(ur) DIMENSIONS

The Journal of Undergradate Research
in Natural Sciences and Mathematics

Volume 26 | Spring 2025

Marks of a CSUF graduate from the College of Natural Sciences and Mathematics

GRADUATES FROM THE COLLEGE OF NATURAL SCIENCES AND MATHEMATICS:

Understand the basic concepts and principles of science and mathematics.

Are experienced in working collectively and collaborating to solve problems.

Communicate both orally and in writing with clarity, precision, and confidence.

Are adept at using computers to do word processing, prepare spreadsheets and graphs, and use presentation software.

Possess skills in information retrieval using library resources and the internet.

Have extensive laboratory, workshop, and field experience where they utilize the scientific method to ask questions, formulate hypotheses, design and conduct experiments, and analyze data.

Appreciate diverse cultures as a result of working side by side with many people in collaborative efforts in the classroom, laboratory, and on research projects.

Have had the opportunity to work individually with faculty members in conducting research and independent projects, often leading to the generation of original data and contributing to the research knowledge base.

Are capable of working with modern equipment, instrumentation, and techniques.

DIMENSIONS

DIMENSIONS: The Journal of Undergraduate Research in Natural Sciences and Mathematics is an official publication of California State University, Fullerton. DIMENSIONS is published annually by CSUF, 800 N. State College Blvd., Fullerton, CA 92834.

ABOUT THE COVER

We would like to give a special thanks to Fernando Del Rosario and his graphic design class. The cover for Volume 26 was designed by Manami Kobayashi. The artist's statement is as follows: "There are departments' icons in a light bulb. Color uses university color for title and the ones (circle and stick) around the light bulb. The light bulb represents inspiration and analysis. The ones around bulb and the right corner also represents light of the bulb".

DIMENSIONS Editorial Staff and NSM Leadership

EDITOR-IN-CHIEF

Jeremy Hansen – Assistant Director for Academic Support and Engagement

EDITORS

Aaron Kim – Mathematics

Jonathan Sarti – Chemistry and Biochemistry

Katie Yang – Biological Science

GRAPHIC DESIGN

Manami Kobayashi – Dimensions 2024-2025 Cover Art

COLLEGE OF NATURAL SCIENCES & MATHEMATICS

Dr. Marie Johnson - Dean

Dr. Nicholas Salzameda - Associate Dean

Dr. Marcelo Tolmasky - Chair, Department of Biological Science

Dr. Niroshika Keppetipola – Chair, Department of Chemistry and Biochemistry

Dr. Adam Woods – Chair, Department of Geological Sciences

Dr. Adam Glesser – Chair, Department of Mathematics

Dr. Ionel Tifrea – Chair, Department of Physics

Letter from the Editor-in-Chief:

Dimensions is a testament to the incredible and insightful research that our undergraduate students complete during their time here at CSU Fullerton. The papers collected in this journal reflect the motivation, inspiration, and knowledge of Titans on their journey to study this world and improve their communities. For most of these students, this is a steppingstone on their educational path as they eventually move on to graduate school or careers outside academia.

Dimensions would like to thank the College of Natural Sciences and Mathematics faculty, for their guidance of these students' research and for encouraging them to submit their papers. Your research inspires your students to get involved in research labs and become better scholars. We also would like to thank the NSM Deans' Office for supporting this publication and their assistance in getting the word out to students and faculty about this opportunity. And as Editor-in-Chief, I would like to thank our student editors and submitters for continuing to support the mission of *Dimensions* through their involvement.

This journal would not exist without curiosity and the willingness to share research with the campus community. If you are reading this, no matter where you are in your educational journey, we hope that it inspires you to complete your own research project. Share this publication with your peers, your family, and your community. Get involved with *Dimensions*, apply to be a student editor or encourage others to submit their research papers. Most of all, be curious, and chase answers to questions you have. When you find those answers, *Dimensions* will be here to help you bring visibility to your work and push the field you study further forward.

Table of Contents

83 - Methods for Finding the Second Moment of Insurance Payment

- Author: Aaron Kim
- Advisor: Dr. Justin Nguyen

104 – Beyond Correlation: An Analysis of Risk in the S&P 500 Index

- Author: Alejandro Reyes
- Advisor: Dr. Matheus B. Gurrero

Biological Sciences

The Impact of Microbes on the Evolution of *Drosophila melanogaster* Development Time

Burgundy Davison, Sonia Garcia, Shelly Li, Urja Nandu

Advisor: Dr. Parvin Shahrestani

Abstract

Microbiome communities that inhabit animals or exist on their surfaces have historically been overlooked in laboratory evolution studies, yet they are essential for understanding the evolution of life history traits. Microbiota influence host traits, for example, affecting the developmental time of *Drosophila melanogaster*. Direct selection on development can result in divergence of this trait among laboratory populations and divergence of the host's microbiota. However, it remains unclear whether changes to development time evolve before changes in the host's microbiota or if alterations in the host's microbiota occur before the evolution of the host's life history traits. This project focuses on altering the host's developmental time through selection pressure while manipulating their microbiota through supplementation with acetic acid and lactic acid bacteria**.** This will be done by measuring bacterial abundance and development time before and after imposing selection pressure for fast development in different microbial conditions. Generation 0 data included in this study show that the development time of A-type flies is quicker than that of C-type flies. It also demonstrates that Acetic Acid Bacteria are more abundant in A-type flies whereas Lactic Acid Bacteria are more abundant in C-type flies. Anticipated results suggest that acetic acid-producing bacteria will alleviate selection pressure on the host genotype, leading to a slowdown in the evolution of fast development. Conversely, we expect that lactic acid-producing bacteria will buffer selection pressure on the host genotype,

accelerating the evolution of fast development. These findings will help elucidate how the host's evolution impacts and relies on the microbiota.

Introduction

Adaptive evolution occurs through natural selection, with the key requirements of heritable phenotypic differences within the population resulting in fitness differences. Within adaptive evolution, variation may result in advantageous and adverse alleles that lead to selection of phenotypes and accordingly, their subsequent genotypes. Adaptive evolution can occur in the wild or in a laboratory, where it is known as experimental evolution (Garland, T. & Rose, M. R., 2009). Within a laboratory setting, experimental evolution is a common practice used for phenotype-to-genotype mapping. Organisms used in experimental evolution are held under controlled conditions and environments, where variables such as temperature, humidity, and food can impact results (Garland, T., & Rose, M. R. 2009). The study of the microbiota within animal experimental evolution is often ignored and under-explored, leading to our inquiry on adaptive evolution within *Drosophila melanogaster* in relation to its microbiota.

D. melanogaster, alongside humans, is one of the most extensively studied organisms in biology due to sharing around 75% of the disease-causing genes within humans (Mirzoyan et al. 2019). *D. melanogaster* is considered a model organism for experimental, genomic, and evolutionary studies as they have a rapid life cycle, high reproductive rates, and are extremely versatile (Mirzoyan et al. 2019). The specific *D. melanogaster* populations used in this study come from the Drosophila Experimental Population Resource, also known as the Rose populations (e.g. Rose et al. 2004). Specifically, we used the C-type and the A-type *D. melanogaster.* C-type *D. melanogaster* have a 28-day generational cycle (one generation cycle is from one egg to the next egg), are slow developing, and long-lived. A-type *D. melanogaster* has

a 10-day generation cycle, are fast developing, and short-lived (Graves et al. 2017). Both C-type and A-type *D. melanogaster* contain five replicate populations, with the C-type being referred to as CO_{1-5} and the A-type is referred to as ACO_{1-5} (Graves et al. 2017)

The microbiota of an organism (microorganisms that live within and on the fly) are known to influence the host's traits (Matthews et al. 2021, & Gould et al. 2018). The components of the microbiota are also heavily influenced on *D. melanogaster* development time. Due to the strong influence the microbiota holds on the *D. melanogaster* development time, the experimental manipulation of this factor is the focus of this study. The two predominant bacterial types found within the *D. melanogaster* microbiota are lactic acid bacteria (LAB), and acetic acid bacteria (AAB) (Chaston et al. 2016). The A-type populations are abundant in both LAB and AAB, whereas C-type populations are less abundant in both bacteria (Walters et al. 2019). Our study seeks to determine if microbiota or genotype has stronger effects on the evolution of the organism's life history traits.

With continuous research, the question of the microbiota function and its overall ability to relieve or buffer selection pressure on the host genotype arises. We work to address this question through the laboratory selection of both the faster and slower developing flies in conventional and supplemented conditions with AAB and LAB. We hypothesize that the development time of the A-type and C-type populations of *Drosophila melanogaster* will have varying response times to the different selections based on their bacterial abundance. This study can help lead to a better understanding of the interactions between bacteria and development time of *D. melanogaster.*

Methods

Experimental Design

To investigate our hypothesis, we use experimental evolution to study the evolutionary relationship between development time and the microbiota in *D. melanogaster* populations. This experiment is done under three conditions: conventional, supplemented with AAB, and supplemented with LAB. For the first and last generations, we measure the host's ability to control its microbiota and its development time. The last generation is determined once the flies have reached A-type conditions. These are measured compared to the control and experimental populations. By measuring their bacterial abundance and development time, we gain further insight into the role that microbes play in the response to selection.

Selection

Drosophila melanogaster populations are kept on 24-hour light cycles on a banana molasses diet. The A-type flies (ACO) are on 10-day generation cycles, and the C-type flies (CO) are on 28-day generation cycles. nACO populations (new ACO populations) develop from CO ancestors by transferring the first 20% of emerged flies in each generation from rearing vials to cages. Then, they are allowed to lay eggs for 18 hours for the next generation. This continues until A-type flies are achieved. This is done under three different conditions. The first experimental population is kept under conventional conditions (nACO) with no bacterial manipulation. The second experimental population (nACO-AAB) is supplemented with acetic acid bacteria (*Acetobacter* species) at every generation. The third experimental population (nACO-LAB) is supplemented with lactic acid bacteria (*Lactobacillus* species) at every generation. Each of the ancestral and experimental populations has five replicates (ACO_{1-5}, CO_{1-5}, $nACO_{1-5}$, $nACO\text{-}AAB_{1-5}$, $nACO\text{-}LAB_{1-5}$).

Phenotyping

Bacterial Preparation

Bacterial stocks of *Acetobacter pasteurianus (Ap), Acetobacter tropicalis (At), Lactobacillus brevis (Lb),* and *Lactobacillus plantarum (Lp)* are stored separately in glycerol solution diluted at -80℃. Approximately two days before inoculation, bacterial stocks are streaked onto Criterion Lactobacilli MRS agar plates. Per Liter of Pure Life Distilled Water, 20g dextrose, 10g meat peptone, 10g beef peptone, 5g yeast extract, 5g sodium acetate, 2g disodium phosphate, 2 grams ammonium citrate, 1g tween 80, 0.1g magnesium sulfate, and 0.05g manganese sulfate is added to make agar plates. *Acetobacter* species grow under aerobic conditions, while *Lactobacillus* species are placed in tightly sealed containers with added CO_2 to create anaerobic conditions. A day before inoculation, a single colony from each species is isolated from the agar plates and mono-inoculated into sterile Criterion Lactobacilli MRS broth tubes before being vortexed. The *Acetobacter* species are incubated while shaking at 30°C, and the *Lactobacillus* species are incubated stagnant at 30°C. A day later (24 hours), 1mL of each bacterial suspension is vortexed and pipetted into cuvettes to receive absorbance readings on the spectrophotometer. Using 1.7 mL microcentrifuge tubes, the bacterial suspensions are diluted to receive a normalization absorbance reading of OD600=1, vortexed, and later centrifuged. The supernatant is removed, and the pellet is resuspended in PBS. After vortexing the pellet in PBS, the bacteria is mixed to create different treatments. *500 µL of Acetobacter pasteurianus* and *500 µL* of *Acetobacter tropicalis* is added to a microcentrifuge tube to create the AAB treatment. *500 µL of Lactobacillus brevis* and *500 µL* of *Lactobacillus plantarum* are added to a separate microcentrifuge tube to create the LAB treatment. The LAB+AAB treatment consists of 250 µL of each of the 4 bacteria species mixed in one microcentrifuge tube. All treatments are then placed in the fridge until inoculation.

Dechorionation

Half-filled banana food plates are placed inside the cages with a yeast paste in a micro-petri dish 6 hours before dechorionation. After 6 hours, the eggs laid by adult flies are washed off the food plate with autoclaved DI water into a sterile bushing. The bushing is washed in 0.6% bleach solution for 2.5 minutes outside of the sterile hood. Then, the bushing is placed in a new cup filled with 0.6% bleach, and the washing process is repeated for another 2.5 minutes. The bushing is then placed inside the sterile hood, where a third bleach is done with a new solution for another 2.5 minutes. Then, autoclaved DI water is used to wash the bushing three times by dunking it three times at each wash. Finally, once the last wash is done, the eggs that are on the walls of the bushing are transferred to sterile food with a paintbrush. Solutions are discarded after every wash.

Bacterial Inoculation

After dechorionation in the first and last generation, eggs from the CO, ACO, nACO, nACO-AAB, and nACO-LAB are inoculated with 50μL of the 4 species bacterial mixture described in the "selection" section above. We will inoculate eggs from the CO, ACO, nACO, nACO-AAB and nACO-LAB population with 100μL of *Acetobacter pasteurianus* and *Lactobacillus brevis* directly into each tube. The caps of the tubes are left a quarter turned, kept at room temperature with a 24-hour light in an incubator.

Homogenization & Plating

We prepare sterilized microcentrifuge tubes by filling them with 125 µL of sterile mMRS broth and 0.1 mL of ceramic mixing beads. Adult flies are removed from their respective food vials in the incubator and anesthetize them 5 days post-eclosion with CO_2. Five females and five males are chosen randomly from vials of each population and placed separately into the microcentrifuge tubes. The flies go through two rounds of homogenization at 6.5 m/s, with each

round being 60 seconds. We then plate the homogenate on mMRS agar plates and incubate it at 30°C for 48 hours. After 48 hours, we remove the plates from the incubator and count the grown colonies under a microscope. We estimate the number of Colony Forming Units (CFU) using the equation:

$$CFU = colonies\ counted \left(\frac{dilutions}{\mu L\ plated}\right) x \left(\frac{voume\ of\ fly\ homogenete\ total}{\text{number of flies homogenized}}\right).$$

Development Time

In the first and last generation, from each population, we will collect a sample of 10 vials with 60 eggs in each to measure development time. The last generation is determined once the flies have reached A-type conditions. Once the flies emerge from the pupal casing, the flies are checked every 6 hours. Freshly enclosed flies are removed from the vials, which are sexed and counted.

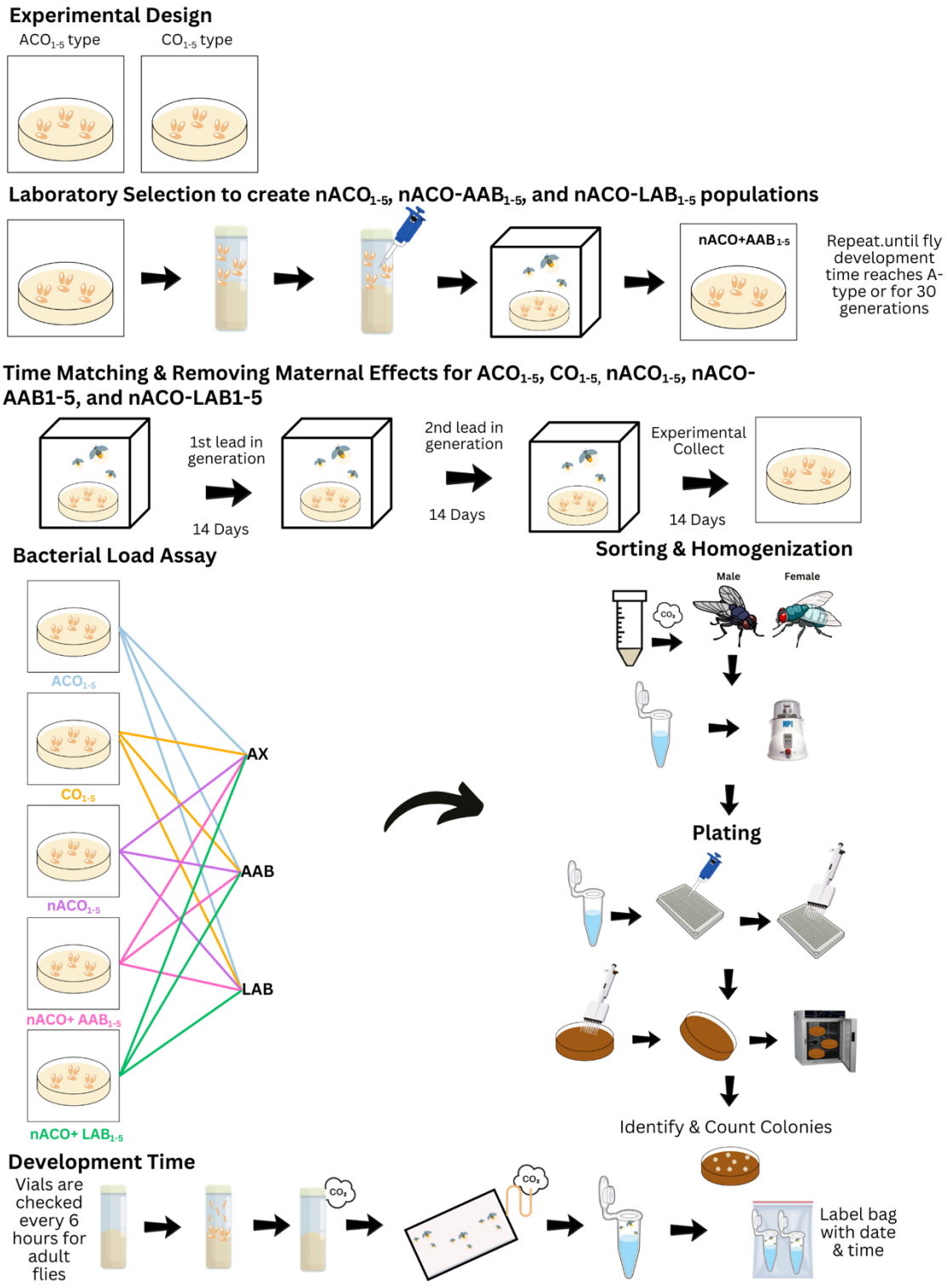


Figure *1*. The overall experimental design shows the control populations (ACO_{1-5} and CO_{1-5}) are laboratory selected to create experimental populations: $nACO_{1-5}$, nACO-AAB_{1-5}, and nACO-LAB_{1-5}. To time match different populations, flies are treated with two 14-day lead-in generations before experimental collection. Bacterial load is quantified by sorting, homogenizing, plating,

and colony counting of male and female flies. Development time is determined by counting vials every 6 hours for emergence of adults to evaluate the microbiota effects on host traits. Selection is continued until the development time equivalent to A-type flies or for 30 generations.

Results

Preliminary results

Generation 0 data from conventional ACO and CO populations were used to determine their bacterial abundance and the development time to predict experimental population results. At Generation 0 CO populations had more bacterial abundance overall compared to ACO populations. However, Figure 2 demonstrates that AAB is more abundant in ACO populations, whereas LAB is more abundant in CO populations.

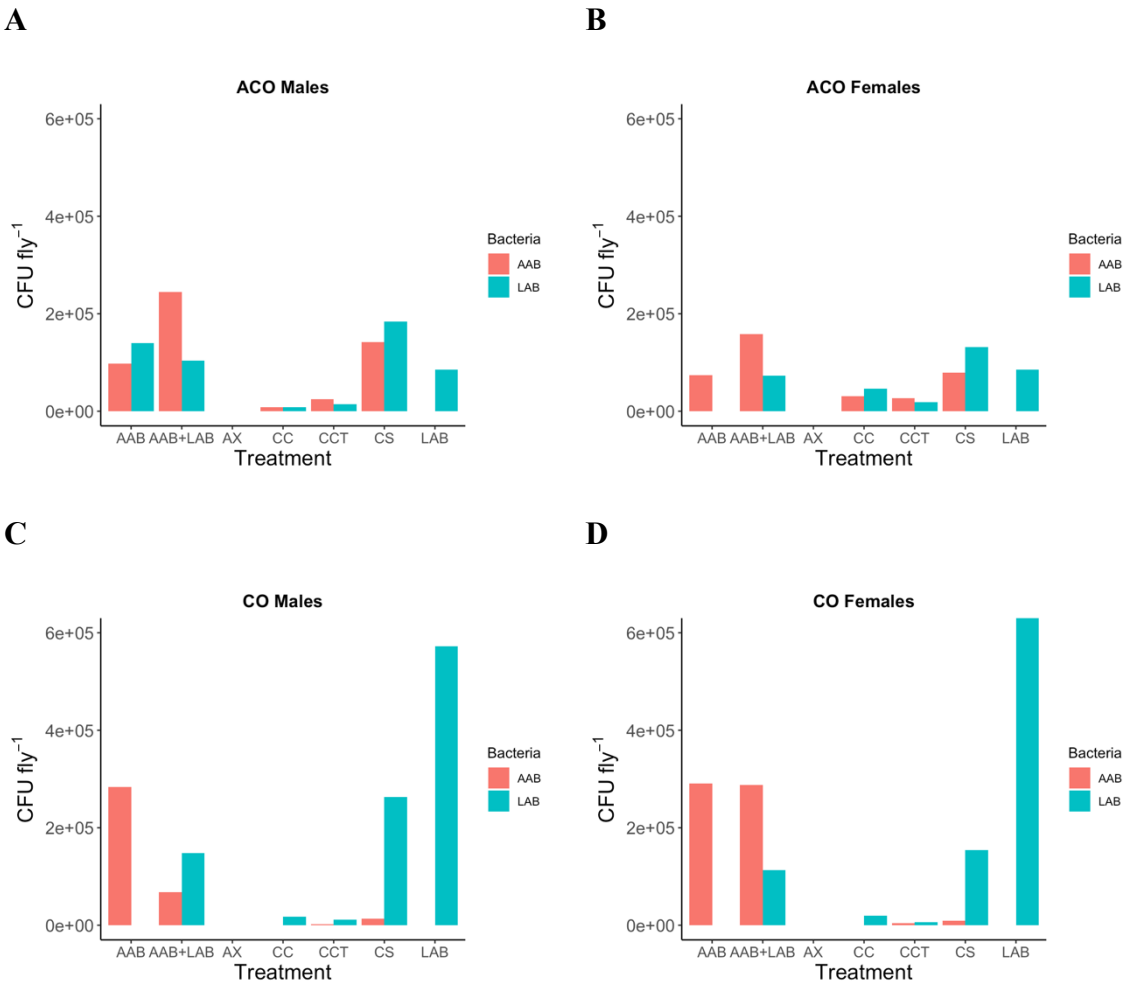


***Fig 2. D. melanogaster* Acetic Acid Bacteria and Lactic Acid Bacteria Abundance for Generation 0 using the CFU equation per fly for conventional conditions.** A) ACO males demonstrating more abundance for AAB. Total sample size per treatment is AAB=53, AAB+LAB=75, AX=75, CC=75, CCT=52, CS=21, LAB=75. B) ACO females demonstrating more abundance for AAB. Total sample size per treatment is AAB=53, AAB+LAB=75, AX=75, CC=72, CCT=53, CS=26, LAB=75. C) CO males demonstrating more abundance for LAB. Total sample size per treatment is AAB=75, AAB+LAB=63, AX=75, CC=71, CCT=75, CS=75, LAB=75. D) CO females demonstrating more abundance for LAB. Total sample size per treatment is AAB=75, AAB+LAB=60, AX=75, CC=66, CCT=75, CS=75, LAB=75.

Generation 0 data also demonstrates that flies under conventional conditions develop more quickly for A-type flies compared to C-type flies. Figure 3A demonstrates that ACO flies take fewer hours for flies to emerge as adults compared to CO flies which take longer to develop as seen in Figure 3B. This aligns with previous research by Chippindale et al. 1997 which indicates that ACO populations develop a lot quicker compared to CO. The research also shows that ACO populations are quickly developing, but short-lived lived and CO populations are long developing and long-lived (Chippindale et al. 1997). The difference in bacterial abundance for the two populations and their development time help explain the evolutionary relationship between microbiota and life history traits.

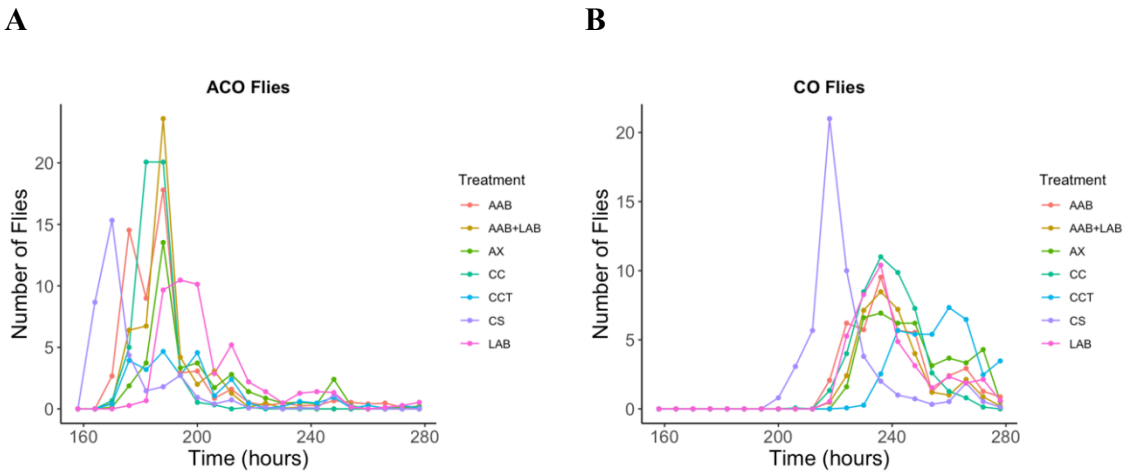


***Figure 3. D. melanogaster* development time for Generation 0 under conventional conditions.** A) Development time for ACO populations show flies under CS treatment are developing quicker compared to flies under LAB treatment which develop slower. AAB+LAB treatment shows the largest number of flies emerged as adults compared to CCT treatment showing the least number of flies emerged. B) Development time for CO populations show flies under CS treatment are emerging into adult flies the quickest compared to CCT flies developing the slowest. CS treatment also shows the largest number of flies emerged as adults compared to AX treatment showing the least number of flies emerged.

Discussion

The preliminary results for Generation 0 fruit flies show that under conventional conditions, A-type flies develop more quickly than C-type flies. Given the differences in development time and the bacterial abundance between A-type and C-type populations, we can infer that as flies evolve, their bacterial abundance will also change. Using Generation 0 as the baseline of the project, we can compare changes in the phenotype and genotype of different populations throughout the experiment. It is expected that by Generation 15, the conventionally evolved nACO populations would have the same bacterial abundance as A-type flies. Compared to conventionally evolved flies, the additional supplementation of AAB and LAB is expected to show a difference in response times for nACO populations to achieve A-type bacterial abundance.

The acetic acid-supplemented populations (nACO+AAB) are expected to have a slower response to selection for fast development, as the selection pressure is relieved. This means that the population supplemented with AAB will take longer time to show changes in life history phenotype. Under constant supplementation with acetic acid, nACO populations will evolve the ability to regulate their AAB abundance more rapidly. The lactic acid supplemented populations (nACO + LAB) are expected to respond more quickly to selection for fast development, since LAB buffers selection pressure. Under constant supplementation of lactic acid, the populations will evolve the ability to control their LAB abundance more gradually. These predicted results

indicate that the supplementation of AAB and LAB influences the response to rapid selection, showing that the microbiome affects life history phenotypes in fruit flies.

These findings should introduce new insights into the interactions between microbiomes and life history traits, which have previously been shown to affect nutritional phenotypes (Chaston et al., 2016). Since earlier studies indicated a trade-off between lifespan and reproduction, our projected outcomes suggest that bacteria may reduce the necessity for a fly's genotype to evolve, as some selection pressure is mitigated by the bacterial supplementation (Mathews et al., 2021; Gould et al., 2018). If the results reveal a significant difference as predicted, it will provide compelling evidence supporting the strong impact of bacteria and corroborate previous investigations. Comparisons with similar studies may address the long-standing question of which factor is more crucial, genotypes or microbiomes. Understanding the mechanisms driving the evolution of bacteria and fruit flies offers valuable insights into the broader dynamics of evolution. This knowledge can help expand practical applications of fast adaptive evolution in healthcare, agriculture, and conservation. Given that the results could have potential applications in other fields, it is essential to complete the project.

References

Chaston, J. M., Dobson, A. J., Newell, P. D., & Douglas, A. E. (2016). Host Genetic Control of the Microbiota Mediates the Drosophila Nutritional Phenotype. *Applied and Environmental Microbiology*, *82*(2), 671-679. https://doi.org/doi:10.1128/AEM.03301-15

Chippindale, A. K., Alipaz, J. A., Chen, H. W., & Rose, M. R. (1997). EXPERIMENTAL EVOLUTION OF ACCELERATED DEVELOPMENT IN DROSOPHILA. 1. DEVELOPMENTAL SPEED AND LARVAL SURVIVAL. *Evolution; international journal of organic evolution*, *51*(5), 1536–1551. https://doi.org/10.1111/j.1558-5646.1997.tb01477.x

Garland, T., & Rose, M. R. (Eds.). (2009). *Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments* (1st ed.). University of California Press. http://www.jstor.org/stable/10.1525/j.ctt1ppqbc

Gould, A. L., Zhang, V., Lamberti, L., Jones, E. W., Obadia, B., Korasidis, N., Gavryushkin, A., Carlson, J. M., Beerenwinkel, N., & Ludington, W. B. (2018). Microbiome interactions shape host fitness. *Proceedings of the National Academy of Sciences*, *115*(51), E11951-E11960. https://doi.org/doi:10.1073/pnas.1809349115

Graves, J. L., Jr, Hertweck, K. L., Phillips, M. A., Han, M. V., Cabral, L. G., Barter, T. T., Greer, L. F., Burke, M. K., Mueller, L. D., & Rose, M. R. (2017). Genomics of Parallel Experimental Evolution in Drosophila. *Molecular biology and evolution*, *34*(4), 831–842. https://doi.org/10.1093/molbev/msw282

Matthews, M. K., Malcolm, J., & Chaston, J. M. (2021). Microbiota Influences Fitness and Timing of Reproduction in the Fruit Fly Drosophila melanogaster. *Microbiology Spectrum*, *9*(2), e00034-00021. https://doi.org/doi:10.1128/Spectrum.00034-21

Mirzoyan, Z., Sollazzo, M., Allocca, M., Valenza, A.M., Grifoni, D., Bellosta, P. (2019). Drosophila melanogaster: A Model Organism to Study Cancer. Frontiers, 10. https://doi.org/https://doi.org/10.3389/fgene.2019.00051

Rose, M., Passananti, H. B., & Matos, M. (2004). *Methuselah flies: A Case Study in the Evolution of Aging*. World Scientific. https://doi.org/10.1142/5457

Walters, A. W., Hughes, R. C., Call, T. B., Walker, C. J., Wilcox, H., Petersen, S. C., Rudman, S. M., Newell, P. D., Douglas, A. E., Schmidt, P. S., & Chaston, J. M. (2020). The microbiota influences the Drosophila melanogaster life history strategy. Molecular Ecology, 29(3), 639-653. https://doi.org/https://doi.org/10.1111/mec.15344

Wong, A. C., Chaston, J. M., & Douglas, A. E. (2013). The inconstant gut microbiota of Drosophila species revealed by 16S rRNA gene analysis. *Isme j*, *7*(10), 1922-1932. https://doi.org/10.1038/ismej.2013.86

A Biological Lens on Breast Tumor Classification Using Machine Learning Techniques

Vivian Duong, Sabrina Ly

Advisor: Dr. Sunny Le

California State University, Fullerton

Abstract

Breast cancer remains one of the most prevalent and life-threatening cancers among women worldwide. Early and accurate classification of tumors as benign or malignant is essential for effective treatment. In this study, we applied logistic regression modeling and the cross-validation method to the Breast Cancer Wisconsin (Diagnostic) dataset to investigate the classification of tumors based on nuclear features. After addressing multicollinearity and standardizing variables, our final model achieved high predictive accuracy (97.28% on a test set and 92.79% through leave-one-out cross-validation). Significant predictors included smoothness, texture, radius, compactness, and symmetry. As undergraduate students majoring in biology, this research deepened our understanding of statistical modeling in the biomedical context and demonstrated the powerful interdisciplinarity between data science and medicine.

Key words: breast tumor classification, cross validation, machine learning, cancer research

Introduction

As undergraduate students preparing to enter the medical field, we were drawn to studying breast cancer due to its prevalence and impact on public health. Recognizing the importance of early detection and diagnosis, we sought to explore how data science and statistical modeling can contribute to medical advancements. Through our coursework in statistics, we developed an appreciation for how medicine can be greatly supported by mathematical and statistical modeling and decided to explore these interdisciplinary areas. The course also inspired us to undertake a project that integrates our growing skills in data analysis with a critical biomedical issue - breast cancer classification and data science. By conducting this research, we intend to increase our knowledge of breast cancer diagnosis and classification and improve our understanding of data analysis and coding.

According to the American Cancer Society, the most common cancer diagnosed among women in the US is breast cancer. Breast cancer arises from uncontrolled cell growth in breast tissue, which frequently starts in milk ducts or glands. About 1 in 8 women (13.1%) get diagnosed with invasive breast cancer in their lifetime, and 2.3% pass away from the disease, making breast cancer a serious public health concern in the United States (American Cancer Society, 2024).

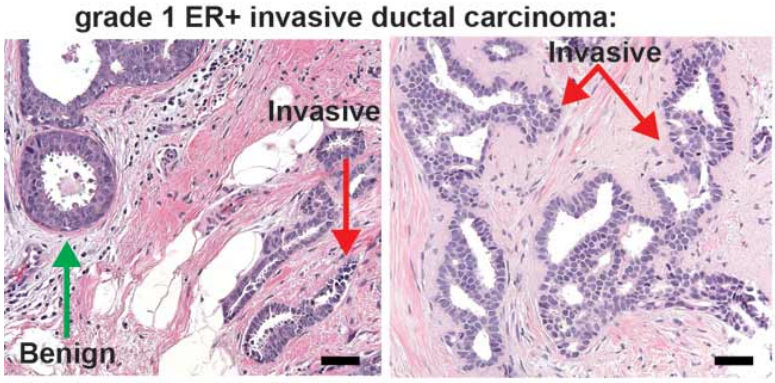


Figure 1. Types of breast cancer. The most common include non-invasive forms like ductal carcinoma in situ, where cancerous cells are contained, and invasive forms that spread outside of ducts or glands (Perry, 2014).

A clinical breast exam is usually the first step in the diagnosis process. Doctors use imaging techniques such as mammograms, ultrasounds, or MRIs to identify irregularities in size, shape, or density. If a suspicious mass is found, a biopsy is performed to determine whether the tumor is benign or malignant by examining tissue samples under a microscope (American Cancer Society, 2024).

Benign tumors are non-cancerous, slow-growing, and do not spread to other parts of the body. At the cellular level, the nuclei of benign tumors are usually smoother, smaller, and more uniform. Malignant tumors, on the other hand, are cancerous, fast-growing, and metastasize or spread to other body parts. The nuclei of malignant tumors tend to have larger radii, irregular textures, higher concavity, and more concave points due to abnormal growth (Raha et al., 2024).

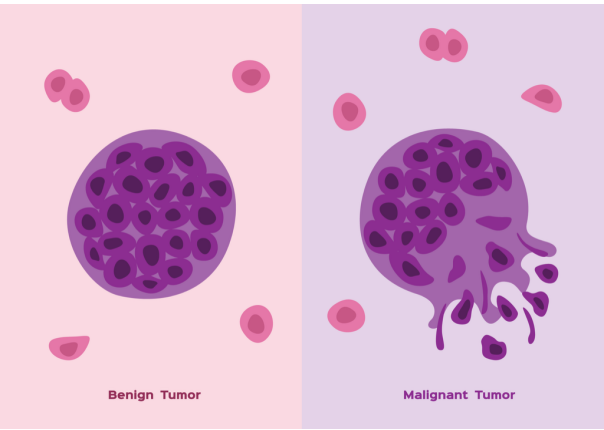


Figure 2. Visual comparison of benign (left) and malignant (right) tumor cell structures (Admac Oncology, 2021).

Additionally, digital mammograms play a crucial role in early detection. However, identifying and correctly classifying benign and malignant patterns in digital mammograms is challenging for radiologists as the appearance of such tumors is subtle and varied, especially in the early stages. Studies such as Sakai et al. (2020) have shown that even experienced radiologists can misclassify tumors based on visual features alone.

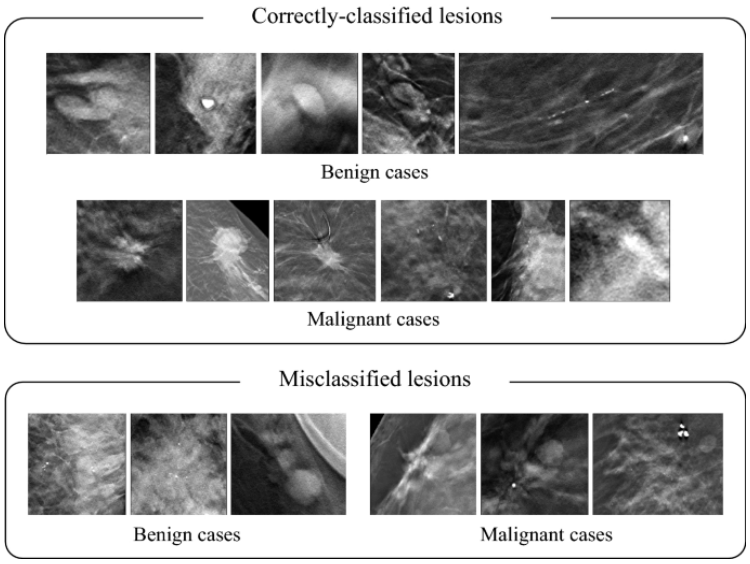


Figure 3. Ultrasound images of correctly classified and misclassified breast lesions for both benign and malignant cases. Visual overlap in features causes misclassification to occur (Sakai et al., 2020).

To support radiologists, machine learning techniques have been developed to improve classification accuracy. Typically, radiologists characterize masses by their location, size, shape, and margins when judging the likelihood of being cancerous. Benign masses are compact and limited to a circular or oval shape, while irregularly shaped masses with ill-defined margins are generally suggestive of malignancy. Verma et al. (2010) proposed that a soft, cluster-based direct learning classifier can significantly enhance and aid radiologists in accurately classifying suspicious areas and diagnosing breast cancer. Their model achieved over 97% classification accuracy, outperforming traditional methods and highlighting the potential of machine learning in diagnostic imaging.

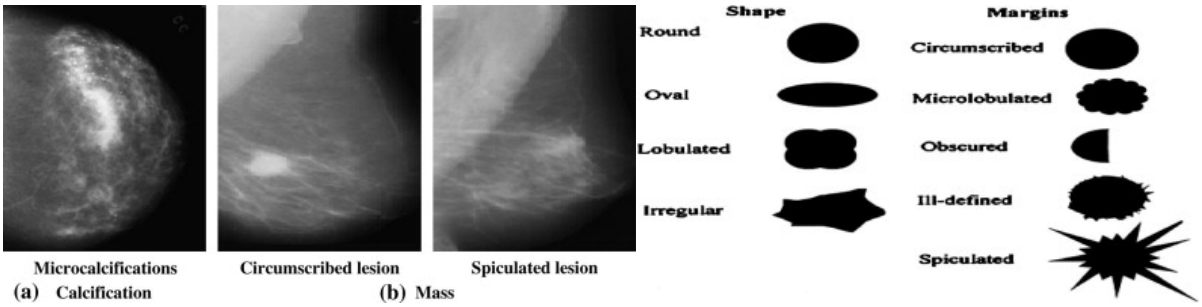


Figure 4. Modified image demonstrating breast abnormalities in mammograms, and different mass shapes and margins (Verma et al., 2010).

Other research suggests that tumor characteristics, being highly correlated with nuclear features, may be predicted at the cellular level using machine learning, a form of artificial intelligence that enables computers to learn through algorithms without explicit programming (IBM, 2025). The study by Kalaiyarasi et al. (2020) explores the classification of tumors using machine learning algorithms to improve medical diagnosis and aid physicians in making decisions about cancer. The authors emphasize the significance of machine learning-based early detection, describing how machine learning algorithms can be used to automate tumor identification. Similarly to our study, the authors stressed the importance of data preprocessing and feature extraction based on correlation.

This study aims to address the question: How can tumors be classified as benign or malignant using data analysis techniques? We apply logistic regression and assess model performance using cross-validation to ensure the robustness and reliability of our models.

Methodology

For this study, we selected the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository, a trusted source for datasets from various fields. This dataset was chosen for its suitability for tumor classification, compatibility with logistic regression, and the absence of missing values, which allowed for efficient preprocessing and analysis. Thus, this

dataset is ideally suited for our research as it offers a strong foundation for analyzing whether tumors are benign or malignant based on various characteristics.

We began by examining the dataset's overall tumor type distribution. Figure 4 below displays the distribution of tumor diagnoses, in which 33% of the tumors are malignant, while roughly 67% are benign. This distribution falls within the range (20-40%) that is considered mildly imbalanced, which is acceptable for training machine learning models (Google).

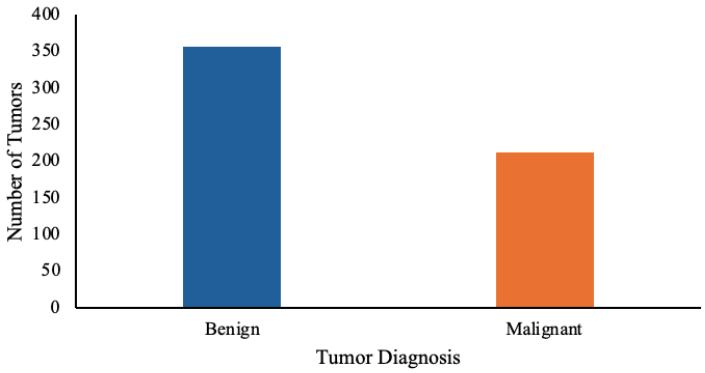


Figure 4: Tumor counts in the dataset.

The dataset includes 10 numerical variables that measure the nuclear features of tumor cells, which will be used as predictor variables for classifying tumor types (Table 1).

Table 1: *Independent Variable List*

Variable	Type and Level
Area	Numerical
Compactness	Numerical
Concave Points	Numerical
Concavity	Numerical
Fractal Dimension	Numerical
Perimeter	Numerical
Radius	Numerical
Smoothness	Numerical
Symmetry	Numerical
Texture	Numerical

All variables are numerical, and larger values are generally associated with malignancy. We conducted an exploratory data analysis (EDA) to explore this relationship by creating numerical summaries for each variable. Table 2 presents each nuclear feature's means and standard deviations (SD) for benign and malignant tumors.

Table 2: *Mean and SD of Tumor Nuclear Features*

Benign	Area	Compactness	Concave Points	Concavity	Fractal Dimension	Perimeter	Radius	Smoothness	Symmetry	Texture
Mean	462.7902	0.0801	0.0257	0.0461	0.0629	78.0754	12.1465	0.0925	0.1742	17.9148
SD	134.2871	0.0337	0.0159	0.0434	0.0067	11.8074	1.7805	0.0134	0.0248	3.9951

Malignant	Area	Compactness	Concave Points	Concavity	Fractal Dimension	Perimeter	Radius	Smoothness	Symmetry	Texture
Mean	978.3764	0.1452	0.0880	0.1608	0.0627	28.0826	17.4628	0.1029	0.1929	21.6049
SD	367.9380	0.0540	0.0344	0.0750	0.0076	21.8547	3.2040	0.0126	0.0276	3.7795

As expected, most nuclear features have higher average values in malignant tumors. However, some variables, such as perimeter, showed exceptions where benign tumors (mean = 28.083) had unexpectedly high values (malignant mean = 21.855), warranting further investigation.

To investigate the presence of outliers in the dataset, we calculated five-number summaries (minimum, Q1, median, Q3, and maximum) for each variable by type of tumor (Table 3 and Table 4). For benign tumors, values beyond the lower limit should be of no concern; however, values greater than the upper limit may be an issue when determining the "cut-off" value in classifying whether a tumor is malignant or benign based on that specific nuclear feature. If higher values are indicators of malignancy, observed benign tumor cells with unusually large areas may be predicted as malignant in the classification model, if considered a

stand-alone factor in determining tumor type. Thus, outliers may influence the accuracy of the classification model and should be carefully examined before creating the model.

Table 3: *Five Number Summary of Benign Tumors*

	Area	Compactness	Concave Points	Concavity	Fractal Dimension	Perimeter	Radius	Smoothness	Symmetry	Texture
LB	9.89	12.39	62.78	278.2	0.0737	0.0309	-0.0046	0.0052	0.1425	0.0542
min	6.981	9.71	43.79	143.5	0.0526	0.0194	0	0	0.1060	0.0519
Q1	11.08	15.15	70.87	378.2	0.0831	0.0556	0.0203	0.0150	0.1580	0.0585
median	12.18	17	78.01	451.1	0.0914	0.0728	0.0351	0.0227	0.1735	0.0614
Q3	13.37	19.76	86.1	551.1	0.1007	0.0976	0.0600	0.0325	0.1890	0.0658
max	17.85	33.81	114.6	992.1	0.1634	0.2239	0.4108	0.0853	0.2743	0.0958
UB	14.47	21.61	93.24	624	0.1090	0.1147	0.0748	0.0402	0.2045	0.0686

Table 4: *Five Number Summary of Malignant Tumors*

	Area	Compactness	Concave Points	Concavity	Fractal Dimension	Perimeter	Radius	Smoothness	Symmetry	Texture
LB	12.81	17.0225	83.02	433.55	0.0853	0.0696	0.0578	0.0477	0.1541	0.0511
min	10.95	10.38	71.9	361.6	0.0737	0.0461	0.0240	0.0203	0.1308	0.0500
Q1	15.075	19.3275	98.745	705.3	0.0940	0.1096	0.1095	0.0646	0.1741	0.0566
median	17.325	21.46	114.2	932	0.1022	0.1324	0.1514	0.0863	0.1899	0.0616
Q3	19.59	23.765	129.925	1203.75	0.1109	0.1724	0.2031	0.1032	0.2099	0.0671
max	28.11	39.28	188.5	2501	0.1447	0.3454	0.4268	0.2012	0.3040	0.0974
UB	21.84	25.8975	145.38	1430.45	0.1191	0.1952	0.2449	0.1248	0.2257	0.0721

Next, we examined multicollinearity among the predictors by computing a correlation matrix (Figure 5). Here, multicollinearity involves the presence of high correlations among predictor variables, which can reduce regression model reliability by raising the coefficient

estimate variance. For example, tumors with a larger radius also have a higher perimeter mean, resulting in a stronger linear relationship between these variables. This creates redundancy and makes it more challenging to determine the specific effect of each variable. To address the issue of multicollinearity in our dataset, we computed the correlation matrix between the 10 predictor variables and identified pairs of variables with a correlation coefficient greater than 0.8. The size and color intensity of the circles reflect the strength and direction of the correlations, where stronger positive correlations are indicated by deeper blue circles. We removed one variable from each highly correlated pair, keeping only the first variable in each pair, to reduce redundancy. By doing so, multicollinearity is reduced, making the model more stable and effective in classifying tumors.

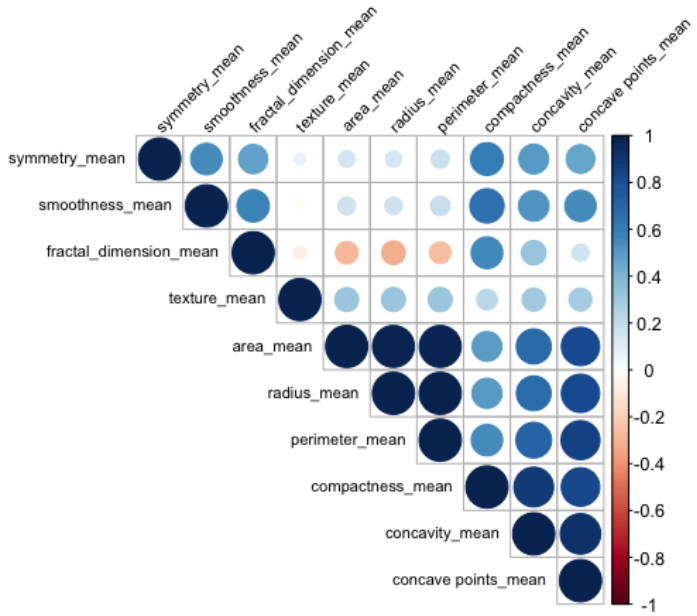


Figure 5. Correlation matrix of predictors from the dataset.

Finally, we applied logistic regression to classify breast cancer tumors as either benign or malignant based on the nuclear features of tumors. Logistic regression is well-suited for binary classification problems in which the outcome variable has two possible categories— in this case,

benign (0) or malignant (1). The goal of logistic regression is to find the best-fitting curve that estimates the probability of an outcome based on predictor values such as tumor area, radius, symmetry, or other nuclear features. The model uses the logistic or sigmoid function to produce an S-shaped curve, to ensure predicted probabilities are bound between 0 and 1. The Maximum Likelihood Estimation (MLE) determines the logistic regression model's coefficients (intercepts and slopes). MLE identifies the set of coefficients that maximizes the likelihood of observing the given data, ensuring that the model's predicted probability for tumor classifications closely aligns with actual outcomes in the dataset.

To evaluate the logistic regression model's performance, we assessed the significance of each predictor through its p-values and created a confusion matrix to calculate classification accuracy. We also applied leave-one-out cross-validation (LOOCV) to further validate the model on unseen data and assess generalizability. Overall, this methodology allowed us to develop a model capable of predicting tumor types based on nuclear features with high accuracy and reliability.

Results

Following the correction of multicollinearity, a logistic regression model was implemented to determine whether a tumor was benign or malignant using standardized predictors. The initial model included all ten nuclear features, and statistical significance was assessed by calculating the p-values for each predictor. Table 5 displays the coefficient estimates and the corresponding p-values for each variable in the initial model to identify the predictors strongly related to tumor classification.

Table 5: *Initial Model Significant Values*

Coefficients:	Estimate	Pr(>|z|)
(Intercept)	-7.260	0.567
radius_mean	-2.049	0.5813
texture_mean	0.385	2.5e-09
perimeter_mean	-0.072	0.885
area_mean	0.0398	0.017
smoothness_mean	76.432	0.017
compactness_mean	-1.462	0.9427
concavity_mean	8.469	0.297
concave points_mean	66.822	0.0192
symmetry_mean	16.278	0.126
fractal_dimension_mean	-68.337	0.424

Predictors such as texture_mean, area_mean, and smoothness_mean were among the variables that showed statistically significant p-values ($p < 0.05$), indicating a possible association with malignancy classification. However, variables such as radius_mean, perimeter_mean, and symmetry_mean had higher p-values, suggesting that their associations with tumor diagnosis were not strong or significant.

To strengthen the model's reliability and clarity, we removed either highly correlated or statistically insignificant variables. The remaining predictors in the model were centered and scaled. The final model included the six variables: radius_mean, texture_mean, smoothness_mean, compactness_mean, symmetry_mean, and fractal_dimension_mean. Table 6 shows the p-values and coefficient estimates for the final model. Among these predictors, malignancy was strongly associated with radius_mean, texture_mean, smoothness_mean, and symmetry_mean (all $p < 0.05$), but not compactness_mean or fractal_dimension_mean.

Table 6: *Final Model Significant Values*

Coefficients:	Estimate	Pr(>|z|)
(Intercept)	-37.564	7.11e-14
radius_mean	1.235	7.31e-14
texture_mean	0.326	1.48e-07
smoothness_mean	77.556	0.0014
compactness_mean	13.312	0.053
symmetry_mean	22.616	0.0340

Residual deviance: 146.12 on 416 degrees of freedom AIC: 158.12

The final model demonstrated high predictive performance, with a low residual deviance (146.12, df = 416) and AIC (158.12), indicating the model reliably differentiated benign and malignant diagnoses (Table 6). A confusion matrix generated from a 74/26 training/test split showed an overall accuracy of 97.28%, correctly classifying 89 benign and 54 malignant tumors, with only 4 misclassifications (Table 7). The 74% training and 26% testing division was determined arbitrarily, within the commonly accepted range of 70-80% for training data, balancing the need for model learning with reliability of performance evaluation (Gholamy et al., 2018).

Table 7: *Final Model Confusion Matrix*

	Actual	
Predicted	B	M
B	89	1
M	3	54

Accuracy: 0.9728

LOOCV further confirmed the model's reliability, achieving an accuracy of 92.79% (Table 8). The LOOCV confusion matrix shows that the model correctly predicted 341 benign and 187 malignant tumors, with 41 misclassifications overall. Classification was based on a

probability threshold (alpha) of 0.5, where tumors with predicted probabilities above 0.5 were labeled malignant.

Table 8: *LOOCV Confusion Matrix*

	Actual	
Predicted	B	M
B	341	25
M	16	187

Accuracy: 0.9279

To assess variable contributions beyond p-values, we fitted the final model on the full dataset and examined feature importance. Smoothness_mean had the highest importance score (86.909), followed by fractal_dimension_mean (33.546), and the least being texture_mean (0.337). Notably, fractal_dimension_mean – despite being statistically insignificant – showed a substantial influence on model predictions, suggesting that feature importance may capture predictive value not always reflected by significance testing alone.

Discussion

The final logistic regression model achieved high classification accuracy, both on the test data (97.28%) and in cross-validation (92.79%), indicating its strong ability to differentiate between benign and malignant tumors. Among the predictors, smoothness_mean emerged as the most influential features, as shown by both statistical significance and feature importance. Interestingly, some discrepancies were noted between p-value significance and feature importance rankings. For instance, fractal_dimension_mean was determined to be the second greatest important feature in prediction outcomes, despite the fact that it was considered statistically insignificant based on its p-value ($p > 0.05$). This suggests that relying solely on p-values for variable selection may overlook valuable predictive information. On the other hand, texture_mean, which had a highly significant p-value, contributed relatively little to model

predictions when considered alongside other variables. This emphasizes the importance of evaluating multiple metrics – such as coefficient size, p-value, and feature importance – when determining a variable's overall impact on classification accuracy.

Evaluating the performance of the final model after cross-validation, an accuracy of 0.9279 indicates that it is successful in correctly classifying tumors roughly 93% of the time based on the features chosen. These odds are fairly good, especially considering that the applications of this model would be to the diagnosis of breast tumors, which can lead to better patient prognosis if it can be identified early and accurately for treatment. However, it should be noted that using an alpha threshold of 0.5 assumes equal costs for false positives and false negatives. In clinical settings, misclassifying a malignant tumor as benign poses a greater risk than the reverse. Adjusting the classification threshold to prioritize sensitivity could reduce the risk of overlooking potentially harmful tumors, though at the expense of more false positives.

Overall, the findings highlight the effectiveness of logistic regression in tumor classification while revealing opportunities for refining variable selection methods and optimizing classification thresholds for practical applications.

Limitations and Future Research

While our logistic regression model demonstrated high accuracy, several limitations must be acknowledged. First, the Breast Cancer Wisconsin (Diagnostic) dataset was initially collected in 1993. Developments in pathology, medical imaging, and our knowledge of tumors over the last three decades may have led to new discoveries on how tumors develop and are classified. As a result, models developed using this dataset may not fully reflect recent diagnostic standards or newer tumor characteristics. Second, the dataset is cross-sectional and static, meaning it only contains one-time measurements of the nuclear characteristics of each tumor. Tumor features

such as size, texture, and shape may evolve over time in hospitals, especially when treatment varies. Due to the lack of longitudinal data, this study cannot account for the potential effects of these features' development on classification accuracy.

Third, the dataset includes each nuclear feature's standard deviation (SD) and worst (highest) values, which could offer additional predictive power. Including these variables in future models may improve classification accuracy and robustness.

Last, our model was developed and validated on a single dataset. Although LOOCV provides a reliable internal estimate, external validation on independent datasets is essential to ensure generalizability across diverse patient populations and clinical settings. Future research should explore incorporating additional feature types, validating the model across datasets, and optimizing classification thresholds to better match real-world clinical needs.

Conclusion

This study demonstrated the effectiveness of logistic regression in classifying breast tumors as benign or malignant based on nuclear features, achieving high accuracy both on test data and through cross-validation. Our findings highlight the valuable role that data-driven approaches can play in supporting early cancer diagnosis and improving patient outcomes.

As undergraduate students, this project allowed us to bridge the gap between statistical theory and real-world biomedical applications. Through conducting this research, we strengthened our skills in data analysis, coding, and critical evaluation of model performance — experiences that will be essential as we pursue future careers in medicine and research. Moving forward, we hope to build on this foundation by incorporating more advanced methods and working with up-to-date clinical datasets to further enhance diagnostic accuracy.

Acknowledgements

We sincerely express our gratitude to UROC for their funding that allowed us to begin our data science journey. Your support contributed greatly to our learning and experience.

We especially would like to express our gratitude and appreciation to our mentor, Dr. Sunny Le. We thank you for performing the coding for the various tests in this paper. We also thank you for your guidance and knowledge that strengthened our data analysis skills. As students, we are extremely grateful for your keen eye in spotting our potential to grow and encouraging us further as we would not have gained these valuable experiences without it.

References

Admac Oncology (2021). *What are the differences between malignant and benign tumours?*. *ADMAC ONCOLOGY*. https://www.admaconcology.com/2021/06/10/malignant-vs-benign-tumor-know-the-differences/

American Cancer Society (2024). *Breast Cancer Facts & Figures 2024-2025*. Atlanta: American Cancer Society, Inc.

Datasets: Imbalanced datasets. Google. https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation*. Int J Innov Res Sci Eng Technol. 5(3):411-419.

IBM. (2025). *What is machine learning (ML)?* https://www.ibm.com/think/topics/machine-learning

Kalaiyarasi, M., Dhanasekar, R., Ram, S. S., & Vaishnavi, P. (2020). *Classification of benign or malignant tumor using machine learning*. IOP Conference Series: Materials Science and Engineering (Vol. 995, No. 1, p. 012028). IOP Publishing.

Patel, A. (2020). *Benign vs malignant tumors*. JAMA oncology, *6*(9), 1488-1488.

Perry, C. (2014) *Tugging on the 'malignant' switch*. Harvard John A. Paulson School of Engineering and Applied Sciences. https://seas.harvard.edu/news/2014/06/tugging-malignant-switch

Raha, A. D., Gain, M., Hassan, Md. M., Bairagi, A. K., Dihan, F. J., Adhikary, A., Hossain, Md. B., Murad, S. A., Al-Shehari, T., Kadrie, M., & Alsadhan, N. A. (2024). *Modeling and predictive analytics of breast cancer using Ensemble Learning Techniques: An explainable artificial intelligence approach*. Computers, Materials & Continua, *81*(3), 4033–4048. https://doi.org/10.32604/cmc.2024.057415

Sakai, A., Onishi, Y., Matsui, M., Adachi, H., Termoto, A., Saito, K., & Fujita, H. (2020). *A method for the automated classification of benign and malignant masses on digital breast tomosynthesis images using machine learning and radiomic features*. Radiol Phys Technol 13, 27–36. https://doi.org/10.1007/s12194-019-00543-5

Verma, B., McLeod, P., & Klevansky, A. (2010). *Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer.* Expert systems with applications, *37*(4), 3344-3351.

Chemistry and Biochemistry

Development of fluorescent methods of determining Dissociation Constant (Kd) for ssDNA-aptamer-based biosensors

Amanda Reyes and Dr. Stevan Pecic

Department of Chemistry and Biochemistry, California State University, Fullerton

Abstract

Biosensors are designed devices that come in contact with an analyte to provide a measurable signal. Recently, there has been research interests in developing single stranded DNA (ssDNA) aptamers as biosensors. In our lab we use a biochemistry technique called SELEX (Systematic Evolution of Ligands by Exponential Enrichment) to identify ssDNA aptamer candidates. Once SELEX yields possible aptamers their dissociation constants, Kd, is evaluated through fluorescent assay analysis. Most accurate testing of aptamers is using the 6-carboxyfluorescein (6-FAM)/DABCYL system. However, this process is expensive and labor intensive. Thus, there is a need for a fast and robust screening fluorescent assay to determine best aptamer candidates for 6-FAM/DABCYL testing. In this project we aim to develop a fluorescent assay method by using fluorescent dyes, Thiazole Orange (TO) and Thioflavin T (ThT), to determine the Kd for single stranded DNA aptamers. Using a commercially available ssDNA library of 36 random nucleotides flanked by two primers and through multiple rounds of SELEX, we isolated several aptamers for Kavain and Aldosterone. We have finished ThT and TO fluorescent testing on each aptamer, and we compared each dye's ability to bind to the ssDNA. From these experiments, we concluded that ThT is a better dye to use when testing preliminary ssDNA aptamer's Kd. We expect that our study will enable development of the more advanced biosensors for detection of challenging small molecules of interest.

Introduction

Aptamers are short oligonucleotide sequences that recognize or bind to a target with high affinity and specificity they can be either single stranded DNA (ssDNA) or RNA or peptides [1]. They are isolated from the process of Systematic Evolution of Ligands by Exponential Enrichment (SELEX). Aptamers are similar to antibodies however they have unique characteristics that allow them to surpass antibodies capabilities such as straight forward chemical synthesis, cost effectiveness, and adaptability [2]. RNA aptamers targeted for amino acids can distinguish between different polarities, side chains, and different shapes. Specifically, the aptamer, Tyr 1, binds to amino acid L-Tyrosine has been shown to be stereo-specific [3]. Additionally, an aptamer called enantioselective can distinguish L-arginine from D-arginine with a 12,000-fold stronger affinity for the L conformation of the amino acid [4]. However, RNA libraries are generated from double stranded DNA; therefore, after each round of PCR, the library is reverse transcriptase is used to regenerate the library, and this step can introduce synthesis mistakes in the RNA sequence [1].

DNA aptamers have high stability, which allows the molecules to have a longer shelf life and can be stored at room temperature [5]. Aptamer identification occurs without the use of live animals though in vitro processes. Since the identification is going through *in vitro* selection, the aptamers can be used to identify molecules toxic to animals, or molecules produce an immunogenic response [6]. To quantify an aptamer's quality, they are often measured by their cross reactivity and their binding affinity. Binding affinity is reported as the dissociation constant (K_d) and ideally has a range of picomolar to micromolar.

The process to identify an aptamer is called SELEX (Fig. 1). This process requires the immobilization of the DNA library on to the column through non-covalent methods but through the use of a capture. A capture is a complementary base sequence that binds to a primer on the

library. Biotin is attached to the capture sequence because biotin strongly interacts with the medium streptavidin agarose. This method does not require a structural change to either the ssDNA or the target but instead relies on the structure switching capabilities of DNA and the weak hydrogen interactions with the complementary capture sequence. Target is introduced, and the library and target are eluted as the aptamer-target complex [7, 8]. Target concentration is decreased every few rounds of SELEX to increase aptamer affinity. [9]. The library is then introduced to a target to create the aptamer-target complex. Then the aptamer-target complex is eluted and collected for PCR amplification. PCR amplification will result in double stranded DNA, which will need to be reduced to single strands. After the strand separation step, the library is used in a new round to further enrich the library. This process may continue for 5-20 rounds of SELEX. After the last round of SELEX, the aptamer candidates are cloned and sequences for characterization [1, 10].

Results and Discussion

In this work, we used 14 rounds of SELEX and 4 rounds of counter SELEX to identify an aptamer for the target kavalactone, kavain. Our library consisted of 10^{14} unique ssDNA sequences with a random region of 36 nucleotides bookended between two primers (Table 1). We maintained the concentration of kavain until round 14 and introduced counter target, Pipermethystine (Fig. 2). Through immobilizing the library on a column, the target was added and based on structure switching capabilities aptamer- target complex was eluted and collected. Each round the quality of DNA was monitored using gel electrophoresis to determine how many rounds of PCR should be used to amplify our library. Through evolutionary pressure we isolated several possible aptamer sequences (Table 2). These sequences tested with Thioflavin T (ThT) did not produce an aptamer that resulted in a low K_d. We hypothesized that this may be due the target kavain's structure not

containing epitopes or recognizable functional groups. A similar problem was seen in the research to find an aptamer for glucose. The research team chose to create an aptamer that will bind to glucose and a receptor that changes conformational shape when interacting with glucose. Researchers used Shinkai's receptor bound to glucose as a target for SELEX. Shinkai's receptor was chosen due to the modification to override the preference the receptor would have for fructose and instead bind to glucose. The receptor when in contact with glucose then undergoes a conformational change, which allows a slight new arrangement of epitopes. The protocol included using receptor and target complex for SELEX until the library pool showed a high enough affinity for the complex. Then counter SELEX was run with the Shinkai receptor used as a counter-target, which reduces the aptamer's cross reactivity. This resulted in an aptamer capable of identifying glucose [11]. Therefore, future directions will change the target from kavain alone to a protein that is highly selective for kavain and then using counter-SELEX against the receptor to isolate an aptamer that binds specifically to kavain. As well as testing those sequences with the 6-FAM Dabcyl system and ThT to determine their dissociation constants.

We pivoted to creating a new protocol for fluorescent assays for determining a dissociation constant. A new preliminary step will rapidly narrow the large pool of potential ideal aptamer sequences. Reducing the costs to run the more expensive and more accurate 6-carboxyfluorscein (6-FAM) and 4-((4-(dimethylamino) phenyl) diazenyl)benzoic acid] (DABCYL) quenching method. Using previously published and unpublished aldosterone aptamer sequences we decided to create a preliminary step using fluorescent dyes Thiazole Orange (TO) or Thioflavin T (ThT). Using preciously published and unpublished aptamer sequences allowed us to validate our preliminary step by comparing dissociation constants from ThT or TO with the 6-FAM Dabcyl

system. The two sequences tested resulted in a K_d of 33 nM and 338 nM were respectively aptamers 14 and 23 seen in Table 3. Each sequence was truncated for optimal binding.

These assays show that the good fluorescent dye for preliminary testing dissociation constant of aptamers is Thioflavin T. This dye gave the best results in that the K_d values are smaller values and best correlate to the 6-FAM Dabcyl system that indicated aptamer 14 was an ideal aptamer for aldosterone. ThT also gave the best results for dissociation curve (Fig. 3). These tests indicate that ThT can be accurate enough to narrow the list of potential aptamers. Thiazole Orange showed less consistency with dissociation values in relation to the previous 6-FAM tests, the graphs produced do not follow the curve to indicate decreased binding, and the trends from secondary analysis are less correlated than with ThT data. This may be an indication that TO binds more tightly to DNA so that it cannot structure switch to bind with its target.

Conclusion

In the future, we will optimize the ThT assay by measuring different concentrations of dye and aptamer. We will also need to further validate the test by using more previously published aptamers that were tested using the 6-FAM Dabcyl system. With the rising interest in biosensors as aptamers a new preliminary step can be useful in cutting down time on testing and narrow the field of aptamer sequences so that only the most ideal candidates are tested reducing costs. Our results show, ThT is the best cost-effective fluorescent dye comparatively to the most sensitive and accurate standard, 6-carboxyfluorscein (6-FAM) and 4-((4-(dimethylamino) phenyl) diazenyl)benzoic acid] (DABCYL) quenching method. This preliminary step would precede the most precise assay 6-FAM Dabcyl to narrow aptamer candidates. With this work more aptamers can be tested and therefore aid in publishing new sequences for a widening array of targets.

Tables and Figures

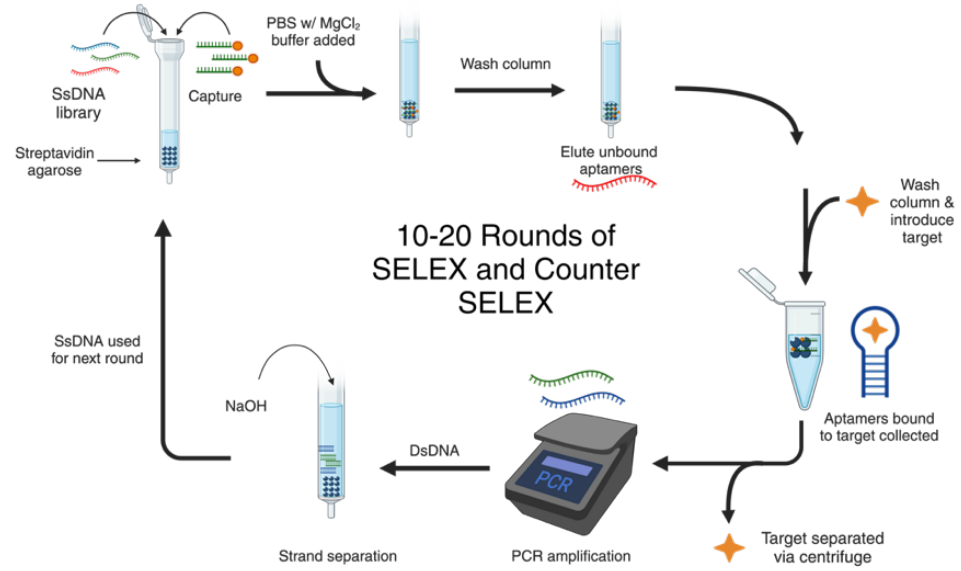


Figure 1. Shows SELEX schematic process for each round.

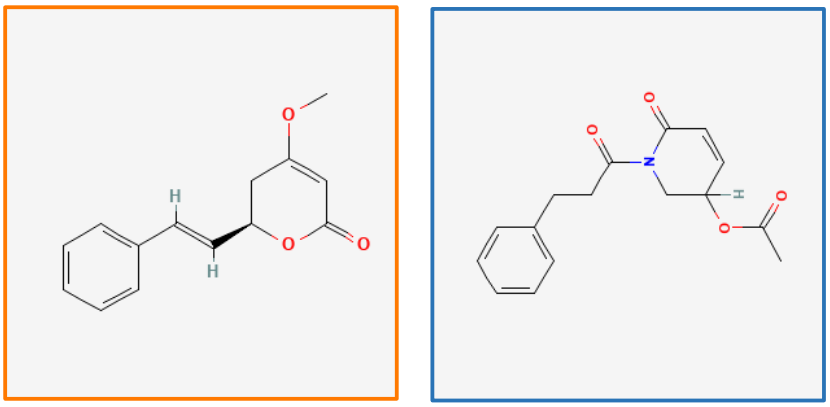


Figure 2. Chemical strucutres of Kavain and Pipermethystine.

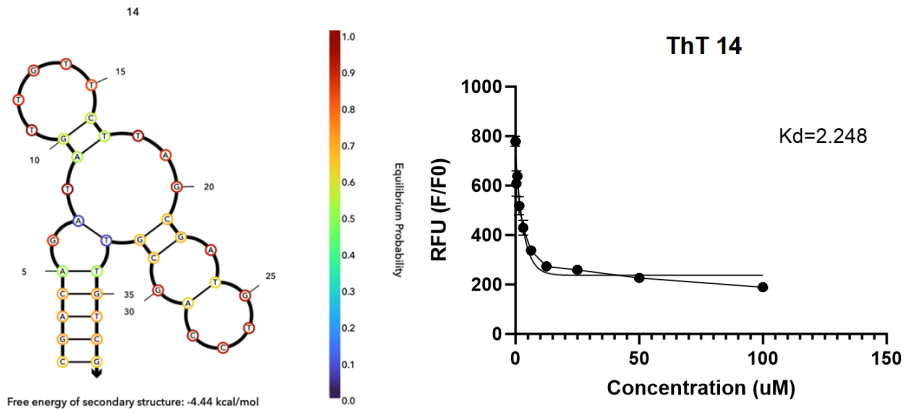


Figure 3. 2D Structure of Aldosterone 14 aptamer (NuPack) and ThT fluorescent assay results.

Table 1. Shows sequences used for expirement.

Primers	***Sequence***
Primer 1(forward)	GGA GGC TCT CGG GAC GAC
Primer 2(reverse)	CTG TAA ATC CTA AAG GCG GGA CGA C
P2_biot: /5Biosg/	CTG TAA ATC CTA AAG GCG GGA CGA C

Table 2. Shows sequences resulting from SELEX.

Number	*Sequence (without primers)*
1	GGAGGCTCTCGGGACGACCACAGGATCGTTGGTACACTGGTTCGGGGAGTCGTCCCGCCTTTAGGATTTACAG
4	GGAGGCTCTCGGGACGACGATGCGTCCTTGCTCTCTCTTTCGTGTCGTGTCGTCCCGCCTTTAGGATTTACAG
5	GGAGGCTCTCGGGACGACGGCATGCAAGGTTCTTTGTTTCGTGCTTCAGTCGTCCCGCCTTTAGGATTTACAG
6	GGAGGCTCTCGGGACGACGGGGCACTTTGTTGTTTTGTTTTTTGTCCTGTCGTCCCGCCTTTAGGATTTACAG
7	GGAGGCTCTCGGGACGACCAGGAGCTTCATTGTGTACTTTTCTGCTGTGTCGTCCCGCCTTTAGGATTTACAG
8	GGAGGCTCTCGGGACGACCGGCGTAGTTATCTTTCGGTCTTTTGTCGTGTCGTCCCGCCTTTAGGATTTACAG
9	GGAGGCTCTCGGGACGACCGAAGAGGGGGTGTGTTTGTGTTGGTCTAGGTCGTCCCGCCTTTAGGATTTACAG
10	GGAGGCTCTCGGGACGACGGGCATGTGTTATCTACCCTCCTTTAGCCAGTCGTCCCGCCTTTAGGATTTACAG
11	GGAGGCTCTCGGGACGACCATAGTGCAGTTCTTCGTAGTTTGTTGCTGGTCGTCCCGCCTTTAGGATTTACAG
12	GGAGGCTCTCGGGACGACCACGTGCTCTCTTGTTTTCTTGTTTAGTGTGTCGTCCCGCCTTTAGGATTTACAG
14	GGAGGCTCTCGGGACGACCATGATGCCTCTTTCCCTTTTTGCTCGTGCGTCGTCCCGCCTTTAGGATTTACAG
15	GGAGGCTCTCGGGACGACGGGGGGTTCTGTCTTGTTCTATTTACCGCTGTCGTCCCGCCTTTAGGATTTACAG
16	GGAGGCTCTCGGGACGACGGGTGCAGTTGTTCGTTCTTCTTTCACCTTGTCGTCCCGCCTTTAGGATTTACAG
17	GGAGGCTCTCGGGACGACGGGCAGGTGGCTGGTAGGGTAGGGCTGTATGTCGTCCCGCCTTTAGGATTTACAG
19	GGAGGCTCTCGGGACGACGGCCGTGTTGTCTTTTCTCGCTTTTGGATCGTCGTCCCGCCTTTAGGATTTACAG

20	GGAGGCTCTCGGGACGACCAACGGCTTGTACTTCTTTCGTTTTGTTGTGTCGTCCCGCCTTTAGGATTTACAG
22	GGAGGCTCTCGGGACGACGGAACAGGCACGTGGAATGGTGTACTTGAGGTCGTCCCGCCTTTAGGATTTACAG
23	GGAGGCTCTCGGGACGACCATGCCGGATGTGGTGATCGTGTTGCTTGTGTCGTCCCGCCTTTAGGATTTACAG
25	GGAGGCTCTCGGGACGACCAGCTCAGAGGATGCTCCGGTGTGTGTTGCGTCGTCCCGCCTTTAGGATTTACAG
27	GGAGGCTCTCGGGACGACGGGTCACCGATGTATCATGTGTCGGGGATGGTCGTCCCGCCTTTAGGATTTACAG

Table 3. Shows aldosterone aptamers tested and their resulting dissociation constants.

Number ALDO	*SEQ*	*ThT K_d (nM)*	*TO K_d (nM)*
14	CGACAGATAGTTGTTCTTAGCGATGTCCAGCGTTGTCG	2.248	29.13
15	ACGACGACAGTGCCTTGATATACGTTGGGCTGGTAGTCGT	26.2	33698
05	GACGACGAGACCCTTGATACTTCAAGCGTCAGCGAAGTCGTC	0.7306	0.4091
	CTGGGGTGCAATTCTTATGTAACGGGTCCCGG	3994244	60.57
17	CGACCGGAGTGTCTTAGTGTATAACTGAGTTTAGGTCG	29.57	6345158
23	CGACGGTAGGTAGGCCAACTGGGTATTTACTGGTGTCG	281.8	179.4
16	CGACAAATCTTTGGTGTGATGGATGTTGTCTTGTGTCG	2632540	1739195
22	ACGACGACCGGTCCCTGTTGTTAGGAATAGGGGGAGTCGT	4945384	3002228
06	ACGACGATGGGCCCCGATATGTACATTTCGGGTGAGTCGT	0.3785	0.3008
13	GGACGACGGAGACCCCATGTTTTCTGGTGTCAGCGTAGTCGTCC	225.7	83.76
10	ACGACGGAGGGTCCCGTTGTTCTACGATGGGTTTAGTCGT	36.68	47414
02	CGACGGGACACTTTGTATGTAAAGTGAGGTTCACGTCG	107369	2948616
07	ACGACGGGTGAGGTTCTTTATAACTGATGGGCCCTGTCGT	3332006	5084175

References

1. Chinchilla-Cárdenas, D.J., et al., *Current developments of SELEX technologies and prospects in the aptamer selection with clinical applications.* Journal of Genetic Engineering and Biotechnology, 2024. **22**(3): p. 100400.
2. Silwal, A.P., et al., *Aptamer-Assisted DNA SELEX: Dual-Site Targeting of a Single Protein.* ACS Biomaterials Science & Engineering, 2025.
3. Mannironi, C., et al., *Molecular recognition of amino acids by RNA aptamers: the evolution into an L-tyrosine binder of a dopamine-binding RNA motif.* Rna, 2000. **6**(4): p. 520-7.
4. Ku, T.H., et al., *Nucleic Acid Aptamers: An Emerging Tool for Biotechnology and Biomedical Sensing.* Sensors (Basel), 2015. **15**(7): p. 16281-313.
5. Cai, R., et al., *Systematic bio-fabrication of aptamers and their applications in engineering biology.* Systems Microbiology and Biomanufacturing, 2023. **3**(2): p. 223-245.
6. Qian, S., et al., *Aptamers from random sequence space: Accomplishments, gaps and future considerations.* Analytica Chimica Acta, 2022. **1196**: p. 339511.
7. Zhao, Y., A.Z. Li, and J. Liu, *Capture-SELEX for Chloramphenicol Binding Aptamers for Labeled and Label-Free Fluorescence Sensing.* Environment & Health, 2023. **1**(2): p. 102-109.
8. Yang, K.-A., R. Pei, and M.N. Stojanovic, *In vitro selection and amplification protocols for isolation of aptameric sensors for small molecules.* Methods, 2016. **106**: p. 58-65.
9. Wang, J., et al., *Influence of target concentration and background binding on in vitro selection of affinity reagents.* PLoS One, 2012. **7**(8): p. e43940.
10. Jauset-Rubio, M., et al., *One-Pot SELEX: Identification of Specific Aptamers against Diverse Steroid Targets in One Selection.* ACS Omega, 2019. **4**(23): p. 20188-20196.
11. Yang, K.A., et al., *Recognition and sensing of low-epitope targets via ternary complexes with oligonucleotides and synthetic receptors.* Nat Chem, 2014. **6**(11): p. 1003-8.

Geology

Cement paragenesis of septarian concretions of the Holz Shale

Ms. Jamie Hoffman
Faculty Mentor: Professor Sean Loyd
Geology (B.Sc.), 2025, Department of Geological Sciences
714-356-5236
jlhoffman@csu.fullerton.edu

Abstract

The late Cretaceous Holz Shale hosts calcite concretions of ellipsoidal morphology up to 1 m in diameter. Some of these concretions exhibit septarian veins that are filled with multiple generations of cement. Septarian varieties of Holz Shale concretions exhibit up to four distinct calcite phases including 1) an initial gray cemented shale body, 2) a light brown early fringe outer phase, 3) a dark brown late fringe inner phase, and 4) a latest course crystalline vein filling white spar phase. Cement phases were analyzed for their carbon isotope composition ($\delta^{13}C$). The phase-specific $\delta^{13}C$ ranges are as follows: 1) body, 0 to -16‰, 2) light brown early fringe, 0.5 to -11‰, 3) dark brown late fringe, -8‰, 4) crystalline vein phase, -10 and -18‰. These changing compositions suggest that each exhibited a unique formation pathway with shifting contributions from organic carbon. The $\delta^{13}C$ ranges for each phase imply fractional contributions of organic matter that range from 0.08 to 0.7 with body phases incorporating the least organic-derived carbon and spar phases incorporating the most organic-derived carbon. The paragenetic sequence and δ13C-derived organic carbon contributions reveal that successive cement phases precipitated at deeper depths, primarily within the sulfate reduction zone. These findings provide further support concretions (and authigenic carbonates in general) form across a range of mineralization depths through progressive cementation.

Introduction

Past studies on concretions have concluded ways to determine their timing and formation mechanisms (Mozley & Burns, 1993). There remain many gaps in knowledge for the study of concretions. One major issue is the lack of modern analogues, as no modern concretions are identical to ancient concretions (Coleman, 1993). Modern concretions have only been found in nonmarine environments like marshes, while many ancient concretions were formed in marine settings, further complicating our understanding of concretion formation mechanisms (Loyd and Berelson, 2016). Concretions vary greatly in shape, size, and chemical composition (e.g., Mozley and Burns, 1993). The causes of these variations are not fully understood (Coleman, 1993). It is believed that existing interpretations of concretions may underestimate the depth and duration of concretion growth (Raiswell & Fisher, 2000). Septarian concretions are concretions with branching veins that are often filled with diagenetic minerals such as calcite (Rainswell & Fisher, 2000). Septarian concretion formation is also not completely understood. The origin of septarian fissures is still debated but may be the result of tensile fractures (Astin, 1986) or dehydration (Duck, 1995; Raiswell & Fisher, 2000).

It is important to continue to study septarian concretions. Septarian concretions record the environmental conditions at the time of formation, providing insights into how waters evolve through progressive diagenesis. Concretions provide a proxy reservoir for shallow diagenetic environments and provide information on transient processes that would not be accessible otherwise. Indeed, carbonate concretions have been used to reconstruct changes in shallow diagenetic environments even through deep geological time (e.g., Loyd et al., 2023). Organic matter is a common constituent of sediments, and its degradation leads to the release of carbon as dissolved inorganic carbon. This dissolved inorganic carbon drives concretion formation. As

such, concretion formation represents the product of organic matter degradation. With increasing amounts of CO_2 in the atmosphere, it is essential to more completely understand the mechanisms whereby natural systems sequester carbon. With the increasing effects of climate change, it has never been more important to create a greater understanding of environments that exhibit extensive carbon cycling.

Geologic Background

The Holz Shale is the uppermost member of the Ladd Formation which crops out in the Santa Ana Mountains of California (Figure 1, Buck and Bottjer, 1985). Diagnostic fossils indicate that the Holz Shale formed during the Upper Cretaceous, approximately 70 million years ago (Buck and Bottjer, 1985). The Holz Shale concretions range from ~2 to 100 cm in diameter and make up ~4-5% of the outcrop area. The presence of chute deposits, gully deposits, specific fossils, and turbidites indicates that the Holz Shale formed in a near-shore marine environment (Buck and Bottjer, 1985). The Holz Shale was likely deposited within an anoxic to sulfidic environment as evidenced by the presence of laminated sediments, absence of bioturbation, and evidence of thin-shelled bivalve fossils (Buck and Bottjer, 1985).

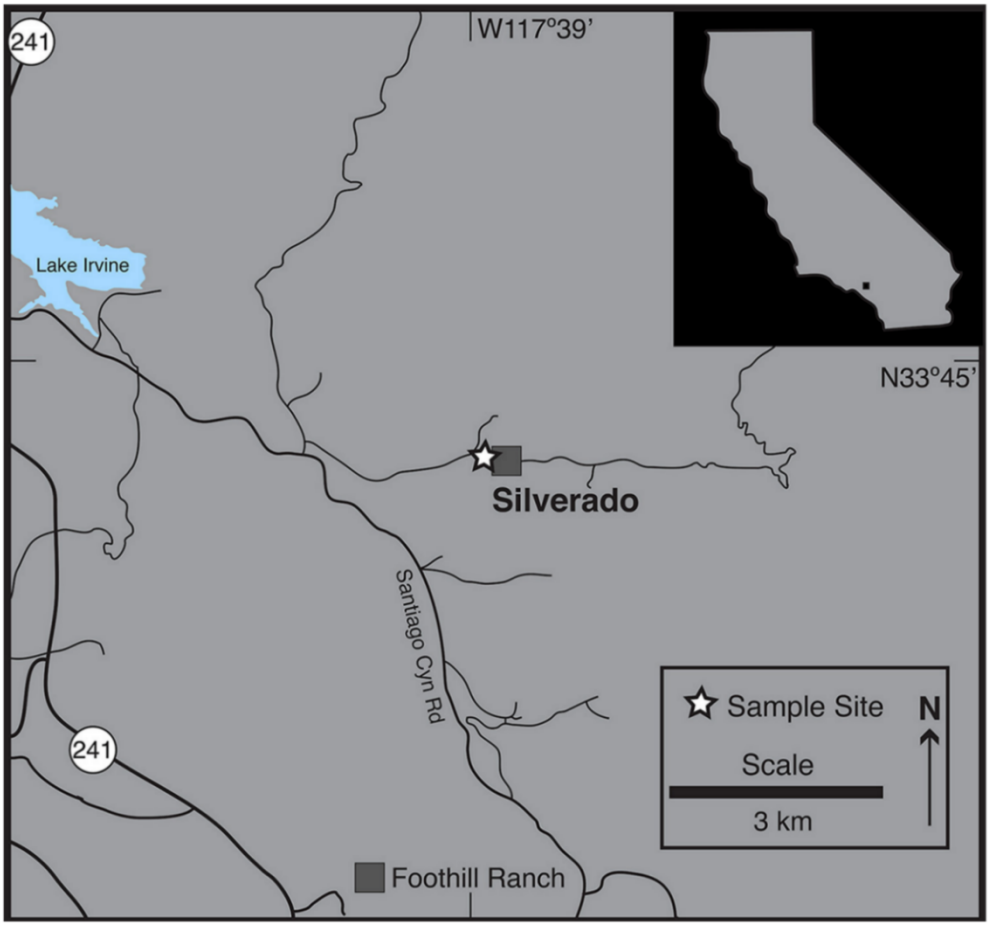


Figure 1 Study Site location in the Santa Ana Mountains, California. Figure after Loyd et al., 2017.

Research Question and Hypothesis

The goal of this study is to better understand the formation conditions of Holz Shale septarian concretions through petrographic and carbon geochemical analysis. Petrographic analysis allows identification of the different cement phases that occur in the septarian concretions. If phases express significantly different carbon isotope compositions, then different mechanisms of formation and/or variations in carbon sources can be inferred (e.g., Coleman, 1993). This study's hypothesis is that the different phases of Holz Shale septarian concretions formed via unique pathways and/or incorporate carbon from different sources, as indicated by unique $\delta^{13}C$ ranges.

Materials and Methods

Seven samples of Holz Shale septarian concretions were taken. Where present, each phase was drilled and powdered using a Dremel Rotary Tool. Multiple drill sites were collected for each phase. Produced powders were weighed and collected into Exetainer vials. The sampled powder from each drill site was run in triplicate. After powdered samples were placed in Exetainer vials, the ambient atmosphere was removed by vacuum. 10% phosphoric acid was added into each vial to convert the powdered carbonate to CO_2 gas. The CO_2 was passed into a Picarro G2121-*i* Cavity Ringdown Spectrometer via Automate Carbonate Preparation Device and measured for total inorganic carbon (TIC) content and carbon isotope composition ($\delta^{13}C$). TIC contents are reported as weight percent (wt%), and $\delta^{13}C$ values are reported in permil (‰) relative to the VPDB standard.

Results

Four calcite cement phases are identified in Holz Shale Septarian concretions. They include the body, outer fringe, inner fringe, and spar phases as shown by Figure 2. Most of the concretions are composed of gray cemented shale and have septarian veins lined with fringing cements and filled with white sparry calcite. Within most veins lies an outer fringe, inner fringe, and vein-occluding spar phase; not all veins contain all phases, and sometimes only show inner fringe cements. The outer fringe or early fringe, is light brown and formed first as indicated by its position closest to the initial body phase. The inner fringe or late fringe is dark brown and formed after the outer fringe and body phases. The last phase to form is the white crystalline calcite spar phase.

After the data was processed, $\delta^{13}C$ and total inorganic carbon weight percent were calculated for each phase. The body phase has the largest variability with $\delta^{13}C$ ranging from 0 to -16‰ and TIC from 7-12 wt.%. The early fringe has a $\delta^{13}C$ range from +0.5 to -11‰ with a TIC range from 7-12 wt.%. The late fringe has a $\delta^{13}C$ range from -8.5 to -15‰ and a TIC range from 8-10 wt.%. The spar phase has a $\delta^{13}C$ range between -10 and -18‰ and a TIC content of around 12 wt.%.

Petrographic photos of the Holz Shale give additional insight into phase spatial relationships and reveal a potential latest-stage dolomite phase shown in Figure 3. At the microscopic level, phases appear to have filled in directly after the sparry calcite phase has been found in septarian concretions from other locales (Thyne and Boles, 1979).

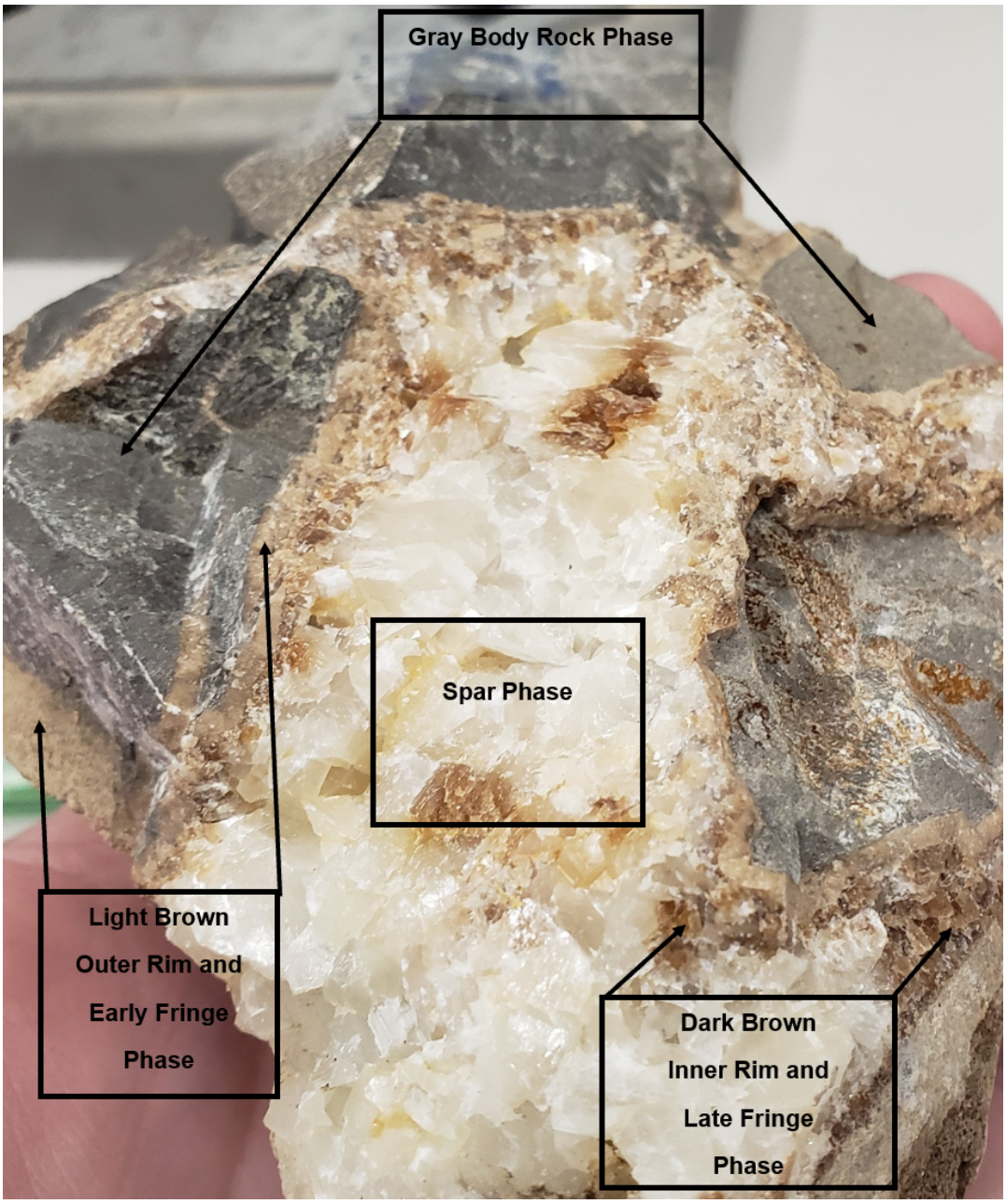
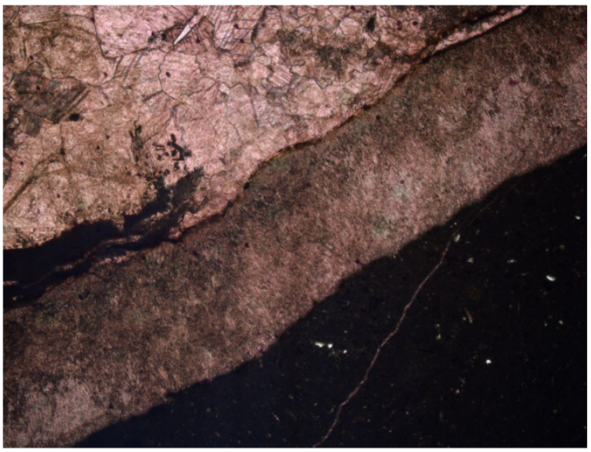
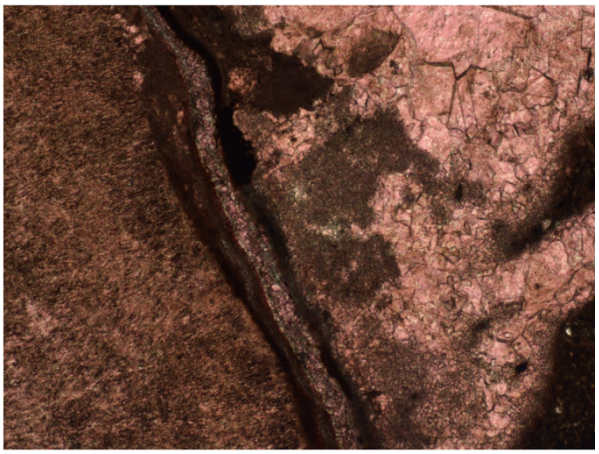


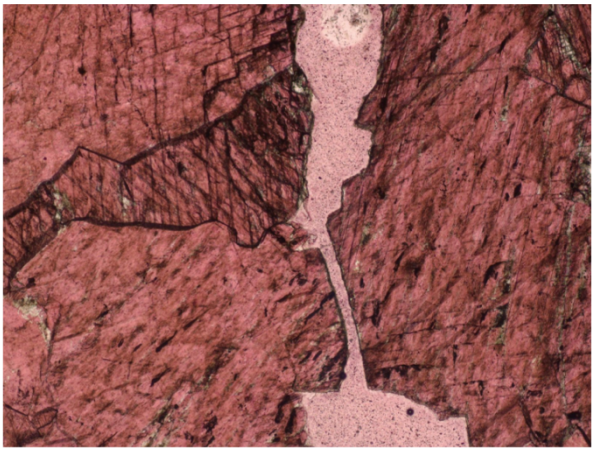
Figure 2: Photograph of Holz Shale septarian concretion with each of the four phases labeled. The four phases are a gray body rock, a light brown outer rim phase, a dark brown inner rim phase, and a crystalline spar phase.



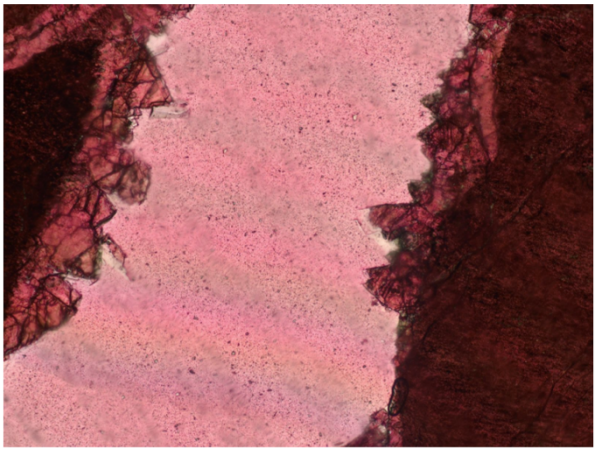
HSC1-4 PPL 2x: Multiple Phases



HSC1-4 PPL 5x: Border between phases



HSC 6-8 PPL 5x: Dolomite within Spar



HSC 6-9 PPL 2x: Dolomite within Spar

Figure 3: Petrographic photos of the Holz Shale under plain polarized light. Width of view for upper left and lower right is 4464 μm, width of view for upper right and lower left is 1936 μm.

Discussion

The paragenesis of the Holz Shale Septarian concretions began with the body cement followed by the outer fringe phase, the inner fringe phase, and lastly the vein-occluding crystalline phase, as indicated by Figure 2. The body must have formed first and then fractured, followed by the formation of the early outer light brown fringe phase. Fractures were further filled with a second layer of dark brown fringing cement. Lastly a crystalline calcite spar phase formed, including septarian veins.

Figure 4 displays each of the six samples with their corresponding TIC contents and $\delta^{13}C$ compositions. Figure 5 groups the six samples of the Holz Shale septarian concretion by phase, to allow more direct data comparison. Figure 5 shows that most $\delta^{13}C$ values for the spar phase are similar, and most are clustered together between -10 and -18‰ and TIC around 12 wt.%. As most spar phases have a similar range, it is likely that the spar phases from different concretions formed under similar conditions. The $\delta^{13}C$ and TIC values for other phases are less distinct and show some overlap.

Figure 5 displays that the late fringe cements express $\delta^{13}C$ values of ~ -8 to -15‰ and TIC between 8-10 wt.%. Although most of the data for late fringe are tightly grouped, a few data points of late fringe are outside of this range. It is possible that most late fringe formed together under the same conditions or formation mechanisms, and that some of the late fringe formed under different conditions.

Figure 5 does not indicate a clear correlation between total inorganic carbon contents and isotope compositions for the early fringe. The early fringe displays wide ranges in TIC and $\delta^{13}C$. The wide ranges for the early fringe cements imply a range of conditions for formation pathways. It is important to note that some of the early fringe data cluster together with 7-8 wt.

% TIC and -7 to -5‰ $\delta^{13}C$. This cluster of data points could indicate that some of the early fringe from different concretions formed at about the same time and under similar conditions.

Like the early fringe, Figure 5 does not show a clear correlation between TIC and $\delta^{13}C$ for the body phase. Like the early fringe, the body phases exhibit appreciable data ranges. The body phases range from 7- 12 wt.% TIC and from 0 to -16‰ $\delta^{13}C$. It is important to note that there is little overlap between the spar and body phases, further suggesting that there was a time gap or significant change in conditions between the times each was forming.

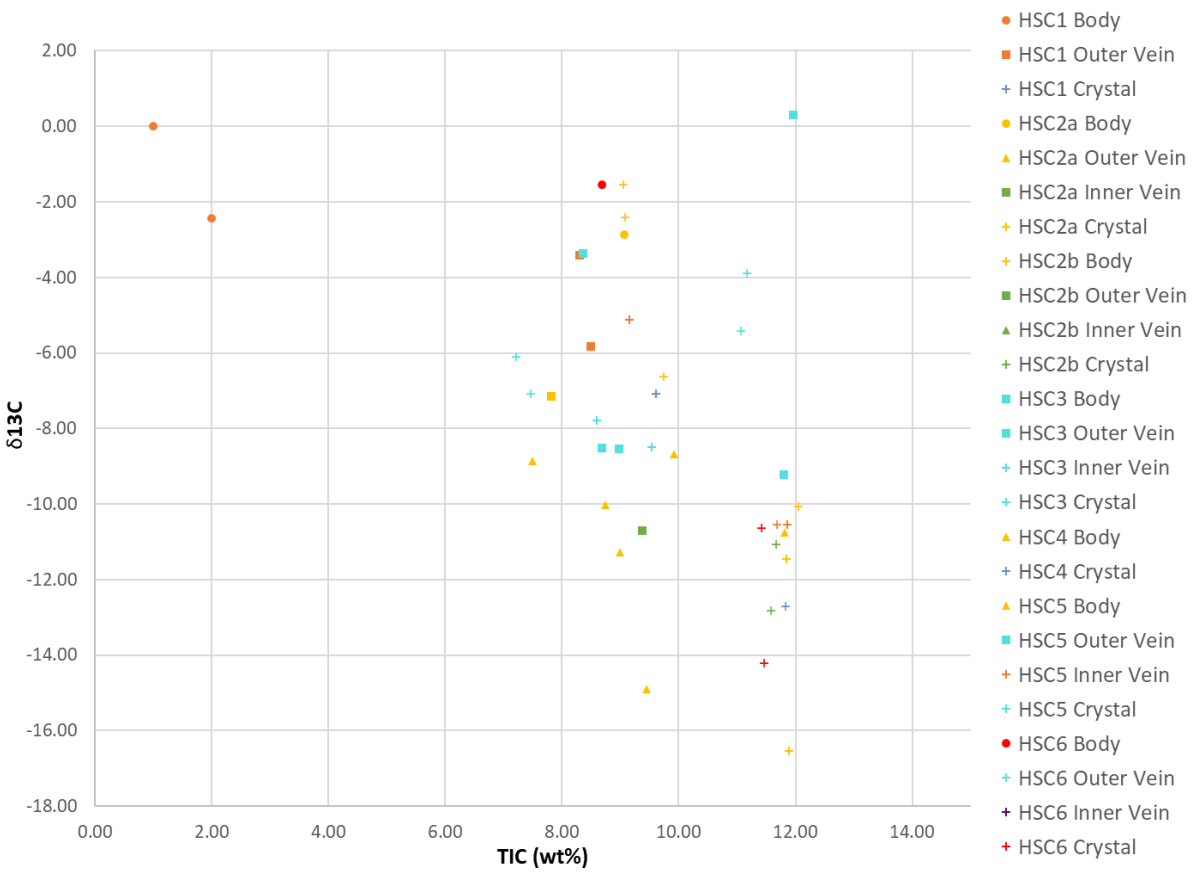


Figure 4: $\delta^{13}C$ compared to TIC of Holz Shale septarian concretions.

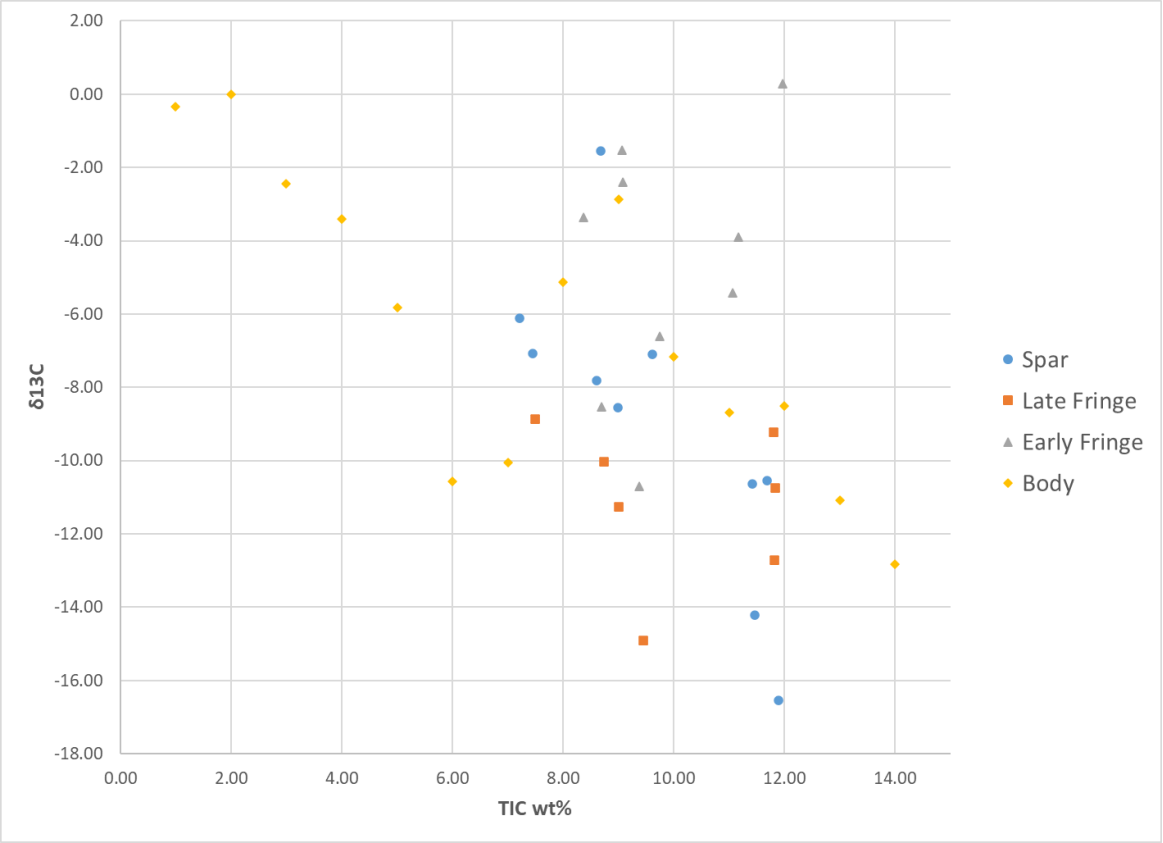


Figure 5: $\delta^{13}C$ compared to TIC of Holz Shale septarian concretions highlighting each phase.

Although the body and spar cements express disparate data ranges, most of the phases have some overlap (Figure 5). The overlap among the phases could indicate that formation conditions changed slowly over time, and that cement phases precipitated under somewhat similar but evolving diagenetic conditions. Indeed, the pore water profile shown in Figure 6 demonstrates that dissolved inorganic carbon expresses a smooth, consistent change with depth. In shallow sediments, this change is manifested as a decrease that represents an increase in the contribution of organic-degradation-derived inorganic carbon with depth and through time, primarily resulting from sulfate reduction diagenetic pathways.

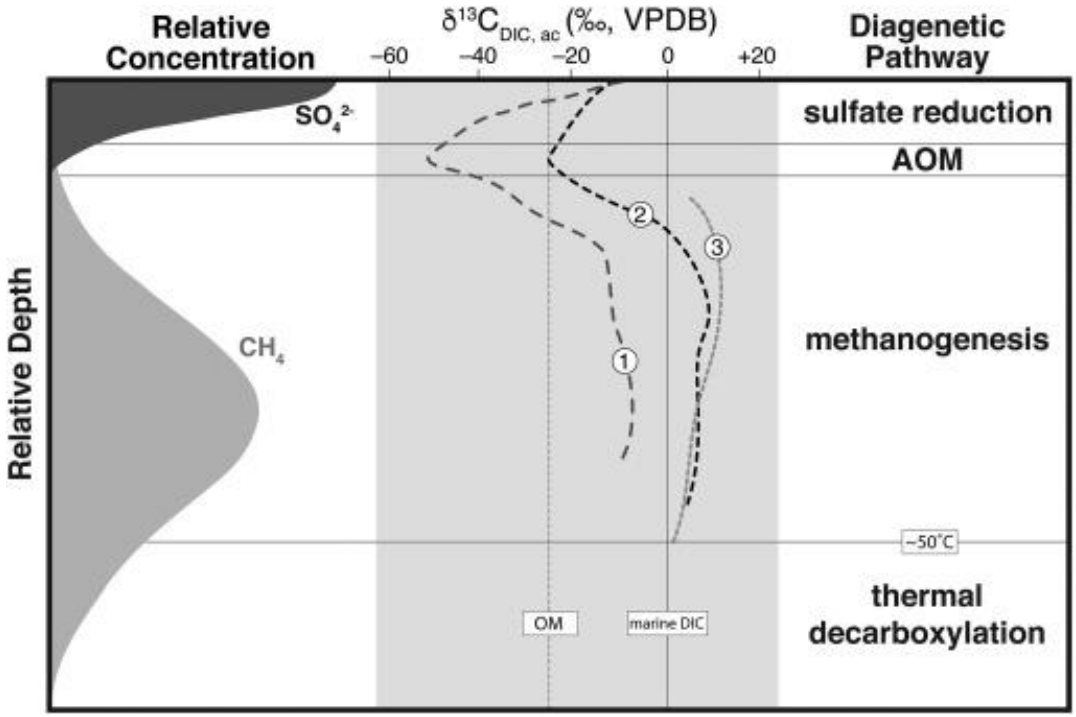


Figure 6: Summarized marine pore water profile. Middle column depicts changing $\delta^{13}C$ compositions that correlate to depth changes and associated diagenetic pathways. Figure from Loyd and Smirnoff (2021).

Figure 7 displays a range of the fractional contribution of organic matter to the four Holz Shale septarian concretion phases. The fractional contribution of organic matter or F_{org} is calculated using the following equation:

$$F_{org} = (\delta^{13}C_{cement} - \delta^{13}C_{sw}) / (\delta^{13}C_{org} - \delta^{13}C_{sw}).$$

Where $\delta^{13}C_{cement}$, $\delta^{13}C_{sw}$, and $\delta^{13}C_{org}$ are the isotope compositions of the concretion cements, late Cretaceous seawater, and organic matter, respectively. Late Cretaceous seawater $\delta^{13}C$ was ~ +1.8‰ (Veizer et al., 1999). The organic matter $\delta^{13}C$ value is -25‰, consistent with Holz Shale-contained organic matter data presented by Loyd (2017). The concretion's cement $\delta^{13}C$ values are those that were measured here. This equation assumes that the dissolved inorganic carbon in Holz Shale pore waters was derived from seawater and organic matter (converted upon diagenetic degradation) and does not take into consideration potential contributions from other carbon sources (e.g., methane).

After the body phase formed and then fractured, the early fringe could precipitate within the fracture. After the early fringe formed, the late fringe could precipitate within the remaining fracture void space. Lastly, after the late fringe formed, the crystalline spar phase precipitated and potentially completely occluded the fracture. As a period must have elapsed between the formation of the body phase and the spar phase, it makes sense that these phases are most geochemically divergent. Furthermore, it is likely that the organic material contributions to and formation mechanisms of these two phases were different.

The formula to determine F_{org} is dependent on the $\delta^{13}C$ value of the concretion cement. As such there is a clear separation between ranges of F_{org} of body and spar phases. The variation

between F_{org} of each of the phases of the Holz Shale indicates different carbon sources. There is a consistent trend of increasing dissolved organic carbon with increasing depth in marine sediments, driven by the addition of remineralized organic carbon. The $\delta^{13}C$ of each phase is different, suggesting that the $\delta^{13}C$ value was decreasing with time and thus depth. This trend is consistent with an increase in organic matter contribution to the dissolved inorganic carbon pool with depth. Therefore, each of the four phases of the Holz Shale formed with progressive burial of host sediments, an assertion supported by Figure 7.

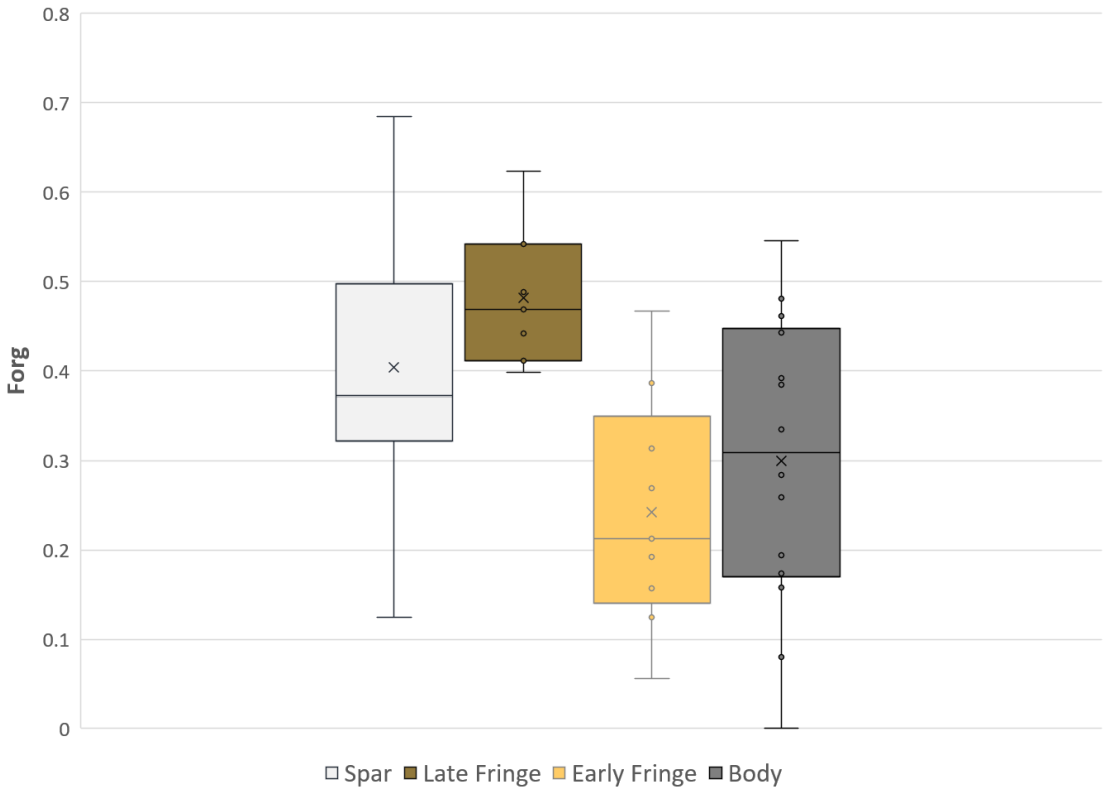


Figure 7- Box and whisker plot of the fractional contribution of organic matter (F_{org}) of each of the four phases of the Holz Shale septarian concretions.

Figures 6 and 8 can be used to determine the diagenetic pathway of each phase based on their $\delta^{13}C$ values. The body phase $\delta^{13}C$ ranges from ~0 to -16‰, which is consistent with the diagenetic pathway of sulfate reduction (Figure 6). The early fringe phase $\delta^{13}C$ ranges from 0.5 to -11‰, which likewise agrees with the zone of sulfate reduction. The late fringe phase $\delta^{13}C$ is about -8‰, which is also attributable to sulfate reduction. The spar phase $\delta^{13}C$ ranges from -10 to -18‰, which may also reflect sulfate reduction with a minor contribution from the AOM pathway. As some data ranges are large, the exact diagenetic pathway cannot be determined by $\delta^{13}C$ alone, but the diagenetic pathway was likely limited to these two possibilities. The occurrence of pyrite within Holz Shale concretions supports a sulfate reduction origin (Loyd et al., 2012, Loyd, 2017).

The continuous depletion in $\delta^{13}C$ compositions across the different phases within Holz Shale septarian concretions implies that the precipitation of cements was progressive. This supports the idea that concretions and authentic marine carbonates experience progressive formation with burial (e.g., Loyd and Smirnoff, 2022). Perhaps most intriguing, however, is the relationship between progressive cementation and septarian fracturing. The data presented here suggest that septarian fracturing occurred relatively early in the paragenetic sequence, and that all of the cementation occurred at relatively shallow sediment depths. This has broad implications for the mechanisms driving septarian fracturing in other units as well.

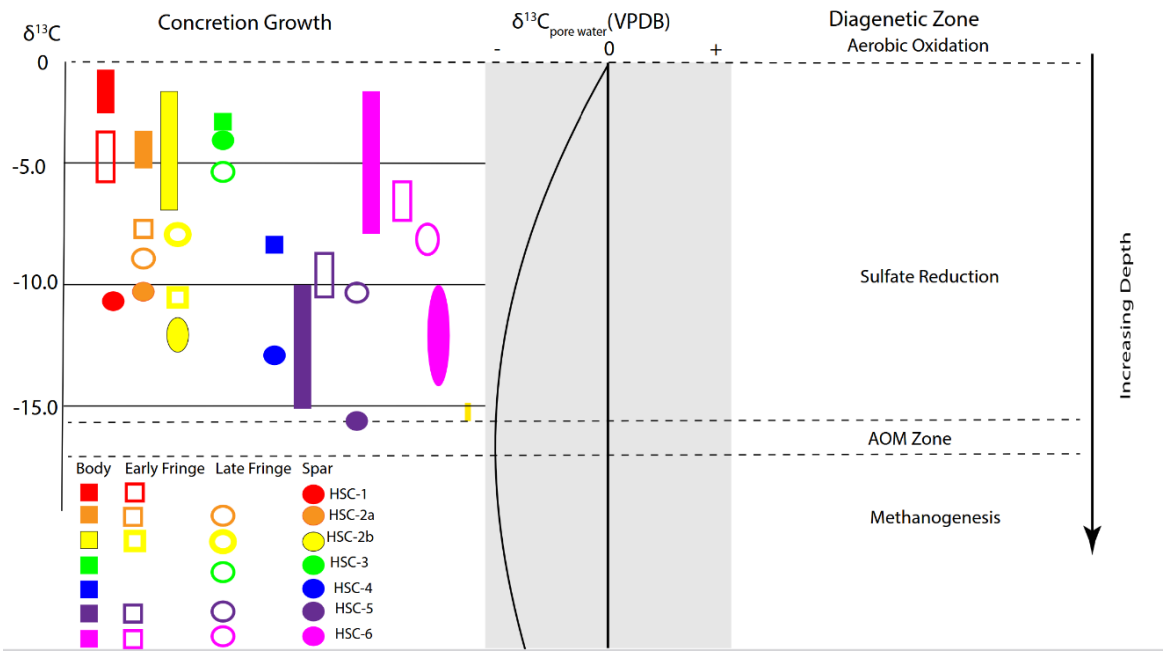


Figure 8: $\delta^{13}C$ ranges for each sample and interpreted diagenetic zone of formation.

Conclusions

The original hypothesis was not supported, as the $\delta^{13}C$ variations across the Holz Shale were not a result of unique formation pathways of each phase. $\delta^{13}C$ values showed overlap across phases. However, F_{org} analysis indicates an increasing contribution of organic matter-derived carbon in later paragenetic phases. This indicates that each successive phase formed at increasing depth, with increasing organic matter contributions. Despite increasing contributions from organic matter and increasing depth of formation, all of the phases within the septarian concretions of the Holz Shale precipitated at shallow depths within the sulfate reduction and perhaps anaerobic oxidation of methane zones.

Acknowledgements

Thank you for all of the teaching, editing, and comments from Dr. Sean Loyd and thank you to all the support and encouragement from my sister, Lauren Minor!

References

Astin, T.R., 1986*,* Septarian crack formation in carbonate concretions from shales and mudstones: *Clays and Clay Minerals, v. 21, p. 617–631.*

Buck, S. P., & Bottje, D. J. (1985). "Continental slope deposits from a late cretaceous, tectonically active margin, Southern California". *SEPM Journal of Sedimentary Research*, *Vol. 55*, 843–853. https://doi.org/10.1306/212f8819-2b24-11d7-8648000102c1865d

Coleman, Max L. "Microbial Processes: Controls on the Shape and Composition of Carbonate Concretions." *Marine Geology*, vol. 113, no. 1-2, 26 May 1993, pp. 127–140., https://doi.org/10.1016/0025-3227(93)90154-n.

Duck, R.W., 1995, Subaqueous shrinkage cracks and early sediment fabrics preserved in Pleistocene calcareous concretions*: Geological Society of London, Journal, v. 152, p. 151–156.*

Hicks, K.S., et al. "Origin of Diagenetic Carbonate Minerals Recovered from the New Jersey Continental Slope." *Proceedings of the Ocean Drilling Program, 150 Scientific Results*, 1996, https://doi.org/10.2973/odp.proc.sr.150.023.1996.

Kelts, K., and J.A. McKenzie. "Diagenetic Dolomite Formation in Quaternary Anoxic Diatomaceous Muds of Deep Sea Drilling Project Leg 64, Gulf of California." *Initial Reports of the Deep Sea Drilling Project, 64*, 1982, https://doi.org/10.2973/dsdp.proc.64.110.1982.

Loyd, S. J., Corsetti, F. A., Eiler, J. M., & Tripati, A. K. (2012). Determining the diagenetic conditions of concretion formation: assessing temperatures and pore waters using clumped isotopes. Journal of Sedimentary Research, 82(12), 1006-1016.

Loyd, S., & Berelson, W. M. (2016). *The modern record of "concretionary" carbonate: Reassessing a discrepancy between modern sediments and the geologic record*. CoLab. https://colab.ws/articles/10.1016%2Fj.chemgeo.2015.11.009

Loyd, Sean J. "Preservation of Overmature, Ancient, Sedimentary Organic Matter in Carbonate Concretions during Outcrop Weathering." *Geobiology*, vol. 15, no. 1, 2017, pp. 146–157., https://pubmed.ncbi.nlm.nih.gov/27384615/.

Loyd, Sean J, and Marissa N Smirnoff. "Progressive Formation of Authigenic Carbonate with Depth in Siliciclastic Marine Sediments Including Substantial Formation in Sediments Experiencing Methanogenesis." *Chemical Geology*, Elsevier, 17 Feb. 2022, https://reader.elsevier.com/reader/sd/pii/S0009254122000699?token=EB7F5D6EE0DB9A74E943483EDAEC2B8A16E0354208E33BC14977D2F50565D29EDCD6DB02BB54F8D877F42B67F00E04B1&originRegion=us-east-1&originCreation=20220407191952.

Loyd, S. J., Meister, P., Liu, B., Nichols, K., Corsetti, F. A., Raiswell, R., Berelson, W., Shields, G., Houndslow, M., Waldron, J. W. F., Westrick-Snapp, B., & Hoffman, J. (2023, May 1). *Temporal evolution of shallow marine diagenetic environments: Insights from carbonate concretions*. Geochimica et Cosmochimica Acta. https://www.sciencedirect.com/science/article/pii/S0016703723001904?via%3Dihub

Luo, Mingming, et al. "Horizontal and Vertical Zoning of Carbonate Dissolution in China." *Geomorphology*, vol. 322, 2018, pp. 66–75., https://doi.org/10.1016/j.geomorph.2018.08.039.

Mozley, Peter S, and Stephen J Burns. "Oxygen and Carbon Isotopic Composition of Marine Carbonate Concretions: An Overview." *Journal of Sedimentary Petrology*, Vol. 63, 1 Jan. 1993, pp. 73–83., https://doi.org/10.1306/d4267a91-2b26-11d7-8648000102c1865d.

Mavromatis, Vasileios, et al. "Formation of Carbonate Concretions in Surface Sediments of Two Mud Mounds, Offshore Costa Rica: A Stable Isotope Study." *International Journal of Earth Sciences*, vol. 103, no. 7, 2012, pp. 1831–1844., https://doi.org/10.1007/s00531-012-0843-7.

Polgári, Márta, et al. "Characterization and 10 Be Content of Iron Carbonate Concretions for Genetic Aspects – Weathering, Desert Varnish or Burning: Rim Effects in Iron Carbonate Concretions." Journal of Environmental Radioactivity, vol. 173, 2017, pp. 58–69., https://doi.org/10.1016/j.jenvrad.2016.11.005.

Raiswell, R., and Q. J. Fisher. "Mudrock-Hosted Carbonate Concretions: A Review of Growth Mechanisms and Their Influence on Chemical and Isotopic Composition." *Journal of the Geological Society*, vol. 157, no. 1, 1 Jan. 2000, pp. 239–251., https://doi.org/10.1144/jgs.157.1.239.

Schultz, J.L, et al. "Tracking Calcium in the San Joaquin Basin, California: A Strontium Isotopic Study of Carbonate Cements at North Coles Levee." *Geochimica Et Cosmochimica Acta*, vol. 53, no. 8, 1989, pp. 1991–1999., https://doi.org/10.1016/0016-7037(89)90319-0.

Thyne, G.D., and Boles, J.R., 1989, Isotopic evidence for origin of the Moeraki septarian concretions, New Zealand: Journal of Sedimentary Petrology, v. 59, p. 272–279.

Torres, Marta E., et al. "Silicate Weathering in Anoxic Marine Sediment as a Requirement for Authigenic Carbonate Burial." *Earth-Science Reviews*, vol. 200, 2020, p. 102960., https://doi.org/10.1016/j.earscirev.2019.102960.

Veizer, Jan, et al. "87Srr 86Sr, d 13C and d 18O evolution of Phanerozoic seawater." Chemical Geology, vol. 161, 12 Oct. 1998.

Yoshida, H., et al. "Fe-Oxide Concretions Formed by Interacting Carbonate and Acidic Waters on Earth and Mars." *Science Advances*, vol. 4, no. 12, 2018, https://doi.org/10.1126/sciadv.aau0872.

Mathematics

Bits and Primes: Exploring Digits of Prime Numbers in Binary

Brianna Castillo
Bobby Orozco
Advisor: Mr. Francisco Zepeda

Abstract

This paper explores the patterns that can be observed in the digits of prime numbers in binary. To analyze these patterns, we looked at subsets of prime numbers in binary. Our research found that patterns in the digits of prime numbers in binary exist for Mersenne Primes, Sexy Pair Primes, Fermat Primes, and Germain Primes. The patterns we observe in these subsets give us a broader understanding of primes. These patterns can give us insight into properties of prime numbers and lead to new conjectures about prime number theory. We conjectured the following: the lengths of Mersenne primes, a property of the relation between Sexy Prime Pairs, the number of zeros in Fermat primes, and the form of Germain Primes. Learning more about prime numbers can enhance theoretical and practical applications in mathematics.

1 Introduction

Patterns of prime numbers in base 10 have been extensively researched. This research has led to a better understanding of the properties and relationships of prime numbers. Previous works such as [3] and [4] have explored prime distribution in base 10 and explain the asymptotic density of primes. These works also cover the various patterns amongst primes such as Sophie Germain primes and Mersenne Primes, both of which are special classes of prime numbers that have specific relationships to other prime numbers. [2] conducted further exploration in the applications of prime numbers in fields such as cryptography. These works also explore sieves and probabilistic methods used to study and generate primes. Although this research has brought forth great discoveries and conjectures about prime numbers, it focuses only on observations made in base 10. Extending our research of primes to other bases, such as binary, can lead to new conjectures about patterns of primes numbers.

We asked the question: *What patterns can be observed in the digits of binary representations of prime numbers?*

Analyzing prime numbers in binary form can lead to new discoveries or conjectures about the properties or distribution of primes. This extends our

current knowledge on prime number properties because the binary representation of primes may reveal patterns not seen in other bases. By observing the patterns of the distribution of primes in binary, we can learn more about how prime gaps behave. The binary representation of prime numbers can simplify examination of certain properties such as their divisibility and parity. This can also help with primality testing, used to determine whether a number is prime, and prime generation which is the process of producing primes. Due to the nature of binary representation, analyzing primes in binary can be beneficial in computational applications. This can allow us to create more compact data and will enhance our understanding of the mathematical properties of prime numbers.

2 Background

Our conjectures focus on the patterns that may occur in the binary representation of prime numbers. To address these, we looked at the digits of different types of primes in binary and observed to see if any patterns occurred. We focused on looking for patterns in some prime numbers rather than patterns for all prime numbers. Analyzing these patterns may help us identify and make note of patterns that occur in all prime numbers.

Let's recall our definition of a prime number defined in [1].

Prime Numbers. *A prime number, p, is a positive integer greater than* 1*, such that the only factors of p are* 1 *and p.*

Note that 1 is not prime. Also, the representation for the number "1" in binary is the same as in base 10, written also as "1."

Now, let's consider other special primes as defined in [1], [5], and [6].

Palindrome Primes. *Palindrome Primes are prime numbers in any base such that the reverse order of that number's digits is also prime.*

Mersenne Primes. *Mersenne Primes are prime numbers written of the form*

$$M_n = 2^n - 1,$$

where n must be a prime number. Note that not all values of n will generate Mersenne Primes.

Sexy Primes. *Sexy Primes are pairs of prime numbers (p, q), such that $q - p =$* 6*.*

Fermat Primes. *Fermat Primes are prime numbers of the form*

$$F_n = 2^{2n} + 1,$$

where n is a non-negative integer.

Germain Primes. *Germain Primes are prime numbers contained in the set of primes,* $\mathbb{P}$*, such that*

$$\mathbb{P}_G = \{p \in \mathbb{P} : p \text{ and } 2p+1 \in \mathbb{P}\}.$$

Twin Prime Pairs. *Twin Prime Pairs are pairs of prime numbers,* $(p, p+2)$*, such that* p *and* $p+2$ *are both prime.*

3 Exploration Findings

In this section, we will present the patterns that we found in the digits for different types of prime numbers in binary. For this paper, a subscript of 2 denotes a number is expressed in binary, and the subscript of 10 denotes a number is expressed in base 10.

3.1 Mersenne Primes in Binary

The first type of primes that we will look at in binary are the Mersenne Primes defined in §2.

3.1.1 Case 1: n = 2

$M_2 = 2^2 - 1 = 3_{10} = 11_2$. Thus, 3_{10} is a Mersenne prime.

3.1.2 Case 2: n = 3

$M_3 = 2^3 - 1 = 7_{10} = 111_2$. Thus, 7_{10} is a Mersenne prime.

3.1.3 Case 3: n = 5

$M_5 = 2^5 - 1 = 31_{10} = 1111_2$. Thus, 31_{10} is a Mersenne prime.

3.1.4 Case 4: n = 7

$M_7 = 2^7 - 1 = 127_{10} = 11111_2$. Thus, 127_{10} is a Mersenne prime.

3.1.5 Case 5: n = 11

$M_{11} = 2^{11} - 1 = 2047_{10} = 111111_2$, which is not prime. Thus, 2047_{10} is <u>not</u> a Mersenne prime.

3.1.6 Case 6: n = 13

$M_{13} = 2^{13} - 1 = 8191_{10} = 1111111_2$. Thus, 8191_{10} is a Mersenne prime.

Theorem 3.1. *Let* p *be a prime number, where* $p \leq 13_{10}$*. Then, the Mersenne Primes will have the binary digits of* 1*'s a number of* p *times.*

3.2 Conjecture for Mersenne Primes

We believe that for all Mersenne Primes in binary will be written as

$$M_n = \underbrace{111\cdots1_2}_{n \text{ times}}.$$

3.3 Sexy Primes in Binary

We now consider a different set of prime numbers called Sexy Primes, defined in §2. Consider Table 1.

Table 1: Sexy Prime Pairs

Decimal	Binary
(5, 11)	(101, 1011)
(7, 13)	(111, 1101)
(11, 17)	(1011, 10001)
(13, 19)	(1101, 10011)
(17, 23)	(10001, 10111)
(31, 37)	(11111, 100101)
(37, 43)	(100101, 101011)
(41, 47)	(101001, 101111)
(47, 53)	(101111, 110101)

3.4 Reversing with Digits

Let's consider the reverse of these prime digits in binary to determine if they are prime themselves and whether they belong to the set of Sexy Prime Pairs. Note that being in the set does not necessarily mean that these numbers are pairs with each other.

- 101_2 reversed is still 101_2, so we're done.
- 1011_2 reversed is 1101_2, which are both prime and are both in the set of Sexy Prime pairs.
- 111_2 reversed is still 111_2, so we're done.
- 1011_2 reversed is 1101_2, which are both prime and are both in the set of Sexy Prime pairs.
- 10001_2 reversed is still 10001_2, so we're done.
- 10011_2 reversed is 11001_2, which are both prime and are both in the set of Sexy Prime pairs.
- 10111_2 reversed is 11101_2, which are both prime and are both in the set of Sexy Prime pairs.

- 101011_2 reversed is 110101_2, which are both prime and are both in the set of Sexy Prime pairs.

We see that these are all Palindrome Primes, which occur when the reverse order of the digits are prime as well as the forward direction, as explained in §2.

3.5 Conjecture for Sexy Prime Pairs

We deduce that every decimal representation of Sexy Prime pairs in binary have a pattern where the reverse order of digits are in the set of Sexy Prime pairs too.

For example, let's consider the decimal Sexy Prime Pair $(17, 23)$. In binary, this is represented as $(10001, 10111)$. If we reverse the order of digits of this binary Sexy Prime pair we have 10001_2 and 11101_2. Both these values are prime and in the set of Sexy Prime pairs.

3.6 Fermat Primes

Let's consider Fermat Primes, defined in §2. The first few Fermat primes are:

$$\begin{aligned}
F_0 &= 2^{2^0} + 1 = 3_{10} = 11_2, \\
F_1 &= 2^{2^1} + 1 = 5_{10} = 101_2, \\
F_2 &= 2^{2^2} + 1 = 17_{10} = 10001_2, \\
F_3 &= 2^{2^3} + 1 = 257_{10} = 100000001_2, \\
F_4 &= 2^{2^4} + 1 = 65537_{10} = 10000000000000001_2.
\end{aligned}$$

These are all the known Fermat primes. These numbers are also palindromic primes. Notably, each Fermat prime begins and ends with a "1," with varying numbers of zeros in between. Specifically, we observe the following number of zeros:

- For $n = 0$: 0 zeros
- For $n = 1$: 1 zero
- For $n = 2$: 3 zeros
- For $n = 3$: 7 zeros
- For $n = 4$: 15 zeros

3.7 Conjecture for Fermat Primes

We conjecture that the number of zeros between the ones in the binary representation is given by the formula

$$2^n - 1.$$

Beyond $n = 4$, it has been shown that F_n is composite for $n = 5, 6, 7, 8$, and higher values, so we do not consider these although it remains an open question whether any additional Fermat primes exist for larger values of n.

3.8 Germain Primes in Binary

We will look at a specific subset of prime numbers called Germain Primes, defined in §2.

3.8.1 First Sequence Case

Consider the sequence starting at 2_{10} with the terms growing by a recursive formula $a_{i+1} = 2a_i + 1$. We get the sequence (in base 10): $2, 5, 11, 23, 47, 95, \ldots$ However, $95 \notin \mathbb{P}$, which means that $47 \notin \mathbb{P}_G$. Thus, this first sequence is finite. In particular, the sequence (in decimal) is

$$2_{10}, 5_{10}, 11_{10}, 23_{10},$$

which converts to

$$10_2, 101_2, 1011_2, 10111_2.$$

Now, let's consider more cases.

3.8.2 Second Sequence Case

Similarly, consider the sequence starting at 3_{10} with terms growing by the same recursive formula, $a_{i+1} = 2a_i + 1$. So the sequence we have is

$$3_{10}, 7_{10}, 15_{10}, \ldots$$

Since $15_{10} \notin \mathbb{P}$, then $7_{10} \notin \mathbb{P}_G$. Therefore, this sequence is 3_{10}, which is finite. Note that this sequence case, in binary, is 11_2.

3.8.3 Third Sequence Case

Next, let's consider the prime 29_{10}, which is Germain. Now consider the sequence starting at 29_{10} with the same recursive formula, $a_{i+1} = 2a_i + 1$. So the sequence is

$$29_{10}, 59_{10}, 119_{10}, \ldots$$

Since $119_{10} \notin \mathbb{P}$, then $59_{10} \notin \mathbb{P}_{\mathbb{G}}$. Therefore, the sequence is 29_{10}, which is finite. Note that this sequence case, in binary, is 11101_2.

3.8.4 Fourth Sequence Case

Finally, consider the prime 41_{10}, which is also Germain. Using the same recursive formula, $a_{i+1} = 2a_i + 1$, the sequence is

$$41_{10}, 83_{10}, 167_{10}, 335_{10}, \ldots$$

Since $335_{10} \notin \mathbb{P}$, then $167_{10} \notin \mathbb{P}_{\mathbb{G}}$. Therefore, the finite sequence is $41_{10}, 83_{10}$, which in binary is $101001_2, 1010011_2$.

Theorem 3.2. *Let p be a prime number in base 10 such that $p \leq 83_{10}$. Let p_g be the primes numbers p that are Germain Primes. These Germain Primes have the binary representation where the consecutive terms in the sequence will add a digit of "1" at the end of the number.*

3.9 Conjecture for Germain Primes

In binary, we observe that primes add a "1" to the sequence each time. This occurs because multiplying a prime number p by 2 shifts it binary representation left, adding a "0" at the end. When we add "1", we convert that 0 to a 1. This works similarly to decimal: when we multiply by ten, we add a zero in the ones place.

3.10 Twin Prime Pairs

Now, we will consider Twin Prime Pairs, defined in §2. Consider Table 2 of such pairs in decimal and binary. We don't see any distinct patterns in these pairs, but we encourage researchers to further analyze patterns among binary representation of prime numbers.

Table 2: Twin Prime Pairs

Decimal	Binary
(3, 5)	(11, 101)
(5, 7)	(101, 111)
(11, 13)	(1011, 1101)
(17, 19)	(10001, 10011)
(29, 31)	(11101, 11111)
(41, 43)	(101001, 101011)
(59, 61)	(111011, 111101)
(71, 73)	(1000111, 1001001)

4 Conclusion

We considered what patterns can be observed in the digits of binary representations of prime numbers. We explored various prime numbers in binary to identify digit patterns. By examining smaller subsets, we aimed for a broader understanding of primes. We conjectured that Mersenne Primes are always a string of 1s that are of length n, where n is the term number. We also conjectured that Sexy Prime pairs and their reversed digits are also in the set of Sexy Prime pairs. We conjectured Fermat Primes have $2^n - 1$ zeros between the 1s at the ends of their binary representations. Finally, we conjectured that consecutive Germain Primes add a "1" at the end of the proceeding Germain Prime in the sequence. We did not state a conjecture for Twin Prime Pairs.

Observing the digits of prime numbers in binary and in any base can help us create theorems or conjectures about the digits of prime numbers and their properties. Mathematicians have been trying to understand prime numbers and patterns that are hidden within prime numbers for centuries.

Areas for future research include Fermat Primes and Twin Prime Pairs (§3.6 and §3.10). There are only five Fermat numbers known to be prime for $n = 0, 1, 2, 3, 4$. Other Fermat numbers have been found, but are composite (i.e., not prime). Twin Prime Pairs are directly connected to the Twin Primes Conjecture, which states that there are an infinite number of pairs of Twin Primes; this remains an open question. Not only can this conjecture be an area of continued research, but possible patterns could emerge in the binary representation of Twin Prime Pairs.

References

[1] D.M. Burton. Elementary Number Theory. Mcgraw-Hill, seventh ed., 2011.

[2] J.W. Porras Ferreira, Study about the pattern of prime numbers. *Recent Advances in Mathematical Research and Computer Science*, **9**, 83-93, 2022.

[3] R.B. Ghandi, A.M. Patel, M. Patel, Prime numbers and their analysis, *International Journal of Emerging Technologies and Innovative Research*, **7**, 466-470, 2020.

[4] A. Granville, Prime number patterns, *The American Mathematical Monthly*, **115**, 279-296, 2008.

[5] H. Ibstedt, Palindrome studies (Part I) The palindrome concept and its applications to prime numbers, *Scientia Magna*, **2**, 101-116, 2006.

[6] M. Ndiaye, Origin of Sexy Prime Numbers, Origin of cousin prime numbers, equations from supposedly prime numbers, origin of the Mersenne number, origin of the Fermat number. *Advances in Pure Mathematics*, **14**, no. 5, 321–332, 2024.

Methods for Finding the Second Moment of Insurance Payment

Aaron Kim
Department of Mathematics
California State University
Fullerton CA 92831, USA
aaronskim@fullerton.edu

Justin Nguyen
Department of Mathematics
California State University
Fullerton CA 92831, USA
20jnguyen39@csu.fullerton.edu

May 1, 2025

Abstract: While the formula for evaluating the first moment of insurance payment are widely available, the formula for evaluating its the second moment and hence variance are quite limited. Within this project, we are investigating several methods for this purpose, We then apply these methods when loss follows either a uniform, exponential, gamma, or Pareto distribution.

1 Introduction

Let X denote the loss amount, d represent deductible and Y^L be the insurance payment when a loss occurs. Thus

$$Y_L = \begin{cases} 0 & X \leq d \\ X - d & X > d. \end{cases}$$

Mathematically, we can write $Y_L = \max(0, X - d)$. This random variable is also known as (insurance) *payment per loss*. It is straight forward to see that $Y_L = X - X \wedge d$. In this formulation, the first moment of Y_L can be readily obtained. In fact, $E(Y_L) = E(X) - E(X \wedge d)$. The formula for $E(X \wedge d)$ are widely available when X follows either a uniform, exponential or pareto distribution. See, for example, "Loss Models" by Panjer and et el. This book however does not report such formula when X follows a gamma distribution. We will present the formula in this case along with our main results in calculating the second moment of Y_L.

2 Second Moment of Y_L - First Method

Let u be the limit for the insurance policy, then

$$Y_L = \begin{cases} 0 & X \leq d \\ X - d & d < X \leq u \\ u - d & X > u. \end{cases}.$$

Hence, $Y_L = (X \wedge u) - (X \wedge d)$ and

$$\begin{aligned} Y_L^2 &= [(X \wedge u) - (X \wedge d)]^2 \\ &= (X \wedge u)^2 - 2(X \wedge u)(X \wedge d) + (X \wedge d)^2 \\ &= (X \wedge u)^2 + (X \wedge d)^2 - 2(X \wedge d)[X \wedge u - X \wedge d] - 2(X \wedge d)^2 \\ &= (X \wedge u)^2 - (X \wedge d)^2 - 2(X \wedge d)[X \wedge u - X \wedge d]. \end{aligned}$$

Since we are only considering the case where the insurance policies do not have a limit, $u = \infty$ and

$$\begin{aligned} 2(X \wedge d)(X \wedge u - X \wedge d) &= 2(X \wedge d)(X - X \wedge d) \\ &= 2X(0) + 2d(X - d) \\ &= 2d(X - d) \\ &= 2dY_L. \end{aligned}$$

Therefore, we have

$$E(Y_L^2) = E(X^2) - E[(X \wedge d)^2] - 2dE(Y_L).$$

Mathematically, we can also write

$$\begin{aligned} Y_L^2 &= (X - X \wedge d)^2 \\ &= X^2 - 2X(X \wedge d) + (X \wedge d)^2. \end{aligned}$$

Therefore,

$$\begin{aligned} E(Y_L^2) &= E[(X - X \wedge d)^2] \\ &= E(X^2) - 2E[X(X \wedge d)] + E[(X \wedge d)^2]. \end{aligned}$$

2.1 Uniform Distribution

For a loss modeled by the uniform distribution $X \sim unif(0, b)$, its density function will be $f(x) = \frac{1}{b}$. Thus,

$$\begin{aligned} E(X^2) &= \int_0^b x^2 \frac{1}{b} dx \\ &= \frac{1}{b} \frac{x^3}{3} |_0^b \\ &= \frac{1}{3b} b^3 \\ &= \frac{b^2}{3}. \end{aligned}$$

$$\begin{aligned} E[X(X \wedge d)] &= \int_0^b x(x \wedge d) \frac{1}{b} dx \\ &= \int_0^d x^2 \frac{1}{b} dx + \int_d^b x d \frac{1}{b} dx \\ &= \frac{1}{b} [\frac{x^3}{3} |_0^d + d \frac{x^2}{2} |_d^b] \\ &= \frac{1}{b} [\frac{d^3}{3} + \frac{d(b^2 - d^2)}{2}]. \end{aligned}$$

$$
\begin{aligned}
E[(X \wedge d)^2] &= \int_0^b (x \wedge d)^2 \frac{1}{b} dx \\
&= \int_0^d \frac{x^2}{b} dx + \int_d^b \frac{d^2}{b} dx \\
&= \frac{1}{b}[\frac{d^3}{3} + d^2(b-d)] \\
&= \frac{1}{b}\frac{d^3 + 3d^2(b-d)}{3} \\
&= \frac{3bd^2 - 2d^3}{3b}.
\end{aligned}
$$

With these values calculated, we can now determine $E(Y_L^2)$.

$$
\begin{aligned}
E(Y_L^2) &= E(X^2) - 2E[X(X \wedge d)] + E[(X \wedge d)^2] \\
&= \frac{b^2}{3} - \frac{2}{b}[\frac{d^3}{3} + \frac{d(b^2 - d^2)}{2}] + \frac{3bd^2 - 2d^3}{3b} \\
&= \frac{b^2}{3} - \frac{1}{b}(\frac{2d^3}{3} + b^2 d - d^3) + \frac{3bd^2 - 2d^3}{3b} \\
&= \frac{b^2}{3} - \frac{2d^3 + 3b^2 - 3d^3}{3b} + \frac{3bd^2 - 2d^3}{3b} \\
&= \frac{1}{3b}(b^3 - 2d^3 - 3b^2 d + 3d^3 + 3bd^2 - 2d^3) \\
&= \frac{1}{3b}(b^3 - 3b^2 d + 3bd^2 - d^3) \\
&= \frac{1}{3b}(b-d)^3.
\end{aligned}
$$

2.2 Exponential Distribution

$X \sim exp(\theta)$, $f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$

$$
\begin{aligned}
E(X^2) &= \int_0^\infty x^2 \frac{1}{\theta} e^{-\frac{x}{\theta}} dx \\
&= -x^2 e^{-\frac{x}{\theta}} - 2x\theta e^{-\frac{x}{\theta}} - 2\theta^2 e^{-\frac{x}{\theta}} |_0^\infty \\
&= 0 - (-2\theta) = 2\theta.
\end{aligned}
$$

$$
\begin{aligned}
E[(X \wedge d)^2] &= \int_0^d x^2 \frac{1}{\theta} e^{-\frac{x}{\theta}} + \int_d^\infty d^2 \frac{1}{\theta} e^{-\frac{x}{\theta}} dx \\
&= I_1 + I_2 \\
&= 2\theta^2 - 2\theta d e^{-\frac{d}{\theta}} - 2\theta^2 e^{-\frac{d}{\theta}}
\end{aligned}
$$

$$
\begin{aligned}
I_1 &= -x^2 e^{-\frac{x}{\theta}} - 2x\theta e^{-\frac{x}{\theta}} - 2\theta^2 e^{-\frac{x}{\theta}} |_0^d \\
&= -d^2 e^{-\frac{d}{\theta}} - 2d\theta e^{-\frac{d}{\theta}} - 2\theta^2 e^{-\frac{d}{\theta}} + 2\theta^2
\end{aligned}
$$

$$
I_2 = -d^2 e^{-\frac{x}{\theta}} |_d^\infty = d^2 e^{-\frac{d}{\theta}}
$$

We know that the expected value of the first moment is $E(Y_L) = \theta e^{-\frac{d}{\theta}}$. Therefore,

$$
\begin{aligned}
E(Y_L^2) &= E(X^2) - E[(X \wedge d)^2] - 2dE(Y_L) \\
&= 2\theta^2 - (2\theta^2 - 2\theta d e^{-\frac{d}{\theta}} - 2\theta^2 e^{-\frac{d}{\theta}}) - 2\theta d e^{-\frac{d}{\theta}} \\
&= 2\theta^2 e^{-\frac{d}{\theta}}.
\end{aligned}
$$

2.3 Pareto Distribution

$X \sim Pareto(\alpha, \theta)$, $f(x) = \frac{\alpha\theta^\alpha}{(\theta+x)^{\alpha+1}}$

$$
\begin{aligned}
E(X^2) &= \int_0^\infty x^2 \frac{\alpha\theta^\alpha}{(\theta+x)^{\alpha+1}} dx \\
&= \alpha\theta^\alpha \int_\theta^\infty \frac{u^2 - 2\theta u + \theta^2}{u^{\alpha+1}} du \\
&= \alpha\theta^\alpha [\int_\theta^\infty u^{1-\alpha} du - 2\theta \int_\theta^\infty u^{-\alpha} du + \theta^2 \int_\theta^\infty u^{-\alpha-1} du] \\
&= \alpha\theta^\alpha [-\frac{u^{2-\alpha}}{\alpha-2}|_\theta^\infty + 2\theta \frac{u^{1-\alpha}}{\alpha-1}|_\theta^\infty - \theta^2 \frac{u^{-\alpha}}{\alpha}|_\theta^\infty] \\
&= \alpha\theta^\alpha [\frac{\theta^{2-\alpha}}{\alpha-2} - \frac{2\theta^{2-\alpha}}{\alpha-1} + \frac{\theta^{2-\alpha}}{\alpha}] \\
&= \theta^\alpha [\frac{2\theta^{2-\alpha}}{(\alpha-2)(\alpha-1)}] \\
&= \frac{2\theta^2}{(\alpha-2)(\alpha-1)}.
\end{aligned}
$$

$$
\begin{aligned}
E[(X \wedge d)^2] &= \int_0^d x^2 \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx + \int_d^\infty d^2 \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx \\
&= I_1 + I_2.
\end{aligned}
$$

$$
\begin{aligned}
I_1 &= \alpha\theta^\alpha \int_0^d \frac{x^2}{(x+\theta)^{\alpha+1}} dx \\
&= \alpha\theta^\alpha \int_\theta^{d+\theta} \frac{u^2 - 2\theta u + \theta^2}{u^{\alpha+1}} du \\
&= \alpha\theta^\alpha [\int_\theta^{d+\theta} u^{1-\alpha} du - 2\theta \int_\theta^{d+\theta} u^{-\alpha} du + \theta^2 \int_\theta^{d+\theta} u^{-\alpha-1} du] \\
&= \alpha\theta^\alpha [-\frac{1}{\alpha-2} u^{2-\alpha}|_\theta^{d+\theta} + \frac{2\theta}{\alpha-1} u^{1-\alpha}|_\theta^{d+\theta} - \frac{\theta^2}{\alpha} u^{-\alpha}|_\theta^{d+\theta}] \\
&= \alpha\theta^\alpha [\frac{\theta^{2-\alpha} - (d+\theta)^{2-\alpha}}{\alpha-2} + \frac{2\theta(d+\theta)^{1-\alpha} - 2\theta^{2-\alpha}}{\alpha-1} + \frac{\theta^{2-\alpha} - \theta^2(d+\theta)^{-\alpha}}{\alpha}] \\
&= \frac{2\theta^2}{(\alpha-2)(\alpha-1)} + \theta^\alpha [\frac{\alpha d^2 - 2\alpha\theta d - \alpha^2 d^2 - 2\theta^2}{(\alpha-2)(\alpha-1)(d+\theta)^\alpha}].
\end{aligned}
$$

$$
\begin{aligned}
I_2 &= d^2 \int_d^\infty \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx \\
&= d^2 S_X(d) \\
&= d^2 (\frac{\theta}{d+\theta})^\alpha \\
&= \theta^\alpha [\frac{\alpha^2 d^2 - 3\alpha d^2 + 2d^2}{(\alpha-2)(\alpha-1)(d+\theta)^\alpha}].
\end{aligned}
$$

$$
\begin{aligned}
E[(X \wedge d)^2] &= I_1 + I_2 \\
&= \frac{2\theta^2}{(\alpha-2)(\alpha-1)} + 2\theta^\alpha [\frac{d^2 - \alpha d^2 - \alpha\theta d - \theta^2}{(\alpha-2)(\alpha-1)(d+\theta)^\alpha}].
\end{aligned}
$$

$$
\begin{aligned}
E(Y_L) &= E(X) - E(X \wedge d) \\
&= \frac{\theta}{\alpha - 1} - \frac{\theta}{\alpha - 1}[1 - (\frac{\theta}{d+\theta})^{\alpha-1}] \\
&= \frac{\theta}{\alpha - 1}(\frac{\theta}{d+\theta})^{\alpha-1} \\
&= \theta^{\alpha}[\frac{d+\theta}{(\alpha-1)(d+\theta)^{\alpha}}] \\
&= \theta^{\alpha}[\frac{\alpha d - 2d + \alpha\theta - 2\theta}{(\alpha-2)(\alpha-1)(d+\theta)^{\alpha}}].
\end{aligned}
$$

$$
\begin{aligned}
E(Y_L^2) &= E(X^2) - E[(X \wedge d)^2] - 2dE(Y_L) \\
&= 2\theta^{\alpha}[\frac{(d+\theta)^2}{(\alpha-2)(\alpha-1)(d+\theta)^{\alpha}}] \\
&= \frac{2\theta^2}{(\alpha-2)(\alpha-1)}(\frac{\theta}{d+\theta})^{\alpha-2}.
\end{aligned}
$$

2.4 Gamma Distribution

If $X \sim gam(\alpha, \theta)$, then $f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}$.

$$
\begin{aligned}
E(X^2) &= \int_0^{\infty} x^2 \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)} dx \\
&= \theta^2\alpha(\alpha+1)\int_0^{\infty} \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+2}\Gamma(\alpha+2)} dx \\
&= \theta^2\alpha(\alpha+1).
\end{aligned}
$$

$E[(X \wedge d)^2] = \int_0^d x^2 f(x)dx + \int_d^{\infty} d^2 f(x)dx = I_1 + I_2.$

$$
\begin{aligned}
I_1 &= \int_0^d x^2 \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha \Gamma(\alpha)} dx \\
&= \int_0^d \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^\alpha \Gamma(\alpha)} dx \\
&= \frac{\theta^{\alpha+2}\Gamma(\alpha+2)}{\theta^\alpha \Gamma(\alpha)} \int_0^d \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+2}\Gamma(\alpha+2)} dx \\
&= \theta^2\alpha(\alpha+1) \int_0^d \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+2}\Gamma(\alpha+2)} dx \\
&= \theta^2\alpha(\alpha+1)P(X^* \leq d) \\
&= \theta^2\alpha(\alpha+1)F_{X^*}(d) \\
&= \theta^2\alpha(\alpha+1)[1 - P(Y \leq \alpha+1)] \\
&= \theta^2\alpha(\alpha+1) - \theta^2\alpha(\alpha+1)P(Y \leq \alpha+1).
\end{aligned}
$$

Note that here, $X^* \sim gam(\alpha+2, \theta)$, and $Y \sim Poi(\lambda = \frac{d}{\theta})$.

$$
\begin{aligned}
I_2 &= d^2 \int_d^\infty f(x) dx \\
&= d^2 S_X(d) \\
&= d^2 P(Y \leq \alpha - 1).
\end{aligned}
$$

Note that here, $S_X(d)$ is the survival function of the function X beyond the value of the deductible d. Thus,

$$
\begin{aligned}
E[(X \wedge d)^2] &= I_1 + I_2 \\
&= \theta^2\alpha(\alpha+1) - \theta^2\alpha(\alpha+1)P(Y \leq \alpha+1) + d^2 P(Y \leq \alpha - 1).
\end{aligned}
$$

$$
\begin{aligned}
E(X) &= \int_0^\infty x \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha \Gamma(\alpha)} \\
&= \frac{\theta^{\alpha+1}\Gamma(\alpha+1)}{\theta^\alpha \Gamma(\alpha)} \int_0^\infty \frac{x^{(\alpha+1)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1}\Gamma(\alpha+1)} \\
&= \theta\alpha.
\end{aligned}
$$

$$\begin{aligned}
E(Y_L) &= E(X) - E(X \wedge d) \\
&= \theta\alpha - [\alpha\theta[1 - P(Y \leq \alpha)] + dP(Y \leq \alpha - 1)] \\
&= \theta\alpha - \alpha\theta[1 - P(Y \leq \alpha)] - dP(Y \leq \alpha - 1) \\
&= \theta\alpha P(Y \leq \alpha) - dP(Y \leq \alpha - 1).
\end{aligned}$$

$$\begin{aligned}
E(Y_L^2) &= E(X^2) - E[(X \wedge d)^2] - 2dE(Y_L) \\
&= \theta^2\alpha(\alpha+1)P(Y \leq \alpha + 1) - 2d\theta\alpha P(Y \leq \alpha) + d^2P(Y \leq \alpha - 1).
\end{aligned}$$

3 Second Moment of Y_L - Second Method

Let $I_{(X>d)}$ denote a dummy variable that is 1 when the condition $X > d$ is met and 0 otherwise. We can see that

$$(X-d)^2 I_{(X>d)} = \begin{cases} 0 & X \leq d \\ (X-d)^2 & X > d \end{cases} = Y_L^2.$$

Thus $E(Y_L^2) = \int_{\mathcal{D}} I_{(X>d)}(x-d)^2 f(x)dx$.

3.1 Uniform Distribution

$X \sim unif(0, b)$, $f(x) = \frac{1}{b}$

$$\begin{aligned}
E(Y_L^2) &= \int_0^b (x-d)^2 I_{(X>d)} \frac{1}{b} dx \\
&= \frac{1}{b}\int_d^b (x-d)^2 dx \\
&= \frac{1}{b}\int_0^{b-d} u^2 du \\
&= \frac{1}{b}\frac{1}{3}u^3|_0^{b-d} \\
&= \frac{1}{3b}(b-d)^3.
\end{aligned}$$

3.2 Exponential Distribution

$X \sim exp(\theta)$, $f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}$

$$\begin{aligned}
E(Y_L^2) &= \int_0^\infty (x-d)^2 I_{(X>d)} f(x) dx \\
&= \int_d^\infty (x-d)^2 \frac{1}{\theta} e^{-\frac{x}{\theta}} dx \\
&= \int_0^d y^2 \frac{1}{\theta} e^{-\frac{(y+d)}{\theta}} dy \\
&= e^{-\frac{d}{\theta}} \int_0^\infty \frac{1}{\theta} y^2 e^{-\frac{y}{\theta}} dy \\
&= e^{-\frac{d}{\theta}} (-y^2 e^{-\frac{y}{\theta}} - 2y\theta e^{-\frac{y}{\theta}} - 2\theta^2 e^{-\frac{y}{\theta}} |_0^\infty) \\
&= e^{-\frac{d}{\theta}} (0 + 2\theta^2) \\
&= 2\theta^2 e^{-\frac{d}{\theta}}.
\end{aligned}$$

3.3 Pareto Distribution

$X \sim Pareto(\alpha, \theta)$, $f(x) = \frac{\alpha\theta^\alpha}{(\theta+x)^{\alpha+1}}$

$$
\begin{aligned}
E(Y_L^2) &= \int_0^\infty (x-d)^2 I_{(x>d)} \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} dx \\
&= \alpha\theta^\alpha \int_d^\infty \frac{(x-d)^2}{(x+\theta)^{\alpha+1}} dx \\
&= \alpha\theta^\alpha \int_{d+\theta}^\infty \frac{(u-\theta-d)^2}{u^{\alpha+1}} du \\
&= \alpha\theta^\alpha [\int_{d+\theta}^\infty u^{1-\alpha} du - 2(d+\theta)\int_{d+\theta}^\infty u^{-\alpha} du + (d+\theta)^2 \int_{d+\theta}^\infty u^{-1-\alpha} du] \\
&= \alpha\theta^\alpha [\frac{1}{2-\alpha} u^{2-\alpha}|_{d+\theta}^\infty - \frac{2(d+\theta)}{1-\alpha} u^{1-\alpha}|_{d+\theta}^\infty - \frac{(d+\theta)^2}{\alpha} u^{-\alpha}|_{d+\theta}^\infty] \\
&= \alpha\theta^\alpha [\frac{(d+\theta)^{2-\alpha}}{(\alpha-2)} - \frac{2(d+\theta)^{2-\alpha}}{(\alpha-1)} + \frac{(d+\theta)^{2-\alpha}}{\alpha}] \\
&= \alpha\theta^\alpha [\frac{2(d+\theta)^{2-\alpha}}{(\alpha-2)(\alpha-1)\alpha}] \\
&= \theta^2 [\frac{2\theta^{\alpha-2}}{(\alpha-2)(\alpha-1)(d+\theta)^{\alpha-2}}] \\
&= \frac{2\theta^2}{(\alpha-2)(\alpha-1)} (\frac{\theta}{d+\theta})^{\alpha-2}.
\end{aligned}
$$

3.4 Gamma Distribution

$X \sim Gamma(\alpha, \theta)$, $f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha \Gamma(\alpha)}$

$$
\begin{aligned}
E(Y_L^2) &= \int_0^\infty (x-d)^2 I_{(x>d)} \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)} dx \\
&= \int_d^\infty (x-d)^2 \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)} dx \\
&= \int_d^\infty \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)} dx - 2d\int_d^\infty \frac{x^{(\alpha+1)-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)} dx + d^2 \int_d^\infty \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^\alpha\Gamma(\alpha)} dx \\
&= I_1 - 2dI_2 + d^2 I_3.
\end{aligned}
$$

$$
\begin{aligned}
I_1 &= \int_d^{\infty} \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}dx \\
&= \frac{\theta^{\alpha+2}\Gamma(\alpha+2)}{\theta^{\alpha}\Gamma(\alpha)}\int_d^{\infty} \frac{x^{(\alpha+2)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+2}\Gamma(\alpha+2)}dx \\
&= \alpha(\alpha+1)\theta^2 S_{X^*}(d), X^* \sim Gamma(\alpha^* = \alpha+2, \theta^* = \theta) \\
&= \alpha(\alpha+1)\theta^2 F_Y(\alpha+1), Y \sim Poisson(\lambda = \frac{d}{\theta}) \\
&= \alpha(\alpha+1)\theta^2 \Sigma_{k=0}^{\alpha+1} P(Y=k) \\
&= \alpha(\alpha+1)\theta^2 P(Y \leq \alpha+1).
\end{aligned}
$$

$$
\begin{aligned}
I_2 &= \int_d^{\infty} \frac{x^{(\alpha+1)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}dx \\
&= \frac{\theta^{\alpha+1}\Gamma(\alpha+1)}{\theta^{\alpha}\Gamma(\alpha)}\int_d^{\infty} \frac{x^{(\alpha+1)-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha+1}\Gamma(\alpha+1)}dx \\
&= \alpha\theta S_{X^{**}}(d), X^{**} \sim Gamma(\alpha^{**} = \alpha+1, \theta^{**} = \theta) \\
&= \alpha\theta F_Y(\alpha), Y \sim Poisson(\lambda = \frac{d}{\theta}) \\
&= \alpha\theta \Sigma_{k=0}^{\alpha} P(Y=k) \\
&= \alpha\theta P(Y \leq \alpha).
\end{aligned}
$$

$$
\begin{aligned}
I_3 &= \int_d^{\infty} \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}dx \\
&= S_X(d), X \sim Gamma(\alpha, \theta) \\
&= F_Y(\alpha-1), Y \sim Poisson(\lambda = \frac{d}{\theta}) \\
&= \Sigma_{k=0}^{\alpha-1} P(Y=k) \\
&= P(Y \leq \alpha-1).
\end{aligned}
$$

$$
E(Y_L^2) = \alpha(\alpha+1)\theta^2 P(Y \leq \alpha+1) - 2d\alpha\theta P(Y \leq \alpha) + d^2 P(Y \leq \alpha-1).
$$

4 Second Moment of Y_L - Third Method

Let $E(Y_P^2)$ be the second moment of insurance payment, given that the loss exceeds the deductible. Thus, $E(Y_P^2) = E(Y_L^2|X > d)$. In addition to this, by the law of total expectation we can see that

$$\begin{aligned} E(Y_L^2) &= E(Y_L^2|X \leq d)P(X \leq d) + E(Y_L^2|X > d)P(X > d) \\ &= E[(X-d)^2|X > d]P(X > d) \\ &= E(Y_P^2)P(X > d). \end{aligned}$$

Additionally, we can find the CDF of Y_P

$$\begin{aligned} F_{Y_P}(y) &= P(Y_P \leq y) \\ &= P(X - d \leq y|X > d) \\ &= P(X \leq y + d|X > d) \\ &= \frac{P(X \leq y + d \cap X > d)}{P(X > d)} \\ &= \frac{P(d \leq X \leq y + d)}{P(X > d)} \\ &= \frac{F_X(y+d) - F_X(d)}{S_X(d)}. \end{aligned}$$

If X is a random variable where $X \sim unif(0, b)$, then we can see

$$\begin{aligned} F_{Y_P}(y) &= \frac{F_X(y+d) - F_X(d)}{S_X(d)} \\ &= \frac{\frac{y+d}{b} - \frac{d}{b}}{1 - \frac{d}{b}} \\ &= \frac{\frac{y}{b}}{\frac{b-d}{b}} \\ &= \frac{y}{b-d}. \\ &\therefore Y_P \sim unif(0, b-d) \end{aligned}$$

If X is a random variable where $X \sim exp(\theta)$, then we can see

$$\begin{aligned} F_{Y_P}(y) &= \frac{F_X(y+d) - F_X(d)}{S_X(d)} \\ &= \frac{1 - e^{-\frac{y+d}{\theta}} - 1 + e^{-\frac{d}{\theta}}}{e^{-\frac{d}{\theta}}} \\ &= \frac{e^{-\frac{d}{\theta}} - e^{-\frac{y+d}{\theta}}}{e^{-\frac{d}{\theta}}} \\ &= \frac{e^{-\frac{d}{\theta}}(1 - e^{-\frac{y}{\theta}})}{e^{-\frac{d}{\theta}}} \\ &= 1 - e^{-\frac{y}{\theta}}. \\ &\therefore Y_P \sim exp(\theta) \end{aligned}$$

If X is a random variable where $X \sim pareto(\alpha, \theta)$, then we can see

$$\begin{aligned} F_{Y_P}(y) &= \frac{F_X(y+d) - F_X(d)}{S_X(d)} \\ &= \frac{1 - (\frac{\theta}{\theta+y+d})^\alpha - 1 + (\frac{\theta}{\theta+d})^\alpha}{(\frac{\theta}{\theta+d})^\alpha} \\ &= \frac{(\frac{\theta}{\theta+d})^\alpha - (\frac{\theta}{\theta+y+d})^\alpha}{(\frac{\theta}{\theta+d})^\alpha} \\ &= 1 - (\frac{\theta+d}{\theta+d+y})^\alpha. \\ &\therefore Y_P \sim pareto(\alpha, \theta+d) \end{aligned}$$

4.1 Uniform Distribution

If $X \sim unif(0, b)$, then $Y_P \sim unif(0, b-d)$.

$$E(Y_L^2) = E(Y_P^2)P(X > d).$$

$$\begin{aligned} E(Y_p^2) &= \int_0^{b-d} y^2 \frac{1}{b-d} dy \\ &= \frac{1}{b-d}\frac{1}{3} y^3 |_0^{b-d} \\ &= \frac{(b-d)^2}{3}. \end{aligned}$$

$$\begin{aligned} P(X > d) &= \int_d^b \frac{1}{b} dx \\ &= \frac{b-d}{b}. \end{aligned}$$

$$\begin{aligned} E(Y_L^2) &= E(Y_p^2) P(X > d) \\ &= \frac{(b-d)^2}{3}\frac{b-d}{b} \\ &= \frac{1}{3b}(b-d)^3. \end{aligned}$$

4.2 Exponential Distribution

If $X \sim exp(\theta)$, then $Y_p \sim exp(\theta)$ as well, as seen above. Thus,

$$E(Y_L^2) = E(Y_P^2) P(X > d).$$

$$\begin{aligned} E(Y_p^2) &= \int_0^{\infty} y^2 \frac{1}{\theta} e^{-\frac{y}{\theta}} dy \\ &= -y^2 e^{-\frac{y}{\theta}} - 2y\theta e^{-\frac{y}{\theta}} 2\theta^2 e^{-\frac{y}{\theta}} |_0^{\infty} \\ &= 2\theta^2. \end{aligned}$$

$$\begin{aligned} P(X > d) &= \int_d^{\infty} \frac{1}{\theta} e^{-\frac{x}{\theta}} dx \\ &= -e^{\frac{x}{\theta}} |_d^{\infty} \\ &= e^{-\frac{d}{\theta}}. \end{aligned}$$

$$\begin{aligned} E(Y_L^2) &= E(Y_p^2)P(X > d) \\ &= 2\theta^2 e^{-\frac{d}{\theta}}. \end{aligned}$$

4.3 Pareto Distribution

If $X \sim Pareto(\alpha, \theta)$, then $Y_p \sim Pareto(\alpha, \theta + d)$.

$$E(Y_L^2) = E(Y_P^2)P(X > d)$$

$$\begin{aligned} E(Y_P^2) &= V(Y_P) + E^2(Y_P) \\ &= (\frac{\theta + d}{\alpha - 1})^2(\frac{\alpha}{\alpha - 2}) + (\frac{\theta + d}{\alpha - 1})^2 \\ &= (\frac{\alpha}{\alpha - 2} + 1)(\frac{\theta + d}{\alpha - 1})^2 \\ &= \frac{2(\alpha - 1)}{\alpha - 2}(\frac{\theta + d}{\alpha - 1})^2 \\ &= \frac{2(\theta + d)^2}{(\alpha - 2)(\alpha - 1)}. \end{aligned}$$

$$\begin{aligned} P(X > d) &= S_X(d) \\ &= (\frac{\theta}{\theta + d})^\alpha. \end{aligned}$$

$$\begin{aligned} E(Y_L^2) &= \frac{2(\theta + d)^2}{(\alpha - 2)(\alpha - 1)}(\frac{\theta}{\theta + d})^\alpha \\ &= \frac{2\theta^2}{(\alpha - 2)(\alpha - 1)}(\frac{\theta}{\theta + d})^{\alpha - 2}. \end{aligned}$$

4.4 Gamma Distribution

If $X \sim Gamma(\alpha, \theta)$ then

$$\begin{aligned} F_{Y_P}(y) &= \frac{F_X(y+d) - F_X(d)}{S_X(d)} \\ &= \frac{S_{W^*}(\alpha-1) - S_W(\alpha-1)}{F_W(\alpha-1)}, W^* \sim Poisson(\lambda^* = \frac{y+d}{\theta}) \;\; W \sim Poisson(\lambda = \frac{d}{\theta}) \\ &= \frac{P(W \leq \alpha-1) - P(W^* \leq \alpha-1)}{P(W \leq \alpha-1)} \\ &= 1 - \frac{P(W^* \leq \alpha-1)}{P(W \leq \alpha-1)}. \end{aligned}$$

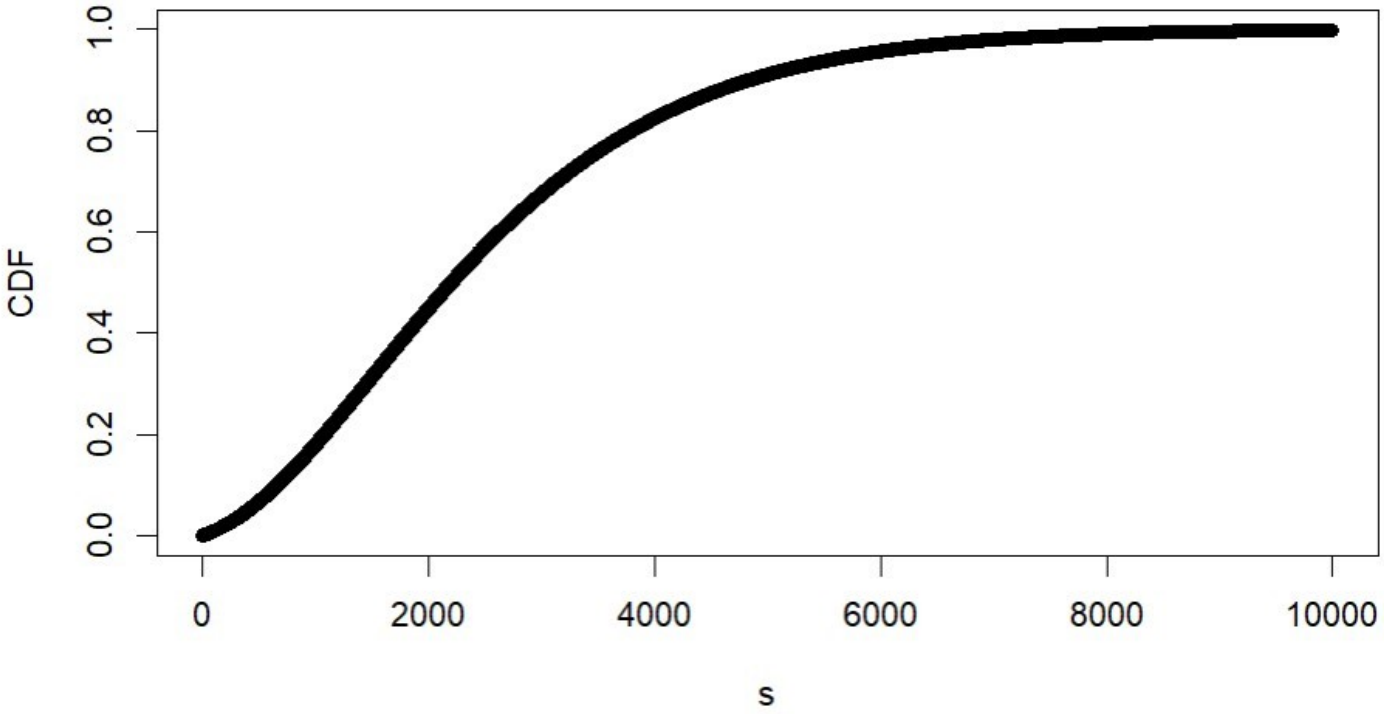
Unfortunately, for this method there is no concise formula for $E(Y_P^2)$, when the loss follows a gamma distribution. This is mainly due to the fact that the Y_P of a gamma distribution cannot be expressed in the form of a recognizable distribution. As seen above, the $F_{Y_p}(y)$ does not match any distribution's density function.

Although we were not able to derive a formula for $E(Y_P^2)$ using this method, we still found a result worth sharing. Since $F_{Y_P}(y) = 1 - \frac{P(W^* \leq \alpha-1)}{P(W \leq \alpha-1)}$ where W and W^* are Poisson distributions with the respective expected values of $\frac{d}{\theta}$ and $\frac{y+d}{\theta}$. We were able to create and define a function in R which gives us the cumulative probability of Y_P when X, the loss, follows a gamma distribution.

```
‘‘‘{r}
pgamma.yp <- function(a, t, d, y){

  b = (d + y)/t
  c = d/t

  1 - (ppois(a - 1, b)/ppois(a - 1, c)
       )
  }
pgamma.yp(3, 1000, 500, 2000)

s <- seq(from = 0, to = 10000, by = 1)
```

```
CDF <- pgamma.yp(3, 1000, 500, s)

plot(s, CDF)
```

In this R chunk, we are defining the function with the parameters "a" as α, "t" as θ, "d" as the deductible, and "y" as the point where we want to find the cumulative probability up to. Next, we test the function by plugging in 3 for α, 1000 for θ, 500 for d, and 2000 for y, which gives us back the probability of 0.4482485. After that, we create a sequence from 0 to 10000 in increments of 1, which we will use as input for y to make 10001 different cumulative probabilities. Lastly, we plot the sequence on the x-axis and the corresponding probabilities on the y-axis to visually display the cumulative distribution function.


CDF
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
s

We can verify that the cumulative distribution function is valid by checking if it is monotonically increasing, if it converges to a probability of 1 as $s \to \infty$, and if the probability is 0 when $s \to -\infty$. Based on the graph, we can see that the cumulative probability is 0 when n = 0. Additionally, we can see that the function is monotonically increasing. Lastly, we can observe that as $S \to \infty$, the cumulative probability approaches 1.

5 Conclusion

During this research, three different methods were explored to calculate the second moment of insurance payout Y_L, in situations where the loss amount followed a uniform, exponential, gamma or Pareto distribution. The first method relied on using the existing formula $Y_L = (X \wedge u) - (X \wedge d)$, then squaring both sides and expanding. As an extension of the most common method used to find the first moment of the insurance payout, this is conceptually straightforward. However, it is computationally very heavy, with some distributions relying on integration by parts to be fully integrated. This can make the actual use of this first method impractical.

The second method used the indicator variable, a variable that was set equal to 1 when condition $X > d$ is met, and was set equal to 0 when it was not. Although indicator variables are more commonly seen in data analysis and regression modeling, their usage here makes the integration process much easier, streamlining the computational process of finding the second moment for the various distributions.

The third method, which involves creating a new variable Y_p and relating it to Y_L, requires the introduction of several new concepts. It involves the creation of an entirely new variable, then finding the CDF of said variable and attempting to match it to an existing distribution. This conceptual difficulty, however, is rewarded by being the easiest method to use computationally. For some distributions (uniform, Pareto), using Y_p allows $E(Y_L^2)$ to be calculated without needing to use an integral at all. It should be noted that this method does not result in finding a common distribution associated with the CDF of gamma's associated Y_p. This makes the third method less useful for finding the second moment of the gamma distribution specifically compared to the other methods found.

The development of these new methods should make it much easier to evaluate the variance of insurance payments when they take on the four dif-

ferent distributions evaluated here. This will make it easier to evaluate and analyze the worth of various insurance packages. Additionally, offering more and easier ways to calculate the second moment should make learning about them in classroom settings more approachable. Moment-generating functions are a key concept that appears repeatedly in upper-division statistics courses, and these new methods will ideally make learning and computing them easier for students unfamiliar with them. Further research could be done by investigating the validity of these new methods on other common continuous distributions, such as the beta distribution.

Beyond Correlation:

An Analysis of Risk in the S&P 500 Index

Alejandro Reyes*[1]

Advisor: Matheus B. Guerrero[2]

[1]College of Engineering and Computer Science, CSUF

[2]College of Natural Sciences and Mathematics, CSUF

Abstract

This paper investigates the evolving patterns of risk and dependence between US equities in the S&P 500 Index, focusing on three major sectors. Communication Services, Consumer Discretionary, and Consumer Staples. Using GARCH-filtered log returns, we compare three complementary dependence measures, Spearman correlation, mutual information, and tail dependence coefficient, under both stable market conditions and the COVID-19-induced downturn. Our findings reveal that while traditional measures capture general co-movement, they often underestimate hidden systemic risks that emerge during crises. In particular, tail dependence analysis uncovers critical vulnerabilities that are not apparent through correlation or information-based metrics alone. Furthermore, cross-sector analysis highlights how diversification benefits erode when market stress escalates, as correlations and joint crash risks rise across traditionally independent industries. This study emphasizes the need for dynamic and tail-sensitive approaches in portfolio construction and risk management, providing practical insights for investors navigating increasingly volatile markets.

*Corresponding author: mralexreyes99@csu.fullerton.edu

1 Introduction

Over the past two decades, the U.S. stock market has experienced numerous significant crises. The 2020 market recession, defined by the COVID-19 pandemic, brought about massive layoffs across the globe and led to a rapid economic decline [Falk et al., 2021]. As the world shut down in favor of quarantine measures, supply chains were disrupted, and travel restrictions were implemented, stalling economic growth. Consumer markets were dramatically impacted, as people could not venture out to purchase goods and services. Simultaneously, the communication services sector experienced high volatility following the widespread adoption of remote work [Baker et al., 2020]. These abrupt shifts created a market environment filled with uncertainty, where traditional risk assessment methods proved limited in providing complete information regarding asset relationships. In addition, the COVID-19 pandemic caused unprecedented economic shocks, with U.S. GDP contracting 9.1% in Q2 2020 and unemployment peaking at 14.8% [U.S. Bureau of Economic Analysis (BEA), 2020]. This context underscores the need for **robust** risk assessment tools.

In order to effectively minimize risk within a stock portfolio, it is crucial to understand these shocks and their impact on statistical measures. This paper delves into the financial periods between 2016 and 2024 [OECD, 2023], providing a foundation in stock market analysis and extremal (or tail) dependence to promote portfolio diversification. In today's age of media saturation and misinformation, terms like "portfolio diversification" and "risk management" are frequently mentioned but are often not clearly defined [Markowitz, 1952]. For instance, portfolio diversification's limitations became evident during the 2020 crisis when previously uncorrelated assets moved in tandem [Center on Budget and Policy Priorities (CBPP), 2021]. Traditional correlation measures, as given by Myers and Sirois [2014], failed to predict this phenomenon, motivating our investigation into tail dependence [Embrechts et al., 1997]. Hence, this paper seeks to clearly define these concepts and demonstrate practical approaches to implementing them through statistical measures of dependence between two stocks.

Although portfolio diversification is widely discussed, its importance cannot be overstated. Diversification aims to minimize losses and smooth out stock returns during stable periods. However, diversifying investments during financial crises can be difficult, as some stock tickers may exhibit high correlations, leading to simultaneous declines across a portfolio [Cheng et al., 2022]. Tail dependence plays a critical role in risk management and portfolio diversification. Unlike traditional measures such as correlation, tail dependence provides valuable insights into the probability of extreme co-movements between assets during market stress [Ergen, 2014]. Traditional risk assessment models, like value at risk (VaR) and expected shortfalls, often overlook these dependencies, which can result in substantial financial losses during crises. Extreme value copulas offer a powerful tool to identify and model these extremal dependencies, providing a more accurate portfolio risk assessment [Frees and Valdez, 1998].

Our analysis focuses on the S&P 500 Index, a benchmark comprising 500 leading U.S. companies across 11 sectors. Each sector responds uniquely to market conditions [Nagarajan, 2021]; for example, consumer staples (e.g., household goods) tend to exhibit stability during downturns, whereas consumer discretionary stocks (e.g., luxury goods) and communication services (e.g., remote-work technologies) face higher volatility. We identify how (tail) dependencies evolve across market regimes by examining these sectoral dynamics before, during, and after the COVID-19 crisis.

This paper contributes to the field by comparing three dependence measures (Spearman correlation, mutual information (MI), and the tail dependence coefficient(TDC)) across multiple sectors of the S&P 500, using GARCH-filtered returns to account for volatility clustering and provide sector-specific insights for building resilient portfolios.

This paper proceeds as follows. Section 2 details our methodology, including volatility modeling via GARCH models and copula-based dependence measures. Section 3 presents our empirical findings, highlighting sector-specific tail risks and discussing practical portfolio diversification applications. Section 5 concludes.

2 Methods

This section describes how we quantify and compare dependence among stock returns under various market conditions. We begin by defining log returns, then introduce three key measures of dependence: Spearman correlation, mutual information (MI), and the tail dependence coefficient (TDC). Next, we describe how copulas provide a flexible way to model dependence—especially in the tails—and how we use GARCH modeling to handle volatility clustering in financial data.

Disclaimer: This chapter relies on the foundational work of [Bollerslev, 1986] for the modeling of volatility using GARCH processes, a standard approach for capturing time-varying variance in financial time series. For the theoretical underpinnings of extreme value theory and its application to financial risk, we draw on [Embrechts et al., 1997], a seminal reference in the field. The analysis of dependence structures, particularly in the tails, builds on the framework of copulas as introduced in [Nelsen, 2006]. Additionally, the concept and computation of mutual information, crucial for detecting both linear and non-linear dependencies between assets, follow the principles established by [Cover and Thomas, 2006]. Together, these references provide the statistical and probabilistic foundation necessary for the models and methodologies developed throughout this chapter.

2.1 Log Returns

To understand how stocks behave under different market conditions, we begin by computing the **log returns** of each stock in our dataset.

Let P_t denote the price of the stock at time t. We define the *log return* r_t as $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right)$.

Log returns are preferred in financial analysis because they are time-additive and scale-invariant, which simplifies comparisons and modeling.

Figure 1 illustrates the evolution of Apple Inc. (AAPL) share prices and their log returns. The left

panel shows the closing price in U.S. dollars, revealing general trends and volatility over the specified period. The right panel plots the log returns, which both handle large price changes more gracefully than raw percentage returns and tend to have more symmetric distributions—features that aid in risk modeling and time-series forecasting.

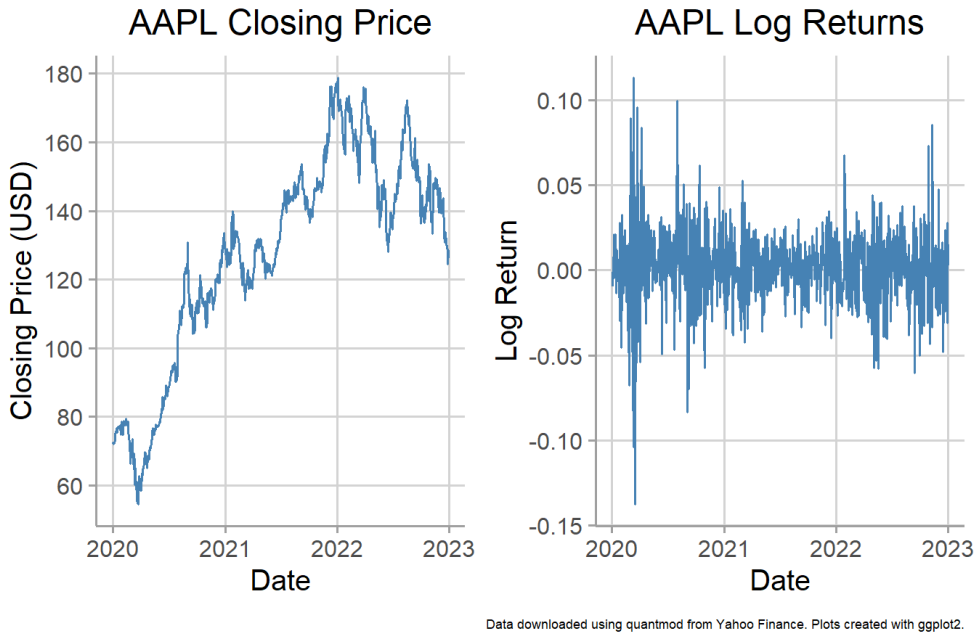


Figure 1: AAPL daily closing price (left panel) from 2020 to 2023 and corresponding log returns (right panel). Data were obtained from Yahoo Finance using the `quantmod` R package, and the figure was generated using `ggplot2`.

2.2 Dependence Measures

Once the log returns are computed, we analyze the dependence structure between a pair of stocks using three complementary approaches:

i. **Spearman Correlation:** a rank-based measure of monotonic dependence.

ii. **Mutual Information:** a measure that captures both linear and nonlinear dependencies.

iii. **Tail Dependence Coefficient:** a measure that focuses on the probability of extreme co-movements between stocks during market stress.

To improve the reliability of our dependence measures—especially for assessing extreme events—we first adjust for volatility in the log returns using Generalized Autoregressive Conditional Heteroskedasticity (GARCH) modeling. In the case of TDC, an additional step is required: before computing it, we fit the best copula to each pair of stocks' (or GARCH residuals') log returns. This extra step ensures that the estimated tail behavior accurately reflects the underlying dependence structure, as different copula families capture distinct tail properties.

Spearman Correlation Spearman correlation, denoted by ρ, measures the *rank-based* association between two variables. It is computed by replacing the raw observations with their respective ranks and then calculating the standard Pearson correlation of these ranks. Spearman correlation is non-parametric and robust to outliers but does not capture non-monotonic or extreme-tail relationships. Formally,

$$\rho \;=\; \frac{\mathrm{Cov}\big(R(X),\, R(Y)\big)}{\sigma_{R(X)}\,\sigma_{R(Y)}},$$

where $R(X)$ and $R(Y)$ denote the rank transformations of X and Y, respectively, and $\sigma_{R(X)}$ is the standard deviation of $R(X)$.

Mutual Information Mutual information measures how much knowing one random variable reduces uncertainty about another. It is defined as

$$I(X;Y) \;=\; \iint p(x,y)\,\ln\!\left(\frac{p(x,y)}{p(x)\,p(y)}\right)\,dx\,dy,$$

where $p(x,y)$ is the joint probability density function of X and Y, and $p(x)$, $p(y)$ are their marginal densities. Because MI can detect *nonlinear* dependencies, it may reveal relationships missed by Spearman correlation. However, it can be computationally intensive and often requires data discretization, which may introduce bias.

Tail Dependence Coefficient The Tail Dependence Coefficient, sometimes referred to as the *Chi measure*, focuses on *extreme co-movements* (e.g., when both stocks experience large losses simultaneously). It is defined separately for the lower tail, λ_L, and the upper tail, λ_U:

$$\lambda_L \;=\; \lim_{q\to 0^+} P\Big(X \le F_X^{-1}(q) \;\Big|\; Y \le F_Y^{-1}(q)\Big),$$

$$\lambda_U \;=\; \lim_{q\to 1^-} P\Big(X > F_X^{-1}(q) \;\Big|\; Y > F_Y^{-1}(q)\Big),$$

where F_X^{-1} and F_Y^{-1} are the quantile functions of X and Y, respectively. A higher TDC indicates a greater propensity for both assets to crash (or rally) together, making it especially important for *risk management* during crisis periods.

In summary, these complementary measures—capturing monotonic, nonlinear, and extreme-tail dependencies—offer a robust framework for tailored risk analysis.

2.3 Copulas

A *copula* is a CDF that joins the marginal distributions of random variables into a joint distribution, thereby separating individual (marginal) behavior from the *dependence structure*. Formally, if $F_X(x)$ and $F_Y(y)$ are the marginal CDFs for X and Y, then a copula C satisfies

$$H(x,y) \;=\; P(X \le x,\, Y \le y) \;=\; C\big(F_X(x),\, F_Y(y)\big).$$

In this study, we use the `VineCopula` package in R to automatically select and fit the best copula for each pair of stocks (i.e., each pair of X and Y). This step is crucial *before* computing the Tail Dependence Coefficient because different copula families capture different types of tail behavior. By selecting the copula that best fits each pair, we obtain a more accurate estimate of the joint extremes than would be possible with assumption-based approaches. Hence, the TDC computed is tailored to the actual behavior of each pair's data, leading to a more reliable risk measure.

Figure 2 illustrates an example of a *fitted copula* capturing the joint dependence of GOOGL and AMZN log returns. The color-coded contour regions represent areas of higher probability density, providing a top-down perspective on how the two stocks' returns co-move across a range of market conditions. Notably, this approach can reveal nuances—such as asymmetric or tail-heavy dependencies—that ordinary correlation measures may overlook.

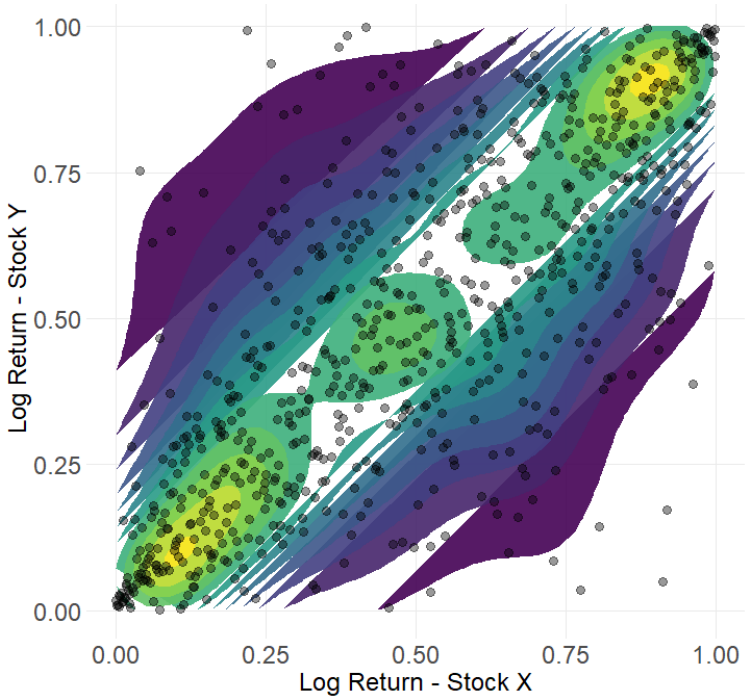


Figure 2: Estimated bivariate copula density for GOOGL vs. AMZN log returns. The color scale and contour lines highlight regions of higher joint density, revealing the stocks' non-linear dependence structure that may not be apparent through correlation alone.

2.4 Detecting Volatility Clustering with GARCH

Financial time series often exhibit *volatility clustering*, where periods of high volatility tend to follow other high-volatility periods (and vice versa). To address this, we adopt the GARCH model—specifically the GARCH(1,1) model—which, due to its simple yet robust structure, is often sufficient to capture the volatility clustering observed in a wide range of financial time series [**?**]. In

the GARCH(1,1) model, the observed return y_t is given by:

$$y_t = \mu + \varepsilon_t,$$

$$\varepsilon_t = \sigma_t \, z_t, \quad z_t \sim N(0,1),$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \, \varepsilon_{t-1}^2 + \beta_1 \, \sigma_{t-1}^2,$$

where

- μ is the mean return,
- ε_t is the innovation or shock at time t,
- σ_t^2 is the conditional variance (volatility) at time t,
- α_0, α_1, and β_1 are parameters.

We begin by applying an *ARCH test* on each stock's log returns $\{r_t\}$ to detect volatility clustering. If the test is significant, we fit a GARCH(1,1) model and extract the *GARCH residuals*, $e = \varepsilon_t / \sigma_t$, which serve as volatility-adjusted returns for subsequent dependence analysis. If the ARCH test does not indicate significant clustering, we use the raw log returns directly.

Figure 3 demonstrates how volatility clustering manifests in a simulated time series. Two calm segments—where returns fluctuate within a narrow band—are separated by a central, high-volatility regime with large spikes in magnitude. This pattern mirrors real-world financial markets, where periods of relative stability are often followed by stretches of persistently elevated volatility. Such clusters underscore the need for models that dynamically capture changes in market risk, such as GARCH(1,1) and its variants.

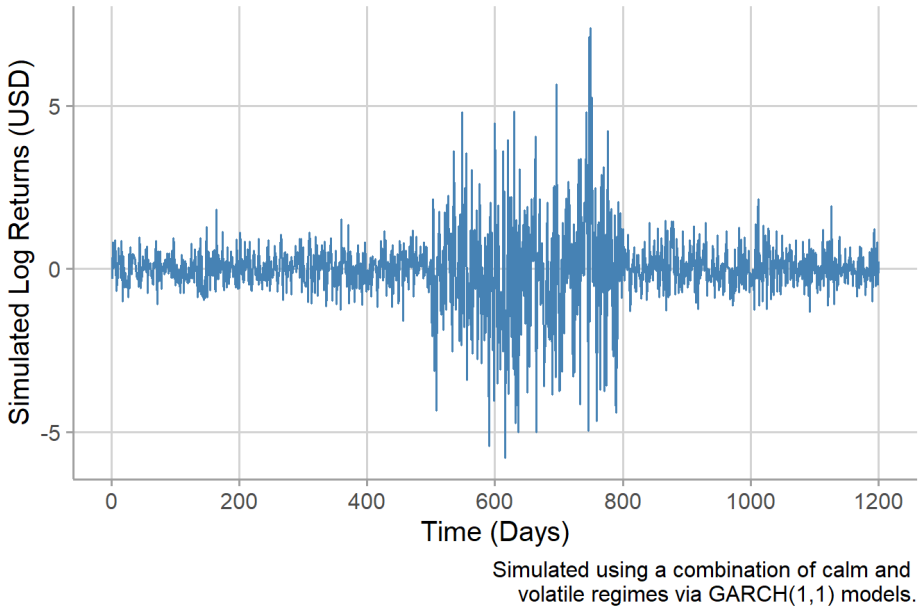


Figure 3: Simulated log returns illustrating volatility clustering: the first 500 observations and final 400 observations depict relatively calm periods, while the central 300 observations show heightened volatility. Such transitions between calm and volatile regimes highlight the persistent, clustered nature of volatility in financial time series.

3 Exploratory Data Analysis

Here, we analyzed stock tickers from three key prominent sectors in the S&P 500 Index: Communication Services, Consumer Discretionary, and Consumer Staples. Twelve stocks were carefully selected within each sector based on popularity, share price, and overall holding percentage.

We began our analysis by plotting the scaled closing prices for the selected stock tickers from January 1, 2016, through January 1, 2024. This interval includes distinct market regimes: a "Stable Market" period, during which stock market trends moved closely together, and a subsequent period starting in early 2020 marked by significant market fluctuations due to the COVID-19 pandemic, called here "Pandemic and Downturn." Since we selected twelve stocks per sector, the visualizations highlight three stocks, in particular, to better illustrate changes in closing prices over time and underscore the influence of notable companies.

From the scaled closing prices, it was evident that fluctuations would occur. However, to ade-

quately assess risk, we also calculated the log returns for each stock and visualized these data. Log returns were chosen for multiple reasons. Although each stock is included in the S&P 500 Index, their prices differ considerably. Some tickers have significantly higher prices, making direct comparisons using standard returns potentially misleading or unfair. By calculating log returns, all stock returns could be compared uniformly on the same scale.

3.1 Communication Services

Figure 4 consists of two panels that illustrate the performance of three major stocks in the Communication Services sector, Google (GOOGL), Meta (META), and Netflix (NFLX) over several years. The left panel displays the scaled closing prices of these stocks from 2016 through 2023, while the right panel shows their corresponding log returns for the same period. Colored shading distinguishes key market phases: the "stable market" and the "pandemic and downturn."

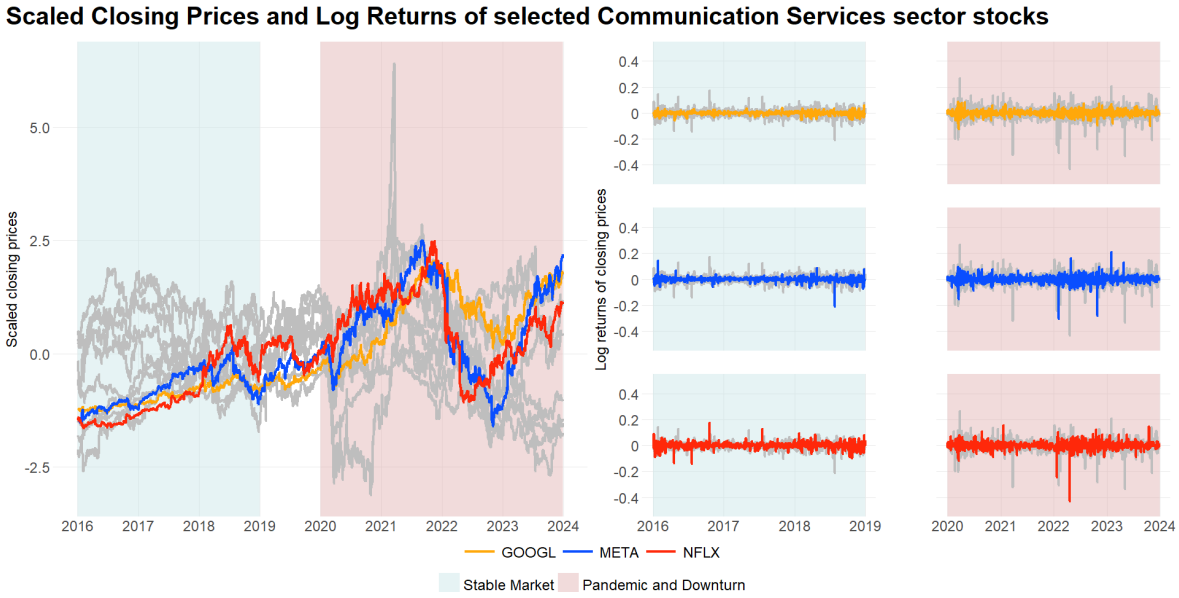


Figure 4: Scaled closing prices and log returns for selected stocks in the Communication Services sector (Google, Meta, and Netflix) from January 1, 2016, through January 1, 2024. Colored shading distinguishes different market phases: "stable market" and "pandemic and downturn."

During the stable market phase, stocks in the Communication Services sector exhibited relative growth in their scaled closing prices. Although minor shifts occurred, an overall upward trend is

clearly observable. The log returns plot (right panel) similarly reflects this pattern, with returns consistently centered around zero and characterized by minor fluctuations. This indicates that investors had relatively stable expectations for future earnings and market conditions during this period.

In the pandemic phase, the impact of COVID-19 becomes evident. All stocks in our analysis experienced sharp declines, reaching, for some, their lowest points since 2019. The volatility in log returns intensified, showing substantial negative returns, especially pronounced during 2022. Interestingly, GOOGL, META, and NFLX experienced partial recoveries prior to 2022, including rapid closing price increases, a behavior also reflected in their log returns. Such recoveries may be attributed to consumer adjustments, as quarantine measures and remote work practices were widely implemented. Nevertheless, the year 2022 brought yet another significant downturn across all stocks, including those companies that had previously recovered strongly, underscoring the market's vulnerability to shocks.

3.2 Consumer Discretionary

Figure 5 displays the scaled closing prices and corresponding log returns for selected stocks within the Consumer Discretionary sector. During the stable market period, we observed a similar growth pattern in the Communication Services sector. In this case, the closing prices for all stocks were more closely aligned. The highlighted stocks, Amazon (AMZN), MGM Resorts (MGM), and Tesla (TSLA), show log returns centered around zero and behaved similarly during this period.

The onset of the pandemic marks a dramatic change in the behavior of these stocks. TSLA, which was relatively steady before the pandemic, exhibited exponential growth that continued through much of the pandemic until 2022. AMZN also experienced significant growth, primarily driven by increased consumer reliance on e-commerce to purchase goods. In contrast, MGM experienced the most significant initial decline in share price, coinciding with COVID-19 and the resulting quarantine measures, as the company's business rapidly contracted when patrons were re-

stricted from non-essential activities. In subsequent years, MGM gradually adapted to regulatory changes and substantially recovered from these early losses. However, 2022 proved challenging for all companies within this sector, as each stock experienced notable declines in share prices.

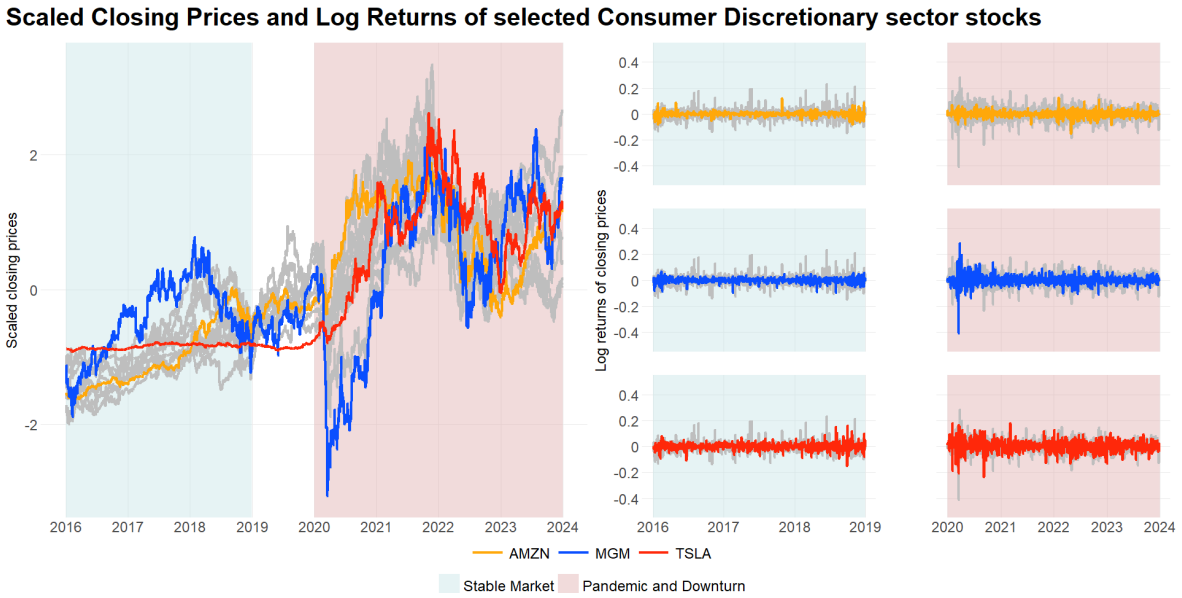


Figure 5: Scaled closing prices and log returns for selected stocks in the discretionary consumer sector (Amazon, MGM Resorts, and Tesla) from January 1, 2016, through January 1, 2024. Colored shading distinguishes different market phases: "stable market" and "pandemic and downturn."

These movements are clearly reflected in the log returns, illustrating significant variations in investor profits. Upon closer inspection, larger deviations from zero are particularly apparent for MGM, indicating multiple periods of substantial losses. TSLA, despite appearing stable based on closing prices alone, demonstrates pronounced risk to investors as periods of high returns are intertwined with significant downturns. AMZN exhibits a unique pattern compared to the other two companies: Starting with minor fluctuations, it subsequently displays an expanded variability in the returns. In general, the pandemic period highlights how inconsistent share prices contribute to elevated risk for investors in these companies.

3.3 Consumer Staples

Figure 6 presents the scaled closing prices and corresponding log returns for selected stocks within the Consumer Staples sector. In this sector, behaviors partly reflect trends observed in other sectors, but deviate in some notable aspects. The highlights in this figure are Walmart (WMT), Coca-Cola (KO), and Costco (COST), which were chosen to illustrate the changes throughout our study period. During the stable market phase, most stocks in this sector fluctuate around the -1 scaled closing price point, but overall show a steady upward trend.

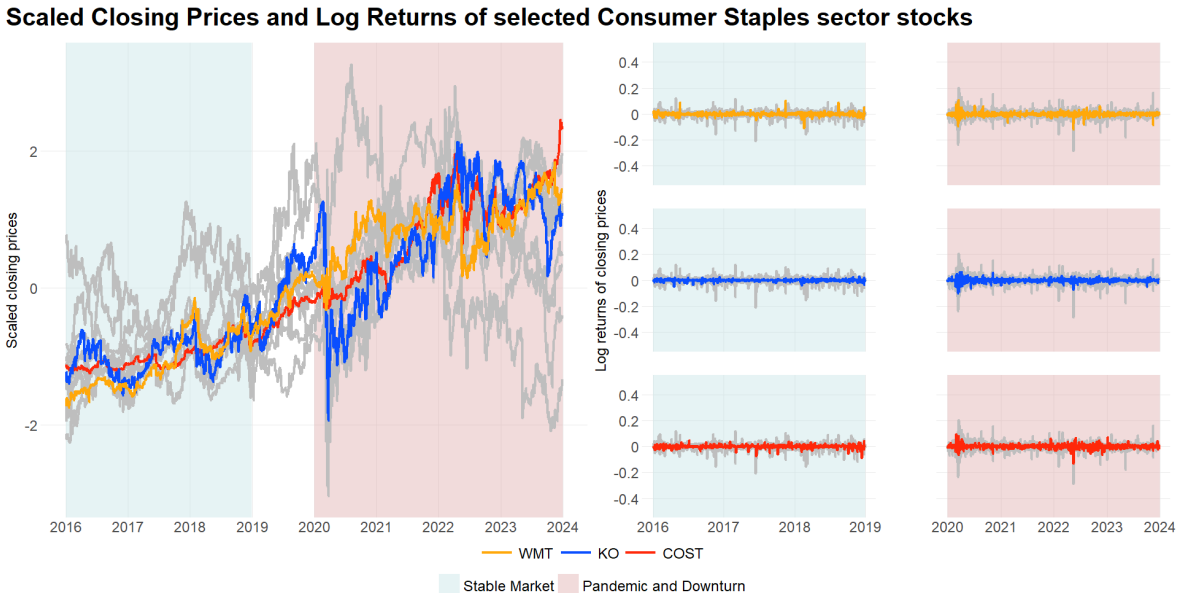


Figure 6: Scaled closing prices and log returns for selected stocks in the Consumer Staples sector (Walmart, Coca-Cola, and Costco) from January 1, 2016 to January 1, 2024. Colored shading distinguishes different market phases: "stable market" and "pandemic and downturn."

As in other sectors discussed previously, the pandemic, although not uniformly severe, affected many companies. Losses are observable for WMT and COST, but these declines were not as pronounced as for KO, which experienced a substantial drop in the share price. Despite the pandemic period, where consumer travel and mobility were significantly restricted, the ability of major grocery markets to provide essential goods largely shielded companies like COST and WMT from severe downturns, a resilience confirmed by their log returns. This resilience is also ob-

served, to some extent, in the log returns of KO despite its initial decline.

All three highlighted stocks recovered from their pandemic-era price drops, continuing to exhibit growth into 2024, with only a few companies experiencing downturns starting around 2022. The log returns confirm that although market shocks temporarily affected some companies, most recovered, returning to behaviors similar to those observed during the stable market period.

4 Results and Discussion

This section highlights the metrics computed for all pairwise stocks within their respective sectors and concludes with findings regarding extreme-value dependence. Initially, we created a function to test for ARCH effects, assessing whether stocks exhibit volatility clustering and, if necessary, to fit a GARCH(1,1) model. From this model, we obtained the GARCH residuals to perform subsequent analyses. The R function used for this process is provided in Appendix A.

If no ARCH effects were detected, the original log returns were retained for analysis; otherwise, GARCH residuals were used. Interestingly, only three stocks—META, NFLX, and TGT—failed the ARCH test, meaning that only these stocks continued to carry out their original log returns.

We computed three dependence measures using the residuals obtained from the GARCH modeling: Spearman correlation, mutual information (MI), and tail dependency coefficient (TDC). The results for each measure are discussed below.

4.1 Sector Analysis

4.1.1 Spearman Correlation Within Sectors

The Spearman correlation plots, provided in Figure 7, reveal a clear shift in interstock dependencies between stable and pandemic periods across the Communication Services, Consumer Discretionary, and Consumer Staples sectors.

During the stable period, most pairs of stocks exhibit mild to moderate correlations, with Spearman's ρ values generally ranging from 0.2 to 0.5. However, several standout pairs consistently show stronger rank-based relationships, suggesting that these stocks tended to move in sync regarding their rankings even under calm market conditions. Notable examples include GOOGL and META ($\rho = 0.70$), and VZ and T ($\rho = 0.70$) within Communication Services; LOW and HD ($\rho = 0.75$) within Consumer Discretionary; KO and PEP ($\rho = 0.71$) within Consumer Staples. These relationships likely reflect underlying business similarities or shared industry dynamics.

Interestingly, some cross-industry correlations emerge even during the stable period, such as PEP and HD ($\rho = 0.59$). This highlights that certain relationships are structurally embedded rather than purely driven by external shocks.

In contrast, correlations intensified across all three sectors during the pandemic period. Systemic pressures such as supply chain disruptions, changes in consumer behavior, and coordinated monetary responses diminished idiosyncratic stock movements and led to elevated correlations.

Certain strong relationships are further strengthened. For example, LOW and HD increased from $\rho = 0.75$ to $\rho = 0.84$. At the same time, other previously moderate correlations, such as NKE-SBUX and DIS-WBD, moved into the high-dependence range, contributing to a collapse in portfolio diversification opportunities.

4.1.2 MI Within Sectors

The MI plots in Figure 8 follow a trend similar to that observed with the Spearman correlation, although important differences emerge. Like the Spearman correlation, MI shows an increase in dependence during the pandemic compared to the stable period. However, MI also captures subtle, non-monotonic relationships that the Spearman correlation may miss.

Several pairs of stocks, such as GOOGL-META, LOW-HD, and KO-PEP, maintain strong dependence on both measures, reinforcing the robustness of their interconnected behaviors. Interest-

ingly, MI reveals relationships where Spearman's correlation appears to be weak. For example, pairs such as DIS-NFLX or TSLA-AMZN show low Spearman correlation but moderate MI values, suggesting more complex nonlinear interdependence not driven solely by rank order. This shows the strength of MI in detecting hidden structures within stock relationships.

4.1.3 TDC Within Sectors

The TDC provides additional insights that complement—and sometimes challenge—what is observed through Spearman correlation and MI; see Figure 9. TDC focuses on joint extreme events, the probability that two stocks experience simultaneous extreme losses.

One major finding is the identification of high-tail dependence in stock pairs that exhibit only mild or moderate dependence on Spearman or MI measures. For example, stock pairs such as DIS–EA, TSLA–BBY, and KO–WMT display noticeably elevated TDC values despite showing modest Spearman or MI values. This suggests that these stocks are much more likely to crash together under crisis than traditional metrics imply.

This phenomenon highlights the danger of false negatives in traditional dependency analysis. Investors relying solely on correlation or MI could mistakenly assume diversification benefits when, in reality, hidden co-crash risks remain. Misjudging this risk could lead to a substantial underestimation of portfolio vulnerabilities during stress events.

In contrast, some pairs with high Spearman and MI values do not exhibit elevated TDC scores. For example, although GOOGL–META are highly correlated in general movements, their tail dependence is relatively modest, implying a lower joint crash risk than their average co-movement would suggest.

Finally, it is important to acknowledge that high dependence scores are partly expected within the same sector. Companies such as HD and LOW, direct competitors offering almost identical products, naturally exhibit similar behavior across all measures, Spearman, MI, and TDC.

Overall, while intra-sector dependencies are high, it is crucial for investors seeking proper diversification to extend their portfolios across different sectors. The next section, cross-sector analysis, offers a broader view of diversification opportunities.

4.2 Cross-Sector Analysis

While the previous findings provide valuable insights into intra-sector dependencies, higher dependence scores among stocks within the same sector are expected due to shared market dynamics, industry trends, and similar business models. For instance, Home Depot (HD) and Lowe's (LOW) are direct competitors offering nearly identical products and services, resulting in financial performance that tends to move in tandem—a pattern consistently observed across all three dependence measures: Spearman correlation, MI, and TDC.

To develop a more comprehensive understanding of co-movement patterns and to better assess diversification potential, we extend our analysis beyond individual sectors. By computing dependence metrics across sectors and examining their behavior under different market regimes, we can uncover hidden systemic risks and interdependencies that may not be apparent within sector boundaries. This cross-sector analysis is particularly valuable for investors seeking to build diversified portfolios, as it highlights how exposures in one industry can unexpectedly become correlated with exposures in another during periods of market stress.

4.2.1 Spearman Correlation Across Sectors

During the stable period, the heatmap in Figure 10 reveals distinct sectoral clustering. Stocks within the same sector exhibit stronger correlations, reflected by the warmer diagonal blocks separated by vertical lines. For example, GOOGL–META and DIS–NFLX within Communication Services, HD–LOW within Consumer Discretionary, and KO–PEP within Consumer Staples all display strong intra-sector relationships, driven by shared exposure to similar consumer bases, industries, or market forces.

Cross-sector correlations, by contrast, remain lower during the stable period. The cooler colors outside the sector blocks indicate that sectoral diversification tends to be effective under normal market conditions.

During the pandemic period, however, this sectoral structure begins to erode. The heatmap becomes broadly warmer, with correlations increasing both within and across sectors. Market-wide shocks—such as supply chain disruptions, monetary policy interventions, and shifts in consumer behavior—caused stocks across different sectors to move more closely in tandem. Communication Services stocks, in particular, became more correlated with other sectors due to the growing reliance on digital platforms during the pandemic.

Cross-Sector Observations A key observation is the contagion-like effect observed during the pandemic. Cross-sector correlations that were previously mild intensified:

- DIS (entertainment) and WMT (retail), which were previously only modestly correlated, exhibited stronger co-movement.
- TSLA (technology/auto) began correlating more strongly with companies such as PEP and MCD, reflecting a broader convergence in market responses.

These patterns suggest that diversification strategies based solely on sector allocation may fail during periods of crisis, when correlations across sectors tend to converge.

4.2.2 MI Across Sectors

Unlike Spearman correlation, MI captures both linear and non-linear dependencies without assuming monotonicity. As shown in Figure 11, MI detects subtle cross-sector relationships even during the stable period that are missed by Spearman correlation. For instance, pairs such as

DIS–NFLX and TSLA–AMZN exhibit non-linear dependencies likely linked to shared trends in technology adoption, consumer sentiment, or market expectations.

During the pandemic period, MI matrices show a noticeable increase in inter-sector dependencies. Although the intensification is not as uniform as that observed with Spearman correlation, the MI heatmap indicates that stocks from different sectors began to share more information, reflecting broad exposure to systemic market forces.

However, MI has limitations. While it captures the presence of dependency, it does not distinguish between mild, moderate, or extreme co-movements. As a result, although MI reveals the existence of interdependencies, it does not convey the severity of joint movements under stress. This limitation highlights the importance of incorporating tail-sensitive measures, such as the TDC, into the analysis.

4.2.3 TDC Across Sectors

During the stable period, the TDC matrix shown in Figure 12 remains relatively muted, although certain stock pairs exhibit elevated tail dependence despite displaying weak correlation or mutual information values. For instance, pairs such as MCD–SBUX and DIS–EA demonstrate higher TDC values, revealing vulnerabilities to joint extreme losses that are not apparent through traditional dependence metrics. These cases represent false negatives from the perspective of conventional dependency measures.

During the pandemic period, however, tail dependence increases dramatically across the entire matrix. Stocks across sectors—including pairs that were previously only weakly related—began exhibiting high probabilities of simultaneous extreme losses. Notable examples include TSLA–-WMT and DIS–KO, which underscore the systemic nature of crisis events and the breakdown of sectoral distinctions during periods of market stress.

4.3 Synthesis and Implications

The combination of Spearman correlation, MI, and the TDC provides a multifaceted understanding of inter-stock dependencies:

- **Spearman correlation** captures typical rank-based co-movement and is useful for analyzing day-to-day behavior.
- **MI** reveals more complex, non-linear relationships, particularly when traditional correlation measures fail.
- **TDC** exposes joint behavior under extreme conditions, making it essential for assessing downside risk and portfolio contagion.

False Negatives and Hidden Risk A key insight from this analysis is the identification of false negatives in traditional dependence metrics. Several stock pairs exhibit decreasing correlations from the stable to the pandemic period, yet their TDC values increase, revealing hidden co-crash risks:

1. TSN – COST (Tyson Foods – Costco)
2. SYY – T (Sysco – AT&T)
3. PEP – MCD (PepsiCo – McDonald's)
4. SYY – DIS (Sysco – Disney)
5. CLX – MAR (Clorox – Marriott)
6. MNST – TGT (Monster Beverage – Target)
7. ETSY – T (Etsy – AT&T)

These examples highlight the importance of complementing correlation and mutual information with tail-based metrics such as the TDC when constructing portfolios and conducting risk analysis. Relying solely on average-case dependencies can lead to a significant underestimation of exposure to systemic shocks.

Conclusion Cross-sector analysis demonstrates that during crises, asset behaviors tend to converge. Stocks that appear independent under normal conditions can become tightly linked under stress, substantially eroding the benefits of diversification. Incorporating tail dependence into risk models provides investors with a deeper and more realistic understanding of market behavior during extreme events, enabling the construction of more resilient portfolio strategies.

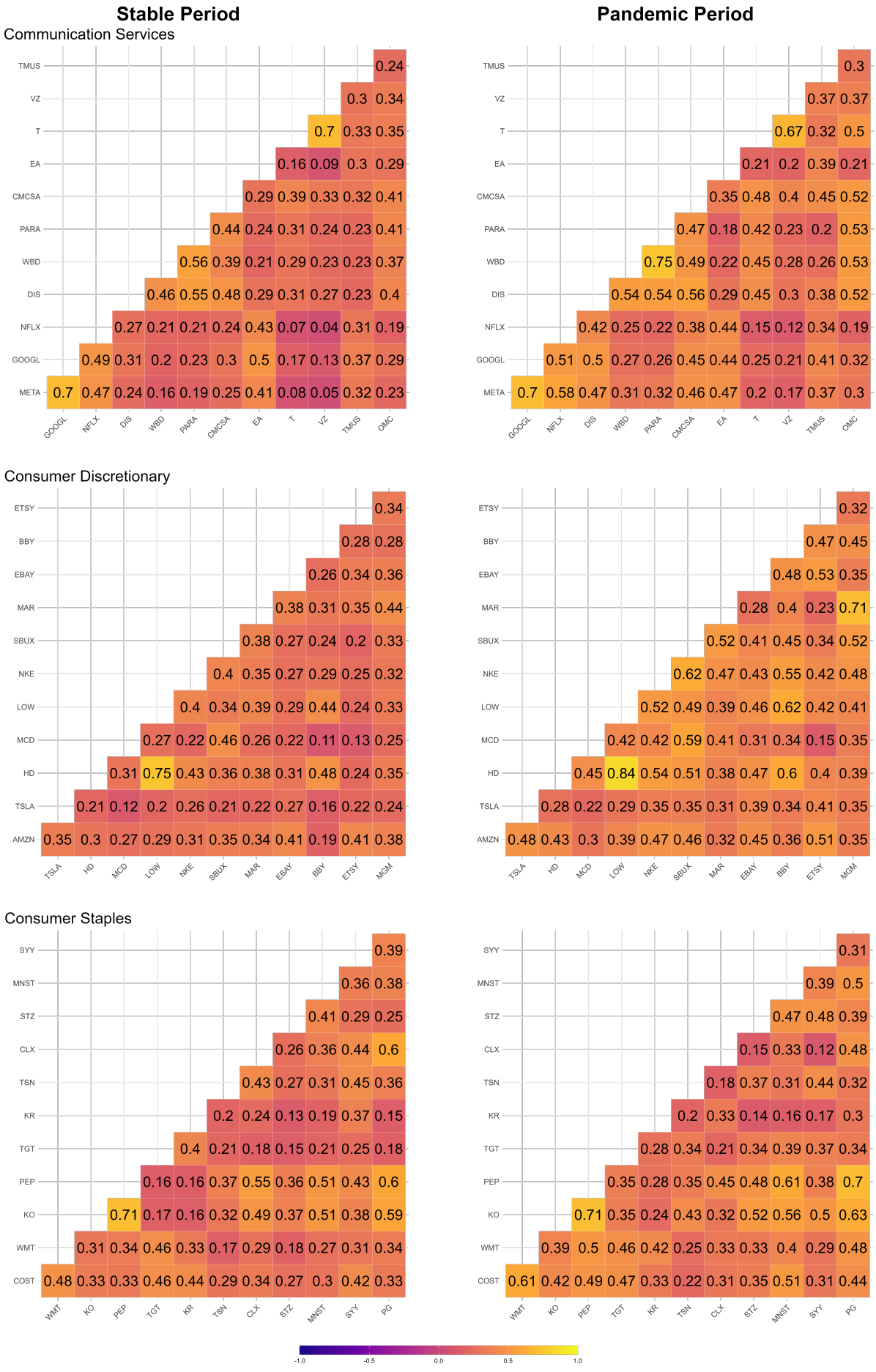


Figure 7: Spearman correlation across three S&P 500 sectors for stable and pandemic periods.

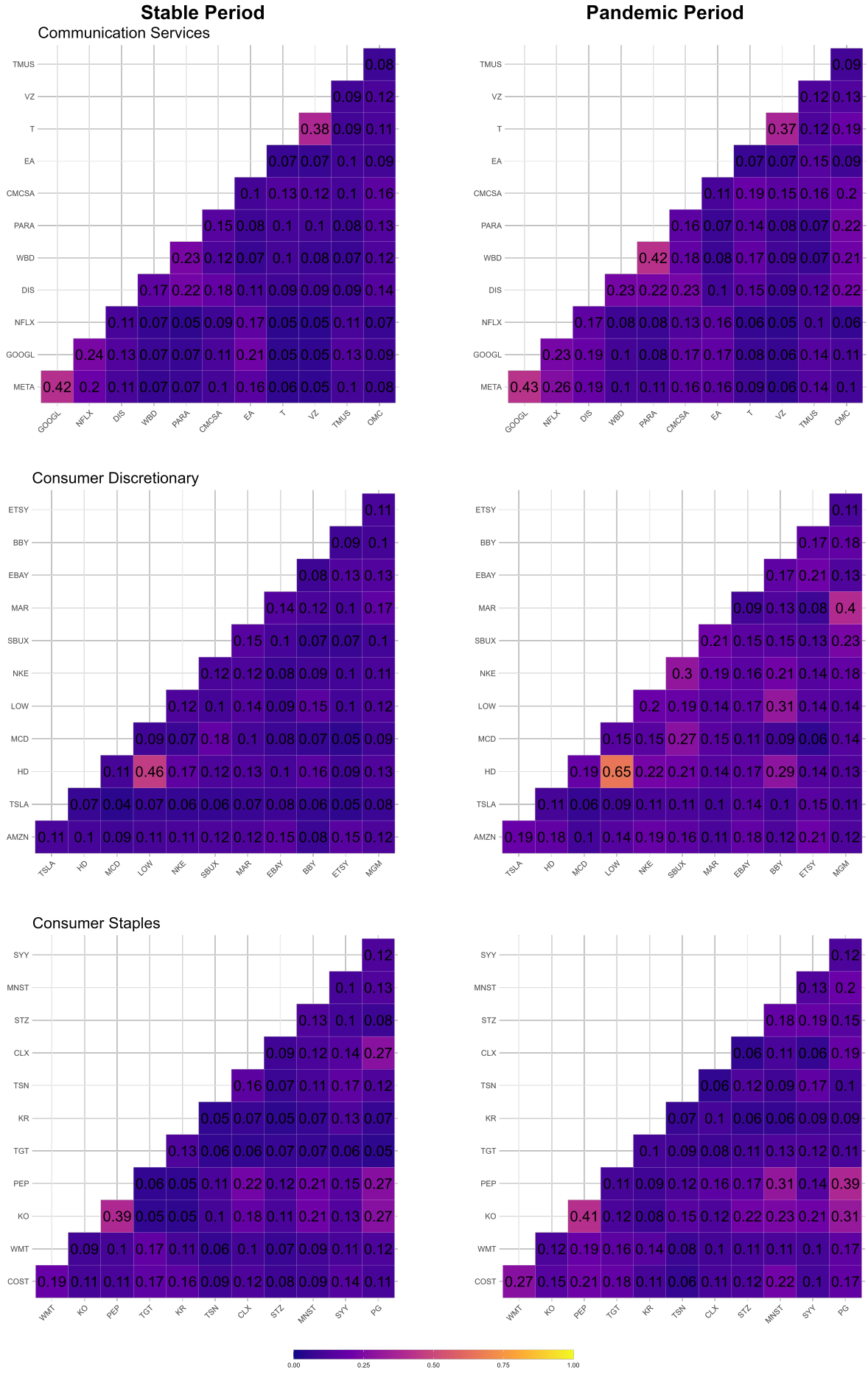


Figure 8: MI across three S&P 500 sectors for stable and pandemic periods.

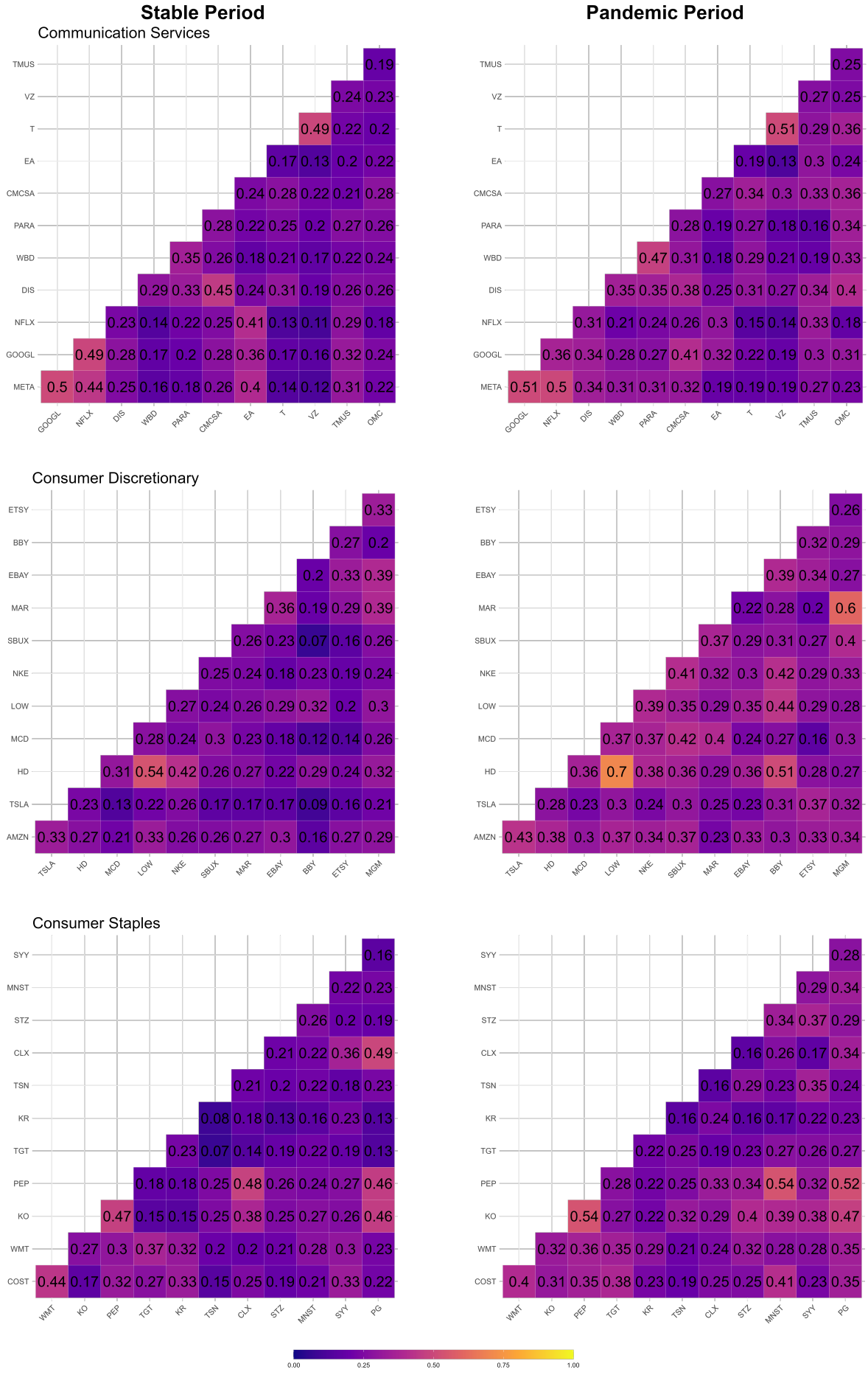


Figure 9: TDC across three S&P 500 sectors for stable and pandemic periods.

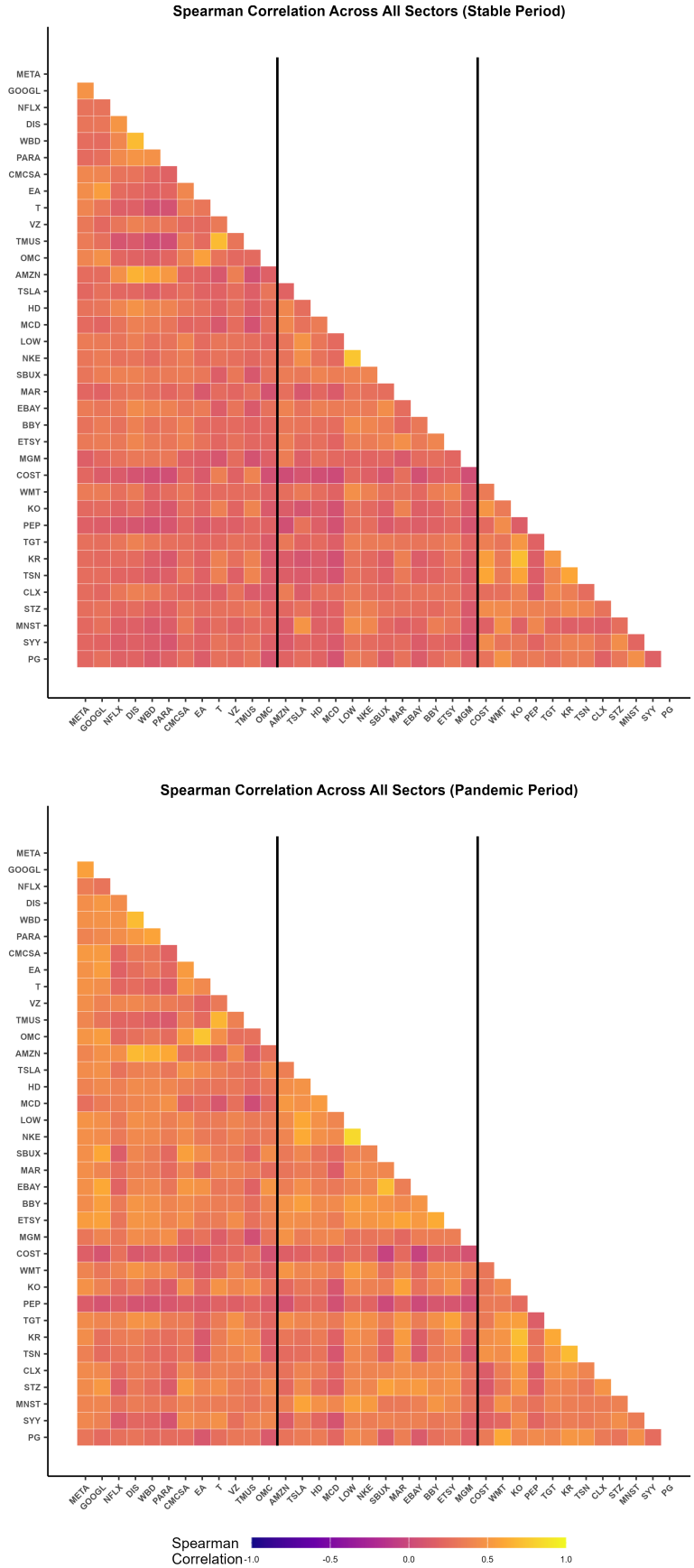


Figure 10: Spearman correlation matrix for all sectors during stable and pandemic periods.

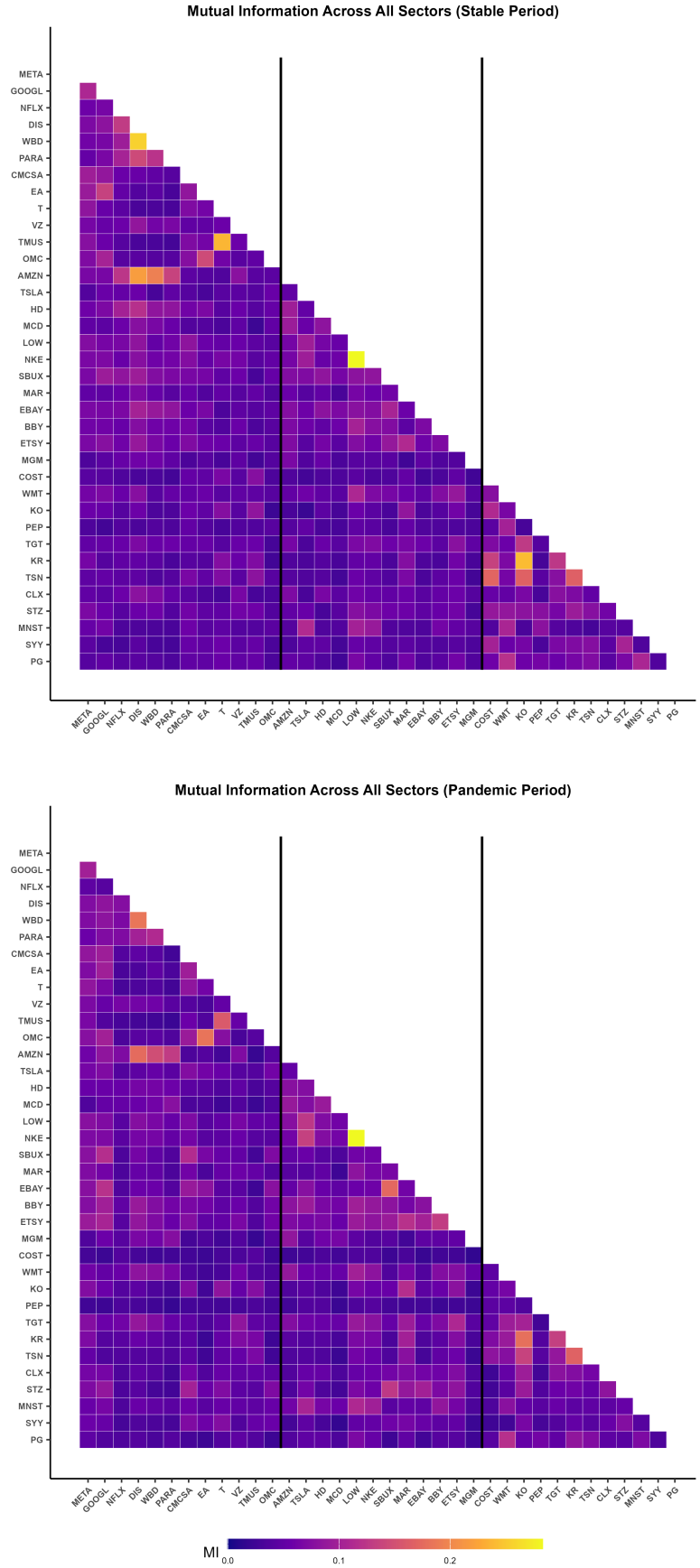


Figure 11: MI matrix across all sectors during stable and pandemic periods.

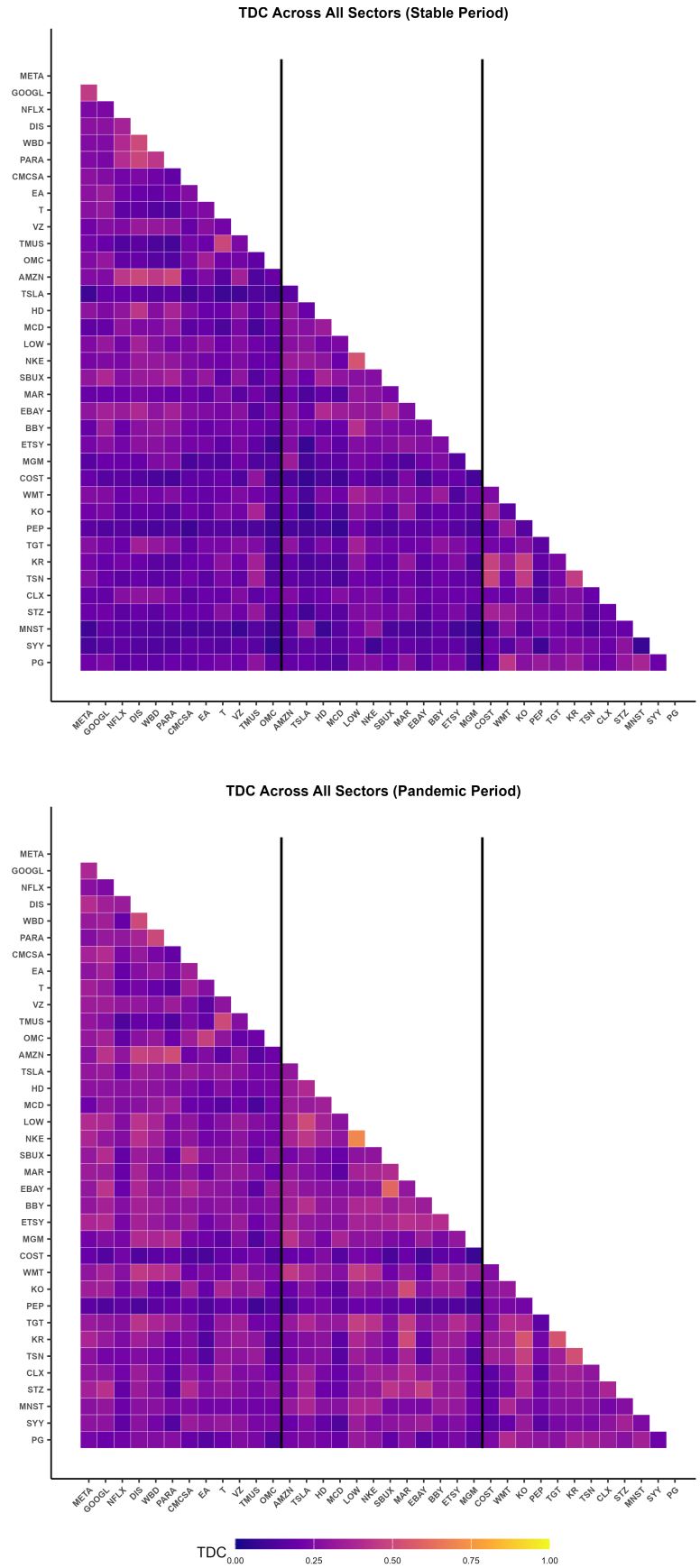


Figure 12: TDC matrix across all sectors during stable and pandemic periods.

5 Future Work

Several avenues exist to expand and deepen the research presented in this study. First, the analysis could be extended to incorporate additional historical periods of market stress, such as the 2008 financial crisis. Including past market shocks would allow for a broader understanding of how dependencies between assets evolve under different types of crises.

Second, the methodology could be applied to international markets or alternative asset classes, such as cryptocurrencies. Since this study focused exclusively on U.S. equities within the S&P 500, analyzing other markets would provide insight into whether the observed dependency structures and tail risks generalize beyond U.S.-based stocks.

Finally, future research could explore the application of machine learning models that adapt to changing market conditions. For instance, integrating dynamic copula models or regime-switching frameworks could better capture time-varying dependencies, thereby enhancing portfolio diversification strategies. These extensions would provide investors with more flexible and adaptive tools for managing risk in increasingly complex and interconnected financial markets.

References

Scott R. Baker, Nicholas Bloom, Steven J. Davis, Kyle J. Kost, Marco C. Sammon, and Tasaneeya Viratyosin. The unprecedented stock market impact of covid-19. Technical Report 26945, National Bureau of Economic Research, 2020.

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

Center on Budget and Policy Priorities (CBPP). Tracking the recovery from the pandemic recession. *CBPP Research*, 2021. URL `https://www.cbpp.org/research/economy/tracking-the-recovery-from-the-pandemic-recession`.

Tingting Cheng, Junli Liu, Wenying Yao, and Albert Bo Zhao. The impact of covid-19 pandemic on the volatility connectedness network of global stock market. *Pacific-Basin Finance Journal*, 71:101678, 2022.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. doi: 10.1002/047174882X.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events: for Insurance and Finance*, volume 33 of *Stochastic Modelling and Applied Probability*. Springer, 1997. doi: 10.1007/978-3-642-33483-2.

Ibrahim Ergen. Tail dependence and diversification benefits in emerging market stocks: an extreme value theory approach. *Applied Economics*, 46(19):2215–2227, 2014. doi: 10.1080/00036846.2014.899678.

Gene Falk, Paul D. Romero, Isaac A. Nicchitta, and Emma C. Nyhof. Unemployment rates during the covid-19 pandemic: In brief, 2021. URL `https://crsreports.congress.gov/product/pdf/R/R46554/9`.

Edward W. Frees and Emiliano A. Valdez. Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25, 1998. doi: 10.1080/10920277.1998.10595667.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi: 10.1111/j.1540-6261.1952.tb01525.x.

Leann Myers and Maria J. Sirois. Spearman correlation coefficients, differences between. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, 2014. doi: 10.1002/9781118445112.stat02802.

Sachin Nagarajan. These sectors performed best and worst in the pandemic. *Morningstar*, 2021. URL https://www.morningstar.com/markets/these-sectors-performed-best-worst-pandemic.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, 2 edition, 2006. doi: 10.1007/0-387-28678-0.

OECD. Global real gross domestic product (gdp) growth after covid-19 from 2019 with forecast until 2024, 2023. URL https://www.statista.com/statistics/1102889/covid-19-forecasted-global-real-gdp-growth/.

U.S. Bureau of Economic Analysis (BEA). Gross domestic product, 2nd quarter 2020 (second estimate); corporate profits, 2nd quarter 2020 (preliminary estimate), 2020. URL https://www.bea.gov/sites/default/files/2020-08/tech2q20_2nd.pdf. News Release BEA 20-41.

A Appendix

A.1 GARCH Residual Computation Code

The following R function was used to detect ARCH effects, fit GARCH(1,1) models where necessary, and extract residuals for subsequent analysis.

```
#' @param x A numerical vector containing the time series of log returns.
#' @param lags The number of lags for the ARCH test (default is 5).
#' @param alpha The significance level for the ARCH test (default is 0.05).
#'
#' @return If no ARCH effects are found, the function returns the original time series.
#' If ARCH effects are detected, it returns the residuals of the fitted GARCH(1,1) model.
#'
#' @examples
#' # Example with simulated GARCH(1,1) data:
#' n <- 1000
#' x <- as.numeric(fitted(ugarchpath(ugarchspec(), n.sim = n)))
#' result <- test_and_fit_garch(x)
#' print(result)
#'
#' @import rugarch
#' @import FinTS
#' @export
test_and_fit_garch <- function(x, lags = 5, alpha = 0.05) {

  # Step 1: Perform the ARCH test using FinTS::ArchTest
  arch_test <- FinTS::ArchTest(x, lags = lags)

  # Extract the p-value from the test
  p_value <- arch_test$p.value

  # Step 2: Decision rule based on p-value
```

```
  if (p_value >= alpha) {
    # If the p-value is greater than alpha, no ARCH effects are detected
    print("No significant ARCH effects. A GARCH model might not be necessary.")
    return(x)  # Return the original time series
  } else {
    # If the p-value is less than alpha, ARCH effects are detected
    print("There is evidence of ARCH effects. A GARCH model may be necessary.")

    # Step 3: Fit the GARCH(1,1) model to the data 'x'
    garch_spec <- ugarchspec(
      variance.model = list(model = "sGARCH", garchOrder = c(1, 1)),
      mean.model = list(armaOrder = c(0, 0), include.mean = TRUE),
      distribution.model = "norm"
    )

    # Fit the model
    garch_fit <- ugarchfit(spec = garch_spec, data = x)

    # Step 4: Extract the residuals
    garch_residuals <- residuals(garch_fit)

    # Return the residuals of the GARCH(1,1) model
    return(as.numeric(garch_residuals))
  }
}

# List of data frames for each sector
sectors_log_returns <- list(
  communication = communication_period,
  discretionary = discretionary_period,
  staples = staples_period
)
```

```
# Initialize an empty data frame to store the GARCH results
garch_results <- data.frame(
  date = as.Date(character()),
  sector = character(),
  symbol = character(),
  residual = numeric(),
  passed_garch_test = logical()
)

# Loop through each sector
for (sector_name in names(sectors_log_returns)) {

  # Get the data for the current sector
  sector_data <- sectors_log_returns[[sector_name]]

  # Get the unique stock symbols in this sector
  stock_symbols <- unique(sector_data$symbol)

  # Loop through each stock
  for (stock in stock_symbols) {

    # Extract the log returns and dates for the current stock
    stock_data <- sector_data[sector_data$symbol == stock, ]
    dates <- stock_data$date  # Assuming 'date' column exists
    log_returns <- stock_data$close_lgret

    # Run the GARCH test and fit function
    garch_result <- test_and_fit_garch(log_returns)

    # Determine if the GARCH model was fitted
    passed_garch_test <- !identical(log_returns, garch_result)

    # Add each entry as a separate row to the garch_results data frame
```

```
    for (i in seq_along(dates)) {
      garch_results <- rbind(
        garch_results,
        data.frame(
          date = dates[i],
          sector = sector_name,
          symbol = stock,
          residual = garch_result[i],
          passed_garch_test = passed_garch_test
        )
      )
    }
  }
}

# head(garch_results)
```

Listing 1: R function for ARCH testing and GARCH(1,1) model fitting