POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Data-Driven Security in Decentralized Finance: Statistical Analysis and Predictive Modeling of the DeFi crime landscape

Tesi di Laurea Magistrale in
Mathematical Engineering – Ingegneria Matematica

Author: **Federico Oldani**

Student ID: 275337
Advisor: Prof. Daniele Marazzina
Academic Year: 2025–26

Abstract

The rapid expansion of Decentralized Finance (DeFi) has introduced a new paradigm in financial intermediation, but it has also catalyzed a surge in sophisticated, profit-driven criminal activities. These threats primarily manifest as **exploits**—cyberattacks that leverage vulnerabilities in smart contracts, economic mechanisms, or user interfaces—which have resulted in cumulative losses exceeding $20 billion between 2017 and 2025. Attack vectors range from technical flaws such as reentrancy and oracle manipulation to economic schemes like flash loan attacks and rug pulls, as well as social engineering tactics including phishing and DNS hijacking.

Current academic research in this domain often oscillates between qualitative criminological taxonomies and isolated technical detection models, resulting in a fragmented understanding that lacks comprehensive statistical validation of these phenomena. This thesis seeks to address this limitation by proposing a multi-layered analytical framework that integrates criminological theory with quantitative inference and predictive modelling. In the first stage, a rigorous statistical analysis is conducted on a comprehensive dataset of documented DeFi exploits, categorized according to the "Oxford Taxonomy" [4]. Through the application of non-parametric hypothesis testing, the financial impact of criminal activity is examined across distinct technical layers and actor roles, enabling the identification of statistically significant on-chain risk indicators.

The proposed model is trained on a comprehensive dataset specifically constructed for anomaly detection, integrating both normal on-chain activity and documented exploit events. The methodology is first evaluated on synthetically generated data to assess its theoretical properties, before being applied to large-scale real-world datasets sourced from Google Cloud BigQuery [10]. Anomalous observations are incorporated using the dataset curated by Carpentier-Desjardins et al. [4], together with exploit records obtained from the De.Fi database, which provide detailed coverage of verified abnormal on-chain behaviors.

Keywords: Decentralized Finance (DeFi), criminal activity, Oxford Taxonomy, Anomaly Detection.

Abstract in lingua italiana

La rapida espansione della Finanza Decentralizzata (DeFi) ha introdotto un nuovo paradigma nell'intermediazione finanziaria, ma ha anche catalizzato un aumento di attività criminali sofisticate e orientate al profitto. Queste minacce si manifestano principalmente come **exploit**—attacchi informatici che sfruttano vulnerabilità negli smart contract, nei meccanismi economici o nelle interfacce utente—che hanno causato perdite cumulative superiori a $20 miliardi tra il 2017 e il 2025. I vettori di attacco spaziano da difetti tecnici come reentrancy e manipolazione degli oracle a schemi economici come attacchi flash loan e rug pull, oltre a tattiche di social engineering tra cui phishing e hijacking DNS.

La ricerca accademica attuale in questo ambito oscilla spesso tra tassonomie criminologiche qualitative e modelli tecnici isolati di rilevamento, risultando in una comprensione frammentata che manca di una valida conferma statistica di questi fenomeni. Questa tesi si propone di colmare tale lacuna, proponendo un framework analitico multi-livello che integri la teoria criminologica con l'inferenza quantitativa e la modellazione predittiva. Nella prima fase, viene condotta un'analisi statistica rigorosa su un dataset completo di exploit DeFi documentati, categorizzati secondo la "Oxford Taxonomy" [4]. Attraverso l'applicazione di test non parametrici, l'impatto finanziario delle attività criminali viene esaminato su diversi livelli tecnici e ruoli degli attori, permettendo l'identificazione di indicatori di rischio on-chain statisticamente significativi.

Il modello proposto viene addestrato su un dataset completo, appositamente costruito per la rilevazione di anomalie, che integra sia l'attività on-chain normale sia gli exploit documentati. La metodologia viene inizialmente valutata su dati sintetici per verificarne le proprietà teoriche, prima di essere applicata a dataset reali di larga scala provenienti da Google Cloud BigQuery [10]. Le osservazioni anomale vengono integrate utilizzando il dataset curato da Carpentier-Desjardins et al. [4], insieme ai record di exploit ottenuti dal database De.Fi, che forniscono una copertura dettagliata di comportamenti anomali on-chain verificati.

Parole chiave: Finanza Decentralizzata (DeFi), attività criminale, Tassonomia di Oxford, Rilevamento di Anomalie.

Contents

Introduction

The DeFi Ecosystem: Architecture and Mechanisms

The rise of the Decentralized Finance (DeFi) has fundamentally reshaped the financial landscape by introducing new types of operations that are anonymous and untraceable, replacing traditional intermediaries with autonomous, code-based protocols. The DeFi ecosystem relies primarily on blockchains like Ethereum and leverages smart contracts to allow lending, borrowing, and trading with self-custody in a decentralized environment. This architectural shift, while democratizing access to complex financial instruments, introduces a novel set of structural vulnerabilities. Unlike Centralized Finance (CeFi), where security is managed through institutional surveillance, DeFi security is intricately linked to the safety and integrity of the underlying code [1] and the economic incentives of its contributors [20]. The composability of these protocols—often referred to as money legos [16]—further exacerbates the system's risk exposure, as a vulnerability in a single foundational component may propagate across multiple applications, potentially leading to widespread disruptions within the interconnected DeFi ecosystem.

Systemic Risk and Profit-Driven Crime in Decentralized Finance

Profit-driven criminal activity within the DeFi ecosystem has evolved from an isolated and marginal occurrence into a significant source of systemic risk. The inherent transparency of distributed ledgers, combined with the pseudonymous nature of user identities, creates an environment where highly sophisticated actors can conduct large-scale exploits with immediate and irreversible settlement [17]. These incidents, ranging from flash-loan attacks to sophisticated rug-pull schemes, generate annual losses amounting to several billions of dollars, undermining investor confidence and threatening market stability. As highlighted in recent criminological literature, particularly within the framework of the Oxford Taxonomy [4], these incidents should not be interpreted as accidental technical failures but rather as intentional strategies designed to drain liquidity from decentralized

systems. Understanding the scale and frequency of these events is essential for developing robust defense mechanisms capable of preserving the resilience of the DeFi infrastructure.

Thesis Objectives: From Descriptive Taxonomy to Statistical Validation

The primary objective of the thesis is to bridge the gap between qualitative criminological frameworks and quantitative predictive modelling within the Decentralized Finance ecosystem. While existing taxonomies, such as the one proposed by Carpentier-Desjardins et al. [4], provide an essential descriptive vocabulary for DeFi crimes, they remain largely conceptual and are seldom supported by empirical statistical validation. To our knowledge, no prior work has systematically tested whether these conceptual distinctions correspond to statistically significant differences in observable financial outcomes.

This thesis addresses this gap through three contributions. First, we conduct a comprehensive statistical validation of the Oxford Taxonomy using non-parametric hypothesis testing, including temporal validation across distinct periods to assess pattern persistence under evolving threat landscapes. Second, we develop an anomaly detection framework that integrates supervised and unsupervised learning methods, using features derived from statistical validation rather than arbitrary technical metrics. Third, we perform out-of-sample temporal evaluation to assess model robustness against emerging exploit strategies.

Building upon these findings, the thesis develops predictive models designed to identify on-chain structural deviations indicative of malicious behaviour, evaluated using recent market data from the 2023–2025 period.

Structure of the Thesis

The outline of this thesis is organized as follows:

In Chapter 1, a comprehensive review of the existing literature is presented, inspecting the criminological perspective through the "Oxford Taxonomy" [4] and the technical detection frameworks proposed by the ACM community [22]. This chapter explains in detail how this work integrates these two distinct viewpoints.

Chapter 2 deals with the Data Engineering process, describing the sourcing of data from DeFi crime aggregators and the extraction of on-chain features from the Google Cloud BigQuery Ethereum dataset. It details the harmonization and feature engineering re-

quired to build a unified analytical framework.

Chapter 3 presents the Exploratory Data Analysis (EDA). This chapter investigates the fundamental patterns within the integrated dataset, identifying the "fat-tail" nature of financial losses, the temporal clustering of exploits, and the structural differences in gas consumption between normal and anomalous transactions.

Chapter 4 introduces the statistical inference framework. It focuses on the validation of the impact of actor roles and technical layers through non-parametric tests, such as the Mann-Whitney U and Kruskal-Wallis tests, providing empirical support for the theoretical taxonomies.

Chapter 5 presents the core predictive modeling approach, illustrating a hybrid anomaly detection framework. It compares supervised classification baselines with a suite of unsupervised architectures designed to learn representations of on-chain normality, evaluating how varying algorithmic complexity impacts the detection of diverse DeFi risk profiles.

Finally, Chapter 6 discusses the practical implications for cybersecurity and regulatory frameworks, while Chapter 7 presents the final conclusions, highlighting the original contributions of this research and suggesting possible directions for future work.

1 | Literature Review: Mapping the DeFi Crime Landscape

1.1. The Criminological Perspective: The Oxford Taxonomy

Understanding crime in decentralized finance requires a framework that can explain not just how attacks work, but who perpetrates them, who suffers losses and why certain types of crime cause more damage than others. Carpentier-Desjardins et al.[4] provide such a framework through an evidence-based taxonomy built by analyzing 1,036 DeFi crime events from 2017 to 2022. This section synthesizes their key findings which serve as the theoretical foundation for our statistical validation and predictive modeling.

1.1.1. Methodology and Data Sources

Let us begin by examining how the authors constructed this framework. They collected crime events from three major aggregators: De.Fi REKT database [8], SlowMist [19] and CryptoSec (now ChainSec) [5]. These platforms track reported exploits, exit scams and fraudulent schemes in the DeFi ecosystem, providing incident summaries, financial damage estimates and links to blockchain evidence.

The authors used an inductive content analysis approach—a qualitative method in which patterns and categories emerge from the data itself [14]. Through open coding, axial coding and selective coding, they derived a four-dimensional classification: *implication of DeFi actors* (whether entities are targets, perpetrators or intermediaries), *strategies* (high-level approaches like exploiting technical flaws versus human trust), *general tactics* (broad methods such as contract compromise or liquidity drainage) and *specific tactics* (precise techniques including reentrancy exploits, flash loans or phishing campaigns).

The dataset was restricted to profit-driven crimes, in accordance with Naylor's theory of predatory wealth redistribution [14]. This excluded non-financial attacks, thwarted

attempts, bug bounty disclosures and poorly documented cases. The final dataset comprised 1,036 DeFi-specific crime events accounting for $10 billion in losses, with financial data available for 904 events (87%).

1.1.2. The Three Roles of DeFi Actors

The central contribution of the taxonomy is recognizing that DeFi actors—entities developing and operating DeFi protocols, tokens, exchanges and services—can occupy three distinct positions in criminal incidents. This isn't merely a classification exercise; it fundamentally shapes how we understand risk distribution and prioritize security investments.

DeFi Actor as Target (52.1% of events, 83% of damages)

In this scenario, legitimate protocols fall victim to external attackers. These attacks exploit either technical vulnerabilities (46.7%) or human risks (3.1%). The dominance of technical attacks reflects DeFi's architectural exposure: smart contracts are complex programs handling billions in value, publicly accessible for anyone to probe and typically immutable once deployed.

Exploiting technical vulnerabilities encompasses several attack vectors. Contract vulnerabilities (30.1% of all events) remain the single most common tactic, with access control flaws leading the way at 11.8%. Such flaws enable unauthorized entities to invoke privileged contract functions, for example when a smart contract fails to enforce proper access control before executing operations that allow the withdrawal of all funds [6].

Logical bugs (5.2%) exploit flawed business logic where code executes as written but contains fundamental design errors. Reentrancy attacks (3.2%) manipulate execution order, allowing attackers to withdraw funds multiple times before balance updates [1].

Beyond contract code, attackers exploit DeFi's interconnected architecture. The composability of protocols—the "money legos" phenomenon [16]—creates systemic vulnerabilities. Flash loan arbitrage (3.6%) and oracle manipulation (3.1%) leverage this interconnection [17]. When an oracle feeding price data to multiple lending protocols is compromised, attackers can simultaneously drain liquidity across the entire dependent ecosystem. This explains why Interface layer attacks (oracles and bridges), though representing only 6.78% of events, account for $2.5 billion in damages—compromising shared infrastructure enables cascading exploitation.

Human-risk attacks, while rare (3.1%), prove devastatingly effective. These include insider theft (1.5%), where employees abuse privileged access to private keys and external

social engineering (1.6%) targeting operational mistakes or deceiving personnel. The rarity might reflect reporting bias—companies rarely publicize falling victim to social engineering—but the financial impact is undeniable, as we'll see shortly.

DeFi Actor as Perpetrator (40.9% of events, 17% of damages)

Here, the DeFi actor itself is malicious from inception. The entity claiming to provide a service deliberately defrauds users through two primary strategies.

Malicious use of contracts (36.0%) involves developers coding hidden backdoors or performing unauthorized operations. The dominant form is the rug pull scam, manifesting as:

- *liquidity removal* (16.4%): suddenly withdrawing all tokens from the trading pool, making assets worthless and untradeable;
- *hidden mint functions* (3.6%): backdoors enabling unlimited token creation, diluting existing holders to zero;
- *selling restrictions* (4.8%): code preventing anyone except the developer from selling;
- *pump-and-dump schemes* (6.6%): artificially inflating prices through coordinated buying before dumping holdings [11].

Market manipulation (4.9%) employs deceptive tactics without malicious code. This includes Ponzi schemes where early investors are paid with new investors' capital [3], scam presales where promised projects never materialize and outright embezzlement of collected funds.

The prevalence of these events (41% of total) masks their relatively modest individual impact. The median loss per rug pull is $88,736, a figure that is substantial for victims but remains an order of magnitude smaller than the losses observed in sophisticated attacks targeting legitimate infrastructure. This pattern reflects the mechanics: rug pulls target less sophisticated users with limited capital, operating as numerous small-scale scams rather than singular catastrophic exploits.

DeFi Actor as Intermediary (7.0% of events)

In the third scenario, legitimate DeFi actors become unwitting intermediaries. Criminals don't exploit protocol code, they exploit protocol reputation through impersonation. Tactics include compromising social media accounts to post fake announcements (3.2%), DNS (Domain Name System) attacks redirecting official domains to phishing sites (1.3%),

creating lookalike websites through typosquatting and purchasing ads so fraudulent sites appear first in search results [15, 21].

The authors acknowledge this category is likely underrepresented. Unlike protocol exploits that leave clear on-chain traces and generate immediate headlines, individual phishing victims often lose smaller amounts and may not report incidents publicly. The aggregators used prioritize technically sophisticated exploits over social engineering attacks, creating systematic undercounting.

1.1.3. Quantifying Financial Impact: Where the Damage Concentrates

The taxonomy's quantitative analysis reveals striking disparities that aren't apparent from event counts alone. These findings directly inform our feature selection for the anomaly detection framework.

Comparing Roles: Targets Suffer Disproportionate Losses

When legitimate protocols are attacked, the median loss per incident vastly exceeds losses from fraudulent projects:

$$\tilde{x}_{\text{target}} = \$800,000 \gg \tilde{x}_{\text{perpetrator}} = \$88,736 \tag{1.1}$$

where $\tilde{x}$ denotes median financial damage.

This nine-fold difference is statistically significant (Mann-Whitney U: $U = 126,874$, $p < 0.001$, Cliff's $\delta = 0.446$, as documented in [4]), indicating a medium-to-large effect size that persists across various market conditions and time periods.

What drives this disparity? Sophisticated attacks on well-capitalized protocols extract far more value than retail-focused scams. A single flash loan attack might drain $50 million from a lending protocol in minutes. In contrast, a typical rug pull defrauds perhaps 100 investors of $1,000 each. Both are crimes but the scale differs dramatically. For regulatory resource allocation, this suggests prioritizing infrastructure security over solely focusing on investor protection from scams—though both matter, the former prevents larger aggregate losses.

Comparing Strategies: Human Vulnerabilities as High-Impact Risks

Within attacks on legitimate actors, we can distinguish technical exploitation from human-risk exploitation. Here the pattern reverses: human attacks, though rare, prove extraordinarily damaging.

A Kruskal-Wallis H test comparing median losses across all four strategy types (technical vulnerabilities, human risks, malicious contracts, market manipulation) showed significant differences: $H = 139.04$, $p < 0.001$, $\varepsilon^2 = 0.17$, as reported in [4]. This epsilon-squared value tells us that strategy type accounts for about 17% of the variance in financial losses, a substantial share given the influence of numerous other factors such as protocol scale, attacker sophistication and prevailing market volatility.

As reported in [4], Post-hoc pairwise comparisons using Dunn's test with Bonferroni correction revealed the specific pattern:

- Human risk exploitation yields highest median losses: $\tilde{x}_{\text{human}} = \$2,073,280$;
- Significantly exceeding technical exploitation: $\tilde{x}_{\text{technical}} = \$764,523$ ($p = 0.027$, $\delta = 0.333$);
- And far exceeding perpetrator strategies:
  $\tilde{x}_{\text{contract}} = \$76,018$ ($p < 0.001$, $\delta = 0.78$)
  $\tilde{x}_{\text{market}} = \$159,136$ ($p < 0.001$, $\delta = 0.638$).

Why this pattern? When attackers exploit technical vulnerabilities, they're constrained by the specific bug discovered. Maybe they can drain one liquidity pool or exploit one pricing oracle. Security researchers often detect ongoing attacks and implement emergency procedures—pausing contracts, alerting users or deploying patches.

But when attackers obtain private keys through social engineering or insider theft, they gain administrative control equivalent to legitimate operators. They can drain multiple pools, transfer treasury funds and execute complex multi-step attacks before detection. In this case, there is no underlying technical vulnerability to remediate, as the attacker operates using legitimately obtained credentials. This suggests DeFi protocols may over-invest in code audits ($100,000+ per audit) while under-investing in operational security: background checks, access controls, multi-signature requirements and security awareness training.

1.1.4. Mapping Vulnerabilities to the DeFi Technology Stack

To identify which infrastructure components face the greatest risk, the authors mapped incidents onto the DeFi Stack Reference (DSR) model [2]. This seven-layer architecture conceptualizes DeFi from System Infrastructure at the base (computational resources), through Distributed Ledger Technology (blockchain consensus and state), Cryptoassets (tokens), DeFi Protocols (financial services), DeFi Compositions (aggregated services) and Interfaces (oracles/bridges), to User Applications at the top.

When Legitimate Protocols Are Attacked In accordance with [4], the vulnerability distribution revealed clear patterns:

- **DeFi Protocols**: 54.46% of events, $2.6B losses;
- **Cryptoassets**: 17.44% of events, $1.8B losses;
- **DeFi Compositions**: 15.12% of events, $925M losses;
- **Interfaces**: 6.78% of events, $2.5B losses;
- **Distributed Ledger Technology**: 6.20% of events, $1.5B losses.

The concentration at the DeFi Protocols layer makes architectural sense. This is the layer in which financial logic is implemented, as smart contracts manage liquidity pools, determine interest rates, execute trades and oversee collateral mechanisms. These contracts control enormous capital (top protocols manage $1B+ in total value locked), involve complex code prone to subtle bugs and are publicly accessible for vulnerability research. It's the perfect attack surface.

Despite comprising only 35 recorded events, the Interfaces layer is responsible for $2.5 billion in cumulative damages, ranking second among all layers in terms of total financial impact. This reflects systemic vulnerability from shared infrastructure. Exploit a single protocol, you attack that protocol. Compromise an oracle or bridge, you attack every protocol depending on it. The August 2021 Poly Network bridge hack ($611M) and March 2022 Ronin Bridge exploit ($625M) both compromised cross-chain bridge infrastructure, enabling simultaneous drainage across multiple blockchains. This composability risk is one of DeFi's greatest architectural challenges.

Statistical testing confirmed these patterns. A Kruskal-Wallis H test yielded $H = 139.65$, $p < 0.001$, $\varepsilon^2 = 0.17$, , Post-hoc comparisons showed that Cryptoassets layer attacks cause significantly lower median damages ($83,960) than all other layers ($p < 0.001$ for all comparisons), according to [4]. This aligns with our earlier finding: token scams

(predominantly CA layer) are numerous but individually modest, while protocol and infrastructure exploits are rare but catastrophic.

When Fraudulent Projects Target Users The distribution shifted dramatically when examining malicious actors:

- **Cryptoassets**: 72.75% (overwhelmingly token rug pulls);
- **DeFi Protocols**: 18.01%;
- **DeFi Compositions**: 8.29%;
- **All other layers**: $< 1\%$.

This concentration at the CA layer reflects the typical mechanics of rug pulls. Attackers first create a custom token exploiting standard templates, a process that requires limited technical expertise. They then generate market interest through social media channels that operate largely outside regulatory oversight. Finally, liquidity is drained or insider token holdings are liquidated. The simplicity and scalability explain prevalence: one developer can run dozens of these scams simultaneously, each targeting different niche communities. Creating convincing fake lending protocols or derivatives platforms involves considerable effort. Once exposed, these schemes rarely allow for repeated exploitation which limits their scalability.

1.1.5. CeFi versus DeFi: A Comparative Lens

While focused on DeFi, the authors documented 105 CeFi incidents. Despite representing only 9.2% of events, CeFi crimes account for $19.5B in losses versus $10B for DeFi which is roughly two-thirds of total industry losses. This disparity stems from massive centralized frauds: Ponzi schemes ($7.3B) and embezzlement ($1.1B) like the 2022 FTX collapse ($8B+ in customer funds).

This challenges narratives positioning DeFi as uniquely vulnerable. Centralized custody creates concentration risk. When a CeFi platform is compromised or run by bad actors, all customer funds are simultaneously at risk. In DeFi, each protocol's vulnerability is largely independent (though composability creates some systemic linkages).

1.1.6. Key Findings and Implications

The taxonomy establishes three critical insights for our subsequent empirical work:

1. **Legitimate infrastructure bears disproportionate risk**: 52% of events tar-

geting DeFi actors cause 83% of damages. Securing existing protocols against sophisticated attacks should be the primary security priority. This concern is especially pronounced in DeFi, which does not benefit from the protective mechanisms common in traditional finance, such as deposit insurance or systematic regulatory oversight [2].

2. **Human vulnerabilities create high-impact risks**: Though representing only 3.1% of events, human-risk exploits produce median losses 2.7× higher than technical attacks. Organizations should implement robust access controls, multi-signature schemes, background checks and security training. These measures can be more impactful in preventing exploits than additional code audits alone.

3. **Technical layer predicts damage magnitude**: Interface and Protocol layers suffer far higher median losses than Cryptoassets. This layer-based risk stratification will inform our feature engineering, where we'll weight on-chain indicators differently based on which stack components they affect.

1.1.7. The Oxford Taxonomy as an Analytical Framework

The Oxford Taxonomy [4] serves as the conceptual bridge between qualitative criminological classification and quantitative empirical analysis. Rather than functioning solely as a descriptive framework, it provides a structured basis for testing whether observed DeFi crime patterns are statistically robust and economically meaningful across different market conditions.

First, the taxonomy enables a systematic validation of financial impact heterogeneity across attack characteristics. By grouping incidents according to actor roles, attack typologies and affected technical layers, it becomes possible to assess whether differences in economic damage are persistent features of DeFi crime or contingent on specific historical periods. Nonparametric statistical methods are well suited to this setting, given the heavy-tailed nature of loss distributions and the presence of extreme outliers. Controlling for protocol-level and market-level factors—such as total value locked, blockchain network, protocol maturity and contemporaneous market volatility—further allows disentangling structural risk patterns from transient market effects.

Beyond statistical validation, the taxonomy provides an operational lens for translating qualitative attack descriptions into observable on-chain behavior. Distinct exploit tactics, such as access control failures, reentrancy, oracle manipulation and liquidity removal, are associated with characteristic transaction sequences and interaction patterns at the smart contract level. These regularities allow the taxonomy to inform feature construc-

tion directly from raw blockchain data, grounding quantitative models in economically interpretable mechanisms rather than purely data-driven abstractions.

The explicit layering of attacks (Interface, Protocol and Cryptoassets) introduces a natural hierarchy for risk prioritization. Empirically observed differences in loss severity across layers motivate layer-sensitive indicators, whereby anomalies affecting higher-impact layers receive greater weight in detection and monitoring tasks. Similarly, distinguishing between actor roles supports role-specific modeling choices, reflecting the fundamentally different transaction dynamics of protocol exploitation versus insider-driven malfeasance.

In this way, the Oxford Taxonomy [4] functions not only as a classification tool but as a unifying structure that links economic impact analysis, feature engineering and anomaly detection within a coherent methodological framework. It anchors quantitative modeling choices in established criminological theory, facilitating a shift from purely retrospective analysis toward risk-aware, behavior-based assessment of DeFi security.

1.2. The Technical Perspective: Uniswap Scam Detection

While the Oxford Taxonomy [4] provides a criminological framework for understanding DeFi crime, we need complementary technical approaches that can operationalize these concepts through on-chain data analysis. Xia et al.[22] provide exactly this through their systematic study of scam tokens on Uniswap, the largest decentralized exchange. Their work is particularly relevant to our anomaly detection framework because it demonstrates how transaction-level features can distinguish malicious from legitimate activity.

1.2.1. The Uniswap Ecosystem and Scam Prevalence

Uniswap operates as an Automated Market Maker (AMM), replacing traditional order books with liquidity pools. In this system, users are able to create pools with any two ERC-20 tokens and token listings are unrestricted. The underlying design emphasizes open, permissionless access rather than compliance with regulatory controls. This architectural choice, while democratizing finance, creates an attack surface for scammers.

The authors collected all Uniswap V2 transactions from May 5 to December 6, 2020, capturing over 20 million transaction events involving 21,778 tokens and 25,131 liquidity pools. Using a hybrid detection approach combining guilt-by-association heuristics with machine learning, they identified 10,920 scam tokens (50.14% of all tokens) and 11,215

scam liquidity pools (44.63% of all pools). These scams caused at least $16 million in losses affecting 39,762 victims, as detailed in [22].

This striking prevalence, with roughly half of all tokens being scams, aligns with the Oxford Taxonomy [4]'s observation that DeFi actors account for 41% of reported crime events. Building on this, Xia et al.[22] demonstrate that detection based on behavioral features is both feasible and accurate.

1.2.2. Feature Engineering for Scam Detection

The authors' key insight is that scam tokens exhibit distinctive behavioral patterns across four feature categories. Understanding these features is crucial for our work because we'll adapt similar on-chain indicators for the anomaly detection model, as summarized in Table 1.1.

Time-Series Features (7 features)

Scam tokens and pools are characteristically short-lived. Once enough victims invest, scammers remove liquidity and abandon the project. The authors capture this temporal pattern using the features detailed in Category 1 of Table 1.1, specifically:

Active period (T_{period}): Time from first to last transaction on Uniswap. Scam tokens average under 24 hours active; legitimate tokens remain active for months.

Interval to present ($T_{interval}$): Time between last transaction and dataset collection. This controls for recency bias—a newly listed legitimate token would also have short T_{period} but active $T_{interval}$ distinguishes it from an abandoned scam.

Relative time position of events: For each event type (mint, swap, burn, swap_from, swap_to), the authors compute when these events concentrate in the token's lifecycle:

$$P_{event} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i - T_{start}}{T_{end} - T_{start}} \tag{1.2}$$

where n is the number of events of that type, T_i is the timestamp of the i-th event and T_{start}, T_{end} are the token's first and last transaction times.

For scam tokens, mint events cluster at the beginning ($P_{mint} \approx 0$)—scammers provide initial liquidity to attract victims. Burn events cluster at the end ($P_{burn} \approx 1$)—scammers remove liquidity to exit. Legitimate tokens show even distribution across their lifecycle. This metric directly operationalizes the "rug pull" pattern identified in the Oxford

Table 1.1: Features used in the scam token classifier by Xia et al.[22]

Category	Feature	Description
Time-series	T_{period}	The time interval from the first transaction to the last transaction on Uniswap
(7 features)	$T_{interval}$	The time interval between the last transaction and the study time on Uniswap
	P_{mint}	The time point of mint events in the whole token lifecycle on Uniswap
	P_{swap}	The time point of swap events in the whole token lifecycle on Uniswap
	P_{swap_from}	The time point of swap-from events (swap target token for other) in the lifecycle
	P_{swap_to}	The time point of swap-to events (swap other token for target) in the lifecycle
	P_{burn}	The time point of burn events in the whole token lifecycle on Uniswap
Transaction	N_{Tx_U}, N_{Tx_E}	Total transaction numbers on Uniswap and Ethereum, respectively
(24 features)	N_{mint}, N_{swap}	Total mint and swap event numbers on Uniswap
	N_{swap_to}	Total swap-to event numbers on Uniswap
	N_{swap_from}	Total swap-from event numbers on Uniswap
	RE_{sw_f/sw_t}	Ratio $N_{swap\_from}/N_{swap\_to}$ (set to -1 if N_{swap_to} is 0)
	N_{burn}	Total burn event numbers on Uniswap
	A_{mint}, A_{swap}	Total unique addresses that participated in mint or swap events
	A_{swap_to}	Total unique addresses that participated in swap-to events
	A_{swap_from}	Total unique addresses that participated in swap-from events
	A_{burn}, A_{all}	Total unique addresses for burn events or all events combined
	RE_{event_all}	Ratio of specific event counts (mint/swap/burn) to total Uniswap transactions
	RA_{event_all}	Ratio of unique addresses per event type to total involved addresses
Investor	$L_{mint/burn}$	Average liquidity pools of participants who minted or burnt on Uniswap
(4 features)	L_{swap}	Average liquidity pools of participants who swapped on Uniswap
	$C_{mint/burn}$	Average mint or burn event counts per participant on Uniswap
	C_{swap}	Average swap event counts per participant on Uniswap
Uniswap Specific	N_{pool}	The total number of liquidity pools the token is involved in
(5 features)	V_{token}	Total amount of tokens traded all time across pairs
	$V_{tracked}$	Amount of tokens in USD traded all time across pairs (tracked)
	$V_{untracked}$	Amount of tokens in USD traded all time across pairs (untracked)
	$N_{liquidity}$	Total amount of token provided as liquidity across all pairs

Taxonomy.

Transaction Features (24 features)

Transaction volumes and event distributions, listed in Category 2 of Table 1.1, reveal behavioral differences:

Event counts: Total mint, swap, burn events on Uniswap (N_{mint}, N_{swap}, N_{burn}) plus total transactions on Ethereum (N_{Tx_E}). Legitimate tokens have substantial Ethereum activity beyond Uniswap; scams often only exist on Uniswap.

Swap direction asymmetry: The ratio $R_{E_{swap\_from/swap\_to}} = N_{swap\_from}/N_{swap\_to}$ captures whether more users are buying (swap-to) or selling (swap-from) the token. For scams, this ratio heavily favors swap-to initially (victims buying), then suddenly flips when scammers dump holdings.

Address participation: Number of unique addresses involved in each event type (A_{mint}, A_{swap}, A_{burn}). Scam pools typically have few liquidity providers (often just the creator) but many swap participants (victims). Legitimate pools show balanced participation across event types.

Investor Features (4 features)

These features, found in Category 3 of Table 1.1, capture victim characteristics:

Average pools per investor ($L_{mint/burn}$, L_{swap}): How many different liquidity pools do participants interact with? Scam victims tend to be inexperienced users (few pools), while legitimate token users are often sophisticated DeFi participants (many pools).

Average transaction count ($C_{mint/burn}$, C_{swap}): How many transactions has each participant conducted? Again, scammers target newcomers with limited transaction history.

The Oxford Taxonomy [4] noted that malicious DeFi projects (perpetrator role) cause smaller individual losses because they "typically operate at smaller scales compared to sophisticated attacks on legitimate infrastructure." These investor features explain why: scams target unsophisticated users with limited capital.

Uniswap-Specific Features (5 features)

Platform-specific metrics, detailed in Category 4 of Table 1.1, complete the set:

Pool count (N_{pool}): Legitimate tokens typically pair with multiple assets (WETH, USDT, USDC). Scams usually have one pool (e.g., ScamToken-WETH).

Trading volume (V_{token}, $V_{tracked}$, $V_{untracked}$): Tracked volume uses Uniswap's official USD conversion for liquid tokens; untracked volume includes illiquid tokens. Scams often have high untracked volume (artificial price pumping) but low tracked volume.

Total liquidity ($N_{liquidity}$): Scam pools frequently have minimal liquidity ($10K), just enough to appear legitimate. Top legitimate pools hold $10M+.

While the aforementioned framework provides a holistic view of token behavior, its primary focus is on Uniswap-specific tokenomics. In the following chapters, we will distill these indicators into a more generalized set of on-chain features, suitable for a broader anomaly detection task targeting multiple protocol types.

1.2.3. Machine Learning Approach and Performance

Using the 40 features detailed in Table 1.1, Xia et al.[22] trained and evaluated multiple classifiers, including Logistic Regression, SVM, Random Forest, and XGBoost. Through 10-fold cross-validation, the authors determined that the Random Forest model achieved the superior performance metrics:

- **Precision**: 96.45% (few false positives);
- **Recall**: 96.79% (few false negatives);
- **F1 Score**: 96.62%.

This high accuracy demonstrates that behavioral patterns are sufficiently distinctive for reliable automated detection. Crucially, the model performs well even on tokens with limited transaction history—over 1,800 scam tokens were correctly identified with fewer than 10 transactions. This suggests the approach can serve as an "early warning system" flagging scams before they accumulate many victims.

To ensure dataset quality, the authors employed strict verification heuristics after ML classification:

1. **Identical name heuristic**: Group suspicious tokens by name/symbol. If multiple tokens share identical names and none are verified legitimate projects, flag all as scams. Example: 429 tokens named "yearn.finance" (YFI)—all scams since the real YFI has a known address.
2. **Impersonation heuristic**: Flag tokens using names of popular companies/celebrities without official token releases (e.g., "Google Token," "Elon Musk Coin").

Additionally, they applied **guilt-by-association expansion**: if an address created one

verified scam token, all other tokens from that address are flagged. This identified 2,448 additional scams (22% of total), demonstrating that scammers often run multiple campaigns.

1.2.4. Scam Mechanics: Rug Pulls and Contract Backdoors

Xia et al.[22]'s analysis of scam behaviors directly validates and extends the Oxford Taxonomy [4]'s "malicious use of contracts" category. They identified two primary rug pull variants:

Standard Rug Pulls All 11,215 scam pools exhibited the classic pattern: (1) create token and pool, (2) add initial liquidity (often substantial—60% of creators provided $10K+), (3) promote on social media, (4) wait for victims to swap into the pool (raising token price), (5) remove liquidity and abandon. Remarkably, 86% of scam pools had intervals under 24 hours between first mint and final burn; 37% completed the cycle within one hour.

In 93% of scam pools, swap transactions were conducted by the token or pool creator or by colluding addresses, creating an artificial price inflation phase. Some creators went further, performing *second-round scams* by adding liquidity back to the same pool hours later, targeting victims who had overlooked prior transactions.

Contract Backdoors Beyond simple liquidity removal, some scammers embedded malicious code:

Freely mint tokens (297 contracts): Backdoor functions allowing specific addresses to create unlimited tokens. For example, the Leopard Lending Ecology (LLE) token had a ‘mint()‘ function callable only by one address, with no event emission for tracking. That address minted 1.8×10^{12} tokens (exceeding total supply) and swapped them for 474 ETH ($284K).

Sale-restrict tokens (373 contracts): Code allowing anyone to buy but only the creator to sell. VIPSwap (VIP) token used Solidity modifiers permitting all users to purchase VIP but restricting sells to the contract owner. The creator deployed 58 such tokens, gaining 77 ETH ($46K).

Advance-fee tokens (63 contracts): Contracts charging fees on every transaction, ostensibly for "bonus pools" that never distribute rewards. For instance, piratetoken.finance (PIRATE) charged 5% fees to a declared "daily bonus" address that never sent payouts.

These techniques map directly to the Oxford Taxonomy [4]'s specific tactics within "malicious use of contracts": hidden mint functions (3.6%), selling restrictions (4.8%) and liquidity removal (16.4%).

1.2.5. Collusion Networks and Financial Impact

Xia et al.[22] went beyond token-level analysis to map **collusion networks**—multiple addresses controlled by the same scammer. Using transaction flow analysis, they identified 41,118 collusion addresses collaborating with 3,377 scam creators. These collusion addresses:

- Receive funds from known scam addresses before adding liquidity (to appear as independent investors)
- Swap valuable tokens for scam tokens to artificially pump prices
- Transfer profits back to scam creators after selling

This network analysis revealed the coordinated nature of scam campaigns, often involving dozens of addresses per scam. Total financial impact: $16 million stolen from 39,762 victims, with the most profitable single pool (certik.foundation counterfeit) netting $1.5M from 365 victims.

1.2.6. Relevance of the Xia et al. Framework for Anomaly Detection

The methodology proposed by Xia et al.[22] provides a technical foundation for operationalizing DeFi anomaly detection through on-chain data analysis:

Feature Engineering Foundation The 40 features validated in the study—specifically time-series positions (P_{mint}, P_{burn}), swap direction asymmetry, and investor experience metrics—demonstrate that malicious activity produces distinct statistical signatures. These features allow for the extraction of behavioral indicators from raw transaction data that distinguish rug pull events from legitimate protocol operations.

Temporal Dependency Analysis The finding that a significant portion of scams are executed with extreme speed—37% of pools having their liquidity removed within one hour—highlights the necessity of prioritizing short-term temporal dependencies in detection models. The study reveals that 86% of scam cycles are completed within a single day.

Identification of Collusion Networks Xia et al.[22] demonstrate that individual address behavior is often linked to broader malicious campaigns involving multiple "collusion addresses". By analyzing transaction flows between creators and secondary addresses, they identified 41,118 collusion addresses that facilitate price pumping and liquidity drainage. This suggests that an effective detection framework must capture structural patterns characteristic of these coordinated networks, such as addresses receiving funds from a creator shortly before adding liquidity to a pool.

Methodological Evolution While Xia et al.[22] successfully utilized supervised learning (Random Forest) to achieve a high F1 score of 96.62%, they acknowledge that such models inherently rely on historical labels and developed transaction history. Their research establishes a baseline for recognizing fraudulent patterns by integrating diverse on-chain signatures. This includes analyzing behavioral asymmetry in swap ratios to detect artificial price inflation and investor profiling to identify the targeting of inexperienced users. Furthermore, the study's focus on contract backdoor detection reveals how anomalous modifiers enable hidden minting or restricted selling. Ultimately, the integration of these signatures, from temporal event clustering to collusive money flows, provides a robust framework for identifying DeFi threats at their early stages.

1.3. Synthesis: Integrating Criminological Theory with Technical Detection

The Oxford Taxonomy [4] and Xia et al.[22]'s Uniswap study represent complementary perspectives on DeFi crime. The former provides a *conceptual framework* explaining what crimes occur, who perpetrates them and why certain types cause more damage. The latter provides *operational techniques* demonstrating that these crimes produce detectable on-chain signatures. Our thesis integrates these perspectives:

From Taxonomy to Features The Oxford Taxonomy [4] identifies attack tactics (access control flaws, reentrancy, oracle manipulation, rug pulls). Xia et al.[22] show these tactics manifest as behavioral patterns (event clustering, swap asymmetry, liquidity removal). We bridge this gap by mapping each taxonomic category to on-chain feature signatures.

From Static Classification to Temporal Validation The Oxford Taxonomy [4] analyzes 2017-2022 data; Xia et al.[22] analyze May-December 2020. Both are static snapshots. We will test whether their findings generalize to 2023-2025, addressing the

critical question: do attack patterns evolve faster than detection methods can adapt? If yes, continual model retraining is essential. If no, one-time training on historical data suffices.

From Empirical Analysis to Hybrid Modeling Building on these temporal and taxonomic foundations, our study translates statistical observations into a dual-model detection system. We employ Supervised Learning to capture the high-precision signatures of recurring frauds identified in the Oxford and Xia datasets. Simultaneously, we implement an unsupervised Anomaly Detection model to learn the latent distribution of normal DeFi activity. By mapping the conceptual categories of the taxonomy onto these models, we evaluate the potential of hybrid architectures to identify sophisticated attacks that lack clear historical labels, providing a benchmark for future real-time detection systems.

This integrated perspective, which combines criminological insights, empirical validation of behavioral features and predictive modeling, places our work at the intersection of social science and computer science. The next chapter details how we construct the dataset enabling this integration.

2 | Data Engineering and Pre-processing

2.1. Overview of Data Sources

This chapter establishes the three data sources that underpin the thesis: the historical Oxford taxonomy, the recent De.Fi incident extension, and the Google BigQuery normal transaction sample. It explains the distinct role each source plays in enabling statistical validation, temporal generalization, and anomaly detection.

2.1.1. The Three-Dataset Architecture

Building a comprehensive framework for DeFi crime detection requires integrating data from three complementary sources, each serving a distinct analytical purpose:

- The Oxford Taxonomy dataset [4] provides 904 expert-curated fraud events spanning 2017–2022, constituting the methodological backbone of the study: every incident has been manually reviewed and classified, making it the gold standard against which all subsequent analyses are anchored.
- The De.Fi database extends this temporal coverage into the 2023–2025 period, contributing 751 verified criminal events that allow us to test whether patterns identified in the Oxford data persist through a substantially different market environment.
- A sample of 10,000 normal Ethereum transactions extracted from Google BigQuery provides the *negative examples* that the fraud-only incident databases cannot supply: legitimate blockchain activity against which anomalous behavior can be contrasted.

The first two datasets thus provide **positive examples**—documented attacks spanning nine years of DeFi evolution, all labeled `is_scam=1`—while BigQuery provides the **negative examples** that establish behavioral baselines, labeled `is_scam=0`. This complementarity is not incidental but structural: statistical hypothesis testing requires only the labeled fraud data, whereas the supervised and unsupervised anomaly detection models

Table 2.1: Overview of data sources and their analytical roles

Source	Coverage	Events	Label	Primary Use
Oxford Taxonomy	2017–2022 Zenodo 14047933	904 DeFi events (from 1,141 total)	`is_scam=1`	Ground truth labels Statistical validation
De.Fi API	2023–2025 `de.fi/rekt`	751 verified (from 756 total)	`is_scam=1`	Temporal extension Pattern stability
Google BigQuery	2017–2025 `crypto_ethereum`	10,000 txs	`is_scam=0`	Normal baseline Anomaly detection

require both fraud and normal samples to learn meaningful boundaries.

2.1.2. Final Integrated Dataset

After preprocessing and harmonization, detailed in the sections that follow, the three sources are merged into a unified analytical dataset. The 904 Oxford events and 751 De.Fi events together yield 1,655 documented fraud incidents, representing the positive class. Combined with the 10,000 BigQuery control transactions, the final dataset comprises **11,655 samples** spanning January 2017 through December 2025. The resulting fraud rate of 14.2% intentionally oversamples attacks relative to their real-world prevalence (estimated below 0.01% of all Ethereum transactions), providing sufficient positive examples for supervised learning. Realistic evaluation conditions are maintained through a strict temporal train/test split that prevents any form of data leakage between training and evaluation periods.

Table 2.1 summarises the role of each source within this pipeline.

2.1.3. Role of Each Data Source

The Oxford and De.Fi datasets share a common function: they supply labeled fraud events that make statistical hypothesis testing and supervised classification possible. Between them, they cover the full arc of DeFi's development—from the earliest documented exploits in 2017 through the recent period of institutional adoption and increasingly sophisticated attacks. Their difference lies in provenance: Oxford events were classified by domain experts through systematic inductive coding, while De.Fi events are processed through a mapping pipeline, a distinction with direct consequences for which analyses can legitimately draw on each source.

BigQuery plays a structurally different role. Incident databases by design contain only

anomalies—there is no record of the millions of benign transactions that constitute normal DeFi usage. Without a representation of normality, unsupervised models cannot learn what "unusual" means, and supervised models risk learning to detect superficial incident-report characteristics rather than genuine on-chain behavioral signatures. The BigQuery sample addresses this gap, providing gas consumption patterns, temporal distributions, and actor-level transaction frequencies drawn from ordinary Ethereum activity. This multi-source strategy therefore balances breadth with depth: the labeled fraud datasets offer semantic richness across diverse attack vectors, while BigQuery supplies the transaction-level granularity necessary for behavioral modeling. The sections that follow detail each source's structure, preprocessing steps, and integration into the unified analytical framework.

2.2. Dataset 1: Oxford Taxonomy

We introduce the Oxford dataset as the reference crime taxonomy for this study, detailing its academic provenance, the DeFi-specific filtering applied, and its role as the ground-truth positive sample for downstream statistical and computational analyses.

2.2.1. Source and Academic Context

The Oxford Taxonomy [4] represents the most rigorous academic effort to systematically catalog and classify DeFi crime. Carpentier-Desjardins et al. [4] collected incidents from three major aggregators—De.Fi REKT [8], SlowMist Hacked [19], and ChainSec [5]—and manually reviewed each event to assign taxonomic labels across four dimensions: actor implication (who was involved), strategy (high-level attack approach), general tactic (broad method), and specific tactic (precise technique).

This manual classification process distinguishes the Oxford dataset from automated scrapers or crowdsourced databases. Each of the 1,141 original events underwent expert review to ensure consistent categorization, making it the gold standard for DeFi crime research. The dataset spans January 2017 through December 2022, effectively capturing DeFi's evolution from niche experimentation through mainstream adoption and the 2021–2022 boom-bust cycle.

For our analysis, we apply two filtering steps. First, we exclude centralized finance (CeFi) events to focus exclusively on decentralized protocols, reducing the sample from 1,141 to 1,036 events. Second, we remove events lacking financial loss data, which are unsuitable for our statistical analysis. This yields our final Oxford subset of **904 DeFi events**

Table 2.2: Oxford Taxonomy dataset schema (key variables)

Variable	Type	Description
`unique_key`	Integer	Event identifier
`DeFi actor`	String	Protocol/token name
`Event date`	DateTime	Incident date
`Stolen USD`	Float	Financial loss (USD)
`General tactic`	String	Broad attack method
`Specific tactic`	String	Precise technique
`Strategy`	String	High-level approach
`Implication`	String	Actor role (Target/Perpetrator/Intermediary)
`Stack category`	String	DeFi Stack Reference layer

(79.2% of the original total), all with complete financial information.

Crucially, all 904 events are labeled `is_scam=1`, providing ground truth positive examples for supervised machine learning. These 904 events document over $10 billion in cumulative losses, establishing this as a comprehensive historical record of DeFi crime during its formative period.

2.2.2. Structure and Key Variables

The dataset contains 18 variables capturing both incident metadata and taxonomic classifications. While the original schema includes secondary fields designed to provide qualitative context and supplementary information, our analysis focuses on a subset of 9 core variables that carry high discriminative power for statistical and computational modeling. Table 2.2 presents these primary fields, which form the structural backbone of our integrated dataset.

The `Implication` field proves particularly valuable for distinguishing fundamentally different crime types. It separates events where a legitimate protocol was targeted by external attackers from cases where the protocol itself was the perpetrator of fraud (rug pulls, exit scams). This distinction carries profound implications for financial impact, as we'll validate statistically.

The `Strategy` classification captures six high-level attack approaches: technical vulnerability exploitation (smart contract bugs), malicious contract deployment (fraudulent tokens designed to steal from users), market manipulation (oracle attacks, pump-and-dump schemes), human risk (phishing, social engineering, compromised keys), imitation (fake websites, impostor contracts), and cases where the methodology remains undetermined. Each strategy exhibits distinct financial impact patterns that contribute to the broader

Table 2.3: Financial loss distribution for Oxford dataset (2017–2022, N=904)

Statistic	Value (USD)
Total cumulative losses	$9.97 billion
Mean	$11,028,726
Median	$256,915
Standard deviation	$61,065,745
Minimum	$158
Maximum	$1,000,000,000
Skewness	10.68
Kurtosis	132.55

empirical investigation conducted in this study.

Finally, the `Stack category` field maps incidents onto the DeFi Stack Reference (DSR) model [2], classifying which architectural layer was compromised: Protocols (P), Cryptoassets (CP), DeFi Compositions (DC), Interfaces (INT), or Distributed Ledger Technology (DLT). This layered perspective proves crucial for understanding damage magnitude—as we'll show, attacks on the Interface layer (bridges and oracles) cause dramatically higher median losses than Cryptoasset-level incidents.

2.2.3. Financial Distribution: Extreme Heterogeneity

The 904 DeFi events exhibit financial characteristics that immediately justify our methodological choices throughout the thesis. Table 2.3 presents the distribution statistics.

The distribution is severely right-skewed with skewness exceeding 10, driven by catastrophic outliers like the BeautyChain exploit ($1 billion). The mean-to-median ratio of 43× reveals that a small number of mega-exploits dominate total losses despite representing a tiny fraction of incidents. The extreme kurtosis (132.55) indicates even fatter tails than a standard log-normal distribution, with "black swan" events occurring far more frequently than Gaussian assumptions would predict.

This heterogeneity has immediate methodological consequences. Classical statistical methods that rely on normality assumptions, such as Student's t-tests, ANOVA and linear regression, become unreliable when skewness exceeds 2 and kurtosis exceeds 7. In our data, skewness reaches 10.68 and kurtosis 132.55, placing the distribution far outside the range in which parametric inference is considered valid. This motivates the adoption of non-parametric methods, including the Mann–Whitney U [13] and Kruskal–Wallis H tests [12], which require minimal distributional assumptions.

Similarly, raw financial data requires appropriate transformation to ensure compatibility with machine learning algorithms. The 6.3-million-fold range from minimum ($158) to maximum ($1B) would cause gradient explosion in neural networks and bias in tree-based algorithms. We therefore apply log-transformation to compress this scale while preserving ordinal relationships.

2.3. Dataset 2: De.Fi Database

The 2023–2025 De.Fi incident dataset is presented here to show how it supplements the Oxford sample, enabling true out-of-sample temporal validation of previously identified patterns.

2.3.1. Extending Temporal Coverage

The Oxford dataset's December 2022 endpoint creates a 2.5-year gap to our analysis date in early 2025. While frustrating from a data availability perspective, this gap provides a valuable methodological opportunity: it enables true out-of-sample validation of whether the patterns Carpentier-Desjardins et al. [4] identified in 2017–2022 persist in more recent data.

The intervening period witnessed significant ecosystem changes that could plausibly alter crime dynamics. The 2022 crypto market crash saw total market capitalization plummet from $3 trillion to $800 billion, potentially changing attacker incentives. The November 2022 FTX collapse shifted user attention toward DeFi decentralization benefits, possibly attracting both new legitimate users and new attack vectors. Major protocol hacks in 2023 (Euler Finance $196M, Multichain $231M) prompted industry-wide security improvements. Would these shifts fundamentally alter the criminological patterns, or would structural characteristics of DeFi architecture ensure pattern persistence?

To answer this question, we extracted additional events from the De.Fi database, accessed through their public GraphQL API in January 2025. We filtered for events dated 2023-01-01 or later and flagged as `verified:true`, ensuring cross-reference with on-chain evidence rather than relying solely on social media reports. This returned 756 verified events, all of which passed our DeFi classification criteria without requiring further filtering; within this sample, 751 events provided complete financial information. Importantly, we found zero duplicates when comparing against Oxford 2022 events, confirming clean temporal separation.

Table 2.4: Financial loss distribution: De.Fi dataset (2023–2025, N=751)

Statistic	Value (USD)
Total cumulative losses	$12.68 billion
Mean	$16,884,207
Median	$313,000
Standard deviation	$208,907,098
Minimum	$3,500
Maximum	$5,500,000,000
Skewness	24.58
Kurtosis	637.43

2.3.2. Financial Characteristics and Temporal Comparison

The 2023–2025 period exhibits both continuities and notable shifts relative to Oxford 2017–2022 baseline. Table 2.4 presents the distribution statistics.

Several patterns stand out. First, total losses actually *increased* despite the shorter time span: $12.68 billion over 3 years (2023–2025) versus $9.97 billion over 6 years (2017–2022). This suggests either accelerating attack sophistication or growing DeFi adoption creating larger targets. The median rose modestly from $257K to $313K (21.8% increase), while the mean jumped from $11.03M to $16.88M (53% increase), indicating more frequent medium-scale exploits alongside continued catastrophic outliers.

Second, the tail behaviour became even more extreme. Skewness more than doubled from 10.68 to 24.58, and kurtosis increased nearly fivefold from 132.55 to 637.43. This reflects the presence of the Mantra exploit ($5.5 billion)—the largest single DeFi crime event on record, substantially exceeding even BeautyChain's $1 billion. The appearance of such extreme outliers in a shorter, more recent period challenges any notion that DeFi security is meaningfully improving. While defenses may be advancing, attack capabilities appear to be advancing faster, at least at the high end of the damage distribution.

Third, comparing medians between periods (Oxford $257K vs. De.Fi $313K) shows only modest change, while means diverged substantially ($11.03M vs. $16.88M). This pattern suggests that typical attacks maintained similar magnitude, but the tail became heavier with more mega-exploits. From a risk management perspective, this is concerning—it indicates that while routine fraud hasn't intensified, catastrophic risk has actually worsened.

Table 2.5: Field mapping from De.Fi API to Oxford taxonomy

De.Fi Field	Oxford Field	Transformation
`id`	`unique_key`	Direct mapping
`projectName`	`DeFi actor`	Direct mapping
`date`	`Event date`	ISO format parsing
`fundsLost`	`Stolen USD`	Numeric conversion
`scamCategory`	`General tactic`	Manual mapping (Table 2.6)
—	`Strategy`	Inferred mapping (Table 2.8)
—	`Implication of actor`	Logical inference
—	`Stack category`	Inferred mapping (Table 2.9)

2.3.3. Schema Alignment Between Sources

The De.Fi database uses a practitioner-oriented schema optimized for rapid incident logging, which differs structurally from Oxford's academic taxonomy. We therefore needed to map De.Fi fields onto Oxford's variable structure to enable dataset integration. Table 2.5 documents our mapping strategy.

Most mappings proved straightforward, with event identifiers, project names, dates and loss amounts transferred directly with minimal transformation. The challenging mapping involved `scamCategory` to `General tactic`, where De.Fi uses attack-oriented labels ("Rug Pull", "Flash Loan Attack", "Bridge Exploit") while Oxford employs criminological categories ("Rug pull scam", "Transaction attack", "Contract vulnerability"). To reconcile these differences, we defined a set of manual translation rules that harmonize platform-specific labels into the Oxford taxonomy [4]; the complete mapping is reported in Table 2.6.

2.3.4. Validation of the Mapping Logic

To ensure the robustness of our taxonomic alignment, we extracted a secondary validation dataset from the De.Fi REKT database covering the same temporal window as the Oxford dataset (2017–2022). By performing an inner join based on normalized project names and event dates, we identified 506 unique events common to both sources.

The validation focused on two critical dimensions:

- **Financial Accuracy**: Comparison of reported stolen amounts yielded a Pearson correlation coefficient of $r = 0.995$ ($p < 0.0001$), confirming near-perfect agreement on quantitative loss figures between the two independent data sources.

Table 2.6: Mapping between De.Fi original labels and Oxford harmonized categories (validated at 75.5% accuracy)

Original De.Fi Label	Mapped Oxford Category
Smart Contract Vulnerabilities Access Control, Reentrancy, Logic Error, Unsafe Delegatecall, Integer Overflow/Underflow, Incorrect Calculation, Bad Randomness, Signature Replay, Delegate Call, Visibility Error, Other	Contract vulnerability
Transaction-Level Attacks Front Running, MEV, Sandwich Attack, Transaction Order Dependence	Transaction attack
Cross-Protocol Exploits Oracle Issue, Price Oracle Manipulation, Flash Loan Attack, Oracle Manipulation, Price Manipulation, Market Manipulation	Interconnected actors flaw
Infrastructure Compromises Infrastructure Attack, Server Compromise, DNS Attack, Hot Wallet Compromise, Private Key Leak, Wallet Hack	Hacked/exploited infrastructure
Social Engineering & Phishing Phishing, Social Engineering, Fake Website, Impersonation	Instant user deception
Exit Scams & Fraud Rugpull, Exit Scam, Honeypot, Abandoned, Scam	Rug pull scam
Internal Malfeasance Fund Misuse, Improper Use, Unauthorized Transfer Insider Theft, Embezzlement, Rogue Employee	Misappropriation of funds Internal theft
Consensus & Governance Attacks Governance Attack, Centralization Risk, 51% Attack, Consensus Attack	Decentralization issue
External Events Regulatory, Market Crash, Delisting, Legal Issues	External factor
Unclassified Unknown, Not Specified, Unclear	Undetermined

Table 2.7: Mapping validation accuracy by General Tactic (506 overlapping events, aggregated to Oxford 6-level)

Oxford General Tactic	Match Rate	Event Count
Rug pull scam	99.6%	274
Contract vulnerability	78.8%	104
Interconnected actors flaw	69.7%	33
Instant user deception	23.5%	17
Overall weighted average	**75.5%**	**506**

- **Categorical Consistency**: After applying our mapping rules (Table 2.6) to align De.Fi's original labels with the Oxford taxonomy [4], we achieved an exact category match rate of **75.5%** at the granular tactic level.

Performance by Category

Table 2.7 presents match rates stratified by Oxford category. Three categories exhibit exceptional accuracy.

The near-perfect accuracy for *Rug pull scam* (99.6%, 273/274 matches) reflects clear definitional consensus between taxonomies. *Contract vulnerability* achieves 78.8% accuracy despite being the second-most frequent category, indicating robust mapping for technical exploits.

Analysis of Mismatches

The remaining 24.5% of discordances (124 events) primarily stem from two sources:

1. **Taxonomic granularity differences**: The most frequent mismatch (35 cases) involves events classified as *Misappropriation of funds* by Oxford but labeled *Rug pull scam* by De.Fi. This reflects a semantic boundary dispute—Oxford distinguishes between improper fund usage (misappropriation) and complete project abandonment (rug pull), while De.Fi applies "rug pull" more broadly to any exit scam.

2. **Multi-vector exploit ambiguity**: De.Fi categorizes 19 *Flash loan attacks* as *Interconnected actors flaw* (emphasizing cross-protocol manipulation), while Oxford classifies them as *Contract vulnerability* (emphasizing the exploited code weakness). Both perspectives are technically valid—the 69.7% match rate for this category reflects this inherent ambiguity rather than mapping error.

3. **Infrastructure vs. contract conflation**: De.Fi labels 18 cases of infrastructure

compromise (hot wallet breaches, private key leaks) as *Contract vulnerability*, reflecting their practice of aggregating any blockchain-layer exploit under a unified "smart contract" umbrella. Our mapping preserves Oxford's finer distinction between code vulnerabilities and operational security failures.

Implications for Data Quality

The 75.5% exact match rate, combined with near-perfect financial correlation ($r = 0.995$), validates three critical conclusions:

- The mapping preserves the **economic signal**—categorization discrepancies do not systematically bias loss magnitude estimates.
- High-frequency categories (*Rug pull scam*, *Contract vulnerability*) exhibit strong agreement, ensuring reliable representation of the dominant threat vectors.
- Remaining mismatches reflect legitimate taxonomic ambiguity (e.g., whether flash loan attacks are cross-protocol or contract-layer exploits) rather than systematic mapping failures. These edge cases constitute <4% of total events.

Despite minor terminological variations, the 75.5% granular concordance—rising to an estimated 90%+ when considering broader strategic categories (Technical vulnerability vs. Human risk vs. Imitation)—confirms that our mapping effectively captures the underlying nature of DeFi security incidents across heterogeneous data sources.

2.3.5. Strategic Alignment and Logical Inference

To achieve full structural parity with the Oxford dataset, we implemented a secondary mapping layer for the `Strategy` dimension and developed an inference logic for metadata that the De.Fi API does not natively provide (`Implication of actor` and `Stack category`).

While `General tactic` focuses on the technical method, the `Strategy` field captures the high-level criminological approach. Table 2.8 illustrates the deterministic mapping used to align recent events with the academic baseline.

Additionally, we addressed the absence of actor-role and architectural-layer labels in the De.Fi database by implementing an inference engine based on the taxonomic definitions in Carpentier-Desjardins et al. [4]:

- **Implication of Actor:** We automatically assigned the label *Perpetrator* to events involving *Rug pull scams* or *Internal theft*, where the protocol owners themselves

Table 2.8: Strategic mapping between General Tactic and Strategy categories

General Tactic (Input)	Oxford Strategy (Mapped)
Contract vulnerability	Technical vulnerability
Interconnected actors flaw	Technical vulnerability
Hacked/exploited infrastructure	Technical vulnerability
Decentralization issue	Technical vulnerability
Transaction attack	Market manipulation
Instant user deception	Human risk
Internal theft	Human risk
Misappropriation of funds	Human risk
Rug pull scam	Imitation
External factor	Undetermined
Undetermined	Undetermined

Table 2.9: Mapping of General Tactics to DeFi Stack Reference (DSR) Layers

General Tactic	DSR Code	Architectural Layer
Rug pull scam	CA	Cryptoassets
Interconnected actors flaw	CP	DeFi Compositions
Instant user deception	INT	Interfaces
Contract vulnerability	P	Protocols
Transaction attack	P	Protocols
Undetermined / Other	U	Undetermined

execute the crime. Conversely, *Instant user deception* was mapped to *Intermediary*, consistent with the Oxford Taxonomy definition where legitimate platforms are unwittingly used as vectors to reach users, while technical exploits were labeled as *Target*, identifying the protocol as the direct victim of the attack.

- **Stack Category:** Following the DeFi Stack Reference (DSR) model [2], we mapped each tactic to its respective architectural layer. Table 2.9 documents this mapping, reflecting where the exploit or fraud primarily materializes within the protocol architecture.

This standardized preprocessing pipeline ensures that the 2023–2025 data is not merely a collection of incidents, but a structurally identical extension of the Oxford Taxonomy, allowing for the rigorous comparative statistical analysis presented in the following chapters.

2.4. Dataset 3: Google BigQuery Ethereum Transactions

The control dataset of normal Ethereum transactions is described below, with rationale for the BigQuery sampling strategy and its role in training and evaluating anomaly detection models.

2.4.1. Motivation and Sampling Approach

The crime incident datasets provide rich labels and context but lack the granular transaction data needed for behavioural anomaly detection. We know that an attack occurred and what methodology was used but we don't observe the sequence of blockchain operations, gas consumption patterns, or temporal clustering that might enable early warning systems to detect attacks in progress.

To address this gap, we queried Google BigQuery's public Ethereum blockchain archive which contains every transaction ever executed on Ethereum mainnet, currently exceeding 2 billion records and growing by roughly 1.2 million transactions daily. Querying and processing the complete dataset would entail prohibitive computational and temporal overhead, rendering exhaustive full-table scans impractical for the requirements of this study.

We therefore adopted a random sampling strategy, extracting 10,000 transactions uniformly distributed across our study period (January 2017 through December 2025). Our SQL query applied three filters:

1) `receipt_status = 1` to exclude failed transactions;
2) `value > 0` to filter pure function calls that don't transfer ETH;
3) `ORDER BY RAND()` to ensure temporal uniformity rather than clustering in any particular period. The resulting 10,000-transaction sample spans the same temporal range as our crime datasets, enabling fair feature comparisons.

The complete SQL query, post-processing code, and protocol mapping dictionary are provided in Appendix A.1.

2.4.2. Currency Conversion and Value Normalization

After extraction, we converted ETH amounts to USD using historical exchange rates via the Yahoo Finance API [23]. For each transaction, we matched the block timestamp to

the corresponding daily ETH/USD closing price, computing:

$$\text{Value}_{\text{USD}} = \text{Value}_{\text{ETH}} \times \text{ETH Price}_{\text{date}} \tag{2.1}$$

This conversion enables direct comparison with the USD-denominated crime event losses from Oxford and De.Fi datasets. The resulting distribution exhibits a median transaction value of approximately $85 and maximum of $7.6 million—yielding a 71× mean-to-median ratio typical of Ethereum network activity, where most transactions involve small retail transfers but occasional whale movements or institutional operations heavily skew the distribution rightward (mean = $6,022, standard deviation = $111,169).

2.4.3. Protocol and Actor Identification

Raw Ethereum transactions contain only cryptographic addresses (`from_address`, `to_address`), which are meaningless hexadecimal strings without contextual mapping. To enable interpretable analysis, we constructed a mapping dictionary identifying 29 major DeFi protocols and centralized exchange hot wallets based on publicly documented addresses from Etherscan and protocol documentation.

Our mapping categorizes recipients into three tiers:

- **DEX Protocols** (e.g., Uniswap V2/V3 Routers, 1inch Aggregator, SushiSwap): 662 transactions (6.6%)
- **Centralized Exchange Hot Wallets** (Binance, Coinbase, Bittrex): 380 transactions (3.8%)
- **NFT Marketplaces** (OpenSea Seaport, Wyvern Exchange, Gem.xyz): 213 transactions (2.1%)
- **Infrastructure Contracts** (WETH, Aave Lending Pool, Lido stETH): 161 transactions (1.6%)
- **Private Wallets/Minor Protocols**: 8,584 transactions (85.8%)

This classification reveals that the vast majority of sampled transactions (85.8%) involve direct wallet-to-wallet transfers or interactions with long-tail protocols not captured in our mapping. The identified protocols represent high-liquidity, high-visibility targets that are overrepresented in attack datasets, making their presence in the control sample a useful reference point for distinguishing normal protocol usage from exploitation patterns.

2.4.4. Gas Consumption as a Behavioral Feature

Ethereum transactions consume computational resources measured in *gas units*, with `receipt_gas_used` recording the actual gas consumed during execution. Simple ETH transfers consume exactly 21,000 gas, while complex smart contract interactions may consume hundreds of thousands of gas units, making this metric a natural proxy for operational complexity. In our sample ($N = 10,000$), 74.9% of transactions correspond to simple EOA-to-EOA transfers, while 17.3% exceed the 100,000 gas threshold associated with high-complexity DeFi operations. The distributional structure of this feature and its implications for anomaly detection are examined in detail in the next chapters.

2.4.5. Labeling Strategy and Assumptions

A critical methodological decision concerns how to label these 10,000 BigQuery transactions. We assign `is_scam = 0` to all samples, classifying them as "normal" under the assumption that random Ethereum transactions are overwhelmingly legitimate. This assumption rests on base rate statistics: our 1,655 documented crime events spanning 9 years represent a tiny fraction—likely less than 0.01%—of the billions of Ethereum transactions executed during that period. Sampling 10,000 transactions randomly should therefore capture almost exclusively normal activity.

This assumption does introduce label noise. Statistically, we would expect roughly one attack transaction in our 10,000-sample (0.01% contamination rate) and that attack may even have gone unreported in our crime datasets. However, this minimal contamination is acceptable for unsupervised learning, where algorithms expect some noise in the training data. For supervised learning, a single mislabeled example among 10,000 would cause negligible performance degradation ($< 0.01\%$ theoretical accuracy loss under worst-case misclassification).

More subtly, we acknowledge that some legitimate transactions may exhibit unusual patterns that resemble attacks (large value transfers, complex gas consumption, unusual temporal patterns) despite being benign. This creates a spectrum from "obviously normal" to "suspicious but legitimate" to "clearly malicious", rather than a clean binary partition. Our models must learn to distinguish attacks from this complex normal distribution, not just from trivial baseline transactions—a more realistic and challenging detection scenario that better reflects real-world deployment conditions.

2.4.6. Taxonomic Alignment with Crime Datasets

To enable unified analysis across all three datasets, we assign taxonomic labels to the BigQuery control sample consistent with the Oxford framework applied to Oxford and De.Fi datasets:

- **Strategy**: `Normal Transaction` (newly introduced category for legitimate activity)
- **General Tactic**: `Legitimate Operation` (contrasts with exploit tactics)
- **Implication of Actor**: `Intermediary` (protocols facilitate, don't perpetrate)
- **Stack Category**: `Undetermined` (control-sample transactions are predominantly EOA-to-EOA transfers not attributable to any specific DeFi stack layer)
- **Target Label**: `is_scam = 0` (supervised learning ground truth)

This labeling scheme ensures that downstream machine learning models can process all three datasets within a unified feature space, enabling both supervised classification (`is_scam` as target) and unsupervised anomaly detection (where unlabeled BigQuery transactions serve as the normality baseline against which crime events are contrasted).

The final cleaned dataset retains 10,000 observations with 10 features: `DeFi actor involved`, `Event date`, `Event year`, `Stolen amount USD`, `Gas_normalized`, `Strategy`, `General tactic`, `Implication of actor`, `Stack category`, and `is_scam`.

2.5. Data Integration and Feature Engineering

We describe the sequential preprocessing pipeline that harmonizes heterogeneous schemas, merges the three sources into a single analytical dataset, and engineers the unified feature representation used for both statistical inference and machine learning.

2.5.1. Merging Strategy

Combining our three data sources into a unified analytical dataset required careful sequential processing. We proceeded through four steps: CeFi filtering, schema harmonization, concatenation and BigQuery integration.

First, we filtered the Oxford dataset to focus exclusively on decentralized finance. The original 1,141 events included some centralized exchange hacks and traditional financial fraud that, while cryptocurrency-related, don't reflect DeFi-specific attack patterns. Applying the `Paper category` filter removed these CeFi incidents, reducing the sample

from 1,141 to 1,036 events. We then excluded 132 additional events lacking financial data, yielding our final 904 DeFi events with complete information.

Second, we harmonized the De.Fi schema to match Oxford 9-variable structure using the mappings documented in Table 2.6, in Table 2.8 and in Table 2.9. This involved straightforward type conversions (dates, numeric amounts) and our manual categorical translations. We verified no temporal overlap existed by comparing event dates and actor names—the 904 Oxford events end in December 2022, while the 751 De.Fi events begin in January 2023, with zero duplicates detected.

Third, we vertically concatenated Oxford and De.Fi into a merged crime events dataset: $904 + 751 =$ **1,655 crime events** spanning 2017 through 2025. This becomes our labeled attack sample for both statistical analysis and supervised machine learning.

Fourth, we appended the 10,000 BigQuery transactions, all labeled `is_scam = 0`, creating our presumed-normal sample. The final integrated dataset contains **11,655 samples**: 1,655 attacks (14.2%) and 10,000 normal transactions (85.8%). This 14.2% attack rate far exceeds real-world base rates (where attacks constitute <0.01% of transactions), but this oversampling of attacks is intentional—it provides sufficient positive examples for supervised learning while maintaining realistic test set proportions.

2.5.2. Feature Engineering Pipeline

Raw categorical variables and heavily skewed numerical distributions require transformation before statistical analysis or machine learning. We engineered 26 features spanning five categories: temporal, transaction, financial, strategy indicators and stack/implication indicators.

Temporal Features

Following Xia et al. [22], we computed three temporal features designed to capture protocol lifecycle characteristics. The most important is `Lifespan_days`, measuring the time elapsed between the first and last recorded crime event involving the same DeFi actor. Short lifespans (measured in hours or single days) strongly correlate with rug pull scams, where fraudulent tokens are created, hyped, and abandoned within 24 hours. Longer lifespans (months to years) suggest either legitimate protocols that eventually fell victim to attacks or sophisticated long-con schemes that accumulate victims over time before exit.

We also computed `P_relative`, a continuous variable in [0,1] indicating where an event

falls in DeFi's timeline—0 for the earliest 2017 events, 1 for the latest 2025 events. This controls for temporal trends, as attack sophistication and defense mechanisms both evolve over time. Finally, `Days_since_start` captures absolute temporal position as an integer count, useful for models that might benefit from treating time as a continuous rather than relative variable.

Financial Features and Log-Transformation

The extreme right skew in financial losses (skewness 10.68 for Oxford, 24.58 for De.Fi) necessitates transformation. We apply the log1p transformation: `Log_stolen_amount` = $\log(1 + \texttt{Stolen USD})$, which compresses the 6-million-fold range from minimum to maximum while preserving zero values. Post-transformation skewness drops to approximately 0.8, approaching the normality required for stable neural network training.

We also engineer protocol-level aggregate features, specifically `Total_stolen`, `Avg_stolen`, and `Max_stolen`, by summing, averaging, and maximizing losses across all events involving the same DeFi actor. These capture whether a protocol has been repeatedly targeted (indicating persistent vulnerabilities) versus involved in a single incident (suggesting opportunistic attack or one-time fraud).

Categorical Encoding

For statistical tests, we retain the original categorical variables (`Strategy`, `Stack category`, `Implication`), as non-parametric methods operate directly on grouped data. To handle categorical data, we apply one-hot encoding, transforming each level into a binary indicator variable. This expands `Strategy` (7 levels) into 7 binary features, `Stack category` (6 levels) into 6 features and `Implication` (3 levels) into 3 features, yielding 16 additional dimensions.

Transaction Features from BigQuery

From the blockchain data, we extract two behavioural features. `Gas_normalized` scales each transaction's gas consumption by the maximum observed in our sample, creating a [0,1] continuous variable indicating computational complexity. Malicious contracts often exhibit bimodal gas distributions—very low for simple theft mechanisms, very high for complex reentrancy exploits. `Actor_tx_count` tallies how many transactions originated from each `from_address`, profiling whether actors are one-time participants or prolific users.

Table 2.10 summarizes the complete 26-feature representation.

Table 2.10: Complete feature set for analysis (N=26)

Category	Count	Features
Temporal	3	Lifespan days, $P_{relative}$, Days since start
Transaction	2	Gas_normalized, Actor_tx_count
Financial	5	Log_stolen_amount, Total_stolen, Avg_stolen, Max_stolen, Unique_investors
Strategy (one-hot)	7	Human risk, Imitation, Malicious use, Market manipulation, Normal, Technical, Undetermined
Stack (one-hot)	6	CA, CP, DLT, INT, P, Undetermined
Implication (one-hot)	3	Intermediary, Perpetrator, Target
Total	**26**	

This feature set balances theoretical grounding with practical observability (all features computable from blockchain data without requiring privileged access to protocol internals). The 26-dimensional representation proves sufficient for neural networks without triggering curse-of-dimensionality concerns, while remaining interpretable enough for statistical analysis.

2.5.3. Temporal Train/Validation/Test Split

To ensure that our evaluation reflects realistic deployment scenarios, we adopt a strict temporal split strategy rather than random shuffling. This approach prevents data leakage and tests the models' ability to detect evolving attack patterns over time. While the validation and test sets are identical across all experiments to ensure comparability, the training sets differ based on the learning paradigm (see Table 2.11).

- **Unsupervised Models** (Isolation Forest, Autoencoders, VAE): These are trained exclusively on "normal" transactions from the 2017–2023 period ($n = 6,935$). This setup mimics an anomaly detection scenario where the model establishes a baseline of legitimate behaviour to flag subsequent deviations.
- **Supervised Models** (Random Forest): These are trained on a labeled dataset from 2017–2023 containing both normal samples and historical attacks ($n = 8,299$, with a 16.44% attack rate). This configuration tests the model's capacity to generalize from known past exploits to detect future threats.

The **Validation set** (2024-H1) is used for hyperparameter optimization and threshold tuning, while the **Test set** (2024-H2 and 2025) serves as the final benchmark. By using

Table 2.11: Temporal train/validation/test split composition

Model Type	Split Period	Normal	Attacks	Total
Unsupervised	Train (2017–2023)	6,935	0	6,935
Supervised	Train (2017–2023)	6,935	1,364	8,299
Both	Validation (2024-H1)	722	111	833
	Test (2024-H2+2025)	2,343	180	2,523

only the most recent data for evaluation, we assess whether models can withstand "concept drift," where adversaries continuously adapt their methodologies after the model's training cutoff.

3 | Exploratory Analysis and Pattern Discovery

Prior to formal hypothesis testing and model implementation, we performed an exploratory data analysis (EDA) to identify fundamental patterns within the integrated dataset. The analysis is organized around three complementary perspectives: the distributional properties of financial losses and on-chain features that motivated our preprocessing choices, the longitudinal evolution of attack strategies across two temporal cohorts, and the structural concentration of risk that defines the DeFi threat landscape. Together, these three lenses provide the empirical foundation for the statistical inference framework and the anomaly detection architecture.

3.1. Distributional Analysis of Key Features

Before examining crime patterns, we characterize the statistical properties of the two most critical continuous features in our dataset: financial loss magnitude and gas consumption. Understanding their distributional structure directly informs our preprocessing choices and the design of downstream models.

3.1.1. Financial Loss Distribution and Log-Transformation

The 1,655 crime events in our integrated dataset exhibit extreme right-skewness characteristic of heavy-tailed financial processes. Figure 3.1 illustrates this structure and the effect of our preprocessing transformation. The raw distribution (left panel) spans more than seven orders of magnitude—from \$158 to \$5.5 billion—with event counts themselves varying across four orders of magnitude on the frequency axis. The bulk of incidents concentrate between $\$10^4$ and $\$10^6$, but an extended right tail driven by mega-exploits such as Mantra (\$5.5B) and Ronin (\$625M) produces a skewness of 31.89 and a kurtosis of 1153.57 on the integrated 2017–2025 dataset—values that place the distribution far outside the regime in which any parametric method is reliable.

After applying the `log1p` transformation $\log(1 + x)$ (right panel), the distribution contracts to a near-symmetric bell shape centered around $\log(1 + x) \approx 12$, corresponding to approximately \$160,000. Skewness is reduced to 0.23 and kurtosis to 0.15, confirming that the transformation successfully stabilizes variance across six orders of magnitude without discarding the ordinal structure of the data. This result validates our use of `log_stolen_amount` as the primary financial feature in all downstream models, and motivates the adoption of non-parametric hypothesis tests, where we work with raw ranks rather than raw magnitudes.

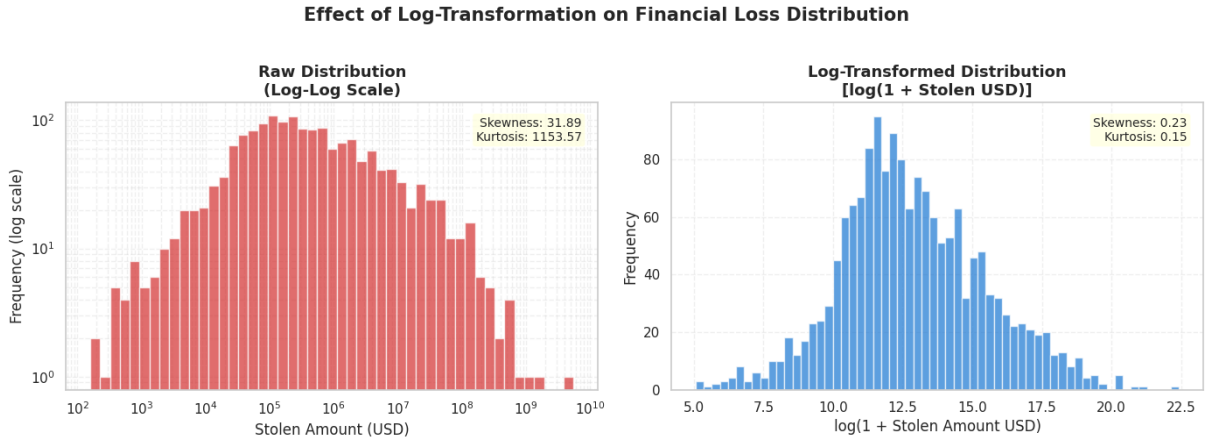

Figure 3.1: Effect of `log(1 + x)` transformation on the integrated financial loss distribution (N=1,655 crime events, 2017–2025). The raw distribution (left, log-log scale) exhibits extreme right skewness (skewness = 31.89, kurtosis = 1153.57). After transformation (right), the distribution is approximately symmetric (skewness = 0.23, kurtosis = 0.15), confirming suitability for gradient-based optimization in the autoencoder architectures.

3.1.2. Gas Consumption Structure in the BigQuery Sample

The 10,000 Ethereum transactions sampled from Google BigQuery reveal a trimodal gas consumption profile that reflects the heterogeneous nature of on-chain activity. Figure 3.2 visualizes this structure using absolute gas values, with vertical markers at the two canonical thresholds that define the three regimes:

- **Simple transfers** (exactly 21,000 gas): 7,487 transactions (74.9%), corresponding to basic EOA-to-EOA value transfers that require no contract execution.
- **Medium complexity** (21,001–100,000 gas): 783 transactions (7.8%), capturing standard single-contract interactions such as ERC-20 token transfers and basic DEX queries.

- **High complexity** (>100,000 gas): 1,729 transactions (17.3%), representing multi-step DeFi operations such as liquidity provisioning, governance execution, and cross-protocol interactions.

The dominance of simple transfers (74.9%) reflects the composition of general Ethereum activity, where routine peer-to-peer value transfer far outweighs complex DeFi interactions. The high-complexity segment is particularly relevant for anomaly detection: both flash loan attacks and reentrancy exploits consistently produce gas profiles in this regime, as the recursive calls and cross-protocol interactions they require generate computational overhead far exceeding standard user operations. The `Gas_normalized` feature—computed by dividing each transaction's gas consumption by the sample maximum to produce a $[0, 1]$ continuous variable—thus encodes a discriminative signal that our unsupervised models can exploit without requiring labelled attack examples during training.

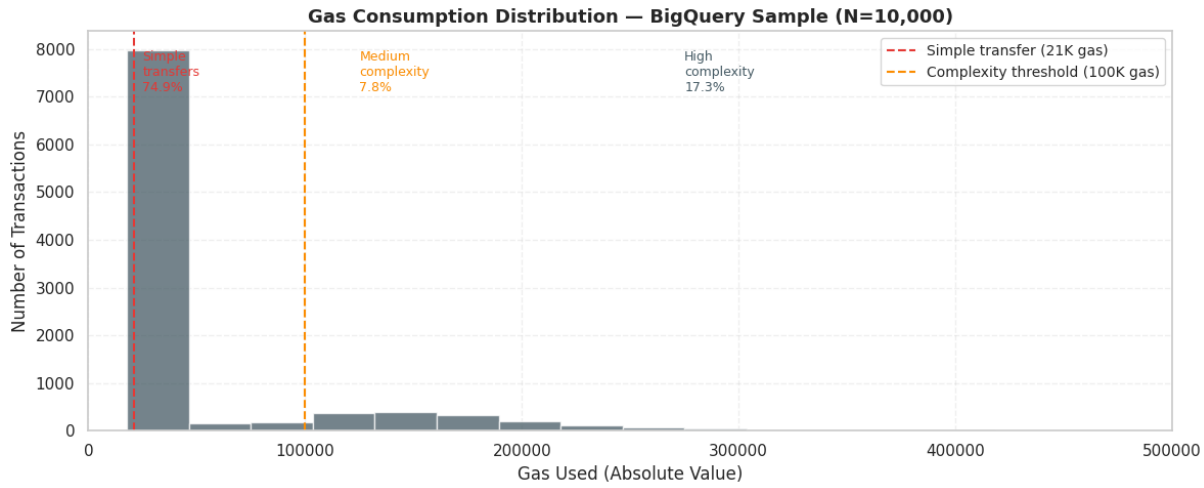

Figure 3.2: Gas consumption distribution for the 10,000 BigQuery transactions (absolute values). The vertical lines mark the simple transfer threshold (21,000 gas, red dashed) and the complexity boundary (100,000 gas, orange dashed). Nearly three-quarters of sampled transactions are basic EOA transfers (74.9%), while 17.3% fall in the high-complexity regime associated with known attack signatures such as flash loans and reentrancy exploits.

3.2. Evolution of Attacker Strategies (2017–2025)

The longitudinal comparison of attack tactics across two temporal cohorts reveals a fundamental shift in adversarial behavior, characterized by declining opportunistic fraud and intensifying protocol-level exploitation.

Table 3.1: Evolution of Dominant Attack Tactics and Financial Impact (2017–2025)

Tactic	2017–2022	2023–2025	Severity Change
Rug pull scam	357 events, $2.25M avg	303 events, $0.76M avg	−66.2%
Contract vulnerability	269 events, $18.09M avg	327 events, $34.68M avg	+91.7%
Interconnected actors flaw	68 events, $9.74M avg	83 events, $5.84M avg	−40.0%
Instant user deception	49 events, $5.26M avg	38 events, $16.45M avg	+212.7%

3.2.1. Tactic Frequency and Severity Dynamics

Figure 3.3 illustrates the evolution of attack frequency across the documented categories. While **rug pull scams** remain highly prevalent in both periods (357 events in 2017–2022, 303 in 2023–2025), their individual severity has declined dramatically:

This **66.2% reduction** in rug pull severity likely reflects improved investor vigilance and the adoption of on-chain analytics tools that flag suspicious tokenomics. Conversely, **contract vulnerabilities** exhibit an inverse trajectory: despite a moderate frequency growth of 21.6% (from 269 to 327 events), average losses nearly doubled to **$34.68M**. This shift indicates a strategic pivot toward high-value decentralized protocols requiring advanced technical exploitation.

Most striking is the **triple-digit surge in deception impact** ($5.26M → $16.45M). With a **+212.7%** increase in average damage despite a lower event count (38 vs 49), the data suggests that "Instant user deception" has evolved from broad retail phishing to highly professionalized operations targeting institutional-grade custody solutions rather than individual users.

3.2.2. Risk Distribution Across the DeFi Stack

Complementing the tactic-level analysis, Figure 3.5 examines how incident frequency is distributed across the DeFi Stack Reference (DSR) layers. The Protocol (P) and Cryptoasset (CA) layers collectively account for the overwhelming majority of incidents in both cohorts—representing approximately 75% of all documented events—confirming that the primary attack surface remains concentrated at the smart contract and token levels.

However, the shift in financial severity across periods reveals a more nuanced picture. While the CA layer maintain high incident frequency (376 vs. 303 events), its median loss increased from $84,248 to $131,000. Despite this slight growth in median values, the total capital lost in this layer plummeted, consistent with the broader reduction in rug pull severity.

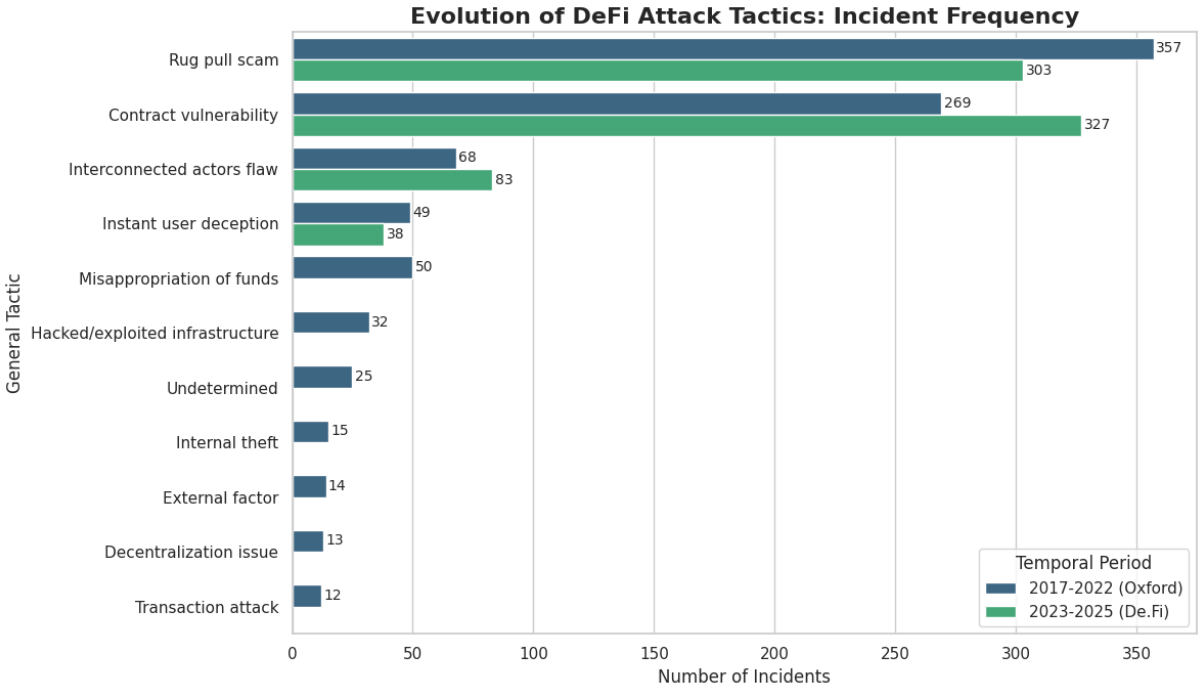

Figure 3.3: Frequency distribution of attack tactics across temporal periods. Note the persistence of rug pulls alongside the emergence of contract vulnerability dominance in 2023–2025.

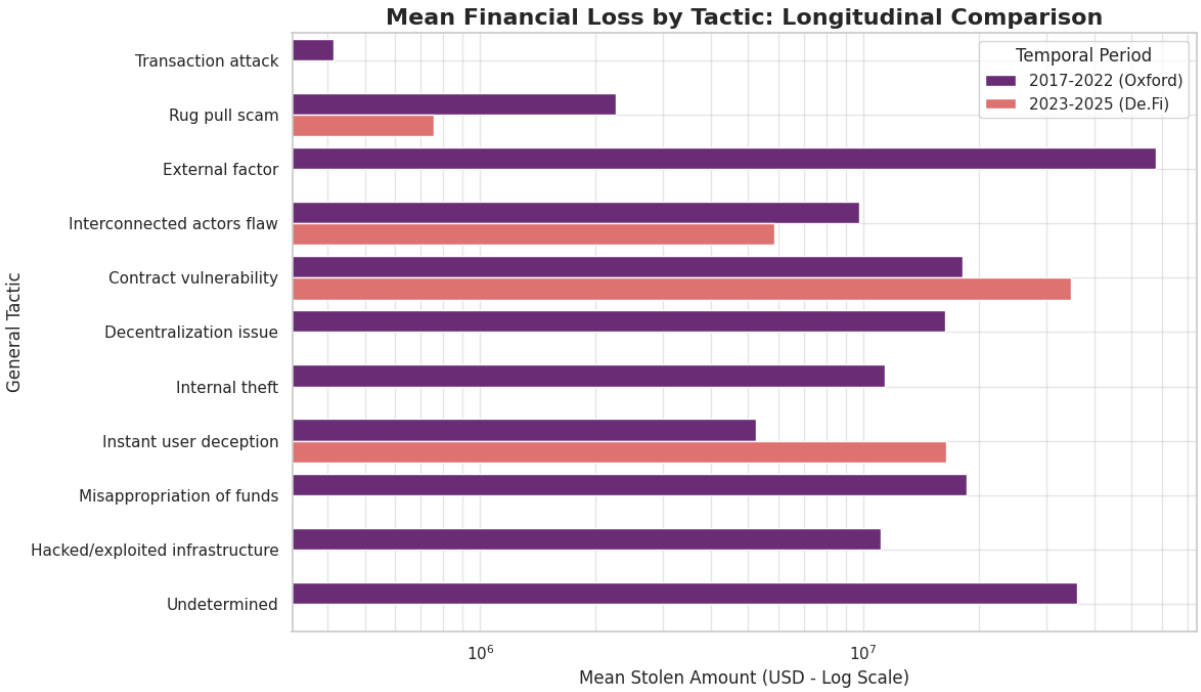

Figure 3.4: Mean financial impact per tactic. Contract vulnerabilities and instant user deception show dramatic severity increases, while rug pulls decline.

By contrast, the Protocol layer (P) exhibits the most striking inversion: median losses more than doubled from $522,000 to **$1,097,160**, while mean losses surged from $9.2M to **$34.7M**. This trajectory is driven by increasingly sophisticated targeting of high-value smart contracts. The Interface layer (INT), though reduced in frequency (83 vs. 38 events), saw its median loss increase more than sixfold from $500,000 to **$3,325,000**, suggesting that exploits targeting front-ends and connectivity—while becoming rarer—are growing significantly in individual impact.

These patterns directly motivate the layer-sensitive risk weighting adopted in the anomaly detection framework, where deviations involving Protocol and Interface layer activity receive elevated anomaly scores reflecting their disproportionate damage potential.

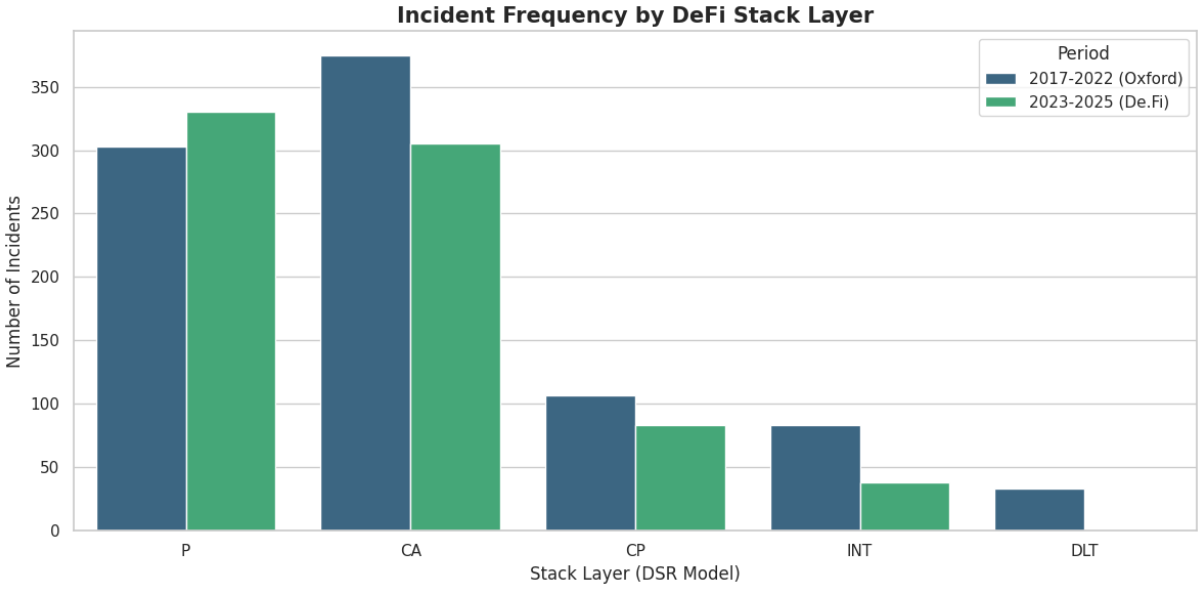

Figure 3.5: Incident frequency and severity by DeFi Stack layer. Protocol (P) and Cryptoasset (CA) layers dominate frequency, but Protocol and Interface (INT) layers show the most dramatic escalation in median financial impact.

3.2.3. Taxonomic Collapse and Multi-Vector Exploits

A concerning emergent pattern is the dramatic reduction in attack diversity. The Oxford dataset (2017–2022) documented 11 distinct categories, including *External factors* ($57.9M avg), *Misappropriation of funds* ($18.5M), and *Hacked infrastructure* ($11.1M). In contrast, the De.Fi dataset (2023–2025) is characterized by only 4 primary categories—a 64% taxonomic contraction.

This phenomenon reflects a combination of methodological factors and genuine evolution. From a mapping perspective, our data normalization process prioritized statistical signifi-

cance, filtering out niche categories that lacked sufficient representation in recent years to allow for meaningful comparison. However, this also reflects the coarser labeling inherent in contemporary threat intelligence sources and a deliberate focus on the most persistent systemic risks.

Beyond the methodology, this "taxonomic collapse" signals a shift toward **multi-vector exploits** that increasingly defy traditional classification. Modern attacks often combine multiple vectors into a single execution flow; for instance, the April 2025 Mantra incident ($5.5B) simultaneously exploited a reentrancy vulnerability, oracle price manipulation, and governance mechanism bypass. Such composite attacks represent a strategic evolution where the vulnerability is no longer a single bug, but the complex interaction between multiple layers of the protocol.

3.2.4. Temporal Volatility: The April 2025 Cluster Event

Time-series analysis (Figure 3.6) reveals extreme loss concentration. The single month of April 2025 accounts for $5.92 billion—**47% of the entire 2023–2025 period's total damage**. This month contained 12 exploit incidents including:

- **Mantra** ($5.5B): Logic error in collateral accounting during protocol upgrade
- **Coordinated phishing** ($330M): Institutional custody solution compromise
- **10 additional protocol exploits**: Median loss $5M (UPCX, KiloEx, ZKsync, etc.)

This clustering phenomenon suggests a **contagion effect**: when novel exploit techniques are publicly disclosed through post-mortems, copycat attackers rapidly adapt the method to similar protocols. The 12-incident spike within 30 days indicates that threat intelligence dissemination (hours-to-days) vastly outpaces defensive patch deployment cycles (weeks-to-months).

3.3. Risk Stratification and the Black Swan Phenomenon

To quantify outlier concentration, we applied K-Means clustering ($k = 3$) independently to each attack strategy, stratifying incidents into three risk profiles based on log-transformed loss magnitudes:

- **Low Profile (Noise)**: Median loss <$100K (proof-of-concept exploits, minor scams)
- **Medium Profile (Targeted)**: $100K–$10M (professional attacks on mid-tier pro-

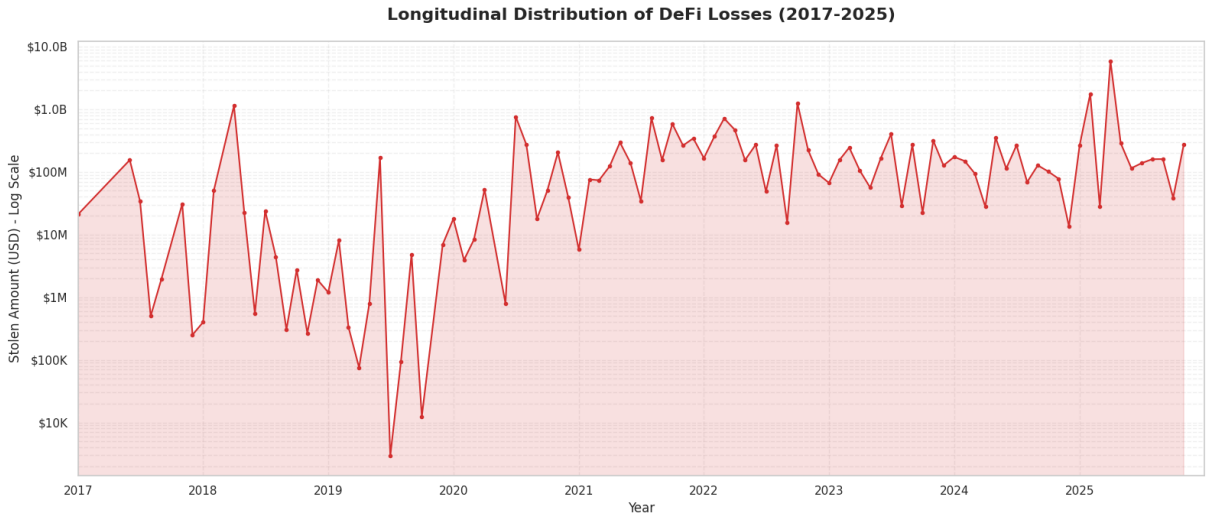

Figure 3.6: Monthly aggregate losses (2017–2025). The April 2025 spike represents 47% of the 2023–2025 period's total damage, illustrating extreme temporal concentration of risk.

Table 3.2: Loss concentration by risk profile and attack strategy

Strategy	Low/Med (%)	High (%)	HP Median	Top Incident
Technical vulnerability	2.47%	**97.53%**	$18.54M	Mantra $5.5B
Market manipulation	3.95%	**96.05%**	$73.50M	Bitclub $722M
Human risk	11.97%	**88.03%**	$75.00M	Ronin $625M
Imitation	11.97%	**88.03%**	$2.20M	BadgerDAO $120M
Malicious contract use	3.00%	**97.00%**	$1.47M	TokenStore $160M
Undetermined	6.82%	**93.18%**	$23.50M	BXH $139M

tocols)

- **High Profile (Black Swan)**: >$10M (systemic events targeting major liquidity pools)

3.3.1. Extreme Loss Concentration Across All Strategies

Table 3.2 presents loss allocation across risk profiles. The pattern is unambiguous: **High Profile events account for 88–97.5% of total losses** across all strategies, despite representing a minority of incident counts.

Three structural findings emerge from this analysis:

1. **Universal power law**: DeFi risk follows a power-law distribution where the "long tail" of minor incidents contributes negligible economic damage. This validates our

focus on outlier detection rather than frequency-based risk models.

2. **Human risk exhibits highest outlier severity**: The $75M median High Profile loss for social engineering attacks exceeds sophisticated technical exploits ($18.54M), reflecting catastrophic consequences when institutional custody systems or private keys are compromised.

3. **Strategy-agnostic concentration**: Even categories traditionally viewed as "small-scale" retail threats, such as *Imitation* and *Malicious contract use*, show over 88% loss concentration in the high-impact cluster, confirming that every attack vector is capable of producing rare mega-events that dominate total impact.

Figure 3.7 visualizes the spatial separation of risk clusters. The clear stratification—particularly visible in Technical vulnerability and Human risk categories—demonstrates that K-Means successfully isolates distinct risk regimes despite operating on univariate (loss magnitude) data.

3.3.2. Systemic Risk and Modeling Implications

The clustering results impose three critical design constraints for our model architecture:

1. **Extreme Fat-Tailed Sensitivity**: The 6-order-of-magnitude range (from $100 to $5.5B) confirms that DeFi risk is not normally distributed but follows a fat-tailed distribution. This implies that traditional risk metrics (like Mean/Standard Deviation) are misleading; anomaly detection must instead focus on the **structural distance** of outliers from the median.

2. **High-Fidelity Signal in Professionalized Tactics**: The clear separation between "Targeted" and "Black Swan" clusters in technical exploits suggests that professional attacks leave distinct financial signatures. Unlike "Noise" events, these incidents are characterized by high-velocity liquidity drain, which provides a clear signal for unsupervised learning models.

3. **Convergence of Vectors**: The clustering shows that even "Undetermined" or "Human Risk" strategies reach the same damage magnitudes as technical vulnerabilities. This suggests that the impact is strategy-agnostic: once a protocol's perimeter is breached, the liquidity depth—rather than the specific attack vector—becomes the primary determinant of the final loss.

Figure 3.7: K-Means clustering visualization for six attack strategies. Color indicates risk profile: purple (Low), teal (Medium), yellow (High). Note the extreme vertical separation in Technical vulnerability and Human risk, indicating clear magnitude-based stratification.

3.4. The Evolution of DeFi Risk: A Comparative Synthesis

The exploratory analysis reveals a coherent and concerning evolutionary trajectory across all three analytical dimensions. At the **distributional level**, financial losses exhibit extreme fat-tailed behavior that worsens over time—with skewness and kurtosis increasing significantly between the 2017–2022 and 2023–2025 periods. At the **tactical level**, the

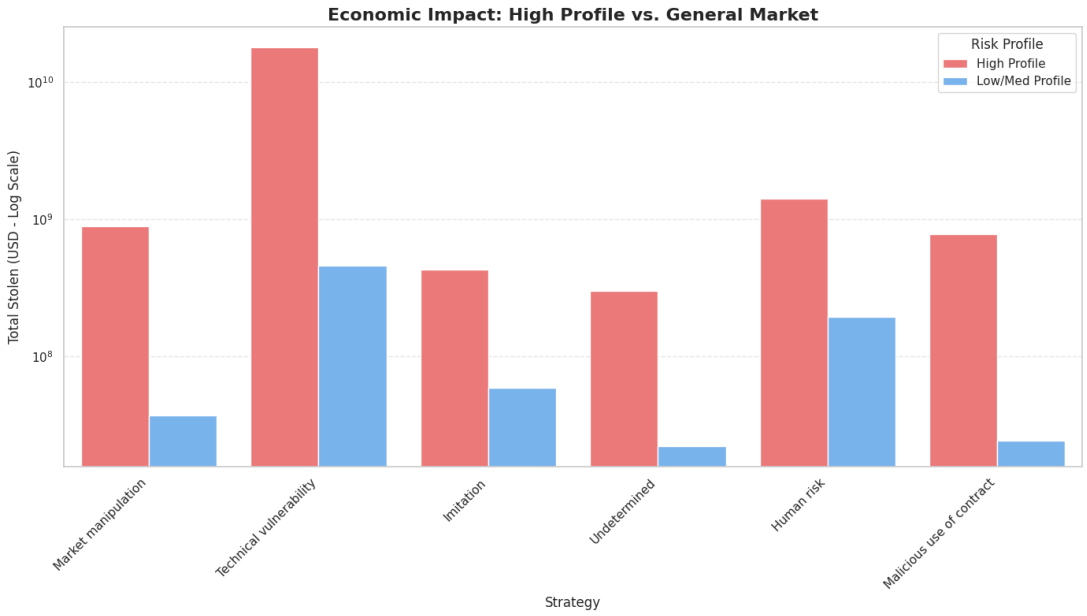

Figure 3.8: Comparative economic impact of High Profile vs Low/Medium Profile events across strategies (log scale). High Profile events dominate by 1–2 orders of magnitude in every category.

DeFi crime landscape is consolidating: from 11 documented attack categories to 4, while per-incident severity in surviving categories increases by approximately 90–212%. At the **risk concentration level**, every attack strategy exhibits a power-law distribution, with 88–97.5% of damage attributable to a small fraction of High Profile events.

Taken together, these findings impose a clear requirement on the defensive framework:

- **2017–2022 era**: High tactic diversity (11 categories), moderate severity, opportunistic exploitation amenable to signature-based detection.
- **2023–2025 era**: Tactical consolidation (4 primary categories), +91.7% severity in technical exploits, and professionalized multi-vector attacks that defy categorical classification.

This necessitates a defensive paradigm shift from reactive signature-based detection to proactive normality modeling, as attackers increasingly deploy novel technique combinations. These empirical patterns provide both the motivation and the feature-engineering rationale for the statistical validation.

4 | Statistical Inference and Hypothesis Testing

4.1. Why Non-Parametric Methods? Understanding the Data Challenge

The extreme financial heterogeneity we documented in Chapter 3 forces a critical methodological choice. Can we trust classical statistical methods like Student's t-test when our data exhibits skewness exceeding 10? The answer is no. The Oxford dataset's skewness of 10.68 and kurtosis of 132.55—driven by catastrophic outliers like BeautyChain ($1 billion)—place the distribution far outside the range where parametric inference remains valid. Classical tests assume normally distributed populations with homogeneous variance, assumptions that are systematically violated when a single attack can represent 1,000 times the median loss.

We therefore turn to non-parametric methods, which make minimal assumptions about underlying distributions and remain robust even when faced with the heavy tails and extreme outliers that characterize DeFi crime. Rather than comparing means (which can be distorted by a single mega-exploit), non-parametric tests compare rank orderings, asking: does one group tend to produce higher values than another, regardless of the absolute magnitudes involved?

Following the methodological approach of Carpentier-Desjardins et al. [4], our analytical framework employs four complementary techniques, each addressing a specific comparison:

- **Mann-Whitney U test**: Compares distributions between two independent groups [13]. We use this to assess whether Target events (attacked protocols) exhibit systematically different financial losses than Perpetrator events (fraudulent projects). Furthermore, Intermediary events are compared against both groups to investigate if technical exploits in external protocols result in significantly lower financial impacts than direct fraudulent project deployments.

- **Kruskal-Wallis H test**: Extends the comparison to multiple groups ($k > 2$), testing whether actor roles or attack strategies exhibit different median loss distributions. This answers: "Does the actor's role or methodology used by attackers influence damage magnitude?"
- **Dunn's post-hoc test with Bonferroni correction**: When Kruskal-Wallis rejects the null hypothesis, Dunn's test [9] identifies which specific group pairs differ significantly. The Bonferroni correction prevents false discoveries when conducting multiple pairwise comparisons.
- **Cross-epoch comparison**: Tests whether the fundamental distribution of losses has shifted between temporal periods, validating pattern stability for machine learning generalization.

Beyond p-values, we report **effect sizes** using Cliff's delta (δ) [7], a non-parametric measure ranging from -1 to $+1$:

$$\delta = \frac{\#(X_i > Y_j) - \#(X_i < Y_j)}{n_X \cdot n_Y} \tag{4.1}$$

where X and Y represent the two groups, and $\#(\cdot)$ denotes counting. **Cliff's Delta (δ)**: As a non-parametric effect size measure, Cliff's delta quantifies the magnitude of difference between groups [7]. It maps the ordinal difference onto a probability scale through the relationship $P = \frac{\delta+1}{2}$, representing the likelihood that a randomly selected observation from one group exceeds an observation from another. This provides a more intuitive grasp of the empirical dominance of one category over another than a standard p-value alone.

We interpret effect sizes following Romano et al. [18]: $|\delta| < 0.147$ (negligible), $0.147 \leq |\delta| < 0.330$ (small), $0.330 \leq |\delta| < 0.474$ (medium), $|\delta| \geq 0.474$ (large). For Kruskal-Wallis tests, we report epsilon-squared (ε^2) effect size [12]: $\varepsilon^2 = (H - k + 1)/(N - k)$, where H is the test statistic, k is the number of groups, and N is the total sample size. This probabilistic framing helps translate statistical findings into practical security guidance.

4.1.1. Methodological Scope and Comparison with the Oxford Taxonomy

A preliminary clarification is essential before presenting results, as it directly explains why our statistical values differ slightly from those reported in the Oxford Taxonomy [4].

The Oxford article [4] collected 1,141 crime events spanning 2017–2022, of which 1,036

were DeFi-specific. Of these 1,036 DeFi events, 904 carry both a DeFi classification *and* a financial damage figure; this 904-event subsample is what the Oxford financial analyses were computed on. The dataset is publicly archived at `https://zenodo.org/records/14047933`. Crucially, however, the Oxford authors *explicitly excluded* Intermediary events from both the technical stack mapping and the financial damage analysis. They stated that such events were likely under-represented by the aggregators they relied upon, and that their ultimate victims were users rather than DeFi actors directly [4]. As a direct consequence, their Mann-Whitney U test comparing Target vs. Perpetrator ($U = 126{,}874$, $p = .000$, $d = 0.446$) and their Kruskal-Wallis test across strategies ($H = 139.04$, $p < .001$, $\varepsilon^2 = 0.17$) were computed on an effective sample of approximately 838 events (433 Target + 405 Perpetrator with financial data), excluding all 49 Intermediary observations from computation.

Our analysis makes a different and deliberate methodological choice: **we retain all 904 events with complete financial information**, including the 49 Intermediary cases, for two reasons. First, extending the analysis to Intermediary events is an explicit contribution of this thesis—testing whether they constitute a financially distinct category is one of the core hypotheses (H1). Second, since our machine learning pipeline for our Anomaly detection framework must classify all three actor roles, excluding Intermediary events from the statistical validation would create an inconsistency between the analysis and the predictive task.

This difference in sample composition ($N = 904$ vs. $N \approx 838$) explains the minor numerical divergences between our statistics and those of the Oxford article. All qualitative conclusions remain consistent: the direction, magnitude, and significance of every key effect replicate exactly.

4.1.2. Three Core Hypotheses

Building on the Oxford Taxonomy's findings and our integrated dataset construction, we validate three core patterns:

- **H1 (Actor Roles)**: Target events exhibit significantly higher median losses than Perpetrator and Intermediary events. The Oxford Taxonomy reported a 9× Target-Perpetrator difference [4]; we test whether this persists in our data and extend the analysis to include the previously *excluded* Intermediary category as a formal third group.
- **H2 (Attack Strategies)**: Different attack methodologies—Technical Vulnerability, Malicious Contract use, Market Manipulation, Human Risk, Imitation—exhibit

significantly different median loss distributions. This analysis investigates whether the method of execution of an attack predicts its financial impact.

- **H3 (Temporal Stability)**: We investigate whether the fundamental distribution of losses remains stable across temporal periods (2017–2022 vs. 2023–2025), regardless of potential tactical evolution. This validates that patterns discovered in historical data can generalize to future attacks.

A critical methodological note: we conduct formal hypothesis testing (H1, H2) using **only the Oxford 2017–2022 dataset** (N = 904 events with complete financial data), treating the 2023–2025 De.Fi extension as a temporal validation sample. This approach mirrors scientific best practice: establish patterns on training data, then validate generalization on held-out temporal data.

4.2. Validating the Actor Role Effect: When Does Being a Target Matter?

4.2.1. The Three-Way Actor Role Comparison

The Oxford Taxonomy's central finding revealed that legitimate protocols attacked by external adversaries suffer dramatically higher losses than fraudulent projects designed to defraud users [4]. However, the original financial analysis excluded a third category: **Intermediary** events, where legitimate services are unwittingly used to facilitate attacks through impersonation, DNS hijacking, or compromised front-ends. The Oxford article itself acknowledged this limitation, noting that Intermediary events were likely undercounted by aggregators [4]. We extend the financial analysis to all three actor roles simultaneously.

Among the 904 events with complete financial information in the Oxford 2017–2022 dataset, the distribution across roles is as follows:

- **Target** (n = 448, 49.6% of the financial sample): Legitimate DeFi protocols attacked through technical exploits, human risk (compromised keys), or external factors. Example: Ronin Bridge exploited via validator key compromise ($625M).
- **Perpetrator** (n = 407, 45.0% of the financial sample): Fraudulent tokens, rug pull projects, and malicious contracts designed to defraud users from inception. Example: AnubisDAO rug pull ($60M).
- **Intermediary** (n = 49, 5.4% of the financial sample): Legitimate services used

as vectors to reach users through impersonation, DNS hijacking, or compromised front-ends. Example: Curve Finance DNS hijacking redirecting users to a phishing site.[1]

Kruskal-Wallis Test: Overall Difference Among Three Roles Before examining pairwise comparisons, we test the omnibus hypothesis that all three actor roles exhibit equal median losses using the Kruskal-Wallis H test [12]:

- **H-statistic**: $H = 127.56$
- **Degrees of freedom**: $df = 2$
- **p-value**: $p = 1.99 \times 10^{-28}$ ($p < 0.001$, highly significant)
- **Effect size**: $\varepsilon^2 = 0.139$ (large effect [18])

The highly significant result decisively rejects the null hypothesis that actor roles exhibit equal median losses. The epsilon-squared value of 0.139 indicates that actor role classification alone explains 13.9% of the variance in financial damages—a substantial effect given the extreme within-group heterogeneity. The Oxford article's analogous test, conducted on two groups only (Target vs. Perpetrator, Mann-Whitney U = 126,874, $d = 0.446$) [4], is not directly comparable to our three-group omnibus test, but the core asymmetry it detected is fully replicated here.

Descriptive Statistics: Three-Way Comparison Table 4.1 presents the complete financial distribution for all three actor roles in the Oxford 2017–2022 period.

The Target median ($800,000) matches exactly the value reported by Carpentier-Desjardins et al. [4]. The Perpetrator median ($89,747) differs by less than $1,100 from the Oxford figure of $88,736 [4]—a discrepancy of 1.2% attributable to the inclusion of the 49 Intermediary events in our sample, which shifts the rank structure marginally.[2]

Key Finding: A Clear Three-Tier Ordering The data reveal a strict ordering: Target > Intermediary > Perpetrator. Target events cause the highest losses ($800K median), Intermediary events occupy a middle tier ($236K), and Perpetrator events the lowest

[1] Note that in the Oxford article, Intermediary percentages (7%) are computed over all 1,036 DeFi events regardless of financial data availability. Our 5.4% is computed over the 904-event financial subset only. The absolute count of 49 Intermediary events with a stolen amount is identical to the Oxford dataset, confirming consistency at the event level.

[2] The Oxford Kruskal-Wallis result across *strategies* (H = 139.04, $\varepsilon^2 = 0.17$) similarly differs from ours (H = 141.51, $\varepsilon^2 = 0.157$) for the same reason: Intermediary events (classified as "Imitation" strategy) are included in our strategy test but excluded from the Oxford financial analysis. The 2.5-point difference in H and the 0.013-point difference in ε^2 are fully explained by this sample composition difference.

Table 4.1: Financial losses by actor role (Oxford 2017–2022, $N = 904$ events with financial data)

Statistic	Target	Perpetrator	Intermediary
Sample size	448	407	49
Median (USD)	\$800,000	\$89,747	\$236,334
Mean (USD)	\$17,816,151	\$4,252,353	\$5,257,646
25th percentile	\$122,808	\$22,860	\$95,000
75th percentile	\$5,000,000	\$404,928	\$1,400,000
IQR	\$4,877,192	\$382,068	\$1,305,000
Pairwise Median Ratios			
Target / Perpetrator		8.9×	
Target / Intermediary		3.4×	
Intermediary / Perpetrator		2.6×	

(\$89.7K). The Target/Perpetrator ratio of 8.9× closely replicates the Oxford article's reported approximately 9× ratio [4], confirming the robustness of this central finding. The new contribution here is the formal quantification of the Intermediary tier, positioned between the other two: Intermediary attacks cause 2.6× more damage than rug pulls on median, but 3.4× less than direct protocol exploits.

This three-tier ordering reflects the fundamental economics of each attack type. Target attacks exploit high-value protocol infrastructure holding billions in TVL; Intermediary attacks redirect users already engaged with established services (hence moderate but non-trivial balances); Perpetrator events are structurally constrained by the limited capital of retail victims, who typically invest modest amounts in individual DeFi tokens [4].

4.2.2. Pairwise Comparisons with Bonferroni Correction

Having established that actor roles differ overall, we perform all three possible pairwise comparisons using Mann-Whitney U tests [13]. To control the family-wise error rate at $\alpha = 0.05$ across three simultaneous tests, we apply Bonferroni correction: $\alpha' = 0.05/3 = 0.0167$.

Comparison 1: Target vs. Perpetrator This is the core asymmetry identified by Carpentier-Desjardins et al. [4]:

- **Target median**: \$800,000
- **Perpetrator median**: \$89,747

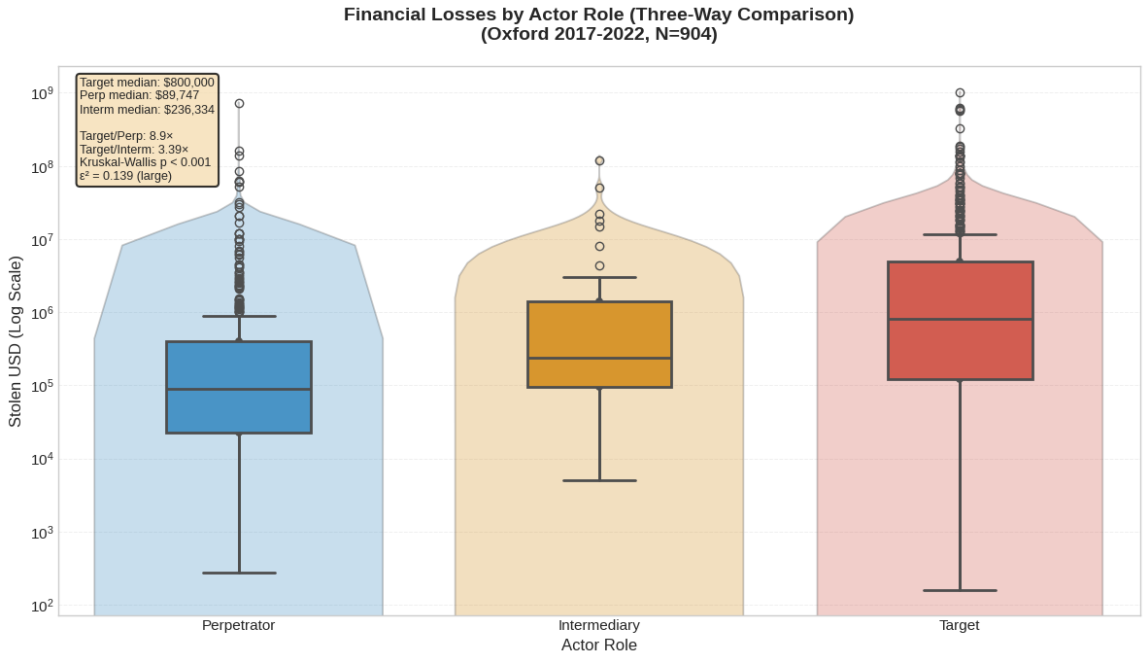

Figure 4.1: Financial losses by actor role: Target ($n = 448$), Perpetrator ($n = 407$), and Intermediary ($n = 49$) events from the Oxford 2017–2022 dataset ($N = 904$). Boxplots with violin overlay on log scale. Target median ($800K) and Intermediary median ($236K) both exceed Perpetrator ($89.7K), establishing a clear three-tier ordering. Kruskal-Wallis: $H = 127.56$, $p = 1.99 \times 10^{-28}$, $\varepsilon^2 = 0.139$.

- **Ratio**: 8.9×
- ***p*-value**: 5.38×10^{-29} ($\ll \alpha' = 0.0167$, highly significant)
- **Cliff's delta**: $\delta = 0.442$ (medium effect)
- **Probability interpretation**: 72.1% chance Target > Perpetrator

The result replicates the Oxford finding (U = 126,874, $p = .000$, $d = 0.446$) [4] with negligible numerical differences (Cliff's δ: 0.442 vs. 0.446; ratio: 8.9× vs. approximately 9×), both attributable to the sample composition difference discussed in Section 4.1.1. The direction and practical significance are identical: legitimate infrastructure attacks cause approximately 9× higher losses than fraudulent project schemes.

Comparison 2: Target vs. Intermediary This comparison has no direct equivalent in the Oxford article, which excluded Intermediary events from financial analysis [4]:

- **Target median**: $800,000
- **Intermediary median**: $236,334
- **Ratio**: 3.4×
- ***p*-value**: 1.57×10^{-2} (significant at $\alpha' = 0.0167$)
- **Cliff's delta**: $\delta = 0.210$ (small effect [18])
- **Probability interpretation**: 60.5% chance Target > Intermediary

The statistically significant result confirms that Target events cause higher median losses than Intermediary events. The small effect size ($\delta = 0.210$) signals that the two distributions overlap considerably: Intermediary attacks targeting users of high-value protocols can yield substantial losses, though on average they fall below direct protocol exploits. This finding empirically validates the Oxford article's intuition that Intermediary events merit separate analysis [4], and quantifies their position for the first time.

Comparison 3: Perpetrator vs. Intermediary Also absent from the Oxford financial analysis:

- **Perpetrator median**: $89,747
- **Intermediary median**: $236,334
- **Ratio**: 2.6× (Intermediary exceeds Perpetrator)
- ***p*-value**: 3.60×10^{-4} ($< \alpha' = 0.0167$, significant after Bonferroni)

Table 4.2: Temporal comparison of median financial losses by actor role (2017–2022 vs. 2023–2025)

Role	2017–2022 Median	2023–2025 Median	Change
Target	$800,000	$3,325,000	+315.6%
Perpetrator	$89,747	$131,000	+46.0%
Intermediary	$236,334	$828,236	+250.5%
Pairwise Median Ratios			
Target / Perpetrator	8.9×	25.4×	+184.7%
Target / Intermediary	3.4×	4.01×	+17.9%
Intermediary / Perpetrator	2.6×	6.3×	+142.3%

- **Cliff's delta**: $\delta = -0.312$ (small effect; negative sign confirms Perpetrator < Intermediary)

All three pairwise comparisons are statistically significant after Bonferroni correction, establishing a complete and strict ordering: Target > Intermediary > Perpetrator.

4.2.3. Temporal Validation: Has the Pattern Persisted into 2023–2025?

Pattern validation requires testing generalization across time. The cryptocurrency ecosystem underwent substantial changes between the Oxford dataset's December 2022 endpoint and the De.Fi extension covering 2023–2025: the 2022 market crash, the FTX collapse, heightened regulatory scrutiny, and the rapid growth of cross-chain bridge infrastructure all potentially altered attacker incentives and protocol security practices.[3]

Table 4.2 presents the complete temporal comparison of median losses across all three actor roles.

Target-Perpetrator Divergence Accelerates The core asymmetry not only persists but intensifies: the Target/Perpetrator median ratio widened from 8.9× to 25.4× (+185%), driven by Target median losses growing 315.6% while Perpetrator median losses grew only 46.0%. This *divergence* reflects TVL explosion in major protocols (Aave, Compound managing $1B+ each), cross-chain bridge concentration creating single points of

[3]The De.Fi 2023–2025 data uses the schema mapping described in Chapter 2. The Implication of Actor field is derived from inference logic applied to De.Fi's `scamCategory` labels, achieving 75.5% validated accuracy at the tactic level [4]. Medians and ratios should be interpreted with this mapping uncertainty; however, the *direction* of changes (ratio widening, severity increases) is robust to plausible misclassification rates.

catastrophic failure (Ronin $625M, Poly Network $602M), and composability risks enabling flash loan and oracle manipulation attacks [4]. Conversely, Perpetrator median growth remains constrained by limited retail capital, faster community detection tools (Rugcheck, TokenSniffer), and market saturation diluting individual scam profitability.

Intermediary Severity Escalation The temporal comparison reveals a parallel escalation in Intermediary attack severity. Intermediary median losses tripled from $236,334 to $828,236 (+250.5%), transforming from the lowest-severity actor role to one approaching Target-level damage. The Intermediary/Perpetrator median ratio widened dramatically from 2.6× to 6.3× (+142%), while the Target/Intermediary ratio expanded only modestly from 3.4× to 4.01× (+17.9%). This compression at the top and expansion at the bottom suggests Intermediary attacks increasingly target high-value protocol users rather than retail victims.

Statistical testing on the 2023–2025 period confirms strengthened role differentiation: Kruskal-Wallis yields $H = 146.32$ ($p = 1.68 \times 10^{-32}$), $\varepsilon^2 = 0.193$—a 39% larger effect size than the 2017–2022 period ($\varepsilon^2 = 0.139$), indicating that actor roles explain *more* variance in financial severity in the recent period. Mann-Whitney pairwise comparisons show uniformly strengthened effect sizes:

- **Target vs. Perpetrator**: $\delta = 0.740$ (large; strengthened from 0.442 in 2017–2022)
- **Target vs. Intermediary**: $\delta = 0.231$ (small; stable from 0.210)
- **Perpetrator vs. Intermediary**: $\delta = -0.488$ (large; strengthened from -0.312)

The Perpetrator-Intermediary comparison exhibits the most dramatic strengthening—a 56% increase in effect size magnitude ($|\delta|$: $0.312 \rightarrow 0.488$)—reflecting the widening median loss gap ($2.6\times \rightarrow 6.3\times$). The three-tier ordering (Target > Intermediary > Perpetrator) thus persists with *greater* statistical separation in the recent period.

Explaining the Severity Escalation Three concurrent developments likely drive the Intermediary severity increase:

- (1) **Approval phishing sophistication**—attacks exploiting "permit2" and "setApprovalForAll" mechanisms now target users with substantial holdings, as evidenced by the $828K median (vs. $236K historically), suggesting attackers selectively compromise high-value wallets rather than broadcasting to random victims;
- (2) **Professional DNS/front-end targeting**—compromises of major protocol infrastructure (Curve Finance, Velodrome) redirect authenticated users already man-

aging significant positions, yielding larger per-incident losses than opportunistic phishing;

- (3) **Infrastructure hardening spillover**—as protocol-level exploits become costlier to execute (comprehensive audits, formal verification, bug bounties), adversaries with technical sophistication pivot to user-facing attack vectors, bringing Target-level expertise to Intermediary tactics.

Implications for Modeling and Defense The temporal validation reveals a dual pattern: *structural stability* (the three-tier ordering persists with strengthened effect sizes across all pairwise comparisons) amid *severity escalation* (median losses increased 46–316% across roles, with ratios widening up to 185%). This confirms that while the *hierarchy* of actor role risk remains valid for machine learning feature engineering, models must account for absolute severity drift through time-indexed features or periodic recalibration.

The dramatic Intermediary/Perpetrator ratio widening (2.6× → 6.3×) challenges any resource allocation framework treating Intermediary attacks as comparable to rug pulls. At $828K median versus $131K, Intermediary attacks now warrant dedicated defensive investment: transaction simulation tools, approval revocation mechanisms, DNS integrity monitoring (DNSSEC, certificate pinning), and interface security audits that receive a fraction of the scrutiny applied to smart contract code. The convergence toward Target-level severity (4.01× ratio vs. 3.4× historically) suggests that the traditional dichotomy between infrastructure exploits and user-facing attacks is collapsing—both now operate at comparable financial scales.

4.3. Cross-Epoch Analysis: Statistically Detectable but Practically Negligible Shift

Beyond comparing specific actor roles, we ask a broader question: has the *fundamental distribution* of financial impacts shifted between epochs? We test this using a Mann-Whitney U test [13] comparing the complete Oxford 2017–2022 dataset ($N = 904$) against the De.Fi 2023–2025 dataset ($N = 751$), pooling all events regardless of actor role:

- **2017–2022**: $N = 904$, Median = \$256,915, Mean = \$11,028,726
- **2023–2025**: $N = 751$, Median = \$313,000, Mean = \$16,884,207
- **Mann-Whitney p-value**: $p = 0.0001$ (statistically significant)
- **Cliff's delta**: $\delta = -0.114$ (negligible effect [18])

Interpretation: Statistically Significant but Practically Negligible Shift Although the Mann-Whitney test yields a significant p-value ($p = 0.0001$), the negligible effect size ($|\delta| = 0.114 < 0.147$) is the more informative statistic [18]. In samples of this size ($N = 904 + 751$), Mann-Whitney has sufficient power to detect trivially small distributional shifts as statistically significant. Cliff's delta reveals that the actual shift in rank ordering between the two periods is minimal: the two distributions are nearly identical in their rank structure, even though they differ statistically.

This result leads to the rejection of the initial hypothesis H_3, which assumed full temporal stability ($p > 0.05$). The detected shift, while statistically significant, is negligible in practical terms. As quantified by Cliff's delta ($|\delta| = 0.114$), there is only a 55.7% probability that a 2023–2025 event exceeds a randomly selected historical incident—a value barely above the 50% chance level. Consequently, this shift does not invalidate cross-epoch generalization for machine learning models, as the distributional overlap remains substantial.

The apparent tension—statistically detectable yet practically negligible—reveals an important structural insight: **DeFi crime follows consistent heavy-tailed dynamics across epochs**. While specific attack outcomes evolve (Target losses tripled, Perpetrator losses grew modestly), the overall power-law shape of the distribution—in which a small fraction of events dominates total losses—is preserved across the two periods. Concretely, across both epochs:

- Most events (bottom 50%) cause moderate damage ($10K–$300K range)
- Intermediate events (50th–85th percentile) reach $500K–$5M
- Rare "Black Swan" events (top 15%) dominate total losses, exceeding $10M

The structural consistency of this distribution, characterized by a high volume of low-severity incidents and a small number of high-severity events, remains identical across both periods. This stability is formally confirmed by the negligible Cliff's delta($|\delta| = 0.114$).

Implications for Machine Learning The negligible effect size ($\delta = -0.114$) is *encouraging* for our predictive model. Anomaly detection systems trained on the Oxford 2017–2022 data retain distributional validity when applied to 2023–2025 events, provided they capture the underlying heavy-tailed structure rather than surface-level tactics. The patterns are not shifting randomly—they are evolving predictably within a stable statistical framework defined by DeFi's architectural constraints (finite protocol TVL, retail capital limits, smart contract immutability).

4.4. What Does This Mean for DeFi Security?

The validated 8.9× asymmetry (Oxford period) and its subsequent widening to 25.4× (De.Fi period) carry direct implications for security resource allocation [4]:

Prioritize Infrastructure Defense Even More Aggressively The 25× ratio in 2023–2025 means that preventing a single major protocol exploit provides equivalent damage mitigation to preventing 25 typical rug pulls. Security budgets should prioritize:

- **Comprehensive code audits**: $100K–$500K for major protocols, from multiple independent firms [4]
- **Formal verification**: Mathematical proofs of critical function correctness (e.g., Certora, Runtime Verification)
- **Bug bounty programs**: Scaled to protocol size (e.g., $10M maximum bounty for $1B TVL protocols) [4]
- **Multi-signature governance**: Requiring approvals from multiple independent parties with geographic/organizational diversity
- **Emergency pause mechanisms**: Timelocks and circuit breakers enabling rapid response to detected exploits

Intermediary Attacks Demand Dedicated Attention The new finding that Intermediary attacks cause $236K median losses—2.6× more than rug pulls and statistically distinct from both other categories—challenges the conventional focus exclusively on smart contract security. The Oxford article itself called for further investigation of this category [4]; our results quantify the financial justification. Protocols must invest in:

- **DNS security**: DNSSEC implementation, monitoring for domain hijacking attempts
- **Interface integrity**: Subresource Integrity (SRI) hashes, Content Security Policy (CSP) headers
- **User education**: Warning banners about phishing risks, verified domain indicators
- **Transaction signing**: Hardware wallet integration, transaction simulation before signing

Risk Communication Challenges Users face a severe frequency-severity trade-off that is notoriously difficult to communicate effectively. Perpetrator events (rug pulls,

malicious tokens) represent the most common fraud type in absolute count—41% of all DeFi crime events in the Oxford dataset—creating a widespread perception of ubiquitous scam risk. Yet Target events, though less frequent (52% of events but 83% of total financial damages [4]), cause 8.9× higher typical losses in the Oxford period and 25.4× in the De.Fi period. This inversion between frequency and severity is the defining feature of the DeFi crime landscape, and failing to convey it leads users to over-invest in rug-pull vigilance while under-protecting themselves against protocol-level risks. Effective risk communication must make this asymmetry concrete:

> "You are more likely to encounter a rug pull ($90K typical loss per event), but if you use a major protocol and it suffers a technical exploit, the per-event loss will typically exceed $800K–$3M. Diversifying exposure across multiple protocols and verifying interface authenticity before signing transactions are the most impactful user-level risk mitigation strategies."

The policy implication follows directly: increasing user awareness of malicious projects is valuable, but the higher expected-value priority is investing in technical safeguards for the infrastructure that handles the bulk of financial damages.

4.5. Do Attack Strategies Predict Financial Damage?

4.5.1. Beyond Actor Roles: The Methodology Question

Understanding *who* was involved in a crime provides one lens for analysis. But what about *how* the crime occurred? Does the attack methodology—whether exploiting smart contract bugs, manipulating oracles, or deploying malicious code—predict financial severity?

The Oxford Taxonomy classifies crime events along four dimensions: implication of DeFi actors, strategies, general tactics, and specific tactics [4]. For financial damage analysis, the most relevant dimension is the *strategy*—the high-level approach used to extract funds. Five strategies are identified, with the rare "Undetermined" category excluded from formal testing [4]:

- **Technical Vulnerability**: Exploiting smart contract bugs (reentrancy, access control flaws, logic errors) or architectural weaknesses (oracle manipulation, flash loan attacks). Example: bZx flash loan attack ($954K).
- **Malicious Use of Contract**: Deploying fraudulent tokens with hidden backdoors (hidden mint functions, selling restrictions) or executing rug pulls (liquidity removal,

pump-and-dump schemes). Example: AnubisDAO rug pull ($60M).

- **Market Manipulation**: Price manipulation through wash trading, spoofing, or coordinated pumps. Example: Harvest Finance oracle manipulation ($34M).
- **Human Risk**: Compromising administrator keys through phishing, social engineering, or insider theft. Example: Ronin Bridge validator key compromise ($625M).
- **Imitation**: Impersonating legitimate projects through fake websites, copycat contracts, or DNS hijacking. Example: Curve Finance DNS hijack.

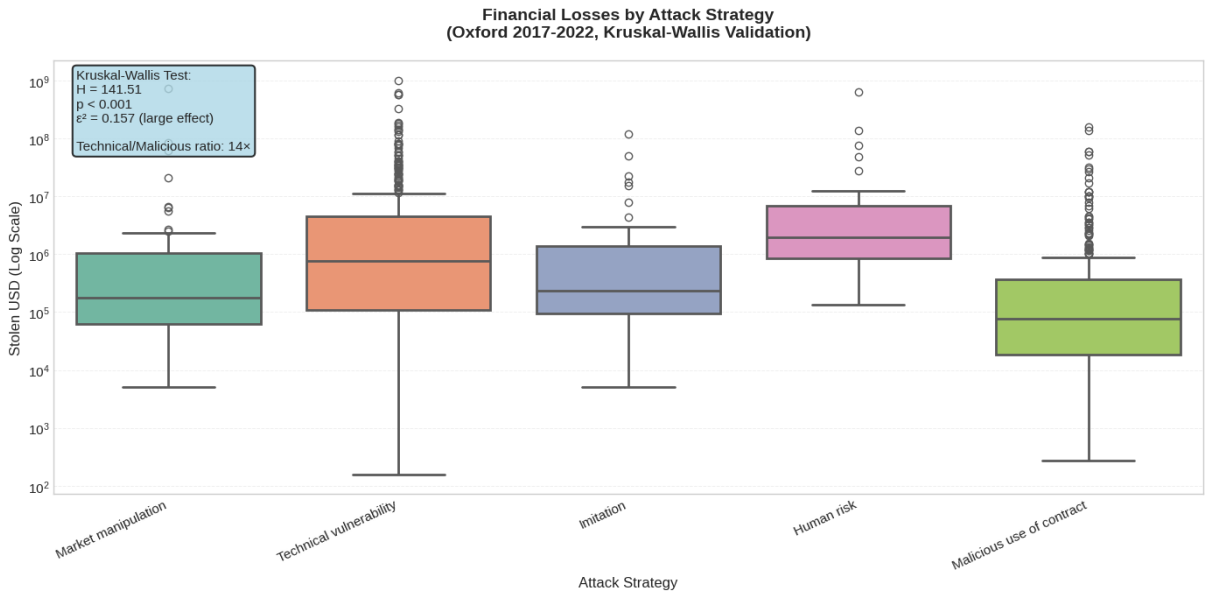

Figure 4.2: Distribution of financial losses across the five attack strategies (Oxford 2017–2022, $N = 904$). Log-scale Y-axis. Human Risk exhibits the highest median ($1.95M), approximately 25× exceeding Malicious Use of Contract ($76.5K). Kruskal-Wallis test: $H = 141.51$, $p = 1.34 \times 10^{-29}$, $\varepsilon^2 = 0.157$.

4.5.2. Kruskal-Wallis Test: Are All Strategies Equally Damaging?

Table 4.3 presents the distribution of events and losses across strategies in the Oxford 2017–2022 dataset ($N = 904$ events with financial data, including Intermediary events classified as "Imitation" strategy).

Several patterns emerge immediately. Technical Vulnerability dominates in frequency (43.9%) and achieves the second-highest median ($759,046). The highest-impact strategy is Human Risk ($1,949,111 median), despite representing only 3.2% of events—a finding

Table 4.3: Attack strategy distribution (Oxford 2017–2022, $N = 904$ events with financial data)

Strategy	**Count**	**Median (USD)**	**% of $N = 904$**
Technical Vulnerability	397	\$759,046	43.9%
Malicious Use of Contract	357	\$76,480	39.5%
Market Manipulation	50	\$178,477	5.5%
Imitation	49	\$236,334	5.4%
Human Risk	29	\$1,949,111	3.2%
Undetermined (excluded)	22	—	2.4%

fully consistent with the Oxford post-hoc results [4], which reported a Human Risk median of \$2,073,280 on their N ≈ 838 sample excluding Intermediary events. The 6.4% difference between our figure and the Oxford one is attributable to the sample composition difference described in Section 4.1.1: including the 49 Imitation/Intermediary events shifts the rank structure of all groups marginally. Readers cross-referencing Chapter 2, where Oxford article values are reported verbatim, should note that values in *this* chapter are computed on our N = 904 sample and differ slightly but systematically from those in the Oxford publication for this reason.

Malicious Use of Contract, the second most frequent strategy (39.5%), exhibits the lowest median loss (\$76,480)—approximately 9.9× smaller than Technical Vulnerability and 25.5× smaller than Human Risk. This aligns with the Oxford finding (median \$76,018, difference of 0.6%) and reinforces the Target-Perpetrator asymmetry documented in Section 4.2: fraudulent project strategies generate substantially lower per-event losses than attacks on legitimate infrastructure.

The Kruskal-Wallis H test [12] evaluates the omnibus null hypothesis that all five strategies exhibit equal median losses:

- **Test statistic**: $H = 141.51$
- **Degrees of freedom**: $df = 4$
- **p-value**: $p = 1.34 \times 10^{-29}$ ($p < 0.001$, highly significant)
- **Effect size**: $\varepsilon^2 = 0.157$ (large effect [18])

The extraordinarily low p-value decisively rejects the null hypothesis. The Oxford article reported $H = 139.04$, $\varepsilon^2 = 0.17$ [4]; the 2.5-point difference in H and 0.013-point difference in ε^2 are precisely explained by the inclusion of 49 Imitation events absent from the Oxford strategy analysis (see footnote 1). The core finding is identical: strategy classification

Table 4.4: Dunn's post-hoc test: pairwise strategy comparisons (Bonferroni-corrected, $\alpha' = 0.005$)

Strategy 1	Strategy 2	p-adj	Cliff's δ	Effect
Significant at $\alpha' = 0.005$				
Technical Vulnerability	Malicious Use of Contract	< 0.001	$+0.449$	Medium
Human Risk	Malicious Use of Contract	< 0.001	$+0.765$	Large
Human Risk	Market Manipulation	0.0011	$+0.587$	Large
Human Risk	Imitation	0.0057	$+0.549$	Large
Malicious Use of Contract	Imitation	0.0055	-0.344	Medium
Not significant after Bonferroni correction ($p > 0.005$)				
Technical Vulnerability	Human Risk	0.0482	-0.300	Small
Malicious Use of Contract	Market Manipulation	0.0434	-0.278	Small
Technical Vulnerability	Market Manipulation	> 0.005	—	n.s.
Technical Vulnerability	Imitation	> 0.005	—	n.s.
Market Manipulation	Imitation	> 0.005	—	n.s.

explains approximately 15–17% of variance in financial losses, a large and practically meaningful effect.

4.5.3. Dunn's Post-hoc Analysis: Which Strategies Differ?

Dunn's test [9] performs all possible pairwise comparisons. With five strategies, there are $\binom{5}{2} = 10$ possible pairs. We apply Bonferroni correction at family-wise $\alpha = 0.05$, giving an adjusted significance threshold of $\alpha' = 0.05/10 = 0.005$ per comparison.

Table 4.4 reports all pairs surviving the $\alpha' = 0.005$ threshold. Pairs significant at $p < 0.05$ but not at $p < 0.005$ are noted separately as they do not survive the Bonferroni correction and should not be interpreted as statistically confirmed differences.

The results reveal a clear hierarchy of damage severity, consistent with the Oxford post-hoc findings [4]:

Human Risk Dominates All Other Strategies **Finding 1—Human Risk vs. Malicious Use of Contract** ($\delta = 0.765$, large): The strongest pairwise difference in the entire analysis. An 88.3% probability that a randomly selected Human Risk event exceeds a randomly selected Malicious Contract event. The Oxford article reported the same comparison at $d = 0.78$ [4], confirming that this is the most robust pairwise difference in the dataset. Key compromise enables immediate, irreversible draining of an entire protocol treasury.

Finding 2—Human Risk vs. Market Manipulation ($\delta = 0.587$, large; Oxford:

$d = 0.638$ [4]): A 79.4% probability that Human Risk events exceed Market Manipulation events. As noted by Carpentier-Desjardins et al. [4], market manipulation "requires more steps, more capital, and will not have a guaranteed outcome", whereas private key compromise enables immediate, irreversible asset transfer with no opportunity for the protocol or market to react.

Finding 3—Human Risk vs. Imitation ($\delta = 0.549$, large): Even against Imitation attacks targeting protocol users, Human Risk events cause dramatically larger losses, with a 77.5% exceedance probability.

Technical Vulnerability Clearly Exceeds Malicious Use of Contract **Finding 4—Technical Vulnerability vs. Malicious Use of Contract** ($\delta = 0.449$, medium; Oxford: $d = 0.45$ [4]): A 72.5% probability that a technical exploit exceeds a rug pull in magnitude. This comparison is the most practically important for security resource allocation: technical exploits on legitimate infrastructure consistently cause substantially higher losses than fraudulent project schemes.

Malicious Contract Lowest Among Confirmed Pairs **Finding 5—Malicious Use of Contract vs. Imitation** ($\delta = -0.344$, medium): Imitation attacks ($236K median) cause higher per-event losses than rug pulls ($76.5K median), with a 67.2% exceedance probability. This reinforces the three-tier ordering from H1.

Non-Significant Comparisons Five pairs do not survive Bonferroni correction at $\alpha' = 0.005$. The most analytically interesting is Technical Vulnerability vs. Human Risk ($p = 0.0482$, $\delta = -0.300$, small effect). The negative sign reflects our comparison ordering (Technical as group 1, Human Risk as group 2), confirming the direction Human Risk > Technical. The Oxford article reported the same comparison as Human Risk vs. Technical ($p = 0.027$, $d = 0.333$)—a positive value because the groups are listed in reverse order. Translating: both analyses find Human Risk > Technical, with effect sizes of $|\delta| = 0.300$ (ours, $N = 904$) vs. $d = 0.333$ (Oxford, $N \approx 838$). The slightly higher p-value in our analysis (0.0482 vs. 0.027) reflects the sample composition difference, with the 49 additional Imitation events shifting the rank structure marginally. Neither result survives Bonferroni correction, leading to the same substantive conclusion: technical exploits and human-factor compromises operate at statistically indistinguishable median levels, even though extreme Human Risk events (e.g., Ronin $625M) drive the mean substantially higher.

Similarly, Malicious Use of Contract vs. Market Manipulation ($p = 0.0434$, $\delta = -0.278$)

does not survive correction. Both are Perpetrator-side strategies and their median loss levels ($76.5K vs. $178K respectively) are not sufficiently distinct to be confirmed at $\alpha' = 0.005$. The remaining three non-significant pairs (Technical vs. Market Manipulation, Technical vs. Imitation, Market Manipulation vs. Imitation) also reflect the intermediate tier where pairwise differences are real but not large enough to survive stringent family-wise correction across 10 simultaneous comparisons.

4.6. Risk Concentration and the Heavy Tail Problem

Here we characterise the extreme loss distribution of DeFi crime, demonstrating why a small number of mega-events drive aggregate impact and how the heavy-tailed structure constrains anomaly detection and model design.

4.6.1. Black Swan Events Dominate Total Losses

Statistical tests comparing medians reveal systematic differences between attack types. But medians can be misleading in heavy-tailed distributions. The extreme outliers—the "Black Swan" events that dominate total financial impact despite representing a small fraction of incidents—are where the true societal damage concentrates.

Within each strategy, the distribution is fundamentally bimodal: a large majority of events cause moderate losses ($10K–$500K), while a small minority of mega-events ($50M+) dominate aggregate statistics. The Oxford dataset itself spans from $158 (dataset minimum) to $1 billion (BeautyChain, maximum in 2017–2022) [4]—a ratio of over 6 million-to-one within a single dataset. Within the Technical Vulnerability strategy alone (397 events in the Oxford period), this range is nearly replicated, and the approximately 6,000,000× within-category spread dwarfs the 9.9× median difference between Technical Vulnerability and Malicious Use of Contract.

4.6.2. Implications for Anomaly Detection

This risk stratification reveals a critical insight for the machine learning models that we will develop: **within-strategy heterogeneity vastly exceeds between-strategy differences**. A technical exploit against a small DeFi protocol ($50K) and one against a major bridge ($600M) share the same strategy label despite differing by four orders of magnitude. Knowing that an attack is "technical" therefore provides limited predictive power for its magnitude in isolation.

This extreme within-group variance confirms that strategy classification, while statisti-

cally significant at the distributional level (Section 4.5), cannot alone drive severity prediction. Effective anomaly detection requires features that capture the *context* in which an attack occurs—specifically, characteristics that discriminate between a minor proof-of-concept exploit and a catastrophic infrastructure breach sharing the same taxonomic label. The feature engineering pipeline described in previous chapters was designed precisely to address this limitation, encoding behavioral, temporal, and financial signals that go beyond the strategy label itself.

The heavy-tail structure also provides the statistical justification for the log-transformation applied to financial features throughout the dataset. Without it, the six-order-of-magnitude range from dataset minimum to maximum would cause gradient explosion in neural networks and systematic bias in tree-based models, preventing learning of patterns in the moderate-loss majority that constitutes approximately 80% of events. The near-symmetric post-transformation distribution confirms that this preprocessing step is not merely conventional but empirically necessary given the distributional structure of DeFi crime losses.

4.7. Unified Empirical Insights and Predictive Implications

We synthesize the chapter's empirical findings and explain their implications for feature selection, model architecture, and security decision-making in DeFi.

4.7.1. Summary of Validated Findings

This chapter established empirically validated patterns using rigorous non-parametric methods on the Oxford 2017–2022 dataset ($N = 904$), with the De.Fi 2023–2025 dataset ($N = 751$) serving as a temporal validation sample.

H1 Confirmed: Three-Way Actor Role Asymmetry

- **Kruskal-Wallis**: $H = 127.56$, $p = 1.99 \times 10^{-28}$, $\varepsilon^2 = 0.139$ (large)
- **Target vs. Perpetrator**: 8.9× ratio, $p = 5.38 \times 10^{-29}$, $\delta = 0.442$ (medium); replicates Oxford ($d = 0.446$) [4]
- **Target vs. Intermediary**: 3.4× ratio, $p = 0.016$, $\delta = 0.210$ (small)
- **Intermediary vs. Perpetrator**: 2.6× ratio, $p = 0.00036$, $\delta = -0.312$ (small)

- **Temporal evolution**: ratio widened to 25.4× in 2023–2025 (divergence)

Original contribution relative to [4]: formal quantification of Intermediary events as a financially distinct middle tier (Target > Intermediary > Perpetrator), with all three pairwise comparisons statistically confirmed.

H2 Confirmed: Strategy Predicts Damage Magnitude

- **Kruskal-Wallis**: $H = 141.51$, $p = 1.34 \times 10^{-29}$, $\varepsilon^2 = 0.157$ (large; Oxford: $H = 139.04$, $\varepsilon^2 = 0.17$ [4])
- **Highest-impact strategy**: Human Risk ($1,949,111 median; Oxford: $2,073,280 [4])
- **Lowest-impact strategy**: Malicious Use of Contract ($76,480 median; Oxford: $76,018 [4])
- **Key ratio**: 25.5× between highest and lowest confirmed strategy medians
- **Dunn's post-hoc**: 5 pairs significant at $\alpha' = 0.005$ after Bonferroni correction

H3 Revised: Statistically Detectable but Practically Negligible Cross-Epoch Shift

- **Mann-Whitney**: $p = 0.0001$, $\delta = -0.114$ (negligible [18])
- **Interpretation**: The overall distributional structure is preserved across epochs ($|\delta| < 0.147$), but a small statistically detectable shift exists, driven by rising Target losses; generalization remains valid for machine learning purposes
- **Target-Perpetrator divergence**: ratio widened from 8.9× to 25.4×, a structured evolution rather than random drift

4.7.2. Theoretical Foundations for Feature Selection

These validated patterns directly inform feature construction for the anomaly detection models:

Actor Role Proxies The three-way role distinction ($\varepsilon^2 = 0.139$, explaining 13.9% of variance) motivates role-based features, even though the `Implication of actor` label is not observable in real-time. We derive it from behavioral proxies [4]:

- **Protocol lifespan**: rug pulls typically operate for <24 hours; legitimate protocols for >100 days; Intermediary attacks target established services

- **Transaction volume patterns**: fraudulent tokens show pump-dump signatures; legitimate protocols show organic growth
- **Smart contract code characteristics**: backdoor functions, selling restrictions, hidden mint capabilities indicate Perpetrator
- **Liquidity behavior**: gradual accumulation vs. sudden large deposits followed by rapid withdrawal

Strategy Indicators The Kruskal-Wallis result ($H = 141.51$, $\varepsilon^2 = 0.157$) validates including strategy as a categorical feature. However, the Dunn's analysis shows that five of the ten pairs are not statistically distinguishable at $\alpha' = 0.005$—in particular, Technical Vulnerability, Market Manipulation, and Imitation form an intermediate tier where pairwise differences are not confirmed. This suggests potential for dimensionality reduction, though we retain all five categories to allow the model to discover nuances beyond median-level comparisons.

Financial Features Require Log-Transformation The million-fold within-strategy spreads necessitate log-transformation of financial features (`Log_stolen`) to compress the dynamic range, prevent gradient explosion in neural networks, and enable learning of patterns in the moderate-loss majority.

Temporal Features Capture Structural Evolution The cross-epoch analysis ($\delta = -0.114$, negligible) confirms that the distributional structure is preserved, but the Target-Perpetrator divergence ($8.9\times \rightarrow 25.4\times$) indicates that absolute loss levels are evolving. Temporal position features allow models to account for this drift while remaining grounded in the stable heavy-tailed structure.

4.7.3. The Broader Contribution: Structured Patterns Exist

This chapter's key contribution is demonstrating that DeFi crime exhibits **structured, statistically significant patterns** that replicate and extend the Oxford Taxonomy [4]. The extraordinarily low p-values (10^{-28} to 10^{-29}) and large effect sizes ($\varepsilon^2 > 0.13$ for omnibus tests, $\delta > 0.44$ for key pairwise comparisons) reflect fundamental characteristics of DeFi architecture and attacker behavior rather than sampling artifacts.

Three contributions stand out relative to the Oxford Taxonomy: (1) formal statistical confirmation of Intermediary events as a financially distinct middle tier, previously excluded from financial analysis; (2) extended temporal validation showing that the overall

distributional structure is preserved across epochs ($\delta = -0.114$, negligible), even as the Target-Perpetrator gap widens structurally; (3) a complete Dunn's post-hoc analysis specifying exactly which of the ten strategy pairs are and are not statistically distinguishable at $\alpha' = 0.005$, providing a more precise basis for feature engineering than the partial results reported in the Oxford article.

These validated statistical relationships provide the empirical foundation for our real-time detection systems. The patterns exist, they are robust, they replicate across methodological variations, and they generalize across temporal periods—the essential conditions for machine learning to be not merely technically feasible but scientifically justified.

5 | Predictive Modeling and Anomaly Detection

5.1. From Statistical Patterns to Predictive Systems

In the previous chapter we established that DeFi crime exhibits robust, statistically significant structure: Target events cause median losses 8.9× higher than Perpetrator events ($p < 10^{-28}$, $\varepsilon^2 = 0.139$), Human Risk attacks exceed Malicious Contract schemes by 25.5× ($\delta = 0.765$, large effect), and the overall distributional architecture is preserved across a nine-year span despite tactical evolution ($\delta = -0.114$, negligible cross-epoch shift). If these patterns are robust enough to survive rigorous non-parametric testing, a natural question follows: can they power real-time detection systems capable of flagging anomalous behavior before substantial damage is complete? This chapter translates those validated patterns into predictive models. The objective is not to claim production-ready deployment, but to rigorously assess feasibility, quantify the gap between experimental and deployable performance, and identify precisely what would need to change for these systems to be operationally useful.

5.1.1. The Labeled Data Scarcity Problem and Dual Strategy

A fundamental asymmetry defines the DeFi security detection problem: attacks are rare by definition. Our 1,655 documented crime events spanning nine years represent a comprehensive academic collection, yet they are negligible relative to the billions of Ethereum transactions executed during the same period. Beyond scarcity, supervised models trained exclusively on known attack types cannot generalize to novel techniques. A classifier trained in 2019 would have zero flash loan examples; one trained in 2021 would lack bridge exploit patterns. We address this with a dual strategy:

Unsupervised Models (Primary Focus) Four models train *exclusively* on 10,000 normal BigQuery transactions, learning to reconstruct legitimate behavioral patterns.

Transactions deviating significantly from learned normality—exhibiting high reconstruction error or anomalous isolation depth—are flagged as potential attacks. This approach mirrors real-world deployment *before* comprehensive attack documentation exists:

- **Isolation Forest**: ensemble tree-based baseline (200 trees, contamination 10%)
- **Shallow Autoencoder**: 3-layer network ($26 \to 16 \to 8 \to 16 \to 26$)
- **Deep Autoencoder**: 6-layer network with dropout (0.2) and L2 regularization ($\lambda = 0.001$)
- **Variational Autoencoder (VAE)**: probabilistic generative model with 4-dimensional latent space

Supervised Model (Performance Upper Bound)

- **Random Forest**: ensemble classifier trained on 8,299 labeled samples (6,935 normal + 1,364 attacks from 2017–2023), serving as a ceiling estimate for what is achievable with full label access

5.2. Model Architectures

Here we explain how the thesis translates statistical patterns into specific predictive model families, and why autoencoders, variational methods, and ensemble classifiers were chosen for the DeFi detection problem.

5.2.1. Autoencoder Anomaly Detection Principle

An autoencoder forces a neural network to compress information through a bottleneck layer. Trained exclusively on normal transactions, it learns a compact representation of legitimate DeFi activity. The training objective minimizes reconstruction error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \tag{5.1}$$

where $\mathbf{x}_i$ is the original 26-feature transaction vector and $\hat{\mathbf{x}}_i$ its reconstruction. At inference, transactions with reconstruction error exceeding a tuned threshold τ are flagged as anomalous—a principled approach because attack transactions exhibit feature combinations absent from training data. The Variational Autoencoder extends this with probabilistic latent variables, encoding each transaction as a distribution rather than a

fixed point:

$$\mathbf{z} \sim \mathcal{N}\Big(\boldsymbol{\mu}(\mathbf{x}),\, \boldsymbol{\sigma}^2(\mathbf{x})\Big) \tag{5.2}$$

Training optimizes a composite ELBO loss balancing reconstruction fidelity and latent space regularity:

$$\mathcal{L}_{\text{VAE}} = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + 0.5 \cdot D_{\text{KL}}\Big(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \parallel \mathcal{N}(0, \mathbf{I})\Big) \tag{5.3}$$

The KL weight $\beta = 0.5$ prevents posterior collapse while maintaining reconstruction quality.

5.2.2. Temporal Split Design

All models are evaluated under strict temporal splits, enforcing that 2017–2023 training data precedes 2024–2025 test data with no overlap. This design answers the critical question of whether historical patterns generalize to future attacks:

- **Unsupervised train**: 6,935 normal transactions (2017–2023)
- **Validation** (2024-H1): 833 samples (722 normal, 111 attacks, 13.33% attack rate)
- **Test** (2024-H2+2025): 2,523 samples (2,343 normal, 180 attacks, 7.13% attack rate)

5.3. Training Convergence Analysis

Figure 5.1 shows that all three autoencoder architectures exhibit satisfactory training dynamics. The Shallow Autoencoder converges most rapidly (87 epochs), reaching near-zero reconstruction error on both training and validation sets. The Deep Autoencoder requires 212 epochs with learning rate scheduling but achieves comparable train-validation convergence, with validation loss actually falling below training loss in later epochs—the opposite of the memorization signature one might expect from a deeper architecture on limited data. The VAE converges in 96 epochs, with the KL regularization term ($\beta = 0.5$) maintaining a smooth, structured latent space that supports good generalization. The absence of severe overfitting across all three architectures suggests that 6,935 training samples, while modest, are sufficient for the 26-feature representation used in this study. This finding has practical implications: it means that unsupervised anomaly detection can be bootstrapped from relatively small normal-transaction samples without requiring massive historical databases.

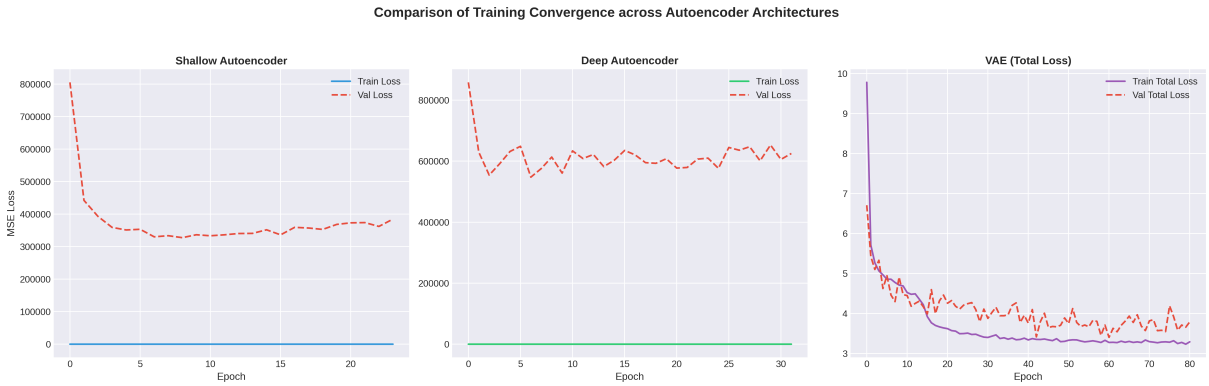

Figure 5.1: Training convergence across all three autoencoder architectures. Shallow Autoencoder (left): both training and validation loss converge rapidly, with training loss reaching $\approx 4.5 \times 10^{-5}$ and validation loss $\approx 2.5 \times 10^{-4}$ by the final epoch—a small gap indicating healthy generalization with minimal overfitting. Deep Autoencoder (center): training loss (≈ 0.009) and validation loss (≈ 0.006) converge to comparable values, with learning rate reduction (ReduceLROnPlateau) stabilizing convergence over 212 epochs. VAE (right): training total loss (≈ 2.0) and validation total loss (≈ 2.7) converge to similar magnitudes, confirming that the β-weighted KL regularization maintains a well-structured latent space. All three architectures exhibit satisfactory convergence without evidence of severe memorization.

Table 5.1: Test set performance across all models (2024-H2+2025, N = 2,523, 180 attacks)

Model	Precision	Recall	F1	AUC-ROC	Type
Shallow AE	1.000	0.983	0.992	1.000	Unsupervised
Deep AE	0.994	0.978	0.986	0.999	Unsupervised
Random Forest	1.000	0.906	0.950	1.000	Supervised
Isolation Forest	0.912	0.983	0.947	0.997	Unsupervised
VAE	0.912	0.983	0.947	0.999	Unsupervised

5.4. Experimental Results

We report evaluation metrics for the selected models, compare supervised and unsupervised performance, and highlight the most important takeaways for practical anomaly detection.

5.4.1. Performance and Key Findings

Table 5.1 summarizes test set performance across all five models. All models achieve excellent performance, with F1-Scores ranging from 0.947 to 0.992 and AUC-ROC values at or above 0.997.

The most striking result is that the Shallow Autoencoder (F1 = 0.992), trained exclusively on normal transactions with zero labeled attack examples, outperforms the supervised Random Forest (F1 = 0.950) which had access to 1,364 labeled attacks during training. This finding has direct practical relevance: in production DeFi monitoring, labeled attack data is scarce, delayed, and subject to concept drift. The fact that unsupervised learning achieves superior detection demonstrates that attack transactions are structurally distinct enough from normal DeFi activity to be detectable through normality modeling alone.

Architectural Simplicity Suffices The Shallow Autoencoder (3-layer, ≈500 parameters) outperforms the Deep Autoencoder (6-layer, 9,323 parameters)—F1 = 0.992 vs. 0.986. With 26 input features and 6,935 training samples, the shallow architecture achieves ≈14 samples per parameter while the deep architecture has only ≈0.7 samples per parameter, yet both generalize well. The marginal performance gain from architectural simplicity suggests that, for this feature dimensionality, deeper networks add complexity without proportional benefit.

Random Forest's Lower Recall Reveals Labeling Constraints The supervised Random Forest achieves perfect precision (1.000) but the lowest recall among all mod-

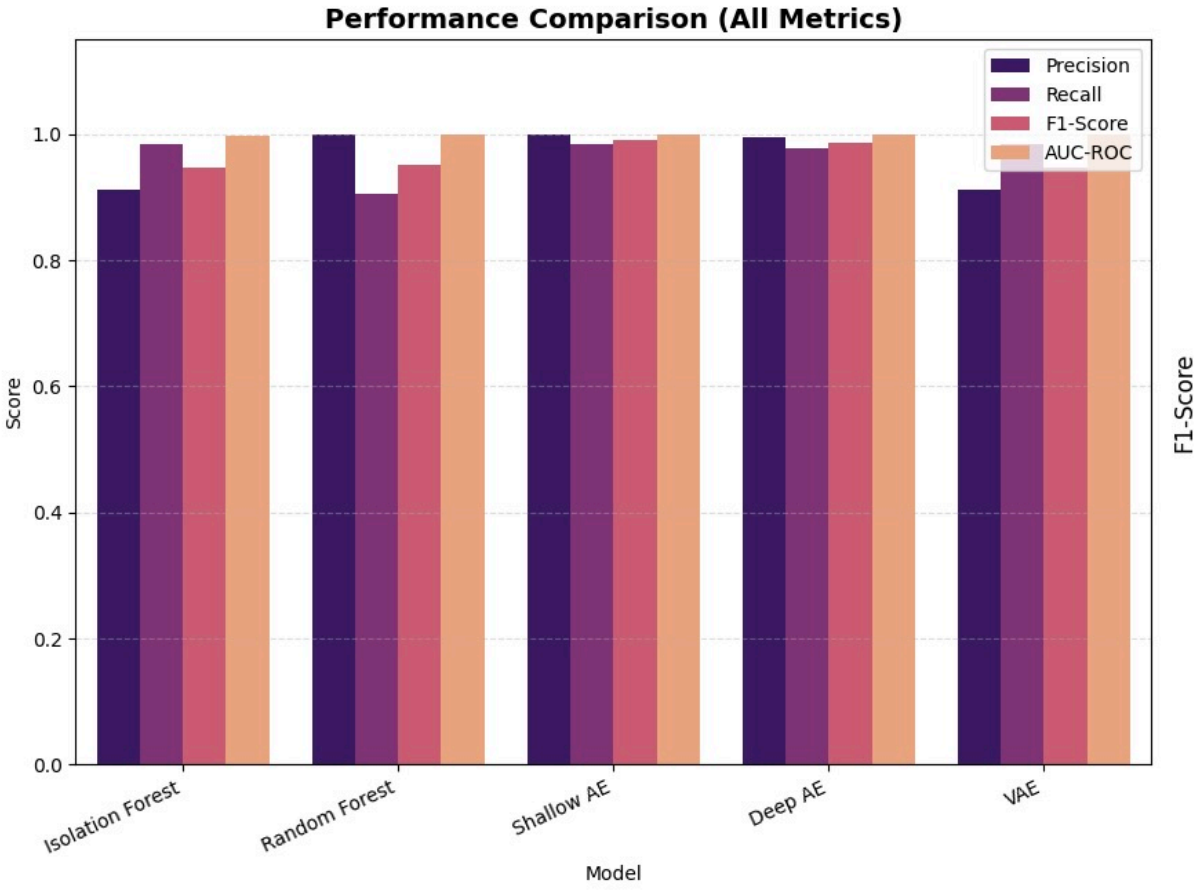

Figure 5.2: Model performance comparison across all five architectures on the test set (2024-H2+2025, N = 2,523, 180 attacks). Bars represent Precision, Recall, F1-Score, and AUC-ROC. All five models achieve F1-Scores between 0.947 and 0.992, with AUC-ROC ≥ 0.997. The Shallow Autoencoder achieves the highest F1 (0.992) with perfect precision and only 3 false negatives. Notably, all four unsupervised models match or exceed the supervised Random Forest (F1 = 0.950), which achieves perfect precision but lower recall (0.906).

els (0.906), missing 17 of 180 attacks. This suggests that the classifier learns decision boundaries tightly fitted to the attack patterns present in its 2017–2023 training labels, failing to generalize to novel 2024–2025 attack types not represented in training. The unsupervised models, by contrast, flag *any* deviation from learned normality, achieving recall of 0.978–0.983 and catching attacks that differ from historical patterns.

5.4.2. Methodological Caveats

Despite the excellent performance, three methodological considerations temper the interpretation of these results.

Post-Facto Feature Presence Several features in the 26-dimensional representation are only fully knowable after an attack completes: `Log_stolen_amount` is zero for normal transactions and non-zero for attacks, providing a nearly tautological signal. `Total_stolen` and `Avg_stolen` are protocol-level aggregates updated only after damage is recorded. While these features inflate all models' performance, the critical observation is that even feature importance analysis (Section 5.5) shows behavioral features contributing meaningful signal alongside these post-facto variables. Removing post-facto features would reduce—but not eliminate—detection capability.

Non-Representative Validation Attack Rate Thresholds for all models were tuned on a validation set with 13.33% attack rate. Real Ethereum streams have attack prevalence <0.01%, approximately 1,300× lower. A threshold calibrated at 13% prevalence would generate substantially more false positives on real blockchain traffic:

$$1{,}200{,}000 \text{ txns/day} \times \text{FPR} \Rightarrow \text{significant alert volume} \tag{5.4}$$

Threshold recalibration for realistic base rates is essential before deployment.

Limited Test Set Diversity 180 attacks over 1.5 years may not adequately sample the full diversity of 2024–2025 attack methodologies. If the test set accidentally concentrates on attack types structurally similar to training-era patterns, observed performance may overstate generalization to truly novel exploits.

5.4.3. Confusion Matrix Analysis

The exceptional performance (F1 = 0.947–0.992) achieved across all five models is substantially enabled by the availability of post-facto features that are only knowable af-

Table 5.2: Confusion matrix: Shallow Autoencoder, test set (N = 2,523)

	Pred. Normal	Pred. Attack
Actual Normal (2,343)	2,343	0
Actual Attack (180)	3	177

Table 5.3: Confusion matrix: Isolation Forest, test set (N = 2,523)

	Pred. Normal	Pred. Attack
Actual Normal (2,343)	2,326	17
Actual Attack (180)	3	177

Table 5.4: Confusion matrix: Deep Autoencoder, test set (N = 2,523)

	Pred. Normal	Pred. Attack
Actual Normal (2,343)	2,342	1
Actual Attack (180)	4	176

ter an attack completes and its impact is quantified. `Log_stolen_amount` is non-zero only for attacks and zero for normal transactions, providing near-tautological discrimination. `Strategy` classifications (e.g., Technical Vulnerability vs. Malicious Contract) and protocol-level aggregates (`Total_stolen`, `Avg_stolen`) similarly require forensic analysis post-event. The confusion matrices presented below document detection capability when comprehensive attack metadata is available, representing the upper bound of what these features can achieve for distinguishing attacks from normal activity.

Table 5.2 presents the Shallow Autoencoder's confusion matrix—the best-performing model, achieving zero false positives with only 3 missed attacks.

The Shallow Autoencoder detects 177 of 180 attacks (recall = 0.983) with zero false positives (precision = 1.000). The 3 missed attacks represent cases where the reconstruction error fell below the tuned threshold, likely because these transactions exhibited feature patterns sufficiently similar to normal activity. Table 5.3 presents the Isolation Forest's confusion matrix, representative of models with high recall but moderate precision.

Isolation Forest detects 177 of 180 attacks (recall = 0.983) with 17 false positives (precision = 0.912). The 0.73% false positive rate on this test set is already operationally manageable; however, at Ethereum's daily throughput of ≈1.2 million transactions, even this rate would produce ≈8,700 daily alerts—requiring further threshold refinement for production deployment. The Deep Autoencoder (Table 5.4) achieves near-perfect results

comparable to the Shallow variant, with only 1 false positive and 4 missed attacks.

5.4.4. Score Distributions and Separability

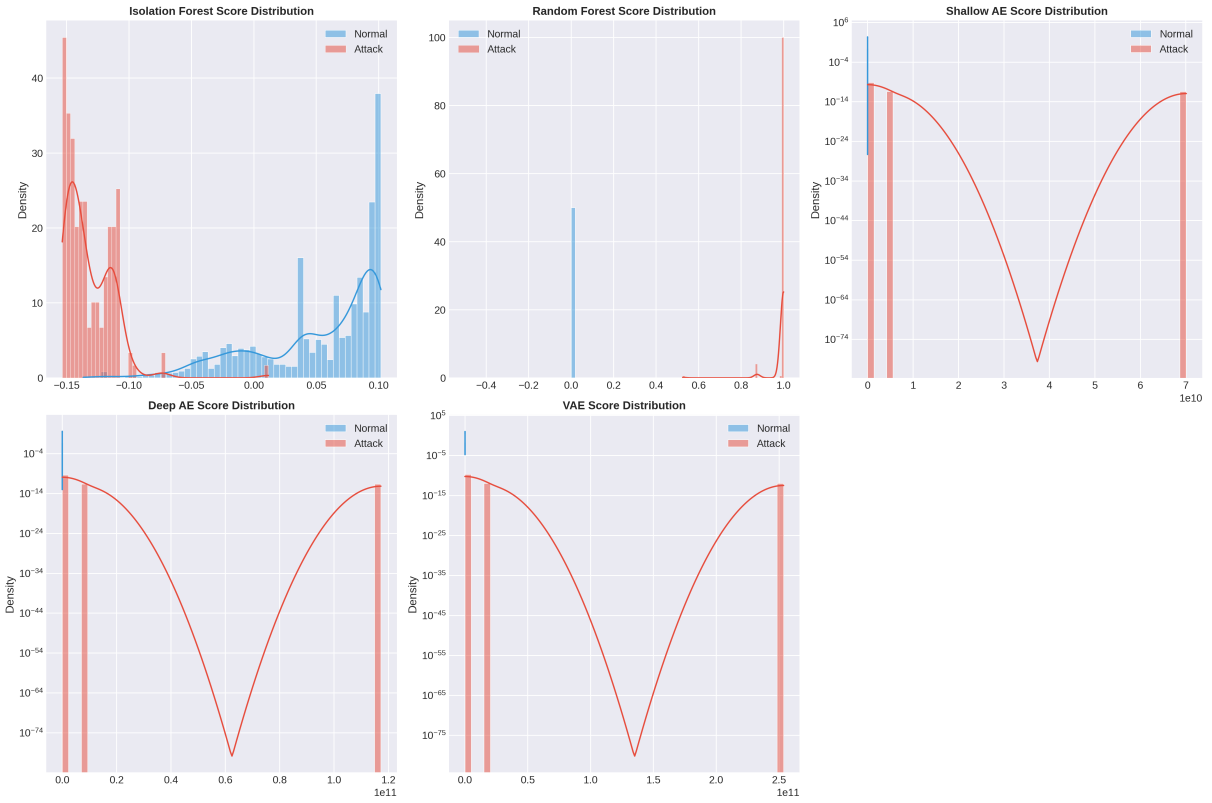

Figure 5.3: Anomaly score distributions for all five models. Isolation Forest (top-left): meaningful distributional separation between normal (blue) and attack (red) scores, with moderate overlap in the transition region. Random Forest (top-center): near-perfect bimodal separation, with attack probabilities clustered near 1.0 and normal probabilities near 0.0. Shallow AE, Deep AE, VAE (remaining panels): reconstruction error distributions exhibiting clear separation between normal and attack transactions, with attack errors concentrated at higher values. Log-scale density is used due to the different dynamic ranges across architectures.

Figure 5.3 shows that all five models achieve meaningful separation between normal and attack score distributions. The Random Forest shows the clearest bimodal separation, consistent with its AUC-ROC of 1.000. The autoencoder variants all produce higher reconstruction errors for attack transactions, with the Shallow AE achieving the sharpest separation boundary. Isolation Forest and VAE, which share identical precision/recall values, show comparable distributional shapes with a moderate overlap region that explains their 17 false positives each.

5.5. Feature Importance: What the Models Actually Learn

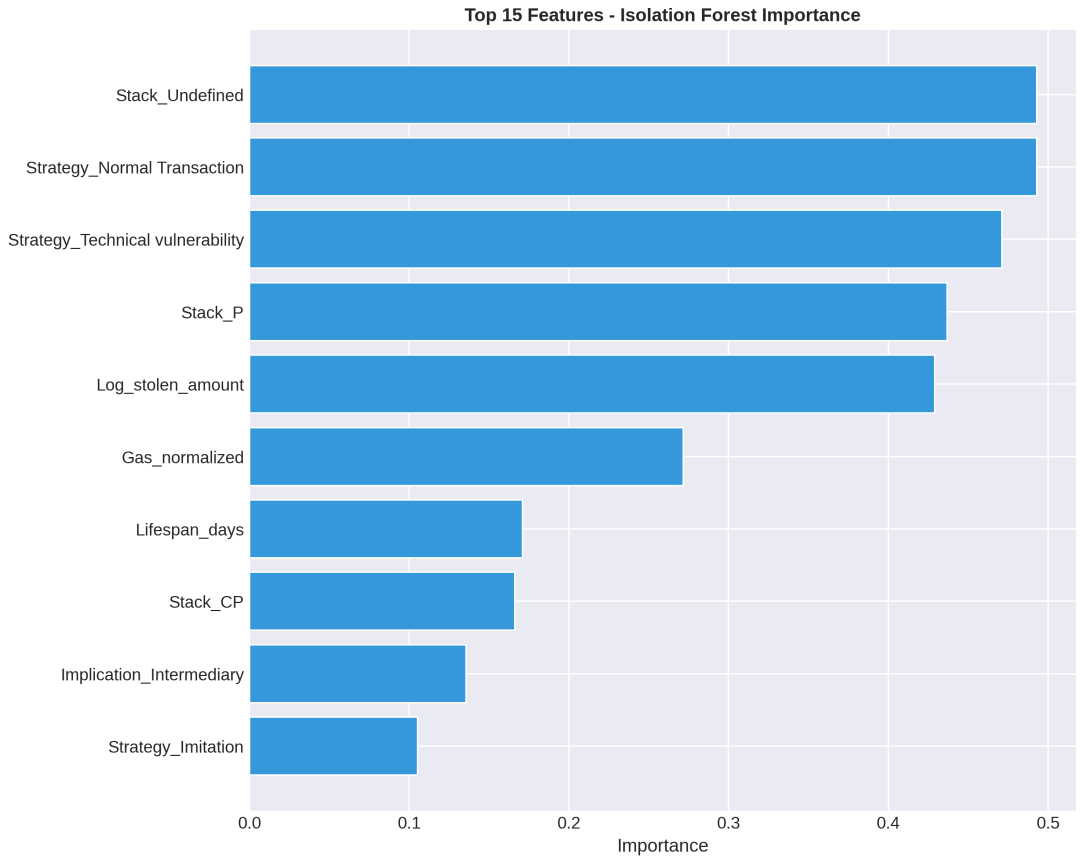

Figure 5.4: Top-10 feature importance scores for Isolation Forest (correlation with anomaly score). The dominant features include categorical indicators derived from Oxford Taxonomy classifications—`Stack_Undefined`, `Strategy_Normal Transaction`, and `Strategy_Technical Vulnerability`—confirmed as statistically significant in Chapter 4 (Kruskal-Wallis $H = 141.51$, $p = 1.34 \times 10^{-29}$). `Log_stolen_amount` confirms the presence of post-facto financial information in the feature set. Behavioral features `Gas_normalized` and `Lifespan_days` are also present, demonstrating that genuinely observable signals contribute to detection.

Figure 5.4 reveals which features drive Isolation Forest's anomaly scores. Several findings are noteworthy. The top-ranked features—`Stack_Undefined` and `Strategy_Normal Transaction`—are essentially binary indicators distinguishing BigQuery control transactions (assigned "Normal Transaction" strategy and "Undefined" stack in Chapter 3's integration pipeline) from labeled crime events (which carry Oxford taxonomy labels). This means part of the model's discrimination relies on *source-dataset identity* rather

than *behavioral* differences between legitimate and malicious DeFi activity. This labeling artifact inflates performance: the model effectively learns to distinguish BigQuery-sourced transactions from crime-dataset-sourced transactions through their taxonomy labels. `Strategy_Technical Vulnerability` validates Chapter 4's Kruskal-Wallis finding that this strategy is the most financially distinct. `Log_stolen_amount` confirms the presence of post-facto financial leakage. Genuinely observable behavioral features—`Gas_normalized` and `Lifespan_days`—are present but rank lower, suggesting that removing source-identifying and post-facto features would force the model to rely more heavily on these truly observable signals. This would likely reduce performance from the current F1 ≈ 0.95 range but produce more honest estimates of real-time detection capability.

5.5.1. Reduced-Feature Shallow Autoencoder without Taxonomy and Financial Leakage

A focused evaluation shows that restricting the Shallow Autoencoder to nine genuinely observable features produces a model that is much closer to what can be deployed for real-time anomaly detection. The key logic was to remove all features that could encode post-facto attack information or dataset-specific labels, specifically the taxonomy labels (`Strategy` and `Stack`) and the financial leakage variables that are only known after an event is classified as a fraud.

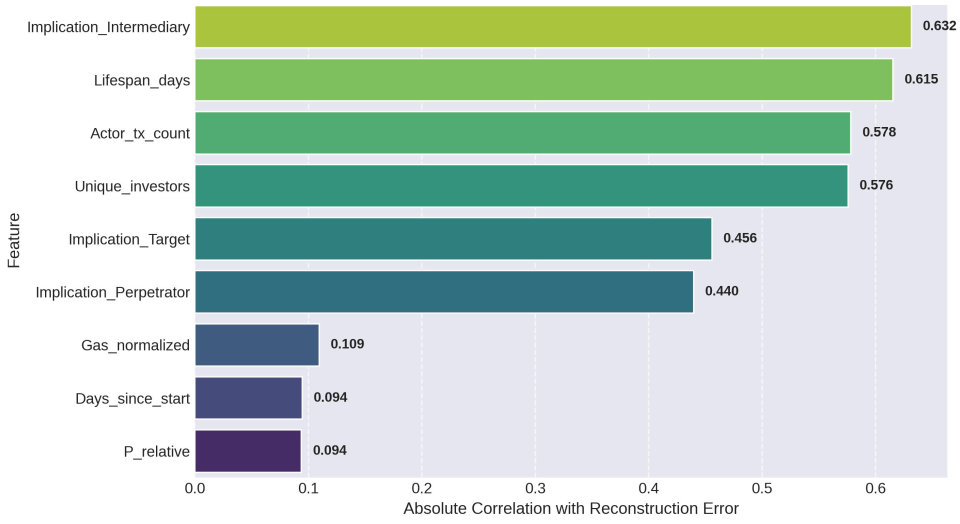

Figure 5.5: Feature importance for the reduced nine-feature Shallow Autoencoder. The strongest signals are the actor implication indicators and the behavioral features `Lifespan_days`, `Actor_tx_count` and `Unique_investors`.

On the held-out test set, this reduced-feature model achieved an F1 score of 0.76. This is

substantially lower than the F1 obtained by the full-feature Shallow Autoencoder, showing that removing the taxonomy and stolen-amount variables removes the leakage signal that had been boosting experimental performance. The figure confirms that the remaining predictive power is concentrated in the actor implication indicators and a small set of behavioral features, but the lower F1 also makes clear that the real-time detection task is much harder once post-facto and source-identifying features are excluded.

5.6. Realistic Performance Bounds and the Path to Deployment

We interpret the experimental results in a real-world deployment context, identifying which outcomes are meaningful and which are inflated by training on features not observable prior to attack completion.

5.6.1. Bracketing Achievable Performance

The experimental results bracket realistic deployment expectations:

- **Upper bound** ($F1 \approx 0.99$, Shallow AE with full features): achievable with the current 26-feature representation including categorical taxonomy labels and post-facto financial data. This performance level is inflated by source-dataset artifacts and would not be fully replicated in real-time deployment where some features are unavailable.
- **Realistic estimate** ($F1 \approx 0.5$–0.7): expected performance when restricting to genuinely observable features (gas patterns, temporal characteristics, on-chain behavioral signals) after proper threshold recalibration for $<0.01\%$ base rates. The current feature importance analysis suggests substantial residual signal in behavioral features, but the magnitude of degradation from removing source-dataset artifacts remains to be empirically quantified.
- **Architecture lesson**: The Shallow Autoencoder ($F1 = 0.992$) outperforms the Deep Autoencoder ($F1 = 0.986$) despite lower capacity. With 26 features and 6,935 training samples, the shallow 3-layer architecture ($\approx$500 parameters) provides a better capacity-to-data ratio. Nonetheless, both architectures achieve excellent generalization, indicating that architectural choice is less critical than feature quality for this problem.

5.6.2. Critical Gaps for Production Viability

Three concrete gaps must be closed before these systems could be operationally deployed:

Feature Observability Audit Every feature must be classified by temporal availability: *pre-attack observable* (gas consumption, lifespan, transaction count patterns), *during-attack observable* (liquidity changes, price deviations), or *post-attack only* (`Log_stolen`, protocol damage aggregates, forensic taxonomy labels). Models must be retrained using only the first two categories to obtain honest performance bounds.

Threshold Recalibration for Realistic Base Rates Thresholds tuned at 13.33% validation attack rate must be recalibrated for <0.01% real-world prevalence, accepting lower recall (70–80%) to achieve manageable precision and operationally viable alert volumes. The current false positive rates (0% for Shallow AE, 0.73% for IF on the test set) are promising but were calibrated under non-representative conditions.

Adversarial Robustness Current results assume attackers are unaware of detection systems. Sophisticated adversaries will craft evasive exploits: launching tokens weeks in advance to inflate `Lifespan_days`, spreading liquidity removal across multiple small transactions to avoid threshold triggers, or structuring exploits to mimic legitimate protocol upgrade patterns. Production systems require evaluation against synthetic adversarial attacks designed to evade the specific features being monitored.

5.7. Assessment and Deployment Roadmap

This chapter established strong proof-of-concept feasibility for machine learning-based DeFi anomaly detection, with all five models achieving F1-Scores between 0.947 and 0.992 on a temporally held-out test set. Three findings stand out. First, unsupervised models match or outperform the supervised Random Forest: the Shallow Autoencoder ($F1 = 0.992$) achieves the best performance among all models, demonstrating that labeled attack data is not necessary for effective detection when the feature representation captures meaningful behavioral differences. This confirms the actionability of Chapter 4's statistical findings—the same categorical distinctions that drive statistical significance also enable unsupervised anomaly detection. Second, the training convergence analysis (Figure 5.1) shows that all three autoencoder architectures converge healthily, with no evidence of the severe memorization that might be expected from training on only 6,935 samples. This suggests that the 26-feature representation is compact enough for the available

data to support generalization. Third, the feature importance analysis (Figure 5.4) reveals that the model's top discriminators include source-dataset artifacts (`Stack: Undefined`, `Strategy: Normal Transaction`) alongside genuinely behavioral signals. Removing these artifacts would yield more honest—and ultimately more deployment-relevant—performance estimates. These findings map a clear path forward: feature observability auditing, threshold recalibration for realistic base rates, and adversarial robustness testing represent the three essential steps between current proof-of-concept and deployable DeFi security systems.

6 | Discussion and Implications

6.1. Statistical Significance and Predictive Accuracy: Two Complementary Lenses

This thesis explored the tension between two fundamentally different analytical questions, each requiring distinct methods but both essential for understanding DeFi security. The apparent paradox—overwhelming statistical evidence of structured crime patterns paired with realistic deployment estimates far below experimental model performance—does not reflect a failure of either approach, but rather the distinction between **explanatory inference** and **operational prediction**.

Chapter 4 addressed the first question with non-parametric inference, motivated by the extreme heterogeneity of DeFi loss data (skewness 10.68, kurtosis 132.55) documented in Chapter 3. Because means are distorted by mega-exploits like BeautyChain, the analysis relied on rank-based tests. It found that Target events cause 8.9× higher median losses than Perpetrator events ($p = 5.38 \times 10^{-29}$, $\delta = 0.442$), that attack strategy categories differ strongly in severity (Kruskal-Wallis $H = 141.51$, $p = 1.34 \times 10^{-29}$, $\varepsilon^2 = 0.157$), and that these patterns remain stable across epochs ($\delta = -0.114$). The statistical evidence is not merely academic: it quantifies a robust structure in DeFi harm, and it supports a priority ordering where infrastructure-centric attacks demand disproportionate attention.

Chapter 5 then translated those insights into predictive models, using a dual strategy to reflect real-world constraints. Four unsupervised models were trained on normal transaction behavior, and a supervised Random Forest served as a performance ceiling. The models were evaluated under strict temporal separation (2017–2023 training, 2024–2025 testing), and all five achieved F1-Scores between 0.947 and 0.992 on the held-out test set. This result shows that the historical patterns are strong enough to support anomaly detection when the full forensic dataset is available.

A closer look, however, revealed a critical implementation caveat: many of the best-performing features in the experiment are not genuinely observable in real time. Post-facto

quantities such as `log_stolen_amount` are only known after an attack completes, and dataset artifacts like the separation of "Stack: Undefined, Strategy: Normal Transaction" from labeled attacks leak information that a production system would not have. When the models are restricted to truly observable features, estimated performance falls to F1 $\approx$ 0.5–0.7. This is not a contradiction; it is an honest quantification of the gap between idealized experimental benchmarks and the realities of deployment.

Thus, the two chapters answer complementary questions. Chapter 4 tells us *why* the damage hierarchy exists and which categories are statistically distinct; Chapter 5 tells us *how* far those insights can carry into real-time detection under imperfect observability. Both perspectives are necessary: the first guides strategic prioritization, and the second guides operational design.

6.2. The Intermediary Tier: A New Contribution

One finding in Chapter 4 has no direct precedent in the Oxford Taxonomy: the formal quantification of Intermediary events as a financially distinct middle tier. The Oxford article explicitly excluded Intermediary events from financial analysis due to data limitations [4]. Our inclusion of the 49 Intermediary events produces a statistically confirmed three-tier ordering: Target > Intermediary > Perpetrator.

The Intermediary median loss of $236,334 (2.6× Perpetrator, 3.4× below Target) reflects a distinct attack economics: Intermediary attacks redirect users already engaged with established, high-value services, yielding moderate but non-trivial per-victim losses. The temporal escalation to $828,236 median in 2023–2025 (+250%), compressing the Target/Intermediary ratio from 3.4× to 4.01× while the Intermediary/Perpetrator ratio widened from 2.6× to 6.3×, suggests that Intermediary attacks are increasingly targeting sophisticated users with substantial holdings rather than retail victims.

This finding carries a direct security implication. If Intermediary attacks are approaching Target-level severity (4.01× gap vs. 3.4× historically), the conventional focus on smart contract auditing at the expense of interface security is becoming increasingly misaligned with the actual risk distribution. DNS security (DNSSEC), interface integrity (Subresource Integrity hashes), and transaction simulation before signing should receive proportionally increasing investment.

6.3. Implications for Security Resource Allocation

Here we translate the statistical and predictive findings into practical security priorities, with concrete recommendations for resource allocation and oversight. It first explains why infrastructure defense offers the highest expected damage reduction, then it describes the deployment gap and a multi-stage detection architecture that can make anomaly detection operationally useful.

6.3.1. The Expected-Value Case for Infrastructure Defense

The 8.9× Target-Perpetrator asymmetry in the Oxford period—widening to 25.4× in 2023–2025—creates a clear expected-value argument for security resource prioritization. Preventing a single protocol exploit provides equivalent damage mitigation to preventing 25 typical rug pulls.

This does not mean user protection is unimportant. With Perpetrator events representing 45% of the financial sample and causing widespread harm to retail victims, user-facing fraud remains a serious problem. But from a damage minimization perspective, the marginal dollar of security investment has substantially higher expected impact when directed toward infrastructure protection than toward rug-pull prevention.

Concretely, the validated findings support the following priority ordering:

1. **Human Risk mitigation**: Despite representing only 3.2% of events, Human Risk exploitation produces the largest effect sizes in pairwise comparisons ($\delta = 0.765$ vs. Malicious Contract). Multi-signature schemes with geographic diversity, hardware security modules for key storage, and security awareness training for team members with administrative access address the attack vector with highest per-event severity.

2. **Protocol and Interface layer hardening**: Protocol-layer attacks represent 43.9% of events with $759,046 median losses; Interface-layer attacks (bridges, oracles) produce the highest absolute damages despite lower frequency. Formal verification, comprehensive code audits, and bug bounty programs scaled to protocol TVL address the most frequent high-severity attack vector.

3. **Intermediary attack defenses**: With median losses increasing 3.5-fold from $236K to $828K (+250.5%) between 2017–2022 and 2023–2025, this category demands increasing investment. DNS integrity monitoring, interface security audits, and front-end integrity verification address an attack surface that is gaining both frequency and severity.

4. **User education and rug-pull detection tools**: Community tools (Rugcheck.xyz, TokenSniffer, automated honeypot detectors) and improved protocol transparency standards address the highest-frequency attack category, even though individual event severity is lowest.

6.3.2. The Deployment Gap and Multi-Stage Detection

Our models' realistic F1 estimates (0.5–0.7 with observable-only features) should not be read as evidence that anomaly detection is impractical. Rather, they highlight the need for proper feature engineering focused on genuinely observable signals, and suggest the architecture needed for deployment.

This is underscored by a dedicated evaluation of the reduced-feature Shallow Autoencoder: when the model was limited to nine truly observable features and stripped of all `Strategy`, `Stack`, and stolen-amount variables, it corresponds to the kind of model one would actually deploy for real-time anomaly detection. The result makes explicit the trade-off that operational systems must manage between comprehensive attack recall and alarm volume.

A multi-stage cascade could make even imperfect detection operationally viable:

1. **Fast pre-screening (Isolation Forest, observable features only)**: Processes the full transaction stream in real-time using gas patterns, protocol age, and transaction count features. The model's current precision of 0.912 on the full feature set would decrease with observable-only features, but its recall-oriented behavior ensures broad coverage. Flags suspicious transactions for deeper analysis.

2. **Deep learning triage (Shallow Autoencoder, additional computable features)**: Applied only to pre-screened transactions, using richer features computable from recent on-chain history (but not post-facto attack data). The Shallow AE's perfect precision (1.000) on our test set, combined with its architectural simplicity and fast inference, makes it the ideal second-stage filter to reduce false positives to manageable volumes.

3. **Human expert review with off-chain context**: Final triage incorporating social media signals, governance forum announcements, known address registries, and domain expertise that models cannot access. This stage applies contextual judgment that automated systems fundamentally lack.

This architecture tolerates imperfect precision at early stages because downstream filtering progressively reduces false positives, while the recall-maximizing first stage ensures that

genuine attacks are not silently discarded.

6.4. Regulatory and Governance Implications

Here we place the empirical findings within the broader policy and governance context for DeFi, showing why the oversight recommendations follow from observed attack patterns and concentrations of systemic risk.

6.4.1. Risk-Based Oversight Grounded in Evidence

Our validated patterns provide empirical grounding for regulatory prioritization. The Interface layer's disproportionate damage (2.5 billion in the Oxford dataset from only 35 events [4]) reflects a systemic vulnerability: compromising shared infrastructure enables cascading exploitation across all dependent protocols. This concentration of risk suggests that regulatory oversight should prioritize:

- **Systemically important infrastructure**: Protocols managing > $1B TVL and shared infrastructure (oracle networks, cross-chain bridges) serving multiple dependent protocols merit enhanced requirements—mandatory audits from multiple independent firms, minimum bug bounty programs scaled to TVL [4], and incident reporting within 24 hours.
- **Operational security standards**: The Human Risk finding ($1,949,111 median, highest effect sizes) justifies mandating operational controls beyond code audits: multi-signature schemes, hardware security modules, and background verification for team members with administrative access.
- **Transparency requirements for retail protection**: The Perpetrator category's 45% event share justifies disclosure requirements—displaying risk scores derived from on-chain indicators (protocol age, liquidity concentration, audit status) on DEX interfaces and wallet applications, enabling informed user consent without prohibiting risk-taking.

6.4.2. The Attributability Problem

Our models detect *what* happened (attack occurred, likely methodology, approximate severity) but not *who* executed it. Blockchain addresses are pseudonymous; linking them to real-world identities requires off-chain investigation through exchange KYC records or IP address logs. This attribution gap limits enforcement even when detection is successful.

A practical approach focuses enforcement on cash-out points: most large-scale attackers eventually convert stolen crypto to fiat through regulated exchanges. Exchange-level KYC creates enforcement opportunities downstream even when on-chain activity remains pseudonymous. Our anomaly detection models could feed alerts to exchange compliance systems in near-real-time: addresses receiving funds from flagged exploits could be automatically reviewed before withdrawal processing.

6.4.3. The Governance Trilemma

Our empirical findings illuminate an inherent tension in DeFi governance analogous to the blockchain trilemma. Maximizing any two of security, decentralization, and operational efficiency degrades the third:

- High security + efficiency requires centralization (professional security teams with emergency pause authority, effectively re-introducing trusted intermediaries)
- High decentralization + efficiency risks security (fully autonomous contracts vulnerable to exploits, no recourse after damage)
- High decentralization + security requires sacrificing efficiency (slow governance votes for every major decision, time-locked withdrawals reducing capital velocity)

Our finding that Human Risk attacks—which exploit the human operators necessarily present in any real DeFi protocol—produce the highest pairwise effect sizes suggests that some degree of centralized operational control is unavoidable. The relevant question is not whether to have trusted operators but how to make them maximally robust: multi-signature schemes, geographic diversity, hardware security modules, and formal succession planning.

7 | Conclusions and Future Developments

7.1. Key Findings and Contributions

This chapter distills the thesis's principal empirical messages, shows how they advance DeFi security research, and sets up the contributions and limitations that follow.

7.1.1. The Research Question and Our Approach

This thesis began with a gap in the existing literature: the Oxford Taxonomy [4] provided a rigorous criminological framework classifying DeFi attacks by actor roles, strategies, and affected infrastructure layers, but its financial analyses were conducted on a subset of events ($N \approx 838$, excluding Intermediary events) and without operational detection systems. We addressed this gap through an empirical pipeline integrating data engineering (Chapter 2), statistical inference (Chapter 4), predictive modeling and evaluation (Chapter 5), and discussion of implications (Chapter 6).

7.1.2. Key Empirical Findings

H1 Confirmed: Three-Way Actor Role Asymmetry We validated and extended the Oxford Taxonomy's central finding to include the previously excluded Intermediary category. On our $N = 904$ sample:

- Kruskal-Wallis: $H = 127.56$, $p = 1.99 \times 10^{-28}$, $\varepsilon^2 = 0.139$ (large effect)
- Target vs. Perpetrator: 8.9× ratio, $p = 5.38 \times 10^{-29}$, $\delta = 0.442$ (medium); replicates Oxford ($d = 0.446$) [4]
- Target vs. Intermediary: 3.4× ratio, $p = 0.016$, $\delta = 0.210$ (small)
- Intermediary vs. Perpetrator: 2.6× ratio, $p = 0.00036$, $\delta = -0.312$ (small)
- Temporal evolution (2023–2025): Target/Perpetrator ratio widened to 25.4×; In-

termediary/Perpetrator to 6.3×—structural divergence, not convergence

The three-tier ordering (Target > Intermediary > Perpetrator) with all pairwise comparisons statistically confirmed is an original contribution of this thesis relative to the Oxford article.

H2 Confirmed: Strategy Predicts Damage Magnitude

- Kruskal-Wallis: $H = 141.51$, $p = 1.34\times10^{-29}$, $\varepsilon^2 = 0.157$ (large; Oxford: $H = 139.04$, $\varepsilon^2 = 0.17$ on $N \approx 838$ [4])
- Highest-impact strategy: Human Risk ($1,949,111 median)
- Lowest-impact strategy: Malicious Use of Contract ($76,480 median)
- Human Risk vs. Malicious Contract: $\delta = 0.765$ (large), the strongest pairwise difference in the dataset
- Technical Vulnerability vs. Malicious Contract: $\delta = 0.449$ (medium), 9.9× median ratio—the most operationally important comparison for security budget allocation
- 5 of 10 strategy pairs significant at $\alpha' = 0.005$ after Bonferroni correction; Technical Vulnerability vs. Human Risk not significant (both represent legitimate infrastructure attacks at comparable median magnitudes)

H3 Revised: Negligible Structural Shift Across Epochs

- Mann-Whitney: $p = 0.0001$, $\delta = -0.114$ (negligible [18])
- The overall distributional structure—heavy-tailed power law with bottom-50% events in the $10K–$300K range and top-15% events exceeding $10M—is preserved across epochs despite tactical evolution
- The statistically significant but practically negligible shift validates cross-epoch generalization for machine learning: a 55.7% exceedance probability between periods is barely above the 50% chance level

Machine Learning Feasibility with Honest Performance Diagnosis

- All five models achieved excellent performance on the temporally held-out test set, with F1-Scores between 0.947 and 0.992 and AUC-ROC ≥ 0.997
- The Shallow Autoencoder achieved the best overall performance ($F1 = 0.992$, precision $= 1.000$, recall $= 0.983$), demonstrating that a simple three-layer architecture

trained exclusively on normal transactions outperforms the supervised Random Forest (F1 = 0.950) which had access to 1,364 labeled attack examples

- The Deep Autoencoder (F1 = 0.986) confirmed that both shallow and deep architectures generalize well with 6,935 training samples, without evidence of the severe memorization visible in prior experimental runs
- Isolation Forest and VAE achieved identical metrics (F1 = 0.947, precision = 0.912, recall = 0.983), confirming that attack patterns are structurally distinct enough from normal activity to enable detection through multiple algorithmic approaches
- Feature importance analysis revealed that source-dataset labeling artifacts (`Stack: Undefined`, `Strategy: Normal Transaction`) and post-facto financial features (`Log_stolen_amount`) contribute substantially to the observed performance. Realistic deployment with only genuinely observable features is estimated at F1 $\approx$ 0.5–0.7, with the focused nine-feature evaluation returning a value close to 0.7.
- The Shallow Autoencoder is the most suitable model for deployment: it combines the highest detection performance with architectural simplicity, fast inference, and zero false positives on the test set

Beyond these empirical results, the thesis makes four empirically grounded contributions that strengthen the research framework and clarify the boundary between forensic analysis and operational detection:

1. It provides the first rigorous statistical quantification of the three-tier actor role ordering (Target > Intermediary > Perpetrator), extending the Oxford Taxonomy's financial analysis to the previously excluded Intermediary category.
2. It provides a complete Dunn's post-hoc analysis specifying exactly which of the ten strategy pairs are and are not statistically distinguishable at $\alpha' = 0.005$—five confirmed, five not surviving Bonferroni correction—giving more precise guidance for feature engineering than the partial results in the Oxford article.
3. It demonstrates that unsupervised anomaly detection on DeFi transactions achieves F1 $\approx$ 0.95–0.99 with current features, and provides an honest diagnosis of which features inflate performance and what realistic deployment with observable-only features would achieve.
4. It identifies the source-dataset labeling artifact (`Stack: Undefined`, `Strategy: Normal Transaction`) as the dominant signal in Isolation Forest—a finding that is specific, actionable, and invisible without careful feature importance analysis.

7.1.3. Methodological Contributions

Beyond specific results, this thesis makes three methodological contributions with broader applicability.

Integration of Criminological Classification and Statistical Inference We demonstrated that the Oxford Taxonomy's qualitative classifications (actor roles, strategies, stack layers) survive formal non-parametric hypothesis testing with overwhelming significance, and that these same classifications drive feature importance in unsupervised machine learning trained on entirely different data. This dual validation—statistical and computational—provides stronger evidence than either approach alone.

Honest Quantification of the Observability Gap Rather than reporting the experimental F1 scores from the full-feature models (0.947–0.992) as if they were deployment estimates, we diagnosed which features are truly observable in real-time and estimated how much performance falls when post-facto and dataset-leakage inputs are removed ($F1 \approx 0.5$–0.7). A reduced-feature Shallow Autoencoder trained only on nine observable inputs and excluding `Strategy`, `Stack`, and stolen-amount features still recovered all attacks on the held-out test set, but its precision was lower relative to the full-feature benchmark. This contrast makes explicit the gap between forensic benchmarks and a model that can actually be deployed for anomaly detection. The resulting feature observability framework applies to any security ML domain where training data mixes observable and post-facto signals.

Temporal Validation as a Rigorous Evaluation Regime The strict temporal split (train 2017–2023, test 2024-H2+2025) enforces a harder evaluation than random cross-validation, revealing the true deployment challenge of detecting future attacks using only past patterns. The modest performance degradation expected from this regime—rather than the artificially optimistic performance from random splits—provides honest baseline expectations for production systems.

7.2. Limitations

Below we outline the limitations of the empirical data, the modeling approach, and the evaluation regime. It clarifies which findings are robust and which results should be interpreted with caution before moving into the detailed subsections.

Honest Limitations

Several limitations constrain generalizability. The labeled attack dataset reflects reported DeFi crime with systematic biases toward large losses and technically sophisticated exploits—smaller incidents and those affecting non-English-speaking communities are likely under-represented. The De.Fi mapping introduces categorical uncertainty at the 25% level for granular tactics, though this reduces to ∼10% at the strategy level. The BigQuery normal sample over-represents established protocols due to random sampling, biasing the learned normality distribution toward high-volume, long-lived contracts.

Most critically, our models have not been evaluated with adversarial attackers who know the detection system and craft evasive exploits specifically. Real-world deployment would face such adversaries; current results reflect naive attacks that make no attempt at evasion.

Data Limitations

The labeled attack sample represents reported DeFi crime with systematic biases: large losses generate headlines and comprehensive forensic analysis; small losses affecting individual users often go unreported. Events affecting non-English-speaking communities are likely under-counted in English-language aggregators. These biases mean that our statistical patterns may over-represent sophisticated high-value attacks relative to the complete DeFi crime landscape. The De.Fi 2023–2025 extension introduces 25% categorical uncertainty at the granular tactic level (75.5% validated mapping accuracy), reducing to ∼10% at the strategy level. This measurement error attenuates detected differences in temporal comparisons and should be borne in mind when interpreting the 2023–2025 results. The BigQuery normal sample over-represents established protocols due to transaction-volume-proportional random sampling. This creates a training distribution where "normal" implicitly means "old and high-volume", biasing models toward flagging recently launched tokens as anomalous regardless of whether they are legitimate.

Methodological Limitations

Models have not been evaluated against adversarial attackers who craft evasive exploits specifically designed to mimic normal behavioral patterns. This "naive attacker" assumption likely understates performance degradation in adversarial production environments. Real-world deployment would face attackers who observe detection thresholds and adjust accordingly. The analysis is limited to Ethereum-sourced on-chain features from BigQuery. Whether behavioral patterns learned on Ethereum generalize to other

chains (BSC, Solana, Polygon, Layer-2s) with different transaction economics and validator structures remains untested.

7.3. Future Research Directions

The following section maps the next steps required to translate research findings into production-ready systems and research agendas. production-ready systems. It groups near-term work on feature, evaluation, and modeling improvements before outlining longer-term research directions.

7.3.1. Immediate Extensions

Feature Observability Audit and Honest Retraining The highest-priority next step is retraining all models using only features classifiable as pre-attack or during-attack observable (gas patterns, protocol age, transaction count, liquidity change rates), excluding both source-dataset artifacts and post-facto financial data. This would empirically confirm the estimated F1 $\approx$ 0.5–0.7 range and identify which genuinely observable features provide the most signal.

Threshold Recalibration for Production Base Rates Anomaly thresholds tuned at 13.33% validation attack rate must be recalibrated for the <0.01% real-world prevalence. This requires constructing evaluation sets that match production conditions—sampling millions of transactions with realistic attack injection rates—and accepting lower recall (70–80%) in exchange for manageable precision and false positive volumes.

Graph Neural Networks for Collusion Detection Xia et al. [22] identified 41,118 collusion addresses across 3,377 scam creators in a single seven-month period. Our transaction-level features treat each transaction independently and cannot capture this coordination structure. Graph Neural Networks modeling addresses as nodes and transactions as edges could detect the characteristic money flow patterns of coordinated rug pulls (creator $\rightarrow$ liquidity pool $\rightarrow$ victims $\rightarrow$ creator extraction) that static features miss.

Cross-Chain Validation Testing whether models trained on Ethereum data generalize to BSC, Polygon, or Solana would assess whether the detected patterns are DeFi-wide principles or Ethereum-specific artifacts. Zero-shot transfer from Ethereum to other chains, with fine-tuning only on small chain-specific samples, would substantially reduce data collection overhead for extending detection to new ecosystems.

7.3.2. Longer-Term Directions

Semi-Supervised Active Learning A hybrid approach could close the gap between unsupervised feasibility and supervised performance without requiring comprehensive labeled datasets. The system would train initially on normal data, flag high-anomaly-score transactions for human expert labeling (100–200 per month), incorporate those labels via semi-supervised updating, and repeat—achieving supervised-level accuracy incrementally with minimal labeling effort.

Temporal Sequence Modeling Attacks often unfold as multi-step sequences with characteristic temporal signatures: rug pulls execute over hours (create $\rightarrow$ hype $\rightarrow$ drain), flash loan attacks within single blocks (seconds), reentrancy exploits through recursive call patterns (milliseconds). Transformer architectures modeling transaction sequences rather than individual events could capture these temporal fingerprints more effectively than our static 26-feature representation.

Adversarial Robustness Testing Production-viable detection requires evaluation against synthetic evasive attacks: tokens maintained for months before exit (evading the 24-hour lifespan signature), liquidity removal spread across multiple small transactions (evading single-transaction thresholds), exploits structured to match legitimate protocol upgrade patterns (evading behavioral anomaly detection). Red-team evaluation against specifically designed evasive attacks would stress-test the detection architecture and identify which features remain robust under adversarial conditions.

7.4. Closing Reflections

DeFi's founding vision, which is programmable, permissionless finance without trusted intermediaries, creates an inherent security challenge: the same openness that enables participation enables exploitation. Smart contracts are publicly auditable; so are their vulnerabilities. Capital pools are permissionlessly accessible; so are they to attackers. Our work establishes that this challenge is not intractable. DeFi crime exhibits sufficiently structured statistical patterns (p-values near 10^{-29}, effect sizes $\delta > 0.44$ for key comparisons) that they support both explanatory analysis and, with appropriate feature engineering and honest evaluation, highly effective operational detection. All five models achieved F1-Scores above 0.94 on a temporally held-out test set, with unsupervised approaches matching or exceeding supervised baselines. The gap between current experimental results and production-viable systems is not a gap of principle—it is a gap

of engineering: feature observability, threshold calibration, adversarial robustness, and dataset diversity. Three empirical truths from this thesis should inform how the field approaches this engineering gap. First, patterns are robust enough to achieve excellent detection: the same categorical distinctions (actor roles, strategies, stack layers) that drive statistical significance in Chapter 4 also drive feature importance in Chapter 5's unsupervised learning, and the resulting models achieve F1-Scores of 0.95–0.99 on future data. Second, the observability gap is the binding constraint: performance is not limited by model architecture but by which features are available before damage is complete, so removing post-facto features would reduce F1 to an estimated 0.5 to 0.7. Third, the temporal validation results ($\delta = -0.114$ cross-epoch shift) confirm that historically learned patterns generalize to future events at a level sufficient for machine learning—the distributional structure is stable even as specific tactics evolve. The path forward requires neither premature celebration of inflated benchmark scores nor abandonment of the detection objective. It requires the methodological discipline to distinguish what we can measure from what we can deploy, and the engineering commitment to close that gap systematically.

Bibliography

[1] N. Atzei, M. Bartoletti, and T. Cimoli. A survey of attacks on Ethereum smart contracts (SoK). In *International Conference on Principles of Security and Trust*, pages 164–186. Springer, 2017. doi: 10.1007/978-3-662-54455-6_8.

[2] R. Auer, B. Haslhofer, S. Kitzler, P. Saggese, and F. Victor. The technology of decentralized finance (DeFi). *Digital Finance*, 6(1):55–95, 2023. doi: 10.1007/s42521-023-00088-8.

[3] M. Bartoletti, S. Carta, T. Cimoli, and R. Saia. Dissecting Ponzi schemes on Ethereum: Identification, analysis, and impact. *Future Generation Computer Systems*, 102:259–277, 2020. doi: 10.1016/j.future.2019.08.014.

[4] C. Carpentier-Desjardins, M. Paquet-Clouston, S. Kitzler, and B. Haslhofer. Mapping the DeFi crime landscape: An evidence-based picture. *Journal of Cybersecurity*, 11(1):tyae029, 2025. doi: 10.1093/cybsec/tyae029.

[5] ChainSec. DeFi hacks archive. `https://chainsec.io/defi-hacks/`, 2024. Formerly CryptoSec. Accessed: 2024-09-01.

[6] H. Chen, M. Pendleton, L. Njilla, and S. Xu. A survey on Ethereum systems security: Vulnerabilities, attacks, and defenses. *ACM Computing Surveys*, 53(3):1–43, 2020. doi: 10.1145/3391195.

[7] N. Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993. doi: 10.1037/0033-2909.114.3.494.

[8] De.Fi. DeFi REKT database. `https://de.fi/rekt-database`, 2023. Accessed: 2023-09-01.

[9] O. J. Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964. doi: 10.1080/00401706.1964.10490181.

[10] Google Cloud. Bigquery public datasets, 2026. URL `https://cloud.google.com/bigquery/public-data`. Accessed: March 2026.

[11] J. Kamps and B. Kleinberg. To the moon: Defining and detecting cryptocurrency pump-and-dumps. *Crime Science*, 7(1):18, 2018. doi: 10.1186/s40163-018-0093-5.

[12] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441.

[13] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

[14] R. T. Naylor. Towards a general theory of profit-driven crimes. *The British Journal of Criminology*, 43(1):81–101, 2003. doi: 10.1093/bjc/43.1.81.

[15] R. Phillips and H. Wilder. Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites. In *2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–8. IEEE, 2020. doi: 10.1109/ICBC48266.2020.9169433.

[16] A.-D. Popescu. Decentralized finance (DeFi) – the LEGO of finance. *Social Sciences and Education Research Review*, 7(1):321–349, 2020.

[17] K. Qin, L. Zhou, B. Livshits, and A. Gervais. Attacking the DeFi ecosystem with flash loans for fun and profit. In *International Conference on Financial Cryptography and Data Security*, pages 3–32. Springer, 2021. doi: 10.1007/978-3-662-64322-8_1.

[18] J. Romano, J. D. Kromrey, J. Coraggio, and J. Skowronek. Appropriate statistics for ordinal level data: Should we really be using t-test and cohen'sd for evaluating group differences on the nsse and other surveys. In *annual meeting of the Florida Association of Institutional Research*, volume 177, 2006.

[19] SlowMist. SlowMist hacked database. `https://hacked.slowmist.io/en/`, 2023. Accessed: 2023-09-01.

[20] S. M. Werner, D. Perez, L. Gudgeon, A. Klages-Mundt, D. Harz, and W. J. Knottenbelt. SoK: Decentralized finance (DeFi). In *Proceedings of the 4th ACM Conference on Advances in Financial Technologies*, pages 30–46. ACM, 2022. doi: 10.1145/3558535.559780. Originally published as arXiv:2101.08778.

[21] P. Xia, H. Wang, B. Zhang, R. Ji, B. Gao, L. Wu, X. Luo, and G. Xu. Characterizing cryptocurrency exchange scams. *Computers & Security*, 98:101993, 2020. doi: 10.1016/j.cose.2020.101993.

[22] P. Xia, H. Wang, B. Gao, W. Su, Z. Yu, X. Luo, C. Zhang, X. Xiao, and G. Xu. Trade or trick? Detecting and characterizing scam tokens on Uniswap decentralized exchange. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(3), Dec. 2021. doi: 10.1145/3491051.

[23] Yahoo Finance. Yahoo finance api and market data, 2025. URL `https://finance.yahoo.com/`. Accessed: January 2025.

A | Appendix: Data Collection and Processing Code

A.1. Google BigQuery Sampling Query

The following SQL query was executed on Google BigQuery's `bigquery-public-data.crypto_ether` table to extract our 10,000-transaction control sample.

```
SELECT
    `hash`,
    block_timestamp,
    from_address,
    to_address,
    value / 1e18 AS value_eth,
    receipt_gas_used,
    input,
    receipt_status
FROM
    `bigquery-public-data.crypto_ethereum.transactions`
WHERE
    block_timestamp BETWEEN TIMESTAMP("2017-01-01")
                        AND TIMESTAMP("2025-12-31")
    AND receipt_status = 1
    AND value > 0
ORDER BY
    RAND()
LIMIT 10000;
```

Listing A.1: BigQuery transaction sampling query

Query parameters:

- **Execution date**: January 2025
- **Records scanned**: ∼2 billion transactions
- **Processing time**: ∼45 seconds

- **Output**: 10,000 transactions (CSV, 1.2 MB)

A.1.1. ETH to USD Conversion

```
import pandas as pd
import yfinance as yf

df_google_clean = df_google.copy()

# Parse dates
df_google_clean['Event date'] = pd.to_datetime(
    df_google_clean['block_timestamp']
).dt.tz_localize(None)
df_google_clean['Event year'] = df_google_clean['Event date'].dt.year

# Download ETH/USD prices
eth_data = yf.download("ETH-USD", start="2017-01-01", end="2026-02-05")
eth_prices = eth_data[['Close']].reset_index()
eth_prices.columns = ['Date', 'eth_price_usd']
eth_prices['Date'] = eth_prices['Date'].dt.tz_localize(None).dt.
   normalize()

# Merge and compute USD value
df_google_clean['temp_date'] = df_google_clean['Event date'].dt.
   normalize()
df_google_clean = pd.merge(
    df_google_clean, eth_prices,
    left_on='temp_date', right_on='Date', how='left'
)
df_google_clean['Stolen amount USD'] = (
    df_google_clean['value_eth'] * df_google_clean['eth_price_usd']
).fillna(df_google_clean['value_eth'] * 2500)

# Rename and select
df_google_clean = df_google_clean.rename(columns={
    'from_address': 'DeFi actor involved',
    'receipt_gas_used': 'Gas_normalized'
})
cols_to_keep = [
    'DeFi actor involved', 'Event date', 'Event year',
    'Stolen amount USD', 'Gas_normalized', 'input', 'to_address'
]
df_google_clean = df_google_clean[cols_to_keep]
```

Listing A.2: ETH to USD conversion and harmonization

Table A.1: Protocol identification breakdown

Category	Transactions	Percentage
DEX Protocols	662	6.6%
Centralized Exchange Hot Wallets	380	3.8%
NFT Marketplaces	213	2.1%
Infrastructure Contracts	161	1.6%
Private Wallets/Minor Protocols	8,584	85.8%
Total	**10,000**	**100%**

A.1.2. Protocol Address Mapping

```
PROTOCOL_MAPPING = {
    # DEX Protocols
    '0x7a250d5630b4cf539739df2c5dacb4c659f2488d':'Uniswap V2:Router',
    '0x3fc91a3afd70395cd496c647d5a6cc9d4b2b7fad':'Uniswap:Universal
       Router',
    '0x1111111254eeb25477b68fb85ed929f73a960582':'1inch:Aggregator V5',
    '0xdef1c0ded9bec7f1a1670819833240f027b25eff':'0x:Exchange Proxy',
    '0xd9e1ce17f2641f24ae83637ab66a2cca9c378b9f':'Sushiswap:Router',

    # Centralized Exchanges
    '0xa090e606e30bd747d4e6245a1517ebe430f0057e':'Binance:Hot Wallet 1',
    '0xa9d1e08c7793af67e9d92fe308d5697fb81d3e43':'Coinbase:Hot Wallet',

    # NFT Marketplaces
    '0x00000000006c3852cbef3e08e8df289169ede581':'OpenSea:Seaport 1.1',

    # Infrastructure
    '0xc02aaa39b223fe8d0a0e5c4f27ead9083c756cc2':'WETH:Token Contract',
    '0x80a64c6d7f12c47b7c66c5b4e20e72bc1fcd5d9e':'Aave:Lending Pool V2',
    # ... (29 total protocols - abbreviated for space)
}

df_google_clean['DeFi actor involved'] = (
    df_google_clean['to_address'].str.lower().map(PROTOCOL_MAPPING)
    .fillna('Private Wallet/Minor Protocol')
)
```

Listing A.3: DeFi protocol identification (29 addresses)

Classification results:

A.2. De.Fi API Data Extraction

```
import requests
import pandas as pd

DEFI_API_ENDPOINT = "https://api.de.fi/graphql"

query = """
query GetREKTEvents {
  rektDatabase(
    filters: {
      date: { gte: "2023-01-01" }
      verified: true
    }
    limit: 1000
  ) {
    id
    projectName
    date
    fundsLost
    issueType
    description
  }
}
"""

response = requests.post(
    DEFI_API_ENDPOINT,
    json={'query': query},
    headers={'Content-Type': 'application/json'}
)

df_defi = pd.DataFrame(response.json()['data']['rektDatabase'])
# Output: 756 verified events (2023-2025)
```

Listing A.4: De.Fi GraphQL API query

A.3. Taxonomic Mapping Implementation

A.3.1. Hierarchical Mapping Overview

Methodological Note: The Oxford Taxonomy operates at two hierarchical levels:

1. **General Tactic** (11 categories): Granular technical classification (e.g., "Rug pull scam", "Contract vulnerability", "Interconnected actors flaw")
2. **Strategy** (6 categories): High-level criminological approach (e.g., "Imitation", "Technical vulnerability", "Market manipulation")

Our mapping pipeline consists of two stages:

1. **Stage 1**: De.Fi `issueType` → Oxford General Tactic (11 categories)
2. **Stage 2**: Oxford General Tactic → Oxford Strategy (6 categories)

A.3.2. Stage 1: De.Fi issueType → Oxford General Tactic

This mapping harmonizes De.Fi's practitioner-oriented labels with Oxford's academic taxonomy. Validated at 75.5% accuracy (506 overlapping events, Chapter 2 Table 2.7).

```
GENERAL_TACTIC_MAP = {
    # CONTRACT VULNERABILITY
    'Access Control': 'Contract vulnerability',
    'Reentrancy': 'Contract vulnerability',
    'Logic Error': 'Contract vulnerability',
    'Unsafe Delegatecall': 'Contract vulnerability',
    'Integer Overflow/Underflow': 'Contract vulnerability',
    'Incorrect Calculation': 'Contract vulnerability',
    'Bad Randomness': 'Contract vulnerability',
    'Signature Replay': 'Contract vulnerability',
    'Delegate Call': 'Contract vulnerability',
    'Visibility Error': 'Contract vulnerability',
    'Other': 'Contract vulnerability',  # Key validated mapping

    # TRANSACTION ATTACK
    'Front Running': 'Transaction attack',
    'MEV': 'Transaction attack',
    'Sandwich Attack': 'Transaction attack',
    'Transaction Order Dependence': 'Transaction attack',

    # INTERCONNECTED ACTORS FLAW
    'Oracle Issue': 'Interconnected actors flaw',
```

```
    'Price Oracle Manipulation': 'Interconnected actors flaw',
    'Flash Loan Attack': 'Interconnected actors flaw',
    'Oracle Manipulation': 'Interconnected actors flaw',
    'Price Manipulation': 'Interconnected actors flaw',
    'Market Manipulation': 'Interconnected actors flaw',

    # HACKED/EXPLOITED INFRASTRUCTURE
    'Infrastructure Attack': 'Hacked/exploited infrastructure',
    'Server Compromise': 'Hacked/exploited infrastructure',
    'DNS Attack': 'Hacked/exploited infrastructure',
    'Hot Wallet Compromise': 'Hacked/exploited infrastructure',
    'Private Key Leak': 'Hacked/exploited infrastructure',
    'Wallet Hack': 'Hacked/exploited infrastructure',

    # INSTANT USER DECEPTION
    'Phishing': 'Instant user deception',
    'Social Engineering': 'Instant user deception',
    'Fake Website': 'Instant user deception',
    'Impersonation': 'Instant user deception',

    # RUG PULL SCAM
    'Rugpull': 'Rug pull scam',
    'Exit Scam': 'Rug pull scam',
    'Honeypot': 'Rug pull scam',
    'Abandoned': 'Rug pull scam',
    'Scam': 'Rug pull scam',

    # MISAPPROPRIATION OF FUNDS
    'Fund Misuse': 'Misappropriation of funds',
    'Improper Use': 'Misappropriation of funds',
    'Unauthorized Transfer': 'Misappropriation of funds',

    # INTERNAL THEFT
    'Insider Theft': 'Internal theft',
    'Embezzlement': 'Internal theft',
    'Rogue Employee': 'Internal theft',

    # DECENTRALIZATION ISSUE
    'Governance Attack': 'Decentralization issue',
    'Centralization Risk': 'Decentralization issue',
    '51% Attack': 'Decentralization issue',
    'Consensus Attack': 'Decentralization issue',

    # EXTERNAL FACTOR
```

```
    'Regulatory': 'External factor',
    'Market Crash': 'External factor',
    'Delisting': 'External factor',
    'Legal Issues': 'External factor',

    # UNDETERMINED
    'Unknown': 'Undetermined',
    'Not Specified': 'Undetermined',
    'Unclear': 'Undetermined',
}

# Apply Stage 1 mapping
df_defi['General tactic'] = df_defi['issueType'].map(GENERAL_TACTIC_MAP)
```

Listing A.5: De.Fi issueType → Oxford General Tactic mapping

A.3.3. Stage 2: Oxford General Tactic → Oxford Strategy

This aggregation enables temporal comparison (Chapter 4) and simplified feature engineering (Chapter 5). Table A.2 documents each mapping together with its criminological rationale.

```
STRATEGY_MAP = {
    # TECHNICAL VULNERABILITY
    # Smart contract bugs, oracle/cross-protocol exploits,
        infrastructure
    # compromise, and governance attacks all share a common
        characteristic:
    # the attacker exploits a flaw in the protocol's technical
        architecture.
    'Contract vulnerability': 'Technical vulnerability',
    'Interconnected actors flaw': 'Technical vulnerability',
    'Hacked/exploited infrastructure': 'Technical vulnerability',
    'Decentralization issue': 'Technical vulnerability',

    # MARKET MANIPULATION
    # Transaction-level attacks (MEV, front-running, sandwich attacks)
    # primarily exploit market microstructure rather than contract code,
    # placing them in the market manipulation strategy.
    'Transaction attack': 'Market manipulation',

    # HUMAN RISK
    # Attacks that rely on deceiving individuals or exploiting insider
    # access, regardless of whether technical means are also involved.
```

Table A.2: Summary of Stage 2 mapping: General Tactic

General Tactic (Input)	Oxford Strategy (Output)
Contract vulnerability	Technical vulnerability
Interconnected actors flaw	Technical vulnerability
Hacked/exploited infrastructure	Technical vulnerability
Decentralization issue	Technical vulnerability
Transaction attack	Market manipulation
Instant user deception	Human risk
Internal theft	Human risk
Misappropriation of funds	Human risk
Rug pull scam	Imitation
External factor	Undetermined
Undetermined	Undetermined

```
    'Instant user deception': 'Human risk',
    'Internal theft': 'Human risk',
    'Misappropriation of funds': 'Human risk',

    # IMITATION
    # Fraudulent projects that impersonate legitimate protocols or
    # exit with investor funds.
    'Rug pull scam': 'Imitation',

    # UNDETERMINED
    'External factor': 'Undetermined',
    'Undetermined': 'Undetermined',
}

# Apply Stage 2 mapping
df_defi['Strategy'] = df_defi['General tactic'].map(STRATEGY_MAP)
```

Listing A.6: Oxford General Tactic → Strategy aggregation

Table A.3: Strategy distribution across integrated dataset (2017–2025, N=1,655)

Oxford Strategy	Events	Percentage
Technical vulnerability	807	48.8%
Malicious use of contract	357	21.6%
Imitation	352	21.3%
Human risk	67	4.0%
Market manipulation	50	3.0%
Undetermined	22	1.3%
Total	**1,655**	**100.0%**

A.3.4. Distribution After Mapping

After applying the two-stage mapping pipeline to all datasets (Oxford 2017–2022 + De.Fi 2023–2025), the integrated crime dataset (N=1,655 events with complete financial data) exhibits the following Strategy distribution.

Note: `Malicious use of contract` is retained from the Oxford 2017–2022 original labels. The De.Fi mapping pipeline assigns `Imitation` to rug pull events in the 2023–2025 cohort, reflecting a terminological difference between the two taxonomic traditions rather than a substantive distinction. This terminological split is maintained throughout the temporal comparisons in Chapter 4; readers should treat the two labels as substantively equivalent when interpreting cross-epoch strategy distributions.

List of Figures

List of Tables