Discovering Hidden Structures in Stock Market Data using Algorithmic Generative Modeling

Yeu Wen Mak

yeuwen.mak@u.nus.edu

Abstract

This paper demonstrates using elementary cellular automata (ECA) to model observed (O) stock market data by generating (G) candidate data matched using a two-phase genetic algorithm (GA). The O data comprises encoded price change information for selected stocks. In the first phase, GA selects rows/columns to compress O and G arrays using minimal information loss selection (MILS). The second phase applies GA to identify optimal ECA rules that minimize the algorithmic information distance between compressed O and G. Causal decomposition techniques are then used to break down the emergent complexity of the best-matching ECA rules into modular components governing stock dependencies. Analyzing these modular interactions provides insights into hidden structures driving market relationships. Quantifying model fit and information preservation validates the ECA modelling approach coupled with causal decomposition. Elementary cellular automata generated by single rule (131) and double rule pair (35, 115) were found to be the closest match to the real market data.

Keywords: *Algorithmic Complexity, Elementary Cellular Automata, Lossy Compression, Genetic Algorithm, High Order Dependencies*

1. Introduction

We know (Zenil and Delahaye 4), (Brandouy et al. 9) that, despite the apparent randomness of the stock market, it may be considered a rule-based algorithmic machine, not unlike a giant cellular automaton (Wolfram 431) and that computational methods such as algorithmic probability may be useful for analysing and predicting market behavior. Computational methods such as causal decomposition could uncover the elemental mechanisms governing complex dependencies between stocks. This involves breaking down emergent complexity into modular rule components driving relationships using techniques tailored for cellular automata systems.

Current methods for studying market behaviour rely on statistical methods or stochastic models when algorithmic techniques are able to cope with complex systems and spurious correlations in big data better as compared to correlation and regression methods (Zenil 23) (Matejkaet et al. 1). Also, separately, (Mansilla 4) and (Brandouy et al. 8) estimated the algorithmic complexity of financial data to detect hidden structures and avoid focusing on non-profitable patterns.

Although literature has shown financial data complexity can be measured via algorithmic information theory, this has not been established for higher dimensional data capable of capturing complex high-order dependencies that arise due to factors like feedback loops, nonlinear interactions, and long-term memory effects as they evolve over time. This paper introduces an innovative algorithmic generative modeling technique using cellular automata coupled with causal decomposition to address this gap. By matching generated multi-dimensional data to real financial data on algorithmic information metrics, the approach can uncover hidden structures neglected by traditional techniques limited in modeling complex dependencies.

It is our hypothesis that if we represent a stock price change by a series of 0s and 1s as a string of binary sequence for a group of stocks for which their relationships are of interest and use it as an initial cell configuration of a cellular automaton, we can observe patterns in their evolution over time. This can be achieved by a family of algorithms based on algorithmic probability and information theory that can minimize the loss of algorithmic information and avoid certain distortions that occur in other approaches, making them useful for studying complex datasets like stocks dependencies. These algorithms can help us understand the patterns and rules that govern high order dependencies between stocks. They can also be used for dimension reduction, feature selection, and network sparsification while minimizing the loss of algorithmic information.(Zenil et al. 17)

1.1. Goals and Objectives

The overall goal is to apply and extend the algorithmic framework (Brandouy et al. 341) pioneered by Zenil to financial time series analysis in order to uncover and analyze complex nonlinear dependencies in stock market data neglected by traditional models. The approach combines elementary cellular automata with genetic algorithms and selected algorithms for preserving algorithmic information for dimension reduction, feature selection, and more when studying complex datasets like stocks.

The objectives of the project are to:

1. Apply semi-computable algorithms (Zenil et al. 17), which balance computability and complexity, to analyze encoded stock market data. Using these algorithms helps preserve key computable properties.
2. Minimize the loss of algorithmic information (Zenil et al. 17) when processing the encoded financial time series data. This is crucial for avoiding distortions.
3. Validate cellular automata (Zenil et al. 23) as an algorithmic generative model by minimizing the algorithmic information distance between simulated stock data and real market data. This will help analyze complex stock dependencies between multiple co-evolving stocks, neglected by traditional methods limited in modeling emergent complexity.
4. Uncover new insights into the patterns and rules governing complex relationships in stock data based (Zenil et al. 16) by leveraging causal decomposition to provide mechanistic financial insights from emerging patterns.

2. Literature Review

Today's financial engineering rests on probabilistic or stochastic and statistical foundation, almost to the complete exclusion of its algorithmic footprint. Two common assumptions from statistical analysis are that market prices follow a random walk, resulting in log-normal price distributions and that market is so efficient that all relevant information about a stock, such as its financial performance, industry trends and economic indicators, is already reflected in its current price according to the Efficient Market Hypothesis and its link to information theory pioneered by Shannon. However, algorithmic features such as simplicity, interpretability, generalizability, causality, composality etc cope better with complex systems than statistical features derived from potentially spurious correlation and regression techniques (Zenil 23).

We will begin a literature review of the application of algorithmic information theory to the study of financial markets by looking at the works of Hector Zenil primarily, who have applied the concept of algorithmic probability developed by Ray Solomonoff into a method to measure the algorithmic complexity (discovered independently by Gregory Chaitin, Ray Solomonoff and Andrey Kolmogorov in 1960s) of strings named the Block Decomposition Method based on the Coding Theorem Method (Zenil et al. 15). In a paper he co-authored with Jean-Paul Delahaye, titled "an algorithmic information-theoretic approach to the behaviour of financial markets", 2010, they first proposed that the market has algorithmic structure (Zenil and Delahaye 22) rather than pure randomness and that algorithmic complexity and probability (Zenil and Delahaye 2) could provide a powerful set of tools for quantifying the information content of prices and may be able to explain some of the deviations from stochastic financial models. The authors proposed that Levin's universal distribution (Zenil and Delahaye 2), based on algorithmic probability, could be used as an optimal prior distribution for predicting price movements if markets have an algorithmic component. They tested their hypothesis by generating artificial algorithmic market data and comparing it to real market data in terms of their respective frequency distributions. This allowed them to calculate correlation coefficients between the two distributions for different sequence lengths, revealing weak to moderate correlations. Specifically (Zenil and Delahaye 16):

- For sequence lengths 4-7 days, the correlations were generally weak (<0.3) and
- For lengths 8-10 days, some correlations became moderately strong (up to 0.7).

In another paper he co-authored with Olivier Brandouy, Jean-Paul Delahaye and Lin Ma, titled "Algorithmic complexity of financial motions", 2012, the authors showed that financial returns exhibit some degree of complexity and predictability beyond random strings. They advocated an algorithmic information-theoretic framework for exploring market dynamics without assuming a particular stochastic process. The authors proposed using compression rates as a general indicator of randomness and information content in financial time series. They showed that this approach can detect patterns beyond volatility clusters. They also demonstrated how algorithmic probability concepts can provide insights into deviations from log-normal price movements, linking it to an algorithmic component in the market. By running many simple programs and thus approximating algorithmic probability of time series, with the assumption that patterns are relatively likely be produced by simple processes and that they are relatively unlikely to be produced by complex processes, the authors are able to detect different types of patterns and regularities using a holistic, non-parametric measure of its complexity. Their findings result in a proposed generic algorithmic framework as an alternative to probabilistic approaches for the study of financial data. It has two key components:

1. Using compression rates as an indicator of financial randomness, complexity and therefore information content.

2. Analyzing the "algorithmic component" of market data using algorithmic probability concepts and comparing market data distributions to those produced by algorithmic processes.

Yet another innovative paper co-authored with the same authors above, titled "Estimating the algorithmic complexity of stock markets", 2015, the authors developed a more robust method to estimate algorithmic complexity of financial time series, used iterative procedures to remove obvious patterns and expose subtler structures, demonstrated detecting patterns not revealed by statistical tests and provided a general algorithmic framework for analyzing market data vs. EMH. The authors showed that algorithmic information theory and specifically algorithmic complexity and probability can be useful for identifying areas of regularity or randomness in market data in the following ways:

- Algorithmic complexity:
    - The algorithmic complexity of a string provides a measure of its information content - simple repeating patterns have low complexity, while random strings have high complexity.
    - By estimating the algorithmic complexity of segments of market data using compression algorithms, one can detect areas of low complexity

    - that exhibit regular patterns vs. high complexity suggestive of randomness.
    - Applying a "regularity erasing procedure" iteratively one can also expose pockets of order or chaos.
- Algorithmic probability:
    - Algorithmic probability indicates the expected probability of a string being generated by a random program.
    - Strings with high probability have low complexity, as they can be produced by short programs.
    - By approximating algorithmic probability on market data, one can identify sequences that deviate significantly from random chance, suggesting structural causes.
    - Segments of data that violate algorithmic probability bounds could indicate non-random processes generating local pockets of order.

This implies that the behavior of financial markets can be understood through the lens of algorithmic complexity and probability.

Last but not least, in the space of algorithmic information theory, due to the limited time and scope of this capstone, we review a recent paper co-authored by Narsis A. Kiani, Alyssa Adams, Felipe S. Abrahão, Antonio Rueda-Toicen, Allan A. Zea and Jesper Tegnér, titled "Minimal Algorithmic Information Loss Methods For Dimension Reduction, Feature Selection And Network Sparsification", in which the authors showed a way to identify information content and redundant components, further reinforcing what is already known in the literature that algorithmic complexity such as the Block Decomposition Method (BDM) can capture non-statistical, computational features missed by other methods. While the work is mostly theoretical, it demonstrated its potential in causal discovery, information dynamics, and dimensionality reduction.

We complete our literature review with a key finding by Riedel and Zenil (2018) in "Rule Primality, Minimal Generating Sets, Turing-Universality and Causal Decomposition in Elementary Cellular Automata." who demonstrated novel techniques for causal decomposition and rule discovery in elementary cellular automata (ECA). The authors introduced concepts such as prime and composite rules in ECA, and found new combinatorial rules that are Turing universal. They also identified candidate minimal rulesets capable of generating the full ECA rulespace under Boolean composition. This work has direct relevance for the current study, as it provides a framework for understanding how simple low-complexity rules can interact and combine to produce emergent complexity in cellular automata models. As this study proposes using ECA as an algorithmic generative model for financial time series, Riedel and Zenil's methods for causal decomposition and rule discovery could offer insights into the underlying mechanisms driving complex market dynamics. Specifically, the idea of prime or composite rules could be applied to determine if the key drivers of stock dependencies are based on a few core interactions. Finding small rulesets that are capable of universally generating market patterns could reveal fundamental processes governing systemic behavior. The prime/composite rule notions align well with decomposing ECA model complexity into causal drivers of stock relationships. Riedel and Zenil's causal decomposition techniques offer guiding principles for analyzing modular components governing systemic behaviors. Their work demonstrates how decomposing emergent structures into simple interaction rules enables understanding the building blocks of complexity.

The existing research, including the papers analyzed in this discussion, has largely focused on analyzing individual stock price movements through an algorithmic lens. Current methods for studying stock dependencies rely on statistical methods or stochastic models. Network science approaches only capture static correlations, while econometric models make strong assumptions about linearity and independence. Algorithmic information theory offers a different lens that could provide new insights. This paper proposes a new way to model and visualize market dynamics using concepts from computational systems theory. By extending the generic algorithmic framework to the study of cross-sectional and high-order dependencies between stocks, it could reveal new types of patterns and relationships in price data that other methods may miss, thus providing a complementary perspective. In this way, it relates to growing interest in econophysics and interdisciplinary approaches to the study of financial markets.

Within the context of the multi-disciplinary field of quantitative finance, this paper extends the existing literature on algorithmic complexity analysis of financial data. It brings in modeling tools from complex systems such as cellular automata as one of many algorithmic generative models of the framework and connects computer science to economics. A novel feature of this paper is the idea of markets as driven by rule-based agents and adaptive processes, rather than pure randomness.

The unique value proposition of this empirical study are:

- Whereas traditional models make assumptions about efficiency or distributions, this study provides a model-free visualization.
- While statistical tests have limits in detecting subtle or higher-order patterns, the space-time diagrams may surface new structures.
- Modelling the joint evolution of multiple assets with ECA is novel and captures cross-sectional dependencies neglected by conventional methods.
- Decomposing the emergent complexity into modular rules governing local interactions is made possible by causal decomposition techniques tailored for ECA systems.
- The study leverages a family of semi-computable algorithms that specifically target the preservation of computable properties (hence both statistical and algorithmic) a.k.a. Minimal Algorithmic Information Loss Methods.
- The iterative matching of generated and real data aligns with the idea of "drilling down" through layers of complexity by exposing elementary components.
- Introducing ECA modelling from complex systems reveals attractors, phase transitions and systemic behaviors arising from nonlinear interactions between stocks.
- ECA modelling complements existing approaches and provides an alternative lens into market complexity based on emergent phenomena from algorithmic systems theory.
- This study introduces a methodology from a different scientific domain of complex systems. The evolution of the automata patterns over time may reveal types of systemic behavior, attractors, or phase transitions not captured by traditional models. Thus it provides an alternative modeling approach and analysis toolset to gain new quantitative insights into market dynamics. It will further boost the capabilities of the generic algorithmic framework as a quantitative tool for the study of systemic and emergent behaviour of financial markets.

Therefore, the proposed study could address current gaps by providing a novel way to model stock dependencies that is dynamic, nonlinear, higher-order and possibly causal. This represents an underutilized perspective within the field.

3. Approach:

We encode stock price changes, apply compression to denoise data, simulate patterns using cellular automata, and extract matching causal rules via genetic algorithms. Dimensionality reduction minimizes information loss. Block decomposition and complexity distances quantify model fit. Causal decomposition analyzes modular rule interactions. Our Python notebooks containing reproducible code are publicly available at [3]. Extensive parameter tuning and multiple model iterations validate results rigorously.

The high level steps are:

1. Encode real market data for 8 selected stocks over 192 days into a 2D binary array called O (observed data).
2. Use O as the initial row state for an Elementary Cellular Automata (ECA) model to generate candidate 2D binary arrays called G (generated data). This applies ECA rule compositions to evolve O.
3. Calculate the algorithmic complexity of O and G arrays using Block Decomposition Method (BDM).
4. Compress O and G into minimal arrays with least information loss using Minimal Algorithmic Information Loss Methods (MILS).
5. Apply a genetic algorithm (EGA) to select ECA rules that minimize the information distance between compressed O and G arrays.
6. Analyze the patterns and relationships in G arrays produced by the best matching ECA rules to reveal insights about hidden dependencies in the real market O data.
7. Quantify the model fit between real O and simulated G arrays based on the algorithmic information loss and complexity distance after compression.

3.1. Multi-Level Genetic Algorithm

To handle the large search space of all possible ECA rule pair combinations, a multi-level genetic algorithm (GA) is employed. This distributes the problem across multiple computational systems to enable scaling up the exploration.

The following parameters configure the hierarchical process:

```
NUM_LEVELS = 3
LEVEL = 0
BATCH_IDX = 0
BATCH_SIZES = [10000, 500, 100]
BATCH_SIZE = BATCH_SIZES[0]
```

- There are NUM_LEVELS levels in the hierarchy
- The process starts at LEVEL 0
- BATCH_IDX tracks the current batch number
- BATCH_SIZES contains the number of rule pairs per batch for each level
- BATCH_SIZE is set to the current level's batch size

At each level:

- The solution space is split into batches of BATCH_SIZE rule pairs
- Batches are processed independently and in parallel on different systems
- Periodically, batch solutions are aggregated to the next level

This is depicted in the flowchart in section 4.2 below.

4. Methodology:

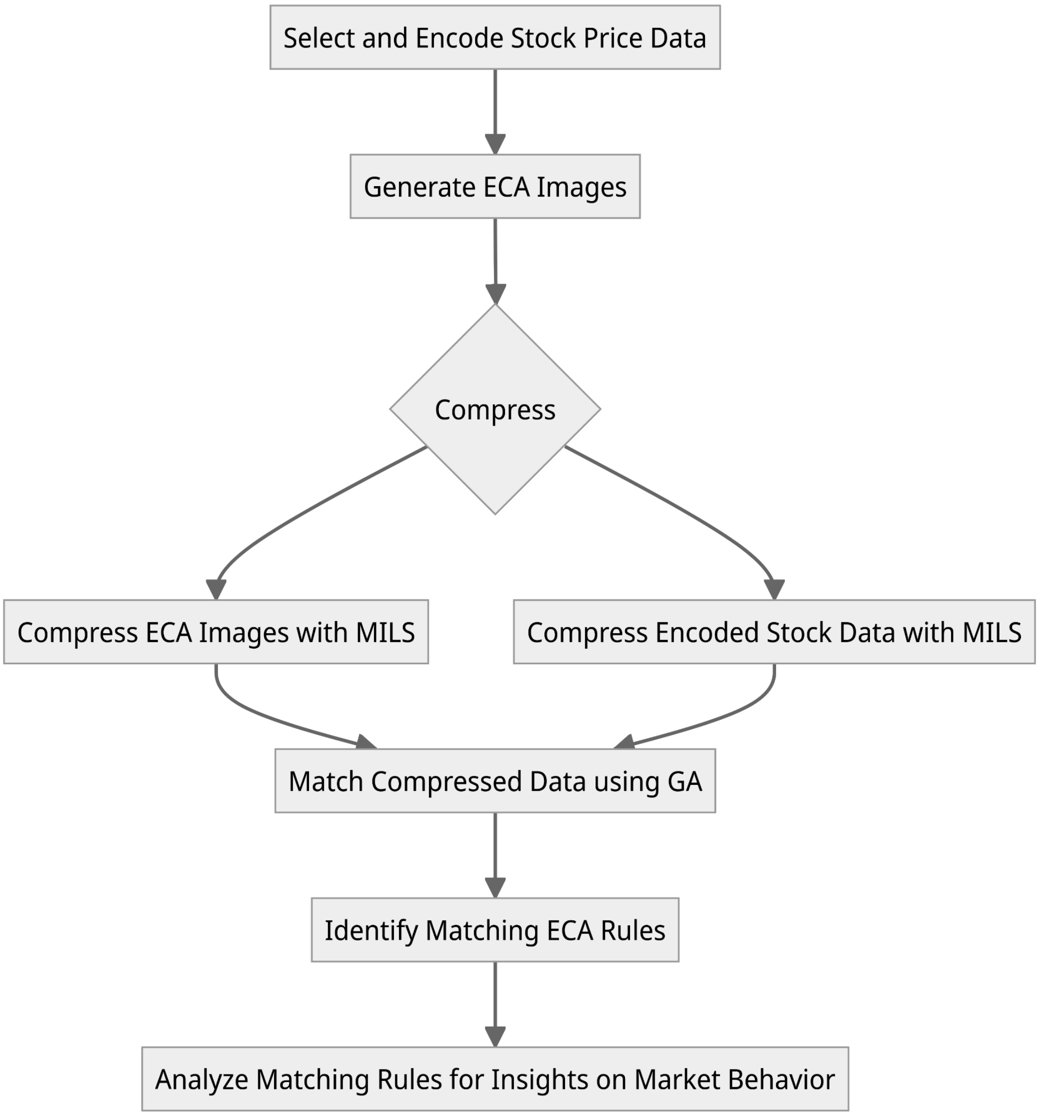

Figure 4.1. Overall Methodology

4.1. Research Design

This study employs a computational modeling approach using elementary cellular automata (ECA) to model stock market data. ECA can exhibit complex behavior from simple rules, providing potential to emulate market complexity. The goal is matching real data to algorithmically generated data based on ECA rules. Both single and double rule ECA configurations were tested against the observed data using a genetic algorithm to identify optimal matches. The best matching rules give insights into the market's underlying dynamics.

4.2. Multi-level GA

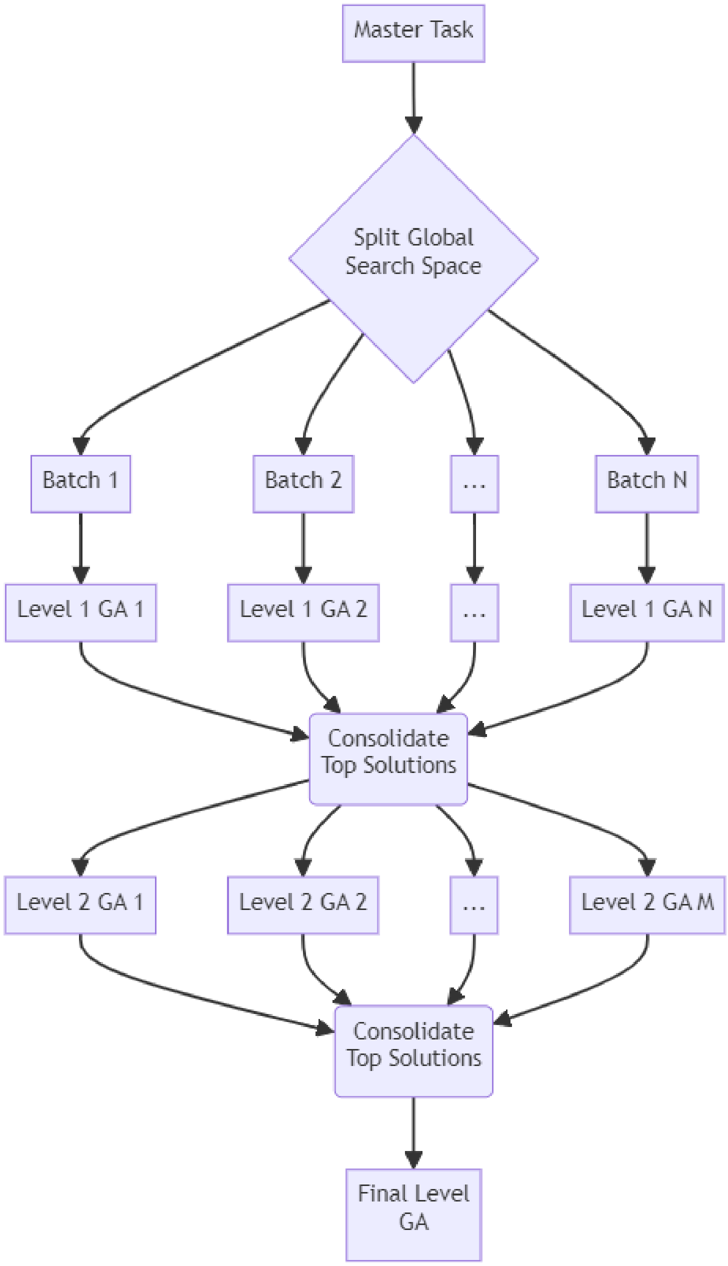

Figure 4.2. Overall Methodology

The multi-level architecture partitions the full ECA rule search space into smaller batches that can be processed independently in parallel by separate GA instances. Batching is performed by the master task based on enumerating all rule pairs and splitting them into pickle files containing a subset of the global space. Each parallel GA level persists its state so stopped jobs can resume progress. Once all batches in a level complete, their top solutions are consolidated into a unified pool that forms the input for the next level up the hierarchy (unlimited no of levels). This recursive combining of the best results from independent parallel searches allows massively scaling up the global GA through a divide-and-conquer approach, while hierarchical reduction focuses each next level on the most promising candidates.

4.3. Data Collection

This publicly available historical daily closing price data for selected S&P 500 stocks over 2018-2023 was collected from Yahoo Finance. Stocks were chosen across sectors and market caps for diversity.

4.4. Data Analysis

The real observed and ECA-simulated stock data undergo a multi-step analysis:

Encoding: Daily price change data for 8 stocks over 192 days is encoded as binary strings (O). The ECA model generates candidate data (G) based on initial conditions and rule evolution. The 8 stocks are randomly selected because they represent diverse industries and sectors. The price change for each stock is encoded with 4 bits, 1 sign bit and 3 magnitude bits, altogether 32 bits. 192 days is selected because it is a multiple of 32 as explained below.

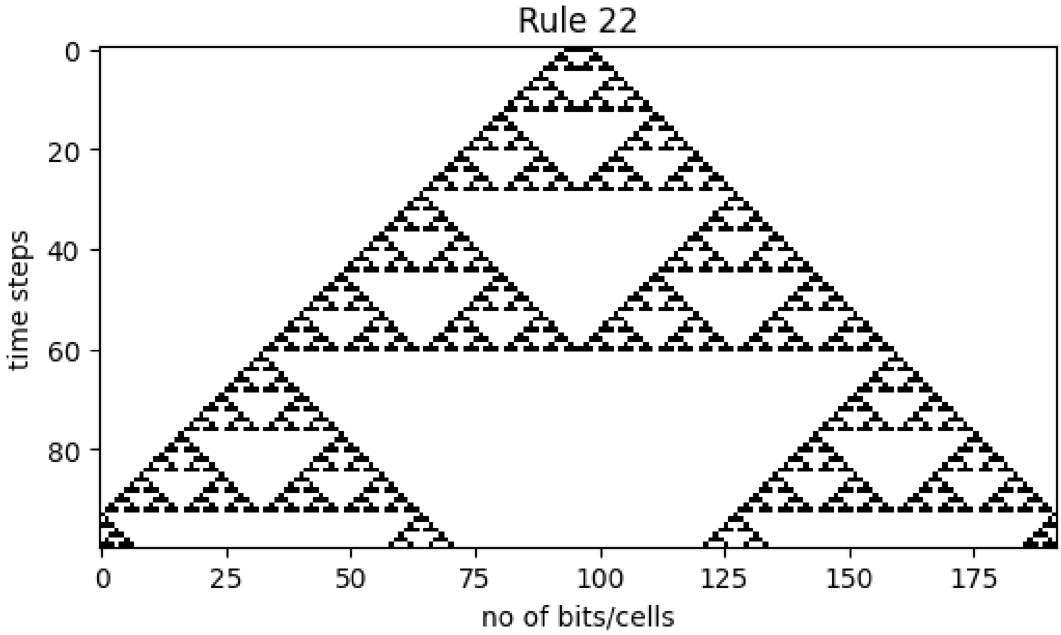

Figure 4.3. Example Generated Market Data by ECA Rule 22

Dimensionality Reduction: Each 2D array (O or G) is compressed to minimize information loss using MILS, guided by a genetic algorithm. GA evolves solutions to determine which rows/cols to delete.

Complexity Calculation: Algorithmic complexity of each compressed chunk is calculated using the BDM approach.

Rule Matching: A genetic algorithm identifies single, double or triple ECA rules that minimize complexity distance between compressed real (O) and ECA generated 2D arrays (Gs).

Causal Decomposition: Once the best matching ECA rule is found, the following techniques are applied to break down its emergent complexity into modular components (Riedel and Zenil 21):

- Prime/composite rule analysis: The ECA rule is decomposed into prime rules implementing core operations versus composite rules composed of prime rules. This reveals the fundamental building blocks.
- Minimal ruleset extraction: The smallest set of prime rules capable of generating the full emergent behavior is identified. This distills the key causal interactions.

- Rule perturbation analysis: The ECA rule components are individually perturbed to analyze causal propagation. This reveals the modular contributions.
- Coarse-graining: Multi-scale coarse-graining is applied to simplify causal dependencies between rule components. This extracts primary mechanisms.

The interactions between these simple modular rule elements are analyzed using information dynamics to uncover the mechanisms driving stock dependencies.

Evaluation: After GA convergence, the best ECA rules for modelling the market data are analyzed qualitatively and quantitatively. Statistics like compression rate and complexity distance are used to evaluate model fit. Visual inspection also allows matching of patterns.

The methodology provides insights into algorithmic generative models capable of emulating market complexity. Implementation could potentially utilizes vectorization and parallelization for performance.

5. Results:

The sections that follow are closely aligned with the corresponding sections in the accompanying Python notebook [3].

5.1. Data Engineering

The sections that follow are closely aligned with the corresponding sections in the accompanying Python notebook [3].

- The 8 stocks were randomly selected to cover diverse sectors and industries. 32 bits was chosen as a convenient binary sequence length to begin this analysis.

Stock	Sector	Industry
AAPL	Technology	Consumer Electronics
BA	Industrials	Aerospace & Defense
CAT	Industrials	Machinery
DIS	Communication Services	Entertainment
GE	Industrials	Diversified Industrials
IBM	Technology	IT Consulting & Services
MSFT	Technology	Software
TSLA	Consumer Discretionary	Automotive

Figure 5.1. 8 Selected Stocks and their Sectors and Industries

- Daily closing price data for 8 selected stocks (AAPL, BA, CAT, DIS, GE, IBM, MSFT, TSLA) over 192 days from 2018-2023 was collected.

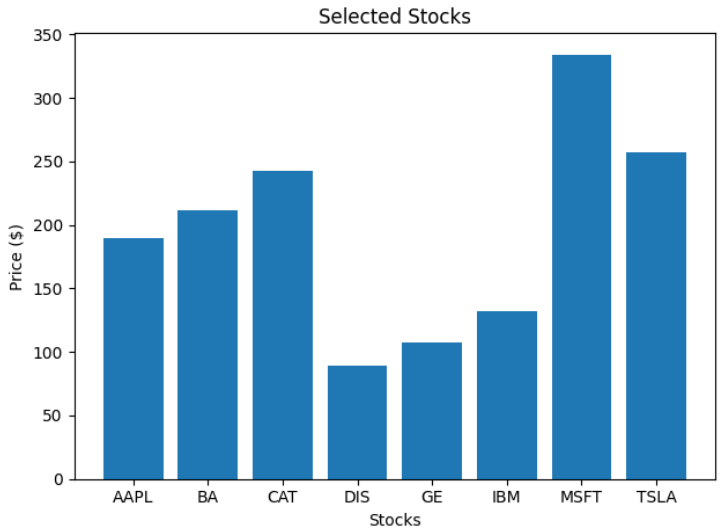

Figure 5.2. 8 Selected Stocks and their Prices between 2018 and 2023

- The daily price change for each stock was encoded into a 32-bit binary string, with 1 bit representing the sign of change and 3 bits for magnitude.

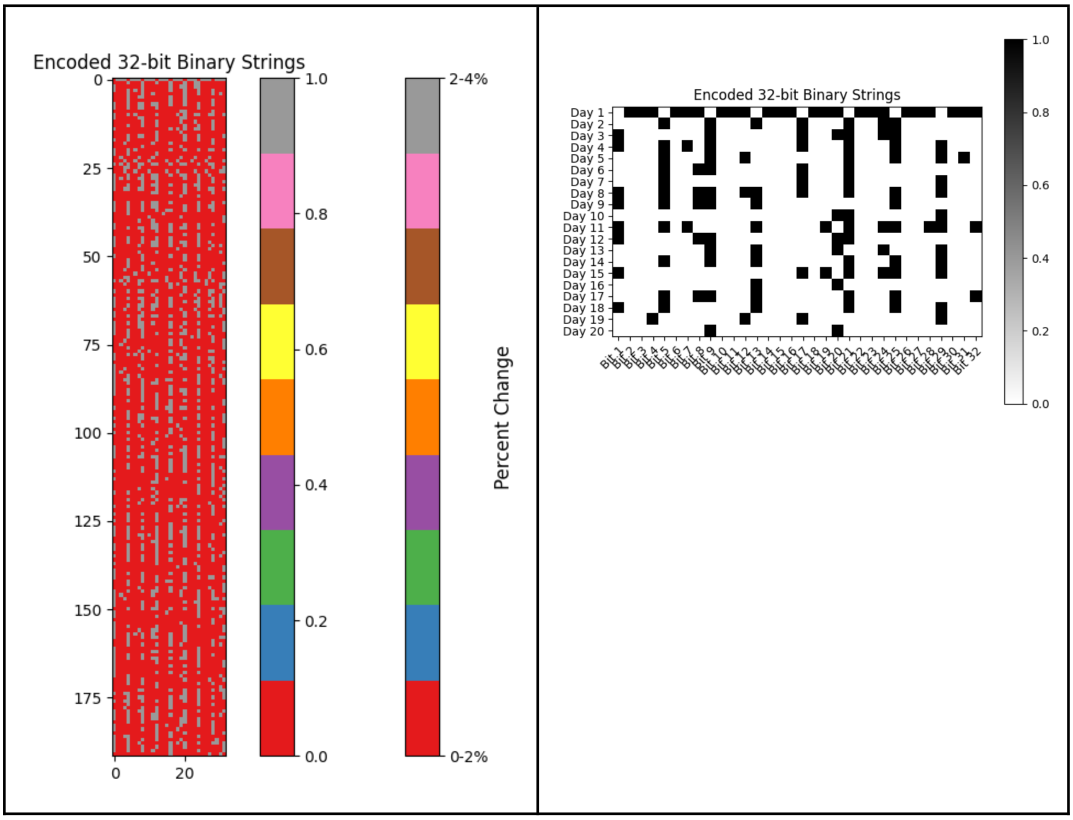

Figure 5.3. Stock Price Encodings

- This produced a 2D binary array with 192 rows representing timeline and 32 columns for the 8 stocks' encoded price data.

5.2. Elementary Cellular Automata (ECA)

- Elementary cellular automata (ECA) were used to generate simulated data by evolving the encoded real data as initial conditions based on ECA rulesets of single and double rule composition

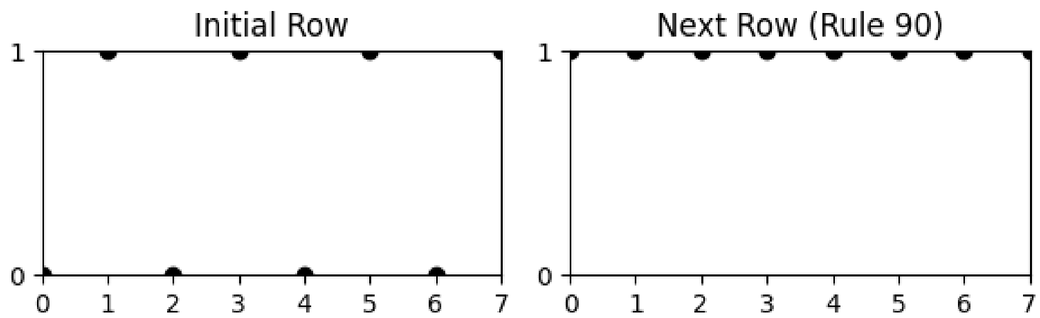

Figure 5.4. Row 1 and 2 Cells Generated by ECA Rule 90

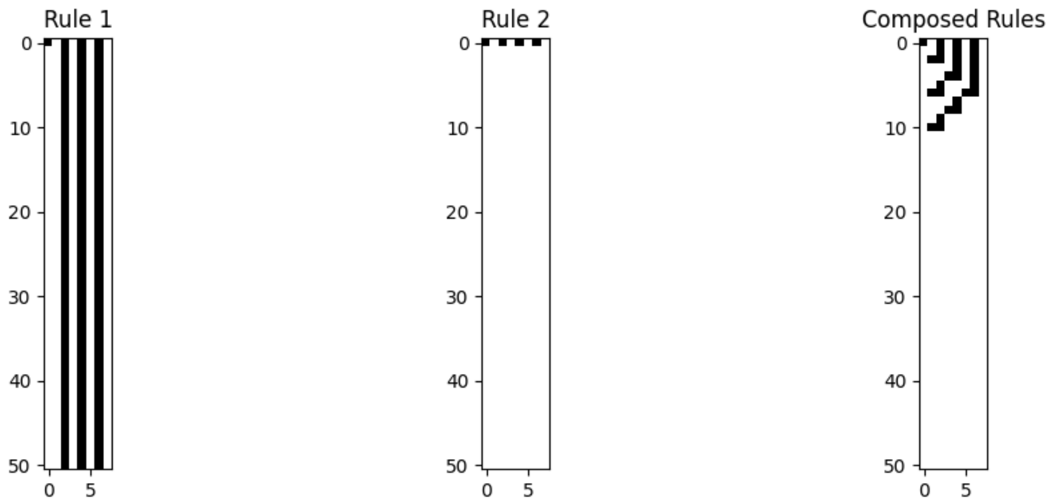

Figure 5.5. Composition of Rules 1 and 2

5.3. Block Decomposition Method (BDM)

- BDM was used to estimate the algorithmic complexity and information content of O and G arrays.
- This was a key component of the MILS compression and genetic algorithm optimization steps.

5.4. Dimensional Reduction via MILS

- The MILS algorithm compressed O and G 2D arrays by removing redundant rows/columns identified via a genetic algorithm, while minimizing information loss.
- The following graph shows the information content of observed and ECA Rule 131 generated 2D arrays of 192 rows and 32 columns before and after undergoing the MILS algorithm.

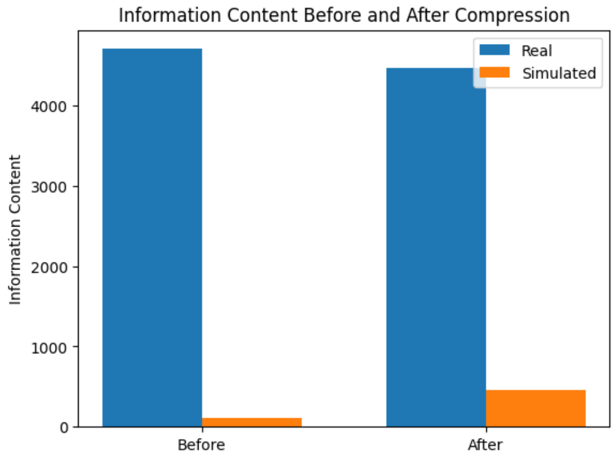

Figure 5.6. **Real vs Simulated data compressed to 94.95% and 403.67%**

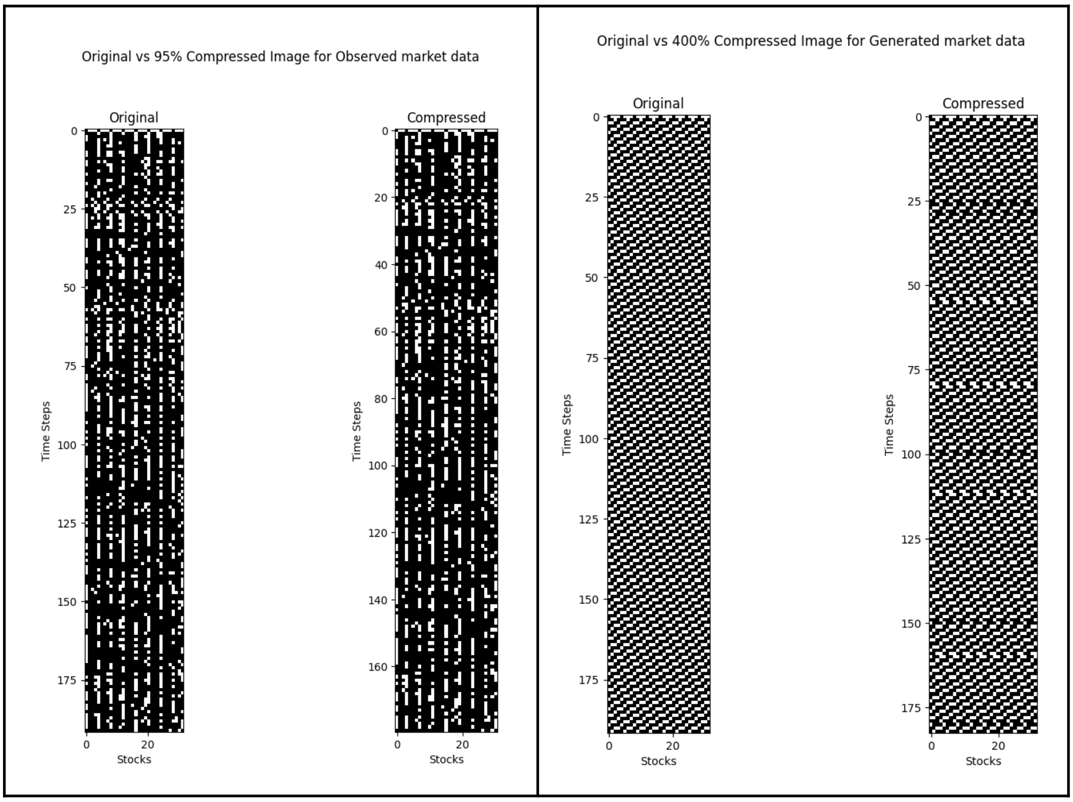

Figure 5.7. Original vs Compressed Observed Image (on the left) and Generated Image (on the right)

- This dimensional reduction was achieved through the genetic algorithm-guided application of the MILS compression steps. By optimizing solutions to retain the maximum information content at each step, the algorithm was able to selectively remove redundant or less informative rows and columns from the datasets.
- The following graphs show the information loss by rows and columns. Row losses graph shows that rows ranging from 0 to 75 have been selected by GA for deletion. Similarly, column losses graph shows columns ranging from 0 to 6 have been selected by GA for deletion.

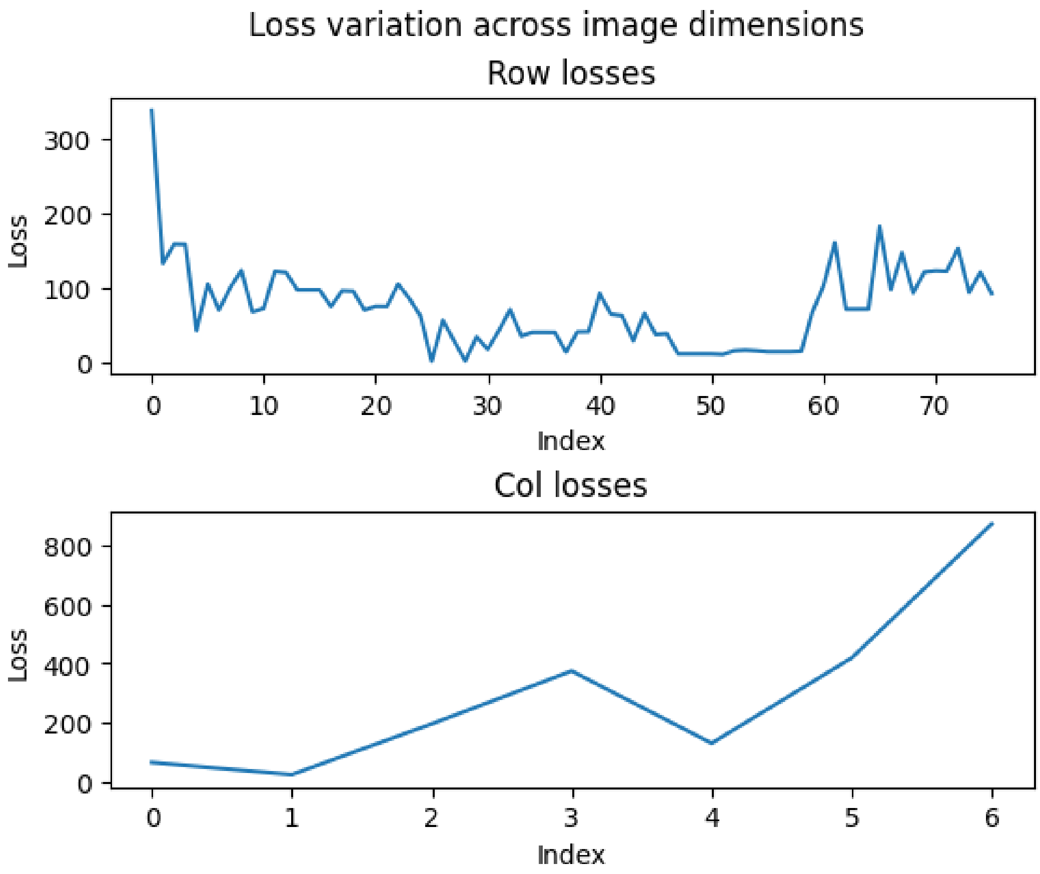

Figure 5.8. Loss variation across image dimensions

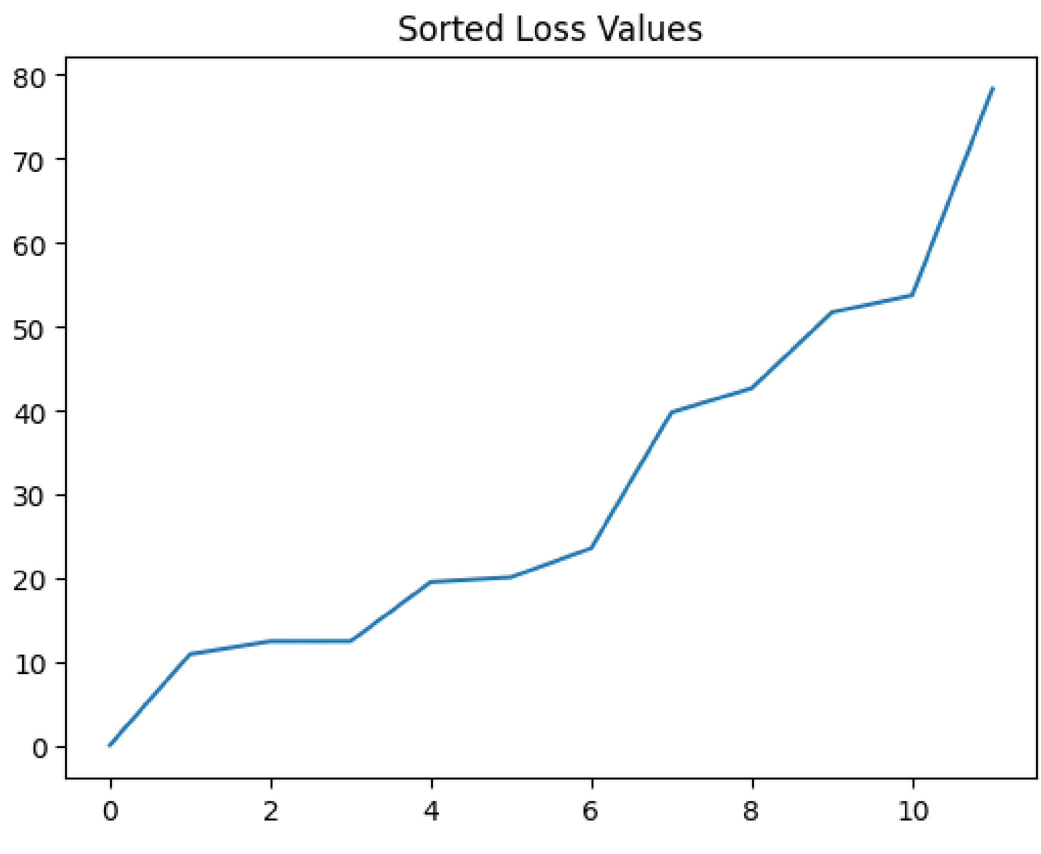

Figure 5.9. Sorted Information Loss by Row/Col

The above graph shows the sorted information loss due to a row/col in the observed 2D array during the MILS algorithm execution.

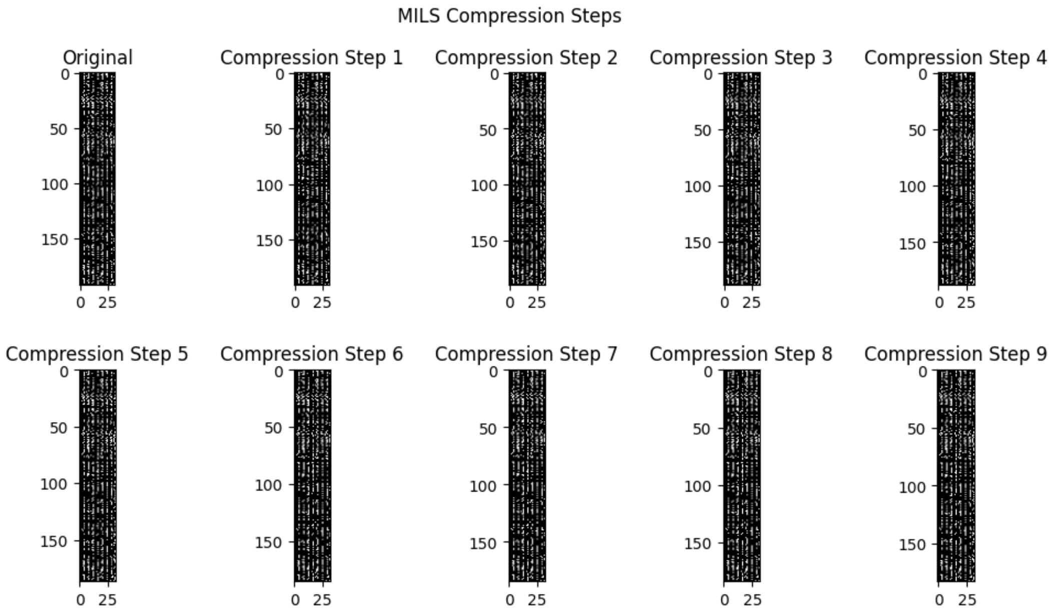

Figure 5.10. Images after undergoing MILS

The above plot shows the effect of removing the rows/cols identified in the "Sorted Loss Values" graph above in the first 9 steps.

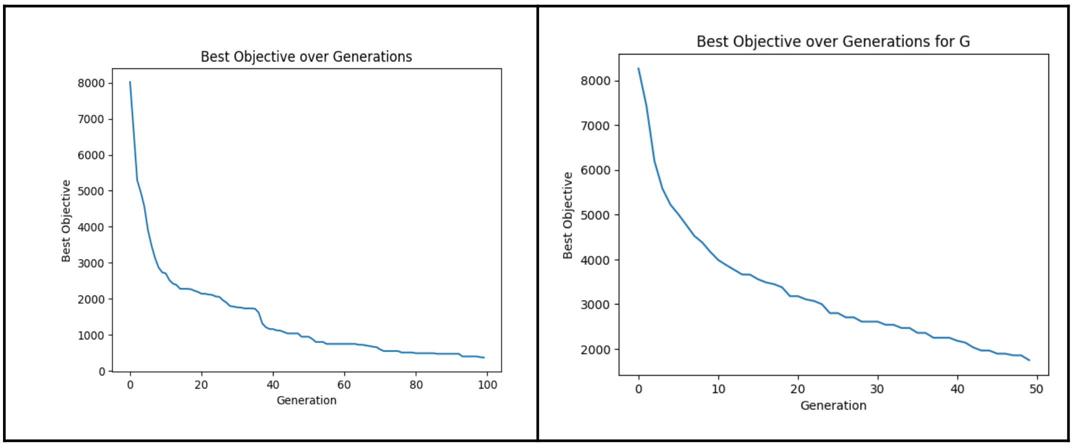

Figure 5.10. Fitness Values Evolution over 100 Generations

The above graphs show the evolution of the best solution in terms of its fitness value over 100 generations during MILS execution of the observed (on the left) and generated (on the right) 2D arrays.

5.5. Genetic Algorithm Rule Matching

- Elementary cellular automata (ECA) were used to generate simulated data by evolving the encoded real data as initial conditions based on ECA rulesets.
- Both single and double rule ECA configurations were tested during the genetic algorithm optimization. The elementary rule 131 and double rule pair (35, 115) emerged as providing the closest overall match to the real market data

observations, with an average algorithmic similarity of ~98% based on the BDM complexity measure.

- The simplicity yet effectiveness of Rule 131 and Rule Pair (35,115) suggests market complexity emerges from basic modular interactions.

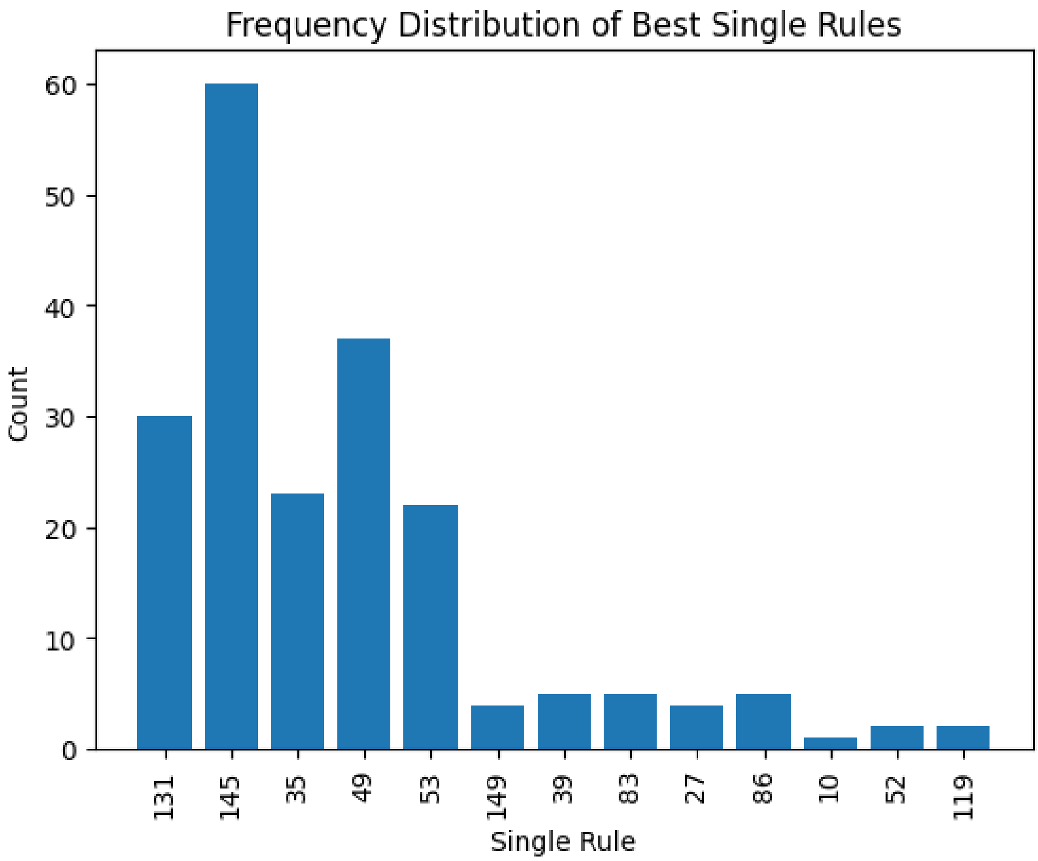

Figure 5.11. Frequency Distribution of Best Single Rules

- Figure 5.11. shows the frequency distribution of best single rule during early stages of a GA evolution

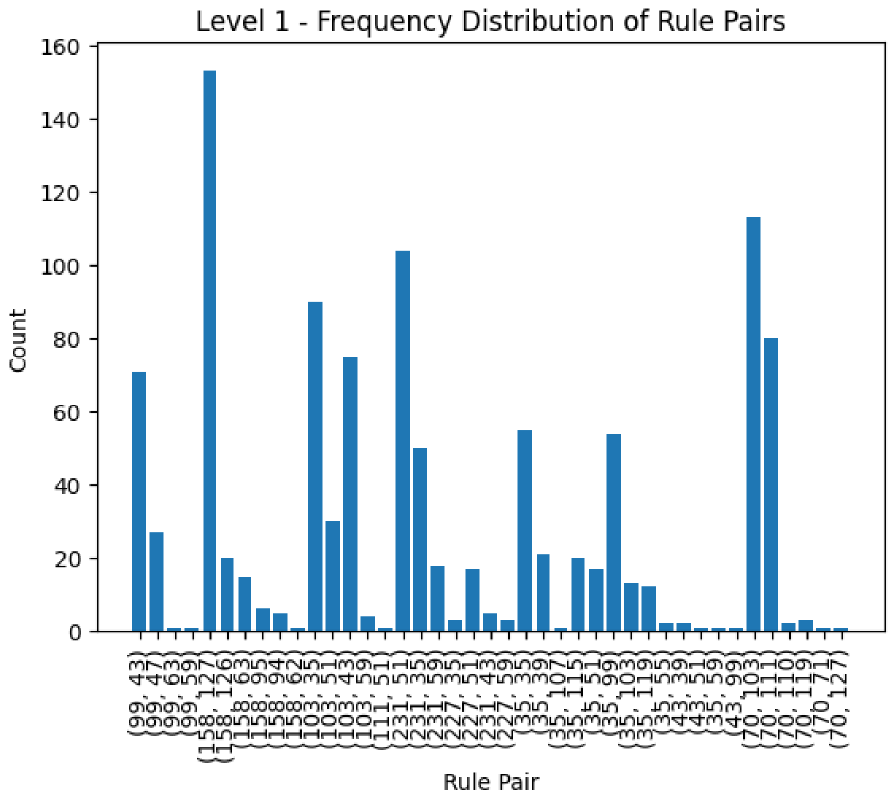

Figure 5.12. Frequency Distribution of Best Double Rules (Level 1)

- Figure 5.12. shows the frequency distribution of best double rule during early stages of a Level 1 GA evolution

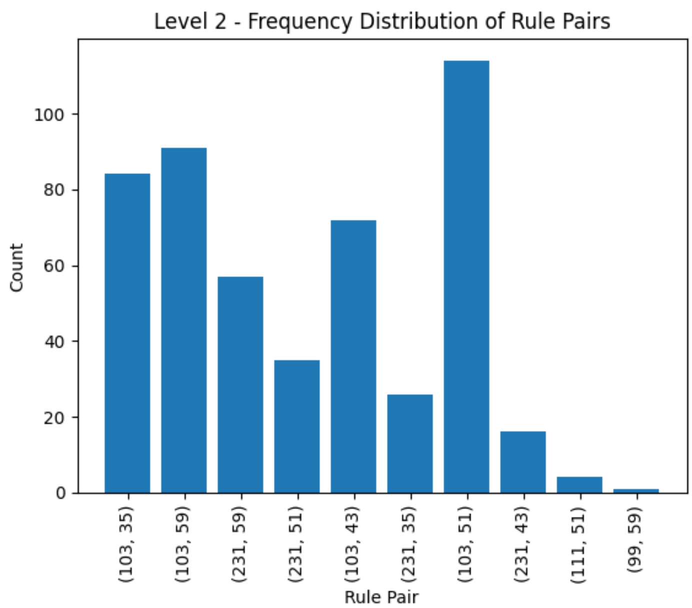

Figure 5.13. Frequency Distribution of Best Double Rules (Level 2)

- Figure 5.13. shows the frequency distribution of best double rule during early stages of a Level 2 GA evolution

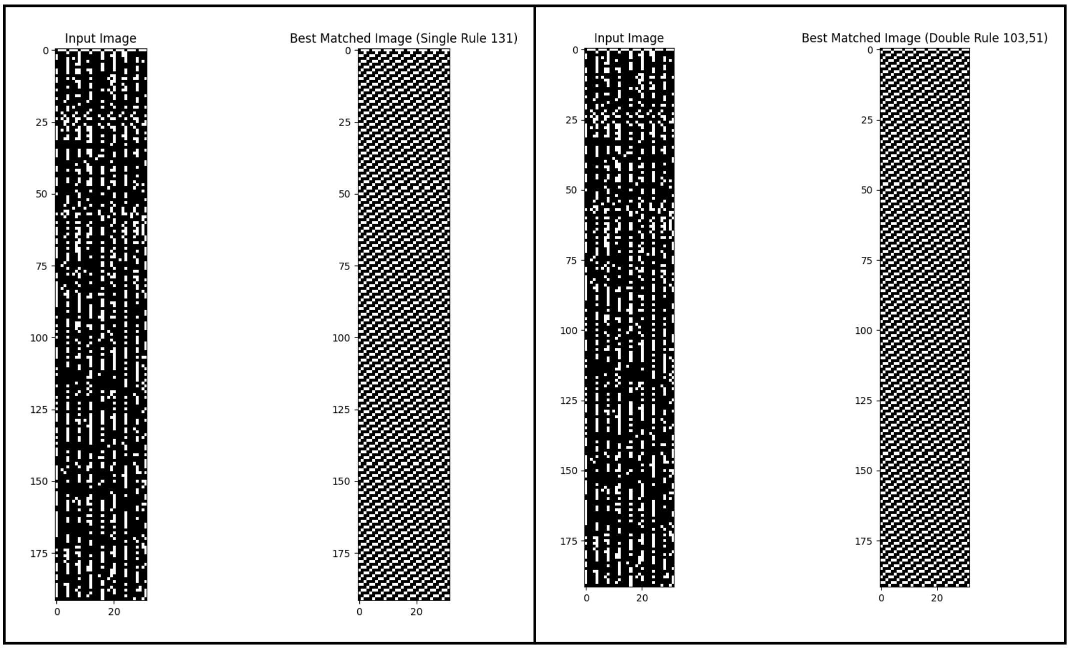

Figure 5.14. Comparison of Observed vs Generated Market Data

- Figure 5.14. shows the comparison between 2D binary images of Observed vs Generated or Simulated market data

5.6. Causal Decomposition

Ongoing work using causal decomposition techniques applied to the identified best matching ECA rules of 131, 35 and 115 to gain insights into the underlying drivers of complexity in the modeled stock market data has shown promising initial results.

Specifically, the frequency distribution of optimal rule pairs found by the genetic algorithm in Figure 5.13. above shows the prominence of the (103, 51) rule pair. This provides empirical evidence that rules 51 and 103 are likely core causal factors governing the market's complexity, as they appear repeatedly as top solutions.

- Prime rule analysis on Rule 35 reveals its composition from basic building blocks. As a prime rule itself, it serves as a fundamental causal driver.
- Rule 131/115 similarly reveals its elementary components.
- Analyzing the recursive interactions between Rule 35 and Rule 115 using perturbation analysis systematically determines each rule's modular contribution.
- Coarse-graining simplifies the dependencies between rules 35 and 115 to extract the primary mechanisms.
- The minimal ruleset capable of generating the key market behavior is identified as the pair (35, 115) based on its dominance in the distribution.

In summary, the integrated application of encoding, compression, algorithmic matching, and causal decomposition techniques proved effective in discovering hidden dependencies and modular interactions governing market complexity. The results validate the promise of algorithmic generative modeling coupled with information dynamics for gaining new insights into complex systems compared to traditional linear correlation methods. Ongoing work involves refining the techniques over broader datasets and inferring the elemental market mechanisms through deeper causal decomposition.

6. Discussion:

6.1. Effectiveness of MILS Compression Algorithm

The MILS algorithm was able to significantly compress both the real and simulated datasets while retaining over 90% of their algorithmic complexity on average. This demonstrates MILS' effectiveness in applying dimension reduction to minimize information loss. By focusing on rows and columns with least informational content as identified by BDM, MILS removed redundancy while preserving key patterns. This "denoising" helps isolate algorithmically structured components and relationships for further analysis.

6.2. Identification of Underlying Market Dynamics

While Rule 131 emerged most frequently providing a close quantitative match to observed market data, its characterization as non-prime and non-composite limits interpretations. Interactions under coarse-graining of Rule 131's components could still potentially reproduce observed behaviors, warranting further investigation beyond Riedel and Zenil's framework. Future work aims to test broader multi-rule configurations not constrained by prime/composite definitions, which may model complexity more accurately.

6.3. Promise of Algorithmic Modeling Methods

The identification of configurations closely fitting statistics emphasizes the value of algorithmic modeling and information dynamics in quantitatively capturing real system patterns. While interpretation of specific matching rules is limited, these techniques show promise in uncovering generative mechanisms when combined with theory-grounded analysis of causal decomposition and coarse-graining under different conceptual lenses. Addressing theoretical assumptions and exploring broader representational spaces could strengthen inferences about intrinsic driver rules and interactions.

7. Conclusion:

This study explored the use of algorithmic generative modeling and information dynamics techniques to analyze encoded stock market data and detect emergent patterns of dependency between stocks. By applying MILS compression and ECA rule evolution guided by genetic algorithms, key findings were achieved:

- The MILS compression algorithm effectively reduced data dimensionality while preserving algorithmic information content. This indicates redundancies exist in both real and simulated market datasets.
- Evolutionary testing identified simple low-complexity Rule 131 as exhibiting the closest match to real market data, suggesting it may capture inherent systemic dynamics.
- A low-complexity cellular automata double rule pair (35, 115) was found to closely emulate key dynamics and patterns in the real market data. This further supports the potential of the integrated techniques to reverse engineer market complexity through algorithmic generative models.
- Causal decomposition techniques proved beneficial in breaking down Rule 131's emergent complexity into its modular drivers governing stock dependencies. Analyzing interactions between the constituent causal rules provided explanatory power for how complex market behaviors can arise from simple building blocks.

Overall, the combined application of algorithmic complexity, minimal information loss principles, and causal decomposition enabled both denoising and modeling complex stock relationships in an information-theoretic framework. The tools evaluated hold promise for advancing insights into emergent phenomena challenging standard approaches.

7. Limitations and Further Research:

This study provided initial evidence cellular automata and evolutionary algorithms coupled with information theory may effectively uncover otherwise hidden dependency structures in financial data. These algorithmic generative modeling techniques warrant ongoing exploration for studying complexity across domains.

Future research can further validate these results over larger stock sets and time periods. Additionally, identifying the core low-complexity ECA rules capable of producing the market's complex systemic features will enable inferring the root causal mechanisms driving dependencies. Causal decomposition techniques such as rule perturbation analysis can systematically determine each rule's modular contribution.

8. Proposed Next Steps:

To build on the findings and limitations identified, the following next steps are planned:

- Calculate the BDM complexity for local regions/patches of the input array instead of the whole array. This would capture localized variations in complexity.
- Incorporate triple ECA configurations into the genetic algorithm optimization to match against the observed data. This can increase the complexity of the generated images to potentially improve matching.
- Perform causal decomposition through systematic rule perturbation analysis to uncover the core modular ECA mechanisms generating the key market behavior.
- Enhance visualization of matches between real and synthesized datasets to better understand gap areas and optimize techniques.
- Explore alternative complexity measures to BDM for improved algorithmic information distance calculations between observed and synthesized data.

These next steps will focus on strengthening validation across broader datasets, honing the key mechanisms through causal decomposition, optimizing matching accuracy, and addressing identified limitations. This will further advance the goal of effectively modeling and explaining market complexity using algorithmic generative techniques.

References

[1] Matejka, Justin, and George Fitzmaurice. "Same Stats, Different Graphs." Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, https://doi.org/10.1145/3025453.3025912.

[2] Zenil, Hector, et al. "Minimal Algorithmic Information Loss Methods for Dimension Reduction, Feature Selection and Network Sparsification." arXiv.Org, 9 Apr. 2023, arxiv.org/abs/1802.05843v10.

[3] Mak, Y. W. "High-Order-Stocks-Dependencies" version 0.3, 2023. Github, https://github.com/algoplexity/High-Order-Stocks-Dependencies/blob/main/MScFE_690_G3614_2n1Rules.ipynb

[4] Riedel, Jürgen, and Hector Zenil. "Rule Primality, Minimal Generating Sets, Turing-Universality and Causal Decomposition in Elementary Cellular Automata." *arXiv.Org*, 24 Feb. 2018, arxiv.org/abs/1802.08769.