 the next day; the redemption panel shows the fifteen-hour dark window caused by the contract's own zero-price guard. No public API serves this resolution.

Published accounts of this event operate at hourly granularity (Saengchote, 2021); the two-wave structure — a recoverable morning drawdown, a +67.3% relief bounce, then an unrecoverable second wave — is visible only below the minute, and so is the fifteen-hour window during which the redemption contract reverted every transaction because the share price had rounded to zero.

Two mechanical features turned a price shock into an uncontrolled spiral. TITAN was priced for redemption off a ten-minute time-weighted average of a single thin pair, so when spot fell faster than that average could track, a redeemer received more TITAN than the contemporaneous market value of the IRON surrendered and could sell it immediately. And, decisively, the TITAN paid out on redemption was minted with no supply cap: supply rose from a pre-crash float of 104.7 million tokens to approximately 35.0 trillion, a factor near 335,000, with 127.7 million TITAN minted in a single minute at the peak. This is the single design choice this internship's own forensic reconstruction identifies as fatal.

The first feature is a testable condition rather than a narrative, and testing it is what the thirty-second reconstruction was built for. On each bar the redemption loop is open when the premium of the stale average over the live spot price exceeds the redemption fee:

$$\pi_t \;=\; \frac{P_s^{\mathrm{TWAP}}(t)}{P_s^{\mathrm{spot}}(t)} \;-\; 1 \;>\; \varphi_r, \qquad \varphi_r = 0.40\%, \tag{4.3}$$

which held on **53.2% of the collapse bars**, 2,099 of 3,943, with a mean premium of 134.4% where the loop was open. Equation (4.3) cannot be evaluated on an hourly series at all, because the averaging window it compares against is ten minutes long. The difference between a description of the collapse and a measurement of its cause is entirely a difference in sampling resolution.

Early-warning signals were largely silent. Seven of eight pre-registered signals gave no warning in the thirty days before the collapse, and liquidity-provider inflows were in fact accelerating into it, chasing a roughly 50,000% annualized yield. Only the backing-cover ratio carried real information, decaying monotonically from 38.5× to 32.5× in the run-up — but never crossing any pre-set alarm level, so the warning was in its slope rather than its level. This negative result is reported as the case study's own central finding, not softened to make the analysis look more predictive than it was. It is also the first evidence this report produces on RQ3, and it arrives from the single event whose data is most complete — the least favourable possible setting for an optimistic answer.

An autopsy on its own establishes what happened, not what mattered. Table 2.1 compares Iron Finance against FRAX V1 on five variables that differ simultaneously, and that comparison exists only because the survivor's history had to be reconstructed from scratch as well — there is no literature on FRAX V1 of the kind that exists for the casualty, precisely because it did not collapse (Kazemian et al., 2020). The FRAX pipeline writes into the same four generic house tables built for Iron Finance without adding a single new one. The result is a two-year Ethereum reconstruction holding **969,501 observations with zero imputed rows**, reconciling exactly against archival total supply for both tokens. The share token FXS peaks at 1.008× its genesis supply — set beside Iron Finance's factor near 335,000, the entire comparison in one line — and its collateral ratio ratchets *up* under stress, from a 99.75% genesis to an 82.0% low and back to 92.75%, where Iron Finance's fell.

4.7 What Phase 1 established

The specification's first gate asked whether historical reconstruction was realistic on schedule, with an explicit fallback to a narrower top-three-asset scope. The programme continued at full multi-protocol scope rather than descoping — practical evidence that the gate's criteria were met in substance, even where the roadmap's own formal sign-off checkbox for the specific meeting was not separately recorded.

Phase 1 produced an instrument, not a result, and an instrument is judged by what it makes measurable and by where it stops. It makes RQ1 answerable: the variance decomposition needs the collateral, traded and oracle legs of one mechanism on a single clock at their own native cadence,

and needs the analyst to know per row which observations are real, so without the four-layer contract and the refusal encoded in Equation (4.1) a tighter tracking error is indistinguishable from a coarser data source. RQ2 is the same argument at a different timescale: a lead–lag claim is a claim about ordering, and ordering is recoverable only at the resolution the chain itself emits, which is why the Iron Finance reconstruction is thirty seconds and not one hour. It supplies the first evidence on RQ3, and that evidence is negative. And it bounds RQ4 rather than answering it: the dependency graph of Chapter 6 is built from the collateral, oracle and pool tables described above, so the network it can represent is exactly the network those tables contain, and no modelling choice can add an edge that was never emitted on-chain — a Phase-1 property with a Phase-3 consequence.

What Phase 1 did not establish is anything about behaviour. It establishes that approximately 77 GiB of reconstructed history is internally consistent, that its central supply series reconcile to the cent against independent contract reads, and that its gaps are genuine rather than manufactured. Everything downstream inherits those guarantees and their boundaries too: no off-chain order flow, no reserve composition for the fiat-backed assets whose reserves are held off-chain by design, no recovery of parameters that predate the earliest contract in a scanned lineage.

Chapter 5

Phase 2: Tracking Error Decomposition and the Implied-Volatility Smirk Engine

Phase 2 ran two workstreams that look unrelated and are not. The first looks *inside* the mechanism and asks where a peg's observed deviation comes from: which part of the variance belongs to the collateral, which to the venue the token trades on, and which to the oracle that reports its price. The second looks *outside* it, at what a liquid options market charges for downside protection, and asks whether that price moves before stress arrives. One decomposes a quantity that has already happened, the other tries to anticipate one that has not.

The first workstream is this report's answer to RQ1. What Phase 2 adds to the formal statement of Chapter 2 is an estimator applied identically across a fiat-backed stablecoin, an over-collateralized debt position, a fractional-algorithmic design, a shared debt pool and a perpetual-futures index, so the three components mean the same thing everywhere. The second workstream is one half of RQ3, and it carried the internship's second hard gate. The specification's hypothesis was that the twenty-five-delta risk reversal of Equation (2.14) steepens two to seven days before systemic stress. That hypothesis was tested to a conclusion: it produced a headline result, a better-instrumented version of the same test then withdrew half of it, and this chapter reports the withdrawal at the same length as the claim.

5.1 Multi-level tracking-error decomposition

The estimator takes three aligned series — the synthetic's own returns a_t, the collateral's returns c_t, and the peg residual p_t — and returns six numbers per window:

$$\hat{\sigma}_C = \mathrm{sd}(c), \qquad \hat{\sigma}_S = \mathrm{sd}(a), \qquad \hat{\sigma}_P = \mathrm{sd}(p), \tag{5.1}$$

$$\widehat{\mathrm{Cov}}_{CS} = \mathrm{Cov}(c,a), \qquad \widehat{\mathrm{Cov}}_{CP} = \mathrm{Cov}(c,p), \qquad \widehat{\mathrm{Cov}}_{SP} = \mathrm{Cov}(a,p). \tag{5.2}$$

One convention changes how the output must be read: the returned variance is that of the mid-side residual plus the peg gap, and therefore does *not* contain $\hat{\sigma}_C$. The collateral term enters as a driver and as a covariance partner, not as an additive share of the measured deviation. Equation (2.6) is the conceptual statement; what the code computes is narrower, and conflating the two would overstate how much of the observed tracking error the collateral channel mechanically accounts for.

The decomposition is comparable across protocols only if the mapping from each mechanism onto $(\sigma_C, \sigma_S, \sigma_P)$ is stated explicitly, because the mapping is not obvious and in two cases does not exist: UST has no vaulted collateral, so σ_{LUNA} substitutes and no third component exists, and Vertex perpetuals margin in USDC only, so there is no synthetic leg to decompose. Where a component is absent it is substituted or left undefined rather than computed anyway.

sETH is the clearest instance of the decomposition doing what it was built to do. Its three components differ by an order of magnitude: the collateral channel dominates at roughly 0.025 per day, the oracle-versus-exchange residual sits near 0.005, and the synthetic's own venue contributes about 0.0026, over 1,276 windows (Figure 5.1). Reported as one number that asset looks moderately volatile; decomposed, almost all of its variance is a property of the collateral backing it, which points a designer at the collateral policy and away from the trading venue. The same panel supports the causal reading — Granger tests put $\sigma_C \to \sigma_S$ at $p = 2.3 \times 10^{-6}$ and $\sigma_P \to \sigma_C$ at $p = 5.2 \times 10^{-4}$ — while a Johansen test of the three-variable system (Johansen, 1991) fails to reject at every rank: the components move together in the short run without sharing a long-run anchor.

A fractional-algorithmic design has no single collateral leg, so FRAX needs one extra object. What matters is not how volatile the share token is but how much of that volatility the peg is exposed to, which the collateral ratio sets directly (Kazemian et al., 2020):

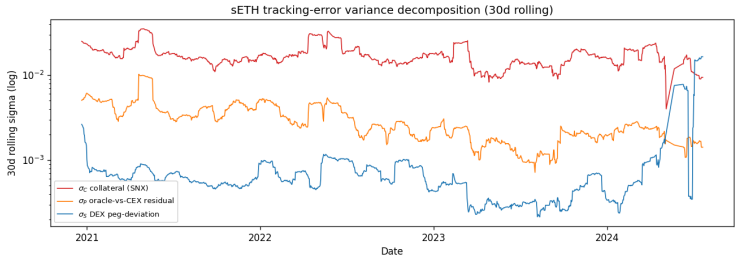

Figure 5.1: The multi-level decomposition applied to sETH: tracking error split into collateral-side σ_C, DEX peg-deviation σ_S and oracle-versus-exchange residual σ_P, on a rolling thirty-day window. The collateral channel dominates by an order of magnitude; the synthetic's own venue channel is the smallest of the three.

$$\text{peg\_risk}_t \;=\; (1 - \text{CR}_t) \times \sigma_t^{\text{Synth}}. \tag{5.3}$$

On FRAX V1, $(1 - \text{CR})$ has a median of 14.25% and never exceeds 18%, so a share-token volatility whose median is 1.04% per day and whose worst single day is 19.76% arrives at the peg with median 0.139 — an order of magnitude smaller than the raw input. This is the quantitative version of the claim in Section 2.3: the collateral ratio is not a solvency statistic, it is the gain on a transmission channel, and it is the variable a designer can turn.

Two decisions made before any model is fitted moved results more than any subsequent modelling choice. The first is the realised-variance estimator: at minute and hourly frequency an observed on-chain price carries a microstructure component whose squared increments accumulate without bound as the interval shrinks (Zhang et al., 2005; Barndorff-Nielsen et al., 2008), and taking the naive estimator would have overstated volatility throughout, propagating directly into Equation (5.1). The second is the choice of reference series, and it is more consequential because it is usually invisible. In March 2023 DAI's oracle bottomed at −1112 bp and its Curve pool at −1717 bp: the price of record understated the traded dislocation by more than six hundred basis points, and a monitoring system reading only the oracle would have measured the smaller number. There is no neutral choice here, so this project states the reference for every number it publishes and, where both exist, publishes both.

5.1.1 The conditional decomposition, and a negative result that repeats

Alongside the variance attribution, tracking error was regressed on the structural drivers a designer can act on — Equation (2.5) — instantiated per protocol, estimated with HAC standard errors (Newey and West, 1987; MacKinnon and White, 1985) and read after a Benjamini–Hochberg adjustment (Benjamini and Hochberg, 1995). The result is negative, and it repeats with unusual consistency across eight independently built pipelines (Table 5.1).

Panel	**Sampling**	$\boldsymbol{R^2}$
Vertex perpetuals (3 products)	minute	< 0.002
Synthetix sUSD ($N = 1{,}277$)	daily	0.0011
Synthetix commodity and index cohort	daily	< 0.01
MakerDAO DAI ($N = 2{,}024$)	daily	0.0173
Synthetix equity cohort (8 synths)	daily	< 0.11
Ethena USDe ($N = 686$)	daily	0.1210
Iron Finance ($N = 332$, living period)	hourly	0.173
FRAX V1 ($N = 13{,}890$)	hourly	0.341

Table 5.1: Explained variance of the conditional tracking-error regression, instantiated per protocol. The four canonical structural drivers explain almost nothing on most of the corpus; the two exceptions at the bottom are the two partial-collateral designs.

On half the panels the canonical drivers of collateral volatility, liquidity, funding and gas explain under two percent of the variation, and on the Synthetix cohorts no individual covariate is significant

at all. That is worth stating plainly, because the specification treats this regression as a workhorse and across five unrelated protocols it is not one.

The two exceptions are informative precisely because they are not random. FRAX V1 reaches $R^2 = 0.341$ and Iron Finance 0.173, and in both the binding coefficient is the same thing: the realised volatility of the endogenous share token, at +34.33 bp of |TE| per standard deviation on FRAX ($t = 7.54$, HAC $p = 4.6 \times 10^{-14}$) and +14.75 bp on Iron Finance ($p = 0.014$, the only significant driver there). The structural drivers bind exactly where the mechanism routes peg risk through a volatile leg it issues itself — an independent confirmation of Equation (5.3). And the counterpoint showing the negative is about the choice of covariate rather than the method is not in the table: on Lido's stETH after withdrawals were activated, the discount regresses on withdrawal-queue latency at −1.28 bp per day of queue, $R^2 = 0.43$, $p \approx 1 \times 10^{-87}$. When the covariate *is* the mechanism — the length of the redemption rail itself — the regression explains forty-three percent of the deviation. The four canonical drivers are simply the wrong four for most of these protocols.

Running one estimator across many protocols has a side benefit: an output that cannot be right becomes visible. Fitting a vector autoregression to sUSD's three-component panel failed on a singular covariance matrix, because σ_S and σ_P were exact duplicate columns. The peg residual had been defined as $\log p_t$ itself, already the series used for σ_S. The fitting script now flags collinearity and degrades to a bivariate system rather than reporting a fit that is not identified, but the upstream definition has not been corrected, so any sUSD σ_P in this corpus should be read as a duplicate of σ_S. The mechanism that caught it is the point: the defect had sat in a single-protocol pipeline without symptoms, and became detectable only when the same estimator was pointed at a second protocol whose output it had to resemble. Uniformity of method is a correctness check, not only a comparability convenience.

5.2 The implied-volatility smirk engine

5.2.1 A costly assumption, corrected

The specification's original data-sourcing assumption pointed to a paid historical-options vendor at an estimated few hundred dollars. That estimate was wrong by roughly two orders of magnitude — the vendor's actual historical tier costs on the order of $30,000 — which would have made the workstream infeasible. The fix was not a cheaper vendor but a free path entirely: Deribit exposes a public charting endpoint that serves daily premium data for expired option contract codes at no cost (Deribit, 2026), and resolves any generated instrument code, including options pruned from the public catalogue years ago. Historical instrument names are therefore generated rather than looked up: every Friday expiry from 2019-12 to 2026-06, crossed with a strike grid around the observed forward, yields **888,334 reconstructed per-strike legs** for no fee and no API key. Each leg's daily settlement premium is inverted against the coin-denominated Black-76 form of Equation (2.13) (Black, 1976) by a bracketed root-find. The originally planned paid source was never necessary, once someone checked what the exchange's own public infrastructure already exposed.

The risk reversal of Equation (2.14) is computed by interpolating implied volatility linearly *in delta space*, with no extrapolation: if the observed delta range on a given day does not bracket the target, that day returns no value rather than a manufactured one. A second family of shape measures — slope and curvature of a quadratic fitted to the out-of-the-money branch — was computed and then deliberately dropped from the scored pipeline during the audit, because it inherits exactly the same reconstruction caveats without adding independent information, so keeping it would have manufactured corroboration out of three correlated views of one measurement.

5.2.2 The scoring protocol

The test universe is the constraint that shapes everything else. Liquid daily option markets exist only for Bitcoin and Ethereum, so their smirk can lead an event only insofar as that event produces a realised-volatility spike in those two assets. The correct universe is therefore systemic volatility-stress events, not the stablecoin-depeg catalogue alone. Sixty-eight events were catalogued and date-verified across 2020–2026, of which sixty are smirk-testable; the remaining eight are retained as negative controls rather than deleted, which sharpens the false-positive count instead of flattering it. Every date is anchored on the volatility *onset* rather than the price trough — Terra at the 2022-05-07 Curve drain rather than the 05-09 capitulation — because a trough-anchored label would fall after the early-warning window and bias the test against its own hypothesis.

The detector is a trailing, causal, signed rolling z-score with a thirty-day window and a threshold of two, so no future information enters a fire decision; the label is strictly forward-looking:

$$y_d \;=\; \mathbf{1}[\,\exists\, k \in \{1,\ldots,7\} \;:\; d+k \in \mathcal{E}\,]\,, \tag{5.4}$$

where $\mathcal{E}$ is the set of testable event dates. A signal that merely *coincides* with stress scores no true positive; only genuine lead is rewarded. Two baselines are scored on identical days through the *same* detector, so the comparison isolates shape against level: the exchange's own thirty-day implied-volatility index (Fernando, 2021; Britten-Jones and Neuberger, 2000), a single at-the-money scalar blind to the wing by construction, and a matched-random predictor firing on a prevalence-matched fraction of grid days. If a smirk-blind scalar warns as well as the per-strike skew does, the smile is redundant and the workstream has no product.

Inference rests on one resampling unit. Stress events arrive in bursts, so the effective sample is not the number of days but the number of independent clusters; single-linkage merging at a seven-day gap gives fifty-two clusters, of which **only thirteen carry a true positive at all**. Every interval below is a cluster block-bootstrap over that unit (Efron, 1979; Künsch, 1989), paired tests are McNemar's exact test (McNemar, 1947), and multiplicity is controlled by Benjamini–Hochberg (Benjamini and Hochberg, 1995). One honesty note: this is a *retrospective* study with a look-ahead-free detector, not a formally pre-registered trial. What was frozen, before the definitive bootstrap ran, is the rule mapping the resulting intervals onto a verdict — including the wording of the two negative verdicts — so the gate call could not be chosen after seeing which one it would produce.

5.2.3 A result caught and corrected before it shipped

An initial analysis, completed early in Phase 2, found that the 25-delta skew led Bitcoin and Ethereum volatility-stress events by five to six days, beating baseline signals at an F_1 of 0.263 on Bitcoin — a promising "lean go" on a five-event sample. Before it was finalized, a systematic review of the underlying data uncovered a real defect: an earlier version of the fetching script silently emptied a time window whenever a network drop interrupted it mid-fetch, but still marked that window as successfully completed, so a later completeness check saw no missing data at all. Part of the Ethereum options backbone in that initial result had been silently corrupted this way.

The pipeline was hardened to fail loudly — a connection error is now retried with exponential backoff and then *halts* the run with its resumable state preserved, while an HTTP 400, meaning the instrument never existed, is the one case allowed to return no data — and a finalization step now refuses to produce a result unless every expected window is genuinely present. Rebuilding on the complete sixty-event panel added over half a million implied-volatility rows the corrupted run had recorded as successfully empty. This is presented deliberately as an example, not an embarrassment: a fast, exciting early result was not taken at face value, the pipeline that produced it was checked, a real bug was found, and the corrected result is the one this report and the gate decision are built on.

5.2.4 A second audit, and the claim it withdrew

The rebuilt result was then put through a second, deliberately adversarial review — this time not of the plumbing but of the claim itself. It found two problems, and both cost the workstream its headline.

The first was in the data. The recovered implied volatilities included a population of legs pinned at the solver's own upper bracket of 999.7 volatility points, and the raw skew reached +848 points on Bitcoin and +746 on Ethereum — values with no economic meaning, produced by inverting a thin or stale quote against a bracket wide enough to accommodate almost anything. Those legs were not inert: **seven of the nineteen Bitcoin true positives in the pre-audit result were driven by untraded legs** quoting at eight to eleven times the same-day volatility index. The fix was specified in writing, with its sensitivity grid, *before* its effect on the result was inspected: lower the inversion bracket from 10 to 5 at extraction, and drop any leg at the old solver ceiling or quoting above five times the contemporaneous index. It removes about 1.4% of Bitcoin legs. Its effect is asymmetric and instructive: on Bitcoin the skew loses seven of nineteen true positives while keeping essentially all its false ones, so precision falls from 0.333 to 0.227; on Ethereum the cleaning cuts false positives from 61 to 42 and precision *rises* from 0.228 to 0.288. The Bitcoin collapse is stable across every threshold in the pre-specified sweep, which rules out its being a tuned knob. The plain reading is that a meaningful part of the original Bitcoin precision had been built on artefacts.

The second problem was in the inference, and it is the more important. The lead-time statistic behind the headline "leads 3.4 days on Bitcoin, 4.1 days on Ethereum" searches only $k \in \{1, \ldots, 7\}$ — the same interval Equation (5.4) uses to define a positive — so any true positive's lead lies in $[1, 7]$ *by construction*. The diagnostic that settles it is that the matched-random predictor's own mean lead is 4.0 days, the exact midpoint of the search interval: a signal with no information whatsoever reports a lead in the same range. The statistic was measuring the width of the window, not the anticipation of the signal, and it was demoted to a descriptive quantity with no inference attached.

Anticipation was then tested directly, by a circular-shift permutation: all event dates are shifted by a single random offset within the grid span and precision recomputed against a frozen fire set, two thousand times, which asks whether the signal's fires precede *real* events more than they precede arbitrary dates. They do not. The permutation returns $p = 0.186$ on Bitcoin and $p = 0.098$ on Ethereum, and after false-discovery-rate control $q = 0.248$ and $q = 0.208$. On this panel there is no evidence that the skew anticipates stress. The anticipatory pillar of the original hypothesis falls, and this report states it as a withdrawal rather than folding it into a caveat.

5.2.5 Final result and gate decision

The definitive detector result, on the continuous de-outliered panel of 2,004 scored Bitcoin days and 1,732 Ethereum days, does not say what a one-line summary would say. The skew beats the matched-random floor on both underlyings — F_1 of 0.076 against 0.046 on Bitcoin, 0.089 against 0.055 on Ethereum. Against the smirk-blind volatility index the outcome is *asymmetric*: on Bitcoin the skew loses on F_1, 0.076 against 0.111, and is the less precise of the two, 0.227 against 0.264; on Ethereum the skew *wins*, 0.089 against 0.079, and is the more precise, 0.254 against 0.188. A blanket statement that the skew does not beat the index would be wrong as written, and so would its opposite. The right statement is the one the paired test supports: McNemar's exact test returns $p = 0.72$ on Bitcoin and $p = 0.104$ on Ethereum, so on this sample the two are statistically indistinguishable as thresholded alarms and the direction of each margin is not something to build on.

The top block of Table 5.2 is uniformly negative. Every interval on a precision difference spans zero, including against the matched-random floor, so at an effective sample of thirteen true-positive-bearing clusters the thresholded detector establishes no separable edge over either baseline, and the anticipation test does not reject. The pre-registered verdict for this half is therefore the one it was written to be able to say: *no established edge over baselines*.

Quantity	BTC	ETH
(skew − random) precision, 95% cluster CI	$[-0.086,\ 0.229]$	$[-0.074,\ 0.242]$
(skew − index) precision, 95% cluster CI	$[-0.171,\ 0.146]$	$[-0.040,\ 0.180]$
Anticipation permutation p (q_{BH})	0.186 (0.248)	0.098 (0.208)
McNemar exact p, skew versus index	0.72	0.104
Incremental logistic coefficient on z^{skew}	+0.336	+0.235
95% pairs-cluster bootstrap interval	$[0.142,\ 0.535]$	$[0.101,\ 0.357]$
bootstrap p (q_{BH})	0.0040 (0.0040)	< 0.0020 ($<$ 0.0040)
Coefficient on z^{DVOL} in the same fit	−0.048	+0.046
Strict-lag variant, predictor at $d-1$, target $[d+3, d+7]$	+0.236 ($p = 0.012$)	+0.201 ($p < 0.002$)
Leave-one-cluster-out range of the skew coefficient	$[+0.305,\ +0.381]$	$[+0.208,\ +0.254]$

Table 5.2: The inference layer, all resampled on the same unit of fifty-two merged event-window clusters. The top block concerns the thresholded detector and establishes nothing; the bottom block concerns the graded covariate of Equation (5.5) and is robust on every diagnostic applied to it.

The bottom block is where something survives, and it is a different claim about a different object. Rather than thresholding the skew, its standardised value is entered as a graded covariate in a logistic model that already contains the volatility index:

$$\Pr(y_d = 1 \,|\, z_d^{\mathrm{skew}},\, z_d^{\mathrm{DVOL}}) \;=\; \Lambda\Big(\beta_0 + \beta_1\, z_d^{\mathrm{skew}} + \beta_2\, z_d^{\mathrm{DVOL}}\Big)\,, \tag{5.5}$$

with standard errors clustered on the merged event windows. The skew coefficient is positive, of similar size on both underlyings, and its cluster-bootstrap interval excludes zero on both; it survives false-discovery-rate control; it survives a strict-lag variant in which the predictor is dated the day before a target window that does not open until three days out, so four days separate covariate from outcome; and it survives leaving out any one of the contributing clusters. The volatility index's own coefficient in the same fit is null on both underlyings and the two regressors are close to orthogonal — so this is not the skew re-expressing the level (Figure 5.2).

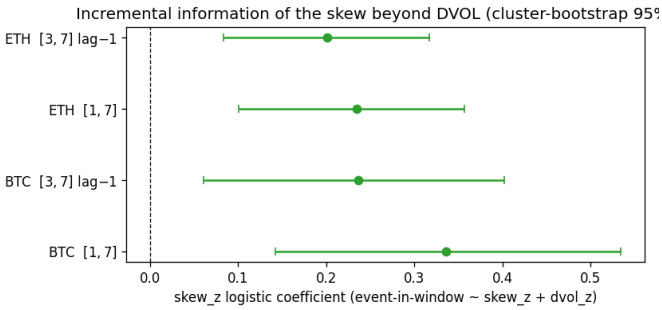

Figure 5.2: The load-bearing positive result of the Phase 2 gate: the incremental-information coefficient of the standardised twenty-five-delta skew in the logistic model of Equation (5.5), which already controls for the smirk-blind volatility index, per underlying and for the strictly-lagged variant, with cluster-bootstrap 95% intervals. All four exclude zero. What the figure does not show — and what the detector could not establish — is any edge over that same index as an alarm, or any evidence of anticipation. The two claims are separate, and only this one survived.

The Phase 2 gate verdict is accordingly a qualified, weak "go", and it is now precise about what it is a go for. The skew carries genuine incremental information about near-future stress over and above a smirk-blind volatility index, so it earns inclusion in a composite early-warning signal for the V2 design as a *continuous* covariate fused with other inputs. It does not earn standalone deployment as a thresholded trigger: as an alarm it is statistically indistinguishable from a trivial volatility threshold, and there is no evidence it anticipates. Any live use must consume the graded score in a fused model rather than a fixed cutpoint, subject to three hardening conditions recorded

before that could happen — a graded probabilistic mapping in place of the binary detector, an independent per-strike chain to ground-truth the reconstructed wings, and a larger effective event count than thirteen informative clusters.

The other outcome of this workstream is not a statistic. The engine works, it reproduces from four commands, and it runs for nothing: an entire per-strike crypto implied-volatility history, reconstructed daily from 2020 to 2026, on a public endpoint plus a root-find. That the answer it returned was mostly negative does not make the instrument less useful, and it is deliverable D2 in Table 2, delivered complete.

5.3 What Phase 2 established

RQ1 asked whether peg deviation can be decomposed into interpretable components. The answer is yes, with two qualifications that are themselves results. The decomposition is portable: the same estimator ran across a shared debt pool, an over-collateralized debt position, a fractional-algorithmic design and a perpetuals venue, and on sETH it attributes an order of magnitude more variance to the collateral channel than to the trading venue — an actionable statement a single tracking-error number cannot make. The first qualification is that the components must be mapped onto each mechanism explicitly, and for two of the six protocols one of them does not exist. The second is that the *conditional* decomposition mostly fails: on half the panels of Table 5.1 the canonical structural drivers explain under two percent of the variation, and where they do bind the binding covariate is the same one, exactly the channel Equation (5.3) describes. Attribution across components is available; explanation from market aggregates is not, unless the covariate is the mechanism itself, as the stETH withdrawal queue is.

RQ3 asked whether a depeg is detectable before it happens. Phase 2 answers the options-market half, and the answer is a heavily qualified no with one durable positive component. The twenty-five-delta skew is not a usable standalone alarm: on sixty events and thirteen informative clusters, a thresholded skew is statistically indistinguishable from a trivial volatility threshold — worse on Bitcoin, better on Ethereum, separable from neither — and every precision interval spans zero. The claim that it anticipates stress by three to four days was withdrawn: the statistic that produced it could not have returned anything else, and a direct permutation test does not reject. What survives is narrower. Entered as a graded covariate alongside the volatility index, the skew carries incremental information about stress in the next one to seven days that is positive on both underlyings, robust to a four-day predictor gap, and stable to dropping any single event cluster. That is a component of a composite signal, not a circuit breaker.

Two things carry forward beyond either research question. The first is that both halves of this chapter reached their conclusion by auditing the author's own work rather than defending it — a network bug that faked completeness, an inversion bracket that manufactured true positives, a lead-time statistic that measured its own search window, and a duplicated column in a variance decomposition. Each was found after a result had already been written down, and each changed it. The second is the negative result of Table 5.1, which recurs across five unrelated protocols and is the first of the two large cross-cutting negatives in this report. The second is in Chapter 6, it concerns machine learning, and it points the same way.

Chapter 6

Phase 3: Advanced Research

Phase 3 applied four largely independent toolkits to the dataset built in Phases 1 and 2: econometrics, machine learning, reinforcement learning and graph theory. Each answers a different question. The econometrics of Section 6.1 answers **RQ2**, on the direction of causality inside a peg mechanism and on whether liquidation cascades are measurably self-sustaining. Sections 6.2 and 6.3 are the on-chain half of **RQ3**, on whether a depeg is detectable in advance at a false-alarm rate an operator would accept; the options-implied half was settled in Chapter 5. Section 6.4 is the whole of **RQ4**. **RQ1** is not re-opened: the three-component decomposition of Equation (2.6) enters as an input, supplying the series the vector autoregressions are estimated on.

The apparatus was reviewed in Chapter 2.6 and is not re-derived; what this chapter adds is the instantiation. Estimation conventions, hyperparameter inventories, per-asset result matrices and diagnostics are in the sub-report corpus (Appendix A.2). Each workstream carries its own decision rule, fixed before the results were seen, and is reported against that rule rather than against its own internal metrics. Three of the four produced a headline negative result, reported at the same length as the positive one. The second half of Phase 3 in the internal specification — designing and prototyping Derivalink's V2 protocol — was not reached within the sixteen-week window.

6.1 Econometrics and causality

Every level series first passed a paired stationarity screen — augmented Dickey–Fuller (Dickey and Fuller, 1979) and KPSS (Kwiatkowski et al., 1992) — and a series the two tests disagreed on was reported as inconclusive rather than pushed into a verdict it could not support. Standard errors are heteroskedasticity- and autocorrelation-consistent throughout (Newey and West, 1987), with HC3 (MacKinnon and White, 1985) where only heteroskedasticity was at issue; the residual diagnostics recorded in the sub-report corpus (Appendix A.2) are why this is not boilerplate. Because a study of this shape tests dozens of directed pairs at once, significance is read on Benjamini–Hochberg adjusted q-values (Benjamini and Hochberg, 1995) over an explicitly declared family.

Cointegration. Across the dataset, 31 of 39 tested pairs cointegrate at the 5% level: even though each series wanders individually, they move together in the long run, as a healthy peg mechanism should. The eight exceptions all account for themselves: three are sUSD against outside stablecoins, which Synthetix never arbitrages directly because sUSD sits inside a shared debt pool, so the absence of a long-run relation is the mechanism speaking rather than the test failing; four are Mirror and equity synthetics with thin off-chain arbitrage pathways; and one is a deliberate falsification check, LUSD against ETH — a dollar stablecoin correctly failing to cointegrate with a volatile asset, included so the procedure can be shown to return a negative. Two caveats belong with that headline: the 39 are Engle–Granger pairs across the whole harvest rather than 39 clean synthetic-versus-own-underlying pairs, and the residual-based procedure establishes that a long-run relation exists, not what enforces it.

The Johansen system run (Johansen, 1991) is where the order-of-integration warning of Section 2.6 became a concrete correction rather than a caveat. DAI's $(\sigma_C, \sigma_S, \sigma_P)$ system on $n = 2173$ daily windows returns trace statistics of 639.6, 241.9 and 111.0 against critical values 29.8/15.5/3.84 — a cointegrating rank equal to the full dimension of the system. That was *not* reported as a finding: augmented Dickey–Fuller rejects the unit root on all three series at $p < 10^{-3}$, so all three are I(0), and a rank equal to the system dimension is the mechanically expected output rather than evidence of a long-run equilibrium. The prescribed replacement is a vector autoregression in levels with HAC standard errors. On the same apparatus sETH's system *fails* to reject at every rank ($28.05 < 29.80$; $9.46 < 15.49$; $1.22 < 3.84$), a genuine contrast with DAI rather than an artefact. One data-quality defect surfaced here and is disclosed rather than quietly dropped, because it changes what one of the report's own inputs means: in sUSD's three-component system σ_S and σ_P are exact duplicate

columns, traced to an upstream bug in the decomposition step, and the fit was degraded to a bivariate system rather than reporting a three-component result that does not exist.

Self-excitation. A Hawkes comparison between DAI's Black Thursday cascade (March 2020) and Iron Finance's June 2021 collapse puts a number on a difference the qualitative narrative only suggested. DAI's branching ratio is 0.756 — sub-critical, each liquidation triggering on average fewer than one further liquidation, so the cascade is self-limiting. Iron Finance's is 0.968 — near-critical, a cascade one step from self-sustaining — fitted on 9,231 redemption events over the 72-hour collapse, all twelve multi-start optimisations converging to the same parameters. The calm-period baseline of 0.387 matters as much as the two headline numbers, because it shows the estimator is measuring a regime and not a constant of the protocol. The mechanism behind the gap is structural: DAI's ratio is consistent with an over-collateralized, capped liquidation mechanism in which each liquidation extinguishes a bounded amount of debt, against IRON's uncapped reflexive minting, which had no such brake (Saengchote, 2021). One comparability warning travels with those numbers: Iron Finance's fit used the kernel $\alpha\beta e^{-\beta\Delta}$, where $n = \alpha$ outright, and DAI's the kernel $\alpha e^{-\beta\Delta}$, where $n = \alpha/\beta$. They are placed side by side only because both were restated in the convention of Equation (2.10) first; comparing raw α values across parametrisations gives a meaningless answer. A fit attempted on FRAX's eleven events did not converge at all, which is the sample-size limit of Section 2.6 in practice.

A bivariate test between DAI and LUSD found no genuine cross-excitation: over the only window in which the two liquidation streams overlap, both cross terms come back at effectively zero (5.3×10^{-7} and 2.4×10^{-5}) against self-excitation ratios of 0.834 and 0.645. Both protocols reacted to the same external shock; neither triggered the other. That null is reportable only because of the failure that preceded it. With naive starting values, all five multi-starts converged to a degenerate no-cross-excitation optimum whose likelihood was provably worse than at the known true parameters on simulated data built with deliberate cross-excitation — the optimiser could not detect the effect the test claimed to be testing for. Warm-starting each dimension's self-terms from an independent univariate fit fixed it, and the corrected result was re-run from five very different starting guesses, all converging to the same optimum. A fitted zero that has not survived that test is not evidence of anything.

Direction. The finding is that the same direction recurs across protocols analysed independently, by separate pipelines: collateral-side and peg-side shocks lead synthetic-side deviation, and the reverse direction is weaker or absent. DAI's $\sigma_{\text{Peg}} \to \sigma_{\text{Synth}}$ link at $p = 7.2 \times 10^{-30}$ is the strongest Granger result anywhere in the corpus, and the same direction reappears in LUSD (2.9×10^{-10}), sETH (2.3×10^{-6}), stETH (2×10^{-6}) and crvUSD. FRAX supplies the cleanest illustration, because it separates significance from direction: both directions between the peg and its share token are significant by any threshold, but the peg-to-share direction reaches $p_{\text{FDR}} = 1.6 \times 10^{-204}$ while the share-to-peg direction — the one that would indicate the reflexive feedback that destroyed Iron Finance — tops out at 1.1×10^{-45}. That asymmetry is the survivor's signature, and the exact inverse of the casualty's. None of this is causality in the interventional sense, and the chapter does not claim it is: Granger causality is predictive precedence, and where a common external shock hits both components within one sampling interval, precedence can be an artefact of which component is measured faster. The defence is replication across protocols with different mechanisms and different pipelines, which is why the pattern is reported rather than six separate results.

Tail dependence. Three readings come out of the copula fits. sETH against ETH is the control case — Kendall $\hat{\tau} = 0.954$ with lower and upper tail coefficients of 0.955 and 0.946, symmetric and near-unity, which is what a working oracle-tracked synthetic should look like. stETH's discount change against the ETH return is the opposite regime, weak association with no meaningful asymmetry. And DAI's stablecoin pairs are asymmetric *upward*, $\hat{\lambda}_U > \hat{\lambda}_L$ in both cases: the complex rallies together more reliably than it breaks together, a result that survives essentially unchanged on GJR-GARCH-standardised residuals (Patton, 2006). A deliberately retained small-sample failure is reported alongside it — a UST–LUNA coefficient of 0.33 resting on *two* realised co-exceedances, with a bootstrap interval of roughly [0, 0.83].

6.2 Machine learning

Three tiers were tested against a pre-registered early-warning task under identical walk-forward cross-validation: supervised classifiers (logistic regression, Random Forest (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017)), unsupervised detectors (Isolation Forest (Liu et al., 2008), an autoencoder (Hinton and Salakhutdinov, 2006; Sakurada and Yairi, 2014), a one-class SVM (Schölkopf et al., 2001)), and two recurrent architectures (Hochreiter and Schmidhuber, 1997; Cho et al., 2014). The label is the forward exceedance

$$d(t) = \frac{P_t - p^{\text{peg}}}{p^{\text{peg}}} \times 10^4, \qquad y_\tau(t) = \mathbf{1}\left[\max_{s\in(t,\,t+H]} |d(s)| > \tau\right], \qquad H = 24\,\text{h}, \tag{6.1}$$

with every verdict rendered at a single declared gate, $\tau^\star = 25$ bp. Four properties were fixed before any model was fitted: the window is strictly future, excluding t itself; the final twenty-four rows of every series carry a null and never a zero; labels are computed on hourly closes so a one-minute wick cannot flip a target; and the thresholds are exogenous, reusing the band convention of the existing USDC sub-report and the Chainlink oracle's own deviation trigger. That is the guard against threshold mining, and it is why the gate is 25 bp even though the results at 10 bp are more flattering to the models.

The panel is six assets on nineteen to twenty causal features each, cut into expanding walk-forward folds with a twenty-four-hour purge — not a safety margin but an exact guard equal to the label horizon — plus a twenty-four-hour embargo. The origin expands rather than slides deliberately: a sliding window would eventually push Terra, FTX and SVB out of training, and on a problem whose effective sample is of order fifteen episodes, discarding the three most informative ones is indefensible. The result is 47 folds with a common test start chosen so Terra falls in the first test block and FTX and SVB in the second; both major crises are out-of-sample. The full design is drawn fold by fold in the sub-report corpus (Appendix A.2), and it also exposes the ceiling: USDC's test blocks from May 2024 onward contain *zero* positive hours at 25 bp, so the asset's event recall ultimately rests on fifteen independent out-of-sample episodes rather than on its thirty thousand pooled rows.

The baseline that won. Against the learned models stands a persistence baseline that simply scores the present, $|d(t)|$. It has no fitted parameters, consumes no training data, and is the hardest thing in this chapter to beat — peg deviation is strongly autocorrelated, so if $|d(t)| = 30$ bp the probability of a 25 bp exceedance in the next twenty-four hours is close to one without any model at all. Part of that advantage is tautological, since an episode is *defined* by $|d| > 25$ bp and persistence alerts exactly when $|d| \geq 25$ bp, which is also why its median lead time is systematically zero. The correction is the *onset slice*: restricting evaluation to rows still inside the band, which turns nowcasting into forecasting. On USDC at the gate, persistence falls from 0.303 to 0.125 when the tautology is removed — and XGBoost falls from 0.075 to 0.036. The advantage is more than halved; it does not vanish.

The single clearest negative result of the whole internship came from this tier: the persistence baseline beat every one of five supervised model families on every one of six assets, thirty out of thirty asset–model pairs without exception (Table 6.1). This was checked, not assumed — the same comparison was run identically for each model family as it was added, specifically to avoid the common failure mode of quietly discarding an inconvenient comparison.

Read within rows, the table contains results that make the negative verdict more credible rather than less. Random Forest beats XGBoost on five of six assets, which is not the ordering a practitioner would expect, and the two sequence models post their best relative showings exactly where the single-row models had least to work with. The only baseline-beating results anywhere in the sweep sit *outside* the declared gate — Random Forest at 10 bp on DAI, 0.789 against persistence's 0.597 — and are reported as information, not as a changed verdict, because the gate was declared in advance. One honest limitation belongs here: the thirty comparisons rest on point estimates, with no formal test of equal predictive accuracy (Diebold and Mariano, 1995). The

Asset	n	pos.	XGB	RF	LGBM	LSTM	GRU	pers.
USDC	30,720	495	0.075	0.065	0.037	0.023	0.017	**0.303**
USDT	36,202	796	0.120	0.156	0.059	0.171	0.183	**0.608**
DAI	30,648	967	0.037	0.063	0.046	0.098	0.028	**0.328**
LUSD*	25,136	14,936	0.973	0.980	0.970	0.821	0.853	**0.982**
crvUSD*	17,544	3,675	0.933	0.955	0.860	0.870	0.948	**0.956**
stETH*	36,202	7,906	0.948	0.965	0.945	0.858	0.933	**0.974**

Table 6.1: Pooled out-of-sample PR-AUC at the 25 bp gate, test blocks only. Thirty model–asset pairs, zero wins for any learned model. *The near-saturated values for LUSD, crvUSD and stETH are a base-rate artefact, not evidence of learning: their test blocks carry 59.4%, 20.9% and 21.8% positive rows, and at those prevalences almost any monotone score ranks well. For those three the discriminating comparison is the episode level, not PR-AUC.

verdict is safe because the gaps are large and consistent in one direction, not because they were tested.

At the episode level the picture is sharper. Persistence reaches 100% episodic recall on five of six assets, usually at the lowest cost — on USDT it catches all thirteen covered episodes at 0.04 false alerts per thirty days. But its lead time is structurally zero, and where genuine anticipation appears it comes from the models: XGBoost on USDC at −48 h, GRU on USDT at −26 h, and the tree/GRU trio on crvUSD at −47 to −48 h while holding 95.7–100% recall. The secondary regression head points the same way, with Spearman correlation between predicted and realised forward amplitude positive on every one of the 47 folds. There *is* forward information in these features; what the models cannot do is convert it into a better ranking of threshold crossings than the current deviation already provides. Feature attribution says why: on the 25 bp USDC head the current deviation enters only seventh, while rolling and cross-sectional transformations of that same quantity take the first four places. The trees are, to a large extent, re-deriving the baseline they cannot beat.

Anomaly detection fared better. Isolation Forest clears its own pre-registered bar on four of six assets, at episodic recalls of 100% on USDT, USDC and DAI and 93.9% on crvUSD, all inside the false-alert budget; it degrades to 74.1% on stETH and 55.6% on LUSD. USDC is the only configuration in the workstream where a model beats a baseline on both axes at once, catching all thirty-three covered episodes while spending 2.5 times fewer false alerts; on crvUSD it wins by thirty-nine recall points *and* anticipates by a median of −37 hours against the baseline's zero — the only genuine anticipation measured anywhere in the module, and mechanistically explicable, since crvUSD drifts below par progressively as its PegKeepers saturate where USDC breaks suddenly on exogenous news. Two qualifications keep this honest: at the looser 50 bp threshold the advantage evaporates for all four dollar pegs, so the edge is specifically on *subtle* stress rather than large ruptures; and the autoencoder's tempting 0.53 false-alert rate on stETH is not efficiency but the arithmetic consequence of a rarely-crossed threshold that lets seventy percent of episodes through.

The mechanism behind the four-of-six boundary is the one generalisation this workstream supports. Ordering the six assets by base rate — USDT 3.8%, USDC 4.5%, DAI 7.6%, crvUSD 18.4%, stETH 30.1%, LUSD 66.5% — reproduces the ordering of Isolation Forest's recall exactly (100, 100, 100, 93.9, 74.1, 55.6%), monotonically and without exception, with the boundary falling between 18% and 30%. Anomaly detection assumes abnormal is rare; when a peg is structurally biased by design, the assumption is false and the method degrades in the way its own premise predicts. That is a usable design rule: check the base rate before choosing the detector.

A late addition — a sentiment feature built from a public financial-news dataset scored with a crypto-specific language model (Devlin et al., 2019; Araci, 2019) — more than doubled the best supervised model's predictive power on USDC, +107% for XGBoost (0.0749 → 0.1555) and +121% for the GRU. That is a real and substantial effect, and even so the improved score still did not beat persistence's 0.2953, so the project-wide verdict stood. Two caveats travel with it at the same length as the gain. The corpus of 23,301 articles ends in May 2025 and the number of *relevant* articles falls off a cliff away from the majors — 319 for USDC but 4 for crvUSD, all four predating its own

launch — so the feature is a USDC-and-USDT result and was tested as one. And an independent check of the scorer against the source corpus's own crowd labels gives an agreement rate of 36.3% against a chance-level baseline of 36.7%, statistically indistinguishable from an independent draw. The ablation, not that sanity check, is what carries the result, and both are reported because a reader is entitled to both.

One observation constrains everything after it. On the flagship SVB episode — USDC's 25 bp breach from 2023-03-11 with a trough of −1,373 bp (Federal Deposit Insurance Corporation, 2023; Circle Internet Financial, 2023) — the first Isolation Forest alert fires two hours ahead, and XGBoost's operating point is crossed in the same hour. Neither sees it days in advance, and that is the expected result rather than a disappointment: nothing in these on-chain features moves before the news of a bank failure propagates. It is a concrete statement of the limits of on-chain-only early warning, and the reason Chapter 7 argues for a composite signal.

6.3 A reinforcement-learning monitoring pilot

A separate experiment asked whether a reinforcement-learning agent could learn to raise a graduated alert level ahead of real stress episodes, rather than classify a fixed horizon. A deep Q-network (Mnih et al., 2015) was trained against a purpose-built replay environment on USDC, with a pre-registered gate fixed before any result was seen.

The internal specification asks for an agent whose actions are real protocol interventions pushed to Derivalink's V2 contracts. That is not executable, for the plain reason that no V2 contract exists. The action space was therefore substituted, and the substitution is declared rather than absorbed: the agent emits an alert level on four levels, re-bucketing the same exogenous thresholds the label already uses, so no new threshold is invented anywhere in the pilot. The algorithm choice was also revised before any training ran, from PPO (Schulman et al., 2017) to a deep Q-network, because USDC's usable history is tens of thousands of hourly rows rather than millions and the environment is a deterministic historical replay rather than a sampleable simulator.

The gate was written into the evaluation script before that script was ever run: the agent must catch at least as many of the holdout fold's eight real episodes as *every* baseline, and must not have a worse false-alert rate than the hysteresis baseline. On the FTX-and-SVB holdout that resolves to 8/8 episodes at ≤ 0.00 false alerts per thirty days, because the naive current-deviation baseline catches all eight at zero cost.

The agent failed that gate three times, in three different ways (Figure 6.1). V1, on the original reward, caught **zero of eight** — tied with the weakest baseline and clearly beaten by the naive rule. V2 reweighted the training reward to penalize missed alerts roughly seventeen and a half times more heavily, raising λ_{under} from 2.0 to 35.074, which lifted recall to **three of eight** at 0.83 false alerts per thirty days — still a clear no-go. V3 pooled training data across four related stablecoins, to test whether the problem was simply too little data for one asset: it did reach **eight of eight**, with a genuine **forty-six-hour median lead time**, a real result — but at 25.70 **false alerts per thirty days** against a declared budget of two. This, too, is a no-go, reported as one, with no production deployment attempted.

Each failure was diagnosed before the next run, from design and data rather than by retuning until the number moved. V1's silence has an arithmetic explanation: a 2:1 asymmetry between under- and over-alerting is far too weak against a base rate below three percent, so a policy that stays at Normal rarely pays the penalty of 2 and avoids paying 1 on an overwhelming majority of calm hours — "stay calm" dominates in expectation. That is the reinforcement-learning analogue of training a classifier on severe imbalance with an imbalance-blind objective. The fix was itself pre-registered, tied to the label distribution alone rather than to any observed policy behaviour:

$$\lambda_{\text{under}} = \frac{\lambda_{\text{over}}}{p}, \qquad p = \Pr(L_t \geq 1) \ \text{ on training folds only,} \tag{6.2}$$

which on $p = 0.028511$ gives $\lambda_{\text{under}} = 35.074$. The holdout fold's own base rate of roughly forty percent was deliberately *not* used to set it.

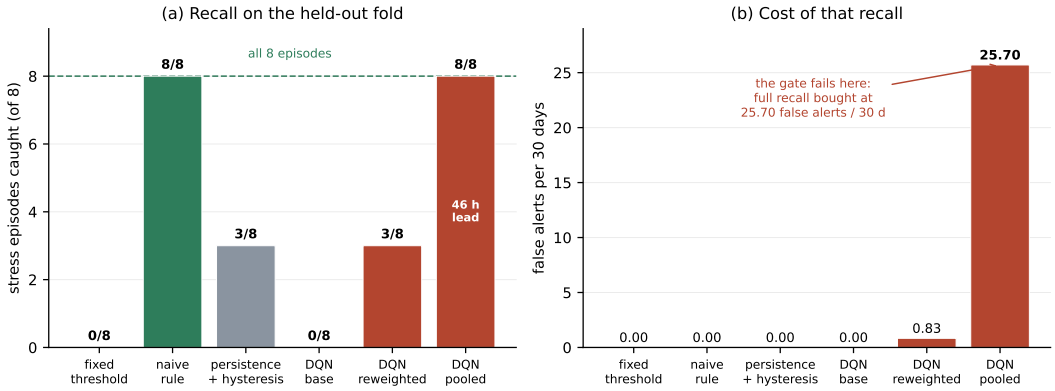

Figure 6.1: The reinforcement-learning pilot against its own pre-registered gate. Left: stress episodes caught on the held-out fold, for three baselines and three successive DQN runs. Right: what that recall costs in false alerts per thirty days. The base agent catches nothing; reweighting the reward buys three of eight at 0.83 false alerts per month; pooling four assets reaches full recall with a genuine 46-hour median lead, but at 25.70 false alerts per month. All three runs are NO-GO, and the third fails on the clause the gate was written to protect.

V3 changed exactly one lever — the training population, to a pooled table of 115,114 rows over four dollar pegs — leaving reward weights, budget and holdout fold identical, so all three runs are directly comparable. And it fails the gate on the clause the gate exists to protect. The per-asset alert-worthy rates in the pooled population are USDC 2.85%, DAI 17.47%, USDT 22.28% and crvUSD 41.41%, for a pooled rate of 19.35%; by Equation (6.2) the correct penalty for that mixture is $1/0.1935 \approx 5.17$, against the 35.07 actually used — roughly 6.8 times too aggressive. The shared policy learned an escalation habit suited to DAI's and crvUSD's far more frequent stress regimes and applied it wholesale to USDC's genuinely rare one. That was verified from the base rates rather than guessed, and it was deliberately *not* corrected and re-run, for the same pre-registration reason that governed V1: retuning after seeing an unfavourable pre-registered result would destroy the point of having pre-registered it. The next experiment is recorded rather than silently attempted.

What the three runs establish together is narrower and more useful than any of them alone. The recall/false-alert tradeoff moved; it never disappeared. Pooling solved exactly the problem it targeted and revealed a different one that was previously invisible: a single scalar reward weight cannot simultaneously fit assets whose stress base rates differ by more than an order of magnitude, 2.85% against 41.41%. That is a concrete, transferable design constraint on any multi-asset alerting agent, and it is the pilot's real deliverable.

6.4 Graph theory: a systemic-risk model tested against three collapses

The final workstream built a graph-theoretic model of systemic risk in direct response to the specification's own question: do protocols that are structurally more central actually suffer more when shocked? A weighted, directed graph was built from real on-chain data — nodes for protocols, synthetic assets, collateral positions, oracles and DEX pools; edges for collateralization weighted by USD value, oracle provision weighted by update frequency, and pool hosting weighted by liquidity, every edge pointing into the asset node along the real economic dependency direction and expiring if its source row is more than ninety days stale. Two of the five edge types the specification enumerates were not built — one requires an event table with no populated partition anywhere on disk, the other a synthetic backed by another synthetic, which has no confirmed instance in this dataset — and those gaps are reported as gaps in the specification's own formalisation rather than patched by inventing edge types outside the agreed schema. Because oracle weight is a count of updates while collateral and pool weights are dollars, the three are not commensurable, so three projections of the same graph exist, one per question. Centrality, community detection (Blondel

et al., 2008; Traag et al., 2019; von Luxburg, 2007) and a from-scratch implementation of DebtRank (Battiston et al., 2012) were computed with NetworkX (Hagberg et al., 2008).

The node value v_i of Equation (2.12) is the one modelling choice the source literature leaves open: interbank studies use balance-sheet size, and on-chain there is no balance sheet, so the proxy used here is total incident edge weight, disclosed as a documented simplification rather than presented as equivalent. The implementation was verified before it was trusted: three graphs with hand-computable answers — a two-node full dependency, a three-node chain and a three-node *cycle* — return exactly 0.5, 0.75 and 0.6667, the third being the one that matters, since it proves the state machine terminates and does not double-count a shock that loops back to its origin.

Three historical replays. The central test method was to build the graph as it stood the evening before a real collapse, calibrate a shock to the real observed magnitude of that collapse — not an arbitrary worst case — propagate it, and compare the prediction to what genuinely happened next.

Event	Collateral regime	Nodes / edges	Predicted	Predicted / observed	Bias
DAI, March 2023	Full collateral, 31 diversified legs	35 / 33	−415 bp	24.1–37.3%	Undershoot 2.7–4.1×
UST, May 2022	Algorithmic, 1 undiversified edge	3 / 1	100.0% distress	130.7%	Overshoot 1.3×
Iron Finance, June 2021	Partial collateral, 2 legs	18 / 16	8.8% distress	21.3–36.8%	Undershoot 2.7–4.7×

Table 6.2: The three historical replays: model prediction as a percentage of the real observed loss. The ranges come from there being more than one defensible observed reference — for DAI, its oracle trough of −1112 bp against its Curve-pool trough of −1717 bp; for Iron Finance, an intraday trough of 41.4% against a settled loss of 24.0%.

Every replay is directionally correct: in each case the model flags exactly the right node as seriously at risk and lands within the right order of magnitude (Table 6.2). The two undershoots land in almost the same range despite very different collateral structures — thirty-one diversified legs against two — which suggests the roughly three-to-four-times understatement is a general property of a static, single-round propagation model facing a real, multi-day collapse with dynamics it cannot see, rather than something specific to any one protocol. UST's overshoot is explained cleanly by a single undiversified collateral edge with nothing to dilute a near-total collapse: the model predicts total loss, where the real UST kept 23.5 cents at its trough because of the Luna Foundation Guard's attempted reserve deployment and the simple fact that a death spiral still takes days.

The Iron Finance replay revealed a modelling subtlety invisible with fewer protocols in scope. Its two DEX pools carry combined liquidity of $1.36 billion against total collateral of $749 million — pool depth nearly double the backing — so with both in the value projection the shock is silently diluted across quantities measuring different things. Cutting the model to collateral edges only raises the prediction from 8.8% to 24.9% distress, which is 103.6% of the real settled loss. That is neither a coincidence nor a fitted parameter: the algorithmic leg's share of backing, $186.4/749.2 \approx 24.9\%$, is by construction close to the share implied by the collateral ratio itself, $100\% - 75.15\% \approx 24.85\%$, and IRON settling near its effective collateral ratio is exactly the mechanism the forensic reconstruction of Chapter 4 describes. The convention was deliberately left unchanged in the propagation module — it remains correct for DAI and UST, and the ambiguity needs a generic resolution rather than a per-protocol patch — and the collateral-only cut is reported alongside the primary result, never substituted for it. What licenses all three replays is a check on that same snapshot: the collateral ratio reconstructed independently from the graph, 75.12%, sits within 0.02 percentage points of the 75.15% recorded on-chain at the same instant.

Generalizing beyond three protocols. A final pass built one graph spanning every tracked asset with real edge data on disk — fifty-nine assets, discovered automatically rather than named by

hand, with zero changes to any algorithm. The result runs against the question's original framing: the real graph is overwhelmingly a disconnected forest of isolated per-protocol stars, not a densely interconnected system.

Two different quantities describe that sparsity and must not be confused. The graph has 110 nodes and 39 edges, and separates into **72 connected components**, of which 48 are single nodes with no on-chain edge to anything at all. Community detection then subdivides some of those further, and Louvain, Leiden and spectral clustering independently agree on **93 communities** across the same 110 nodes. Seventy-two is the number of pieces the graph falls into; ninety-three is the number of groups the modularity objective prefers within them. Only three components are non-trivial: the largest holds nine nodes around two bridged variants of Terra's UST, and the next holds eight around DAI. Every other protocol in scope — Iron Finance, FRAX, eleven Mirror synthetics, ten Synthetix equity synthetics, Ethena, Resolv, GLP — is its own fully isolated star.

Genuine on-chain interconnection between distinct synthetic-asset protocols — as opposed to a shared external trigger, or shared holders, which this schema cannot see at all — is therefore rare in this dataset. The specification's question can be answered *within* a single protocol's local graph, where the answer across all three replays is a clear yes; not across protocols, because there is usually no real edge for such a comparison to run along. A concrete caution surfaced in the same snapshot: a bridged, wrapped form of the long-dead UST token ties DAI for the single highest-connected node in the entire fifty-nine-asset graph, purely because several exchanges still quote a residual market for it years after its peg collapsed. Structural centrality, measured this way, tracks how many venues still trade an asset — a different question from whether that asset is economically healthy.

6.5 Synthesis: what Phase 3 established

Read together, the four workstreams point in the same direction. Simple, transparent, real-data-grounded models — a persistence baseline, an isolation forest, a graph-propagation model calibrated to one real shock — consistently do a genuinely useful, auditable job of flagging that something is wrong in roughly the right place and roughly the right order of magnitude. More sophisticated models built on the same data consistently failed to add value once judged against a disciplined, pre-registered standard rather than their own internal metrics. This is not an argument against sophistication in general; it is a specific, repeated, empirical finding about this problem on this dataset, worth taking at face value precisely because it was not the result any of these experiments set out to find.

What Phase 3 does not establish is worth stating with the same precision. None of the thirty model comparisons carries a formal test of equal predictive accuracy; the graph workstream has no confidence intervals anywhere, and three replays are not enough to generalise a bias statistically; the reinforcement-learning verdict rests on eight episodes in one fold; and the multi-protocol graph can see only the mechanical channel of contagion, never the behavioural one — shared holders, correlated panic, a flight to quality — which Chapter 2 showed carried at least as much of the March 2023 shock as the mechanical channel did.

Chapter 7

Results and Discussion

Chapters 4–6 each judged one workstream against its own pre-registered bar. This chapter reads them across each other: the patterns below are visible only because roughly twenty heterogeneous protocols went through one methodology, and Section 7.2 turns them into the design decisions Derivalink actually faces.

One construction note. An adversarial audit run after this report's first draft found that the apparent lead of the smirk over stress events was bounded by the search window that measured it, and the anticipation claim was withdrawn. The corrected statement is weaker than the one it replaces and is used throughout, because a result that survives its own author's audit is the only kind worth putting in front of a design team.

7.1 What sixteen weeks of real depeg data say, together

Collateral structure predicts the direction of a systemic-risk model's error, even if not its size. Writing ρ for the ratio of the model's DebtRank score to the fractional loss the market actually settled at,

$$\rho \;=\; \frac{R}{D_{\text{obs}}}, \qquad D_{\text{obs}} \;=\; \frac{P_{\text{pre}} - P_{\text{trough}}}{P_{\text{pre}}}, \tag{7.1}$$

so $\rho = 1$ lands exactly on the realised damage, the three replays of Table 6.2 give $\rho = 0.373$ and 0.241 for DAI, $\rho = 1.307$ for UST and $\rho = 0.213$ and 0.368 for Iron Finance. The model undershoots diversified or partially diversified collateral and overshoots a single undiversified edge. The two undershoots land in almost the same band despite structures with nothing in common — thirty-one diversified legs against two, one of them endogenous — which makes the three-to-four-fold understatement a property of the model class rather than of any one protocol: DebtRank propagates a continuous distress variable in one static round (Battiston et al., 2012; Eisenberg and Noe, 2001), with no panic, no multi-day redemption dynamics and no circuit breaker. What is new is that the bias's *sign* is predictable from a structural feature the designer controls, so a composite early-warning signal should expect the asymmetry rather than treat a systemic-risk score as protocol-agnostically calibrated.

Simple, transparent baselines were hard to beat throughout, and that fact is itself the finding. Persistence beat five supervised families on every asset tested, thirty comparisons for thirty losses; a naive current-deviation rule beat a purpose-built reinforcement-learning agent on its own pre-registered gate; the smirk carries incremental information beyond a volatility-index baseline but does not anticipate and establishes no edge as a detector. The margins are wide, not marginal, and the ceiling is a property of the label rather than of any model — USDC has 73 lifetime episodes at the 25 bp threshold, roughly fifteen of them independent out of sample. Sophistication is not thereby ruled out: the sentiment feature and the pooled agent both improved measurably on their simpler counterparts, and each had to clear the same bar as the simple baseline, and most did not.

The reinforcement-learning result adds a mechanism worth stating on its own, because it generalises past this dataset. The reward silently encodes a prevalence, since calibrating the missed-alert penalty ties it to the base rate of alert-worthy states in training. On USDC's $p = 0.028511$ that raised the penalty from 2.0 to 35.074 and recall from zero of eight episodes to three; pooling four assets reached eight of eight with a genuine 46-hour median lead, but the pooled population has $p' = 0.1935$, so the reward inherited from USDC is $p'/p \approx 6.8$ times too aggressive for the mixture it now faces, and false alerts rise to 25.70 per thirty days against 0.00 for the naive rule. That is not a tuning detail; it is why the third run fails the exact clause the gate was written to protect, and it is a transferable constraint on any multi-asset alerting agent.

Graph centrality, oracle presence and pool liquidity are not interchangeable proxies for economic health, and conflating them produces real, quantifiable errors. On the multi-protocol snapshot of 110 nodes and 39 edges, a bridged, long-dead form of UST ties DAI for the highest in-degree in the entire graph, because seven decentralized exchanges still report fresh residual liquidity for a token whose peg died in May 2022 (Liu et al., 2023; Briola et al., 2023). Any influence metric (Page et al., 1999; Katz, 1953; Freeman, 1978) ranks it accordingly, and is not wrong to: it measures how many venues still quote the asset, which is not the quantity a risk score needs. The Iron Finance half is quantitatively sharper — its two pools carried \$1.36 B of combined liquidity against \$749 M of total collateral, so a convention treating both edge types alike dilutes the shock across a denominator the collateral never contained, and cutting the model to collateral edges alone moves the prediction from $R = 8.8\%$ to 24.9% against a settled loss of 24.0%, $\rho = 1.036$ in the sense of Equation (7.1).

The ambient market covariates a designer would instinctively regress tracking error on explain almost none of it; the variables that do bind are the ones the mechanism itself creates. Equation (2.5) was fitted independently in five unrelated sub-reports that never cited one another and returned essentially nothing every time. crvUSD says the same in the lead–lag direction: no stress channel Granger-leads its peg at lag one — PegKeeper (Curve Finance, 2023) debt change $p = 0.84$, soft-liquidation market count $p = 0.37$, collateral drawdown $p = 0.62$ — and none survives multiple-testing adjustment. The three exceptions share one property: the binding regressor is endogenous to the design. In Iron Finance the only significant driver is the realised volatility of TITAN, the share token the redemption leg mints; in FRAX the same channel dominates at $R^2 = 0.341$, alongside a protective -13.79 bp from cumulative algorithmic-market-operation deployment (Kazemian et al., 2020). Both are fractional designs in which the volatile leg sits structurally *inside* the peg mechanism, as the deleveraging-spiral models predict (Klages-Mundt and Minca, 2022). The third is the strongest structural regression in the entire corpus: Lido's post-withdrawal stETH discount on the withdrawal queue's finalisation latency, -1.28 bp per day at $R^2 = 0.43$ and $p \approx 1 \times 10^{-87}$. The queue is not a market variable the protocol is exposed to; it is a parameter the protocol sets.

Redemption is in fact the mechanism this corpus measures most consistently, in four independent places and in both directions. Turning it *on* collapsed stETH's discount distribution from a mean absolute deviation of 87.0 bp and a median of -42.9 bp to 6.1 bp and -4.1 bp, the mean-reversion half-life from 23.1 to 4.7 days and the worst daily discount from -636.8 to -61.3 bp. A hard redemption right at par produces LUSD's asymmetric peg by construction: a premium 74.2% of the time, with redemption intensity 7.17 times higher on below-peg days (Liquity AG, 2021). Turning it *off* does the reverse without bound — Synthetix's SIP-169 deprecation pinned a cohort of synths to a frozen redemption rate while their underlyings kept moving (Momtazi, 2021), giving a factor of about nine in root-mean-square tracking error against a physically-backed equivalent, 0.66% against 5.8% — and Iron Finance's post-mortem records the same failure as an accident rather than a policy, a positive-share-price guard reverting and freezing redemptions at the point where they were the only remaining exit (Iron Finance, 2021). This is the primary-market arbitrage channel of Lyons and Viswanath-Natraj (2023) measured from four directions, and the lever with the largest observed effect on peg quality anywhere in this corpus.

7.2 Implications for Derivalink's V2 design

The specification's V2 design questions map onto these findings without the specification work itself having been done. Two follow directly: a dynamic, volatility-responsive collateral ratio, from the Iron-Finance-versus-FRAX comparison; and a composite, multi-signal circuit breaker rather than any single metric, because neither component is safe to gate a production freeze alone. Four further questions can now be given a number rather than an intuition.

Which way the collateral ratio moves under stress is the survivor/casualty variable. Through the May-2022 contagion FRAX's ratio ratcheted *up* by 8.0 percentage points, 85.25% to

93.25%; Iron Finance's effective ratio was driven *down* by 25.7 points, 99.8% to 74.1%, while its share token inflated. The dilution differs by five orders of magnitude (Figure 7.1). The fee structure points the same way and is cheap to copy: FRAX charged 0.95% to mint against 0.45% to redeem, a +0.50-point asymmetry that taxes the loop where Iron Finance's ordering rewarded it, and the redeem-and-dump premium exceeded the redemption fee on 53.2% of collapse bars (Saengchote, 2021; Clements, 2021). A hard cap on the volatile leg's mint rate is the cheapest structural defence this corpus supports.

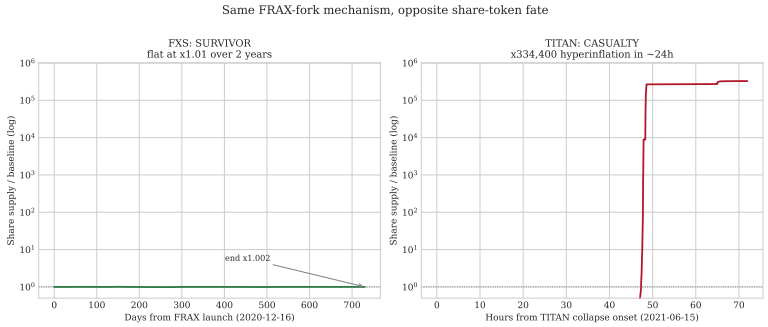

Figure 7.1: The single design decision that separated survivor from casualty. Normalised share-token supply on a log scale: FRAX's FXS is effectively flat over its whole life (1.008× peak dilution) while IRON's TITAN inflates by roughly 334,400× in about twenty-four hours. Uncapped minting of the volatile leg is the fatal flaw.

A circuit breaker must not read a single push oracle. Chainlink's update model fires on a heartbeat or a deviation threshold, whichever comes first (Chainlink Labs, 2024; Eskandari et al., 2021), so a feed is structurally blind to stress below its own trigger. On USDC, of 1,898 overlapping days five carried a centralized-exchange absolute deviation above 25 bp and the oracle stayed pinned on three of them — 60% of genuine stress days invisible to a protocol reading only the feed. During the SVB window the oracle floored at 0.88 while the exchange tape wicked to 0.81029, a gap of 697 bp, with DAI following through the Peg Stability Module (Circle Internet Financial, 2023; Federal Deposit Insurance Corporation, 2023). That pass-through was linear rather than threshold-shaped ($\beta = 0.91$, $R^2 = 0.99$), so a breaker reading oracle prints alone would have fired late on a contagion path already fully determined.

The binding constraint on an alerting component is the false-alert budget, not recall. Every policy here that reached full recall did so at a stated cost: the naive rule catches eight of eight at 0.00 false alerts per thirty days and zero lead time; Isolation Forest clears its bar on four of six assets at 0.37 to 1.27 (Liu et al., 2008); the pooled agent reaches eight of eight with a real 46-hour lead but 25.70. A V2 breaker specification should state its alert budget before its recall target, because the budget is what every candidate signal in this corpus actually failed on.

Redemption should be treated as a tunable peg parameter, not a static protocol property. The stETH queue regression gives the first term of a design curve Derivalink can set directly: a day of additional queue latency costs 1.28 bp of discount, better identified than anything the ambient-covariate specification produced. One qualification travels with it — the interrupted-time-series break at the operational queue-open date is not significant (+2.2 bp, $p = 0.45$), while a placebo test at the earlier protocol fork carries the larger break. The market priced redeemability when it became certain, not when the queue opened.

7.3 An honest inventory of what did not work

Table 7.1 lists every result in this report that did not clear its own pre-registered bar. A scattered handful of caveats across ten chapters is easy to read past; a single table is not.

Two rows carry more than the table can. The IGARCH row is the cleanest case of a plausible signal that does not separate: a GARCH(1,1) fit to hourly tracking-error changes returns $\alpha + \beta = 1.0000$ exactly for FRAX and for Iron Finance alike (Bollerslev, 1986), so a breaker keyed to

Result	What was tested	Verdict
Supervised ML (5 families)	25 bp early warning, 6 assets, walk-forward PR-AUC vs. persistence	No-go: 0 wins in 30 pairs
Sentiment / NLP features	Same task, USDC, with news features added	Real gain (+107%), still below persistence: 0.156 vs 0.295
Cross-asset pooling	One shared model over 4 assets	Helps 2 assets, hurts 2, including the reference asset
IV smirk as a detector	25Δ skew threshold vs. volatility-index and random baselines, 60 events	No established edge: F1 0.076 vs 0.111 (BTC), 0.089 vs 0.079 (ETH); paired test ties
IV smirk as an anticipator	Circular-shift permutation on the event clock	Not shown: $p = 0.186$ (BTC), 0.098 (ETH)
RL alert agent (base)	Graduated alerting vs. 3 baselines, pre-registered gate	No-go: 0 of 8 episodes
RL alert agent (reweighted)	Same gate, penalty tied to the training base rate	No-go: 3 of 8
RL alert agent (pooled)	Same gate, 4-asset pooled training	No-go on the false-alert clause: 8 of 8 at 25.70 alerts / 30 d
DebtRank (3 replays)	Predicted vs. real settled collapse magnitude	Directional only: ρ from 0.21 to 1.31
Centrality vs. suffering	Do more central protocols suffer more, 59-asset graph	Unanswerable here: a disconnected forest, no edges to compare along
Structural drivers	Equation (2.5) on five unrelated cohorts	No explanatory power: $R^2 < 1\%$ unless the design is the driver
IGARCH as a discriminant	$\alpha + \beta$ on hourly TE changes, survivor vs. casualty	Fails: FRAX and IRON both exactly 1.0000
Iron Finance warning panel	8 pre-registered signals, month before the collapse	7 of 8 silent

Table 7.1: Every pre-registered negative or partial result in this report, in one place.

volatility persistence fires on both or on neither. The Iron Finance row is a pre-registered failure, not a post-hoc one: the one signal of eight that did cross — net liquidity-provider flow, at +$200.1 M on the final pre-crash day against a calm-period mean of +$34.2 M — pointed toward *more* deposits, the opposite of what a run signal is read to look like.

Presenting this table is itself part of this report's argument. A sixteen-week internship whose purpose was to inform a real design decision produced more disciplined negative results than headline positive ones, and that ratio is a sign the evaluation standard was genuinely strict, not a sign the internship underperformed.

7.4 What this chapter establishes

The most transferable output of the sixteen weeks is a calibrated scepticism: baselines that are genuinely hard to beat, biases whose sign is now known, and a short list of design parameters — the direction of the collateral ratio under stress, the cap on the volatile leg, the openness and latency of the redemption channel, the number of independent references a breaker reads — that this corpus can put a number on. One measurement discipline underpins all of it: because the oracle, the pool and the centralized tape disagree by hundreds of basis points exactly when it matters, which series is called "the price" is a modelling decision that has to be declared (Makarov and Schoar, 2020; Petajisto, 2017). Chapter 10 states the four research-question answers formally.

Part III — Assessment and Conclusion

Chapter 8

Skills and Personal Contribution

Part II reported what this internship found. This chapter reports what it taught, under one standard: a skill is claimed here only where it produced a result in Part II that another person could reproduce from the same data. Where a technique was read about and never applied to real data, it does not appear. Table 8.1 restates the claim in a form that can be checked — each technique, the equation that defines it, and the result it produced.

The four research questions each demanded a different toolkit, which is the simplest explanation for the breadth. RQ1 required volatility modelling and a disciplined treatment of sampling frequency; RQ2 the cointegration, Granger and point-process machinery; RQ3 supervised learning, anomaly detection, reinforcement learning and options pricing, together with an evaluation standard strict enough to reject all of them; RQ4 graph theory and a from-scratch implementation of a published contagion model. Underneath all four sat a data-engineering layer that had to exist before any of them could be attempted.

Technique	**Formal statement**	**Where applied, and what it produced**
GARCH, GJR-GARCH, EGARCH; HAR	Eqs. (2.7), (2.8)	Volatility of the peg deviation and its persistence across the corpus; every fit refit from four to five seeds, convergence stability reported as evidence of a global optimum.
Engle–Granger, Johansen, VECM	Eq. (2.9)	31 of 39 tested synthetic/underlying pairs cointegrate at the 5% level (Section 6.1).
Granger causality, with FDR control	Section 2.6	Collateral-side shocks Granger-lead synthetic-side deviation, replicated across several unrelated protocols.
Clayton and Gumbel copulas	Section 2.6	Asymmetric tail dependence between a synthetic and its collateral, where a linear correlation would average the tail away.
Hawkes MLE, univariate and bivariate	Eq. (2.10)	Branching ratio 0.756 for DAI's March 2020 cascade against 0.968 for Iron Finance; a cross-excitation null between DAI and LUSD.
Supervised classifiers and recurrent networks	Eqs. (3.4), (2.11)	Thirty of thirty losses to a persistence baseline, over 47 walk-forward folds on six assets (Section 6.2).
Isolation Forest, one-class SVM, autoencoder	Section 2.6	Isolation Forest clears its own pre-registered bar on four of six assets, and fails on the two whose peg is biased by design.
Deep Q-learning with a shaped alert cost	Eq. (6.2)	Three pre-registered runs against a purpose-built environment, all three reported as no-go (Section 6.3).
DebtRank, implemented from scratch	Eq. (2.12)	Three historical replays against real settled losses (Table 6.2), after verification on hand-computed cases with known answers 0.5, 0.75 and 0.6667.
Centrality and community detection	Section 2.6	93 communities across 110 nodes, agreed independently by three algorithms.
Black-76 inversion, 25-delta risk reversal	Eqs. (2.13), (2.14)	A daily per-strike implied-volatility series reconstructed from public data at zero marginal cost (Chapter 5).

Table 8.1: Techniques claimed as acquired, each with the equation that defines it and the result in Part II that demonstrates it.

The data-engineering layer has the least visible output on that list and the largest share of

the hours: roughly 77 GiB of compressed Parquet across about 1.24 million files, 182 registered upstream sources, and a curated subset of 396,453,410 rows loaded into TimescaleDB — year by year, with the query engine capped at one thread and one gibibyte, because a single-transaction load exhausted the target virtual machine outright. The two techniques that mattered most are unglamorous: an adaptive pager that halves a block range whenever a provider rejects it as too large, then grows it back after each success; and a per-extractor resumable state file whose cursor advances only after a successful atomic write — which is what let a six-year MakerDAO backfill survive six in-flight crashes across four days and still reconcile exactly against on-chain supply.

One habit from that layer transfers and was not obvious in advance. Native sampling frequency was preserved everywhere in the raw layer, with imputation confined to a derived layer carrying explicit provenance columns, because forward-filling an hourly series onto a one-minute grid puts a zero return in fifty-nine minutes out of sixty and biases a log-return tracking error toward zero by a factor of roughly sixty — silently, with no error and no visible artefact. The corresponding estimator refuses to run on an upsampled frame rather than trusting its caller to remember.

What the table does not show is the design work around it: a walk-forward protocol of 47 folds, each fold a distinct train/validation/test triptych whose validation tail was used only for early stopping and for freezing an operating point, never for measuring reported performance, with a 48-hour purge-and-embargo gap sized to the labelling window so that no training row could carry a label computed from a price inside the test block. Designing that protocol, and holding to it when it rejected five model families in a row, is the single most transferable thing this internship produced.

8.1 Working and communication skills

Working solo on an open-ended sixteen-week mandate required a different discipline than a shorter, closely-scoped assignment: breaking an ambitious specification into a living, weekly-updated roadmap; maintaining an auditable trail of atomic commits that made it possible to reconstruct exactly what was done and why at any point (used directly to fact-check this report); and preparing for three structured weekly checkpoints with a mentor whose own time was limited, which meant learning to communicate a week's technical work — including its failures — concisely rather than at length. Writing this report and the roughly twenty technical sub-reports it draws on was a sustained exercise in a register this degree does not otherwise require: presenting a negative or partial result with the same clarity and confidence as a positive one, for a reader who needed to trust the report enough to act on it. That trail is not a bureaucratic detail; it is what makes this report checkable, and several figures in its first draft were corrected because the record and the recollection disagreed.

8.2 Individual contribution

Because this was a solo internship, every result in this report is individually attributable: roughly twenty protocol-level ETL pipelines and their sub-reports, the multi-level tracking-error decomposition, the implied-volatility smirk engine and its data-source correction, the econometrics battery, five machine-learning model families and their evaluation protocol, the reinforcement-learning pilot and its two follow-up experiments, and the graph-theoretic systemic-risk model with its three historical replays, were each designed, built, run and written up by the author, with direction and review from the company mentor at the checkpoints described above and the AI-assisted engineering workflow noted in Chapter 3. That attribution extends to the negative results and to the bugs: the extraction defects, the optimiser failure and the memory crash were the author's to introduce, find and fix, and they are reported here for the same reason the negative results are.

One skill resists tabulation: deciding when a result is good enough to report — what counts as a fair baseline, when a comparison is legitimate, when an inconvenient number has to be published rather than quietly re-run. That is what this internship developed most. The breadth above is

real and so is its cost: no one of these techniques was taken to the depth a specialist would, and whether four toolkits covered adequately beats one covered deeply is an open question rather than a settled one. Chapter 9 takes both up directly.

Chapter 9

Personal Reflection

Chapter 8 accounted for what this internship taught. This chapter asks the harder question: what it got wrong, what that cost, and what a second attempt would do differently — on the view that a reflection chapter reporting only growth is not evidence of any.

Four incidents recur below, because they are the ones that actually changed how the work was done afterwards: an extraction script that silently produced empty results and recorded them as complete; an optimiser that returned a confident zero where a real effect existed; an out-of-memory crash on a dataset that could not be scanned the obvious way; and a budget assumption wrong by two orders of magnitude. Three of the four were caught by the process before anything was published on top of them. One was caught late, after a result built on it had already been written up, and that one is worth the most space.

9.1 What went well

Building the internship around a living, honestly-updated roadmap rather than a plan written once and left untouched mattered more than expected. A sixteen-week solo mandate against an ambitious specification is the kind of project where scope quietly drifts if nothing forces a weekly written reckoning of done against planned; updating that roadmap after every task, with a dated changelog entry and an atomic commit, is directly why this report could be written at all — every number in it was checked against a specific dated record rather than reconstructed from memory. The related decision to validate every new analytical capability on a single well-understood protocol before generalizing it caught real bugs earlier than they would otherwise have surfaced, at the cost of feeling, at the time, like slower progress.

Two specific catches justify that claim rather than merely asserting it. The first is the bivariate Hawkes fit between DAI and LUSD, which returned essentially zero cross-excitation — exactly the kind of clean null that is easy to write up without checking. Before it was written up, the same estimator was run on simulated data containing a known non-zero cross-excitation, and it failed to recover it: every one of five naive starting points converged to a degenerate no-excitation optimum whose likelihood was worse than at the true generating parameters. The fix was to warm-start the bivariate fit from independent univariate fits. The null survived that fix, and only then was it reported. The second catch is a set of price-tape artefacts found by a routine integrity pass before any modelling ran on them: a decimal mismatch in DAI's exchange tape that would have registered as a 6,445 basis-point deviation for a single hour — and would, on magnitude alone, have become that asset's headline depeg episode — and a comparable artefact in stETH's series worth roughly 5,000 basis points. Neither would have raised an error. Both would have produced a confident, wrong, entirely publishable finding.

9.2 Real challenges

Three constraints shaped this internship more than any single technical decision: an 11.4 GB-RAM laptop, free-tier data access, and working entirely solo on an open-ended mandate with no peer to sanity-check an approach informally before a scheduled checkpoint. The first two each have a concrete incident attached, and the incidents are more instructive than the general statement.

The memory ceiling stopped being an abstraction when the Terra/UST replay required scanning an event table holding roughly 48 GB for the UST asset alone, with fourteen event types interleaved in the same partition files: a first unscoped scan crashed the machine outright. The fix was not more memory but a different query — scoping every scan to the exact partitions the analysis window touches, and defaulting to a streaming engine — and that is a better habit than the larger machine which would have hidden the need for it. The data constraint stopped being an abstraction when

the historical-options vendor the specification assumed, budgeted at a few hundred dollars, turned out to quote its historical tier at roughly $30,000. The workstream would have ended there. It did not, because the exchange's own public charting endpoint (Deribit, 2026) serves daily premium data for long-expired contract codes, from which a standard futures-option inversion (Black, 1976) recovers the same series at no cost. The lesson generalizes uncomfortably: the specification's assumption went unchallenged until the paid path became unaffordable, and the free path had been sitting in the exchange's public documentation the entire time.

9.3 What I would do differently

The specification's own sixteen-week, four-phase plan was ambitious enough that reaching Phase 4 in full was always going to be tight once the first three phases were each pursued to the depth this report describes. A version of this internship that scoped Phase 3 narrower from the outset — fewer workstreams, each taken further — might have left room for the V2 specification and prototype work. Whether that trade would genuinely have been better is a real open question rather than a settled one: the breadth actually pursued is what produced the cross-cutting findings of Chapter 7, which a narrower scope would not have supported.

There is a second, smaller thing, and it is the one worth stating plainly. Early in Phase 2 a promising result — the volatility smirk apparently leading Bitcoin and Ethereum stress events by five to six days — was written up and communicated to the mentor before the pipeline that produced it had been audited. It was wrong: an extraction script silently emptied a time window whenever a network drop interrupted it mid-fetch and still recorded that window as complete, so the completeness check downstream saw nothing missing. The bug was found and fixed, the extraction re-run, and the corrected, weaker result recomputed on verified-complete data. That corrected result was then itself revised a second time by an adversarial audit of the same workstream: the apparent lead turned out to be bounded by construction by the window the estimator searched, and a circular-shift permutation test found no significant anticipation ($p = 0.186$ for Bitcoin, 0.098 for Ethereum). What survives is a real but weaker claim. What I would do differently is not the fixing, which worked both times, but the ordering: the promising number went out before the audit rather than after it, and doing that twice would be a habit rather than an accident.

9.4 Growth, and a closing assessment

Coming into this internship with a background in market-making and taker-bot development on a prediction-market platform, the biggest shift was from optimizing a single well-defined trading strategy to designing and defending an evaluation standard for many competing approaches to the same open question — learning to trust a disciplined negative result as much as a positive one, and to treat a bug caught before a result shipped as a success of the process rather than a delay.

Set against the four research questions, this chapter's honest summary is that the internship's methods held up better than its schedule, and its schedule better than its scope. RQ1 and RQ2 were answered, and the constraints that shaped them were engineering problems with engineering answers. RQ3 was answered negatively and revised downward twice, which is uncomfortable to write and is also the clearest single piece of evidence in this report that its evaluation standard was real rather than decorative: a standard that never overturns your own result is not being applied. RQ4 was answered only within a protocol, for a reason external to the method — the on-chain edges a cross-protocol answer would travel along mostly do not exist in this data — and that limitation is reported as a finding rather than buried as a caveat. None of those three outcomes is what the internship set out to produce. Stating them as they are, rather than around them, is the part of this work I would defend first.

Chapter 10

Conclusion and Outlook

This chapter returns to the four research questions fixed at the start (Table 1), answers each with its evidence and its limit in the same breath, scores the work against the seven original deliverables, and states what remains for Derivalink.

The answers are uneven, and the unevenness is the result. RQ1 (measurement) and RQ2 (causality) come back affirmative with quantities attached. RQ3 (prediction) comes back negative, with a consistency four independently designed workstreams did not set out to produce. RQ4 (systemic risk) comes back partly unanswerable, for a reason that is itself the finding: built from real on-chain edges rather than assumed, the dependency network the question presupposes is 110 nodes joined by 39 edges into 72 disconnected components, 48 of them isolated.

Every figure below is read from a computed artefact in the project's own corpus rather than from the narrative of an earlier draft, and where the two disagree the corpus governs. That matters for one headline: the Phase 2 claim that the twenty-five-delta skew *anticipates* stress events was withdrawn under an adversarial audit of the workstream's own pipeline, and the verdict under RQ3 is the post-audit one.

10.1 Answers to the research questions

RQ1 — measurement: yes. One convention applied across roughly twenty protocol families decomposes peg deviation into interpretable components, and the channel that binds differs by design family — the endogenous share leg in fractional designs, the redemption queue in a liquid-staking token, nothing measurable where the peg is enforced elsewhere. Its corollary is a discipline rather than a number: because the oracle, the pool and the centralized tape disagree by hundreds of basis points exactly when it matters, which series is called "the price" is a modelling decision that must be declared.

RQ2 — causality: yes, within a protocol, and the direction is consistent. Collateral-side and peg-side shocks Granger-lead synthetic-side deviation across six protocols analysed by six independently built pipelines, and the self-excitation of liquidation cascades separates the survivor from the casualty on a single scalar. The accompanying null strengthens it: no cross-excitation between DAI and LUSD in their shared stress window, established only after the estimator was shown able to recover a known non-zero effect on simulated data.

RQ3 — prediction: largely no, and this is the report's most consistent finding. Nothing here anticipates a depeg at a horizon and a false-alert rate that would make an automated breaker usable on its own. The one qualified positive — Isolation Forest clearing its bar on four of six assets, degrading precisely on the two whose peg is structurally biased away from par — is a detection result, not a prediction one. What this hands the V2 design is the shape of the constraint rather than a signal: roughly fifteen independent stress episodes per asset is not a sample from which a learned model can be expected to beat a trivial baseline, and model capacity did not change that.

RQ4 — systemic risk: answerable inside a protocol, largely unanswerable across them. Within a single protocol's local dependency graph, structural position does predict suffering, directionally, in all three replays. Across protocols it cannot be tested on this data, and that is a limitation of the available evidence rather than a null result; conflating the two would overstate what was tested.

10.2 Scorecard against the original objectives

Table 10.2 scores the seven original targets against evidence, so that "advanced" is not doing work a number could do instead.

	Verdict	Headline evidence	Principal limit
RQ1	Yes	One convention over ∼20 protocol families; USDC's oracle floors at 0.8800 while its CEX tape wicks to 0.81029, a 697 bp gap; 60% of CEX stress days invisible to that oracle; stETH discount −1.28 bp per queue-day, $R^2 = 0.43$	Estimator-level claim; one degenerate decomposition ($\sigma_S \equiv \sigma_P$ on sUSD) reported as failed
RQ2	Yes, within a protocol	31/39 pairs cointegrate at 5%; collateral side Granger-leads in every protocol tested (p down to 7.2×10^{-30}); Hawkes branching 0.968 (Iron) vs. 0.756 (DAI)	Predictive precedence, not structural causality; 8 in-scope assets have no pipeline
RQ3	Largely no	Persistence beats 5 supervised families 30/30; RL 0/8, then 3/8, then 8/8 at 25.70 false alerts/30 d; smirk F1 0.076 vs. 0.111 (BTC), anticipation permutation $p = 0.186$	Small effective samples (13 clusters, 8 episodes); smirk scored on BTC/ETH stress, not depegs
RQ4	Partially	Within-protocol replays directionally correct (ρ = 24–37%, 131%, 21–37%); across protocols, 110 nodes / 39 edges / 72 components / 48 isolated	Schema sees only collateral, oracle and pool edges; node value is a proxy, not a balance sheet

Table 10.1: The four research questions of Table 1, answered.

Id	Deliverable	Final status	Concrete evidence
D1	Unified database	Substantially delivered	∼77 GiB Parquet / ∼1.24 M files; 396,453,410 rows across 25 TimescaleDB tables, 0 empty
D2	IV smirk engine	Complete	888,334 IV legs; 68 events catalogued, 60 testable; free reconstruction path (Deribit, 2026) replacing a $30,000/year vendor tier; gate verdict issued and audited
D3	Historical depeg audit	Substantially advanced	Forensic reconstructions delivered per protocol (Iron Finance alone: 18.4 M rows, 59-day window), not assembled into the single 40–60 page artefact specified
D4	V2 smart contracts	Not started	Foundry project is an empty scaffold: test-library dependency only, zero project contracts
D5	Final research report	Separate deliverable	Not attempted here; this report is explicitly not its replacement
D6	Academic preprint	Not started	No draft written; publication would also need clearance on material covered by the internal specification
D7	Open dataset release	Not started	The multi-protocol tracking-error series exist on one convention; nothing packaged or licensed for release

Table 10.2: Final scorecard against the seven original deliverables, with the evidence behind each judgment.

That is not a disguised way of saying the internship fell short. Three of the specification's four planned phases — infrastructure, tracking-error and volatility analysis, and the broad Phase 3 programme covering econometrics, machine learning, reinforcement learning and graph theory — were completed across roughly twenty real protocols spanning every major peg-design family in use in DeFi today. The gap is specifically the second half of Phase 3 (V2 specification and Solidity prototyping) and all of Phase 4, and this report says so plainly rather than reframing unfinished work as out of scope from the start.

10.3 What remains for Derivalink

Four pieces of work follow from the four answers, and Chapter 7 is written to be usable as their starting point. *From RQ1*: the reference price should be an explicit, versioned parameter of any V2 monitoring component rather than an implementation detail, because the 697 bp gap between USDC's oracle floor and its exchange tape in March 2023 is exactly the spread a circuit breaker would have had to act inside. *From RQ2*: a dynamic collateral ratio should be driven by the volatility of whichever leg is endogenous to the design, the only covariate that binds in either partial-collateral protocol studied here (Kazemian et al., 2020). *From RQ3*: any early-warning component should carry a persistence baseline scored on identical days as its acceptance test, and a false-alert budget fixed in advance — the pooled reinforcement-learning run reached full recall with a real 46-hour lead and still failed on exactly that clause. *From RQ4*: the propagation model is working, tested software rather than a one-off analysis, and can be run periodically against fresh on-chain data as a bounded, signed estimate rather than a point prediction, carrying forward the decision that mattered most — separating pool depth from collateral backing, the difference between a threefold undershoot and a match within four percent. Two further items are inherited rather than produced: the eight in-scope assets with no pipeline are a data-engineering task of known shape rather than a research question, and D7 is closer than its "not started" status suggests, since the tracking-error series already exist on a single convention and what is missing is packaging and licensing rather than computation.

10.4 Limitations of this work

Beyond the limit stated under each answer above, four constraints bound the whole of this work.

A single analyst. Specification, extraction, analysis and writing were carried out by one person over sixteen weeks, with mentor checkpoints but no peer working the same problem. The compensating discipline was internal — pre-registered gates, hand-verified implementations, single-protocol pilots before generalization, adversarial audits of the project's own results, one of which withdrew a headline — and that is not the same as independent replication.

An 11.4 GB memory ceiling shaped the query strategy, not just the runtime. Analyses needing a large cross-asset panel in memory at once were not attempted, which is a constraint on the question set and not only on wall-clock time.

Free-tier data access shaped coverage. The options work is the sharpest case: the exchange's free tier serves one surface snapshot per month against a commercial daily-history tier at roughly $30,000 per year, so the entire Phase 2 panel exists only because a free reconstruction path was found (Deribit, 2026) — its granularity set by what was free rather than by what the question wanted.

Unfinished artefacts, and scope boundaries that were chosen rather than forced. D4, D6 and D7 have no artefact, and the Mirror Protocol sub-report carries its data-layer results only; where a result does not exist, this report says so rather than interpolating from a neighbouring protocol. Separately, no Transformer architecture was tested, the smirk was evaluated against Bitcoin and Ethereum volatility stress rather than synthetic-asset depegs, and the reinforcement-learning environment was built around one asset's reward calibration and generalized only afterwards.

Each was a deliberate scoping decision under a sixteen-week budget, and each is a place where a different choice could have produced a different answer.

10.5 Closing reflection

This internship set out to answer a genuinely open question — what five years of real, public DeFi history actually says about how and why synthetic assets fail — with evidence rather than anecdote, and honestly enough that a real design team could act on the answer. The clearest single lesson across every workstream is that simple, transparent, auditable models did a consistently useful job on it, while more sophisticated approaches built on the same data had to earn their place against a strict evaluation standard and mostly did not. Reporting that plainly, rather than around it, is this report's own contribution to Derivalink's next phase of work.

Two answers affirmative, one negative, one reporting that the question as posed cannot yet be settled. Peg deviation is measurable and attributable across mechanisms that share no design assumptions, provided the reference is chosen deliberately and the bias term is kept (RQ1). Inside a protocol the causal direction runs from collateral to synthetic, and the intensity of self-excitation in a liquidation cascade separates survivors from casualties by a wide, quantified margin (RQ2). A depeg is not, in this data, predictable early enough and cleanly enough to gate an automated circuit breaker on a single learned signal, and four independent attempts to show otherwise all failed against baselines a less disciplined evaluation would never have scored (RQ3). And the systemic-risk question is answerable inside a protocol, where the model is directionally right with a signed and measurable bias, but not yet across protocols, because the on-chain dependency network that question assumes is today a forest of 72 disconnected components rather than a system (RQ4). That distribution is the honest yield of sixteen weeks, and each of the four is more useful to a protocol designer than a fifth confident claim would have been.

Bibliography

Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003. The reference statement of realised volatility as an estimator of integrated variance.

Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models, 2019. URL https://arxiv.org/abs/1908.10063. arXiv:1908.10063. FinBERT is named in the client's own sentiment scope alongside CryptoBERT, but it was never run here: only the CryptoBERT checkpoint was scored. Cite this entry solely in the sentence that records that scope. Accessed 2026-08-27.

Backed Assets (JE) Limited. xStocks: Tokenized equities (TSLAx, COINx, NVDAx) — product documentation. Backed Assets product documentation, 2025. URL https://xstocks.com. Issuance, redemption and proof-of-reserve backing of the physically-backed tokenized equities. The mechanism reference for the synthetic-versus-physical comparison (sTSLA against TSLAx), which otherwise had no primary source in this file. Accessed 2026-06-09.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1550005, 2015. Estimation practice and the branching-ratio (criticality) reading used for the IRON 0.968 versus DAI 0.756 contrast.

Gurdip Bakshi, Nikunj Kapadia, and Dilip Madan. Stock return characteristics, skew laws, and the differential pricing of individual equity options. *Review of Financial Studies*, 16(1):101–143, 2003. Model-free risk-neutral skewness; the left-tail-fattening hypothesis.

Ole E. Barndorff-Nielsen, Peter Reinhard Hansen, Asger Lunde, and Neil Shephard. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536, 2008. Realised-kernel estimator, `src/te/temporal.py`.

David S. Bates. The crash of '87: Was it expected? the evidence from options markets. *Journal of Finance*, 46(3): 1009–1044, 1991. The core hypothesis: a skew steepening precedes a crash.

Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. DebtRank: Too central to fail? financial networks, the FED and systemic risk. *Scientific Reports*, 2:541, 2012. doi: 10.1038/srep00541. The from-scratch DebtRank contagion model, `src/graphs/propagation.py`.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995. BH-FDR adjustment on the Granger and smirk p-value panels.

Fischer Black. The pricing of commodity contracts. *Journal of Financial Economics*, 3(1–2):167–179, 1976. Black-76; the coin-denominated IV inverter, `src/smirk/black76.py`.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81 (3):637–654, 1973.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008. Louvain community detection, `src/graphs/communities.py`.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. doi: 10.1016/0304-4076(86)90063-1. GARCH(1,1) baseline, `src/te/garch.py`.

Tim Bollerslev, Viktor Todorov, and Sophia Zhengzi Li. Jump tails, extreme dependencies, and the distribution of stock returns. *Journal of Econometrics*, 172(2):307–324, 2013. Jump-tail interpretation of the left-tail smirk.

Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972. Eigenvector centrality, computed on the raw-weight projection and nulled for a snapshot when the power iteration fails to converge.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Antonio Briola, David Vidal-Tomás, Yuanrong Wang, and Tomaso Aste. Anatomy of a stablecoin's failure: The Terra-Luna case. *Finance Research Letters*, 51:103358, 2023. doi: 10.1016/j.frl.2022.103358. Preprint arXiv:2207.13914.

Mark Britten-Jones and Anthony Neuberger. Option prices, implied price processes, and stochastic volatility. *Journal of Finance*, 55(2):839–866, 2000. The model-free implied-variance result underlying the DVOL-style index the gate uses as its baseline.

Peter Carr and Liuren Wu. Stochastic skew in currency options. *Journal of Financial Economics*, 86(1):213–247, 2007. Time-varying 25Δ risk reversal as a state variable – the report's RR_{25} signal.

Peter Carr and Liuren Wu. Variance risk premiums. *Review of Financial Studies*, 22(3):1311–1341, 2009. The implied-minus-realised wedge as priced compensation – the correct reading of a skew signal that leads events without beating a volatility baseline on F1.

Christian Catalini, Alonso de Gortari, and Nihar Shah. Some simple economics of stablecoins. *Annual Review of Financial Economics*, 14:117–135, 2022. Taxonomy paper; the framing citation for the report's peg-regime classification (fiat-reserve / over-collateralised / partial / algorithmic / physically backed).

Celsius Network LLC. Voluntary petition for non-individuals filing for bankruptcy (chapter 11). United States Bankruptcy Court, Southern District of New York; case docket hosted by Stretto, July 2022. URL https://cases.stretto.com/celsius/. Withdrawal freeze 12 June 2022, petition filed 13 July 2022. The forced unwind of a large leveraged stETH position is the proximate driver of the June 2022 discount. Accessed 2026-06-18.

Chainalysis. UST's collapse & the trades that triggered it. Chainalysis Blog, June 2022. URL https://www.chainalysis.com/blog/how-terrausd-collapsed/. On-chain reconstruction of the 7–13 May 2022 depeg, including the Curve UST-3pool drain that opens the cascade. Retained because no issuer post-mortem exists for Terraform Labs. Accessed 2026-05-11.

Chainlink Labs. Chainlink data feeds: Aggregator heartbeat and deviation threshold. Chainlink developer documentation, 2024. URL https://docs.chain.link/data-feeds. Push-oracle update model — a feed updates on a heartbeat or on a deviation trigger, whichever comes first — which governs every on-chain oracle series used here and is the source of the oracle-staleness component of tracking error. Accessed 2026-06-18.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. Introduces the gated recurrent unit (GRU).

Circle Internet Financial. An update on USDC and Silicon Valley Bank. Circle Blog, March 2023. URL https://www.circle.com/blog/an-update-on-usdc-and-silicon-valley-bank. Issuer disclosure of $3.3B of reserves stranded at the failed bank, published 11 March 2023. This disclosure, not any on-chain series, is the origin of the USDC depeg. The post URL now redirects to the Circle blog index rather than to the post itself. Accessed 2026-08-27.

David G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978. The lower-tail copula family fitted by Kendall's-τ inversion ($\theta = 2\tau/(1-\tau)$).

Ryan Clements. Built to fail: The inherent fragility of algorithmic stablecoins. *Wake Forest Law Review Online*, 11: 131, 2021. Cited by the publisher as 11 Wake Forest L. Rev. Online 131 (2021). Start page only: the end page could not be confirmed against the publisher and is not guessed here.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7 (2):174–196, 2009. doi: 10.1093/jjfinec/nbp001. HAR-RV(1,5,22).

Gábor Csárdi and Tamás Nepusz. The `igraph` software package for complex network research. *InterJournal, Complex Systems*, 1695, 2006. Python bindings (`python-igraph`) version 0.11 or later, as pinned in `pyproject.toml`. Backend for the community-detection passes over the multi-protocol graph.

Curve Finance. PegKeepers: Algorithmic stabilisation of the crvUSD peg. Curve technical documentation, 2023. URL https://docs.curve.fi/crvUSD/pegkeeper/. Provide and withdraw mechanism: debt-ceiling-bounded minting of crvUSD into a StableSwap pool when the pool runs rich, and burning when it runs cheap. There is no holder-side redemption, so the peg is an algorithmic balance rather than an arbitrage floor. Accessed 2026-06-22.

Deribit. Deribit API v2 — `public/get_tradingview_chart_data`. Deribit API documentation, 2026. URL https://docs.deribit.com/. The endpoint that serves daily premium OHLC for expired option instruments. It is the free data path behind the daily implied-volatility panel used in the smirk study, in place of a paid historical options vendor. Accessed 2026-06-05.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019. The masked-language-model encoder family the pre-trained sentiment classifier belongs to. Cite it for the architecture only: the project runs the `ElKulako/cryptobert` checkpoint through the `transformers` pipeline and never documents which base model that checkpoint was adapted from, so no lineage claim should be made in the text.

David A. Dickey and Wayne A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366):427–431, 1979. ADF stationarity pre-test.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. The formal test the model-versus-persistence comparisons of the Lot-3 supervised chapter would need. It is cited as a stated limitation: this work compares PR-AUC point estimates against the persistence baseline without any test of the difference.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Larry Eisenberg and Thomas H. Noe. Systemic risk in financial systems. *Management Science*, 47(2):236–249, 2001. The clearing-vector model DebtRank is a response to; positions the graph work in its literature.

Paul Embrechts, Alexander J. McNeil, and Daniel Straumann. Correlation and dependence in risk management: Properties and pitfalls. In M. A. H. Dempster, editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, Cambridge, 2002. Tail-dependence coefficients λ_L, λ_U and why linear correlation is the wrong summary under co-crash risk. Also the source of the elliptical $\rho = \sin(\pi\tau/2)$ inversion used for the Gaussian comparator.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007, 1982. doi: 10.2307/1912773. The ARCH model Bollerslev (1986) generalises, and the origin of the ARCH-LM residual test run here on every fitted volatility model (sUSD: $\chi^2(5) > 10^4$ on raw returns, $p \approx 1$ on standardised residuals).

Robert F. Engle and Clive W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987. doi: 10.2307/1913236. Two-step cointegration test, `src/te/causality.py`.

Shayan Eskandari, Mehdi Salehi, Wanyun Catherine Gu, and Jeremy Clark. SoK: Oracles from the ground truth to market manipulation. In *Proceedings of the 3rd ACM Conference on Advances in Financial Technologies (AFT '21)*, pages 127–141, 2021. doi: 10.1145/3479722.3480994. Systematisation of on-chain oracle designs and their manipulation surface: the academic counterpart to the report's push-feed / TWAP / pull-feed oracle catalogue.

Ethena Labs. Key addresses — Ethena documentation. Ethena protocol documentation, 2024. URL https://docs.ethena.fi/solution-design/key-addresses. Contract registry for the delta-neutral design: the spot custody leg, the minting contract, and the staked sUSDe ERC-4626 vault whose share price, not the dollar peg, is the correct reference for sUSDe. Accessed 2026-05-29.

Farama Foundation. Gymnasium: a standard API for reinforcement learning environments. Open-source software, 2026. URL https://github.com/Farama-Foundation/Gymnasium. Version 1.0 or later, as pinned in `pyproject.toml`. The environment interface the historical-replay alert environment implements (`spaces.Discrete(4)` actions, a `spaces.Box` observation of 30 standardised columns plus the 3-bit thermometer encoding of the previous action). Accessed 2026-08-27.

Federal Deposit Insurance Corporation. FDIC acts to protect all depositors of the former Silicon Valley Bank, Santa Clara, California. FDIC press release, March 2023. URL https://www.fdic.gov/news/press-releases/2023/pr23019.html. The depositor backstop that ended the USDC depeg: announced in a joint Treasury/Federal Reserve/FDIC statement on 12 March 2023 and executed by this FDIC-only release of 13 March 2023, which transferred all deposits to a bridge bank. Accessed 2026-08-27.

Shaun Fernando. DVOL — Deribit implied volatility index. Deribit Insights, March 2021. URL https://insights.deribit.com/exchange-updates/dvol-deribit-implied-volatility-index/. Definition of the 30-day model-free implied-volatility index used as the trivial baseline against which the 25-delta skew signal is scored. Last updated by the publisher on 20 March 2023. Accessed 2026-08-27.

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. Betweenness on the 1/weight distance projection.

Linton C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1978. Degree and closeness, the two remaining metrics the snapshot tables report; closeness is computed on the same 1/weight distance projection, which is why its absolute scale is dollar-magnitude dependent.

Gregory Gadzinski, Alessio Castello, and Florie Mazzorana. Stablecoins: Does design affect stability? *Finance Research Letters*, 53:103611, 2023. doi: 10.1016/j.frl.2022.103611.

Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801, 1993. GJR-GARCH(1,1,1) leverage term.

Gary B. Gorton and Jeffery Y. Zhang. Taming wildcat stablecoins. *University of Chicago Law Review*, 90(3):909–971, 2023. Stablecoins as privately issued money, and depegs as runs on a liability – the frame under which the DAI, UST and USDC case studies are read.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Emil J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960. The upper-tail copula family ($\theta = 1/(1-\tau)$).

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA, 2008. Software version 3.4 or later, as pinned in `pyproject.toml`. Graph construction, centrality and the from-scratch DebtRank propagation run on NetworkX graph objects.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. Univariate and bivariate exponential-kernel Hawkes MLE, `src/te/hawkes.py`.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. The encoder–decoder architecture of the radar's autoencoder (symmetric MLP $n \to n/2 \to n/4 \to n/2 \to n$). It is a dimensionality-reduction paper: cite Sakurada and Yairi (2014) for the use of the reconstruction error as an anomaly score.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Ai Jun Hou, Weining Wang, Cathy Y. H. Chen, and Wolfgang Karl Härdle. Pricing cryptocurrency options. *Journal of Financial Econometrics*, 18(2):250–279, 2020. doi: 10.1093/jjfinec/nbaa006. Crypto-specific option-pricing anchor for the smirk chapter, whose other references are drawn from equity and FX markets.

Iron Finance. Iron Finance post-mortem 17 June 2021. Iron Finance team publication, Medium, June 2021. URL https://ironfinance.medium.com/iron-finance-post-mortem-17-june-2021-6a4e9ccf23f5. The protocol's own account of the 16–17 June 2021 IRON/TITAN collapse, including the positive-share-price guard whose revert froze redemptions entirely. Accessed 2026-08-27.

Harry Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, London, 1997. Copula-family selection and tail dependence.

Søren Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59(6):1551–1580, 1991. Trace and maximum-eigenvalue rank tests, VECM; `src/te/causality.py`.

Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

Sam Kazemian, Kedar Iyer, and Jason Huan. Frax: A fractional-algorithmic stablecoin protocol. Frax Finance protocol whitepaper, 2020. URL https://docs.frax.finance/. Defines the collateral-ratio mint and redeem mechanism — pure-collateral mint at CR = 1, fractional mint burning FXS, and redemption paid in collateral plus newly minted FXS — which Iron Finance reproduced almost verbatim. Cited to the documentation root, which is where the protocol currently serves this material. Accessed 2026-06-16.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3146–3154, 2017.

Ariah Klages-Mundt and Andreea Minca. (In)Stability for the blockchain: Deleveraging spirals and stablecoin attacks. *Cryptoeconomic Systems*, 1(2), 2021. Published 2021-10-22; first released as arXiv:1906.02152 in 2019.

Ariah Klages-Mundt and Andreea Minca. While stability lasts: A stochastic model of noncustodial stablecoins. *Mathematical Finance*, 32(4):943–981, 2022. doi: 10.1111/mafi.12357. Deleveraging-spiral model of noncustodial stablecoins; the theoretical frame for the FRAX/IRON reflexivity analysis and for the response-rate constraint on the collateral ratio. Published version of the 2020 preprint arXiv:2004.01304.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17 (3):1217–1241, 1989. Resampling contiguous blocks rather than individual observations: the justification for the cluster (block) bootstrap that carries the Phase-2 verdict, where the resampling unit is the merged event-window cluster and not the grid day.

Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1–3):159–178, 1992. KPSS. Paired with ADF as the order-of-integration pre-filter that ruled out a coherent I(1) system on the UST panel.

Lido DAO. Lido on Ethereum: Protocol documentation. Lido technical documentation, 2024. URL https://docs.lido.fi/. Rebase accounting and `getPooledEthByShares`: the share-price definition (total pooled ETH over total shares) used to separate the wstETH value from the rebasing stETH peg denominator. Accessed 2026-06-18.

Liquity AG. Redemptions and the LUSD price floor. Liquity protocol documentation, 2021. URL https://docs.liquity.org/. Defines the hard $1 redemption right against the lowest-collateralised troves and the decaying base rate — the mechanical floor that makes LUSD's peg premium-biased rather than symmetric. The v1 FAQ page that carried this description has been retired; the host now serves Liquity V2 documentation, whose redemption rules are not those described here. Accessed 2026-08-27.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17. The one Lot-3 model that cleared its own bar (4 of 6 assets). The sub-sampling size fixed a priori by the project, `max_samples=256`, is the value this paper recommends; `n_estimators=300` is the project's own, above the paper's setting.

Jiageng Liu, Igor Makarov, and Antoinette Schoar. Anatomy of a run: The Terra Luna crash. NBER Working Paper 31160, National Bureau of Economic Research, 2023.

LlamaRisk. Reserve fund subcommittee october 2025 update. Post to the Ethena governance forum on behalf of the Reserve Fund Subcommittee, 3 November 2025, November 2025. URL https://gov.ethenafoundation.com/t/reserve-fund-subcommittee-october-2025-update/709. Official record for the 10–12 October 2025 crash window: the Reserve Fund balance held intact and more than $1.5B of USDe outflows cleared on 10 October, while a venue-specific oracle quoted USDe away from par. Accessed 2026-05-29.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017. The SHAP framework. Cite it for the attribution's semantics, but the Lot-3 chapter runs `shap.TreeExplainer` on an XGBoost booster, i.e. the TreeSHAP algorithm of Lundberg et al. (2020).

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. TreeSHAP: the exact polynomial-time algorithm behind `shap.TreeExplainer`, which is the estimator the Lot-3 feature-attribution figures actually use.

Richard K. Lyons and Ganesh Viswanath-Natraj. What keeps stablecoins stable? *Journal of International Money and Finance*, 131:102777, 2023. doi: 10.1016/j.jimonfin.2022.102777. Primary-market arbitrage as the peg-restoration channel.

James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985. The HC3 estimator, which is the declared cross-sectional standard-error convention of the shared level-1 regression (`fit(cov_type="HC3")`, `src/te/decomp.py`). Cited here because Newey–West covers only the time-series case.

Igor Makarov and Antoinette Schoar. Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2):293–319, 2020. Limits to cross-venue arbitrage – the mechanism behind venue-level TE dispersion.

MakerDAO. Peg Stability Module (PSM) documentation. MakerDAO technical documentation, 2021. URL https://docs.makerdao.com/smart-contract-modules/peg-stability-module. Defines the fee-bounded 1:1 USDC–DAI swap facility, the channel through which the March 2023 USDC depeg propagated into DAI. Following

the Sky rebrand this URL now issues a 307 redirect to https://developers.skyeco.com/ (checked 2026-08-27); it is cited as the MakerDAO-era document of record. Accessed 2026-08-27.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. The paired test used on per-day detector correctness (skew versus DVOL on the same grid days), `src/smirk/stats.py`.

Robert C. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4(1): 141–183, 1973. The continuous-dividend-yield extension. The equity cross-market pilot does not reuse the Black-76 engine: it is a separate Black–Scholes–Merton module carrying a dividend yield q, `src/smirk/equity_iv.py`, and the European inversion is applied to American-exercised listed options with that approximation stated.

Microsoft Corporation. LightGBM: a fast, distributed, high-performance gradient-boosting framework. Open-source software, 2026. URL https://github.com/microsoft/LightGBM. Version 4.0 or later, as pinned in `pyproject.toml`. The algorithm itself is cited separately to Ke et al. (2017). Accessed 2026-08-27.

Jason Milionis, Ciamac C. Moallemi, Tim Roughgarden, and Anthony Lee Zhang. Automated market making and loss-versus-rebalancing, 2022. URL https://arxiv.org/abs/2208.06046. arXiv:2208.06046. Basis of the LVR decomposition of venue-level TE. Accessed 2026-08-27.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. DQN with experience replay and a periodically copied target network – the algorithm of the alert-monitoring pilot, `src/rl/`. The agent itself is not reimplemented: it is the Stable-Baselines3 `DQN` with an `MlpPolicy`, cited separately in the software list.

Jordan Momtazi. SIP-169: Deprecate low volume L1 synths. Synthetix Improvement Proposal 169, Wezen release, July 2021. URL https://sips.synthetix.io/sips/sip-169/. Frozen-redemption deprecation of the low-volume L1 synths — the equity, commodity, index and long-tail crypto cohorts — effective with the Wezen release of 2021-09-16. This governance act is the primary record of the protocol-mortality failure mode: after the freeze a synth's tracking error against a live underlying grows without bound. Accessed 2026-08-27.

Nansen Research. On-chain forensics: Demystifying stETH's "de-peg". Nansen Research report, 29 June 2022, June 2022. URL https://www.nansen.ai/research. Wallet-level reconstruction of the June 2022 stETH discount. No protocol post-mortem exists for this event, since Lido's mechanism did not fail — the discount was a secondary-market and forced-deleveraging phenomenon. The publisher's deep link no longer resolves, so the report cites the research portal. Accessed 2026-08-27.

Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991. doi: 10.2307/2938260. EGARCH.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. HAC standard errors, `src/te/did.py` and every time-series driver regression (bandwidths actually used: 5, 22, 24 and 60 minutes depending on the asset).

Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. The objective function both Louvain and Leiden optimise.

T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979. The recursion $A_i = e^{-\beta(t_i - t_{i-1})}(1 + A_{i-1})$ that makes the exponential-kernel log-likelihood $O(N)$ instead of $O(N^2)$.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. Influence centrality on the raw-weight graph projection, `src/graphs/centrality.py`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019. Software version 2.0 or later, as pinned in `pyproject.toml`. Used for the LSTM, GRU and autoencoder heads.

Andrew J. Patton. Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556, 2006. Copulas on GARCH-standardised residuals: the justification for the DAI robustness pass that refits Clayton/Gumbel on GJR-GARCH(1,1,1) residuals $z_t = (r_t - \mu)/\sigma_t$ before the probability integral transform, because shared volatility regimes inflate Kendall's τ on raw returns.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. Software version 1.8 or later, as pinned in `pyproject.toml`. Supplies the Isolation Forest, One-Class SVM, Random Forest and spectral-clustering implementations, the logistic baseline, and the PR-AUC / Brier / precision-recall-curve metrics on which every model verdict is read.

Antti Petajisto. Inefficiencies in the pricing of exchange-traded funds. *Financial Analysts Journal*, 73(1):24–54, 2017. doi: 10.2469/faj.v73.n1.7. ETF premium/discount dynamics: the closest TradFi analogue to a token trading away from NAV (GLP, xStocks).

Polars Development Team. Polars: a multi-threaded DataFrame library for Rust and Python. Open-source software, 2026. URL https://github.com/pola-rs/polars. Version 1.40 or later, as pinned in `pyproject.toml`. Primary in-memory dataframe engine for all extraction and analysis code. Accessed 2026-08-27.

Kaihua Qin, Liyi Zhou, and Arthur Gervais. Quantifying blockchain extractable value: How dark is the forest? In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 198–214, 2022. The adversarial-execution layer (sandwiching, liquidation racing) underneath every on-chain TE observation.

Mark Raasveldt and Hannes Mühleisen. DuckDB: an embeddable analytical database. In *Proceedings of the 2019 ACM SIGMOD International Conference on Management of Data*, pages 1981–1984, 2019. Software version 1.5 or later, as pinned in `pyproject.toml`; https://duckdb.org. Out-of-core query engine over the Parquet research store.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. Software version 2.0 or later, as pinned in `pyproject.toml`. The DQN agent of the alert-monitoring pilot is this implementation (`MlpPolicy`, replay buffer 50,000, target-network interval 500, ε annealed over the first 30 % of training), not a reimplementation of Mnih et al. (2015); the discount factor is left at the library default and is never set in this project's code.

Resolv Labs. Resolv postmortem: March 22, 2026 incident. Resolv Blog, March 2026. URL https://resolv.xyz/blog/resolv-postmortem-march-22-2026-incident. Protocol disclosure of the compromised credentials controlling the off-chain signer and of the unauthorised USR mints of 22 March 2026. Accessed 2026-08-27.

Richard Roll. A mean/variance analysis of tracking error. *Journal of Portfolio Management*, 18(4):13–22, 1992. The base definition used in this report: TE as the standard deviation of the return differential between the replicating instrument and its benchmark.

Mark Rubinstein. Implied binomial trees. *Journal of Finance*, 49(3):771–818, 1994. Risk-neutral-distribution recovery framing of the smirk.

Kanis Saengchote. A DeFi bank run: Iron Finance, IRON stablecoin, and the fall of TITAN. PIER Discussion Paper 155, Puey Ungphakorn Institute for Economic Research, 2021. URL https://www.pier.or.th/dp/155/.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432. Justifies PR-AUC over AUC-ROC, the choice the ML chapter makes explicitly at its own stated 1–14 % positive base rates in the test blocks (and at the 2.85 % rate on which the RL reward is calibrated).

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. ACM, 2014. Per-row reconstruction MSE as the anomaly score – exactly what `autoencoder_scores` returns, and the step Hinton and Salakhutdinov (2006) does not itself make.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. One-class SVM, `src/ml/models.py`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347. arXiv:1707.06347. The on-policy alternative the RL pilot explicitly rejected in favour of DQN's replay buffer. Accessed 2026-08-27.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, 2010. Software version 0.14 or later, as pinned in `pyproject.toml`. Provides the ADF, Johansen, Granger, VAR/VECM and Newey–West HAC routines used throughout.

Kevin Sheppard. `arch`: Autoregressive conditional heteroskedasticity and other tools for financial econometrics in Python. Open-source software, 2026. URL https://github.com/bashtage/arch. Version 8.0 or later, as pinned in `pyproject.toml`. Provides the GARCH(1,1), EGARCH and GJR-GARCH estimators used in the volatility decomposition. Accessed 2026-08-27.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959. The theorem the whole copula layer rests on; `src/te/copulas.py`.

Solana Foundation. Token-2022 program: Scaled UI amount extension. Solana Program Library documentation, 2025. URL https://spl.solana.com/token-2022/extensions. Defines the on-chain multiplier the dividend natural experiment is built on: a corporate action is applied as a scaling factor on the token amount, not as a cash payment, so the tracking error must be measured against a multiplier-adjusted reference. Accessed 2026-06-09.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

Synthetix. Synthetix V3 documentation. Synthetix technical documentation, 2025. URL https://docs.synthetix.io/v/v3. Shared debt-pool accounting: every staker is counterparty to every synth, so a synth carries no dedicated price feed of its own and its peg question is separable from the underlying feed's fidelity. Accessed 2026-05-27.

Donald L. Thistlethwaite and Donald T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960. Regression-discontinuity design for the Shapella / withdrawal natural experiment.

Timescale, Inc. TimescaleDB: a time-series database built on PostgreSQL. Open-source software, 2026. URL https://github.com/timescale/timescaledb. Hypertable storage engine for the curated live-monitoring database (PostgreSQL 17). Accessed 2026-08-27.

Vincent A. Traag. `leidenalg`: Leiden community detection for Python. Open-source software, 2026. URL https://github.com/vtraag/leidenalg. Version 0.10 or later, as pinned in `pyproject.toml`. Implements the algorithm of Traag, Waltman and van Eck (2019), which is cited separately. Accessed 2026-08-27.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019. Leiden community detection, `src/graphs/communities.py`. The project runs it under `leidenalg.ModularityVertexPartition`, i.e. the modularity objective – not the Constant Potts Model this paper also introduces, and with no resolution parameter set.

U.S. Securities and Exchange Commission. SEC v. Terraform Labs Pte. Ltd. and Do Hyeong Kwon, no. 23-cv-1346. Complaint filed in the United States District Court for the Southern District of New York, February 2023. URL https://www.sec.gov/newsroom/press-releases/2023-32. Official record of the Terra/UST design and its promotion. Accessed 2026-05-11.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. The third community-detection algorithm run alongside Louvain and Leiden, on the precomputed weighted adjacency with k tied to Louvain's own community count.

web3.py Contributors. `web3.py`: a Python library for interacting with Ethereum. Open-source software, Ethereum Foundation, 2026. URL https://github.com/ethereum/web3.py. Version 7.0 or later, as pinned in `pyproject.toml`. The JSON-RPC client behind every on-chain event and state extractor. Accessed 2026-08-27.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. Software version 4.40 or later, as pinned in `pyproject.toml`. The `text-classification` pipeline through which the CryptoBERT checkpoint scores the news corpus, truncating at 128 tokens per the model card.

XGBoost Contributors. XGBoost: Scalable, portable and distributed gradient boosting. Open-source software, Distributed (Deep) Machine Learning Community, 2026. URL https://github.com/dmlc/xgboost. Version 3.0 or later, as pinned in `pyproject.toml`. The algorithm itself is cited separately to Chen and Guestrin (2016). Accessed 2026-08-27.

Yuhang Xing, Xiaoyan Zhang, and Rui Zhao. What does the individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis*, 45(3):641–662, 2010. The predictive-smirk literature the Phase-2 gate tests against a DVOL baseline.

Lan Zhang, Per A. Mykland, and Yacine Aït-Sahalia. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411, 2005. doi: 10.1198/016214505000000169. Two-scales realised variance (TSRV); implemented at `src/te/temporal.py` (equation 64).

Appendices

Appendix A

Inventories and Repository Structure

A.1 Full protocol and asset inventory

Protocol	Peg regime	Coverage notes
MakerDAO (DAI)	Full CDP collateral	6-year backfill, 52 historical collateral legs, March 2023 depeg case study
Iron Finance (IRON/TITAN)	Partial collateral	Full June 2021 collapse forensics, block-level reconstruction
FRAX V1	Partial collateral	Hybrid-era reconstruction, direct comparison against Iron Finance
Terra (UST, LUNA)	Algorithmic	May 2022 collapse, native and bridged variants
Mirror Protocol	Algorithmic (UST-collateralized)	34 mAssets, Band oracle freeze empirically dated
Synthetix	Shared debt pool	Equity, commodity, index and inverse cohorts across 3 contract versions
Circle (USDC)	Fiat-backed	Reference stablecoin, March 2023 SVB depeg
Tether (USDT)	Fiat-backed	Reference stablecoin
Lido (stETH, wstETH)	Liquid staking	June 2022 contagion event
Liquity (LUSD)	Full CDP collateral	Premium-biased peg, redemption-floor mechanics
Curve (crvUSD)	Full CDP collateral	PegKeeper-defended peg
GMX (GLP)	Basket/index-tracking	Perpetuals-linked liquidity index
Ethena (USDe, sUSDe)	Delta-neutral	2024 launch, October 2025 oracle-bug case study
Resolv (USR, stUSR, wstUSR)	Delta-neutral	March 2026 opsec-driven depeg, forensic reconstruction
Vertex Protocol	Perpetuals order book	Top-3 perpetual markets, TE relative to perp-mid reference
Backed Finance (xStocks)	Physically-backed	Tokenized equities, on-chain proof-of-reserve

A.2 Technical reports produced

Each protocol above has a dedicated technical sub-report (data audit, tracking-error analysis, and where relevant a historical-collapse case study), in addition to the cross-cutting syntheses below. All are available on request; none are reproduced here, per the confidentiality note on the cover page. The granular material condensed out of this report during final editing — exhaustive per-asset and per-fold tables, hyperparameter inventories and secondary derivations — lives in these documents.

Report	Covers
IV Smirk gate report	Phase 2 gate decision, full methodology and 60-event panel
Econometrics synthesis (S10)	VAR/Granger/cointegration/copulas/Hawkes across the whole dataset
ML cross-asset comparison	Supervised, anomaly-detection and deep-learning results, all 6 assets
Sentiment ablation study	The sentiment-feature experiment (Section 6.2)
RL alert-monitoring pilot	Base pilot + reweighting + pooling follow-ups (Section 6.3)
Lot-3 AI/ML closing report	Section-by-section answer to the AI/ML briefing documents
DAI / UST / Iron Finance replays	Full methodology and results for each historical replay
Graph theory closing report	Systemic-risk model methodology, all 3 replays, multi-protocol generalization

A.3 Repository structure

```
.
|-- ROADMAP.md          Living roadmap, single source of truth for status
|-- src/                Python package
|   |-- etl/            Extract pipelines, one per protocol
|   |-- te/             Tracking-error computation, sigma-decomposition, HAR
|   |-- smirk/          Implied-volatility smirk engine
|   |-- graphs/         DeFi dependency graph, centrality, DebtRank
|   |-- ml/             Supervised, anomaly-detection and deep-learning models
|   `-- rl/             Reinforcement-learning environment and training
|-- contracts/          Foundry project (V2 prototype, not yet started)
|-- data/               Parquet + DuckDB (gitignored)
|-- reports/            Markdown and LaTeX technical reports
`-- docs/               Schema documentation, ADRs, data-source registry
```