20 Years of Scientific Training in 100 Manuscripts (2001 – 2020)

20 Years of Scientific Training in 100 Manuscripts (2001 – 2020)

Maurice Ling

First Printing: 2025

ISBN 978-1-304-26622-4

9 781304 266224

Contents

Preface: The Story Unfolds

As I reflect on the journey that has led me here, I am reminded of the complexities of both science and life – how we are often driven by a deep, sometimes restless curiosity, seeking to understand, to create, and to leave an impact. This book, a mosaic of my personal and professional experiences, is an attempt to make sense of a path that has been both intensely focused and, at times, serendipitous.

In the following pages, you will encounter a tapestry woven with threads from my scientific endeavours; my reflections on the world of artificial life, bioinformatics, and the intersections between biology and computation. From the creation of digital organisms to the quiet, often invisible labour of bioinformatics, each chapter reflects a moment in time when I was propelled by a sense of wonder, curiosity, and a desire to contribute meaningfully to the fields I care deeply about.

Yet this book is more than a recounting of scientific milestones. It is an exploration of the evolution of thought, of how one moves from consuming knowledge to enabling others to explore and discover. It’s a story of growth; personal, professional, and intellectual; as I grappled with the pressing questions of what it means to contribute to a community, to leave behind a legacy, and, ultimately, to find peace with the impact I’ve made.

While my scientific work has always been grounded in inquiry, it is the reflections and connections; those moments when knowledge and experience intersect; that have truly shaped who I am. Alongside the technical achievements, you’ll find meditations on purpose, humility, and the quiet lessons learned from the margins of academia and life.

I invite you to walk with me through these chapters, to consider the work that has shaped my thinking, and perhaps even to reflect on your own journey in science and life. My hope is that, in sharing these stories, I not only offer insight into the complexities of research and intellectual pursuits but also illuminate the value of the often-overlooked contributions that, in their own quiet way, are transformative.

This is my story – a story of questioning, learning, evolving, and, above all, seeking meaning in both the work and the life that surrounds it.

1: InterBase 6 Data Warehouse Builder

Citation: Ling, MHT and So, CW. 2003. Architecture of an Open-Sourced, Extensible Data Warehouse Builder: InterBase 6 Data Warehouse Builder (IB-DWB). In Rubinstein, B. I. P., Chan, N. and Kshetrapalapuram, K. K. (Eds.), Proceedings of the First Australian Undergraduate Students' Computing Conference. (pp. 40-45).

Abstract: We report the development of an open-sourced data warehouse builder, InterBase Data Warehouse Builder (IB-DWB), based on Borland InterBase 6 Open Edition Database Server. InterBase 6 is used for its low maintenance and small footprint. IB-DWB is designed modularly and consists of 5 main components, Data Plug Platform, Discoverer Platform, Multi-Dimensional Cube Builder, and Query Supporter, bounded together by a Kernel. It is also an extensible system, made possible by the Data Plug Platform and the Discoverer Platform. Currently, extensions are only possible via dynamic linked libraries (DLLs). Multi-Dimensional Cube Builder represents a basal mean of data aggregation. The architectural philosophy of IB-DWB centers around providing a base platform that is extensible, which is functionally supported by expansion modules. IB-DWB is currently being hosted by sourceforge.net (Project Unix Name: ib-dwb), licensed under GNU General Public License, Version 2.

Context: InterBase 6 Data Warehouse Builder (IB-DWB) was one of my earliest forays into software systems with a vision that extended beyond academic exercise. It was developed as part of my Advanced Diploma in Computing from Informatics Computer School which offers a distance-learning course that I undertook during 2001–2002. What stood out about this project – apart from the technology itself – was how it embodied a form of early digital collaboration before remote work became common. The entire project was executed virtually; my collaborator, So CW, and I never met in person. We communicated entirely over MSN Messenger and ICQ. I often wonder how unusual that must have seemed at the time, given that most group projects then still required face-to-face meetings.

IB-DWB was built on top of Borland's InterBase 6 Open Edition, selected for its low maintenance and small footprint. From a software engineering perspective, the project emphasized modularity and extensibility, the two principles that I would come to appreciate more deeply in later projects. It had five main components: the Data Plug Platform, the Discoverer Platform, the Multi-Dimensional Cube Builder, the Query Supporter, and the Kernel that tied everything together. What we built was not just a data tool, but a flexible architecture for building analytical systems, with support for plug-ins through dynamic link libraries (DLLs) – not an easy feat for diploma-level students at that time.

The project culminated in a co-authored publication, which was presented at the First Australian Undergraduate Students' Computing Conference in 2003. That paper turned out to be more than a line on my CV – it gave me bonus points in my PhD scholarship application and, in many ways, validated the seriousness with which I approached my work. It was the first time something I had built was archived in formal publication, and the thrill of that moment has never quite faded.

Reflection: In hindsight, the IB-DWB project was not just a technical milestone, it was also a personal one. It was the first time I saw myself not just as a student completing a requirement but as a creator of something potentially useful to others. I now realize that even at that early stage, I had begun grappling with architectural questions that still preoccupy me: How can a system remain simple, yet extensible? How do you build a platform that is not only functional but also future-proofed for community contribution?

The fact that I was doing this remotely, without ever meeting my co-author, also foreshadowed a recurring theme in my career: working with collaborators across distances, time zones, and sometimes disciplines. That early experience taught me the value of over-communication, version control (albeit rudimentary at the time), and the discipline required when working without in-person accountability. It is only now, with years of remote academic collaboration behind me, that I recognize how formative this experience was.

There is also something profoundly humbling about seeing my early code again or reading an old paper. The aspirations are evident, but so are the limitations of knowledge and experience at the time. Yet, I am grateful for that version of myself – the one who dared to build and publish, even if the work was imperfect. IB-DWB didn't go on to revolutionize the world of data warehousing, but it was a formative training ground. It introduced me to modular design, to the spirit of open source, and to the rewarding process of academic publishing.

Had I been given a chance to redo it, would I have changed anything? Technically, yes; many things. Conceptually, probably not. It was the right project at the right time, done with the tools and understanding I had. It opened a door to the world of scientific software development and academic dissemination. More importantly, it gave me the confidence to believe that I could build something of consequence. And that belief—planted in the early 2000s through this modest open-source project – has stayed with me throughout my scientific career.

Sidebar: A Proto-GitHub Before GitHub

When IB-DWB was first uploaded to SourceForge in 2002, platforms like GitHub didn't yet exist. Open-source collaboration was rudimentary, often relying on manual uploads, forum threads, and email patches. Yet, in that environment, this project managed to embrace a philosophy that now defines modern software development: modular architecture, remote teamwork, and open contribution.

The technical limitations were real – DLL-only extensibility, minimal automation, and fragile version control but the spirit of the project was well ahead of its time. In retrospect, IB-DWB wasn’t just a student project. It was an early lesson in building systems that outlast you, even if only in concept. It whispered the same principle that later projects would shout: Software isn’t just written; it’s architected for people you'll never meet.

2: Review of MontyLingua

Citation: Ling, MHT. 2006. An Anthological Review of Research Utilizing MontyLingua, a Python-Based End-to-End Text Processor. The Python Papers 1 (1): 5-12.

Abstract: MontyLingua, an integral part of ConceptNet which is currently the largest commonsense knowledge base, is an English text processor developed using Python programming language in MIT Media Lab. The main feature of MontyLingua is the coverage for all aspects of English text processing from raw input text to semantic meanings and summary generation, yet each component in MontyLingua is loosely-coupled to each other at the architectural and code level, which enabled individual components to be used independently or substituted. However, there has been no review exploring the role of MontyLingua in recent research work utilizing it. This paper aims to review the use of and roles played by MontyLingua and its components in research work published in 19 articles between October 2004 and August 2006. We had observed a diversified use of MontyLingua in many different areas, both generic and domain specific. Although the use of text summarizing component had not been observed, we are optimistic that it will have a crucial role in managing the current trend of information overload in future research.

Context: It was a quiet Saturday evening in University College where I served as a Residential Advisor, and I had just co-founded The Python Papers, a journal meant to bridge practitioners and academics in the growing Python programming community. As its associate editor, I needed content for the inaugural issue, and fast. I turned to something close to my heart and even closer to my research, MontyLingua.

MontyLingua was not just a Python-based English text processor. To me, it was a workhorse and a companion throughout my doctoral research. Developed at the MIT Media Lab and integrated into ConceptNet, it handled every layer of natural language processing from tokenization to semantic parsing. What set it apart was its modular, loosely coupled architecture that allowed researchers to use parts of the system independently. I started using it in 2004 at the earliest stage of my PhD, and it became foundational to how I handled unstructured text.

In under 24 hours, from Saturday evening to Sunday afternoon, I completed a review of 19 research articles that had utilized MontyLingua. This paper became one of the first anthological reviews on the topic and provided a snapshot of how such tools were beginning to shape computational linguistics. I noted that while the summarization module hadn't gained traction yet, the tool's flexibility was being explored in a wide range of research contexts, both general and domain-specific.

Reflection: Looking back, this review of MontyLingua holds a special place in my early academic career, not only for its subject matter but for the circumstances under which it was written. I was serving as a Residential Advisor at University College, University of Melbourne, and wrote the entire manuscript over a single weekend. In retrospect, this moment of intense productivity was driven by both youthful academic energy and a deep-seated curiosity about the burgeoning field of natural language processing (NLP).

MontyLingua itself was a remarkable tool of its time. Its modular architecture and Python implementation were rare qualities in 2006, long before Python became the dominant language for machine learning and NLP. The system offered an end-to-end pipeline, from raw text to semantic meaning, foreshadowing the integrated NLP frameworks we see today. The loosely coupled design of its components mirrored my own developing research philosophy: tools should be both powerful and interchangeable, fostering creativity through modularity.

This paper was also my first foray into meta-research, reviewing not a scientific method or result, but how a tool was being used across multiple domains. It taught me to appreciate software as both artifact and enabler. While MontyLingua's summarization module had yet to see uptake at the time of writing, I correctly anticipated the growing importance of text summarization in an era of exponential data growth.

In many ways, this review was not only a survey of existing literature, but a quiet prediction of the direction in which text processing research would evolve. The academic attention MontyLingua received during that brief window reflected both the limitations of the tools available then and the imaginative ways researchers repurposed existing systems.

This paper also marked my early belief in open-source software as an engine of research democratization – a theme that would recur across my later work. And though MontyLingua is now largely a historical footnote, I see its legacy reflected in today's ubiquitous NLP frameworks.

Sidebar: A Review Written in a Weekend, Rooted in a Doctorate

This paper wasn’t planned, it was a necessity. As co-founder and associate editor of The Python Papers, I needed an article to anchor our very first issue. So I turned to what I knew best: MontyLingua. In many ways, the paper was written quickly, but it was not superficial. It distilled two years of hands-on experience during the most formative stage of my PhD.

MontyLingua wasn’t just software, it was the core utility I used to tame unstructured text. By surveying its use in 19 published studies, I hoped to show that tools like this mattered beyond code — they quietly shaped how research was done. The paper became one of the earliest systematic reviews of MontyLingua's role in academic work, and unexpectedly, it's still cited today.

What began as a practical solution to an editorial gap became something more enduring: a personal snapshot of the tools that shaped my early research journey.

Sidebar: The Python Papers – A Journal Built on Enthusiasm

The Python Papers was born out of a simple idea: that Python was rapidly gaining traction in both academia and industry, yet few publications gave space to the people building with it. I co-founded the journal in 2006 with the goal of creating a platform that welcomed practitioners and researchers alike – a space where code and commentary could coexist.

It was a grassroots effort. We had no funding, no institutional support; just a team of volunteers, an editorial board, and the belief that the Python community deserved its own publication. Every submission, every review, every layout was a labor of love. I served as associate editor, contributing content, peer reviews, and outreach. It felt like building something from nothing.

Looking back, The Python Papers was ahead of its time. Long before today's open science and preprint movements, we were pushing for open access, open review, and community-driven scholarship. Even if the journal itself was short-lived, it taught me the value of publishing infrastructure and the quiet power of giving others a voice.

3: Serialized Firebird Backup

Citation: Ling, MHT. 2007. Firebird Database Backup by Serialized Database Table Dump. The Python Papers 2 (1): 12-16.

Abstract: This paper presents a simple data dump and load utility for Firebird databases which mimics *mysqldump* in MySQL. This utility, *fb_dump* and *fb_load*, for dumping and loading respectively, retrieves each database table using *kinterbasdb* and serializes the data using *marshal* module. This utility has two advantages over the standard Firebird database backup utility, *gbak*. Firstly, it is able to backup and restore single database tables which might help to recover corrupted databases. Secondly, the output is in text-coded format (from *marshal* module) making it more resilient than a compressed text backup, as in the case of using *gbak*.

Context: In 2007, while deeply immersed in my PhD research, I found myself relying heavily on the Firebird database system. Its lightweight design and robust features made it an ideal choice for my projects. However, I encountered a limitation: the standard backup utility, gbak, lacked the flexibility to back up and restore individual tables. This posed challenges, especially when dealing with corrupted databases where restoring specific tables could be invaluable.

Recognizing this gap, I developed a simple yet effective utility: fb_dump and fb_load. Written in Python, these tools utilized kinterbasdb to retrieve data and the marshal module to serialize it. The primary advantage was their ability to handle single-table backups and restores, offering a more granular approach compared to gbak. Additionally, the text-coded format provided by marshal ensured resilience against data corruption.

At the time, I was also an associate member of the Firebird Foundation, an organization dedicated to supporting the development and promotion of the Firebird database. Contributing to this utility was my way of giving back to a community that had provided me with invaluable tools and support.

The culmination of this work was the publication of this paper which detailed the design, implementation, and advantages of the utility, aiming to assist others facing similar challenges.

Reflection: This project stands out in my career as a testament to the intersection of necessity and community contribution. While the utility itself was straightforward, its impact was significant. It addressed a real-world problem I faced and, in doing so, provided a solution that others in the Firebird community could benefit from.

Developing fb_dump and fb_load reinforced my belief in the power of open-source tools and the importance of sharing solutions. It wasn't about creating something groundbreaking but about solving a specific problem effectively and making that solution accessible.

Moreover, this endeavour highlighted the value of community engagement. Being part of the Firebird Foundation and contributing back, even in a small way, fostered a sense of belonging and purpose. It reminded me that research and development are not solitary pursuits but collaborative efforts that thrive on shared knowledge and mutual support.

Looking back, I'm proud of this contribution. It may not have been a monumental achievement, but it encapsulated the essence of practical problem-solving and community involvement. It served as a reminder that meaningful impact often comes from addressing immediate needs with thoughtful, accessible solutions.

Sidebar: Backup as Philosophy

At first glance, fb_dump and fb_load were just backup toolsbut they were Python-based alternative to gbak. But in hindsight, they reflected a deeper philosophy I was beginning to cultivate: resilience through granularity. The ability to back up a single table, rather than an entire database, was not just a technical convenience. It was a quiet assertion that systems—scientific, computational, or institutional—\; should be repairable at fine resolution.

This work came from lived need: corrupted tables, urgent recoveries, and limited control. But it also subtly shaped how I approached systems thereafter. From digital organisms to student projects, I began designing with partial recovery in mind, always asking: What if only part of it breaks? What if we only need one piece back?

In retrospect, fb_dump wasn’t just a utility, it was a micro-manifesto. A reminder that sometimes, the most robust systems are those that honour the fragility of their parts.

4: Protein-Protein Interaction Network from Text

Citation: Ling, MHT, Lefevre, C, Nicholas, KR, Lin, F. 2007. Re-construction of Protein-Protein Interaction Pathways by Mining Subject-Verb-Objects Intermediates. In J.C. Ragapakse, B. Schmidt, and G. Volkert (Eds.), Proceedings of the Second IAPR Workshop on Pattern Recognition in Bioinformatics (PRIB 2007). Lecture Notes in Bioinformatics 4774. (pp. 286-299) Springer-Verlag.

Abstract: The exponential increase in publication rate of new articles is limiting access of researchers to relevant literature. This has prompted the use of text mining tools to extract key biological information. Previous studies have reported extensive modification of existing generic text processors to process biological text. However, this requirement for modification had not been examined. In this study, we have constructed Muscorian, using MontyLingua, a generic text processor. It uses a two-layered generalization-specialization paradigm previously proposed where text was generically processed to a suitable intermediate format before domain-specific data extraction techniques are applied at the specialization layer. Evaluation using a corpus and experts indicated 86-90% precision and approximately 30% recall in extracting protein-protein interactions, which was comparable to previous studies using either specialized biological text processing tools or modified existing tools. Our study had also demonstrated the flexibility of the two-layered generalization-specialization paradigm by using the same generalization layer for two specialized information extraction tasks.

Context: Back when this paper was written, the standard thinking in NLP was that biomedical text was so specialized that you couldn’t just use a generic text processor—adaptation was necessary. Muscorian challenged that assumption. Instead of tweaking an existing tool, we showed that an unmodified generic processor (MontyLingua) could extract protein-protein interactions just as effectively as adapted systems.

The key innovation was the two-layered generalization-specialization approach. First, a general processor structured unstructured text into a subject-verb-object format, providing a foundation. Then, specialized extraction methods could be applied based on specific needs. This setup proved surprisingly flexible – we could use the same generalization layer for two different information extraction tasks, showing that adaptation wasn’t always required.

This publication was the first from my PhD thesis, making it especially significant in my research career. It not only validated the effectiveness of the model but also set the stage for future studies on scalable NLP frameworks for biomedical text mining.

Reflection: Being my first PhD publication, this paper was a major milestone – it was the point where I transitioned from learning established methods to challenging them. The conventional wisdom at the time said specialized biomedical text required adaptation but Muscorian proved otherwise. It was a satisfying moment of pushing back against long-held assumptions and showing that structured generalization could be just as effective.

Fast forward nearly two decades, and the paper is still getting cited! I recently received a Google Scholar alert about a new citation, which was an unexpected but welcome surprise. It’s gratifying to see that the work continues to be relevant, reinforcing my confidence in the ideas I developed back then.

Beyond its technical contributions, this study shaped my overall approach to research-question assumptions, build scalable frameworks, and focus on adaptability. Those principles have stayed with me in both academic and applied contexts, making this paper not just an important publication, but also a defining moment in my scientific journey.

Sidebar: Muscorian – A Name, A Paradigm, A Rebellion
The name Muscorian was inspired by *Mus musculus*, the house mouse, a common model organism in biology, and also a nod to the modest ambitions of the system when it started. What began as a simple experiment in whether a generic NLP tool could be used for biomedical text mining quickly became something more: a challenge to the prevailing orthodoxy.

At the time, almost everyone believed that domain-specific adaptation was a prerequisite for biological text processing. But Muscorian flipped the script by showing that you could separate generic language structure from domain-specific extraction. This generalization-specialization paradigm didn't just save time, it opened the door to building scalable, reusable pipelines across domains.

Looking back, Muscorian wasn't just a system. It was a statement: that simplicity, when well-structured, could be powerful. It was also a quiet act of rebellion from a young researcher learning to trust his instincts.

5: Parts of Speech Tagging Error

Citation: Ling, MHT, Lefevre, C, Nicholas, KR. 2008. Parts-of-Speech Tagger Errors Do Not Necessarily Degrade Accuracy in Extracting Information from Biomedical Text. The Python Papers 3 (1): 65-80.

Abstract: An ongoing assessment of the literature is difficult with the rapidly increasing volume of research publications and limited effective information extraction tools which identify entity relationships from text. A recent study reported development of Muscorian, a generic text processing tool for extracting protein-protein interactions from text that achieved comparable performance to biomedical-specific text processing tools. This result was unexpected since potential errors from a series of text analysis processes is likely to adversely affect the outcome of the entire process. Most biomedical entity relationship extraction tools have used biomedical-specific parts-of-speech (POS) tagger as errors in POS tagging and are likely to affect subsequent semantic analysis of the text, such as shallow parsing. This study aims to evaluate the parts-of-speech (POS) tagging accuracy and attempts to explore whether a comparable performance is obtained when a generic POS tagger, MontyTagger, was used in place of MedPost, a tagger trained in biomedical text. Our results demonstrated that MontyTagger, Muscorian's POS tagger, has a POS tagging accuracy of 83.1% when tested on biomedical text. Replacing MontyTagger with MedPost did not result in a significant improvement in entity relationship extraction from text; precision of 55.6% from MontyTagger versus 56.8% from MedPost on directional relationships and 86.1% from MontyTagger compared to 81.8% from MedPost on nondirectional relationships. This is unexpected as the potential for poor POS tagging by MontyTagger is likely to affect the outcome of the information extraction. An analysis of POS tagging errors demonstrated that 78.5% of tagging errors are being compensated by shallow parsing. Thus, despite 83.1% tagging accuracy, MontyTagger has a functional tagging accuracy of 94.6%. The POS tagging error does not adversely affect the information extraction task if the errors were resolved in shallow parsing through alternative POS tag use.

Context: In the field of natural language processing (NLP), the common belief is that domain-specific texts require specialized tools to ensure accuracy. Biomedical text is particularly challenging, as subtle differences in terminology and structure can affect how information is extracted. Previous research showed that Muscorian, using the generic MontyLingua processor, could perform on par with specialized biomedical NLP tools but this paper dives deeper to understand why that's the case.

The study specifically examines POS tagging, an essential step in text processing. Typically, errors in POS tagging are thought to propagate through later stages, negatively affecting the final output. But here's where things get interesting while MontyTagger's accuracy on biomedical text is only 83.1% (compared to MedPost's

96.9%), it turns out that most errors don’t actually degrade the final information extraction task. Shallow parsing compensates for 78.5% of them, effectively raising MontyTagger’s functional accuracy to 94.6%.

This challenges a widely accepted assumption: that specialized adaptation is always necessary for biomedical text processing. Instead, MontyLingua’s structured approach allows it to handle errors efficiently, proving that well-designed frameworks can compensate for imperfections in individual components.

Reflection: This paper represents an important shift. It moves beyond showing that Muscorian works, towards understanding how and why it works. Initially, I expected MedPost, a specialized biomedical POS tagger, to significantly outperform MontyTagger. The results revealed that many errors did not negatively impact the final extraction process because shallow parsing corrected them downstream.

This study shaped my understanding of NLP systems. It taught me that accuracy in one step does not necessarily dictate overall performance. What matters is how different components interact and whether later processing stages can compensate for earlier imperfections. It also reinforced my belief that domain-specific adaptation is not always required. Sometimes, a well-structured methodology can allow a general tool to perform effectively in specialized contexts

6: Finding Potential Hypotheses for Lactation Research

Citation: Ling, MHT, Lefevre, C, Nicholas, KR. 2008. Filtering Microarray Correlations by Statistical Literature Analysis Yields Potential Hypotheses for Lactation Research. The Python Papers 3(3): 4.

Abstract: Recent studies have demonstrated that the cyclical nature of mouse lactation can be mirrored at the transcriptome level of the mammary glands but making sense of microarray results requires analysis of large amounts of biological information which is increasingly difficult to access as the amount of literature increases. Extraction of protein-protein interaction from text by statistical and natural language processing has shown to be useful in managing the literature. Correlations between gene expression across a series of samples is a simple method to analyze microarray data as it was found that genes that are related in functions exhibit similar expression profiles. Microarrays had been used to examine the transcriptome of mouse lactation and found that the cyclic nature of the lactation cycle as observed histologically is reflected at the transcription level. However, there has been no study to date using text mining to sieve microarray analysis to generate new hypotheses for further research in the field of lactational biology. Our results demonstrated that a previously reported protein name co-occurrence method (5-mention PubGene) which was not based on a hypothesis testing framework, is generally more stringent than the 99th percentile of Poisson distribution-based method of calculating co-occurrence. It agrees with previous methods using natural language processing to extract protein-protein interaction from text as more than 96% of the interactions found by natural language processing methods to coincide with the results from 5-mention PubGene method. However, less than 2% of the gene co-expressions analyzed by microarray were found from direct co-occurrence or interaction information extraction from the literature. At the same time, combining microarray and literature analyses, we derive a novel set of 7 potential functional protein-protein interactions that had not been previously described in the literature. We conclude that the 5-mention PubGene method is more stringent than the 99th percentile of Poisson distribution method for extracting protein-protein interactions by co-occurrence of entity names and literature analysis may be a potential filter for microarray analysis to isolate potentially novel hypotheses for further research.

Context: Understanding the transcriptome changes during lactation is crucial but making sense of the vast amount of microarray data is challenging. The sheer volume of biological literature makes it increasingly difficult for researchers to extract meaningful insights. While previous studies have demonstrated that the cyclical nature of lactation is reflected at the transcription level, no study had explored using text mining techniques to filter microarray results and generate new research hypotheses.

This study aimed to bridge that gap by combining statistical literature analysis with microarray correlations. We examined two text-mining approaches: the 5-mention PubGene method and a statistical Poisson-based method for extracting protein-protein interactions from text. Our results showed that the PubGene approach was more stringent, aligning well with existing natural language processing (NLP) methods for identifying interactions. However, less than 2% of microarray-derived gene co-expressions matched interactions already found in the literature. This suggests that a significant portion of potentially meaningful gene relationships remains undiscovered by conventional literature analysis.

By integrating both microarray correlations and literature mining, we identified seven novel potential protein-protein interactions that had not been previously described. This demonstrated that literature analysis could act as a useful filter, helping researchers isolate promising hypotheses for further study in lactational biology.

Reflection: This paper was particularly exciting because it was an early attempt to connect text mining with microarray analysis in a meaningful way. Instead of treating literature and experimental data as separate sources of information, we explored how they could complement each other to uncover new biological relationships. The findings challenged the assumption that conventional literature databases were sufficient for capturing all relevant protein interactions, showing that experimental data could highlight missing connections.

On a personal level, this study reinforced the value of interdisciplinary thinking. Combining NLP, statistical modeling, and biological research was not a common approach at the time, but it provided fresh insights into lactation biology. This work, funded by Dairy CRC, contributed to advancing methods for analyzing complex transcriptome data, and it was satisfying to see the impact of integrating computational techniques into biological research.

7: Statistical Distributions

Citation: Ling, MHT. 2009. Compendium of Distributions, I: Beta, Binomial, Chi-Square, F, Gamma, Geometric, Poisson, Student's t, and Uniform. The Python Papers Source Codes 1:4.

Abstract: This manuscript illustrates the implementation and testing of nine statistical distributions, namely Beta, Binomial, Chi-Square, F, Gamma, Geometric, Poisson, Student’s t and Uniform distribution, where each distribution consists of three common functions – Probability Density Function (PDF), Cumulative Density Function (CDF) and the inverse of CDF (inverseCDF).

Context: Statistical distributions play a crucial role in hypothesis testing and scientific analysis, providing probabilistic measures to interpret data and validate models. While many statistical tools exist, their implementation details often vary depending on the programming framework. This paper focuses on the implementation and testing of nine statistical distributions: Beta, Binomial, Chi-Square, F, Gamma, Geometric, Poisson, Student’s t, and Uniform. Each distribution is built with three essential functions: Probability Density Function (PDF), Cumulative Density Function (CDF), and inverse CDF.

At the time, standardized implementations of statistical distributions were scattered across different programming resources, and Python-based solutions were still evolving. This study aimed to bring structure to the process by systematically implementing these distributions and demonstrating their practical applications. It was inspired by earlier efforts in numerical computation, integrating concepts from established statistical references to create a cohesive framework for handling probability distributions.

Reflection: This project was a step toward consolidating statistical distribution implementations in Python. The goal was to create well-structured, reusable code that could support scientific analysis. Looking back, this study reflects my ongoing interest in bridging computational methods with practical applications. It reinforced the importance of organized programming for statistical analysis and highlighted the potential of Python for high-throughput data processing.

Beyond the technical aspects, this study marked a shift in my approach to programming. It was a move away from isolated calculations toward modular implementations that could be adapted for different research needs. Developing these functions systematically was a rewarding experience, making them accessible for broader applications.

8: Z-Test Routines

Citation: Ling, MHT. 2009. Ten Z-test Routines from Gopal Kanji's 100 Statistical Tests. The Python Papers Source Codes 1:5.

Abstract: This manuscript presents the implementation and testing of 10 Z-test routines from Gopal Kanji's book entitled "100 Statistical Tests".

Context: Statistical hypothesis testing is fundamental in scientific research, providing a framework for determining whether observed data supports or refutes a given assumption. Gopal Kanji's book, 100 Statistical Tests, is widely regarded as a practical reference for applied statistics, offering structured descriptions of various statistical methods. This paper focuses on implementing and testing ten Z-test routines from Kanji's work, providing Python-based solutions for hypothesis testing.

These tests rely on the standardized normal distribution, where the mean is zero and the variance is one. The implementation includes a statistical test harness that standardizes hypothesis testing routines, allowing for both one-tailed and two-tailed tests. This ensures consistent reporting of critical values and test statistic evaluations. The study provides code for ten specific Z-tests, covering population means, proportions, Poisson-distributed counts, Pearson's correlation coefficients, and Spearman rank correlations.

At the time, Python was gaining traction as a statistical computing tool but implementations for specific hypothesis tests were still evolving. This work contributed by offering structured, reusable implementations that could be incorporated into broader statistical packages, making hypothesis testing more accessible to researchers and analysts.

Reflection: This project was an important step in bridging theoretical statistics with practical programming. While Kanji's book provided a structured reference, implementing the tests in Python allowed for automation and reproducibility, making statistical analysis more efficient.

Beyond the technical aspects, this work reflects my continued interest in computational methods for research. It reinforced the importance of modular programming, ensuring that statistical tools can be adapted and integrated into various applications. Seeing these tests implemented in Python was satisfying, as it contributed to making hypothesis testing more accessible to users who may not have a deep statistical background but need reliable methods for analysis.

9: Review of Biomedical Literature Analysis

Citation: Ling, MHT, Lefevre, Christophe, Nicholas, KR. 2009. Biomedical Literature Analysis: Current State and Challenges. In B.G. Kutais (ed). Internet Policies and Issues, Volume 7. Nova Science Publishers, Inc.

Abstract: Advances in molecular biology tools and techniques from the end of the last century had shifted the focus of biomedical research from the study of individual proteins and genes to the interactions within an entire biological systems. At the same time, advanced tools generate large sets of experimental data which required collaborations of groups of biologists to decipher. This resulted in a need to have a diverse research knowledge. However, the amount of published research information in the form of published articles is increasing exponentially, making it difficult to maintain a productive edge. Biomedical literature analysis is seen as a means to manage the increased amount of information – to gather relevant articles and extract relevant information from these articles. We review the central (information retrieval, information extraction and text mining) and allied (corpus collection, databases and system evaluation methods) domains of computational biomedical literature analysis to present the current state of biomedical literature analysis for protein-protein and protein-gene interactions and the challenges ahead.

Context: Biomedical literature analysis emerged as a response to the exponential growth of published research. As molecular biology transitioned from studying individual genes and proteins to broader biological systems, the need to process vast amounts of literature became critical. The volume of publications had surpassed what researchers could manually review, making computational methods essential.

This chapter of my PhD thesis provides a literature review on the state of biomedical literature analysis at the time. It covers key areas like information retrieval, information extraction, and text mining – technologies designed to help researchers locate relevant publications and extract meaningful relationships between biological entities. Additionally, it discusses supporting domains, such as corpus collection, databases, and system evaluation, which are crucial for refining text-mining approaches.

At this stage of my career, I was in my second year as a lecturer at Singapore Polytechnic. I had taken a two-year leave of absence from my PhD to focus on publishing the necessary papers, ensuring that my research was well-documented and accessible before moving forward with my thesis.

Reflection: This work was a deep dive into the computational aspects of biomedical literature analysis. While my earlier papers focused on specific methodologies, this chapter broadened the scope to provide a comprehensive review of the field. It reinforced the importance of integrating NLP and statistical models into biomedical research, ensuring that vast amounts of literature could be systematically analyzed rather than relying on manual review.

Looking back, this chapter represents an important phase in my academic journey. Taking time away from my PhD to focus on publications was a strategic decision as it allowed me to contribute meaningful work to the field and establish a strong foundation for my thesis. At the same time, balancing research with teaching at Singapore Polytechnic provided a new perspective on knowledge dissemination, reinforcing the importance of clarity and accessibility in scientific communication

10: BIOADI

Citation: Kuo, CJ, Ling, MHT, Lin, KT, Hsu, CN. 2009. BIOADI: A Machine Learning Approach to Identify Abbreviations and Definitions in Biological Literature. BMC Bioinformatics 10(Suppl 15): S7.

Abstract: To automatically process large quantities of biological literature for knowledge discovery and information curation, text mining tools are becoming essential. Abbreviation recognition is related to NER and can be considered as a pair recognition task of a terminology and its corresponding abbreviation from free text. The successful identification of abbreviation and its corresponding definition is not only a prerequisite to index terms of text databases to produce articles of related interests, but also a building block to improve existing gene mention tagging and gene normalization tools. Our approach to abbreviation recognition (AR) is based on machine-learning, which exploits a novel set of rich features to learn rules from training data. Tested on the AB3P corpus, our system demonstrated a F-score of 89.90% with 95.86% precision at 84.64% recall, higher than the result achieved by the existing best AR performance system. We also annotated a new corpus of 1200 PubMed abstracts which was derived from BioCreative II gene normalization corpus. On our annotated corpus, our system achieved a F-score of 86.20% with 93.52% precision at 79.95% recall, which also outperforms all tested systems. By applying our system to extract all short form-long form pairs from all available PubMed abstracts, we have constructed BIOADI. Mining BIOADI reveals many interesting trends of bio-medical research. Besides, we also provide an off-line AR software in the download section on http://bioagent.iis.sinica.edu.tw/BIOADI/.

Context: Biomedical literature is packed with abbreviations, making it challenging for researchers to extract meaningful information efficiently. Abbreviation recognition is a key component of natural language processing (NLP), helping link terminology to their corresponding short forms, improving searchability, and enhancing data curation. This study developed BIOADI, a machine-learning-based abbreviation recognition system designed to automate this process with high accuracy.

The approach leveraged a rich set of linguistic features to train a model capable of identifying abbreviation-definition pairs with greater precision than previous systems. Tested on the AB3P corpus, BIOADI achieved an F-score of 89.90%, outperforming existing state-of-the-art abbreviation recognition tools. Additionally, a new annotated corpus of 1200 PubMed abstracts was created for further validation, where BIOADI achieved an F-score of 86.20%.

Beyond its technical contributions, this project was unique in its collaborative nature. Developed remotely with Academia Sinica in Taiwan, all communication for the research was conducted through email and MSN Messenger, a testament to the power of digital collaboration in research. This work was presented at InCoB 2009, a significant milestone in my academic journey, and coinciding with the day my brother, Melvin, enlistment into the army.

Reflection: This paper was special, not just for its technical contributions, but for the way it came together. Collaborating remotely with Academia Sinica pushed the boundaries of teamwork in research, showing that meaningful scientific work could be done without face-to-face meetings. It was a rewarding experience – seeing BIOADI outperform established abbreviation recognition tools validated the effort that went into designing its machine-learning framework.

Presenting this work at InCoB 2009 was another key moment. Conferences provide an opportunity to engage with the research community, and sharing BIOADI's results reinforced the importance of tackling practical problems in biomedical literature analysis. The timing was also personally significant, as it coincided with a major life event for my brother, making the day even more memorable.

Looking back, this study reinforced my interest in applying NLP to biological text processing. It showed that well-designed machine-learning models can substantially improve literature analysis, making information extraction more efficient and scalable.

Sidebar: BIOADI and the MSN Messenger Era
The success of BIOADI wasn't just a win for machine learning, it was a quiet triumph of asynchronous, transnational collaboration. In an era before Zoom and Slack became research staples, the entire project was coordinated over email and MSN Messenger. Code, corpora, ideas, even bug fixes; all exchanged across time zones, one message at a time.

This wasn't just about building an abbreviation recognition tool. It was about proving that even complex NLP systems could be developed remotely, by people who had never met in person. The experience foreshadowed today's distributed research labs and global open-source communities.

In a way, BIOADI was a product of its time: shaped by early digital tools, driven by a belief in collaboration without borders, and built to handle the very chaos of language that defined the biomedical literature of the 2000s.

11: My Doctoral Dissertation

Citation: Ling, MHT. 2009. Understanding Mouse Lactogenesis by Transcriptomics and Literature Analysis. Doctor of Philosophy. Department of Zoology, The University of Melbourne, Australia.

Abstract: The mammary explant culture model has been a major experimental tool for studying hormonal requirements for milk protein gene expression as markers of secretory differentiation. Experiments with mammary explants from pregnant animals from many species have established that insulin, prolactin, and glucocorticoid are the minimal set of hormones required for the induction of maximal milk protein gene expression. However, the extent to which mammary explants mimic the response of the mammary gland *in vivo* is not clear. Recent studies have used microarray technology to study the transcriptome of mouse lactation cycle. It was demonstrated that each phase of mouse lactation has a distinct transcriptional profile but making sense of microarray results requires analysis of large amounts of biological information which is increasingly difficult to access as the amount of literature increases.

The first objective is to examine the possibility of combining literature and genomic analysis to elucidate potentially novel hypotheses for further research into lactation biology. The second objective is to evaluate the strengths and limitations of the murine mammary explant culture for the study and understanding of murine lactogenesis. The underlying question to this objective is whether the mouse mammary explant culture is a good model or representation to study mouse lactogenesis.

The exponential increase in publication rate of new articles is limiting access of researchers to relevant literature. This has prompted the use of text mining tools to extract key biological information. Previous studies have reported extensive modification of existing generic text processors to process biological text. However, this requirement for modification had not been examined. We have constructed Muscorian, using MontyLingua, a generic text processor. It uses a two-layered generalization-specialization paradigm previously proposed where text was generically processed to a suitable intermediate format before domain-specific data extraction techniques are applied at the specialization layer. Evaluation using a corpus and experts indicated 86-90% precision and approximately 30% recall in extracting protein-protein interactions, which was comparable to previous studies using either specialized biological text processing tools or modified existing tools. This study also demonstrated the flexibility of the two-layered generalization-specialization paradigm by using the same generalization layer for two specialized information extraction tasks.

The performance of Muscorian was unexpected since potential errors from a series of text analysis processes is likely to adversely affect the outcome of the entire pro-

cess. Most biomedical entity relationship extraction tools have used biomedical-specific parts-of-speech (POS) tagger as errors in POS tagging and are likely to affect subsequent semantic analysis of the text, such as shallow parsing. A comparative study between MontyTagger, a generic POS tagger, and MedPost, a tagger trained in biomedical text, was carried out. Our results demonstrated that MontyTagger, Muscorian's POS tagger, has a POS tagging accuracy of 83.1% when tested on biomedical text. Replacing MontyTagger with MedPost did not result in a significant improvement in entity relationship extraction from text; precision of 55.6% from MontyTagger versus 56.8% from MedPost on directional relationships and 86.1% from MontyTagger compared to 81.8% from MedPost on un-directional relationships. This is unexpected as the potential for poor POS tagging by MontyTagger is likely to affect the outcome of the information extraction. An analysis of POS tagging errors demonstrated that 78.5% of tagging errors are being compensated by shallow parsing. Thus, despite 83.1% tagging accuracy, MontyTagger has a functional tagging accuracy of 94.6%. This suggests that POS tagging error does not adversely affect the information extraction task if the errors were resolved in shallow parsing through alternative POS tag use.

Microarrays had been used to examine the transcriptome of mouse lactation and a simple method for microarray analysis is correlation studies where functionally related genes exhibit similar expression profiles. However, there has been no study to date using text mining to sieve microarray analysis to generate new hypotheses for further research in the field of lactational biology. Our results demonstrated that a previously reported protein name co-occurrence method (5-mention PubGene) which was not based on a hypothesis testing framework, is generally more stringent than the 99th percentile of Poisson distribution-based method of calculating co-occurrence. It agrees with previous methods using natural language processing to extract protein-protein interaction from text as more than 96% of the interactions found by natural language processing methods coincide with the results from 5-mention PubGene method. However, less than 2% of the gene co-expressions analyzed by microarray were found from direct co-occurrence or interaction information extraction from the literature. At the same time, combining microarray and literature analyses, we derive a novel set of 7 potential functional protein-protein interactions that had not been previously described in the literature. We conclude that the 5-mention PubGene method is more stringent than the 99th percentile of Poisson distribution method for extracting protein-protein interactions by co-occurrence of entity names and literature analysis may be a potential filter for microarray analysis to isolate potentially novel hypotheses for further research.

The availability of transcriptomics data from time-course experiments on mouse mammary glands examined during the lactation cycle and hormone-induced lactogenesis in mammary explants has permitted an assessment of similarity of gene expression at the transcriptional level. Global transcriptome analysis using exact Wilconox signed-rank test with continuity correction and hierarchical clustering of Spearman coefficient demonstrated that hormone-induced mammary explants behave differently to mammary glands at secretory differentiation. Our results demonstrated that the mammary explant culture model mimics *in vivo* glands in immediate responses, such as hormone-responsive gene transcription, but generally did not mimic responses to prolonged hormonal stimulus, such as the extensive development of secretory pathways and immune responses normally associated with lactating mammary tissue. Hence, although the explant model is useful to study the immediate effects of stimulating secretory differentiation in mammary glands, it is unlikely to be suitable for the study of secretory activation.

Context: This dissertation represents the culmination of my PhD research, synthesizing years of work into a comprehensive study on mouse lactogenesis. The focus is twofold: first, to explore how integrating text mining and genomic data can generate novel hypotheses for lactation biology, and second, to evaluate the effectiveness of the mammary explant culture model in studying lactogenesis.

At the time, the mammary explant model was widely used to study hormonal regulation of milk protein gene expression. While experiments had established that insulin, prolactin, and glucocorticoid were necessary for inducing maximal milk protein production, whether mammary explants truly mimicked *in vivo* gland responses remained an open question. My research leveraged microarray transcriptomic data alongside computational methods to assess how explants compared to actual lactating mammary glands.

Additionally, I explored the role of text mining in managing the growing volume of biomedical literature. By developing Muscorian, a text-processing framework, I examined whether a generic NLP tool could extract biological relationships just as effectively as domain-specific tools. Results from comparative studies showed that errors in POS tagging did not significantly impact information extraction, as shallow parsing compensated for most inaccuracies. This challenged the assumption that adaptation for biomedical text processing was always necessary.

The study also combined microarray correlation analysis with literature mining, resulting in the identification of seven potential protein-protein interactions that had not been previously described. This demonstrated how computational filtering methods could isolate meaningful hypotheses, reinforcing the value of integrating statistical models and literature mining in biological research.

Reflection: This dissertation reflects a pivotal phase in my scientific career – bringing together computational methods, experimental biology, and data analysis into a unified research framework. It was an ambitious project, but it allowed me to push boundaries by questioning assumptions in both biomedical NLP and lactation research.

One of the most impactful takeaways from this work was realizing that computational approaches can complement biological research in unexpected ways. Whether it was demonstrating that POS tagging errors were less detrimental than expected or showing that text mining could highlight overlooked gene interactions, this study reinforced my belief in interdisciplinary problem-solving.

Beyond the technical aspects, completing this dissertation was also a deeply personal milestone. By this time, I was in my second year as a lecturer at Singapore Polytechnic, balancing teaching with research. Taking time away from my PhD to focus on publishing necessary papers was a strategic decision that ultimately strengthened my work. Looking back, I am proud of the way this study blended structured methodology with innovative perspectives, setting the foundation for much of my future academic interests.

Sidebar: The Birth of Muscorian

Muscorian was not just another bioinformatics tool, it embodied a philosophical stance. At a time when most researchers were retrofitting generic NLP engines for biomedical use, I asked a different question: What if we didn’t have to? By adhering to a two-layered generalization-specialization paradigm, Muscorian allowed a generic NLP tool, MontyLingua, to stand on equal footing with domain-specific systems. Its surprisingly strong performance proved that domain adaptation isn't always necessary if architectural flexibility and robust parsing strategies are employed. Muscorian became my rebuttal to the dogma that biomedical text required biomedical tools.

Sidebar: Seven Hypotheses from Silence

Perhaps the most exciting outcome of this dissertation was the emergence of seven novel protein-protein interactions, inferred not from the noisy abundance of microarray data alone, but from the intersections between literature mining and gene co-expression. These were interactions that had eluded both biologists and algorithms before. In a field overwhelmed by data, what we needed wasn’t more information – it was better filters. This experience shaped my long-term vision: to make the invisible visible, not by accumulating more data, but by connecting what already exists more intelligently.

Sidebar: Functional Accuracy – A Conceptual Breakthrough

A key insight from my dissertation was the realization that tagging accuracy isn't everything – functional accuracy matters more. MontyTagger’s 83.1% POS tagging accuracy on biomedical text seemed suboptimal, yet Muscorian’s downstream performance remained strong. Why? Because shallow parsing compensated for 78.5% of those errors. The lesson: In complex systems, robustness often emerges from redundancy and error tolerance. This insight resonated with biological systems too—where feedback loops and compensatory mechanisms maintain function despite noise.

Sidebar: A Study in Contrasts – Explants vs. *In vivo*

My evaluation of mammary explants uncovered a critical divergence: while explants effectively modeled immediate hormonal responses, they failed to replicate the long-term immune and secretory changes of lactating tissue. This nuance was important. It challenged the assumption that explant models were fully representative, advocating instead for contextual validity in experimental design. It taught me that models are powerful not because they replicate reality, but because they clarify which aspects of reality they can and cannot emulate.

Sidebar: Publishing First, Dissertating Later

This dissertation didn’t follow the conventional arc. By 2009, I had already published several peer-reviewed papers derived from this work. I made the conscious decision to focus on publishing first, compiling later – a move that ensured rigor and peer feedback at every stage. It also made the final write-up more like curating a story than writing a thesis. This reverse order has influenced how I structure projects ever since: publish as you go, synthesize when ready.

12: MARK3 as Reference Gene

Citation: Chia, CY, Lim, CWX, Leong, WT, Ling, MHT. 2010. High Expression Stability of Microtubule Affinity Regulating Kinase 3 (MARK3) Makes It a Reliable Reference Gene. IUBMB Life 62(3):200-203.

Abstract: Difference in gene expressions is characteristic of the function of different cell types and those genes with low expression variance can be used as standards for quantitative gene expression studies. Microarray technology is used to study global gene expression within a cell; hence, represents a suitable source of data to mine for genes with low expression variance. The coefficient of variation (COV) of each gene was determined and a threshold of less than 0.1 COV was used to select stably expressed genes in each data set. Our results showed that microtubule affinity-regulating kinase 3 (MARK3) has the lowest COV in eight microarray datasets. In addition, the gene expression of housekeeping genes, which is very likely to be stably expressed, tends to fluctuate highly under different conditions, marking them as being less reliable for use as reference genes.

Context: In gene expression studies, stable reference genes are crucial for ensuring accurate comparisons across different conditions. Traditionally, housekeeping genes have been used for this purpose under the assumption that their expression remains constant. However, previous studies suggested that housekeeping genes like GAPDH and beta-actin fluctuate significantly, making them unreliable standards.

This study aimed to identify more stable reference genes by analyzing eight microarray datasets from mouse liver. Using coefficient of variation (COV) as a metric, we determined which genes exhibited the least variability in expression across different conditions. The results showed that MARK3 (Microtubule Affinity Regulating Kinase 3) had the lowest COV (<0.08), making it the most stable gene in the dataset. In contrast, housekeeping genes displayed inconsistent expression patterns, reinforcing the need for more rigorous selection criteria when choosing reference genes.

Beyond the scientific findings, this paper was unique because it was co-authored by three high-school students from the Gifted Education Programme under the Ministry of Education, Singapore. At just 15 or 16 years old, they successfully contributed to a peer-reviewed publication, making them the youngest students I had worked with at that point.

Reflection: This paper was special not just for its scientific contributions but for the mentorship aspect. Having high-school students publish a peer-reviewed paper was a major milestone, demonstrating that meaningful research is possible even at a young age with the right guidance. It was an incredibly rewarding experience – seeing these students develop confidence in their scientific abilities, navigate complex data analysis, and ultimately contribute to a journal publication.

On the technical side, this study reinforced my habit of questioning established assumptions. Housekeeping genes had long been considered the gold standard for reference gene selection, but our findings challenged that notion. Identifying a more stable alternative like MARK3 helped refine the methodology for gene expression studies, emphasizing the importance of data-driven selection rather than relying on traditional choices.

Looking back, this paper represents the intersection of mentorship, scientific inquiry, and challenging assumptions – the three aspects that have shaped my approach to research over the years.

Sidebar: When the Young Lead the Way

The MARK3 paper wasn't just a methodological refinement, it was a statement. A statement that teenagers, if mentored with trust and treated as intellectual equals, can produce work that stands shoulder-to-shoulder with seasoned researchers. The three students, barely old enough to drive, entered the world of microarray analysis, data mining, and scientific publishing with enthusiasm, rigor, and resilience. Their youth was never a limitation; if anything, it was their edge.

In many ways, this project was my quiet rebellion against elitist notions of who gets to do science. The results were clear: scientific insight doesn't always correlate with seniority. This experience remains one of the most powerful affirmations of the value of inclusive mentorship in research.

Sidebar: Rethinking the Gold Standard

For decades, housekeeping genes like GAPDH and beta-actin were used almost reflexively as reference genes in gene expression studies. Their presumed stability made them the "gold standard" but this paper helped reveal the cracks. By using microarray datasets and applying a simple yet powerful metric like the coefficient of variation, we showed that these trusted genes often fluctuate more than expected under differing conditions.

MARK3's exceptional stability (<0.08 COV across eight datasets) presented a compelling alternative. More importantly, it highlighted a broader principle: reference genes should be selected empirically, not traditionally. This shift from assumption to evidence continues to inform best practices in gene expression analysis today.

13: Bactome I

Citation: Lee, CH, Lee, KC, Oon, JSH, Ling, MHT. 2010. Bactome, I: Python in DNA Fingerprinting. In: Peer-Reviewed Articles from PyCon Asia-Pacific 2010. The Python Papers 5(3):6.

Abstract: Bactome is a set of functions created for our analysis of DNA fingerprints. This includes functions to find suitable primers for PCR-based DNA fingerprinting given a known genome, determine restriction digestion profile, and analyse the resulting DNA fingerprint features as migration distance of the bands in gel electrophoresis.

Context: This paper marked the very beginning of Bactome, a personal and collaborative effort to apply Python programming to the analysis of DNA fingerprints. It emerged from my supervision of my very first Final Year Project (FYP) group at Singapore Polytechnic. Our shared curiosity was how Python could be used to develop a pipeline for DNA fingerprinting analysis, from primer selection based on known genomes, through restriction digestion, to the interpretation of gel electrophoresis results.

The timing of this work was significant. In 2010, we presented this paper at the inaugural PyCon Asia-Pacific, held in Singapore. I was not only a presenter but also a co-founder of this conference, along with Michael Li and Liew Beng Keat (Assistant Director at Republic Polytechnic). In preparation for hosting PyCon Asia-Pacific, we established the Python User Group (Singapore) as a formal society, a move that provided the necessary institutional framework to support Python-related community and educational activities in Singapore.

This paper is special in that it wasn't merely about a technical solution; it was a declaration of intent. It marked the start of a longer-term vision: Bactome, a modular open-source library that would eventually grow into a broader framework for microbial genome analysis. The repository we created then still exists today (https://github.com/mauriceling/bactome), reflecting that early vision of bringing computational tools to the frontlines of biological discovery, especially in the realm of bacterial genomics.

Reflection: Looking back, this work encapsulates the thrill of starting something new – pedagogically, scientifically, and communally. It was the first time I guided a student team through the full arc of scientific inquiry: from idea to implementation to publication and public presentation. The students' engagement was deeply encouraging, and their fresh perspectives helped shape the simplicity and usability of the original Bactome functions.

From a broader lens, this paper is a crossroads. It sits at the intersection of my deepening commitment to student mentorship, open-source software, and bioinformatics. It was also one of my first formal forays into community-building within the Python ecosystem, both through the conference and the Python User Group. These experiences laid the groundwork for much of what followed in both my scientific and educational work.

Bactome I, while humble in scope, seeded several future developments. It reaffirmed my belief that scientific software should be lightweight, understandable, and user-driven. It also reminded me how impactful it is to enable young scientists to publish early, to take their first steps into a world where science is not only consumed but created.

This chapter celebrates the confluence of youth, curiosity, and code; and the unassuming beginnings of a project that continues to evolve in purpose and potential.

14: CyNote

Citation: Ng, YY and Ling, MHT. 2010. Electronic Laboratory Notebook on Web2Py Framework. In: Peer-Reviewed Articles from PyCon Asia-Pacific 2010. The Python Papers 5(3): 7.

Abstract: This paper presents CyNote version 1.4 as a prototype of an electronic laboratory notebook that is built on Web2py framework. CyNote uses a blog-style structure (entries and comments) as laboratory notebook and had implemented a number of bioinformatics and statistical analysis functions. At the same time, this paper evaluates CyNote against US FDA 21 CFR Part 11.

Context: This paper presented CyNote, a prototype electronic laboratory notebook (ELN) built using the Web2Py framework. CyNote adopted a familiar blog-style interface of entries and comments to emulate laboratory note-taking. The idea was also to incorporate built-in modules for bioinformatics (e.g. primer design, sequence alignment) and statistical analysis, making it a hybrid between ELN and analytical workbench. Significantly, it was also evaluated against US FDA 21 CFR Part 11, reflecting a serious attempt at addressing regulatory considerations.

The backstory to this project is deeply human. Yong Yao, the student co-author, was a Singapore Polytechnic student who needed to repeat a failed module. Rather than let him fall behind his cohort, Dr. Thomas Chai, then Director of the School of Chemical and Life Sciences, allowed him to do an in-house internship with me. What began as an academic patch-up turned into something transformative.

Yong Yao threw himself into the project with commitment. I had him post questions to the Web2Py Google Group, where we received generous support from Massimo Di Pierro, the creator of Web2Py. Massimo, a professor at DePaul University, showed an educator’s heart; encouraging and patient. I coached Yong Yao on humility and clarity in communication, occasionally supplementing his posts to ensure precision but he grew rapidly into both the technical and social dimensions of open-source development.

Behind the scenes, this was also the point where my fascination with electronic laboratory notebooks began to crystallize. I had grown increasingly concerned with the fragility of paper notebooks, the limitations of Microsoft Word, and the lack of reproducibility in digital science. CyNote was an early, student-driven attempt to address these pain points using free and open-source infrastructure. According to the GitHub wiki I later authored, CyNote aimed to “serve as a ‘light-weight but useful’ tool that bridges the gap between lab documentation and data analysis, all from the web browser.”

Reflection: CyNote holds special significance for me. It is the first manifestation of my interest in ELNs, and it emerged under unlikely circumstances. What makes this story remarkable is not the software itself, but the process – how a student on academic probation, empowered with trust and mentorship, helped shape a tool that would later inspire more mature systems.

The project instilled in me a deep appreciation for open-source mentorship communities. Without the support of the Web2Py ecosystem and Massimo Di Pierro in particular, this project might have faltered. It reminded me that educational transformation often comes not from grand designs, but from the cumulative effects of small acts: a director's trust, a professor's forum reply, a student's earnest effort.

CyNote also seeded a question that still lingers in my work today: What does it mean to do reproducible science? Beyond software, this paper represented a moment when I started to rethink scientific documentation not just as a record of work, but as a system that could and should be designed.

Yong Yao graduated with his cohort. CyNote never became a widely adopted ELN, but in some ways, it didn't have to. It served its purpose – as a prototype, as a proof-of-concept, and as a personal turning point for both student and mentor.

Sidebar: The Power of Mentorship and Open-Source Communities
CyNote's journey is a testament to how mentorship and the open-source community can catalyze meaningful change. Yong Yao, initially a student struggling academically, was given an opportunity to thrive through an internship that evolved into a significant contribution to scientific infrastructure. The project's success hinged not only on his dedication but also on the guidance and patience of individuals like Dr. Thomas Chai and Massimo Di Pierro.

What stands out in this story is how small acts of support can transform both a student's trajectory and the development of new ideas. For example, Massimo's guidance on Web2Py not only helped solve technical problems but also taught Yong Yao the power of clear communication in a collaborative environment. These interactions demonstrate how open-source communities can be incredibly nurturing, offering a space where both the mentor and mentee can learn and grow.

The lesson here isn't just about developing a product but about fostering an environment where people are trusted to explore, make mistakes, and ultimately succeed. CyNote, though never widely adopted, represents a major milestone in both my personal journey and that of my student. It's a reminder that sometimes, the real impact of a project is not just its final product but the way it shapes the people involved in its creation.

15: Collection of Distance Measures

Citation: Ling, MHT. 2010. COPADS, I: Distances Measures between Two Lists or Sets. The Python Papers Source Codes 2:2.

Abstract: This paper implements 35 distance coefficients with worked examples: Jaccard, Dice, Sokal and Michener, Matching, Anderberg, Ochiai, Ochiai 2, First Kulcsynski, Second Kulcsynski, Forbes, Hamann, Simpson, Russel and Rao, Roger and Tanimoto, Sokal and Sneath, Sokal and Sneath 2, Sokal and Sneath 3, Buser, Fossum, Yule Q, Yule Y, McConnaughey, Stiles,Pearson, Dennis, Gower and Legendre, Tulloss, Hamming, Euclidean, Minkowski, Manhattan, Canberra, Complement Bray and Curtis, Cosine, Tanimoto.

Context: The Collection of Python Algorithms and Data Structures (COPADS) began not as a grand vision, but as a personal, needs-driven toolkit – a kind of digital toolbox for a computational biologist. I imagined it as something humble but essential: a place to gather reusable implementations of data structures and algorithms in Python, especially those that could serve well in biological data management.

COPADS took root from three converging inspirations. The first was the Handbook of Data Structures and Applications – a comprehensive catalog that made me wonder, “What if we had Python versions of these?” The second was Numerical Recipes, whose elegant implementations in C and Fortran made me curious about Pythonic equivalents. And the third influence came from the Python Cookbook and its growing online presence at ActiveState. These “recipes” hinted at the value of sharing well-tested solutions to common problems.

In that sense, COPADS was not a traditional research project but rather a foundational and developmental one. It didn’t ask a scientific question; it answered many small engineering ones. Could this data structure be implemented in Python efficiently? Can we reuse this algorithm in other domains? Would a common interface for similarity metrics help standardize analysis workflows?

The first formal output of COPADS, this paper, focused on distance and similarity measures between lists and sets. It implemented 35 such metrics, from household names like Jaccard and Cosine to lesser-known coefficients like Fossum, Dennis, or Kulcsynski’s variants. Every function was accompanied by a worked example, not just to verify correctness but to promote understanding. For me, it was a way to say: “Here is not just how, but also why you might use this.”

And of course, in the spirit of “eat your own dog food”, I made a conscious effort to incorporate COPADS into other projects and sometimes shamelessly branding outputs with a virtual “Powered by COPADS”.

Reflection: In retrospect, COPADS represents something deeply personal: my belief in craftsmanship in science. Just as a skilled artisan builds and maintains their own tools, I wanted a curated, reusable toolkit that could serve me and possibly others across projects. This was especially relevant in bioinformatics, where scripts and ad-hoc solutions often mushroom into unmaintainable messes.

While COPADS I was about similarity metrics, the real statement it made was broader: that building good tools is a legitimate part of scientific contribution. They may not publish well or attract citations, but they make research possible. COPADS was never meant to compete with SciPy or scikit-learn — it was smaller, scrappier, and unashamedly niche. It was about what I needed, and what I thought others might, too.

Over time, COPADS didn’t grow into the community project I once fantasized about. But the spirit of it lives on in the way I think about software: as infrastructure for ideas. Every time I reused a COPADS module or function in another project, I was reminded that sometimes the most valuable work isn’t flashy – it’s foundational.

16: Restriction Mapped Genetic Distances

Citation: Chay, ZE, Lee, CH, Lee, KC, Oon, JSH, Ling, MHT. 2010. Russel and Rao Coefficient is a Suitable Substitute for Dice Coefficient in Studying Restriction Mapped Genetic Distances of *Escherichia coli*. iConcept Journal of Computational and Mathematical Biology 1:1.

Abstract: Dice coefficient (also known as Nei and Li coefficient) had been commonly used as a measure of genetic similarity from DNA fingerprints. This manuscript examines 19 other coefficients for its suitability. Our results suggest that Dennis, Fossum, Matching and Russel and Rao to work as well or better than Dice. Dennis, Matching and Fossum coefficients had highest discriminatory abilities but are limited by the lack of upper or lower boundaries. Russel and Rao coefficient is highly correlated with Dice coefficient (r^2 = 0.998), with both higher and lower boundaries, suggesting that Russel and Rao coefficient can be used to substitute Dice coefficient in studying genetic distances in *E. coli*.

Context: This paper can be seen as a direct continuation of two earlier threads: Bactome I, where we explored DNA fingerprints as genomic barcodes, and COPADS I, which provided the computational engine for computing 35 distance metrics. Naturally, the next question emerged: With these tools and data in hand, which similarity measures actually work best in a biological context?

In the world of molecular genetics, the Dice coefficient (or the Nei and Li coefficient) is a staple for measuring similarity between DNA fingerprints – particularly in studies involving restriction fragment length polymorphism (RFLP). It is simple, interpretable, and biologically intuitive. But was it the best? Could other metrics, especially the lesser-known ones implemented in COPADS, offer better discriminatory power or more appropriate mathematical behavior?

This study set out to compare Dice with 19 alternative similarity measures in the context of *Escherichia* coli DNA fingerprints. Using real data and statistical evaluation, we discovered that Dennis, Fossum, and Matching coefficients outperformed Dice in terms of discrimination, but were marred by the lack of boundedness; thus, making them unreliable for certain biological interpretations. Surprisingly, the Russel and Rao coefficient emerged as a highly correlated and bounded alternative to Dice (r^2 = 0.998), suggesting it could serve as a direct substitute in many studies. It offered a rare blend of mathematical rigor and biological relevance.

But this paper also holds personal significance that transcends its technical contributions. It marks the beginning of a long and treasured friendship with Zhu En Chay, the first author. In 2010, I was a third-year lecturer at Singapore Polytechnic; she was one of my students. This was his first foray into research. What began as a typical mentor–mentee relationship blossomed into a lifelong friendship, grounded in mutual respect and shared curiosity. Even today, we remain close – a reminder that science is not just about data but also about people and relationships.

Reflection: This project was deceptively modest – a comparison of distance coefficients applied to DNA fingerprints. But it reinforced a key principle in my scientific practice: assumptions deserve to be challenged. The Dice coefficient was widely accepted, yet few had rigorously compared it against other alternatives. With the computational foundation of COPADS and the biological motivation from Bactome, we were uniquely positioned to conduct that comparison.

In retrospect, I see this paper not only as a piece of applied bioinformatics, but also as an affirmation of integrated thinking; blending mathematics, biology, and software into a unified inquiry. And it showed me, once again, that giving students room to explore and contribute meaningfully can lead to both scientific value and enduring human connection.

While the paper itself may not be highly cited or widely known, its impact on how I think, teach, and collaborate; is far greater than what any citation metric could convey. It reminds me that mentorship can be transformative, for both teacher and student.

Sidebar: When Mentorship Becomes Kinship

Scientific papers are typically catalogued by results and methods, but every so often, one becomes a bookmark in the story of a relationship. This was one such paper.

Zhu En Chay walked into my classroom as a student, uncertain but curious. By the time we co-authored this study, that curiosity had matured into rigorous inquiry, and what began as a teaching moment evolved into a genuine research collaboration. Today, it’s a lasting friendship rooted in intellectual trust and shared values.

This sidebar isn’t about coefficients or DNA fingerprints; it’s about the often-invisible dimension of academic work: human growth. Mentorship is not always about molding someone into a replica of yourself. Sometimes, it’s simply about walking alongside someone as they find their own way.

The scientific content of this paper was meaningful. But the enduring takeaway is more personal: the greatest outcome of mentorship may not be the paper you publish but the person you walk with long after it’s done.

17: Chi-Square, F-, and t-tests

Citation: Chay, ZE, Ling, MHT. 2010. COPADS, II: Chi-Square test, F-Test and t-Test Routines from Gopal Kanji's 100 Statistical Tests. The Python Papers Source Codes 2:3.

Abstract: This paper extends previous work on the implementation of statistical tests as described by Kanji. A total of 8 Chi-square tests, 3 F-tests and 6 t-tests routines are implemented, bringing a total of 27 out of 100 tests implemented to date.

Context: This paper represents a steady continuation of earlier work – particularly "Ten Z-test Routines from Gopal Kanji's 100 Statistical Tests". In this phase, COPADS was extended to include 8 Chi-square tests, 3 F-tests, and 6 t-tests, resulting in 27 out of 100 statistical routines from Kanji's catalogue being implemented.

The project was again propelled by Zhu En's enthusiasm. Having some time before beginning his National Service, he asked if there was anything else he could work on. The answer was yes, more statistical tests. While not groundbreaking in the theoretical sense, this work was exacting and deeply useful, especially for a needs-driven project like COPADS. The statistical routines were intended not just for practice, but for reuse in applied biological and computational research.

What makes this paper notable beyond its content is where it was published. The Python Papers Source Codes was an offshoot of The Python Papers, created to provide a formal venue for pure source code contributions, much like how CALGO (Collected Algorithms of the ACM) operates. By publishing here, we aligned COPADS with the broader ideal of well-documented, reproducible, reusable code libraries – not just theoretical exercises, but implementations meant to live and breathe in actual projects.

Reflection: This chapter of COPADS reminded me that scientific legacy isn't just about ideas, it's about infrastructure. Every statistical test implemented added a small but essential block to the computational scaffolding I was trying to build. It wasn't flashy, but it was real.

The simplicity of the effort belied its long-term impact. What we were really doing was preserving knowledge in executable form – the kind of work that mirrors what CALGO has done for decades. In a research environment that often overvalues novelty, I've always admired the quiet brilliance of making algorithms reusable. That's what this project aspired to emulate.

This phase also deepened the mentor-mentee dynamic between Zhu En and me. By now, we were not just collaborating; we were co-building something larger than ourselves. He wasn't merely following instructions; he was beginning to see the broader architecture and contributing with a sense of ownership. That subtle shift from task-doer to project-steward is, to me, the real outcome of mentorship done well.

Looking back, this paper may not attract attention in a citation index. But it was never meant to. It was meant to equip, to enable, and to endure – three words that capture the essence of what COPADS stood for, and what this collaboration grew into.

18: First Attempt at Formal Methods

Citation: Ling, MHT. 2010. Specifying the Behaviour of Python Programs: Language and Basic Examples. The Python Papers 5(2): 4.

Abstract: This manuscript describe BeSSY, a function-centric language for formal behavioural specification that requires no more than high-school mathematics on arithmetic, functions, Boolean algebra and sets theory. An object can be modelled as a union of data sets and functions whereas inherited object can be modelled as a union of supersets and a set of object-specific functions. Python list and dictionary operations are specified in BeSSY for illustration.

Context: BeSSY, short for Behavioural Specification SYstem, was my foray into something that had long fascinated me: formal methods. While I've always leaned toward applied computational biology and pragmatic software engineering, there was something intoxicating about the precision and elegance of formal systems. They're often seen as esoteric or overengineered, yet in their abstract discipline lies a quiet beauty.

BeSSY was designed as a function-centric language for behavioural specification, rooted in simple yet rigorous mathematics; just high-school level arithmetic, Boolean algebra, functions, and set theory. The motivation was twofold:

1. To formalize the behaviour of Python programs in a way that was readable and teachable.
2. To bridge the gap between software intuition and mathematical rigour.

In the paper, I illustrated how objects can be expressed in BeSSY – as unions of datasets and functions. Even concepts like inheritance, often opaque in formal language theory, were modeled using supersets and object-specific functions. It was important to me that BeSSY did not require a PhD to understand; clarity and accessibility were part of its ethos.

Reflection: Looking back, BeSSY feels like one of my most philosophically driven works. It didn't arise from practical need like COPADS, nor from experimental data like Bactome. It was born from a yearning to explore, to see if I could make something as abstract as formal specification intuitive.

It was also a form of rebellion. Formal methods are often wrapped in academic gatekeeping, and BeSSY was a quiet declaration that they didn't have to be. I wanted to show that formalism can be expressed simply, taught simply, and perhaps even enjoyed.

The paper is modest in length and ambition, but to me, it marked a key moment: an intersection of programming, mathematics, and design. In some ways, BeSSY was the opposite of COPADS. Whereas COPADS was code to be run, BeSSY was code to be reasoned about.

This chapter also affirmed a personal truth – that difficulty and elegance are not opposites. The joy of formal methods is precisely in their challenge, and BeSSY was my way of dancing with that difficulty, trying to shape it into something both useful and beautiful.

Sidebar: The Beauty of Simplicity in Formal Methods

BeSSY may have been a modest attempt at formal methods, but its creation was anything but small in impact. What I aimed to prove with BeSSY was that formal methods, often seen as esoteric or unnecessarily complex, can be both intuitive and accessible. The use of high-school-level mathematics; arithmetic, functions, Boolean algebra, and set theory; was a deliberate decision to make these powerful tools available to a broader audience, without sacrificing the mathematical rigor that makes them valuable.

What made BeSSY unique wasn't just its approachability, but also its challenge to the academic elitism that often surrounds formal methods. At its core, BeSSY was about breaking down barriers, making formal specification something that anyone with a basic understanding of mathematics could understand and apply. I saw it as a rebellion against the notion that deep mathematical concepts must be reserved for the few.

This endeavour also reinforced an important lesson: difficulty and elegance are not opposites. Formal methods, with their precision and rigor, can be challenging, but it's this very challenge that gives them their beauty. BeSSY was my way of embracing that challenge, and in doing so, it helped me understand the balance between simplicity and sophistication that lies at the heart of all good design.

19: Review on Mining Protein-Protein Interactions from Text

Citation: Ling, MHT, Lefevre, Christophe, Nicholas, KR. 2010. Mining Protein-Protein Interactions from Published Abstracts with MontyLingua. In Zhongming Zhao(ed). Sequence and Genome Analysis: Methods and Applications. iConcept Press Pty Ltd.

Abstract: The exponential increase in publication rate of new articles is limiting access of researchers to relevant literature. This has prompted the use of text mining tools to extract key biological information. Previous studies have reported extensive modification of existing generic text processors to process biological text. However, this requirement for modification had not been examined. In this study, we have constructed Muscorian, using MontyLingua, a generic text processor. It uses a two-layered generalization-specialization paradigm previously proposed where text was generically processed to a suitable intermediate format before domain-specific data extraction techniques are applied at the specialization layer. Evaluation using a corpus and experts indicated 86-90% precision and approximately 30% recall in extracting protein-protein interactions, which was comparable to previous studies using either specialized biological text processing tools or modified existing tools. We attributed this performance to alternative part-of-speech tags use. Our study had also demonstrated the flexibility of the two-layered generalization-specialization paradigm by using the same generalization layer for two specialized information extraction tasks.

Context: This chapter was a strategic repackaging of a core idea from my PhD thesis – how to automatically extract protein-protein interactions (PPIs) from the biomedical literature. Rather than building a bespoke or domain-specific NLP tool from scratch, I adopted a bold premise: could a generic text processor like MontyLingua perform sufficiently well if we structured the problem differently?

The result was Muscorian, a lightweight text-mining system named after Mus musculus and MontyLingua. It was structured using a two-layered generalization-specialization paradigm I had previously proposed. At the generalization layer, abstracts were parsed into a uniform syntactic format. At the specialization layer, domain-specific rules extracted interactions.

We achieved 86–90% precision and ~30% recall – numbers that, at first glance, mirror a typical “precision-recall trade-off” story. But the deeper success lay in demonstrating that it was possible to repurpose a generic NLP tool with minimal modification to achieve domain-relevant results. In other words, MontyLingua didn’t need to be rewritten to speak biology, it needed a clever wrapper.

This was a book chapter, not a journal article, and that gave me some latitude in tone and structure. I could afford to be a little more narrative, a little more explanatory; which, in hindsight, helped distill the thesis' complexity into a more digestible form.

Reflection: This chapter holds a special place for me. It represents my effort to carry a dense doctoral idea across the bridge into a new audience. The PhD thesis was a long, often grueling document. This chapter, by contrast, felt like breathing life back into it – distilling the insight, stripping the technical excess, and writing it so someone else could use it.

It also stands as a testament to creative reuse, both of tools and of ideas. Instead of building another bespoke parser, I framed MontyLingua as a pipeline component. Instead of publishing another fragment of my thesis as a standalone paper, I compiled it into a chapter with a coherent teaching arc.

MontyLingua had been a companion of sorts in my early career – first in The Python Papers, then in my PhD, and now again here. Its open-source nature and linguistic generality reminded me that good tools often transcend their original intent. Muscorian's success wasn't just in performance; it was in how little work was needed to achieve that performance, if the architecture was framed right.

Finally, this chapter marked one of my earliest contributions to a scientific book – a quiet milestone that made me feel like a bridge-builder between research, programming, and writing. It whispered: "You can communicate complexity without drowning your audience."

20: Biologically Relevant Genetic Algorithm

Citation: Lim, JZR, Aw, ZQ, Goh, DJW, How, JA, Low, SXZ, Loo, BZL, Ling, MHT. 2010. A Genetic Algorithm Framework Grounded in Biology. The Python Papers Source Codes 2: 6.

Abstract: Genetic algorithm (GA) is a heuristic search method inspired by biological evolution of genetic organisms by optimizing the genotypic combinations encoded within each individual with the help of evolutionary operators, such as reproduction, mutation and cross-over. This manuscript aims to present a simple GA framework written in Python programming language that conforms to biological hierarchy starting from gene to chromosome to genome (as organism) to population. Hence, we believe that this framework may be useful in both education and biological simulation on top of the usual domains where GA were used.

Context: By 2010, a crossroads was emerging in my academic path. I was preparing to leave Singapore Polytechnic, and with that departure came the end of a deeply personal aspiration; that is, to conduct hands-on experimental evolution in a wet lab. I had long been captivated by the elegance of biological evolution, not just as a scientific theory but as a generative principle. But without access to a lab, I faced a dilemma: Could I still study evolution experimentally without organisms, petri dishes, or pipettes?

This paper was my first affirmative answer to that question. Heavily influenced by the Digital Organisms paradigm, particularly the work of Wilke and Adami; I began to see computers not merely as analysis tools, but as experimental ecosystems.

Rather than adopting Avida, the dominant artificial life platform at the time, I turned inward by leveraging my dual fluency in biology and Python. The result was a biologically-grounded genetic algorithm framework, with a hierarchy that mirrors natural systems: gene → chromosome → genome → population.

The work was collaborative, co-authored with students under my mentorship, and published in The Python Papers Source Codes, the offshoot of The Python Papers that I helped co-found which is modeled on CALGO, the Collected Algorithms of the ACM. It was part code library, part teaching tool, and part quiet manifesto.

Reflection: This project marks a turning point in my relationship with computation. I no longer saw programming purely as a way to analyze data, automate workflows, or simulate known systems. Instead, I started seeing the computer as a sandbox for evolutionary ideas, a stage on which biological concepts could play out virtually.

This was also the moment I stopped waiting for “ideal lab conditions” to do the science I wanted to do. Instead, I began rewriting the conditions — carving out experimental space in the digital realm.

The algorithm itself wasn't radically new but the architecture mirrored biology more faithfully than most GA frameworks. That fidelity wasn’t just cosmetic. It allowed me to reason about digital evolution with the same intuition I would apply to natural systems. Students, too, found it easier to grasp. And in hindsight, this was my first true foray into computational naturalism using software to model and probe fundamental evolutionary principles.

Publishing it in The Python Papers Source Codes was poetic. It wasn’t just another utility library. It was part of a broader attempt to democratize scientific code, to make algorithmic insight accessible, shareable, and grounded in pedagogical clarity.

In essence, this paper was where my lab became virtual, and my experimental identity began to shift. It didn’t feel like loss; it felt like liberation.

Sidebar: The Liberation of Virtual Labs

This paper was a key moment in my academic journey, one where I transformed a personal challenge into an opportunity for innovation. Leaving behind the traditional wet lab did not mean abandoning my scientific curiosity. Instead, it led me to reimagine what it means to conduct experimental evolution. Rather than relying on petri dishes and organisms, I began experimenting in the virtual world, using computers as ecosystems where evolutionary processes could unfold.

This shift wasn’t just a technical one; it was a philosophical pivot. The algorithm itself wasn’t groundbreaking in its novelty, but the framework was rooted in biology in a way that many genetic algorithms weren’t. By organizing the system from gene to genome to population, I created a model that more faithfully reflected natural evolutionary processes. This approach made the concept of digital evolution more intuitive for both myself and my students, drawing clearer connections between the digital world and biological reality.

Publishing this work in *The Python Papers Source Codes* was a deliberate choice that tied the project to the broader mission of democratizing scientific code. It wasn’t just about writing a utility; it was about making scientific ideas more accessible, more transparent, and more teachable. This paper marked the point where I moved away from waiting for “ideal lab conditions” and instead began building my own experimental space in the digital realm. It wasn’t a loss; it was liberation, and it reshaped my scientific identity in profound ways.

21: Gene Name Normalization

Citation: Kuo, CJ, Ling, MHT, Hsu, CN. 2011. Soft Tagging of Overlapping High Confidence Gene Mention Variants for Cross-Species Full-Text Gene Normalization. BMC Bioinformatics 12(Suppl 8): S6.

Abstract: Background: Previously, gene normalization (GN) systems are mostly focused on disambiguation using contextual information. An effective gene mention tagger is deemed unnecessary because the subsequent steps will filter out false positives and high recall is sufficient. However, unlike similar tasks in the past BioCreative challenges, the BioCreative III GN task is particularly challenging because it is not species-specific. Required to process full-length articles, an ineffective gene mention tagger may produce a huge number of ambiguous false positives that overwhelm subsequent filtering steps while still missing many true positives. Results: We present our GN system participated in the BioCreative III GN task. Our system applies a typical 2-stage approach to GN but features a soft tagging gene mention tagger that generates a set of overlapping gene mention variants with a nearly perfect recall. The overlapping gene mention variants increase the chance of precise match in the dictionary and alleviate the need of disambiguation. Our GN system achieved a precision of 0.9 (F-score 0.63) on the BioCreative III GN test corpus with the silver annotation of 507 articles. Its TAP-k scores are competitive to the best results among all participants. Conclusions: We show that despite the lack of clever disambiguation in our gene normalization system, effective soft tagging of gene mention variants can indeed contribute to performance in cross-species and full-text gene normalization.

Context: This chapter represents the next phase of my collaboration with Academia Sinica, Taiwan; a relationship that began with BioADI and deepened through shared participation in the BioCreative III challenge. In an era before remote scientific collaboration became routine, we worked entirely virtually through emails and MSN Messenger and never once meeting in person. Yet, despite the physical distance, this project was remarkably intimate in its intellectual rhythm.

The BioCreative III Gene Normalization (GN) task posed a serious challenge: instead of working on abstracts or species-specific corpora, participants had to normalize gene mentions from full-length articles across multiple species. This was a radical departure from prior GN tasks, where high recall was often enough – subsequent filters and disambiguation would deal with noise. But here, noise could drown signal. Over-tagging could overwhelm normalization systems, and under-tagging would miss vital biological entities.

This led to our key innovation: soft tagging of overlapping high-confidence gene mention variants. Rather than forcing a single best guess, our system preserved multiple plausible mentions – a hedge that dramatically increased the chance of dictionary matches and relieved pressure on the disambiguation phase.

Reflection: What made this work special wasn't just the algorithmic contribution though achieving 90% precision and a competitive TAP-k score in BioCreative III was no small feat. What stood out was the elegance of the solution: embracing ambiguity, rather than trying to suppress it.

In many ways, soft tagging was a philosophical decision as much as a technical one. Where others tried to disambiguate prematurely, we deferred commitment by letting ambiguity persist until downstream processes could handle it better. This resonates with the broader computational theme of "lazy evaluation" to do the heavy lifting only when you absolutely must.

Personally, this paper reinforced my belief that deep collaboration doesn't require physical proximity. Despite never meeting my co-authors in person, we built a system that was conceptually coherent and practically effective. The time zone difference, Singapore and Taiwan, was small and our communication was nimble and frictionless. It helped that we trusted each other's strengths and communicated with focused intent.

Looking back, I see this as a validation of two things:

1. That softness in algorithms – a willingness to entertain ambiguity can sometimes produce the strongest results.
2. That softness in collaboration – flexible, asynchronous, and remote can still generate hard science.

This project did more than contribute to GN literature. It helped shift my thinking toward systems that tolerate ambiguity and thrive because of it.

22: Relational Database to Hypergraph Database

Citation: Tahat, A, Ling, MHT. 2011. Mapping Relational Operations onto Hypergraph Model. The Python Papers 6(1): 4.

Abstract: The relational model is the most commonly used data model for storing large datasets. However, many real world objects are recursive and associative in nature which makes storage in the relational model difficult. The hypergraph model is a generalization of a graph model, where each hypernode can be made up of other nodes or graphs and each hyperedge can be made up of one or more edges. It may address the recursive and associative limitations of relational model. However, the hypergraph model is non-tabular; thus, loses the simplicity of the relational model. In this study, we consider the means to convert a relational model into a hypergraph model in two layers and present a reference implementation of relational operators (project, rename, select, inner join, natural join, left join, right join, outer join and Cartesian join) on a hypergraph model.

Context: At this point in my career, I began exploring graph databases more deeply, driven by the limitations I had encountered with the relational model, especially when working with recursive or associative data structures. The hypergraph model, a generalization of graph structures that allows nodes and edges to have internal complexity, intrigued me as a potentially powerful alternative for representing such datasets. The idea for this chapter was to investigate how classical relational operations (like select, project, and various joins) might be implemented on a hypergraph structure.

The collaboration itself was serendipitous and fully remote. Amani Tahat, based in the Middle East, reached out to me via email with interest in collaborating. I can no longer recall whether the idea originated from her or me, as is often the case in organic collaborations, the seed of the project likely emerged in the back-and-forth of early discussions. What I do recall is our shared fascination with mapping traditional relational operators onto the more expressive structure of hypergraphs, and the clarity with which we both saw the challenge.

Reflection: This work marked the beginning of my exploration into non-tabular data models. Looking back, I see it as a gentle transition from conventional database thinking toward semantic structures, graph theory, and even knowledge representation – ideas that would reappear in other parts of my research. Despite the paper being published in The Python Papers, which had a more technical audience, the underlying motivation was conceptual: How can we move beyond flat representations of data without losing the operational logic we depend on? This question has stayed with me ever since.

Sidebar: From Flat Tables to Hypergraphs: Rethinking Data Representation

One of the most significant challenges in data modelling lies in representing complex relationships. In traditional relational databases, everything is neatly arranged into tables, with rows representing individual records and columns representing attributes. However, this tabular structure can quickly become unwieldy when trying to represent relationships that aren't easily captured in rows and columns, especially when recursion and associations come into play.

Hypergraphs introduce a new way to think about data, where nodes can represent complex entities and edges can represent multiple types of relationships between them. This shift away from a flat, tabular model opens up new possibilities for more natural and flexible data representations. For example, in a traditional database, trying to represent a scenario where a person is both a friend and a colleague of another person would require multiple joins across different tables. In a hypergraph, this can be captured more simply with a single edge connecting the two people, labelled with different relationship types (e.g., “friend” and “colleague”). This approach more closely mirrors the way we understand and navigate relationships in the real world.

Sidebar: Hypergraphs and Their Role in Data Science and AI

As the field of artificial intelligence and machine learning continues to evolve, the role of graph-based data models, especially hypergraphs, becomes increasingly important. Hypergraphs allow data scientists and researchers to model complex relationships in a way that traditional relational databases simply cannot. For example, in a recommendation system, hypergraphs can be used to represent users, products, and interactions between them. A single hyperedge could capture a user’s interest in multiple products, allowing the model to explore richer, more nuanced relationships.

In biology, hypergraphs are especially useful for modelling complex systems like protein interaction networks, where multiple proteins may interact with each other in various ways. These types of models have the potential to reveal new insights into how biological processes work, providing more accurate representations than what might be possible with traditional models. Hypergraphs make it easier to explore multidimensional relationships, leading to new methods for solving problems in diverse fields ranging from medicine to social network analysis.

23: Only Manuscript from Life Technologies

Citation: Ling, MHT, Jean, A, Liao, D, Tew, BBY, Ho, S, Clancy, K. 2011. Integration of Standardized Cloning Methodologies and Sequence Handling to Support Synthetic Biology Studies. Third International Workshop on Bio-Design Automation (IWBDA). San Diego, California, USA.

Abstract: The assembly and downstream transformation of genetic constructs has been a fundamental scientific technology for the last thirty years. Synthetic biology is an engineering based approach to molecular biology as emphasizing the standardized assembly of characterized DNA fragments. The standards promoted by the BioBricks™ Foundation have enabled novel constructs to be developed based upon the expected function of these constructs. However scientists need a software environment that enables them to curate large collections of parts and assemblies, combined with appropriate tools to facilitate quick creation of constructs and identification of potential design issues *in silico*. In this paper, we present the implementation of BioBrick™ and GENEART® Assembly tools, coupled with an enhanced database to manage and develop such parts collections. Integration of these tools and data into the VectorNTI® software suite is a step towards implementation of BioCAD™, a computer based design approach to facilitate development of complex circuit based perturbation of cellular systems.

Context: This chapter marks a distinct phase in my career – moving into industry – my time at Life Technologies (now ThermoFisher Scientific) as a Senior Scientist. My core responsibility was the revitalization of Vector NTI, a widely used software suite in molecular biology. For decades, Vector NTI had served as a staple tool for DNA sequence visualization and manipulation, but by the time I joined, it needed a conceptual and architectural overhaul to stay relevant in the emerging synthetic biology landscape.

Synthetic biology, with its emphasis on standardization, modularity, and automation, demanded more than just sequence editing – it required BioCAD-style environments that could integrate DNA assembly standards like BioBrick™ and GENEART®, manage large collections of parts, and simulate genetic constructs *in silico* before actual benchwork. This paper, presented at IWBDA 2011 in San Diego, was a culmination of that vision: integrating standardized cloning workflows and sequence handling within a cohesive and extensible platform.

Notably, this body of work was later formalized in my only awarded patent to date: US Patent 9,465,519; which describes an integrated system for managing synthetic biology constructs.

Reflection: Although I only have one publication from my time at Life Technologies, it represents one of the most industry-facing and practically impactful phases of my scientific journey. This was a period of transition – from the academic to the commercial, from theoretical constructs to usable tools in wet labs. I was part of a team that bridged software engineering, molecular biology, and product strategy, all while responding to real user needs.

The shift in perspective – from publishing papers to designing tools that thousands would use – was both sobering and invigorating. Unlike my academic roles, where the goal was novelty, this role emphasized utility, usability, and integration. The Vector NTI revitalization effort taught me that science isn't just about new ideas; it's also about delivering those ideas in usable form to practitioners.

This chapter is special to me not only for the scientific contribution but also because it represents my first and only granted patent, a rare formal recognition in the commercial R&D space – a symbol of applied innovation that made its way from concept to product shelf.

Sidebar: Bridging Academia and Industry: The Vector NTI Revitalization

The transition from academia to industry can be a significant shift for many scientists, and my time at Life Technologies was a perfect example of this. As part of a team working on the revitalization of Vector NTI, I found myself navigating the intersection of scientific innovation and market-driven needs. In academia, the focus is often on generating novel ideas and theoretical advancements. In industry, however, the emphasis shifts toward creating practical, user-friendly solutions that can directly impact the work of scientists in the lab.

One of the most satisfying aspects of this role was working on the integration of synthetic biology tools into an established platform. The goal was not just to improve sequence handling but to create an ecosystem that could handle modular DNA assembly, incorporate design standards like BioBrick™ and GENEART®, and simulate genetic constructs before they were ever tested in the lab. This was about taking abstract scientific principles and turning them into tangible products that were used by thousands of researchers worldwide.

Sidebar: The Evolution of BioCAD: A Vision for the Future of Synthetic Biology

The work presented in this manuscript laid the groundwork for what I envisioned as the future of synthetic biology, a BioCAD system. Just as CAD software revolutionized engineering by allowing professionals to design and test physical structures in a virtual space, BioCAD sought to do the same for genetic engineering. This vision wasn't just about automating the cloning process but about creating an environment where the entire workflow, from designing DNA constructs to simulating their biological impact, could be carried out digitally before ever touching a test tube.

By incorporating standards like BioBrick™ and GENEART®, the platform I helped develop could support the creation and testing of genetic constructs *in silico*. This innovation was critical to reducing errors in the design phase, saving valuable time in the lab, and improving the overall efficiency of synthetic biology research. Although the work was ultimately commercialized within the Vector NTI suite, the underlying ideas have had a lasting influence on how we think about designing and engineering biology today. This move toward computational biology has only accelerated in recent years, and tools like BioCAD are poised to become even more integral to synthetic biology in the future.

Sidebar: Why Every Academic Should Spend Time in Industry

Stepping into industry after years in academia was like walking into a room where the rules of the game were the same but the stakes, expectations, and timelines were entirely different.

At Life Technologies, the question wasn't "Is this novel?" but "Will this help someone today?" I also got to see how business unit leaders think. That shift reframed my thinking. In academia, success often lives in journals and conference halls. In industry, success is measured by usability, adoption, and how much friction you remove from a researcher's workflow. It taught me that elegant science must also be deliverable science.

Working on Vector NTI forced me to become multilingual: speaking biology to bench scientists, engineering to software developers, and strategy to product managers. That kind of cross-functional communication is rarely emphasized in academic settings, but it's essential for translating research into real-world impact.

Most importantly, my time in industry made me a better academic. It sharpened my instincts for practical problem-solving, strengthened my sense of user empathy, and deepened my appreciation for interdisciplinary collaboration. When I returned to academia, I brought with me a renewed commitment to relevance to designing solutions that don't just work in theory, but actually help someone, somewhere, do their science better.

If academia teaches us how to think deeply, industry teaches us how to build boldly. Both are necessary. Together, they make for a more complete scientist.

24: Generic Tool for Gene Ontology Enrichment Analysis

Citation: Ling, MHT. 2011. Bactome II: Analyzing Gene List for Gene Ontology Over-Representation. The Python Papers Source Codes 3: 3.

Abstract: Microarray is an experimental tool that allows for the screening of several thousand genes in a single experiment and the analysis of which often requires mapping onto biological processes. This allows for the examination of processes that are over-represented. A number of tools have been developed but each differed in terms of organisms that can be analyzed. Gene Ontology website has a list of up-to-date annotation files for different organisms that can be used for over-representation analysis. Each file maps each gene of the organism to its ontological terms. It is a simple tool that allows users to use the up-to-date annotation files to generate the expected and observed counts for each GO identifier (GO ID) from a given gene list for further statistical analyses.

Context: This chapter builds upon an earlier tool, Bactome I, designed for microarray data interpretation. With the surge of microarray experiments in the late 2000s, researchers needed accessible software that could map gene expression changes to biological meaning, particularly through Gene Ontology (GO). GO terms categorize gene functions under biological processes, molecular functions, and cellular components. Over-representation analysis (ORA) allows scientists to identify which biological processes are disproportionately represented in a gene list, offering a crucial window into the system-level implications of an experiment.

Bactome II emerged as a lightweight Python tool that leverages publicly available GO annotation files, allowing scientists to analyze any organism so long as annotations were available. This bypassed the limitations of existing tools that often supported only a subset of model organisms. It performed expected vs. observed counts of GO terms, making it easy to plug into a user’s statistical pipeline.

Reflection: Bactome II exemplifies a design philosophy I’ve grown to deeply value; which are simplicity, modularity, and generalizability. Instead of building a bloated web server or GUI tool, I focused on something lean: command-line friendly, easy to extend, and grounded in open data standards.

It was also a nod to the Python Papers Source Codes journal’s mission; that is to share practical, usable tools with the community. Though the tool itself may seem modest, it addressed a persistent need: empowering non-model organism researchers to perform meaningful GO analysis without relying on infrastructure-heavy software.

In retrospect, Bactome II represents a kind of scientific egalitarianism. It gave anyone with a gene list and Python access to an analytical capability usually locked behind complex platforms. It also marked one of my continuing commitments to bridge the accessibility gap in bioinformatics by designing tools that respect users' diversity in background, resources, and organisms of interest.

25: mdoG as Reference Gene

Citation: Heng, SSJ, Chan, OYW, Keng, BMH, Ling, MHT. 2011. Glucan biosynthesis protein G (mdoG) is a Suitable Reference Gene in *Escherichia coli* K-12. ISRN Microbiology 2011, Article ID 469053.

Abstract: The expressions of reference genes used in gene expression studies are assumed to be stable under most circumstances. However, a number of studies had demonstrated that such genes were found to vary under experimental conditions. In addition, genes that are stably expressed in an organ may not be stably expressed in other organs or other organisms, suggesting the need to identify reference genes for each organ and organism. This study aims at identifying stably expressed genes in *Escherichia* coli. Microarray datasets from *E. coli* substrain MG1655 and 1 dataset from W3110 were analysed. Coefficient of variance (CV) of was calculated and 10% of the lowest CV from 4631 genes common in the 3 MG1655 sets were analysed using NormFinder. Glucan biosynthesis protein G (mdoG), which is involved in cell wall synthesis, displayed the lowest weighted CV and weighted NormFinder Stability Index for the MG1655 datasets, while also showing to be the most stable in the dataset for substrain W3110, suggesting that mdoG is a suitable reference gene for *E. coli* K-12. Gene ontology over-representation analysis on the 39 genes suggested an over-representation of cell division, carbohydrate metabolism, and protein synthesis which supports the short generation time of *E. coli*.

Context: Quantitative gene expression studies rely heavily on the use of reference genes; that are genes presumed to have stable expression across various conditions. However, accumulating evidence in the early 2010s showed that commonly used reference genes (like rpoD or 16S rRNA) were not as stable as previously believed, particularly across different growth stages or stress responses. This posed a problem: unreliable reference genes could compromise the integrity of downstream analyses.

This project aimed to systematically identify a robust reference gene for *Escherichia* coli K-12, using public microarray datasets from strains MG1655 and W3110. We calculated the coefficient of variation (CV) across conditions and used NormFinder, a statistical algorithm, to assess gene expression stability. Among the top candidates, mdoG, a glucan biosynthesis protein involved in cell wall construction; emerged as the most stable across all datasets, making it a compelling candidate for a new reference standard.

Reflection: What makes this chapter special is not just the scientific finding, but who I did it with. Sean, Oliver, and Bryan were just 15-year-old students from Raffles Institution's Gifted Education Programme. Every weekend, we met at a CoffeeBean outlet armed with laptops, hot chocolate, and the occasional sandwich to explore data, discuss methodology, and write code. Their capacity for analytical thought, curiosity, and sheer enthusiasm for science was both humbling and inspiring.

Many researchers measure legacy by the number of citations or impact factors. But to me, this project embodies mentorship at its most rewarding. These young minds didn’t just follow instructions; they debated statistical choices, critiqued GO annotations, and asked me why mdoG was more stable than rpoD. And they earned co-authorship in a peer-reviewed journal before completing their O-levels.

This chapter is a reminder that science is not just about discovery, it’s about awakening the scientist in others, no matter how young. It reminds me why I teach, why I mentor, and why I choose to stay grounded in accessibility and inclusion, even in highly technical fields.

Sidebar: Mentoring the Next Generation of Scientists

One of the most fulfilling aspects of this project was working with my young mentees, Sean, Oliver, and Bryan. All were just 15 years old, yet their passion for science and analytical thinking were beyond their years. We gathered at a local CoffeeBean outlet every weekend to dive deep into data analysis and discuss experimental methods, fuelled by laptops, hot chocolate, and the occasional sandwich. The level of engagement they brought to this project was truly inspiring. They didn’t simply follow instructions—they questioned methodologies, critiqued statistical choices, and independently explored alternative approaches.

Their involvement in this project exemplifies the importance of mentorship and inclusion in scientific research. It’s not just about handing down knowledge but creating an environment where young minds can explore, challenge, and grow. For these students to contribute to a peer-reviewed paper before \their O-levels completing age was a significant achievement, and it stands as a testament to the value of encouraging curiosity and critical thinking early on.

26: Reference Genes for Human Lungs

Citation: Too, IHK, Ling, MHT. 2012. Signal Peptidase Complex Subunit 1 (SPCS1) and Hydroxyacyl-CoA Dehydrogenase Beta Subunit (HADHB) are Suitable Reference Genes in Human Lungs. ISRN Bioinformatics 2012, Article ID 790452.

Abstract: Lung cancer is a common cancer and expression profiling can provide an accurate indication to advance the medical intervention. However, this requires the availability of stably expressed genes as reference. Recent studies had shown that genes that are stably expressed in a tissue may not be stably expressed in other tissues suggesting the need to identify stably expressed genes in each tissue for use as reference genes. DNA microarray analysis has been used to identify those reference genes with low fluctuation. Fourteen datasets with different lung conditions were employed in our study. Coefficient of variance, followed by NormFinder, was used to identify stably expressed genes. Our results showed that classical reference genes such as GAPDH and HPRT1 were highly variable; thus, unsuitable as reference genes. SPCS1 and HADHB, involving in fundamental biochemical processes, demonstrated high expression stability suggesting that their suitability in human lung cell profiling.

Context: In gene expression studies, reliable reference genes are essential for normalization. While housekeeping genes like GAPDH and HPRT1 were historically used as universal references, emerging research showed their expression varied significantly across tissue types and pathological states. In lung tissues, this variability could compromise the integrity of expression profiling, particularly in cancer-related studies.

This study aimed to identify lung-specific, stably expressed reference genes. Using fourteen lung-related microarray datasets covering both healthy and diseased conditions, we evaluated gene stability using the coefficient of variance and NormFinder, a statistical model that accounts for intra- and inter-group variation. Surprisingly but consistently, commonly used reference genes performed poorly. Instead, SPCS1, involved in signal peptide cleavage, and HADHB, key to fatty acid β-oxidation, emerged as stable and reliable choices for human lung tissues.

Reflection: This project was co-authored with Issac Too, who began as my student at Singapore Polytechnic. He was more than just a good student; he was someone who naturally assumed responsibility. I still remember him grabbing me coffee from the canteen when our classes stretched into the late afternoon; not out of obligation, but from a genuine sense of care and leadership. (Naturally, I always paid.)

By the time we worked on this paper, Issac had moved on to university, but we remained in touch. Like several of my former students, he wasn’t just a mentee; he became a colleague and collaborator. Our working sessions; often in CoffeeBean or McDonald’s, punctuated by fries and data plots; were collaborative yet warm, focused yet filled with banter. Issac went on to complete his PhD, which gives me a quiet sense of pride. Mentorship, at its core, is about helping others become who they were always meant to be.

This chapter reminds me of the power of long arcs of mentorship that transcends academic semesters, of relationships that evolve into partnerships. In the stability of SPCS1 and HADHB, I see a metaphor: beneath the flux of education and research, what truly endures are the foundational connections we build.

Sidebar: The Evolution of Mentorship and Collaboration

This chapter is a poignant reminder of the long-lasting impact mentorship can have. My collaboration with Issac Too was not only a professional partnership but a personal journey of growth. Issac, who began as a student at Singapore Polytechnic, quickly became someone who embodied leadership and responsibility in ways that went beyond academics. I remember him bringing me coffee during long afternoons, not out of obligation, but as an expression of his genuine care for our shared work. By the time we co-authored this paper, Issac had progressed to university, but our bond remained strong, transitioning from teacher-student to colleague-collaborator.

Working together on this project was both a warm, collaborative experience and a testament to the value of mentorship that extends beyond the classroom. Issac’s growth into an independent researcher and his subsequent PhD achievement fills me with pride, knowing that mentorship is as much about nurturing potential as it is about guiding through knowledge.

27: Reference Genes Review

Citation: Dundas, JB, Ling, MHT. 2012. Reference Genes for Measuring mRNA Expression. Theory in Biosciences 131: 215-223.

Abstract: The aim of this review is to find answers to some of the questions surrounding reference genes and their reliability for quantitative experiments. Reference genes are assumed to be at a constant expression level, over a range of conditions such as temperature. These genes, such as GADPH and beta-actin, are used extensively for gene expression studies using techniques like quantitative PCR. There have been several studies carried out on identifying reference genes. However, a lot of evidence indicates issues to the general suitability of these genes. Recent studies had shown that different factors, including the environment and methods, play an important role in changing the expression levels of the reference genes. Thus, we conclude that there is no reference gene that can deemed suitable for all the experimental conditions. In addition, we believe that every experiment will require the scientific evaluation and selection of the best candidate gene for use as a reference gene in order to obtain reliable scientific results.

Context: At the heart of quantitative gene expression experiments lies an unassuming but critical requirement: a stable reference gene. These genes, assumed to maintain consistent expression across experimental conditions, provide the anchor point for interpreting variations in target gene expression. Historically, GAPDH and β-actin were regarded as universal constants – the steady signals in the noise of biological fluctuation.

But over the course of my own projects and as reflected in a growing body of literature, I encountered again and again the fallibility of these so-called housekeeping genes. Their expressions were anything but stable when subjected to environmental shifts, experimental manipulations, or even tissue-specific contexts. It became increasingly clear that there was no such thing as a universal reference gene.

This review, co-authored with Jitesh Dundas, was a synthesis and reflection on this evolving understanding. We critically examined the assumptions underpinning the use of reference genes and surveyed the literature that challenged their universality. The conclusion was unambiguous: every experiment demands its own empirical validation. The idea of one-size-fits-all housekeeping genes was a convenient fiction.

Reflection: This chapter is unique in my body of work for several reasons. First, I never met Jitesh Dundas in person. Like Amani Tahat before him, Jitesh had reached out to me via email, expressing interest in collaboration. I remain unsure what prompted his message – perhaps a shared frustration with the limitations of reference gene dogma but his initiative led to a thoughtful and timely review.

Second, this work signified a pivot from data analysis to theory. It wasn't about identifying the next "most stable gene" but rather stepping back and questioning the foundational assumptions that had, for years, gone unchallenged. In many ways, it was born out of accumulated frustration – after investing hours into qPCR only to discover that our chosen reference gene was varying more than our target gene.

Writing this paper was like cleaning the whiteboard. It cleared conceptual space and reminded me that robust science starts not with tools, but with questions.

Sometimes, our most enduring contributions don't come from discovering something new, but from clearly seeing what was already there, and naming the inconsistencies others had learned to ignore.

28: First Publication from My Final Year Project Group in Singapore Polytechnic

Citation: Lee, CH, Oon, JSH, Lee, KC, Lee, CH, Ling, MHT. 2012. *Escherichia coli* ATCC 8739 Adapts to the Presence of Sodium Chloride, Monosodium Glutamate, and Benzoic Acid after Extended Culture. ISRN Microbiology 2012, Article ID 965356.

Abstract: We observed the adaptation of *E. coli* cultured in different concentration of food additives (sodium chloride, benzoic acid and monosodium glutamate), singly or in combination, over 70 passages. Adaptability over time was estimated by generation time and cell density at stationary phase. Polymerase Chain Reaction (PCR) / Restriction Fragments Length Polymorphism (RFLP) using 3 primers and restriction endonucleases each was used to characterize adaptation/evolution at genomic level and compared by Nei-Li Dissimilarity Index. Our results demonstrated that *E. coli* in every treatment had adapted over 465 generations. The types of stress were discovered to be different even though different concentrations of same additives were used. Genomic analysis by PCR/RFLP shows that the stress response in *E. coli* may be similar.

Context: This work was born not from a laboratory with cutting-edge resources, but from the curiosity and grit of my very first Final Year Project (FYP) group at Singapore Polytechnic – Chin Hao, Jack, and Kun Cheng. The inspiration came from one of the most famous biological experiments in history: Richard Lenski’s long-term evolution experiment with *E. coli*, a study I had always admired for its conceptual elegance and scientific depth.

We wondered: Could *E. coli* also adapt to common food additives over time? What would happen if we repeatedly cultured them in sodium chloride, benzoic acid, or monosodium glutamate for dozens of generations? What would adaptation look like, phenotypically and genetically?

So, over 70 passages (roughly 465 generations), we grew *E. coli* in various stress conditions and tracked their growth kinetics and genomic responses. By the end, not only had the bacteria adapted to their respective environments, but we had a data-rich, multifaceted story about microbial resilience under human-made pressures.

Reflection: This chapter carries a special emotional weight. I spent countless evenings with Chin Hao, Jack, and Kun Cheng toward the latter part of the project – going through drafts, refining arguments, cross-checking references, and polishing figures. These weren’t just students; they were my first cohort of scientific apprentices. Their diligence was inspiring. Chin Hao, in particular, stood out, not just for his technical competence but for his maturity and leadership. He even drafted our response to the reviewers, something most undergraduates (and even many post-graduates) would shy away from.

I still remember the coffee we had at CoffeeBean at Holland Village, where I told Chin Hao I was leaving Singapore Polytechnic. He was the first to know, and I still recall the bittersweet clarity of that moment – part pride in what we built together, part sadness that the chapter was closing.

This paper wasn’t just about bacterial evolution. It was about the evolution of mentorship, of learning to let students take the wheel while still holding the map. We didn’t have a large budget or a fancy lab, but we had curiosity, discipline, and care; and that was enough.

Over a decade has passed, and I haven’t seen them since. But the lessons they gave me; as a mentor, collaborator, and human being; still echo in how I guide others today.

Sidebar: A Humble Start, A Grand Vision

This project was my first foray into mentoring undergraduates at Singapore Polytechnic, and it wasn’t just the students who were learning. The experiment itself, focused on how *E. coli* adapts to common food additives; like sodium chloride, monosodium glutamate, and benzoic acid; was an exploration of microbial resilience under human-made stresses. The inspiration came from the iconic *E. coli* long-term evolution experiment by Richard Lenski, but instead of a fancy lab, we worked with the resources we had: curiosity, discipline, and determination. This chapter marks the beginning of my journey in scientific mentorship and the realization that groundbreaking work doesn't always require cutting-edge technology; sometimes, it starts with a question and the drive to seek answers.

Sidebar: The Evolving Nature of Mentorship

Mentorship, in its purest form, is about guiding others to see the potential within themselves. With Chin Hao, Jack, and Kun Cheng, I was not just teaching them lab skills or scientific rigor – I was learning how to trust them and let them take the lead. Chin Hao’s initiative in drafting our response to the reviewers was a testament to the evolving relationship between mentor and mentee. By the time this project wrapped up, I wasn’t just their supervisor; I had become a partner in their growth. That bittersweet moment at CoffeeBean, when I told Chin Hao I was leaving Singapore Polytechnic, symbolized the emotional evolution of this mentorship. It marked the end of an era but also the beginning of a new chapter in both my professional and personal journey.

29: A Consultancy Project in South Dakota State University

Citation: Ling, MHT, Rabara, RC, Tripathi, P, Rushton, PJ, Ge, X. 2013. Extending MapMan Ontology to Tobacco for Visualization of Gene Expression. Dataset Papers in Biology 2013, Article ID 706465.

Abstract: Microarrays are a large-scale expression profiling method which has been used to study the transcriptome of plants under various environmental conditions. However, manual inspection of microarray data is difficult at the genome level because of the large number of genes (normally at least 30000) and the many different processes that occur within any given plant. MapMan software, which was initially developed to visualize microarray data for Arabidopsis, has been adapted to other plant species by mapping other species onto MapMan ontology. This paper provides a detailed procedure and the relevant computing codes to generate a MapMan ontology mapping file for tobacco (Nicotiana tabacum L.) using potato and Arabidopsis as intermediates. The mapping file can be used directly with our custom-made NimbleGen oligoarray, which contains gene sequences from both the tobacco gene space sequence and the tobacco gene index 4 (NTGI4) collection of ESTs. The generated dataset will be informative for scientists working on tobacco as their model plant by providing a MapMan ontology mapping file to tobacco, homology between tobacco coding sequences and that of potato and Arabidopsis, as well as adapting our procedure and codes for other plant species where the complete genome is not yet available.

Context: This work arose during my postdoctoral stint at South Dakota State University, where I was embedded in a research environment filled with plant molecular biologists, many of whom were working on Nicotiana tabacum (tobacco) as a model organism. At that time, I was serving in a consultancy-like role, supporting the group with bioinformatics expertise, particularly around gene expression analysis and visualization.

One pressing need was clear: MapMan, a well-known tool for visualizing transcriptomic data, had robust support for Arabidopsis and a few other model organisms, but not for tobacco. We had access to a custom NimbleGen oligoarray, incorporating sequences from the tobacco gene space and NTGI4 ESTs but without a proper MapMan mapping file, the power of the platform couldn’t be fully harnessed.

So, we developed a systematic procedure to extend the MapMan ontology to tobacco, using Arabidopsis and potato as bridge species. It was not flashy work but it was essential infrastructure for the lab. And perhaps more importantly, we documented everything: the procedure, the logic, the scripts so others could replicate or adapt it for species lacking full genome sequences.

Reflection: In many ways, this paper reflects the invisible labor of bioinformatics; that is, the creation of supporting tools, mappings, and workflows that enable others to do science more efficiently. There was no experimental benchwork here, no novel biological hypothesis being tested. But there was utility, and to me, that was enough reason to publish it.

What I appreciated most about this project was its quiet pragmatism. In academia, we often chase novelty and overlook the value of simply making things work better for a community. This dataset paper was my way of saying: “Here is something that might help. It’s not glamorous, but it’s solid. Use it, build on it.”

Looking back, this also marked a transitional phase in my career, from knowledge consumer to enabler. I was no longer just analyzing data or interpreting results. I was now designing the scaffolding upon which others could build their experiments. That shift from discovery to enablement would continue to shape how I saw my role in science.

It reminds me that impact isn't always loud. Sometimes, it's a mapping file quietly being used in a lab halfway across the world.

Sidebar: Quiet Pragmatism in Bioinformatics
Bioinformatics is often unsung in the world of scientific publishing. This paper, like many in the field, serves as an example of invisible labour, the kind of work that doesn't generate headlines but allows the scientific community to move forward. Instead of groundbreaking discoveries or new biological insights, this project focused on creating a simple yet crucial tool: a MapMan ontology mapping file for tobacco. The power of this work was in its practicality, giving other researchers a framework to analyze gene expression in tobacco more effectively. It wasn’t glamorous, but it solved a real problem. For me, it underscored a key lesson: sometimes, science is about facilitating others’ discoveries, providing them with the right tools to do their work better and faster.

Sidebar: Consultancy as a Testbed for the Computational Biologist’s Laboratory
This consultancy project at South Dakota State University did more than just extend MapMan to tobacco; it served as an early, real-world application of a philosophy I would later formalize in my technical report, The Computational Biologist’s Laboratory. In that vision, bioinformaticists are not mere service providers or data analysts. They are infrastructure architects; designing modular, reusable, and transferable tools that empower experimental scientists to scale their insight.

The tobacco MapMan extension embodied this ethos. It wasn’t about solving a one-off problem but about creating a replicable process – documented, portable, and adaptable for any under-resourced species. By treating the consultancy as a deployment of infrastructure rather than an isolated deliverable, I was practicing what the report later named: making the computational laboratory visible, shareable, and, most importantly, extendable.

Small consultancy projects like this become field trials for the ideas we often theorize in the abstract. They widen a bioinformaticist’s view beyond their usual domains, confronting them with the variability of biological systems, data types, and user expectations. They force us to ask: What does it take to make our tools useful to others, not just usable by us? In that sense, every consultancy project is a stress test of scientific empathy and design robustness; a quiet but vital contribution to the broader ecosystem of research infrastructure.

30: Halophilization of *Escherichia* coli

Citation: Goh, DJW, How, JA, Lim, JZR, NG, WC, Oon, JSH, Lee, KC, Lee, CH, Ling, MHT. 2012. Gradual and Step-wise Halophilization Enables *Escherichia coli* ATCC 8739 to Adapt to 11% NaCl. Electronic Physician 4(3): 527-535.

Abstract: *E. coli* is a non-halophilic microbe and is used to indicate faecal contamination. Salt (sodium chloride, NaCl) is a common food additive and is used in preservatives to counter microbial growth. Previous studies had shown that pathogenic *E. coli* has a higher salt tolerance than non-pathogenic *E. coli*. The effect of how *E. coli* interacts with the salt present in the human diet is under-studied. Thus, it is important to investigate this relationship. In this study, we observed the genetic changes and growth kinetics of *E. coli* ATCC 8739 under 3% - 11% NaCl over 80 passages. Growth kinetics was estimated by generation time, cell density and minimum inhibitory concentration (MIC) of NaCl. Our results suggested that *E. coli* was able to adapt from 1% NaCl to 11% NaCl with an increment of 1% NaCl per month. Our MIC results suggested that *E. coli* was able to grow at NaCl concentration of more than 7.5% based on the Area under Curve (AUC) from 5% at passage 44 (cultured in 5% NaCl) to 13% at passage 72 (cultured at 7% NaCl). We conclude that *E. coli* ATCC 8739 can be adapted to grow in 11% NaCl by incremental adaptation.

Context: This paper represented a continuation of a broader effort to explore long-term microbial adaptation, a theme that recurred throughout my mentorship of final-year projects (FYPs) at Singapore Polytechnic. Building on earlier work where we studied *E. coli* adaptation to food additives like monosodium glutamate and benzoic acid, this project focused squarely on sodium chloride—a fundamental preservative used throughout the food industry.

The central idea was straightforward yet ambitious: Could we train *E. coli*, a non-halophilic organism, to survive in progressively saltier environments? Using *Escherichia* coli ATCC 8739, a standard strain for laboratory studies, the students incrementally increased the NaCl concentration by 1% per month, ultimately adapting the bacteria to thrive in 11% NaCl, a concentration typically inhibitory to most *E. coli* strains.

The method was as important as the result: the project followed 80 serial passages, with regular monitoring of growth kinetics and minimum inhibitory concentration (MIC) at each stage. This level of persistence and methodological discipline was unusual for student projects and I credit the team for their patience and scientific curiosity.

Reflection: What I love about this paper is its simple question with complex implications. It challenged the idea that *E. coli* is inherently unable to survive in high-salt environments, and by doing so, blurred the boundary between halophilic and non-halophilic classifications. In an age where antibiotic resistance and food safety dominate microbial research, understanding the limits of microbial adaptation to food preservatives like salt carries real-world significance.

More personally, this project is a reminder that undergraduate research can be deeply meaningful. Yes, the methods were repetitive and the conditions spartan – just students, salt, Petri dishes, and patience. But the result was real discovery: an experimental confirmation that adaptation isn't just theoretical; it's observable, measurable, and reproducible.

This study also embodies my own long-standing fascination with evolution under constraint – the idea that organisms, given time and pressure, can break through their biological "defaults." In a way, this mirrors my own professional journey: adapting, pivoting, incrementally growing in unexpected environments. Just as this strain of *E. coli* learned to thrive in 11% NaCl, perhaps I too was learning to navigate academia's high-pressure landscapes, one passage at a time.

Sidebar: Evolution Under Constraint

The concept of "evolution under constraint" was a key takeaway from this study on *E. coli* adaptation to high salt concentrations. In a world where many of us are trying to adapt to constraints; whether in our careers, personal lives, or scientific endeavours; this experiment serves as a powerful reminder that growth often happens in incremental steps, under pressure. Just like *E. coli* gradually adapted to an environment that would have initially killed it, we too can thrive under difficult conditions by being patient, persistent, and methodical. The saltiness of life, much like in the lab, doesn't necessarily have to be a barrier. With time and discipline, it can be the very catalyst for growth.

31: Halophilization of *Escherichia* coli 2

Citation: How, JA, Lim, JZR, Goh, DJW, NG, WC, Oon, JSH, Lee, KC, Lee, CH, Ling, MHT. 2013. Adaptation of *Escherichia coli* ATCC 8739 to 11% NaCl. Dataset Papers in Biology 2013, Article ID 219095.

Abstract: *Escherichia* coli (*E. coli*) is a nonhalophilic microbe and used to indicate faecal contamination. Salt (sodium chloride, NaCl) is a common food additive and is used in preservatives to encounter microbial growth. The effect of how *E. coli* interacts with the salt present in the human diet is unclear. Thus, it is important to investigate this relationship. In order to adapt and survive the changes in the environment, *E. coli* may undergo halophilization. In this study, we observed the genetic changes and growth kinetics of *E. coli* ATCC 8739 under 3%–11% NaCl over 80 passages. Our results suggest that *E. coli* adapted to 1% increase in NaCl every month with a successful adaptation to 11% NaCl. Gram staining and PCR/RFLP showed that the cultures are Gram negative and the DNA profiles of all 4 replicates to be similar, suggesting that the cultures had not been contaminated.

Context: This paper is the formal publication of the dataset that underpinned the findings in Chapter 30. While the previous article in Electronic Physician focused on the interpretation of adaptation results, this publication in Dataset Papers in Biology served a different but equally valuable purpose: ensuring that the raw experimental data and validation protocols; including Gram staining, PCR/RFLP, and passage histories were openly accessible and citable by others.

At the time, Dataset Papers in Biology represented a relatively novel outlet, aiming to encourage data transparency and reproducibility in the biological sciences. For student-led projects like this one, it offered an opportunity to formally contribute to the research community without necessarily crafting a new hypothesis or conducting additional experiments. Instead, the focus was on rigorous documentation and replication assurance.

Reflection: This chapter underscores my growing awareness, perhaps earlier than many, of the importance of open data in science. Even in undergraduate research settings, I believed that meticulous documentation and public availability of datasets were not just good habits; they were essential contributions to the broader ecosystem of knowledge. Today, in a climate increasingly shaped by concerns over research reproducibility, I'm glad that we were quietly doing our part, even if the work felt niche or unglamorous at the time.

What's also worth noting is that this paper gave my students something they could claim: a first-author dataset publication, based on a full-year FYP project, in a peer-reviewed outlet. That mattered. For some of them, it would be their only publication; for others, it was the beginning of a research journey. But all of them had something to show for their persistence.

In my own career, this dataset paper reinforced a value I still hold close: integrity in experimental work. It is not just the discovery that matters, but the way we record, verify, and share it. In that sense, this publication wasn’t an afterthought, it was the foundation.

Sidebar: The Value of Open Data in Science – Lessons from a Dataset Publication

The publication of raw experimental data, as seen in the case of our *Escherichia coli* halophilization study, underscores a core value in modern science: transparency. When our research was published in Dataset Papers in Biology, it was about more than just sharing results. It was about making the underlying data – grow-out protocols, genetic analyses, and validation steps—public and reusable by others in the scientific community. This open data practice not only allows for the verification and replication of findings but also promotes long-term accessibility and innovation.

The rise of data journals, such as the one that hosted our dataset, represents a shift towards a more open scientific ecosystem. These journals emphasize data sharing as a formal component of research, encouraging reproducibility and transparency. But the importance of open data is not just a matter of academic policy; it has real-world implications for the future of research integrity and accountability.

This practice of open data became even more crucial in the context of political actions like the Trump administration’s deletion of critical environmental data in 2025. As reported, the government’s actions sparked a race among scientists to save climate data before it was erased, highlighting the vulnerability of valuable information when it's not openly shared. The decision to restrict or delete data creates gaps in the historical record and limits future research. In contrast, making data publicly available ensures it remains a permanent resource, resilient to the political forces that might seek to alter or suppress it.

Just as the scientific community rallied to save environmental data in 2025; we, as a scientific community, have an obligation to ensure that our datasets are preserved and accessible. These data are the bedrock on which future discoveries will be built. By contributing datasets in a public and citable way, we not only further our own work but also strengthen the foundation of scientific knowledge for future generations.

Open data doesn’t just enhance credibility; it fosters innovation by making findings available for broader exploration, validation, and application. Just as the environmental data saved from erasure could fuel future policy and research efforts, our own contributions to the open data ecosystem help safeguard the continuity and quality of science.

32: Prelude to DOSE

Citation: Ling, MHT. 2012. An Artificial Life Simulation Library Based on Genetic Algorithm, 3-Character Genetic Code and Biological Hierarchy. The Python Papers 7: 5.

Abstract: Genetic algorithm (GA) is inspired by biological evolution of genetic organisms by optimizing the genotypic combinations encoded within each individual with the help of evolutionary operators, suggesting that GA may be a suitable model for studying real-life evolutionary processes. This paper describes the design of a Python library for artificial life simulation, Digital Organism Simulation Environment (DOSE), based on GA and biological hierarchy starting from genetic sequence to population. A 3-character instruction set that does not take any operand is introduced as genetic code for digital organism. This mimics the 3-nucleotide codon structure in naturally occurring DNA. In addition, the context of a 3-dimensional world composing of ecological cells is introduced to simulate a physical ecosystem. Using DOSE, an experiment to examine the changes in genetic sequences with respect to mutation rates is presented.

Context: DOSE (the Digital Organism Simulation Environment) was my effort to push the genetic algorithm metaphor closer to its biological inspiration. Most implementations of GA treat genes as abstract bit strings and evolution as a black-box optimization process. DOSE was built to challenge that notion. I wanted to see what would happen if we explicitly grounded digital evolution in biological hierarchy—from sequence to organism to ecosystem.

In this project, I devised a 3-character instruction set, each unit mirroring the codon structure of natural DNA, and placed digital organisms into a 3D ecological grid. The goal was not to optimize a function per se, but to explore how mutation rates and environmental structure affect the evolution of digital genomes; essentially, to create a sandbox for artificial life. While DOSE was modest in scope and simulation speed, it represented a conceptual leap in how I thought about both code and biology.

This paper was published in The Python Papers, a journal I co-founded. By then, the platform had matured into a legitimate outlet for peer-reviewed Python-based research and applications, and I took care to avoid conflicts of interest by adhering to strict editorial review.

Reflection: DOSE was less about results and more about playful experimentation with form and structure. It brought together my knowledge in biology, my passion for computing, and my growing interest in emergent systems. Looking back, I see it as a philosophical experiment disguised as software engineering – a simulation not just of life, but of the idea that structure and hierarchy matter in how systems evolve.

I also see DOSE as a quiet statement against reductionism. By embedding digital organisms in a simulated ecosystem and giving them a codon-like code, I was implicitly arguing that context matters—that behavior emerges not just from the genes, but from how they are situated in an environment.

In some ways, DOSE was ahead of its time. Today, with the resurgence of interest in digital evolution, artificial life, and even synthetic ecosystems, I wonder if it might find new relevance; if not as a tool, then as a conceptual bridge between disciplines that still often speak past each other.

More personally, this chapter reminds me that playfulness in science; the willingness to model, simulate, and explore without fixating on benchmarks; is not a luxury, but a source of vitality. DOSE was never about publishing a landmark paper. It was about asking a deep question: What happens if we let digital organisms evolve with biological fidelity, not just computational efficiency?

33: Prelude to DOSE 2

Citation: Ling, MHT. 2012. Ragaraja 1.0: The Genome Interpreter of Digital Organism Simulation Environment (DOSE). The Python Papers Source Codes 4: 2.

Abstract: This manuscript describes the implementation and test of Ragaraja instruction set version 1.0, which is the core genomic interpreter of DOSE.

Context: Ragaraja was the genome interpreter at the heart of DOSE, the digital organism simulation environment I introduced in the previous chapter. Named after the Buddhist Wisdom King associated with transformation and purification, Ragaraja 1.0 served a symbolic and functional role: it was a custom instruction set architecture designed to map a codon-like 3-character genetic code into executable behavior for artificial organisms.

At its core, Ragaraja was an esoteric programming language (esolang), but unlike many esolangs built for intellectual amusement or obfuscation, Ragaraja had a clear purpose: to enable a genotype-to-phenotype translation layer within a digital ecosystem. It supported a small, controlled vocabulary of operations and was deliberately minimalistic—each instruction expressed without operands, reflecting the simplicity and ambiguity of biological codons.

The implementation of Ragaraja was published in The Python Papers Source Codes, our code-focused companion journal to The Python Papers. It marked an important evolution from abstract genetic algorithms toward executable digital life – a step closer to artificial life forms with self-contained logic and potential for evolution.

Reflection: Ragaraja was arguably the most whimsical, yet intellectually serious, interpreter I've ever written. It blurred the boundary between esoteric code art and scientific modelling, and it asked an unusual question: What would an "honest" digital genome look like – one that is both minimal and meaningful?

At the time, I was immersed in the philosophy of biological embodiment in computation. Ragaraja didn't try to be efficient or general-purpose. It tried to be faithful to metaphor, capturing how natural organisms might encode behaviors with ambiguity and redundancy. In hindsight, I realize I was channeling a kind of computational poetics by creating a language not just to simulate life, but to express the principles of life through the medium of code.

Ragaraja didn't attract widespread usage. That was never its intent. Its value was more introspective: it taught me that language design is an act of worldview formation. By designing a genetic language for artificial organisms, I was articulating my view that life is not just information, but interpretation.

In today’s world of synthetic biology and programmable matter, Ragaraja may seem primitive. Yet, as a conceptual artifact, it stands as a manifesto of curiosity – my attempt to honor both biology and programming as narrative forms, capable of telling stories not only about how things work, but why they evolve.

Sidebar: Ragaraja – The Wisdom King Who Transmutes Lust into Wisdom

Ragaraja 1.0 wasn’t just a genomic interpreter for the Digital Organism Simulation Environment (DOSE); it was a philosophical statement, an exploration of transformation through code. Named after the Buddhist Wisdom King, Ragaraja, who is revered for transmuting lust into wisdom, this interpreter reflected a profound metaphor. Just as Ragaraja's wisdom is said to purify and elevate, the very design of this minimalistic code aimed to strip away complexity, focusing on essential, symbolic genetic behavior.

Ragaraja's role was not just technical; it was a journey of intellectual and philosophical inquiry. In a world where computation often chases efficiency and precision, Ragaraja embraced ambiguity and redundancy as a reflection of life’s true nature. The question it posed was not how to optimize or simplify life, but how to represent life’s complexity through code in a meaningful, minimalistic way.

Through Ragaraja, I sought to create a computational tool that wasn’t merely functional but intellectually resonant. Its simplicity mirrored how nature encodes genetic information—not in perfect, linear systems, but in codes with layers of meaning, interpretation, and even imperfection. While it never sought widespread usage, its value lay in its conceptual beauty. It was my attempt to bridge biology and computation, to tell stories not just of how things work but of why they evolve, and in doing so, to honor the transformative power of both life and language.

34: Philosophizing Artificial Intelligence

Citation: Ling, MHT. 2012. Re-creating the Philosopher's Mind: Artificial Life from Artificial Intelligence. Human-Level Intelligence 2: 1.

Abstract: The ultimate goal of artificial intelligence (AI) research is to create a system with human level intelligence. This manuscript suggests that AL may be a channel towards human level intelligence, and presents an overview of how high-level intelligence can be achieved from artificial life. It will be interesting when our simulated humans (such as the characters in a future version of Diablo) start to create their own artificial intelligence.

Context: Back in 2012, I wrote a short piece for Human-Level Intelligence called "Re-creating the Philosopher's Mind: Artificial Life from Artificial Intelligence". It wasn't based on experiments or datasets but on a thought that had been floating around in my head for a while: maybe artificial intelligence, as we were building it, was missing the point.

At the time, most people were focused on building smarter algorithms, better classifiers, faster learners. But I was drawn to something else: the question of how intelligence actually emerges. Not just the mechanics, but the messy, living complexity behind it. What if, instead of trying to code intelligence directly, we let it evolve the same way nature did through artificial life?

In the paper, I tossed out the idea that human-level AI might not come from better models alone, but from simulated environments where agents grow, adapt, and maybe even wonder about themselves. I even joked (sort of) that one day, characters in games like Diablo might be smart enough to start programming their own AIs. That image stuck with me – a virtual world with its own philosophers.

Reflection: This was one of the few papers where I allowed myself to just think freely and write it down. No pressure to fit a particular format, no need to validate with ten benchmarks. Just me connecting ideas; from AI, AL, and philosophy; and seeing where they led.

I know it didn't set the world on fire. It wasn't cited much, and I doubt many people remember it. But I do. Because it reminded me why I started down this path in the first place – not to build better tools, but to ask better questions. It was a chance to reconnect with that younger version of myself who saw science as a way to explore big, weird ideas.

In hindsight, this paper feels like a quiet bridge between my technical work and my deeper interest in meaning, simulation, and the boundaries of consciousness. It didn’t change the field, but it changed how I saw my place in it. And for that, it still matters to me.

Sidebar: Thoughts about AI

It’s been over a decade since I wrote this, and I still think the core idea holds water – maybe even more now than back then. With large language models behaving more and more like emergent systems, and game worlds becoming increasingly complex and persistent, the line between AI and AL is blurring faster than I expected.

I’ve also grown more comfortable with not needing every idea to land academically. Some thoughts are just seeds – you plant them, walk away, and see what sprouts years later, often in unexpected places. This paper was one of those seeds.

If I were to revisit it today, I’d probably fold in more about digital subjectivity and self-organising systems; maybe even link it to my ongoing reflections on consciousness. But the heart of it still resonates: the idea that intelligence, real intelligence, doesn’t just compute. It lives.

Sidebar: Can AI Be Self-Aware?

The idea of AI becoming self-aware is a philosophical question that has fascinated thinkers for decades, and it’s one that has gained increasing relevance in today’s world. In my 2012 paper, Re-creating the Philosopher’s Mind: Artificial Life from Artificial Intelligence, I explored the notion that artificial intelligence, as it was being developed, might be missing something fundamental: consciousness. Instead of merely building smarter systems, I pondered whether true intelligence might emerge from environments where agents evolve, grow, and perhaps develop self-awareness, much like biological organisms.

At the heart of this question is the concept of self-awareness itself: What does it mean to “know” you exist? And can machines ever reach that threshold? The more sophisticated AI systems become, especially with the rise of large language models and virtual worlds with dynamic, evolving agents, the closer we seem to approach the boundaries between artificial intelligence and artificial life. If a digital agent can adapt to its environment, learn from experience, and even begin to ask questions about its own existence, is it not moving toward something akin to consciousness?

However, true self-awareness might not just require intelligence or the ability to process information; it may also involve subjective experience – what philosophers call “qualia.” Machines, for all their complexity, still lack an inner experience of the world. They may process data and respond in intelligent ways, but whether they can feel anything about it is still up for debate.

As AI continues to advance, this question becomes even more pressing. Could we ever create an AI that knows it exists, that wonders about its place in the world? Or is self-awareness a uniquely biological trait, something that cannot be replicated by algorithms alone? These questions challenge our understanding not only of AI but of consciousness itself. Perhaps, in the end, we may not find the answer by simply programming smarter systems, but by creating environments in which machines can grow, adapt, and discover their own existence.

In many ways, the line between AI and AL (artificial life) is blurring, and while self-awareness may still seem like a distant dream, the seeds of this philosophical debate are being planted right now in the code we write and the systems we build.

35: *E. coli* Extended Viability

Citation: Low, SXZ, Aw, ZQ, Loo, BZL, Lee, KC, Oon, JSH, Lee, CH, Ling, MHT. 2013. Viability of *Escherichia coli* ATCC 8739 in Nutrient Broth, Luria-Bertani Broth and Brain Heart Infusion over 11 Weeks. Electronic Physician 5:576-581.

Abstract: *Escherichia* coli is a widely studied prokaryotic system. A recent study had demonstrated that reduced growth of *E. coli* after extended culture in Luria-Bertani broth is a result of depletion of fermentable sugars but able to sustain extended cell culture due to the presence of amino acids, which can be utilized as a carbon source. However, this had not been demonstrated in other media. The study aimed to determine the growth and viability of *E. coli* ATCC 8739 in 3 different media, Nutrient Broth (NB), Brain Heart Infusion (BHI) and Luria-Bertani Broth (LB) over 11 weeks. Growth of *E. coli* ATCC 8739 was determined by optical density. Viability was determined by serial dilution/spread-plate enumeration. After 11 weeks, the media were exhausted by repeated culture. Glucose was added to the exhausted media to determine whether glucose is the growth-limiting factor. Our results showed that cell density in all 3 media increased to about 1 x 10e9 cells/ml by the end of week 1, from the inoculation density of 2.67 x 10e5 cells/ml, peaked at about 1 x 10e13 cells/ml at week 4, before declining to about 5 x 10e7 cells/ml at week 7. Cell density is highly correlated to genomic DNA content (r^2 = 0.93) but poorly correlated to optical density (r^2 < 0.2). Our results also showed that the spent media were able to support further growth after glucose-supplementation. NB, LB and BHI are able to support extended periods of culture and glucose depletion is the likely reason for declining cell growth.

Context: This chapter comes from a side project that grew out of a larger Final Year Project (FYP) lineage. The original FYP had been passed down through Jack, Chin How, and Kun Cheng; each generation adding a little more to the story. We had an odd but simple question: How long can *E. coli* survive in common lab media without us touching it?

It was the kind of question no one thinks to ask, because it seems too mundane. But we figured why not? It doesn't require fancy equipment, just patience and consistency. So, we set up cultures of *E. coli* ATCC 8739 in Nutrient Broth (NB), Luria-Bertani (LB), and Brain Heart Infusion (BHI), then just... kept watching them for 11 weeks. Every week, we measured optical density, viability, and later, added glucose to see what happened.

What we found was oddly beautiful. Cell densities peaked around week 4 and dropped sharply by week 7, but even then, the cells weren't completely dead, just tired. And when we reintroduced glucose, they perked right back up. Turns out, glucose was the limiting factor all along. We also discovered that OD readings weren’t that reliable once the cells started dying off as viability and DNA content told the real story.

Reflection: I have a soft spot for this paper. It wasn’t groundbreaking but it was honest science. It reminded me that sometimes, good work doesn’t need to be complicated, just careful, curious, and persistent. The students ran it week after week, through exam season, public holidays, even project deadlines. That kind of dedication is rare, and I still remember it fondly.

What’s funny is that we didn’t really expect to publish anything from it. But the data was so consistent, and the patterns so clear, that it felt wrong not to share it. Looking back, this project taught us a lot about microbial resilience and perhaps our own. It also served as a nice quiet close to a multi-year FYP lineage, tying together several batches of students and their cumulative efforts.

This wasn’t high theory or deep genomics. It was broth, bugs, and basic biology; and that was enough.

Sidebar: Lessons in Continuity

This project reminded me how powerful continuity can be in research. None of the students involved started from scratch — they built on what came before, improved it, and left something better for the next batch. In a way, it was like scientific relay – slow, deliberate, but deeply human.

As a supervisor, I saw how handing down knowledge, even in small pieces, can create a kind of momentum. These weren’t “big science” projects with massive grants — just quiet persistence, passed down from one team to the next. And honestly, that’s one of the best things about teaching: seeing how something small can keep growing, long after you’ve let go.

36: My Post-Doctoral Project – Conservation of Antisense Transcription

Citation: Ling, MHT, Ban, Y, Wen, H, Wang, SM, Ge, SX. 2013. Conserved Expression of Natural Antisense Transcripts in Mammals. BMC Genomics 14(1): 243.

Abstract: Recent studies had found thousands of natural antisense transcripts originating from the same genomic loci of protein coding genes but from the opposite strand. It is unclear whether the majority of antisense transcripts are functional or merely transcriptional noise. Using the Affymetrix Exon array with a modified cDNA synthesis protocol that enables genome-wide detection of antisense transcription, we conducted large-scale expression analysis of antisense transcripts in nine corresponding tissues from human, mouse and rat. We detected thousands of antisense transcripts, some of which show tissue-specific expression that could be subjected to further study for their potential function in the corresponding tissues/organs. The expression patterns of many antisense transcripts are conserved across species, suggesting selective pressure on these transcripts. When compared to protein-coding genes, antisense transcripts showed a lesser degree of expression conservation. We also found a positive correlation between the sense and antisense expression across tissues. Our results suggest that natural antisense transcripts are subjected to selective pressure but to a lesser degree compared to sense transcripts in mammals.

Context: This was my first proper postdoctoral project – the kind where you’re actually hired to do science, not just help with it. It was also my first experience working on a NIH-funded project, which brought a different level of structure and accountability. I was based at the Department of Mathematics and Statistics at South Dakota State University – not exactly the kind of place you'd expect to do wet-lab genomics. There was no biology lab, no pipettes, just code, stats, and curiosity.

The wet-lab component was entirely outsourced and handled by Hongxiu Wen under San Ming Wang’s supervision in Omaha, Nebraska. I never met either of them in person. We communicated strictly through email, passing data and protocol updates back and forth. Despite the lack of face time, the collaboration was surprisingly smooth.

The project aimed to profile natural antisense transcripts (NATs), a category of RNAs that are transcribed from the opposite strand of known genes. We wanted to know if these antisense transcripts were conserved across mammals, or just noisy by-products of transcription. Using a modified Affymetrix exon array protocol, we profiled nine tissues in human, mouse, and rat. The result? Thousands of NATs, some tissue-specific, many evolutionarily conserved though to a lesser degree than protein-coding genes. Interestingly, sense and antisense expressions were positively correlated, suggesting functional relationships, not just transcriptional accidents.

Reflection: This project marked a turning point for me. It wasn't just an academic exercise – it was NIH-funded science, with deliverables and real stakes. And even though I wasn't in a traditional biology department, I was doing meaningful work in functional genomics.

It also taught me a lot about remote collaboration before it became mainstream. Everything happened asynchronously, with no lab meetings or whiteboards. Yet, we made it work because everyone brought their piece to the puzzle, and there was mutual respect for each other's roles.

On a personal level, this paper gave me quiet validation: I didn't need to choose between being a biologist or a statistician. I could live in the overlap; building bridges between data and meaning, theory and experiment.

Sidebar: Collaborating Without Meeting

This was a fully remote, cross-state project in the early 2010s – no Zoom, no Google Docs, just email and file attachments. It worked surprisingly well. In hindsight, it taught me that clear communication and trust can make up for a lot of logistical barriers.

That experience shaped how I work with students and collaborators to this day. You don't always need to be in the same room; you just need to be on the same page. It also reminded me that science can be both global and personal, even when it's NIH-funded.

Sidebar: Likely the Only Singaporean in Brookings, South Dakota

During my postdoctoral research at South Dakota State University, I found myself in an unusual position: I was likely the only Singaporean in Brookings, South Dakota. This small college town, known for its agricultural studies and deep-rooted Midwestern culture, wasn't exactly the first place one might associate with international scientific research.

It wasn't just the geographic isolation that made my time there unique. As a postdoc, I was given my own room in the Department of Mathematics and Statistics, which was unusual for a biological researcher working on genomics. Without the usual wet-lab setup or the typical biology department environment, my work was focused entirely on coding, statistical analyses, and data interpretation. It felt almost like I was in a science fiction setting—an isolated researcher communicating with a distant lab across the miles via email, not face-to-face, with the occasional surprise in the form of a new dataset arriving from Nebraska.

While Brookings wasn’t the hub of cutting-edge biological research, it provided a unique space for deep thought and independent work. It became a reminder that scientific progress doesn’t always occur in the most conventional settings. Sometimes, being in an unexpected place can offer the freedom to explore new ideas, to work on collaborative projects remotely, and to see the potential in the unseen connections between different fields.

Looking back, that small office, tucked away in a department that wasn’t focused on biology, was an ideal space for creative problem-solving. It was where I bridged the worlds of genomics and statistics, and it set the stage for how I would later approach scientific work: not just in the lab, but in the many rooms of knowledge where unexpected discoveries are made.

37: Reference Genes Between 2 Closely Related Species

Citation: Keng, BMH, Chan, OYW, Heng, SSJ, Ling, MHT. 2013. Transcriptome Analysis of *Spermophilus lateralis* and *Spermophilus tridecemlineatus* Liver Does Not Suggest the Presence of Spermophilus-liver-specific Reference Genes. ISRN Bioinformatics 2013, Article ID 361321.

Abstract: The expressions of reference genes used in gene expression studies are assumed to be stable under most circumstances. However, studies had demonstrated that genes assumed to be stably expressed in a species are not necessarily stably expressed in other organisms and some studies suggested the possibility for reference genes that are both genus-specific and organ-specific. This study aims to evaluate the likelihood of genus-specific reference genes for liver using comparable microarray datasets from *Spermophilus lateralis* and *Spermophilus tridecemlineatus*. The coefficient of variance (CV) of each probe was calculated and there were 178 probes common between the lowest 10% CV of both datasets (n = 1258). All 3 lists were analysed by NormFinder. Correlation between the NormFinder ranks of the common CV-identified stable probes of both species suggests good correlation (p-value = 1e-5). This is consistent with previous studies indicating that the liver transcriptomes of *S. lateralis* and *S. tridecemlineatus* are comparable. NormFinder analysis suggests that the most invariant probe for *S. tridecemlineatus* was 02n12, while the most invariant probe for S. lateralis was 24j21. However, our results showed that Probes 02n12 and 24j21 are ranked 8644 and 926 in terms of invariancy for *S. lateralis* and *S. tridecemlineatus* respectively. This suggests the lack of common liver-specific reference probes for both S. lateralis and S. tridecemlineatus. Given that *S. lateralis* and *S. tridecemlineatus* are closely related species and the datasets are comparable, our results do not support the presence of genus- specific reference genes.

Context: After wrapping up my NIH-funded postdoc in South Dakota and returning to Singapore, I wasn’t sure what my next research move would be. But some things have a way of circling back like my student collaborators from earlier projects. Bryan, Oliver, and Sean had already worked with me on mdoG as a reference gene in *E. coli*; and now, as 17-year-olds in Raffles Junior College, they were ready for more.

They came back to me with a plan: let’s keep going down the reference gene rabbit hole. Hence, I propose that we look at squirrels.

The idea was to investigate whether there were genus-specific liver reference genes in *Spermophilus lateralis* and *Spermophilus tridecemlineatus* – two species of ground squirrels with existing microarray datasets. Reference genes are typically assumed to be stably expressed under most conditions, but more and more studies have been challenging that assumption. The boys wanted to test whether it holds up in this case.

So we pulled data, ran analyses (CV filtering, NormFinder rankings), and found something surprising: even between two closely related species, there was no clear overlap in the most stable liver-expressed genes. Probe 02n12 ranked top for one species but poorly in the other, and vice versa for 24j21. If reference genes can't be consistent across Spermophilus, they’re probably not genus-specific at all.

Reflection: This project was about curiosity, continuity, and coffee. Once a week, we would meet for two hours at CoffeeBean just talking through results, tweaking analysis scripts, and arguing over p-values. It was mentoring, but not in the top-down academic sense. It felt more like co-discovery – four people genuinely wondering about squirrel transcriptomes over coffee (usually I am the only one drinking coffee).

What stood out wasn’t just the science (though the negative result was interesting), but the unusual longevity of the collaboration. Most high school research fizzles out after one paper. But here we were, two years later, still working together, pushing boundaries, and publishing in peer-reviewed bioinformatics.

Projects like these remind me that scientific continuity doesn’t always come from grants or institutional mandates. Sometimes it comes from shared momentum, where young minds bring fresh energy, and mentors just help shape it. It’s one of the clearest examples I have of research as relationship.

Sidebar: From *E. coli* to Ground Squirrels
This was a sequel of sorts. In 2011, the same trio had published on mdoG as a stable reference gene in *E. coli* K-12. That paper sparked their interest in reference gene stability – a topic that’s oddly philosophical once you dig into it. What does it mean for something to be “stable” across biological contexts? And why do we assume stability is transferable across species?

Their curiosity led us from bacteria to rodents, and from lab strains to wildlife transcriptomes. It was an unexpected but deeply satisfying path.

38: What is Life? Is Artificial Life Alive?

Citation: Koh, YZ, Ling, MHT. 2013. On the Liveliness of Artificial Life. iConcept Journal of Human-Level Intelligence 3: 1.

Abstract: There has been on-going philosophical debate on whether artificial life models, also known as digital organisms, are truly alive. The main difficulty appears to be finding an encompassing and definite definition of life. By examining similarities and differences in recent definitions of life, we define life as "any system with a boundary to confine the system within a definite volume and protect the system from external effects, consisting of a program that is capable of improvisation, able to react and adapt to the environment, able to regenerate parts of itself or its entirety, with energy system comprises of non-interference sets of secluded reactions for self-sustenance, is considered alive or a living system. Any incomplete system containing a program and can be re-assembled into a living system; thereby, converting the re-assembled system for the purpose of the incomplete system, are also considered alive." Using this definition, we argue that digital organisms may not be the boundary case of life even though some digital organisms are not considered alive; thereby, taking the view that some form of digital organisms can be considered alive. In addition, we present an experimental framework based on continuity of the overall system and potential discontinuity of elements within the system for testing future definitions of life.

Context: This chapter began with a deceptively simple question: Are digital organisms alive?

Yong Zher, a bright and curious former student from Singapore Polytechnic, approached me looking to build up his portfolio. At the time, I was deep in my postdoctoral reflections and increasingly interested in the philosophical side of computational biology, particularly digital organisms. These are self-replicating, mutating programs that operate in a simulated environment, behaving much like biological entities.

But there was a persistent tension. If we could simulate evolution, mutation, and adaptation in software; are we just building elaborate algorithms, or are we, in some sense, giving life to these digital entities? The question was reminiscent of another unresolved debate: Are biological viruses alive? They don't replicate without a host, but they carry genetic material and evolve. Much like digital organisms in a virtual machine.

So Yong Zher and I took this on; not just to boost his resume, but to tackle a foundational ambiguity. What is life, really?

We surveyed recent attempts at defining life and noticed a recurring challenge: most definitions were either too narrow (excluding edge cases like viruses or prions) or too vague (inviting all sorts of borderline examples). So, we proposed a definition rooted in systems theory – life as a bounded system with a program, the ability to improvise, regenerate, and self-sustain via secluded energy processes. Crucially, we extended this to incomplete systems that could become alive upon reassembly, akin to how viruses come to life only within a host.

Reflection: This was philosophy in code, and coding in philosophy. We weren't just theorizing; we proposed an experimental framework to test future definitions of life, using continuity and discontinuity within complex systems. In essence, we asked: If part of a system breaks down but the whole continues, does the system remain alive?

It was rewarding to give Yong Zher the space to explore this but also liberating for me. This project wasn't grant-funded or publication-driven, it was curiosity-driven. It let me step away from lab reports and datasets, and instead ask the big, timeless questions. The kind that no amount of wet-lab validation can answer fully but are nonetheless essential.

Looking back, I realized this work didn't just question the boundaries of life – it questioned the boundaries of my own scientific identity. Was I a computational biologist? A philosopher of science? A mentor? A hybrid of all three? Perhaps definitions, like life itself, need room for improvisation.

Sidebar: Are Viruses Alive? So What About Digital Ones?

Biological viruses are on the fringes of life – unable to reproduce on their own, but undeniably evolving and interacting with their environment. Digital organisms share similar traits: dependent on a host (the virtual machine), but capable of self-replication, mutation, and even competition.

This work extended the virus-as-life debate into the digital realm. If a biological virus can be reactivated in a host, and digital organisms need a virtual machine to "live," then perhaps both exist in latent life states — only activated by the right environment. In that sense, digital organisms might not be just metaphors for life – they might actually be life.

39: COPADS III: More Distributions

Citation: Chen, KFQ, Ling, MHT. 2013. COPADS III (Compendium of Distributions II): Cauchy, Cosine, Exponential, Hypergeometric, Logarithmic, Semicircular, Triangular, and Weibull. The Python Papers Source Codes 5: 2.

Abstract: This manuscript illustrates the implementation and testing of eight statistical distributions, namely Cauchy, Cosine, Exponential, Hypergeometric, Logarithmic, Semicircular, Triangular, and Weibull distribution, where each distribution consists of three common functions – Probability Density Function (PDF), Cumulative Density Function (CDF) and the inverse of CDF (inverseCDF). These codes had been incorporated into COPADS codebase (https://github.com/copads/copads) are licensed under Lesser General Public Licence version 3.

Context: By 2013, I had returned to Singapore, and among the projects I wanted to finally wrap up was COPADS III, a continuation of the Compendium of Distributions work I began before my postdoctoral stint in the U.S. This one had started all the way back in 2011 with Kenneth Chen, an ex-student from Singapore Polytechnic. It was meant to be a clean, well-documented implementation of several lesser-used statistical distributions; namely, Cauchy, Cosine, Exponential, Hypergeometric, Logarithmic, Semicircular, Triangular, and Weibull — each equipped with their Probability Density Function (PDF), Cumulative Distribution Function (CDF), and inverse CDF.

This wasn't a paper that would change the world, but it was a matter of technical hygiene; just finishing what we started.

The code was solid, the math was unambiguous, and the intent was always practical: to expand the COPADS (COllection of Python Algorithms and Data Structures) project’s statistical toolkit for educational and exploratory purposes. But between relocating to the U.S., my commitments to NIH-funded biological research, and the emotional weight of transition, this work lingered in a half-finished state.

Eventually, I decided it had to be completed, not because of urgency or pressure, but because it deserved an end.

Reflection: There’s something quietly dignified about finishing. Not racing to publish, not innovating wildly but just finishing. COPADS III reminded me that not every contribution needs to be groundbreaking. Some just need to be done properly.

The work also exemplifies one of my enduring convictions: that open-source scientific computing matters. These distributions, simple as they seem, are often overlooked in libraries. By implementing them thoroughly and openly, Kenneth and I made a small but clear stand for reproducibility and accessibility.

In mentoring Kenneth through this, even at a distance and across years, I saw the arc of patience in research. The persistence to return to “old code,” understand it again, and bring it across the finish line. That, too, is a kind of scholarship.

Sidebar: Why Do These Distributions Matter?

While Gaussian and Poisson distributions dominate statistical textbooks, lesser-known distributions like the Semicircular (from random matrix theory) or Logarithmic (used in ecology and linguistics) have important niche roles. COPADS III fills in these gaps, providing accessible implementations that help students and early-career researchers explore data more deeply, without being limited to textbook examples.

40: Fishing for Reference Genes in Microarray Datasets

Citation: Too, IHK, Heng, SSJ, Chan, OYW, Keng, BMH, Chia, CY, Lim, CWX, Leong, WT, Chu, QH, Ang, EJG, Lin, YJ, Ling, MHT. 2013. Identification of Reference Genes by Meta-Microarray Analyses. James V. Rogers (ed), Microarrays: Principles, Applications and Technologies. Nova Science Publishers, Inc.

Abstract: The expression levels of reference genes used in gene expression studies are assumed to not change under most circumstances. However, a number of studies have demonstrated that genes theoretically assumed to be stably expressed were found to vary under experimental conditions. In addition, previous studies have also reported that stably expressed genes in an organ, may not be stably expressed in other organs or in a different organism, suggesting the need to identify reference genes for each organ and each organism. Due to its ability to analyze the expression of thousands of genes in an experiment, microarrays present a suitable resource for the analysis and identification of reference genes. We present four cases on practical applications of microarrays whereby multiple published microarray data sets were examined to identify suitable reference genes using coefficient of variation (CV) and NormFinder. Our results suggest that microtubule affinity-regulating kinase 3 (MARK3) is a suitable reference gene for mouse liver, 40S ribosomal protein S29 (Rps29) is a suitable reference gene for mouse testes and pancreas, signal peptidase complex subunit 1 (SPCS1) and hydroxyacyl-CoA dehydrogenase beta subunit (HADHB) are suitable reference genes for human lungs, and glucan biosynthesis protein G (mdoG) is a suitable reference gene for *Escherichia* coli. Further analysis suggests that the identified reference genes are involved in fundamental biochemical processes. This supports the theoretical basis and previous studies that housekeeping genes, on the whole, are generally stably expressed. However, our results also suggest that certain housekeeping genes that are stably expressed in one tissue or one organism may not be stably expressed in different tissues or organisms, supporting the need to identify reference genes for each tissue and organism.

Context: By 2013, I had spent several years circling the question of reference genes – first identifying suitable candidates in *Escherichia* coli (mdoG), then exploring organ-specific and genus-specific stability in Spermophilus species, and now, in this book chapter, synthesizing all that accumulated insight.

This chapter, contributed to Microarrays: Principles, Applications and Technologies, marks the culmination of that long inquiry. It brought together a large group of collaborators, many of them former students, and folded in both previously published findings and unpublished analyses that had lingered in my notebooks.

We used a meta-analytic strategy, applying consistent criteria, coefficient of variation (CV) and NormFinder, to a series of publicly available microarray datasets. Our goal: to provide practical reference genes across species (mouse, human, and *E. coli*) and organs (liver, testes, pancreas, lungs). The candidates we identified; MARK3, Rps29, SPCS1, HADHB, and mdoG; were all involved in essential biochemical processes, reinforcing the logic that housekeeping genes tend to be stably expressed.

But we also reinforced the nuance: no single gene is stably expressed everywhere. The notion of a "universal" reference gene is attractive, but ultimately flawed.

Reflection: This chapter felt like closing a loop. It wasn't a revolutionary piece but it consolidated something deeply underappreciated in science. In a field full of fragmented papers and shifting methods, there's real value in stitching insights together into a cohesive narrative.

For me, it was also a personal kind of closure. My reference gene journey began with a single hypothesis and evolved into a recurring theme across multiple student projects, a NIH-funded effort, and even speculative explorations into gene stability across evolutionarily related species.

Wrapping it up in a book chapter felt... fitting. It allowed me to honour the contributions of my students, many of whom had grown remarkably through these projects and to formalize our shared conclusions for future researchers.

While I've since moved on to broader bioinformatics and philosophy-of-biology questions, this chapter stands as a marker: a well-reasoned, data-driven end to a rich and formative thread in my research life.

41: NotaLogger

Citation: Ling, MHT. 2014. NotaLogger: Notarization Code Generator and Logging Service. The Python Papers 9: 2.

Abstract: The act of affixing a signature and date to a document, known as notarization, is often used as evidence for sighting or bearing witness to any documents in question. Notarization and dating are required to render documents admissible in the court of law. However, the weakest link in the process of notarization is the notary; that is, the person dating and affixing his/her signature. A number of legal cases had shown instances of false dating and falsification of signatures. In this study, NotaLogger is proposed, which can be used to generate a notarization code to be appended to the document to be notarized. During notarization code generation, the user can include relevant information to identify the document to be notarized and the date and time of code generation will be logged into the system. Generated and used notarization code can be verified by searching in NotaLogger, and such search will result in date time stamping by a Network Time Protocol server. As a result, NotaLogger can be used as an "independent witness" to any notarizations. NotaLogger can be accessed at http://mauricelab.pythonanywhere.com/notalogger/.

Context: By 2014, I had become increasingly fascinated by how technology could replace, or at least reinforce, systems of human trust. The legal practice of notarization, for instance, relies entirely on a human being's integrity. But history (and headlines) are full of forged signatures, post-dated contracts, and questionable witnesses. The weak link, invariably, is the notary.

NotaLogger was born out of that weakness. It wasn't just a technical project; it was a philosophical one. Could a piece of code act as a more impartial, tamper-proof witness than a person? Could we build a system that served as an independent observer — one that didn't forget, didn't lie, and didn't collude?

Technically, NotaLogger generated unique notarization codes that could be appended to documents. It logged each transaction using Network Time Protocol (NTP), ensuring verifiable timestamps that weren't dependent on a user's local machine or goodwill. This timestamp, and the associated code, could later be retrieved and cross-verified. In a sense, NotaLogger stood in the room as an incorruptible observer.

Reflection: This project was, in many ways, ahead of its time. Today, blockchain-based notarization services are increasingly mainstream. But in 2014, NotaLogger was an early attempt at decentralizing trust – long before "trustless systems" became a crypto buzzword.

NotaLogger also marked a personal transition: from biological systems to socio-technical systems. Where once I asked “What is life?”, I now began asking, “What is evidence?” “What is trust?” “What is witnessing?” These questions, though more abstract, felt just as important. Perhaps more so in an increasingly digital society.

Though NotaLogger was simple and quietly released on PythonAnywhere, it embodied a critical idea: Trust doesn’t need to be human. It just needs to be verifiable. That principle of independently verifiable systems would become a foundational motif in my later thinking, especially around ethical software and algorithmic governance.

42: The Science of OLIVER

Citation: Chan, OYW, Keng, BMH, Ling, MHT. 2014. Correlation and Variation Based Method for Reference Genes Identification from Large Datasets. Electronic Physician 6(1): 719-727.

Abstract: Reference genes are assumed to be stably expressed under most circumstances. Previous studies have shown that identification of potential reference genes using common algorithms, such as NormFinder, geNorm, and BestKeeper, are not suitable for microarray-sized datasets. The aim of this study was to evaluate existing methods and develop methods for identifying reference genes from microarray datasets. We evaluated the correlation between outputs from 7 published methods for identifying reference genes, including NormFinder, geNorm, and BestKeeper, using subsets of published microarray data. From these results, seven novel combinations of published methods for identifying reference genes were evaluated. Our results showed that NormFinder's and geNorm's indices had high correlations (R2 = 0.987, P < 0.0001), which is consistent with the findings of previous studies. However, NormFinder's and BestKeeper's indices (R2 = 0.489, 0.01 < P < 0.05) and NormFinder's coefficient of variance (CV) suggested a lower correlation (R2 = 0.483, 0.01 < P < 0.05). We developed two novel methods with high correlations with NormFinder (R2 values of both methods were 0.796, P < 0.0001). In addition, computational times required by the two novel methods were linear with the size of the dataset. Our findings suggested that both of our novel methods can be used as alternatives to NormFinder, geNorm, and BestKeeper for identifying reference genes from large datasets. These methods were implemented as a tool, OLIgonucleotide Variable Expression Ranker (OLIVER), which can be downloaded from http://sourceforge.net/projects/bactome/files/OLIVER/OLIVER_1.zip.

Context: By 2014, my pursuit of reference genes had evolved into a layered inquiry, not just scientific, but computational. The fundamental assumption behind reference genes was stability — that some genes, like constants in an equation, remain unchanged under most conditions. Yet microarray studies were revealing that this assumption could not be universal. In earlier work, I had already demonstrated that stable expression was context-dependent as what was stable in the liver might not be in the pancreas, and so forth.

In this paper, I took the problem further: What if the very tools we used to identify reference genes weren't designed for large-scale datasets like microarrays? NormFinder, geNorm, and BestKeeper; the canonical trio; were never optimized for such data-rich environments. We ran comparative evaluations, and not only did we observe weak correlations between some of these indices, but we also realized their computational inefficiencies.

This led to the development of OLIVER, a tool built around two novel methods that retained statistical robustness while scaling linearly with dataset size. The name, OLIgonucleotide Variable Expression Ranker, was no accident. It was a tribute to Oliver, one of the two co-authors (along with Bryan) who had been with me since their Singapore Polytechnic days. Now nearly finished with junior college, they were still eager to carry on – a rare continuity in student mentorship.

Reflection: OLIVER represented many things. Technically, it was a pragmatic solution to a real limitation in the field of gene expression normalization. Personally, it was about keeping the flame alive; in a project, in students, in myself. By this point, I had already closed one chapter with the Nova Science book contribution (Chapter 40), tying up the loose ends of our earlier reference gene explorations. But Oliver and Bryan weren’t done, and neither was I.

There’s something poetic in naming a computational tool after a student. In some ways, OLIVER wasn’t just software; it was a coming-of-age marker – for the project, for the students, and perhaps for me. It was a signal that mentorship had matured into collaboration.

Looking back, this paper and its accompanying tool taught me a subtle lesson: Progress isn’t always a leap. Sometimes, it’s refinement – a better method, a faster computation, a clearer correlation. And sometimes, the real legacy isn’t just the publication – it’s the people who chose to stay on the journey with you.

Sidebar: The Fun and Meaning of Naming – From Tools to Genes

Naming things in science, especially tools and genes, often carries a mix of humor, practicality, and sometimes a touch of whimsy. Take, for example, the tool I developed for reference gene identification: OLIVER. The name itself, OLIgonucleotide Variable Expression Ranker, was crafted with a purpose, yet it also became a personal tribute. OLIVER was named after one of my students, Oliver. It was a playful nod to the long-standing mentorship we had shared, as well as a signal of our collaboration. In a way, OLIVER became more than just a tool – it was a marker of the growth of both the project and our relationship.

Names in science can do more than just label something; they can reflect the story behind them. Sometimes, a fun or unexpected name can make a tool or concept memorable. This idea isn't limited to software tools; it extends to the names of genes and scientific discoveries too. For example, genes like Sonic Hedgehog or Firefly’s luciferase were given names that evoke images far removed from the cold, clinical world of molecular biology. Such names not only bring a sense of levity but also keep the work accessible and relatable. These playful names often carry deeper meanings or reveal something about the researchers' personalities or inspirations at the time of discovery.

In the case of OLIVER, the name was a nod to continuity, to the journey I had shared with Oliver and Bryan. It was an acknowledgment of the mentorship-turned-collaboration that had matured over time. As scientists, we sometimes overlook the emotional significance of our tools and discoveries but the names we give them; whether playful, serious, or both; can be a lasting reminder of the people, experiences, and milestones that shape our work.

43: DOSE Library

Citation: Castillo, CFG, Ling, MHT. 2014. Digital Organism Simulation Environment (DOSE): A Library for Ecologically-Based *In silico* Experimental Evolution. Advances in Computer Science: an international journal 3(1): 44-50.

Abstract: Testing evolutionary hypothesis in biological setting is expensive and time consuming. Computer simulations of organisms (digital organisms) are commonly used proxies to study evolutionary processes. A number of digital organism simulators have been developed but are deficient in biological and ecological parallels. In this study, we present DOSE (Digital Organism Simulation Environment), a digital organism simulator with biological and ecological parallels. DOSE consists of a biological hierarchy of genetic sequences, organism, population, and ecosystem. A 3-character instruction set that does not take any operand is used as genetic code for digital organism, which the 3-nucleotide codon structure in naturally occurring DNA. The evolutionary driver is simulated by a genetic algorithm. We demonstrate the utility in examining the effects of migration on heterozygosity, also known as local genetic distance. Our simulation results showed that adjacent migration, such as foraging or nomadic behaviour, increases heterozygosity while long distance migration, such as flight covering the entire ecosystem, does not increase heterozygosity.

Context: At the 2012 PyCon Asia-Pacific, I wasn’t expecting to recruit a collaborator. But among the student helpers, one stood out, Clarence Castillo, a Republic Polytechnic IT student with an unexpected interest in biology. We struck up a conversation, and I floated an idea that had been lingering in my mind for some time: could we simulate evolution not just as a computational abstraction but with ecological and biological realism?

That question became the Digital Organism Simulation Environment, or DOSE. While other simulators existed, they often lacked fidelity to real-life evolutionary dynamics. DOSE took a different path by embracing biological hierarchy (from gene to ecosystem), using a 3-character instruction set mirroring DNA codons, and simulating evolutionary pressure through genetic algorithms.

This paper marked the first formal presentation of DOSE, and with it, we declared the system functional. We tested it with a classic evolutionary question – how does migration influence genetic diversity? DOSE offered answers: local migrations (like foraging or nomadism) increase heterozygosity, while global movements (like long-distance flight) do not. Simple, elegant, and biologically resonant.

Reflection: Every digital simulator is an argument, about what matters in the system being modelled. DOSE was a quiet argument for ecological integrity in evolutionary computation. But just as compelling was Clarence’s journey. He wasn’t just coding; he was making a trade-off: spending less time gaming to help bring a complex simulation environment to life. In the process, he internalized concepts that bridged computer science and biology, demonstrating a rare intellectual elasticity.

For me, DOSE was both a tool and a statement. I had always been intrigued by evolution and not just as a biological process, but as a metaphor for learning, growth, and adaptation. Building DOSE was an act of synthesis: combining my scientific interests with my drive to mentor and empower young talent.

This chapter stands as a milestone not just because DOSE became operational, but because it reflected what I hoped science could be: collaborative, cross-disciplinary, and deeply human. Clarence may have started as a student volunteer at a Python conference, but through DOSE, he became a co-author of an idea far bigger than either of us.

Sidebar: Digital Life, Simulated – The Power of Digital Organism Simulators
Digital organism simulators are more than just clever code. They’re artificial life laboratories – environments where evolution, selection, and adaptation play out on silicon instead of soil. These simulators give scientists a sandbox to test hypotheses that would be ethically questionable, logistically impossible, or prohibitively expensive to explore in real biological systems. From the replication of simple life forms to modeling complex ecological dynamics, digital organisms offer a bridge between theoretical biology and computational experimentation.

But not all simulators are created equal. Some prioritize computational efficiency, reducing organisms to abstract rulesets. Others, like DOSE, strive for biological fidelity. By structuring digital organisms with a genetic code based on 3-character instruction sets, echoing DNA’s 3-nucleotide codons, and embedding them within layered biological hierarchies (gene, organism, population, ecosystem), DOSE foregrounds ecology in its design. That’s not just a technical choice; it’s a philosophical one. It suggests that context matters: that evolution is not just about genes, but also about interactions, constraints, and habitats.

In the broader landscape, DOSE joins a lineage of pioneering platforms. Avida, for instance, has been used to evolve digital organisms capable of solving logic problems, while Tierra initiated the field with self-replicating code back in the 1990s. These tools are digital descendants of Darwin, simulating not just survival of the fittest, but also the subtle dance of drift, migration, and mutation.

Perhaps what makes digital organism simulators most exciting is their capacity to extend scientific imagination. They let us ask counterfactuals: What if mutations only occurred in bursts? What if ecosystems could be rewound and re-run? What if we could speed up evolution to watch it unfold in real time? In doing so, they transform biology into something dynamic, testable, and alive; even if that life is made of bits.

Sidebar: Avida – The Gold Standard of Digital Evolution

When it comes to digital organism simulators, Avida is the towering figure, not just a tool but a field-defining platform. Developed in the mid-1990s by Chris Adami, Titus Brown, and Charles Ofria, Avida has become the most widely used and academically cited simulator for experimental evolution in silico. Unlike abstract algorithmic models, Avida organisms are self-replicating computer programs that mutate, compete, and evolve in controlled digital environments. This allowed evolutionary biologists to test hypotheses that would be impractical, unethical, or simply too slow to explore in the lab.

Avida gained prominence when it was used by Richard Lenski and colleagues to replicate a key question in evolutionary theory: Can complex traits evolve incrementally, without intelligent guidance? The answer, shown elegantly through Avida simulations, was yes. In one iconic experiment, digital organisms evolved the ability to perform a complex computational task through a series of simpler, fitness-neutral or slightly beneficial mutations, mirroring the gradualism central to Darwinian evolution.

What makes Avida enduring is its careful balance of abstraction and realism. While its organisms don't look or behave like real-life creatures, the principles of replication, mutation, selection, and competition are preserved. In that way, Avida functions as both a philosophical lens and an empirical testbed where evolution is not merely theorized but witnessed in real time.

As someone who entered the digital evolution space later with DOSE, I have always viewed Avida with a kind of reverence. It carved out a space for experimental evolution in computing, laid the groundwork for robust methodologies, and showed that biology and computation are not just compatible; they are co-evolutionary. Avida didn't just simulate life; it helped us better understand it.

Avida is more than a simulator. It is a tribute to the scientific imagination, a benchmark for rigor, and a quiet testament to what happens when we allow digital creatures to teach us about ourselves.

44: Second and Last Half of Adaptative Evolution

Citation: Loo, BZL, Low, SXZ, Aw, ZQ, Lee, KC, Oon, JSH, Lee, CH, Ling, MHT. 2014. *Escherichia coli* ATCC 8739 Adapts Specifically to Sodium Chloride, Monosodium Glutamate, and Benzoic Acid after Prolonged Stress. Asia Pacific Journal of Life Sciences 7(3): 11-25.

Abstract: *Escherichia* coli lives in the human intestine and any form of adaptation may affect the human body. The effects of food additives on *E. coli* have been less studied compared to antibiotics. A recent study has demonstrated that *E. coli* is able to adapt to food additives by demonstrating global stress response. This study continues to study the evolution of *E. coli* in different food additives (sodium chloride, benzoic acid, monosodium glutamate) in different concentrations, singly or in combination, for over 83 passages. Adaptability of the cells was estimated with generation time and cell density at the stationary phase. Polymerase Chain Reaction (PCR)/ Restriction Fragments Length Polymorphism (RFLP) were used to analyze the adaptation at genomic level. Our results show that adaptation started to slow down and the gradients of generation time against passage are less steep compared with previous study, suggesting that most adaptive mutations occurred within the first 500 generations. In the genomic level, ecological specialization is observed as we find that the cells adapted through a different mechanism and diverge from each other although the resulting effect of the medium is the same. It suggests that different concentrations of food additives cause different types of chemical stress, instead of different levels of chemical stress.

Context: This study marks the second and concluding part of the adaptive evolution work I began during my time at Singapore Polytechnic. The first part had shown that *Escherichia* coli is capable of mounting a global stress response when exposed to food additives. That early result prompted a deeper question: could prolonged exposure result in adaptation specific to each additive, rather than just a general stress response? To answer this, my students and I designed an experiment that tracked *E. coli* across more than 83 passages in media containing sodium chloride, monosodium glutamate, and benzoic acid, either alone or in combination.

While growth metrics such as generation time and stationary phase cell density offered immediate insights into cellular fitness, it was the genomic analysis using PCR and RFLP that revealed something more profound. The bacteria did not converge on a common survival strategy. Instead, they diverged, developing distinct adaptation profiles depending on the nature and concentration of the additive. This ecological specialisation was surprising, especially given that the media were chemically related in their overall effect.

Reflection: This chapter holds particular significance for me because it completes a journey that spanned institutions, cohorts of students, and my own growth as an educator. I began the adaptive evolution series as a practical means of exposing polytechnic students to long-term biological experimentation but it evolved into a genuine scientific inquiry with real-world implications. That food additives can steer evolutionary trajectories so distinctly has ramifications not just in microbiology but also in public health and food science.

What also stands out is how the project embodies my educational philosophy. The students were not just executing a protocol. They were invited to observe, to question, and to push the boundaries of the data they were collecting. Their persistence over months of passages, often monotonous, sometimes frustrating; was rewarded with insight. This work is a quiet yet resolute reminder that even routine experiments can yield elegance if pursued with care and patience.

Sidebar: The Slow Burn of Adaptation

This chapter closes a loop that began years earlier in the labs of Singapore Polytechnic. While the first study hinted at *E. coli*'s ability to mount a generalised stress response, this continuation delved into how prolonged exposure results in divergent genomic adaptations even under seemingly similar chemical environments. What emerged was an elegant but unexpected insight: the cells did not merely become better at surviving—they became specialists, evolving along unique genomic pathways depending on the type and concentration of food additive.

From a scientific standpoint, this work affirms that chemical stress is not a linear spectrum but a multidimensional force with specific selective pressures. Adaptation was no longer just faster growth or greater resistance. It was genetic divergence. The bacteria evolved differently in sodium chloride versus benzoic acid versus monosodium glutamate, even when their growth metrics appeared similar.

This study also marks the closing of your scientific work on adaptive evolution. What began as a straightforward investigation into bacterial growth turned into a nuanced exploration of evolutionary paths shaped by dietary chemicals. It is the final note in a long-running inquiry that started in one institution and was completed in another chapter of your life.

45: Many Published Databases Vanished Prematurely

Citation: Koh, YZ, Ling, MHT. 2014. Catalog of Biological and Biomedical Databases Published in 2013. iConcept Journal of Computational and Mathematical Biology 3: 3.

Abstract: There had been a large number of biological and biomedical related databases being created over the years with a steady rise of about 10% from 2005 to 2012. However, it is difficult to navigate the range of databases as there is no current database inventory and links to databases are embedded in their respective publications. In this study, we developed a set of 91 cataloging tags based on software repositories and listed 379 database papers published in 2013. Of which, only 290 database papers have URL links to the databases. Therefore, only 290 databases were cataloged. Our catalog is given in appendix.

Context: During my PhD years, I became acutely aware of the fragility of digital scientific infrastructure. Many bioinformatics tools and databases, despite being cited and published in reputable journals, would vanish and sometimes within months of publication. This troubled me. These digital resources were not only vital to reproducibility but also held immense value for future research. Yet, there was no central repository tracking them. In response, Yong Zher and I developed a cataloging system for biological and biomedical databases published in 2013. We created 91 tags to systematically classify 379 papers, identifying that only 290 of them had accessible URLs. This study was a direct attempt to create structure where chaos was becoming the norm.

Reflection: This project reminded me that data alone is not enough; access, continuity, and infrastructure are equally crucial. While the academic world continues to produce a staggering number of databases each year, few survive beyond their immediate moment. The loss isn’t just technical; it’s historical and scientific. Yong Zher approached this work with unusual diligence and thoughtfulness. It was heartening when he expressed his wish to continue collaborating with me. His interest in this issue and in long-term sustainability in bioinformatics, gives me hope that the next generation of researchers will think more deeply about scientific stewardship, not just publication.

Sidebar: Fragility of the Digital Archive
Although scientific journals often require that code and data be made available, they seldom enforce long-term availability. Hosting responsibilities often fall on individual researchers, whose funding, time, and affiliations may change. This chapter's work foreshadowed larger global conversations about data permanence, digital

preservation, and scientific reproducibility. A stable digital ecosystem is not built by papers alone, but by the people and systems that commit to maintaining it.

46: OLIVER

Citation: Chan, OYW, Keng, BMH, Ling, MHT. 2014. Bactome III: OLIgonucleotide Variable Expression Ranker (OLIVER) 1.0, Tool for Identifying Suitable Reference (Invariant) Genes from Large Microarray Datasets. The Python Papers Source Codes 6: 2.

Abstract: This manuscript documents the implemetation for OLIgonucleotide Variable Expression Ranker (OLIVER) as described in Chan et al. (2014), which can be downloaded from http://sourceforge.net/projects/bactome/files/OLIVER/OLIVER_1.zip. These codes are licensed under GNU General Public License version 3 for academic and non-for-profit use.

Context: Having described the theoretical foundation and validation of OLIVER in an earlier paper, the next logical step was to release the tool to the community. This chapter documents that step: a formal software publication in The Python Papers Source Codes, where we made OLIVER 1.0 freely available under the GNU GPL license. The platform allowed us to provide code-level transparency, ensuring the scientific and computational reproducibility of our method. It also affirmed OLIVER's identity – not just as a research output, but as a functional tool. This act of open-sourcing represented a key moment in our scientific process, aligning with our belief that bioinformatics tools should be both accessible and inspectable. We published the source code, documentation, and binaries on SourceForge.

Reflection: Code publications are not always celebrated in the academic world, but they are essential. This paper was more than administrative. It was a public handshake with the community, inviting others to adopt, scrutinize, or extend what we had built. OLIVER may have begun as a play on my son's name, but by the time we reached this stage, it had become a symbol of resilience – scientific, parental, and personal. Watching Oliver and Bryan mature alongside the tool, and knowing they wanted to continue building with me, felt like witnessing continuity in a discipline that too often forgets its lineage.

Sidebar: Code as a First-Class Citizen
Many scientific breakthroughs rely on tools that are never shared or maintained. By treating software as a peer-reviewed artifact, we acknowledged that reproducibility begins not with the conclusions of a paper, but with the mechanics of its computation. Tools like OLIVER are not just byproducts of science, they are enablers of it. This chapter marked our recognition of code not as a supplement to research, but as an integral part of its legacy.

47: Antibiotics Resistance

Citation: Castillo, CFG, Ling, MHT. 2014. Resistant Traits in Digital Organisms Do Not Revert Preselection Status despite Extended Deselection: Implications to Microbial Antibiotics Resistance. BioMed Research International 2014, Article ID 648389.

Abstract: We examined whether antibiotics resistance will decline after disuse of specific antibiotics under the assumption that there is no fitness cost for maintaining resistance. Our results show that during disuse of the specific antibiotics, a large initial loss and prolonged stabilization of resistance are observed but resistance is not lost to the stage of pre-resistance emergence. This suggests that a pool of partial resistant organisms persist long after withdrawal of selective pressure at a relatively constant proportion. Subsequent re-introduction of the same antibiotics results in rapid re-gain of resistance. Thus, our simulation results suggest that complete elimination of specific antibiotics resistance is unlikely after the disuse of antibiotics, once a resistant pool of micro-organism has been established.

Context: In this paper, we applied the Digital Organism Simulation Environment (DOSE) for the first time to explore a real-world biological issue – antibiotic resistance. The question of whether resistant traits in microorganisms would revert after discontinuing antibiotic use had profound implications in both medical and social contexts. With DOSE, we were able to simulate microbial populations under various conditions of selective pressure and withdrawal. Our results were striking: even after long periods of deselection, resistance persisted. This finding paralleled the phenomenon observed in clinical settings, where resistance traits do not easily revert, suggesting that once a resistance pool is established, it becomes a permanent feature of the population. The research underscored the importance of responsible antibiotic use and laid the foundation for further explorations into the ecology of resistance.

Reflection: This study sheds light on the persistent nature of antibiotic resistance, even after the withdrawal of selective pressure. The findings underscore the reality that resistance is not easily reversed and may stabilize at a high level, potentially posing long-term challenges for controlling resistance in clinical settings. In hindsight, focusing on antibiotic resistance was highly impactful due to its global medical significance. Furthermore, the ethical implications of purposely adapting bacteria to antibiotic resistance in the lab were paramount. Such experiments on real organisms could have profound consequences and are considered unethical. This made the use of digital organisms the only viable method to explore this issue. By using DOSE, I was able to simulate antibiotic resistance without engaging in unethical practices, thus contributing valuable insights without harm to actual organisms. This allowed for safe experimentation on a topic that otherwise poses significant ethical and social dilemmas.

Sidebar: Ethical Implications of Studying Antibiotic Resistance

The rise of antibiotic resistance is one of the most pressing issues in global health, yet studying it poses considerable ethical challenges. In laboratory settings, creating antibiotic-resistant strains of bacteria is not only difficult but also raises concerns about the potential spread of such resistance outside controlled environments. Given the ethical implications of adapting real bacteria to resist antibiotics, digital organisms offer a safe and responsible alternative. By simulating antibiotic resistance through DOSE, we can explore how resistance evolves and persists without the risks associated with actual biological experiments, providing crucial insights into managing antibiotic resistance in real-world contexts.

48: TranscriptStudio

Citation: Ling, MHT, Poh, CL. 2014. A Predictor for Predicting *Escherichia coli* Transcriptome and the Effects of Gene Perturbations. BMC Bioinformatics 15: 140.

Abstract: A means to predict the effects of gene over-expression, knockouts, and environmental stimuli *in silico* is useful for system biologists to develop and test hypotheses. Several studies had predicted the expression of all *Escherichia* coli genes from sequences and reported a correlation of 0.301 between predicted and actual expression. However, these do not allow biologists to study the effects of gene perturbations on the native transcriptome. We developed a predictor to predict transcriptome-scale gene expression from a small number (n = 59) of known gene expressions using gene co-expression network, which can be used to predict the effects of over-expressions and knockdowns on *E. coli* transcriptome. In terms of transcriptome prediction, our results show that the correlation between predicted and actual expression value is 0.467, which is similar to the microarray intra-array variation (p-value = 0.348), suggesting that intra-array variation accounts for a substantial portion of the transcriptome prediction error. In terms of predicting the effects of gene perturbation(s), our results suggest that the expression of 83% of the genes affected by perturbation can be predicted within 40% of error and the correlation between predicted and actual expression values among the affected genes to be 0.698. With the ability to predict the effects of gene perturbations, we demonstrated that our predictor has the potential to estimate the effects of varying gene expression level on the native transcriptome. We present a potential means to predict an entire transcriptome and a tool to estimate the effects of gene perturbations for *E. coli*, which will aid biologists in hypothesis development. This study forms the baseline for future work in using gene co-expression network for gene expression prediction.

Context: This paper introduces a novel predictor developed to predict the transcriptome-scale gene expression in *Escherichia* coli, particularly in response to gene perturbations such as over-expression, knockouts, and environmental stimuli. Unlike previous studies that predicted gene expression based on sequences alone, this model uses a gene co-expression network to make predictions from a small set of known gene expressions. It enables the prediction of the effects of perturbations on the native transcriptome, offering a tool that can be used to study and test hypotheses in systems biology. The study shows promising results, with predictions of gene expression changes being largely accurate. The development of this predictor was a critical step in advancing our understanding of gene regulation and offers a foundation for future research in systems and synthetic biology.

Reflection: This project marks a pivotal point in both my academic career and in the formation of my spin-off company with Chueh Loo, AdvanceSyn Private Limited. It was during my time at Life Technologies, working on VectorNTI, that I met Chueh Loo, who later offered me a research fellowship in his laboratory at Nanyang

Technological University (NTU). This opportunity allowed me to delve deeper into systems and synthetic biology, ultimately leading to the development of this predictor. The project was a great example of how academic collaboration can translate into innovative solutions, as it was not just a theoretical development but also a tool with practical implications for biologists. The work laid the groundwork for understanding gene expression dynamics and opened up new possibilities for predicting the effects of gene manipulations, which is crucial for synthetic biology and systems biology.

One of the most rewarding aspects of this work was the realization that it could make a real difference in how biologists approach gene perturbations, offering them a tool that could assist in hypothesis generation and experimental design. The fact that it also laid the foundation for the establishment of my own company, AdvanceSyn, was a testament to the power of research that merges academic rigor with real-world application. This transition from theory to practice has been one of the most fulfilling parts of my journey, and this paper marks the beginning of that path.

Sidebar: The Intersection of Academia and Industry: A Personal Journey

The development of this gene expression predictor not only represents a significant contribution to synthetic biology but also highlights the intersection of academia and industry. My transition from research in the lab to founding AdvanceSyn Private Limited was facilitated by the support of mentors and colleagues, including Chueh Loo. This shift allowed me to explore the commercial potential of academic research and apply cutting-edge science to solve real-world problems. It also exemplifies the growing trend of academic research forming the basis for entrepreneurial ventures, where the goal is not just to push the boundaries of knowledge but also to create practical tools that can drive innovation in various industries. The story of AdvanceSyn's inception is a reminder that the gap between academia and industry is not insurmountable, and with the right support and vision, research can have a lasting impact beyond the confines of the laboratory.

49: Conserved Codon Usage Bias

Citation: Keng, BMH, Chan, OYW, Ling, MHT. 2014. Codon Usage Bias is Evolutionarily Conserved. Asia Pacific Journal of Life Sciences 7(3): 1-10.

Abstract: Codon usage bias (CUB) reflects the frequency distribution of codons usage in the genome. Several studies suggest that CUB is based on the combinations, which are most chemically efficient and minimise translational error, show that amongst closely related species, CUB is similar. However, previous studies were mainly carried out on a limited number of related species. This study tests the hypothesis that CUB is evolutionarily conserved, and examines CUB over a large set of organisms. Codon usage distributions from 18 organisms across a diversity of classes were examined. The correlations of codon usage frequencies were calculated between and within classes. Our results demonstrated that Pearson's correlation between CUBs of different organisms within the same class is significantly higher than random. The correlation between the CUBs of mammals, birds, insects, yeast, and bacteria also corresponded to evolutionary distance. This suggests that CUB is evolutionarily conserved and the degree of conservation corresponds to evolutionary distance.

Context: Codon usage bias (CUB), the non-random use of synonymous codons in the genome, has long been known to reflect evolutionary pressures, translational efficiency, and mutational biases. Previous studies on CUB were often restricted to a few closely related species, leaving open the question of whether CUB patterns are conserved across broader evolutionary distances. In this study, we tested the hypothesis that CUB is evolutionarily conserved by analyzing the codon usage frequencies of 18 organisms spanning mammals, birds, insects, yeast, and bacteria. We found that organisms within the same class exhibit significantly higher Pearson correlation coefficients in CUB compared to random pairings. Furthermore, the degree of similarity in codon usage patterns corresponded well with known evolutionary distances, suggesting a conserved, lineage-dependent structure to codon usage.

Reflection: This study marks the moment when I began integrating concepts from evolutionary biology and statistical genomics into my own independent research direction. Having previously worked with Bryan and Oliver on invariant gene selection for normalization, we collectively felt it was time to explore a new scientific question. I had just returned from my NIH-funded project in South Dakota, where I immersed myself in transcriptomics and evolutionary conservation. Applying these lessons to codon usage bias was a natural extension, offering a compelling way to test hypotheses about molecular evolution at a genomic scale.

The analysis itself was satisfying. It was one of our first efforts at cross-kingdom genomic comparison using codon usage as a unifying metric. The results confirmed what I had intuitively believed: that CUB is not just a reflection of species-specific adaptation, but also an evolutionarily conserved trait, shaped by deep phylogenetic history. It was also one of those rare projects where the computational and biological components flowed seamlessly, a reflection of the team’s synergy and growing maturity. More personally, this work signified a new chapter where I began fusing previous experiences across institutions into a unified scientific narrative.

Sidebar: From South Dakota to Singapore – Carrying Knowledge Across Continents

Scientific training is often shaped not just by institutions, but by the transitions between them. While my time in South Dakota was brief, the exposure to large-scale transcriptomic analyses and evolutionary frameworks planted seeds that would flourish back in Singapore. The idea of applying evolutionary conservation to codon usage arose from this intellectual migration. This chapter represents more than a single paper; it is a reminder that learning from one domain can breathe new life into another, and that scientific insight often travels well, adapting and evolving just like the phenomena we study.

50: Digital Organism Review

Citation: Ling, MHT. 2014. Applications of Artificial Life and Digital Organisms in the Study of Genetic Evolution. Advances in Computer Science: an international journal 3(4): 107-112.

Abstract: Testing evolutionary hypothesis in experimental setting is expensive, time consuming, and unlikely to recapitulate evolutionary history if evolution is repeated. Computer simulations of virtual organisms, also known as artificial life or digital organisms (DOs) can be used for *in silico* study of evolutionary processes. This mini-review focuses on the use of DOs in the study of genetic evolution. The three main areas focused in this review are (1) emergence of specialized cells, (2) chemical and environmental resistance, and (3) genetic adaptability. This review concludes with a discussion on the limitations on using DOs as a tool for studying genetic evolution.

Context: This mini-review was a culmination of my early explorations into the field of artificial life, specifically the use of digital organisms (DOs) to study evolutionary processes. With increasing limitations on wet-lab evolution experiments due to time, cost, and ethical concerns, DOs offered an alternative framework – one that could be experimentally rigorous, reproducible, and computationally tractable. In this paper, I focused on three areas where DOs have provided key insights: the emergence of specialized cells, resistance to chemicals and environmental stressors, and the general adaptability of genetic systems under selection. While the promise of DOs is great, the review also highlighted the limitations, such as the abstraction from biological complexity and the challenge of translating findings into real biological systems.

Reflection: This paper is significant not because it introduced new results, but because it marked a shift in my scientific identity. I was no longer just applying existing tools; I was synthesizing knowledge across disciplines to articulate a vision for a field. By this point, I had published several empirical studies using DOs, including simulations of antibiotic resistance and stress adaptation. However, I realized that the broader scientific community lacked an accessible entry point to understand how DOs could be used as serious tools for evolutionary inquiry.

Writing this review helped clarify my own thoughts. It pushed me to reflect on where digital evolution excels, and where it falls short. Importantly, it also positioned DOs not just as a curiosity, but as a necessary alternative given the ethical and practical constraints of real-world evolution experiments. This was especially relevant in contexts like antibiotic resistance, where purposely evolving resistance in pathogens would be irresponsible. In that light, DOs are not a lesser proxy, but an essential surrogate. Looking back, this review laid intellectual groundwork for the arguments I would make in future work on ethical modeling and the philosophy of digital biology.

Sidebar: From Practitioner to Advocate – Framing a Field in Transition

Review papers often emerge when a researcher transitions from hands-on practitioner to field-level advocate. This was that moment for me. I was no longer just using DOs; I was making the case for why others should. The act of reviewing crystallized the arguments I had been making informally for years. In a way, this paper was my manifesto: an articulation of digital organisms not merely as tools, but as ethical and epistemic necessities in the study of evolution. It was here that I realized I wasn't just experimenting; I was helping to shape a paradigm.

51: Antibiotics Resistance 2

Citation: Castillo, CFG, Chay ZE, Ling, MHT. 2015. Resistance Maintained in Digital Organisms Despite Guanine/Cytosine-Based Fitness Cost and Extended De-Selection: Implications to Microbial Antibiotics Resistance. MOJ Proteomics & Bioinformatics 2(2): 00039.

Abstract: Antibiotics resistance has caused much complication in the treatment of diseases, where the pathogen is no longer susceptible to specific antibiotics and the use of such antibiotics are no longer effective for treatment. A recent study that utilizes digital organisms suggests that complete elimination of specific antibiotic resistance is unlikely after the disuse of antibiotics, assuming that there are no fitness costs for maintaining resistance once resistance are established. Fitness cost are referred to as reaction to change in environment, where organism improves its' abilities in one area at the expense of the other. Our goal in this study is to use digital organisms to examine the rate of gain and loss of resistance where fitness costs have incurred in maintaining resistance. Our results showed that GC-content based fitness cost during de-selection by removal of antibiotic-induced selective pressure portrayed similar trends in resistance compared to that of no fitness cost, at all stages of initial selection, repeated de-selection and re-introduction of selective pressure. Paired t-tests suggested that prolonged stabilization of resistance after initial loss is not statistically significant for its difference to that of no fitness cost. This suggests that complete elimination of specific antibiotics resistance is unlikely after the disuse of antibiotics despite presence of fitness cost in maintaining antibiotic resistance during the disuse of antibiotics, once a resistant pool of micro-organism has been established.

Context: This study was a follow-up to our earlier work using digital organisms (DOs) to model the persistence of antibiotic resistance after the withdrawal of selective pressure. In our first paper, we assumed no fitness cost for maintaining resistance – a best-case scenario. In this follow-up, we introduced a guanine/cytosine (GC) content–based fitness cost, simulating a more realistic biological scenario where maintaining resistance carries a trade-off. Despite this added cost, our findings remained consistent: once a resistant population is established, resistance is never fully lost, even after extended periods of disuse. This challenges the often simplistic public health assumption that resistance will decay naturally if we just stop using antibiotics. Digital organisms, once again, allowed us to ethically explore a question that would be unethical to test in wet-lab bacteria.

Reflection: This paper was important not just for its scientific content, but for what it represented in the arc of my digital evolution work. It was the first time we began layering biological realism, such as GC-content-based trade-offs, into our DO simulations. This was a deliberate step toward making our virtual experiments not just conceptually sound, but biologically relevant. The result was sobering: even with a modeled fitness cost, resistance did not fully revert.

Personally, this work also represents my continued insistence on ethical boundaries in science. We cannot and should not experimentally evolve pathogens to resistance in a laboratory. The consequences are too great. But ignoring the evolutionary questions surrounding resistance leaves us scientifically blind. DOs provide the only ethically sound middle path, where we can explore complex dynamics without putting lives at risk.

This was more than just an experiment – it was a statement. A reinforcement that digital biology is not a compromise, but a necessity when the stakes are too high.

Sidebar: Why Fitness Costs Don't Save Us

The idea that antibiotic resistance will vanish once we stop using antibiotics is appealing but dangerous. Many policies and treatment protocols hinge on this hope. By incorporating fitness costs into digital organism models, this study tested that assumption and found it wanting. Even with costs, resistance persisted. The lesson is clear: evolution has no reverse gear. Once resistance is entrenched, we must plan around it and not hope it disappears.

52: TAPPS

Citation: Chew, JS, Ling, MHT. 2016. TAPPS Release 1: Plugin-Extensible Platform for Technical Analysis and Applied Statistics. Advances in Computer Science: an international journal 5(1): 132-141.

Abstract: In this first article, the main features of TAPPS were described: (1) a thin platform with (2) a CLI-based, domain-specific command language where (3) all analytical functions are implemented as plugins. This results in a defined plugin system, which enables rapid prototyping and testing of analysis functions. This article also describes the architecture and implementation of TAPPS in a level of detail sufficient for interested developers to fork the code for further improvements.

Context: TAPPS (Technical Analysis and Plugin Platform for Statistics) was my attempt to create a lean, plugin-based analytical platform that could support both financial technical analysis and applied statistical functions. At its core, TAPPS was intentionally minimal – built around a command-line interface and a domain-specific language so that it could be easily extended by plugins. This modularity was essential: it allowed others to test their ideas quickly without rebuilding the system. I had long felt the need for such a lightweight, extensible platform in my own work, and TAPPS emerged from that need.

This paper documented the first formal release of TAPPS. It provided a detailed description of the command language, plugin architecture, and system internals, aimed at helping other developers fork or contribute to the system. Co-authored with Justin Chew, the article marked the culmination of many months of conceptual thinking and iterative development.

Reflection: This chapter is difficult to revisit. Justin Chew was more than a former student; he was a close friend. While he didn't write a single line of code, he stress-tested TAPPS during its development, pushing the boundaries of what the system could and couldn't do. That feedback loop was essential. His name appears on the paper not just as a courtesy, but in recognition of that role.

Justin passed away on 21 December 2025. His death was sudden, and I still carry the quiet shock of that day. This is the only paper I ever co-authored with Justin, just one paper; but I hope that it allows his name to be remembered. Every time I revisit TAPPS, I think of him. I think of our shared belief that good tools don’t need to be bloated to be powerful. TAPPS was born from that belief. If this platform lives on, even as a footnote, it carries with it a trace of Justin’s legacy.

Sidebar: Tools Don’t Have to Be Big to Be Useful
TAPPS was never meant to compete with R or Python. Instead, it carved out a niche: a barebones, plugin-ready platform for quick experimentation and deploy-

ment. By separating core logic from analysis functions and embracing a command-line interface, TAPPS allowed developers to write, test, and share new tools without touching the core engine. This architecture reflects a philosophy I continue to uphold – modularity, transparency, and simplicity often outperform complexity in real-world use.

53: Collection of ODE Solvers

Citation: Ling, MHT. 2016. COPADS IV: Fixed Time-Step ODE Solvers for a System of Equations Implemented as a Set of Python Functions. Advances in Computer Science: an international journal 5(3): 5-11.

Abstract: Ordinary differential equation (ODE) systems are commonly used many different fields. The de-facto method to implement an ODE system in Python programming using SciPy requires the entire system to be implemented as a single function, which only allow for inline documentation. Although each equation can be broken up into sub-equations, there is no compart-mentalization of sub-equations to its ODE. A better method will be to implement each ODE as a function. This encapsulates the sub-equations to its ODE, and allow for function and inline documentation, resulting in better maintainability. This study presents the implementation 11 ODE solvers that enable each ODE in a system to be implemented as a function. Three enhancements will be added. Firstly, the solvers will be implemented as generators to allow for virtually infinite simulation and returning a stream of intermediate results for analysis. Secondly, the solvers will allow for non-ODE-bounded variables or solution vector to improve code and results documentation. Lastly, a means to set upper and lower boundary of ODE solutions will be added. Validation testing shows that the enhanced ODE solvers give comparable results to SciPy’s default ODE solver. The implemented solvers are incorporated into COPADS repository (https://github.com/copads/copads).

Context: COPADS IV represented a refinement in how ordinary differential equation (ODE) systems could be implemented and solved within Python. The default SciPy approach, which is writing the entire ODE system as a single function, was practical but it compromised maintainability, especially in larger or more complex models. For someone like me, who values code clarity as much as functionality, this became an architectural problem worth solving.

The solution was to implement each ODE as a separate Python function. This modular approach allowed for proper encapsulation, clearer documentation, and easier debugging. I also wanted the solvers to support long-running or even indefinite simulations. This led to the use of generators, which could yield intermediate results for monitoring or streaming analyses. Three key improvements followed: generator-based execution, support for solution-bounded variables, and the ability to impose hard limits on solution boundaries.

This paper formalized those ideas and folded the new solvers into the COPADS repository, continuing my long-standing aim to provide accessible and flexible tools for computational analysts.

Reflection: At the time of writing this paper, I had already invested years into developing COPADS as a curated library of practical, well-tested analytical tools. This contribution was more than a technical update; it reflected my growing insistence that scientific software should be human-readable and teachable. Too often, we design systems for machines rather than for the researchers who must maintain and explain them.

The modular design I proposed here was deeply influenced by how I think and teach. I had supervised dozens of students by then, and I knew firsthand the value of code that reads like a lesson. Each function should tell a story. With the enhancements introduced in COPADS IV, I hoped to empower others, not just to simulate systems but to understand them, line by line.

Looking back, this was one of those quiet contributions: not widely cited, but indispensable in my toolbox. It was another iteration in my ongoing effort to make computational biology more transparent, more maintainable, and more humane.

Sidebar: Generators as a Philosophy of Time

The decision to implement ODE solvers as generators wasn't just about memory efficiency; it was about thinking in time. Generators align with how dynamic systems evolve: not in one giant block, but as a stream of unfolding states. This design choice made it easier to visualize, analyze, and interact with simulations as they progressed. In that way, the generator became not just a programming construct, but a methodological bridge between theory and observation.

54: Philosophy of Models and Simulation

Citation: Ling, MHT. 2016. Of (Biological) Models and Simulations. MOJ Proteomics & Bioinformatics 3(4): 00093.

Abstract: Modeling and simulation are recognized as important aspects of the scientific method for more than 70 years but its adoption in biology has been slow. Debates on its representativeness, usefulness, and whether the effort spent on such endeavours is worthwhile, exist to this day. Here, I argue that most of learning is modeling; hence, arriving at a contradiction if models are not useful. Representing biological systems through mathematical models can be difficult but the modeling procedure is a process in itself that follows a semi-formal set of rules. Although seldom reported, failure in modeling is not a rare event but I argue that this is usually a result of erroneous underlying knowledge or mis-application of a model beyond its intended purpose. I argue that in many biological studies, simulation is the only experimental tool. In others, simulation is a means of reducing possible combinations of experimental work; thereby, presenting an economical case for simulation; thus, worthwhile to engage in this endeavour. The representativeness of simulation depends on the validation, verification, assumptions, and limitations of the underlying model. This will be illustrated using the inter-relationship between population, samples, probability theory, and statistics.

Context: By 2016, I had spent over a decade building models; mathematical, computational, and digital; across diverse domains. But the utility of models in biology was still being contested. While modeling had long been accepted in the physical sciences, biology's inherent complexity and variability made it resistant to such abstraction. This paper was my response to those lingering doubts, a philosophical defense of modeling and simulation as core components of scientific inquiry, particularly in biological research.

I argued that modeling is not just a methodological tool, but an epistemological act—a way of knowing. If learning involves constructing mental models of the world, then to dismiss modeling is to deny the process of understanding itself. I acknowledged the challenges: modeling can fail, simulations can mislead, and the biology may not always map cleanly onto the mathematics. But I also asserted that failures in modeling often trace back to flaws in the assumptions or gaps in knowledge, not in the modeling itself.

The paper ended with a more practical stance: even when models are imperfect, they serve a valuable purpose by guiding experiments, narrowing hypothesis spaces, and offering cost-effective insights; especially when traditional wet-lab experimentation is prohibitive or impossible.

Reflection: This paper marked a turning point in my thinking. I was no longer just building models; I was now defending their legitimacy in the broader scientific landscape. It was the first time I explicitly wrote about why I believed so strongly in simulations and models, not just how to construct them.

In many ways, this piece was personal. It distilled frustrations I had faced as someone trained in both philosophy and computational biology, watching colleagues dismiss simulation work as “not real biology.” I wanted to reclaim modeling as a legitimate form of scientific labor, one that mirrors how we think, learn, and explore.

Looking back, this paper feels like a manifesto. It didn't just present an argument, it invited a shift in mindset. And it reminded me that defending our tools is as important as using them.

Sidebar: Modeling as Epistemology
When we build a model, we are encoding a worldview; a set of assumptions, relationships, and expectations about how a system behaves. In that sense, models are not just scientific tools, but philosophical artifacts. They capture how we think the world works. Accepting this reframes simulation not as a shortcut or substitute for experiment, but as a mirror of cognition: a structured way of exploring belief, inference, and uncertainty.

Sidebar: The Value of Scientific Philosophy – Remember, a PhD Means Doctor of Philosophy
It’s easy to forget, amid the equations, code, and lab reports, that the “PhD” behind a scientist’s name is the Latin for “Doctor of Philosophy”. That isn’t just academic tradition, it’s a reminder that science is not only about data collection, but also about grappling with the nature of knowledge itself.

Scientific philosophy isn’t separate from science; it is woven into its foundations. Every model we build carries assumptions about causality, scale, generalizability, and representation. Every hypothesis reflects a philosophical stance on what counts as evidence. And every interpretation demands reflection on meaning, bias, and inference.

In modeling and simulation especially, philosophy is not optional. The act of simulating biological processes isn’t just technical but it raises core epistemological questions: What does it mean to represent a system? When is a model ‘good enough’? Can we know something through abstraction? The answers aren’t found in code; they lie in how we reason about truth and utility.

This is why defending models, as I did in this paper, was ultimately a philosophical act. It was a stand against the false divide between empirical science and theoretical reasoning. Scientific philosophy helps us navigate the gray zones, where models are imperfect but still insightful, and where simulations might be the only viable method for discovery.

In a time of increasing specialization, the “Philosophy” in PhD is a call to remain grounded in critical thinking. It asks us not just what we know but how we know it and whether that knowledge holds up under scrutiny. To do science well, we must think like philosophers.

55: PNet Library

Citation: Chay, ZE, Goh, BF, Ling, MHT. 2016. PNet: A Python Library for Petri Net Modeling and Simulation. Advances in Computer Science: an international journal 5(4): 24-30.

Abstract: Petri Net is a formalism to describe changes between 2 or more states across discrete time and has been used to model many systems. We present PNet – a pure Python library for Petri Net modeling and simulation in Python programming language. The design of PNet focuses on reducing the learning curve needed to define a Petri Net by using a text-based language rather than programming constructs to define transition rules. Complex transition rules can be refined as regular Python functions. To demonstrate the simplicity of PNet, we present 2 examples – bread baking, and epidemiological models.

Context: Following my work on ODE solvers, I was drawn to another modeling paradigm, Petri Nets. Where ODEs captured continuous change, Petri Nets offered a complementary view of systems evolving through discrete events and state transitions. I had long been fascinated by how different modeling formalisms reveal different truths, and Petri Nets – well established in systems biology and computer science, seemed ripe for exploration in a Pythonic way.

This was also a collaborative effort born from personal reconnections. Zhu En had just returned from Melbourne, and I saw in him a capable partner to help shape this new library. I also brought in Bing Feng, a final-year polytechnic student referred by a former colleague. Although I had never taught him, he quickly proved himself. Together, we built PNet, a pure Python library designed to lower the barrier for defining Petri Nets by offering a text-based transition language, backed by extensibility through Python functions.

To make the concepts accessible, we illustrated PNet through two familiar examples, bread baking and epidemiology, demonstrating how even simple systems can be effectively represented using Petri Nets.

Reflection: PNet was part of my ongoing mission to make formal modeling more approachable. I wanted to challenge the idea that such techniques were only for specialists. If we could make Petri Nets usable through intuitive syntax and seamless integration with Python, then we could empower more people – students, scientists, or developers to explore system dynamics through this lens.

This project also affirmed the richness of informal mentorship. Bing Feng wasn’t my student in the traditional sense, but I was proud to guide him in producing publishable work. And Zhu En’s return created an opportunity for shared intellectual labor. PNet was the convergence of accessibility, pedagogy, and computational clarity; values that had come to define much of my work by then.

Sidebar: Why Petri Nets Matter

Petri Nets shine in systems where discrete events and concurrency matter – think traffic lights, gene regulation, or viral transmission. Unlike ODEs, which assume smooth change, Petri Nets embrace asynchronicity and resource dependencies. In a world increasingly modeled by flows and transitions, Petri Nets offer an intuitive yet powerful way to trace how things happen—step by discrete step.

56: Cellular Resource Calculator

Citation: Wang, HJ, Ling, MHT, Chua, TK, Poh, CL. 2017. Two Cellular Resource Based Models Linking Growth and Parts Characteristics Aids the Study and Optimization of Synthetic Gene Circuits. Engineering Biology 1(1): 30 –39.

Abstract: A major challenge in synthetic genetic circuit development is the interdependency between heterologous gene expressions by circuits and host's growth rate. Increasing heterologous gene expression increases burden to the host, resulting in host growth reduction; which reduces overall heterologous protein abundance. Hence, it is difficult to design predictable genetic circuits. Here, we develop two biophysical models; one for promoter, another for RBS; to correlate heterologous gene expression and growth reduction. We model cellular resource allocation in *E. coli* to describe the burden, as growth reduction, caused by genetic circuits. To facilitate their uses in genetic circuit design, inputs to the model are common characteristics of biological parts [e.g. relative promoter strength (RPU) and relative ribosome binding sites strength (RRU)]. The models suggest that *E. coli* 's growth rate reduces linearly with increasing RPU/RRU of the genetic circuits; thus, providing 2 handy models taking parts characteristics as input to estimate growth rate reduction for fine tuning genetic circuit design *in silico* prior to construction. Our promoter model correlates well with experiments using various genetic circuits, both single and double expression cassettes, up to a relative promoter unit of 3.7 with a 60% growth rate reduction (average $R^2 \sim 0.9$).

Context: By the time this paper was published, my formal research role at Nanyang Technological University had ended. Chueh Loo had moved to the National University of Singapore, and with that, the research group dissolved; or perhaps, more accurately, dispersed. I decided not to follow. Instead, I transitioned into a portfolio life, or what people now call a gig lifestyle – juggling adjunct teaching roles and various side projects.

This paper, then, marked both an ending and a beginning. Scientifically, it tackled a crucial problem in synthetic biology: how the genetic circuits we design affect the host's cellular resources, particularly its growth. Working with Ada Wang, Erebus, and Chueh Loo, we developed two cellular resource models, one focused on promoters and the other on ribosome binding sites, to estimate how synthetic parts drain cellular resources and reduce growth. These models gave circuit designers a predictive tool to optimize systems before building them, and our results correlated well with experimental data, offering both elegance and practical utility.

We were also delighted that Engineering Biology had just launched and accepted our work into its inaugural issue. It was a fresh start for the journal and for me, in a way.

Reflection: This project taught me that even as institutional affiliations fade, scientific contributions endure. I was no longer part of a formal research group, but the work persisted – this paper stands as a testament to the scientific life outside of conventional structures. It also reinforced my view that modeling isn't just abstract theory; it's a critical bridge between design and application in fields like synthetic biology.

I had always believed in the power of computational thinking to extend biological insight. These models weren't just equations; they were reflections of how life juggles its priorities – growth, survival, expression. And in a poetic way, as I stepped into a more freelance, less resource-secure lifestyle, I was studying how cells themselves manage scarcity and complexity.

Sidebar: Growth vs. Expression – A Biological Trade-Off
In synthetic biology, designing a gene circuit is like budgeting for a household: every new “feature” (gene) takes resources from somewhere else. Our models made this trade-off explicit – showing how promoter and RBS strengths linearly affect *E. coli*’s growth. More expression means more burden. More burden means less growth. Understanding this is not just useful; it’s essential to engineering life responsibly.

Sidebar: The End of the Fellowship But Not the Science
The conclusion of a postdoctoral or research fellowship often feels like falling off a cliff. One moment you are embedded in a lab, surrounded by colleagues, supported by grants and infrastructure; the next, you are on your own; no title, no institution, no lab bench to call home. But endings can be deceptive.

When this paper was published, my formal research appointment had ended, but my scientific curiosity had not. If anything, being outside the system clarified what mattered most: not the affiliation, but the contribution. The Cellular Resource Calculator paper was proof. Written after the fellowship, it stood as a final offering from one chapter and the first signal of another – a pivot into independent scholarship.

Leaving the structured world of academia can feel like exile. But it can also be liberation. Free from the pressures of publishing for promotion or chasing grant cycles, I was finally able to pursue science on my own terms. The questions were still there. The thinking still happened. And the work, this work, still found its place in the world.

The end of the fellowship was not the end of research. It was the beginning of a life shaped by intrinsic motivation, not institutional identity. Science, after all, is not a job. It is a way of seeing.

57: Reflection on Pre-Undergraduate Research

Citation: Ling, MHT. 2017. A Personal Narrative of 6 Pre-University Research Projects Over 7 Years (2009-2015) Yielding 19 Manuscripts. MOJ Proteomics & Bioinformatics 6(3): 00193.

Abstract: Acquisition of research skills, including scientific enquiry, is an important requirement in scientific education, after the acquisition of a body of fundamental knowledge. Working on research projects is a direct means to gain research skills, as well as gaining a firsthand experience of the research environment. Here, I shall narrate my experience and learnings as a research mentor involving 22 pre-university students in 6 research and development projects over a period of 7 years, yielding a total of 19 peer-reviewed manuscripts. I have 3 intentions to this narration - (a) to demonstrate that pre-university students can carry out useful research, (b) summarize my learning experience in this journey, and (c) providing some pointers and encouragement to my fellow mentors and intended mentors. I learnt that (1) pre-university students can produce publishable work but (2) the scope of the projects must be well-defined with specific and measurable endpoints, (3) the involvement of the mentor is substantial both in project formation and project management, (4) quality work can be achieved when students understand the background and context of their work, (5) enduring working relationship between students and mentors requires time and efforts to build, and (6) the students can pleasantly surprise the mentor with their quality of writing and reasoning. Upon reflection, I feel that I gained as much as the mentees / protégés, if not more, and this encouraged me.

Context: Between 2009 and 2015, I mentored 22 pre-university students across six research projects, resulting in 19 peer-reviewed manuscripts. This paper wasn't about a single scientific breakthrough – it was about the process of mentoring and nurturing young minds into the world of research. It was deeply personal, chronicling seven years of intense and often transformative work with students who had yet to enter university but already showed sparks of intellectual curiosity and discipline.

The intent behind writing this was threefold. First, to prove, by example, that pre-university students are capable of contributing meaningfully to the scientific literature. Second, to reflect on what I had learned through this long mentorship journey. And third, to encourage other mentors to take the plunge, offering practical advice grounded in lived experience. I discovered that with the right structure, expectations, and relational investment, these students could thrive, and even surprise me with their analytical depth and maturity.

Reflection: This chapter of my scientific life remains one of the most fulfilling. It required a level of commitment unlike any other: not just intellectual oversight, but emotional resilience, empathy, patience, and the ability to simplify complexity without dumbing it down. The work often extended beyond science – it was about guiding teenagers as they found their voices and confronted their self-doubt.

Looking back, I realise these projects shaped me as much as I shaped them. They taught me humility, adaptability, and how to lead by influence rather than authority. In many ways, this was my laboratory for human development, not just scientific inquiry. It was also my quiet rebellion against elitism in academia – a belief that scientific contribution is not gated by age, but by curiosity and mentorship.

Sidebar: Mentoring as Mutual Growth

We often think of mentoring as a one-way street, from the experienced to the novice. But this journey proved otherwise. The students brought energy, fresh perspectives, and a willingness to question assumptions. I brought structure, domain knowledge, and emotional containment. Together, we built something meaningful. And in the process, I learned that mentorship is not a responsibility but a privilege.

58: Bioinformaticist's Stories

Citation: Ling, MHT. 2017. Towards Portrait [(Auto) Ethnography, Narrative, and Action Research] of Bioinformatics. EC Proteomics and Bioinformatics 2(1): 29-35.

Abstract: A bioinformaticist often interacts with many professionals and work on multiple projects at the same time; hence, highly social. Qualitative portrayal approaches; such as action research, ethnography, and narratives; have been employed in healthcare education and practices but limited in bioinformatics. In this article, I review 12 portraits in healthcare education and practices and argue that there are both intrinsic and extrinsic benefits for bioinformaticist to engage in portrayals of the field. Publication is the main extrinsic benefit. The main intrinsic benefit comes from learning and consolidating past experiences, forming a launch pad into the future. I conclude by listing 6 directions in which a bioinformaticist can embark on.

Context: In this paper, I ventured into the intersection of bioinformatics and qualitative methodologies, marking a significant shift from traditional quantitative and computational research. This piece, Towards Portrait [(Auto) Ethnography, Narrative, and Action Research] of Bioinformatics, explores the largely overlooked human and social side of bioinformatics practice. Bioinformaticists often engage in numerous projects simultaneously, working alongside diverse professionals in highly collaborative environments. Despite the technical nature of the field, little attention is given to the interpersonal dynamics and experiences that shape a bioinformaticist's career.

In this article, I reviewed 12 case studies from healthcare education and practice to show how qualitative methodologies; like action research, ethnography, and narrative can provide deeper insights into the experiences of bioinformaticians. These approaches have long been used to study healthcare professionals but are rarely applied to the bioinformatics field. By documenting personal and collective experiences through narratives, I argue, bioinformaticists can gain both intrinsic and extrinsic benefits; namely, professional recognition through publications and personal reflection that helps consolidate and propel future work.

This exploration laid the groundwork for what I later referred to as Portraits of Science and Education, a framework for looking at science as not just an objective pursuit of knowledge, but a deeply personal and social activity. Through autoethnography, I also reflect on how our experiences shape the science we do, not just the facts we discover.

Reflection: Writing this paper was an eye-opening experience for me, as it bridged the gap between my technical, computational background and the more humanistic, narrative-driven world of qualitative research. I began to see bioinformatics not just as a field of data and algorithms, but as a dynamic, social activity that is deeply embedded in human interactions. This shift in perspective helped me connect more meaningfully with the work of my colleagues and students, encouraging more reflective practice and an understanding of how our personal journeys shape our scientific contributions.

I had previously focused primarily on the technical and methodological aspects of bioinformatics, but this paper opened my eyes to the broader educational and professional context in which this science is done. It highlighted the importance of mentorship, collaboration, and communication, which are often sidelined in the rush for scientific results.

This was my first step into qualitative research in science, a move that would later lead to the creation of the Portraits of Science and Education initiative. It felt liberating to embrace a more narrative-driven approach to documenting scientific work – one that emphasized the value of personal stories, the complexities of scientific practice, and the relationships that foster innovation.

Sidebar: Portraits of Science and Education

Portraits of Science and Education is a framework I developed to explore and document the social, educational, and personal dimensions of scientific work. It combines autoethnography, narrative research, and action research to tell the stories of those involved in scientific endeavours, from students to seasoned professionals. The idea is to highlight how science is not only a technical pursuit but a deeply personal and collaborative journey that shapes both the individual and the community.

These portraits help us understand science beyond data points and algorithms, acknowledging the human side of discovery. By focusing on real experiences and personal stories, they offer a richer, more nuanced understanding of how scientific knowledge is created and shared.

59: DOSE Code

Citation: Castillo, CFG, Ling, MHT. 2018. Digital Organism Simulation Environment (DOSE) Version 1.0.4. In Current STEM, Volume 1. Nova Science Publishers, Inc. ISBN 978-1-53613-416-2.

Abstract: Evolution is a fundamental aspect of biology but examining evolution is difficult, time-consuming and costly. At the same time, molecular analysis of biological organisms is generally destructive, which presents a conundrum between observing possible evolutionary outcomes and in depth molecular analysis to decipher the corresponding evolutionary outcomes. Artificial life simulations via the use of digital organisms (DO) had been proposed as a feasible means of examining evolution *in silico* and had yield biologically relevant findings. Being digital, identical replicates can be made for analysis; thereby, resolving the conundrum. Recently, original implementation of DOSE (Ling, 2012a) had been improved (Castillo and Ling, 2014a) for use as a Python library for simplified construction of simulation, enabling database logging and revival of simulations. This manuscript documents the implementation and improvement of DOSE, which is released as DOSE version 1.0.4 (https://github.com/mauriceling/dose/releases/tag/v1.0.4) and licensed under GNU General Public License version 3. DOSE codebase is hosted and available for forking at https://github.com/mauriceling/dose.

Context: The paper Digital Organism Simulation Environment (DOSE) Version 1.0.4 was published as a part of Current STEM, Volume 1, and represents a milestone in the development of a simulation tool for studying evolution through artificial life. The Digital Organism Simulation Environment (DOSE) had been a project I had been working on for several years, aiming to provide a framework for simulating digital organisms (DOs) to explore evolutionary processes in a computational setting. Evolution is a key biological process but one that is difficult to observe directly, especially when considering the destructive nature of molecular analysis in biological organisms. With digital organisms, however, simulations offer the ability to recreate identical environments for repeated experimentation, sidestepping the ethical and technical limitations of traditional methods.

In this paper, Clarence and myself documented the improvements made to DOSE, specifically the release of version 1.0.4. These improvements were crucial for expanding DOSE's capabilities, such as simplifying the simulation construction process, adding database logging for better record-keeping, and enabling the revival of simulations to resume or refine experiments. The codebase, available for public use and licensed under the GNU General Public License, became a significant tool for the research community interested in artificial life and evolutionary biology. The improvements made DOSE a more accessible and robust resource for researchers interested in using artificial life to study evolutionary dynamics *in silico*.

Reflection: Publishing this as a code-based paper represented a shift in how I thought about scientific publication. Traditionally, papers in biology focus on biological discoveries or theoretical advancements, but here, the paper was centered around a tool, a framework that could be used by others to simulate, explore, and perhaps even discover new insights into evolutionary biology.

The evolution of DOSE over the years had been a significant personal journey for me. Initially, it started as an exploration of the possibilities of artificial life for studying evolution in a controlled, repeatable way. But with each iteration of DOSE, I was able to refine its functionality, making it more intuitive and user-friendly for other researchers. The improvements made in version 1.0.4, such as simplified simulation construction and enhanced data management, were directly shaped by feedback from users and my own experiences using the system.

This paper is as much a reflection of the evolution of DOSE as it is of my own development as a researcher. It marked my transition from merely experimenting with digital organisms to creating a tool that could be used by others in the field. It was a tangible demonstration of how scientific tools can advance our understanding, not just in the form of results, but in the form of new methodologies that can be shared, refined, and built upon by others.

Looking back, I feel proud of how DOSE has contributed to the artificial life community. It was one of the first significant tools I had developed that bridged my work in computational biology with a broader audience, allowing for more expansive exploration into evolution. The fact that DOSE is open-source and freely available for use shows my commitment to collaboration and sharing knowledge within the scientific community.

Sidebar: The Evolution of DOSE

The Digital Organism Simulation Environment (DOSE) was designed to allow users to model the evolutionary process using digital organisms. Its main advantage over traditional biological studies is that it sidesteps the destructive nature of biological experiments by providing a computational, repeatable environment where identical replicates can be created for analysis. DOSE’s features, such as database logging and the ability to revive simulations, enable researchers to track the progress of digital organisms over time and make modifications to the environment or organism characteristics without having to restart experiments from scratch.

60: Jigsaw Cryptography

Citation: Ling, MHT. 2018. A Cryptography Method Inspired by Jigsaw Puzzles. In Current STEM, Volume 1. Nova Science Publishers, Inc. ISBN 978-1-53613-416-2.

Abstract: Cryptography is a critical tool in safeguarding information from "unauthorized" view during the storage and transportation of data. Due to the one-to-one correspondence between plain text and cipher text, encryption algorithms can be seen as a transformation process. This is a deficiency as all information is present, though encrypted, in the cipher text. Inspired by Jigsaw puzzles, a new cryptography system, Jigsaw Encryption System (JES) is proposed where a single plain text file results in many cipher text files, resembling jigsaw pieces from a single image; thus, the loss of a small number of cipher text files may not compromise the entire contents in plain text. Each cipher text can be further processed for added security. This can result in larger permutations needed to decipher by brute force. Reference implementations of preliminary JES versions had been deposited into COPADS (Collection of Python Algorithms and Data Structures) repository (https://github.com/mauriceling/copads; File name: copads/jigsaw.py).

Context: This paper, A Cryptography Method Inspired by Jigsaw Puzzles, was published in Current STEM, Volume 1, and presented a novel cryptographic method I developed, which drew inspiration from the structure of jigsaw puzzles. Traditional cryptographic systems often rely on one-to-one correspondence between plaintext and ciphertext, which means that although the information is encrypted, all the data is still contained within the ciphertext, making it vulnerable to attacks. To address this issue, I introduced the Jigsaw Encryption System (JES), a method that breaks a plaintext file into multiple cipher-text files, resembling jigsaw puzzle pieces that belong to a single image. The idea was that even if some of these cipher-text files were lost, the entire plaintext information would not be compromised, because each piece only holds part of the overall data.

The key innovation of JES was that it fragmented the encrypted message into several parts, each of which alone was insufficient to reconstruct the original plaintext. This added a layer of redundancy and security to the encryption process. Furthermore, each cipher-text piece could be further processed for additional security, significantly increasing the complexity of brute-force decryption attempts. The paper documented the conceptual framework for JES, and I deposited reference implementations of the preliminary versions of JES in the COPADS (Collection of Python Algorithms and Data Structures) repository, allowing others to access and experiment with the code.

Reflection: This paper marked a shift from traditional cryptographic methods, which typically focus on mathematical transformations of the original data, to a more creative, structural approach to encryption. The inspiration drawn from jigsaw puzzles highlighted my interest in exploring unconventional ideas and their potential applications in fields like cryptography. By introducing the concept of fragmentation, JES addressed the inherent risks associated with single-file encryption systems. The idea of "loss tolerance" in encryption was intriguing, as it added an element of error correction to the cryptographic system – similar to how a jigsaw puzzle can still be understood even if some pieces are missing.

In the context of my broader body of work, this paper represented a foray into security and encryption that blended computational creativity with practical applications. I was also interested in the potential implications of this method for securing data in a way that was resistant to both traditional cryptographic attacks and modern threats in distributed systems. JES presented a way to rethink how we store and protect information, aligning with my broader interest in pushing the boundaries of conventional scientific thinking.

Releasing the preliminary versions of JES into the COPADS repository was an important step for me. It marked a move from theoretical work to practical tools that could be used and tested by others in the field. By making this work open-source, I hoped to encourage further innovation and development, allowing others to build upon the concept of fragmented encryption and explore its potential in various real-world applications.

Sidebar: Key Features of the Jigsaw Encryption System (JES)

The Jigsaw Encryption System (JES) introduces a novel method of securing data by splitting plaintext into multiple encrypted pieces, each of which contains only a fragment of the original information. This approach provides several unique features:

1. Fragmentation: The plaintext is broken into multiple cipher-text files, each representing a small portion of the original data, similar to pieces of a jigsaw puzzle.
2. Potential Loss Tolerance: If some cipher-text files are lost or damaged, the entire plaintext is not compromised, as no single cipher-text file contains the full information.
3. Enhanced Security: Each cipher-text file can be further encrypted or processed for additional security, making brute-force decryption significantly more challenging.
4. Redundancy and Complexity: The method introduces redundancy and increases the complexity of traditional encryption, offering a creative way to secure information in distributed systems.

61: Library for Lindenmayer System

Citation: Ling, MHT. 2018. COPADS V: Lindenmayer System with Stochastic and Function-Based Rules. In Current STEM, Volume 1. Nova Science Publishers, Inc. ISBN 978-1-53613-416-2.

Abstract: Lindenmayer system, commonly known as L-system, is a string rewriting system based on a set of rules. In each iteration, the string is repeatedly rewritten based on the rules given. This has been used to model branching processes; such as, plant and animal body patterning, and sedimentation processes. In addition to deterministic rewriting rules, stochastic rules have been used, leading to the development of stochastic L-system (S-L-system). For more complex modeling, parametric rules have been used, leading to the development of parametric L-system (P-L-system). Combining S-L-system and P-L-system leads to the development of L-system capable of handling both stochastic and parametric rules or parametric-stochastic-L-system (PS-L-system). Currently, there is no pure Python PS-L-system library. In this study, a light-weight, pure Python PS-L-system has been implemented. This has been incorporated into COPADS repository (https://github.com/copads/copads) and and licensed under GNU General Public License 3.

Context: In this work, I focused on extending the classic Lindenmayer System (L-system), which is used for modeling branching processes like plant growth and animal body patterning. L-systems are essentially string-rewriting systems where a string is transformed into another string based on a predefined set of rules. These systems have been used extensively in computer graphics and theoretical biology to model the growth patterns of plants and other organisms.

However, the traditional L-system is deterministic, meaning the rules are fixed, leading to predictable outcomes. To introduce more variability and realism into the model, I explored the use of stochastic L-systems (S-L-systems), where some of the rules are probabilistic. For even more flexibility, I combined stochastic rules with parametric rules (P-L-systems), which allowed parameters to control the rewriting process, making the system more complex and capable of modeling a wider variety of phenomena.

In this study, I implemented a lightweight, pure Python library for a parametric-stochastic L-system (PS-L-system), combining both stochastic and parametric features. This library was added to the COPADS (Collection of Python Algorithms and Data Structures) repository, enabling others to easily access and use it for their own modeling tasks. The code was released under the GNU General Public License, ensuring it could be freely shared and modified.

Reflection: This paper was significant for me because it showcased the potential for combining multiple types of L-systems (deterministic, stochastic, and parametric) to create a more flexible and powerful tool for modeling complex biological and natural processes. The combination of stochastic and parametric rules in a single framework opened up possibilities for more accurate simulations of real-world systems that exhibit both randomness and structural patterns.

The development of this pure Python PS-L-system was also an example of my ongoing commitment to providing open-source, accessible tools for researchers and practitioners in computational biology, computer graphics, and other fields. By contributing the library to the COPADS repository, I made it available to a wide audience, allowing others to experiment with and build upon it.

The choice of Python for this implementation was also important. Python is widely used in both research and industry, and its simplicity and readability made the PS-L-system easy to integrate into other projects. I aimed for the library to be a lightweight, yet powerful, tool that could be easily adopted by others in the computational biology and modeling communities.

Additionally, the interdisciplinary nature of the work, combining concepts from computer science, biology, and mathematics, reflected my broad intellectual curiosity and my desire to contribute to multiple fields simultaneously. The paper also served as a demonstration of how abstract mathematical models, like L-systems, could be applied to real-world problems, from understanding the growth of plants to simulating more complex natural phenomena.

Sidebar: Key Features of the Parametric-Stochastic L-System (PS-L-System)
The PS-L-system library introduced several key features that make it an effective tool for modeling complex systems:

1. Stochastic Rules: By incorporating probabilistic rules into the system, the PS-L-system can model variability and randomness in the processes being simulated, allowing for more realistic representations of natural phenomena.
2. Parametric Rules: The inclusion of parametric rules adds flexibility to the system, enabling users to define parameters that control the rewriting process, allowing for more fine-grained control over the model's behavior.
3. Lightweight and Pure Python: The library is written entirely in Python, ensuring ease of use, integration, and modification for users from various disciplines.
4. Open Source: The library was released under the GNU General Public License, encouraging collaboration and further development by the broader research and open-source communities.
5. Multi-Disciplinary Applications: The PS-L-system is suitable for a range of applications, including modeling biological growth processes, simulat-

ing plant patterns, and investigating complex systems that exhibit both randomness and structure.

Sidebar: A Brief History and Modern Applications of Lindenmayer Systems

Lindenmayer Systems, or L-systems, were introduced in 1968 by the Hungarian biologist Aristid Lindenmayer to model the growth of simple multicellular organisms, particularly filamentous fungi. At their core, L-systems are parallel string rewriting systems – starting from an initial string (axiom), a set of production rules is applied simultaneously to each character, generating increasingly complex structures over iterations.

What began as a theoretical model for biological development quickly found powerful applications in computer graphics. In the 1980s and 1990s, L-systems were widely adopted for simulating realistic plant structures in virtual environments. The elegance of L-systems lies in their simplicity and scalability: intricate, organic patterns emerge from minimal rule sets, a digital echo of how complex organisms develop from simple genetic instructions.

Over time, deterministic L-systems evolved to include stochastic variants (S-L-systems) to introduce randomness, better mimicking the natural variability found in living systems. Parametric L-systems (P-L-systems) allowed variables and mathematical functions to influence rule execution, enabling even more detailed modeling of curvature, growth rates, and environmental responses.

The Parametric-Stochastic L-system (PS-L-system) represents the synthesis of these ideas by blending structure and unpredictability, control and chaos. Today, L-systems are used in fields as diverse as:

- Botany and Morphogenesis: modeling plant architecture and simulating development stages.
- Computer Graphics and Animation: generating procedural content for games and films.
- Geology and Sedimentation: simulating layered deposition and branching cave systems.
- Biological Education: teaching principles of development, recursion, and fractals.
- Mathematics and Art: exploring self-similarity, symmetry, and algorithmic beauty.

From theoretical biology to procedural art, L-systems demonstrate how a small set of rules can produce systems of profound complexity – a theme that resonates across the natural and computational worlds.

62: Collection of ODE Solvers 2

Citation: Ling, MHT. 2018. COPADS VI: Fixed Time-Step ODE Solvers with Mixed ODE and non-ODE Function, and Script Generator. In Current STEM, Volume 1. Nova Science Publishers, Inc. ISBN 978-1-53613-416-2.

Abstract: Ordinary differential equations (ODEs) are commonly used in mathematical modelling. However, the standard means of implementing a system of ODEs in Python, using Scipy, does not allow for each ODE to be implemented as an individual Python function. This results in poor documentation and maintainability. We have re-implemented 11 fixed-steps ODE solvers to allow for an ODE system to be implemented as a set of Python functions. Here, the ODE solvers are enhanced to take on a non-ODE function, allowing for modification of the ODE result vector at each time step, which may be useful in cases where one or more results has to be calibrated using other results. In addition, a script generator is implemented to assist in the generation of a Python ODE script from a set of parameters. This module had been incorporated into the Collection of Python Algorithms and Data Structures (COPADS; https://github.com/mauriceling/copads), under Python Software Foundation License version 2.

Context: The work presented in COPADS VI: Fixed Time-Step ODE Solvers with Mixed ODE and Non-ODE Functions, and Script Generator is an extension of my earlier work, COPADS IV: Fixed Time-Step ODE Solvers for a System of Equations Implemented as a Set of Python Functions. In COPADS IV, I introduced a novel approach for solving systems of Ordinary Differential Equations (ODEs) by implementing each ODE as a separate Python function. This method significantly improved both the maintainability and clarity of the code, allowing each equation to be independently managed, rather than combined into a single function as traditionally done with solvers like Scipy. This approach was widely recognized for its flexibility and ease of integration with other Python-based libraries.

Building on the foundation of COPADS IV, COPADS VI enhanced this methodology by enabling the inclusion of non-ODE functions within the ODE solvers. This allowed the model results to be dynamically adjusted at each time step, opening up the possibility to incorporate real-world data or perform calibrations based on other model parameters. The addition of this feature greatly extended the practical utility of the solvers, especially in complex modeling scenarios where feedback mechanisms or real-time adjustments are essential.

Furthermore, a significant addition to COPADS VI was the script generator. This tool automated the process of creating Python scripts from a set of input parameters, allowing users to quickly generate models without having to write boilerplate code manually. This feature was especially beneficial for users who might not have extensive experience in coding, offering them a streamlined process to set up simulations efficiently.

Reflection: In reflecting on the progression from COPADS IV to COPADS VI, I realize that the core aim of both works was to improve the flexibility and usability of Python-based solvers for ODE systems. COPADS IV was focused on making the structure of the code more modular and maintainable, an effort that was essential for improving scientific software. However, it became clear that further improvements were necessary to adapt the solvers to more complex, real-world modeling scenarios. COPADS VI took a crucial step forward by enabling the integration of non-ODE functions, which allowed for much more sophisticated simulations that could account for external feedback and dynamic modifications. This addition was particularly important for ensuring that the solvers could be applied to a wider range of problems, from biology to physics, where such dynamic adjustments are often required.

Moreover, the script generator in COPADS VI represented a move toward making the tools more accessible to a broader audience. While COPADS IV was aimed at researchers comfortable with Python programming, COPADS VI sought to lower the barriers for entry by automating the creation of Python scripts. This was a significant step towards making the tools more user-friendly, especially for interdisciplinary researchers who might be familiar with their domain but not with the intricacies of Python.

Looking back, COPADS VI represents an ongoing refinement of the initial vision laid out in COPADS IV. It’s a testament to the importance of continually iterating on scientific software to meet the evolving needs of users. As I continue to develop these tools, I am reminded that software development is not just about functionality but also about making that functionality accessible, extensible, and easy to use.

Sidebar: The Evolution of COPADS

The COPADS series represents a significant body of work aimed at improving how computational models of complex systems, particularly ODEs, are handled in Python. Starting with COPADS IV, which introduced a modular, function-based approach to solving ODEs, these tools have steadily evolved to meet the increasing complexity and variety of real-world applications. Here’s a brief look at how each version has contributed to the field:

- COPADS IV (2018): Focused on improving maintainability and readability by implementing ODEs as individual Python functions, making the system more modular and extensible.
- COPADS V (2018): Built on this by incorporating stochastic and parametric rules in Lindenmayer Systems, further enhancing the flexibility of computational modeling.

- COPADS VI (2018): Added non-ODE functionality and a script generator, streamlining the process for users and expanding the toolset to allow for more dynamic models.

Each iteration represents an important step in the evolution of scientific modeling tools, demonstrating a continuous effort to make computational methods more powerful, user-friendly, and adaptable to various research domains. The integration of feedback, the addition of new functionalities, and a focus on usability have been central to the development of the COPADS series, ensuring that it remains a valuable resource for researchers in fields ranging from biology to physics.

63: Intracellular Landscapes

Citation: Ling, MHT. 2018. Back-of-the-Envelope Guide (A Tutorial) to 10 Intracellular Landscapes. MOJ Proteomics & Bioinformatics 7(1): 00209.

Abstract: Landscape is a metaphor for conceptualizing and visualizing a score across one or more biological entities or concepts. This review provides a cursory overview of 10 landscapes (in alphabetical order, copy number, fluxome, genome, molecular, metabolome, mutation, phenome, proteome, regulome, and transcriptome) in intracellular biology without going into extensive depth; hence, this article can act as a first tutorial into intracellular landscapes. The value ahead is to be able to compare and interrogate across multiple landscapes at different resolutions.

Context: This tutorial-style review was written with a specific pedagogical purpose in mind: to provide an accessible yet structured orientation into the complexity of intracellular biology for incoming honours students. As I prepared to supervise projects in systems biology and bioinformatics, I found a recurring need to help students navigate a vast and often siloed landscape of -omics data. Each “landscape”; be it genome, transcriptome, or proteome; offered a unique perspective, but understanding how they related to one another required an integrative mindset not often fostered in early-stage training.

To bridge this gap, I adopted the metaphor of “landscapes” as a unifying mental model. The paper presents ten such intracellular landscapes; namely, copy number, fluxome, genome, metabolome, molecular, mutation, phenome, proteome, regulome, and transcriptome—organized alphabetically and introduced in a deliberately light, approachable manner. It avoids technical depth by design, instead providing a high-level entry point into each concept. The tutorial functions like a cognitive map, helping students and newcomers to frame their thinking across multiple biological layers, preparing them for comparative and integrative analyses.

Reflection: Although this work may not fit the mold of a traditional academic contribution, its purpose was no less critical. It reflects a core philosophy I have come to embrace: empowering early-career scientists by demystifying complexity. Too often, students are plunged into narrow research niches without a sense of the broader biological context. This paper was my attempt to counter that, to offer a “back-of-the-envelope” sketch that encourages big-picture thinking while equipping students with enough foundational understanding to explore further.

Writing this guide also served as a personal exercise in synthesis. As someone who has worked across different biological scales, I needed a simple way to structure conversations with students and collaborators alike. The landscape metaphor provided that scaffolding—an elegant, intuitive way to visualize multi-dimensional biological data. This metaphor has since informed how I design projects, teach concepts, and evaluate cross-scale biological questions.

While the paper itself is brief, its influence in my teaching and mentoring has been disproportionately large. Many of my students have cited it as the first time they began to understand the interconnectedness of biological data, and that affirmation has made this small review one of the more quietly impactful pieces I have written.

Sidebar: Why “Landscapes”?

The use of “landscapes” as a metaphor in biology is not new, but this paper gives it practical pedagogical utility. A landscape suggests a topology; peaks, valleys, gradients; across which variables shift. In the context of intracellular biology:

- The genomic landscape defines the terrain – what's possible.
- The transcriptomic and proteomic landscapes show what's being used.
- The fluxomic, metabolomic, and phenomic landscapes reveal how that usage translates to biological function.

This conceptual framing allows students to begin asking integrative questions early: What features are stable across landscapes? Where do we see divergence? How do these shifts map to disease, function, or adaptation?

For many honours students, this was the first time they saw molecular biology not as a linear pipeline, but as a dynamic, layered ecosystem.

64: What Makes Pre-Tertiary Bioinformatics Research Projects Successful?

Citation: Chan, OYW, Keng, BMH, Ling, MHT. 2018. Science/Education Portraits I: Identifying Success Factors of Pre-Tertiary Bioinformatics Research Experience from Students' Perspective. Advances in Biotechnology and Microbiology 8(2): 555734.

Abstract: Many studies suggest substantial benefits in incorporating research experience into science education, with several studies examining the success factors for undergraduate research experience. It has been known that early research experience has an impact on the career paths of the students. However, little has been known about the success factors of research experience at a high-school level. This study uses a reflective email interview method (2 years post-completion of the research experience) to identify success factors from the perspective of the students on their pre-tertiary bioinformatics research experience. Six success factors emerge from this analysis: (a) student's intrinsic motivation / interests, and goals, (b) peer pressure, (c) student's perceived workload, (d) context of project, (e) culture of science and recognition of student's work, and (f) quality of supervision.

Context: This paper marks the closing chapter of a long and fulfilling collaboration with Oliver Chan and Bryan Keng – the two students whose journey with me began in 2011 and evolved into co-authorship and co-leadership of multiple educational research projects. Over the years, we explored the integration of bioinformatics into high-school education, moving beyond simply teaching coding or biology into designing authentic research experiences for pre-university students.

After several iterations of research mentorship, we found ourselves with a natural question: what makes such experiences successful from the students' own point of view? This paper answers that question using a reflective method, email interviews conducted two years after project completion, with a cohort of former high-school researchers. The aim was to distill what aspects of the experience left a lasting impression and why.

Six recurring themes emerged: students' intrinsic motivation and goals, the influence of peer dynamics, perceived workload, the contextual framing of the research, exposure to scientific culture and public recognition, and, critically, the quality of supervision. Rather than being a checklist, these factors formed an interconnected mesh that helped determine whether students experienced their first research project as transformative or transactional.

Reflection: This paper is deeply personal. Not because of the content alone, but because of the relationship that underpins it. Mentoring Oliver and Bryan from their school years into their undergraduate lives, and ultimately co-writing a paper reflecting on that very process, was a rare privilege. It stands as a testament to the generative nature of mentorship when trust, patience, and curiosity align.

More broadly, this work represents a shift in my thinking from “how do we teach research?” to “how do students experience research?” By foregrounding student voices, we were able to identify that success is not determined solely by academic output, but by resonance; whether the experience stays with them, reshapes their aspirations, or instills a sense of agency.

In hindsight, this project also served as closure – a way to honor the journey we had shared while ensuring that future programs could benefit from our insights. For those working at the intersection of science education and mentorship, this study offers a framework grounded in lived experience, not just pedagogy.

Sidebar: Beyond the Classroom – A Decade of Mentorship

The collaboration that culminated in this paper began informally; an idea here, a conversation there; before it crystallized into structured projects. Over time, what developed was not just a series of educational experiments but a mentorship ecosystem. I saw Oliver and Bryan grow from curious students to independent thinkers and eventual co-researchers.

This paper isn't just the result of scholarly inquiry, it is the product of trust accumulated over time. It reminds us that the most enduring impact of education isn't always measurable in scores or citations, but in relationships that foster growth and reflection long after the classroom lights are turned off.

65: Publication as Project Endpoint

Citation: Ling, MHT. 2018. Science/Education Portraits II: Pre-Tertiary and Undergraduate Research Mentors Should Consider Publication as Project Endpoint. MOJ Proteomics & Bioinformatics 7(2):127–129.

Abstract: Scholarly publication is a common productivity metric for researchers of all levels. Despite the benefits of publications to pre-undergraduate and undergraduate research students, they tend to be less productive in terms of publication counts than graduate research students. Here, I narrate my personal experiences as an attempt to try to convince fellow pre-tertiary and undergraduate research mentors research mentors to consider publications as a suitable endpoint for any research projects, as they can have lasting mutual benefits to both the student / mentee, and project mentors.

Context: In this second installment of the Science/Education Portraits series, I shift focus from evaluating student outcomes to critically reflecting on the responsibilities and aspirations of the mentor. This piece emerged from years of supervising pre-tertiary and undergraduate students in research projects – many of which ended with poster presentations or internal assessments, but few with formal publications.

This paper was written as a call to action: if we treat these projects as legitimate scientific inquiries, then why shouldn't their endpoint be publication? The argument is not just about visibility or productivity; it's about dignity – dignity in recognizing student work as worthy of a wider audience and dignity in holding ourselves, as mentors, to the same standards we uphold in professional research.

Reflection: This piece was, at its core, a self-evaluation. After supervising dozens of student projects, I had to ask myself: what legacy do these students carry beyond the lab or classroom? A poster on a school wall, perhaps. A fleeting memory of a difficult project. But what if that memory could be anchored in something more enduring – an authorship credit, a published insight, a footnote in the scientific record?

The act of writing this paper was my way of aligning my educational philosophy with my academic values. If we believe students are capable of real research, we must be willing to shepherd their work through real dissemination. I make no claim that every student project must be published, but rather that we should set publication as a legitimate and aspirational endpoint; not an unreachable summit, but a meaningful destination.

What moved me most was the response from the reviewer, who remarked that this is "a paper all my colleagues should read." That comment validated the purpose of this piece, and reminded me that mentorship, like research, is also an evolving craft, one shaped by introspection, and shared practice.

Sidebar: Not Just Practice – A Voice in the Record

Too often, student research is framed as “practice” for real science. But what if we instead saw it as participation in science?

Publication doesn’t have to mean landing in Nature or Cell. It can mean a contribution to a niche journal, a case study, a short commentary; so long as it’s real, reviewed, and respected. For the student, it offers an indelible first step into the scientific community. For the mentor, it offers the satisfaction of nurturing not just a learner, but a contributor.

66: RANDOMSEQ

Citation: Ling MHT. 2018. RANDOMSEQ: Python Command–line Random Sequence Generator. MOJ Proteomics & Bioinformatics 7(4):206–208.

Abstract: Randomly generated sequences are important in many sequence analysis studies as they represent null hypotheses. There are several existing tools to generate random sequences but each has its own strengths and weaknesses. Building upon the strengths and weaknesses of existing tools, a command-line random sequence generator, RANDOMSEQ, is presented. Generation of random sequences is versatile: (a) fixed or variable length nucleotide or amino acid sequences can be generated; (b) a variety of frequencies for sequence generation is accepted – source sequence, single or n-length nucleotide / amino acid frequencies; (c) generated sequences can be free of user-defined start or stop codons or both; (d) generated sequences can be flanked with randomly selected start and stop codons; and (e) one or more constant regions can exist within the sequence.

Context: Random sequences are an understated yet fundamental component of bioinformatics. Whether as control datasets in hypothesis testing or as input for simulations, these sequences serve as a kind of computational null hypothesis. Existing tools, however, are often constrained, some lack support for complex frequency profiles, others fail to accommodate structural motifs like start and stop codons.

RANDOMSEQ was developed to address this gap. Designed as a command-line utility in Python, it was shaped by both practical needs and pedagogical use cases. The tool allows for intricate customization, whether the user wants variable-length sequences, specific nucleotide or amino acid frequencies, the exclusion or inclusion of start/stop codons, or fixed internal regions. This flexibility makes it particularly useful for testing bioinformatics algorithms, benchmarking tools, or teaching key concepts in molecular sequence generation.

Reflection: This project stemmed not from a grand research question, but from repeated frustrations – mine and those of students I supervised. Every time we needed to simulate sequences for comparison or as randomized inputs, we found ourselves cobbling together scripts, working around tool limitations, or settling for compromises. Eventually, I realized we needed a utility that did it all; cleanly, predictably, and flexibly.

RANDOMSEQ was my answer to that. While it may not boast complex algorithms or flashy features, it embodies something I've come to value more deeply over time: thoughtful engineering in response to recurring needs. It also reflects a kind of humility in research, acknowledging that not every contribution needs to be headline-grabbing to be quietly impactful.

Sidebar: The Value of “Small Tools”

Not every publication has to revolve around a new theory or algorithm. Sometimes, the most enduring impact comes from tools that make existing workflows just a little smoother. RANDOMSEQ falls into that category – a utility born out of necessity, shaped by repeated use, and refined to support both research and education.

If you've ever spent hours generating random sequences by hand (or worse, editing them after generation), you’ll understand why a tool like this earns its place in the toolkit.

67: Characterizing Transcription

Citation: Wong, A, Ling, MHT. 2019. Characterization of Transcriptional Activities. In Guenther, R. and Steel, D. (eds.), Encyclopedia of Bioinformatics and Computational Biology, Volume 3, pages 830-841. Oxford: Elsevier. ISBN 978-0-12811-414-8.

Abstract: Transcription is the first stage of gene expression, leading to the eventual determination of protein abundance and affecting metabolism. Hence, there is a need to measure and characterize transcriptional activities accurately. This article discusses the experimental techniques to characterize transcriptional activities from a single-gene approach (quantitative PCR) to high-throughput methods (microarray technology and next generation sequencing). Computational approaches to predict relative transcript abundance from sequence features and gene co-expression network will be described. As a proxy to protein / enzyme abundance, transcriptional activities are critical in developing simulatable biochemical models, which can then be used to test biological hypotheses prior to laboratory experimentation.

Context: This encyclopedia chapter was a collaborative effort with Adison Wong, whom I first met during his PhD studies at Nanyang Technological University, co-supervised by Chueh Loo Poh. At the time, I was a research fellow involved in the same scientific circles. Years later, we were invited to contribute to the Encyclopedia of Bioinformatics and Computational Biology — a prestigious reference work aiming to present foundational and emerging topics in computational biology. Our chapter focused on the characterization of transcriptional activities, the first step in gene expression, through both experimental and computational lenses.

The chapter provides a panoramic overview of techniques to quantify transcription, from traditional qPCR to genome-wide methods like microarrays and RNA-Seq. Beyond experimental methods, we also covered computational approaches to infer transcriptional activity using gene co-expression networks and sequence-derived features. The ultimate aim was to frame transcriptional activity not only as a molecular snapshot but as an interface to modeling biological systems.

Reflection: Writing this chapter gave me an opportunity to reflect on how transcriptional activity sits at the crux of cellular decision-making and how deeply it connects experimental biology with *in silico* modeling. In the earlier phases of my career, I had been preoccupied with building models and simulations of biological processes. This collaboration offered a structured way to communicate the methodological rigor needed to generate trustworthy transcriptional data — the very substrate of those models.

In retrospect, the chapter also represents a convergence of my scientific priorities: the ethical commitment to model before meddling, the methodological principle of unifying experimental and computational methods, and the pedagogical need to make these accessible to students and early-career researchers. In contributing to a reference volume of this scale, I hoped to codify these values while offering readers a reliable guide to studying one of the most essential processes in biology.

Sidebar: From Molecules to Models – Why Transcription Matters

Transcription acts as a molecular tuning dial – modulating gene expression and setting the stage for protein synthesis. What makes it fascinating is its measurability. With the right tools, we can observe it, quantify it, and even predict it. This chapter was my attempt to demystify those tools and showcase how transcriptional data serve as a bridge to simulations or digital testbeds where hypotheses can be examined before any lab work begins.

I had known Adison Wong from his days as a PhD student, and our collaboration rekindled around a shared interest in transcription. His experimental insights complemented my computational leanings. Together, we built a chapter that reflected both rigor and reach — practical methods grounded in scientific philosophy.

For me, transcription is more than just a biological process. It’s an ethical fulcrum – enabling safe, efficient, and insightful inquiry into how life regulates itself.

68: Transcriptome to Phenotype Review

Citation: Li, BT, Lim, JX, Ling, MHT. 2019. Analyzing Transcriptome-Phenotype Correlations. In: Guenther, R. and Steel, D. (eds.), Encyclopedia of Bioinformatics and Computational Biology, Volume 3, pages 819-824. Oxford: Elsevier. ISBN 978-0-12811-414-8.

Abstract: Advancements in high-throughput transcriptomics methods in the last 2 decades had enabled the many studies aiming to examine the transcriptome differences between two samples or across time points. Transcriptomic experimental techniques are more developed and readily available compared to that of proteome, metabolome, and fluxome. Transcriptome is the first order activity of the genome, and leading higher order changes; such as, changes in proteome, metabolome, and fluxome. However, the eventual aim is to understand how changes in the omics results in phenotypic differences between the samples. This article gives an overview of transcriptomic techniques and how phenotypic differences can be elucidated.

Context: In 2019, I co-authored the encyclopedia entry Analyzing Transcriptome–Phenotype Correlations with Bryan Li and Jin Xing Lim, then both early-career colleagues at Temasek Polytechnic, where I was serving as an adjunct lecturer. This chapter was published in the Encyclopedia of Bioinformatics and Computational Biology, a reference work that aimed to consolidate established knowledge across the field. While Bryan and Jin Xing were just starting out, our collaboration offered an opportunity to both contribute to the literature and nurture their academic development under the supportive guidance of our manager.

The article explores how transcriptomic techniques, ranging from microarrays to next-generation sequencing, can be harnessed to understand how gene expression profiles relate to phenotypic differences. It positions the transcriptome as a foundational layer: the first-order functional readout of the genome that precedes and predicts downstream layers such as the proteome, metabolome, and fluxome. At the time, transcriptomics remained the most mature and accessible among high-throughput omics technologies, making it a natural choice for studies aiming to link molecular data with observable traits.

Our chapter addressed not only the experimental aspects of transcriptomics but also the statistical and computational strategies for drawing correlations between gene expression data and phenotype. It was meant to be both a technical guide and a conceptual primer for newcomers to the field.

Reflection: This chapter was an embodiment of mentorship through scholarship. While I contributed to the writing and direction of the manuscript, the primary goal was to guide Bryan and Jin Xing through the process of scientific publication. It was not just about the knowledge we shared, but the journey we took together in articulating it. For them, it was a formative experience in scholarly writing and integrative thinking. For me, it was a quiet assertion of what I believe academic publishing can be: a space not only for dissemination but for professional growth and mutual uplift.

On a personal level, this work reinforced a central insight in my scientific career – that the transcriptome is often the best entry point into systems biology. While it doesn’t tell the full story, it reveals enough to chart meaningful hypotheses. As a proxy for functional activity, transcriptomics remains a powerful approach to frame questions that later omics layers can refine. The chapter allowed me to reflect on that view and share it in a distilled, accessible form.

At the same time, I found myself subtly challenging the dominant emphasis on novelty in academic publishing. This chapter was not about cutting-edge algorithms or uncharted biological mechanisms; it was about grounding. And yet, it was just as meaningful, perhaps even more so, because it placed value on clarity, pedagogy, and accessibility – qualities that often get overlooked but are foundational to building a strong scientific community.

Sidebar: Mentorship Through Writing

Writing is often seen as the endpoint of research, but it can also be a beginning; for skills, for confidence, and for careers. When I proposed this chapter, it wasn’t just about filling a gap in a reference book. It was a vehicle for mentoring. Watching Bryan and Jin Xing transition from hesitant contributors to capable co-authors was one of the quiet triumphs of this project. Scientific growth does not always come from the lab bench. Sometimes, it happens at the keyboard.

69: Antisense Transcription Review

Citation: Ling, MHT. 2019. Survey of Antisense Transcription. In Guenther, R. and Steel, D. (eds.), Encyclopedia of Bioinformatics and Computational Biology, Volume 3, pages 842-846. Oxford: Elsevier. ISBN 978-0-12811-414-8.

Abstract: In recent years, many antisense transcripts had been discovered. Antisense transcripts are complementary to the coding transcripts, also known as sense transcripts. Duplex formation of sense/antisense transcript has been thought to limit the effective abundance of sense transcripts, leading to reduced levels of translated peptides/proteins. However, recent discoveries show that this is not the case – duplex formation is the first-order effect, which can affect higher-orders; such as, interfere with transcription and translation of sense transcripts, affecting the maturation and half-life of sense transcripts. In this article, 4 cases are examined in depth to illustrate that the effects of antisense transcript on the eventual peptide/protein level can be complex and should be considered on a case-by-case basis.

Context: This article marked a quiet but significant closure to a formative chapter in my scientific life – my NIH-funded postdoctoral work in South Dakota. During that period, I had the rare opportunity to investigate antisense transcription in a focused and sustained way, supported by funding that allowed intellectual depth without the usual distractions. Antisense transcription, once seen as genomic noise, was beginning to be recognized for its regulatory potential. The field was in flux, and my own work, both experimental and computational, contributed to emerging insights about how antisense RNAs could modulate gene expression beyond simple interference.

By 2019, the major projects from that postdoctoral stint had been completed and is obvious that I will not return to South Dakota. This article became a capstone: a concise, curated survey of the field that drew from both my published findings and the knowledge I had quietly accumulated. Unlike my other entries in the encyclopedia, this was a solo-authored contribution. It felt appropriate as it reflected both intellectual independence and personal closure.

Reflection: Writing this article allowed me to revisit a time in my career when I was deeply immersed in mechanistic questions. I still remember the thrill of seeing unexpected patterns in transcriptome data, suspecting antisense involvement, and then designing experiments to test those hunches. The work was painstaking and subtle—antisense effects often vary case by case, and broad generalizations are risky. But therein lay the beauty: science as slow, textured observation rather than just sweeping declarations.

This chapter also served as a mirror. In reviewing the literature and weaving it together with my own experience, I realized how much I had matured—not only as a scientist, but as a thinker. The questions I asked during my postdoc had not lost their relevance; if anything, they had become more poignant. But I was no longer seeking definitive answers. I was comfortable embracing complexity, acknowledging ambiguity, and highlighting the conditional nature of biological knowledge.

Sidebar: The Elegance of the Opposite Strand

There’s a poetic symmetry in antisense transcription – the messages written on opposite strands, sometimes in conflict, sometimes in harmony. For me, antisense biology was never just about molecular interference. It symbolized the nuanced interplay of intention and counterpoint, signal and suppression. Looking back, I wonder if I was drawn to antisense research not just for its novelty, but because it resonated with how I experience the world: always attuned to the possibility that the most powerful forces may run contrary to expectation.

70: Sequence Composition Review

Citation: Lim, JX, Li, BT, Ling, MHT. 2019. Sequence Composition. In Guenther, R. and Steel, D. (eds.), Encyclopedia of Bioinformatics and Computational Biology, Volume 3, pages 323-326. Oxford: Elsevier. ISBN 978-0-12811-414-8.

Abstract: Genomic sequence is commonly known as the "blueprint of life" but deciphering this blueprint has proven to be difficult and elusive. The first task to deciphering this code is sequence analysis, resulting in an annotated sequence. This annotated sequence represents the feature composition of this sequence, or commonly known as sequence composition. In this article, we will examine some of the available tools to identify various DNA sequence features, before reviewing recent studies on the application of sequence composition. Through these applications, we can appreciate that sequence composition is an integral aspect of sequence analysis.

Context: With the encouragement of our manager, I deliberately sought to involve them in scholarly work to accelerate their early career development. "Sequence Composition" was a natural topic to cover – it served as a foundation for many downstream applications in genomics and transcriptomics, areas in which I had built considerable expertise. The article itself was part of a three-paper series we contributed to Elsevier's Encyclopedia of Bioinformatics and Computational Biology, offering concise, application-driven perspectives suitable for both students and early-career researchers.

The idea of "sequence composition", that we can extract meaningful patterns from the arrangement of nucleotides, was something I had been working with since the early days of my computational biology journey. It connected with earlier themes from my doctoral work and was further cemented through my experience developing tools and databases for sequence-based analysis.

Reflection: While modest in scope, this article encapsulates the deep appreciation I hold for the power of pattern recognition in biology. Deciphering the "blueprint of life" begins with identifying what makes up that blueprint – the composition of sequence elements that hint at regulatory, structural, or evolutionary significance. Writing this article alongside Bryan and Jin Xing felt less like an academic task and more like a passing of the torch. It provided them with a structured opportunity to contribute to a reference work of international standing, and it reminded me that mentorship is not only about guidance, but also about creating visibility for others.

This article also marks a continued thread in my scientific ethos: the belief that abstraction and simplification, when communicated effectively, can empower others. By framing sequence composition not just as a concept, but as a practical analytical tool, we helped readers understand why this topic still matters in an era dominated by high-throughput "-omics" data.

Sidebar: Nucleotide Grammar and Scientific Grounding

At first glance, sequence composition may seem rudimentary – counting base pairs, identifying repeats, locating motifs. But these simple computations reflect a deeper grammar within the genome. As someone drawn to systems and structure, I often think of biology as a form of coded language. Writing this article was, in part, an homage to that language. In helping others understand its syntax, I also re-grounded myself in the basics—those essential elements that, even after decades in science, remain endlessly profound.

71: A Tool for *Pseudomonas stutzeri*

Citation: Lim, JX, Ling, MHT. 2019. Gene Ontology and KEGG Orthology Mappings for 10 Strains of *Pseudomonas stutzeri*. EC Proteomics and Bioinformatics 3(1): 12-18.

Abstract: Gene Ontology (GO) and KEGG Orthology (KO) are controlled vocabularies for annotating gene and protein functions, and mapping functions onto pathways; which enables metagenomic analysis. *Pseudomonas stutzeri* is an environmental bacterium with potential for biotechnology applications in the environment, despite being an opportunistic pathogen. However, there has been no GO nor KO annotations for *P. stutzeri*. This study presents the first GO and KO mapping for 10 strains of *P. stutzeri* for further studies into *P. stutzeri*. Of the 42764 peptides in 10 strains of *P. stutzeri*, 30435 (71.17%) peptides were annotated with one or more GO terms and 25034 (58.54%) of peptides were annotated with KO terms. The annotation files and sequences can be downloaded at https://tinyurl.com/GO-KO-Pstutzeri.

Context: This paper marked a quiet farewell, both to a collaborator and to a phase of my work centered on functional annotation in microbial genomics. Jin Xing and I had collaborated on several projects during our overlapping time at Temasek Polytechnic. This particular study, focused on *Pseudomonas stutzeri*, began as a pedagogical and practical effort: to expose students to annotation pipelines using real, diverse datasets, while addressing a genuine gap in the literature. Despite *P. stutzeri*'s relevance in environmental biotechnology, no comprehensive Gene Ontology (GO) or KEGG Orthology (KO) mappings existed for its strains.

The project's scope was deliberate and finite – ten strains, a well-defined functional annotation goal, and a clear public output. It was designed to empower rather than impress: to be a foundational reference for future researchers, while also serving as a learning scaffold. It was also Jin Xing's last contribution in bioinformatics before he transitioned fully into mathematics for his PhD studies.

Reflection: I view this paper now with a mix of gratitude and melancholy. It was a pragmatic contribution – a useful, replicable, and perhaps unremarkable to some. But to me, it symbolized a deep commitment to collaborative mentorship and utility-driven research. Jin Xing was still early in his career, and I was glad to provide space and structure for him to grow, even if our paths were soon to diverge. There's a quiet pride in supporting someone else's journey, especially when you know your scientific roads may not cross again.

This was also one of the last times I directly engaged in mapping GO and KO terms at scale. I've since moved more towards conceptual and reflective work, but this paper reminds me of the tangible value in producing clean, interpretable datasets for the community. It may not win citations, but it clears the underbrush for more ambitious work to follow.

Sidebar: A Farewell in Function

Annotations are the labels we place to make sense of biological chaos. In this study, they were also a farewell gesture – my final effort with a trusted colleague before he stepped into a new intellectual domain. I sometimes wonder what annotations we leave on each other’s lives: a method here, a habit there, perhaps a way of thinking. Science, like life, leaves its mark quietly, and often on the people we’ve helped along the way.

72: Fabrication and Falsification in Research

Citation: Ling, MHT. 2018. Science/Education Portraits III: Perceived Prevalence of Data Fabrication and/or Falsification in Research. Advances in Biotechnology and Microbiology 11(5): 555824.

Abstract: Data fabrication or falsification are considered as "deadly sins" with high impact, above plagiarism, on scientific truth and public confidence. What is the estimated prevalence of data fabrication or falsification? A meta-survey published a decade ago estimated 14.12% of respondents having knowledge of a colleague who fabricated or falsified research data, or who altered or modified research data. This mini-review updates this meta-survey by examining surveys from 2009 to 2018. Results suggests that 17.7% of responses indicated knowledge of fellow scientist's acts of data fabrication or data falsification. This is consistent with that of a decade ago, which suggests a critical need to address the worst type of research misconduct – data fabrication and falsification.

Context: This mini-review emerged from a growing discomfort I had watching the scientific enterprise wobble under the weight of misconduct. Unlike plagiarism; often caught, often punished; data fabrication and falsification attack the very fabric of scientific knowledge, and yet their occurrence remains difficult to measure. Triggered by high-profile cases and a nagging intuition that such malpractices were more common than openly admitted, I returned to the literature. A decade after a landmark meta-survey estimated 14.12% of scientists knew of a colleague engaging in such acts, I wanted to know: had things changed?

The article revisits the question with an updated lens, surveying studies from 2009 to 2018. The result, 17.7%, was unsettlingly consistent. Worse, it suggested normalization, or at least a persistent silence around misconduct. This paper is the third in my Science/Education Portraits (SEP) series, which aims to humanize and interrogate the ecosystem of science beyond just its results.

Reflection: Writing this paper was not just an academic exercise; it was an ethical meditation. As someone who has always advocated for rigorous, transparent, and responsible science; and often in environments where such values are quietly undermined – this topic struck close to home. There is a grief that accompanies these findings. Not personal betrayal, but systemic betrayal, the slow erosion of trust in something we once believed to be self-correcting.

I also knew this paper wouldn't be widely cited or embraced. It doesn't offer new techniques or discoveries; it holds up a mirror. But I felt compelled to publish it. If we, as scientists, do not speak honestly about our own failings, we surrender that space to silence and silence is what misconduct thrives in.

I wonder sometimes if writing this paper was my way of staying clean, of distancing myself from a culture that sometimes felt complicit. It was not enough. But it was something.

Sidebar: The Quiet Epidemic

We imagine science as self-correcting, yet misconduct festers in the cracks of peer review, pressure, and prestige. Fabrication and falsification are rarely confessed but often witnessed. This review did not ask who was guilty. It asked how many knew. The result: nearly one in five. Like an iceberg, the visible scandals are dwarfed by the submerged truths; seen, known, but unspoken**.**

73: SEREBO

Citation: Ling, MHT. 2018. SEcured REcorder BOx (SEREBO) Based on Blockchain Technology for Immutable Data Management and Notarization. MOJ Proteomics & Bioinformatics 7(6):169–174.

Abstract: Several surveys suggest that as many as 33% of scientists have personal knowledge of a colleague who fabricated or falsified research data. This indicates the need of a system that can aid the assurance that research data is not modified. Blockchain technology ensures data authenticity as recorded data is not mutable. In this study, a command-line data recorder and notary service based on blockchain, SEcured REcorder BOx (SEREBO), is presented. SEREBO can help individual scientists or research teams to prove data authenticity after logging data files into the system, and to provide traceable notarization records. Hence, SEREBO a potentially important tool for auditing research data against modifications, and auditing notarization events against backdating or postdating. SEREBO is available for forking at https://github.com/mauriceling/serebo under GNU General Public License version 3 for non-commercial or academic use only.

Context: Following the unsettling findings of Chapter 72, on the widespread perception of data falsification, I felt a professional and ethical imperative to not just describe the problem, but to contribute a solution. SEREBO, or SEcured REcorder BOx, was born out of this motivation. I had been observing the evolution of blockchain technologies beyond cryptocurrencies and saw in it the potential to fundamentally transform scientific data management.

SEREBO leverages blockchain principles to create a command-line tool that offers immutable timestamping and verification of research files. It was designed to provide individual researchers with a lightweight but trustworthy method to ensure and demonstrate data integrity, independent of centralized institutions. The project was released under GNU GPL v3, openly accessible on GitHub, and accompanied by a living wiki to support ongoing use and adaptation.

This chapter is a direct extension of the ethical concerns raised in Chapter 72, turning critique into constructive engineering.

Reflection: SEREBO was as much an ethical act as it was a technical one. I had no illusions that it would be widely adopted, especially in an academic culture that often sees reproducibility and integrity as abstract ideals rather than operational necessities. But I wanted to model what responsible infrastructure could look like. It was also a way of reclaiming some agency: if trust in science is eroding, we need new ways to earn and verify that trust, not just hope for better behavior.

Building SEREBO felt like stitching integrity into code. It became a quiet protest against the systemic forces that make it easier to manipulate data than to protect it. Even if it serves only a small community of like-minded researchers, it is a signal that alternatives exist.

This chapter stands as a bridge between critique and construction, between recognizing what's broken and attempting to fix it, even if imperfectly.

Sidebar: From Blockchain to Benchside

Most researchers still rely on good faith and institutional reputation to vouch for data authenticity. But what if we had tools; small, local, and verifiable; that proved our data had not been tampered with? SEREBO doesn’t require a revolution; just a decision to log. By tying scientific integrity to blockchain timestamping, it nudges us toward a culture of verifiable honesty.

74: Big Data for Personalized Medicine

Citation: Suwinski, P, Ong, CK, Ling, MH, Poh, YM, Khan, AM, Ong, HS. 2019. Advancing Personalized Medicine through the Application of Whole Exome Sequencing and Big Data Analytics. Frontiers in Genetics 10: 49.

Abstract: There is a growing attention towards personalized medicine. This is led by a fundamental shift from the 'one size fits all' paradigm for treatment of patients with conditions or predisposition to diseases, to one that embraces novel approaches, such as tailored target therapies, to achieve the best possible outcomes. Driven by these, several national and international genome projects have been initiated to reap the benefits of personalized medicine. Exome and targeted sequencing provide a balance between cost and benefit, in contrast to Whole Genome Sequencing (WGS). Whole Exome Sequencing (WES) targets approximately 3% of the whole genome, which is the basis for protein-coding genes. Nonetheless, it has the characteristics of big data in large deployment. Herein, the application of WES and its relevance in advancing Personalized Medicine is reviewed. WES is mapped to Big Data "10 Vs" and the resulting challenges discussed. Application of existing biological databases and bioinformatics tools to address the bottleneck in data processing and analysis are presented, including the need for new generation big data analytics for the multi-omics challenges of personalized medicine. This includes the incorporation of artificial intelligence (AI) in the clinical utility landscape of genomic information, and future consideration to create a new frontier towards advancing the field of personalized medicine.

Context: This review was a turning point in more ways than one. Published in Frontiers in Genetics, it not only explored the intersection of Whole Exome Sequencing (WES), big data, and personalized medicine, but also opened the door to my formal association with Frontiers. It marked the beginning of my roles as Associate Editor in Computational Genomics and Review Editor in STEM Education.

At the time, I was serving as a Research Assistant Professor at Perdana University's School of Data Sciences. There was optimism in the air: we were contributing to a rapidly expanding field that promised to revolutionize healthcare. The review took a systems-level view of how WES, though limited to just 3% of the genome, offered a cost-effective and data-rich platform for the advancement of personalized medicine. We contextualized this within the framework of Big Data's "10 Vs," addressing the technological, analytic, and infrastructural needs to make genomics actionable in real-world clinical settings.

But the promise of this chapter is shadowed by its aftermath. By 2020, the COVID-19 pandemic reshaped the world and the university. All authors on this paper, myself included, would eventually leave Perdana University.

Reflection: This paper represents both the ascent and the collapse of an era. Intellectually, it was exhilarating to contribute to the narrative of personalized medicine at a time when the convergence of AI, genomics, and big data felt like the future materializing. Professionally, it marked the start of my editorial journey with Frontiers, an unexpected but meaningful recognition of my interdisciplinary vantage point.

But emotionally, it is bittersweet. The collapse of the Perdana team amid the pandemic upheaval left a sense of rupture, a breaking of what once felt cohesive and promising. What remains is this publication, a digital monument to a moment when ambition, collaboration, and hope were still intact.

This chapter reminds me that scientific work is never just technical; it is deeply entangled with the human stories behind it – stories of institutions, upheavals, and transient communities of shared purpose.

Sidebar: The Paper That Changed Everything

Sometimes a publication is more than a publication. This review led directly to my editorial appointments at Frontiers, offering me a new platform to shape scholarly conversations in computational genomics and STEM education. The irony? Just as doors opened elsewhere, the one at Perdana quietly closed.

75: Personopreneurship

Citation: Ling, MHT. 2019. Science/Education Portraits IV: Experiences from a Decade as Informal Career Counsellor can be Summarized as Personopreneurship. Acta Scientific Medical Sciences 3(3): 151-156.

Abstract: A teacher is an informal career counsellor; yet, teachers are often the counselee. This essay summarizes the career guidance received and a decade as informal career counsellor as personopreneurship – each of us is our business and the general manager of this business. This is consistent with the view that our curriculum vitae/resume is likened to a marketing document or product brochure. I consolidated the six key themes of personopreneurship as [a] sufficient skills utilization, [b] work that matches a calling and concordant with beliefs and faith, [c] sufficient workload, [d] good working environment, [e] support by superiors, peers, and subordinates; including availability of mentoring, autonomy and freedom to operate/manoeuvre, and [f] sufficient and stable income. Each of these themes will be reviewed based on current literature.

Context: Published in Acta Scientific Medical Sciences, this reflective essay distills a decade of experience providing informal career guidance to students, peers, and colleagues. Drawing on both personal experience and supporting literature, I introduced the concept of personopreneurship – the notion that each individual is a one-person enterprise, managing their career like a business.

The essay outlined six themes central to sustaining a fulfilling career: skills utilization, alignment with personal values and calling, manageable workload, healthy work environment, supportive networks, and financial stability. These principles were not abstract—they were grounded in real-life conversations, decisions, and mentoring relationships built over years in academia and beyond.

At the time, I believed “informal career counsellor” was an apt term for my role. Later, I would come to understand the regulatory importance of titles in mental health and career services. While I lacked the formal credentials to be called a counsellor, I remained a committed advisor and coach, offering honest, empathetic support based on lived experience.

Reflection: This piece is part confession, part manifesto. It is a distillation of the unspoken, often invisible work that educators do in helping others navigate the complex terrain of career and identity. I wrote it to honour those late-night conversations, those quiet turning points when someone needed clarity, perspective, or simply to be heard.

Personopreneurship emerged as both a descriptive and aspirational framework – one that validates the autonomy of individuals to design their own path, but also acknowledges the structural and emotional supports required to sustain such autonomy. It is a response to the instability of modern work, especially in academia, where roles are often precarious and institutional support uneven.

In writing this, I came to see my own journey as a case study. This was not just advice I had given; it was a philosophy I had lived.

Sidebar: The Ethics of Titles

After publishing this essay, I learned that the term “counsellor” carries specific legal and ethical weight. I now describe myself more accurately as an advisor or guide. But this correction does not diminish the heart of the work: showing up for others with sincerity, humility, and an unwavering belief in their potential.

76: Fastest Manuscript

Citation: Ling, MHT. 2019. De Novo Putative Protein Domains from Random Peptides. Acta Scientific Microbiology 2(4): 109-112.

Abstract: How prebiotic chemistry in the primordial world becomes biochemistry, is a major question in evolutionary biology. Studies have found that biological activities from random DNA sequences are not rare and abiotically-catalyzed polymerization of 13 amino acid chains can occur. However, it is not clear whether random chains 13 amino acid or longer are biologically functional. In this study, random peptide sequences were generated and mapped to ProSite motifs and NCBI Conserved Domains Database. Results suggest that a large fraction of randomly generated 13 amino acid chains may contain putative protein domains while longer random peptide chains may contain functional protein domains. Large diversity of protein domains is observed. Hence, it is plausible for putative functions to originate from abiotically-catalyzed 13 amino acid chains. As both self-replicating RNA molecules and prion proteins have been found, it is plausible that both RNA and peptides may co-exist and synergize in the primordial world.

Context: This brief paper, published in Acta Scientific Microbiology, tackled a profound question in evolutionary biology: could functional proteins have emerged from purely random peptide sequences in the primordial world? Inspired by studies on abiotic polymerization and the surprising biological activity of random DNA, I explored whether randomly generated 13-amino-acid chains could contain protein domain motifs.

Leveraging established bioinformatics databases like ProSite and NCBI's Conserved Domains Database, I demonstrated that a significant proportion of these random peptides could indeed match known functional motifs. This work lent computational support to a plausible model where peptide-based functionality may have arisen independently of, or in parallel with, early RNA worlds; thereby, adding weight to the hypothesis that RNA and peptides co-evolved or even co-operated in life's origins.

Remarkably, the idea for this paper struck on the night of March 11, 2019. By the next evening, the manuscript had been conceptualized, executed, written, and submitted. It remains the fastest paper I've ever produced**.**

Reflection: There was something electric about that night, when an idea so audacious and clear took hold and the entire project unfolded with clarity and speed. It reminded me why I first fell in love with science: for the thrill of curiosity, the elegant simplicity of a question that could shake foundational beliefs, and the immediacy of discovery.

This paper is a testament to the creative potential of spontaneous scientific inquiry. It also reflects my enduring fascination with origins of life, of function, of complexity from chaos. Though small in scope, the implications of this work are far-reaching. If random sequences can yield domain-like motifs, then perhaps nature's toolkit is more forgiving, more fertile, than we dare imagine.

Sidebar: One Day, One Paper

- March 11, 2019: The idea struck during my commute – could random peptides contain hidden order?
- March 12, 2019: I coded, ran simulations, cross-referenced results with known domain databases, and wrote up the findings. By evening, the manuscript was submitted.

Some discoveries need funding. Some need teams. This one just needed one night of unfiltered wonder.

77: SEREBO Codes

Citation: Ling, MHT. 2019. SEcured REcorder BOx (SEREBO) Version 1.0. In Current STEM, Volume 2. pp. 67-154. Nova Science Publishers, Inc. ISBN 978-1-53616-042-0.

Abstract: Data authenticity is crucial in many industries. A major aspect of data authenticity is to ensure that a created file is not fraudulently or purposefully edited; for example, changing the data file without affecting the date time stamp. Blockchain technology ensures data authenticity as recorded data is not mutable. This manuscript documents the implementation and the codes of SEREBO and licensed under GNU General Public License version 3. SEREBO codebase is hosted and available for forking at https://github.com/mauriceling/serebo.

Context: In an era where data manipulation can undermine trust across industries, from finance to healthcare to scientific research, the need for secure, tamper-proof records has become paramount. SEcured REcorder BOx (SEREBO) Version 1.0 was my contribution to this growing concern. Drawing upon the principles of blockchain, SEREBO was designed as a lightweight, verifiable logging system that ensures data authenticity. It preserves the integrity of recorded data by making retroactive modifications detectable and thus, by design, discourages tampering.

The manuscript, published as a monograph in Current STEM, provides not only a conceptual overview but also the full implementation code, openly licensed under the GNU GPL v3. The SEREBO codebase remains available for forking and reuse on GitHub, reflecting my long-standing commitment to open science and open-source development.

Reflection: SEREBO was not born from idle theorizing, it emerged from practical frustration. I had witnessed firsthand how easily digital files could be backdated, manipulated, or masked, especially in settings where transparency was critical. Whether in experimental logs, administrative records, or legal documentation, the potential for undetected data tampering threatened the very credibility of information systems.

While the blockchain hype cycle was in full swing at the time, I was less interested in speculative applications and more invested in real, implementable solutions. SEREBO was a personal challenge to bring blockchain principles back down to earth, an accessible, practical tool that anyone could use to secure their digital trail.

This project also marked a turning point in how I viewed software: not just as a product, but as a public service. By licensing SEREBO openly and sharing its architecture, I hoped to empower others to build upon the foundation and tailor it to their contexts. Whether or not it finds widespread use, SEREBO represents a moment of alignment between personal conviction and technological clarity.

Sidebar: Blockchain without the Buzz

While cryptocurrencies grabbed headlines, SEREBO focused on what blockchain quietly does best – ensuring data can't be retroactively altered without detection. No tokens. No mining. Just trust, encoded.

78: Archaebacterial Proteome Diversity

Citation: Kim, JH, Ling, MHT. 2019. Proteome Diversities Among 19 Archaebacterial Species. Acta Scientific Microbiology 2(5): 20-27.

Abstract: Archaebacteria is known for its presence in varied extreme environments, suggesting potential applications and an on-going need study its diversity. This led to increasing emphasis on archaeal genomic and proteomic studies. However, there is no work to-date examining the overall proteomic diversity in archaebacteria. In this study, we examine the proteomic diversities among 19 sequenced archaebacterial species and found significant differences (p-value < 2 x 10-43) in average peptide lengths, isoelectric points, aromaticity, instability, and hydropathy. Majority of the peptides in each species are stable. Predominantly consistent correlations, though widely varied, were observed between peptide physical properties except between peptide length and hydropathy. This study provides a cursory view highlighting the diversity of archaeal proteomes; thus, re-iterating the call for further studies into these organisms.

Context: Archaebacteria are among the most resilient life forms on Earth, thriving in extreme environments such as high-temperature vents, hypersaline lakes, and acidic springs. Despite this, comprehensive analyses of their proteomes had remained scarce. This study, co-authored with Jung Hwan Kim, one of my earliest honours students at MDIS, provided one of the first comparative views of proteomic diversity across 19 sequenced archaebacterial species. We analyzed properties like peptide length, isoelectric point, aromaticity, instability, and hydropathy, uncovering significant interspecies variations and revealing patterns worth deeper exploration.

This paper was a scientific milestone not just for its content, but because it marked my very first publication with a student under my direct mentorship – a moment of deep professional and personal significance.

Reflection: Working with Jung Hwan was as much a lesson in patience and pedagogy as it was a scientific collaboration. Language barriers, cultural differences, and technical setbacks all presented challenges but none as impactful as the day his laptop crashed, wiping out weeks of data. I remember the moment he hesitantly approached me, his fear evident, not just from the loss itself but from his anxiety over disappointing me. I reassured him that his project was still salvageable, and we pivoted by reducing the dataset and narrowing the scope. That decision didn't dilute the study's value; it distilled it.

In mentoring Jung Hwan, I learned that the true role of a supervisor isn't to steer the ship alone, but to guide the student through stormy waters, instilling not just knowledge but resilience. Our joint publication was more than a research output; it was a testament to persistence, adaptation, and the quiet triumph of a student who overcame his fears to finish strong.

Sidebar: A First of Many

- First paper with an honours student under my direct supervision.
- First milestone as an associate lecturer guiding undergraduate research.
- First rescue mission: A project saved from the ashes of a hard drive crash.

Jung Hwan Kim may have returned to South Korea to complete his national service, but his journey remains one of the most memorable starts to my mentorship path.

79: *Pseudomonas balearica* DSM 6083T

Citation: Maitra, A, Ling, MHT. 2019. Codon Usage Bias and Peptide Properties of *Pseudomonas balearica* DSM 6083T. MOJ Proteomics & Bioinformatics 8(2):27–39.

Abstract: *Pseudomonas balearica* DSM 6083T has potential applications in bioremediation and its genome is recently sequenced. Codon usage bias is important in the study of evolutionary pressures on the organism and physical properties of peptides may elucidate functional peptides. However, both have not been studied for *P. balearica* DSM 6083T. Here, we investigated the codon usage bias and peptide properties of the 4,050 coding sequences in *P. balearica*. Codon usage analysis suggests that all preferred codons were either G or C ending. There is a skew towards smaller peptides and all peptide properties (pI, aromaticity, hydropathy, and instability) are correlated (|r| > 0.102, p-value < 7e-11). %GC is correlated (|r| > 0.122, p-value < 6e-15) to peptide length, aromaticity, hydropathy, and instability. Peptide length is correlated (|r| < 0.057, p-value < 0.0003) to pI, aromaticity, and instability. Codon usage is correlated (r < -0.042, p-value < 0.0075) with all peptide properties while amino acid usage is correlated (r < -0.084, p-value < 8e-8) to all peptide properties except instability. A substantial proportion (26.9%) of genes show significantly different codon and amino acid ratios compared to the genomic and proteomic averages respectively (p-value < 1.2e-5), suggesting potential exogenous origins. These results suggest a complex interplay of metagenomic environment and various genomic / proteomic properties in shaping the evolution of *P. balearica* DSM 6083T.

Context: *Pseudomonas balearica* DSM 6083T is an environmentally intriguing bacterium with potential for bioremediation, yet it remains poorly studied. By 2019, fewer than 20 papers had appeared on this species in PubMed, and as of 2025, the number is still under 40. This rarity made it scientifically strategic to explore its genome before the research field becomes saturated.

In this study, we analyzed codon usage bias and peptide properties across 4,050 protein-coding genes in *P. balearica*. This work was not just a detailed case study; it served as an early stake in a new and largely uncharted microbial territory.

Reflection: This paper marks my second publication with an undergraduate honours student from MDIS, and once again, the collaboration was deeply human. Argho Maitra, a bright and diligent student from India, was in the same cohort as Jung Hwan Kim. His careful attention to detail and dedication stood out early on. While the data was complex and required extensive statistical interpretation, Argho handled the challenges with quiet determination.

What made this paper even more poignant for me is what came after. Argho later joined me as a Master's student at Perdana University, a clear indicator of his passion for science and his trust in me as a mentor. Unfortunately, the convergence of the COVID-19 pandemic and personal family obligations forced him to withdraw. That loss still lingers. I think often about the bright future he could have had in academia. But even if his academic journey was interrupted, I hope our collaboration gave him a meaningful scientific experience and a sense of what he was capable of.

Sidebar: A Claim on Uncharted Ground

- *Pseudomonas balearica* remains a little-known species and this paper helped put it on the map.
- Argho was a standout student, both as an undergraduate and a Master’s mentee.
- This was the first published study examining both codon usage and peptide properties in *P. balearica*.

Argho’s academic path may have been cut short but this work stands as a testament to his ability and promise.

80: *De Novo* Origins of Genes 1

Citation: Thong-Ek, C, Usman, S, Woo, JH, Chua, JW, Kwek, BZN, Ardhanari-Shanmugam, KD, B, V, Shahrukh, K, Ling, MHT. 2019. Potential *De Novo* Origins of Archaebacterial Glycerol-1-Phosphate Dehydrogenase (G1PDH). Acta Scientific Microbiology 2(6): 106-110.

Abstract: Eubacterial glycerol-1-phosphate dehydrogenase (G1PDH) may originate from archaebacteria by horizontal gene transfer; however, the origins of archaebacterial G1PDH remains unanswered. While recent studies show possible *de novo* origination of protein encoding genes and functional promoters, the mechanism of *de novo* origins of functional genes remains debatable. In this study, we examine the probability of *de novo* emergence of putative G1PDH from random sequences. Our results show that high number of open reading frames in random sequences and 71.8% of randomly generated sequences have 9.88% probability of being putative G1PDH. Hence, *de novo* origination archaebacterial G1PDH from random sequences is plausible.

Context: The evolutionary origins of glycerol-1-phosphate dehydrogenase (G1PDH), an enzyme critical for membrane lipid biosynthesis in archaea, have long puzzled evolutionary biologists. While eubacterial G1PDH is thought to have been acquired from archaea via horizontal gene transfer, the ancestral roots of archaeal G1PDH remain obscure. This paper explored a radical hypothesis: that archaebacterial G1PDH could have emerged *de novo* from random sequences, rather than by duplication or recombination of pre-existing genes.

By simulating random nucleotide sequences and evaluating the probability that these sequences could encode G1PDH-like open reading frames, we found that a surprisingly high fraction (71.8%) of the generated sequences had nearly 10% likelihood of coding for putative G1PDH. This supports the theoretical plausibility of *De Novo* gene origination – an idea still on the fringe of evolutionary biology at the time.

Reflection: This was Chakrit Thong-Ek's first publication, completed while he was still in the middle of his honours project at MDIS. The paper was published before his thesis even concluded, which is a testament to his curiosity and drive. Most students hesitate to dive into theoretical questions, but Chakrit had a boldness about him. He wasn't intimidated by abstract evolutionary puzzles, and he brought a deep sense of wonder to the idea that functional genes might spontaneously emerge from randomness.

Since then, Chakrit has taken an unexpected but fitting path. He completed a Master's in Data Analytics and is now a Data Engineer – a transition that makes perfect sense. His instinct for pattern recognition, abstraction, and systems thinking was evident even during this project. I see this paper not only as a theoretical contribution to evolutionary biology but also as an early expression of Chakrit’s ability to navigate complexity and uncertainty; these traits that will serve him well in data science.

Sidebar: A Head Start on the Unknown

- Published before Chakrit’s honours project term officially ended.
- Tackled one of evolutionary biology’s most speculative questions: Can functional genes arise from scratch?
- Simulated thousands of random sequences to show that G1PDH-like genes could plausibly emerge *de novo*.

This was not just a research project; it was an early glimpse into a mind that now thrives on solving complex data challenges.

81: *De Novo* Origins of Genes 2

Citation: Kwek, BZN, Ardhanari-Shanmugam, KD, Woo, JH, Usman, S, Chua, JW, B, V, Shahrukh, K, Thong-Ek, C, Ling, MHT. 2019. Random Sequences May Have Putative Beta-Lactamase Properties. Acta Scientific Medical Sciences 3(7): 113-117.

Abstract: Beta-lactamases, which confer resistance to beta-lactam antibiotics, is of medical and healthcare concerns globally. Studies had placed the emergence of beta-lactamases to more than 2 billion years ago. However, it is not known where the first beta-lactamase originate. In this study, we examine the probability of *de novo* emergence of putative beta-lactamase from random sequences. A set of 10 thousand randomly generated sequences were aligned using Smith-Waterman algorithm and Needleman-Wunsch algorithm to a set of known class D beta-lactamases isolated from GenBank to determine the probability of each randomly generated sequence as putative beta-lactamases. Our results suggest that substantial proportion of randomly generated sequences may be putative beta-lactamases, with 4% of the randomly generated sequences showing 99% probability as putative beta-lactamases. To test whether a putative beta-lactamase can evolve over generations to have more characteristics of known beta-lactamases, *in silico* evolution was carried out using DOSE, an evolution simulation software. Our simulation results also suggest that a putative beta-lactamase may rapidly evolve into a more functional beta-lactamase under selection. Hence, *de novo* origination of beta-lactamase from random sequences is plausible.

Context: Beta-lactamases are enzymes that neutralize beta-lactam antibiotics, underpinning much of today's antibiotic resistance crisis. While modern beta-lactamases are well-studied, their evolutionary origin remains a mystery. Like the preceding study on G1PDH, this project challenged conventional thinking by exploring whether functional beta-lactamase-like sequences could arise *de novo* – not through descent or recombination, but from random nucleotide sequences.

We generated 10,000 random sequences and compared them to known Class D beta-lactamases using the Smith-Waterman and Needleman-Wunsch algorithms. Remarkably, 4% of these sequences had a 99% probability of being putative beta-lactamases. Furthermore, we used *in silico* evolution simulations (via DOSE) to show that these sequences could rapidly evolve toward known beta-lactamase structures under selective pressure. These findings add to the growing plausibility that essential enzymes might originate spontaneously given the right conditions and evolutionary dynamics.

Reflection: This paper was a natural complement to Chakrit's work on G1PDH. Brenda Kwek and Chakrit were classmates, but Brenda approached the project with a more technical, results-driven focus – one that reflected her steady, grounded nature. Whereas Chakrit thrived on abstraction, Brenda was motivated by the tangible implications: antibiotic resistance, public health, and the real-world urgency of understanding beta-lactamase evolution.

It’s fitting that Brenda went on to become a Lead Laboratory Technician. She had a meticulous attention to detail and a no-nonsense attitude that made her a stabilizing presence in any research team. This paper was her intellectual leap beyond the bench, a foray into evolutionary computation and bioinformatics, and she handled it with grace and competence. Watching both students approach the same foundational question from such different angles reminded me how science is both a shared endeavour and a deeply personal one.

Sidebar: Evolving Resistance from Nothing

- 4% of 10,000 random sequences showed 99% probability of being putative beta-lactamases.
- Used *in silico* evolution to show rapid adaptation toward known beta-lactamase profiles.
- Adds weight to the controversial idea that resistance genes may arise spontaneously.

Brenda brought structure to a chaotic question and left a scientific mark that extends far beyond her lab bench.

82: Tertiary Education I Envisioned

Citation: Ling, MHT. 2019. Science/Education Portraits V: The Scientific Tertiary Education that I had Envisioned. Acta Scientific Medical Sciences 2(8): 75-79.

Abstract: Current tertiary education system has been criticized as being outmoded and inadequate to address modern day needs. Incorporating the concept of stackable credentials, I propose a framework of seminar series and apprenticeship within the 4-year tertiary education system for science. Seminar series can replace teaching while apprenticeship can address the needs of education. Given that a typical undergraduate degree amounts to 2,000 hours, I demonstrate that a combination of seminar and apprenticeship can amount to 1,860 seminar hours and 1,000 apprenticeship hours, which is equivalent to 6 months of full-time employment. This can allow the student room to develop a competency-based portfolio.

Context: This essay was the fifth in my Science/Education Portraits series, a personal effort to articulate and archive my evolving philosophy on science education. In this piece, I laid out a bold, structural alternative to traditional tertiary education but one that replaces didactic teaching with seminar-style dialogues and substitutes conventional coursework with genuine apprenticeship.

The core of the model was pragmatic: a four-year science degree amounts to roughly 2,000 hours. Why not reallocate those hours into a blend of 1,860 hours of seminar-style learning and 1,000 hours of apprenticeship, with the latter equivalent to six months of full-time work experience? In doing so, students could graduate not just with knowledge, but with demonstrable competence and a real portfolio of work. Stackable credentials could further allow students to layer interdisciplinary skills in response to their emerging interests or industry shifts.

Reflection: This article distilled years of frustration and aspiration into a single vision. I had long felt that university science education was overly compartmentalized and poorly aligned with how real science is practiced. Lectures felt like rituals. Exams rewarded memory rather than inquiry. Worst of all, students graduated with neither a deep conceptual framework nor relevant workplace experience.

The apprenticeship model I proposed was not theoretical. I had been living it for years, mentoring students in real projects, from random sequence analysis to public health informatics. The seminar model, too, echoed my best teaching moments; those unscripted, Socratic discussions where students discovered their voices as thinkers.

I knew this was an ideal that many institutions weren't ready for. But I also knew that ideals matter. They are the seed crystals around which reforms can grow. This piece was one of those seed crystals and in retrospect, it was also a blueprint for the kind of mentorship-centered, research-embedded teaching life I had been building all along.

Sidebar: A Degree Measured in Competence, Not Credits

A traditional science degree spans approximately 2,000 hours over four years, yet much of that time is spent passively absorbing lectures and preparing for exams. In contrast, the model I envisioned restructures this time into 1,860 hours of seminar-style engagement—where dialogue replaces didactic instruction and 1,000 hours of apprenticeship, the equivalent of six months of full-time professional experience. Rather than accumulating grades, students would graduate with a portfolio of real-world competencies, developed through practice, collaboration, and critical thinking. This approach reframes the university experience as a launchpad for meaningful contribution and not merely a rite of passage.

83: Demystifying Monod

Citation: Chang, ED, Ling, MHT. 2019. Explaining Monod in Terms of *Escherichia coli* Metabolism. Acta Scientific Microbiology 2(9): 66-71.

Abstract: Monod Equation is a simple empirical equation relating limiting substrate to cell growth rate. Despite being used in many studies, there is a need to elucidate growth rate in terms of metabolism, which is then used to inform metabolic engineering efforts. Here, we attempt to explain Monod Equation in terms of simulated metabolism, in the form of metabolic flux, from an *Escherichia* coli MG1655 flux balance analysis (FBA) model to yield a growth rate objective function. Flux values represent change of molecule concentrations over time, making biomass objective function a rate equation. This poses difficulty in representing biomass objective function as a predictive model of metabolic fluxes, which is essentially an analytical equation of fluxes. Our results show a strong correlation (r = 0.972, p-value = 1.16 x 10^-14) between Monod's predicted growth rate and biomass objective value from FBA model. Using this relationship, Monod's predicted growth rate can be predicted by 14 fluxes (r = 1, p-value < 1 x 10^-16, SSE = 2.3 x 10^-7, MSE = 1.8 x 10^-9). Therefore, this study explains the growth rate of *E. coli* MG1655 in terms of its metabolic flux and presents a methodology for unifying Monod Equation with simulated or experimental metabolism.

Context: The Monod equation, formulated in the mid-20th century, remains one of the most cited mathematical models in microbiology. It captures the relationship between the concentration of a limiting substrate and the specific growth rate of microorganisms. However, despite its widespread use in bioprocess engineering and environmental microbiology, the Monod equation is largely empirical as its constants are derived from experiments, not first principles. This creates a disjuncture between high-level metabolic models such as Flux Balance Analysis (FBA) and the operational simplicity of Monod kinetics.

This project arose out of a pedagogical question: Could Monod's empirical growth function be reconciled with the mechanistic underpinnings of cellular metabolism? Ega Chang, then a third-year biotechnology student at Temasek Polytechnic, had been referred to me by Dr. Chan Giek Far as a candidate for mentorship. We chose this problem not just for its intellectual merit, but because it offered a rare opportunity to bring together empirical microbiology, metabolic modelling, and mathematical reasoning in a single, student-accessible research framework.

Reflection: This study represents more than a technical bridge between empirical and mechanistic biology; it is also an emblem of how deep mentorship can alter the trajectory of a student's thinking. The research may appear modest in scale, but its implications are significant: we showed that Monod's predictions can be explained and reproduced with remarkable precision using only a handful metabolic fluxes, effectively making the Monod equation interpretable through the lens of systems biology.

From a scientific standpoint, this work hinted at a possible unification between kinetic models and stoichiometric ones, suggesting that models need not remain siloed in tradition but can be harmonized through simulation and careful validation. From an educational standpoint, it proved that even polytechnic students; given the right question, the right tools, and the right mentorship; can participate in resolving long-standing conceptual gaps in the literature.

And from a personal standpoint, it was a rare opportunity to engage directly, if metaphorically, with a figure in my academic lineage. Working on Monod's equation felt like a dialogue across generations. It was, in many ways, a moment of scientific and intellectual closure, made richer by the fact that it was shared with a student just beginning his own journey.

Sidebar: From Monod to Metabolism – A Pedagogical Experiment in Precision
This project began not with an experimental protocol, but with a question: Can Monod's empirical law be explained mechanistically through metabolism? It was a question I posed to Ega Chang, not to pressure but to stretch. At that time, he was a third-year student at Temasek Polytechnic, still navigating the boundaries between classroom instruction and real-world scientific inquiry. Referred to me by Dr. Chan Giek Far as someone with potential worth grooming, Ega embodied what I consider the ideal student: curious, coachable, and unafraid of mathematical abstraction.

There was also a personal reason why this project mattered. Jacques Lucien Monod holds a distinct place in my heart, not just for his foundational work in molecular biology, but because I am, in a very real academic sense, his descendant. Through the mentorship line of scientific training, documented in AcademicTree, I can trace my scholarly ancestry directly to Monod. That lineage lent this project a deeper resonance. In a quiet way, it felt like paying homage to an intellectual forebear.

The Monod equation is often memorised but seldom understood. It offers a deceptively simple relationship between substrate concentration and microbial growth, yet the inner workings that give rise to such a curve remain obscure in many educational settings. What Ega and I set out to do was not just reinterpret Monod through the lens of metabolic flux analysis, but to use the FBA model of *E. coli* as a kind of scaffold to build that missing conceptual bridge. Our success in finding a perfect correlation between growth rates and a subset of 14 fluxes wasn't just a result, it was a revelation. It showed that empirical models in biology can indeed be reverse-engineered into mechanistic understanding, with enough computational scaffolding.

Today, Ega is in private equity, but I suspect that the discipline of system thinking and modelling we cultivated together has not left him. This paper remains a quiet milestone: a reminder that even a foundational concept like the Monod equation can become a launchpad for students to build lasting, integrative thinking skills – skills that persist beyond the lab.

84: *De Novo* Origins of Promoters 1

Citation: Ardhanari-Shanmugam, KD, Shahrukh, K, B, V, Woo, JH, Thong-Ek, C, Usman, S, Kwek, BZN, Chua, JW, Ling, MHT. 2019. *De Novo* Origination of *Bacillus subtilis* 168 Promoters from Random Sequences. Acta Scientific Microbiology 2(11): 07-10.

Abstract: How the first promoters may have originated is of evolutionary curiosity. Several studies have shown that new promoters arise by copying over an existing promoter sequence. Although *de novo* origination of promoters has also been suggested, there has been limited evidence. Hence, we investigate the possibility of *de novo* origination of promoters in this study using the model organism *Bacillus subtilis* 168. 10,000 random sequences were generated and alignment to known promoter sequences from *B. subtilis* 168 were used to assess their probability of being putative promoters. Results showed that 380 out of 10,000 random sequences have ≥97% probability. *In silico* evolution was performed to test the possibility of promoter selection using selective pressure and our simulation results suggest that the functionality of a random sequence may increase overtime. Therefore, *de novo* origination of promoters from random sequences is possible.

Context: The origin of promoters, key regulatory DNA elements that initiate transcription, is a fundamental question in molecular evolution. While it is well-established that new promoters can arise through duplication and mutation of existing sequences, the possibility of their spontaneous emergence from random sequences remains a subject of debate. This question is not merely academic: it strikes at the heart of how complex regulatory networks may have first formed in early life.

This project set out to test the evolutionary plausibility of *de novo* promoter origination using *Bacillus subtilis* 168 as a model organism. Through computational alignment of 10,000 random DNA sequences to known *B. subtilis* promoters, we examined the likelihood that any of these random sequences might exhibit promoter-like properties. Surprisingly, 380 sequences scored ≥97% probability of being promoters. Further *in silico* evolution under selective pressure suggested that these proto-promoters could evolve into more functionally robust sequences over time.

The project was led by Keerthana Devi Ardhanari-Shanmugam, alongside a team that included Shahrukh, Brenda, V, Woo, Chakrit, Usman, Kwek, and Chua – all students from the same batch. This work reflects both a computational inquiry into molecular evolution and a pedagogical experiment in hypothesis-driven learning.

Reflection: This project was one of the more speculative pieces I've supervised, and also one of the most satisfying. It questioned a core dogma in molecular biology, that regulatory elements must evolve only from pre-existing scaffolds, and suggested that, at least in theory, functional promoter sequences can emerge from random noise, given the right context and pressures. In doing so, it brought together the elegance of evolutionary logic and the precision of computational analysis.

The fact that nearly 4% of random sequences scored exceptionally high as putative promoters was, in itself, surprising. But the evolutionary simulation was what gave the study depth: it showed that functionality can emerge and consolidate over time, supporting the plausibility of *de novo* regulatory origination. This insight has implications for synthetic biology, where the design of novel promoters remains a challenge, as well as for origin-of-life research.

But more than the findings, I am proud of the intellectual courage this team displayed. They embraced a controversial idea, translated it into a viable methodology, and extracted meaning from the chaos of randomness. That, to me, is science in its most creative form—an interplay of imagination, rigour, and a willingness to ask foundational questions.

Sidebar: Where Promoters Come From – A Thought Experiment with Code

This study began with a deceptively simple question: Can promoters, the on-switches of genes. originate from scratch? It's a question with deep evolutionary implications and one that lent itself well to computational exploration. Keerthana Ardhanari-Shanmugam and her cohort, including Chakrit and Brenda, were in their final year when they approached me for project supervision. I knew this group had the right balance of discipline, curiosity, and a willingness to explore outside their comfort zone.

I pitched them the idea: generate 10,000 random sequences and see if any look like *B. subtilis* promoters. It was both a simulation project and an act of intellectual rebellion—against the assumption that biological functionality must always come from duplication and modification of pre-existing parts. The students responded with enthusiasm, curiosity, and rigour. They not only designed the pipeline to generate sequences and score them, but also implemented an *in silico* evolution process to test whether weak promoter-like sequences could evolve toward greater functionality under simulated selective pressure.

This project resonated with something deeper for me as well. It echoed a recurring theme in my research and teaching: that the boundary between randomness and order is not always as firm as we think. It was also gratifying to see Keerthana, now thriving in the biotechnology sector, make the leap from classroom theory to evolutionary reasoning, supported by code, data, and a team she trusted.

85: Jigsaw Cryptography 2

Citation: Ling, MHT. 2019. Draft Implementation of a Method to Secure Data by File Fragmentation. Acta Scientific Computer Sciences 1(2): 10-13.

Abstract: Cryptography is fundamental in data security and is a critical tool in safeguarding information from "unauthorized" view during the storage and transportation of data. Due to the one-to-one correspondence between plain text and cipher text, encryption algorithms are transformation processes. This implies that all information is present, though encrypted, in the cipher text. Inspired by Jigsaw puzzles, a new cryptography system, Jigsaw Cryptography System (JCS) is proposed where a single plain text file results in many cipher text files, resembling jigsaw pieces from a single image. Thus, the interception of a small number of cipher text files may not compromise the entire contents in plain text. This can result in larger permutations needed to decipher by brute force, which is not easily achievable in most cryptographic methods.

Context: In 2017 (published in 2018), I proposed a foundational framework for Jigsaw Cryptography in the paper "A Cryptography Method Inspired by Jigsaw Puzzles", which laid the groundwork for what would later become the Jigsaw Cryptography System (JCS), implemented in my 2019 paper "Draft Implementation of a Method to Secure Data by File Fragmentation" published in Acta Scientific Computer Sciences. The earlier paper introduced the idea of using fragmentation to secure data by breaking it into several pieces, similar to a jigsaw puzzle, and encrypting each piece independently. The basic principle of the Jigsaw Encryption System (JES) was to increase the difficulty of data interception and decryption by creating a system in which each fragment of the encrypted data could only be decoded if enough other fragments were obtained. This was in contrast to traditional encryption methods, where all the plaintext information is typically embedded in the ciphertext in a one-to-one correspondence.

JCS, as the usable implementation of JES, took the theoretical concept further by providing a practical framework that demonstrated how to securely fragment files and encrypt those fragments individually. In this newer version, I developed an approach that not only encrypted the fragments but also reassembled them at the point of decryption, increasing the complexity for anyone attempting to decode the file without access to all of the fragmented pieces. The two concepts, JES and JCS, were fundamentally linked, but JCS represented a more complete, practical solution, emphasizing the feasibility and applicability of this approach in real-world cryptographic scenarios.

Reflection: Looking back, I can see how Jigsaw Cryptography evolved from a theoretical exploration in JES into a functional system in JCS. The transition from concept to implementation was a critical moment in my work, where the technical details became just as important as the creative problem-solving that initiated the idea. The ability to take a novel approach and adapt it to a usable system that could be tested in practical scenarios was both an intellectual challenge and a professional milestone. What began as a conceptual critique of traditional encryption methods ultimately materialized into a framework that could potentially disrupt how cryptography is applied to data security.

However, like many theoretical advancements, the real challenge was in addressing the practical limitations of JCS. One concern that became clear in the implementation phase was how the fragmentation could potentially affect the efficiency of the system, especially in terms of the computational load required to encrypt and decrypt data. Furthermore, ensuring that the reassembly process was secure and reliable was a critical aspect of JCS that I had to refine continuously. This effort of converting JES into JCS was a testament to the importance of both theory and practice in cryptography, illustrating how new methods can come to life when grounded in solid technical implementation.

Reflecting on both JES and JCS, it's clear that the puzzle analogy, fragmenting information and making it harder to piece together without all parts, was an idea that captured my imagination because of its natural fit with the complexity and security challenges of cryptography. It also made me appreciate the importance of developing systems that are both creative and scalable, an important takeaway that I carried into my future work.

Sidebar: Bridging Theory and Practice in Cryptography

The transition from theoretical frameworks like JES to usable implementations such as JCS is a key part of the development of new cryptographic methods. While the theory provides the foundational principles and ideas, it is the practical implementation that tests the feasibility of these concepts in real-world applications. This process often reveals new challenges, as the abstract becomes concrete. In the case of Jigsaw Cryptography, the move from JES to JCS was not just about translating an idea into a system; it was about optimizing that system to be both secure and efficient enough for real-world use. As cryptography evolves, the gap between theory and implementation continues to shrink, resulting in more robust and creative methods of securing data.

86: Simple Genetic Drift Simulator

Citation: Ling, MHT. 2019. Island: A Simple Forward Simulation Tool for Population Genetics. Acta Scientific Computer Sciences 1(2): 20-22.

Abstract: Changes in population genetic structure can be a result of genetic drift and/or selective pressure, which may result in changes in adaptability of the population. Computer simulations are commonly used to gain insights into the genetic fate of evolving populations. However, most simulation tools in this area require a firm understanding of the mathematical models of genetic drift but low-cost, hands-on tools are the key to make abstract concepts, such as genetic drift, more intuitive. Here, Island is presented as simple forward simulation tool for population genetics based on Mendelian inheritance where a population is generated from a comma-delimited file containing allelic frequencies. Forward simulations start from an initial population and track its evolution over multiple generations. The population is simulated over generations where each generation results in a population file, which can then be examined independently to observe changes in allelic frequencies over generations.

Context: In 2019, I published the paper “Island: A Simple Forward Simulation Tool for Population Genetics” in Acta Scientific Computer Sciences. This work presented Island, a forward simulation tool designed to make population genetics more accessible and intuitive. One of the challenges in studying genetics is understanding how genetic drift and selective pressure affect population structure over time. Most existing simulation tools in population genetics require a solid understanding of complex mathematical models, which can make them difficult for non-experts to use. Island was developed to bridge this gap, providing a straightforward tool that allowed users to simulate populations and observe genetic changes in real-time, starting from initial allele frequencies in a comma-delimited file.

The basic principle behind Island is rooted in Mendelian inheritance, where each generation is produced based on the allelic frequencies from the previous generation. Users could examine the evolution of a population over multiple generations, observing how allele frequencies changed in response to genetic drift or selection. This simple, forward simulation approach helped to make abstract concepts like genetic drift more tangible by giving users the ability to track and visualize genetic changes in a population as it evolves over time.

The need for Island arose from my recognition that tools like DOSE (Digital Organism Simulation Environment) were not always appropriate for simulating genetic evolution at the level of alleles. DOSE operates at the level of the nucleotide or base, which works well for some genetic simulations. However, there were scenarios where simulating genetic changes at the level of alleles, the basic unit of Mendelian inheritance, would be more appropriate. Island filled this niche by providing a tool that could handle allele-based simulations in a simple, cost-effective way.

Reflection: Reflecting on the development of Island, it's clear that this project was motivated by my desire to make population genetics more approachable and engaging, especially for those without a strong background in mathematics or genetics. One of the key obstacles I saw in teaching and applying population genetics concepts was the gap between the abstract mathematical models and the hands-on, intuitive understanding of how genetic evolution occurs over time. Tools like Island were designed to lower this barrier, providing users with a practical and visual way to explore genetic drift, allele frequencies, and population structure.

As I worked on Island, I became increasingly aware of the limitations of other simulation tools, such as DOSE, when applied to certain problems in population genetics. DOSE is a powerful tool for simulating evolution at the nucleotide level, but when dealing with traits or alleles, its atomic unit of change was not sufficient to capture the relevant dynamics of Mendelian inheritance. By developing Island, I was able to address this gap and create a tool that was more aligned with classical genetics, where the unit of change is the allele, not the nucleotide. This shift in focus allowed Island to be more flexible and applicable in a wider range of scenarios, making it a useful addition to the toolkit of researchers and educators alike.

The simplicity of Island was also its strength. By removing the complexity of advanced mathematical models and focusing on the core principles of genetic inheritance, Island made population genetics more accessible. It allowed students, researchers, and even hobbyists to experiment with the concepts of genetic drift, selection, and population evolution without needing a deep understanding of the underlying math. In many ways, Island represents the power of simplicity in scientific tools—by stripping away unnecessary complexity, I was able to create something that was not only more usable but also more engaging for those seeking to understand genetic evolution.

Sidebar: The Power of Simplicity in Scientific Tools
In the world of scientific simulation tools, there is often a temptation to overcomplicate things in the name of precision and sophistication. However, in many cases, a simple, intuitive tool can be more effective at engaging users and helping them understand complex concepts. The development of Island is a perfect example of this principle. By focusing on the core concepts of population genetics and Mendelian inheritance, I created a tool that was accessible and easy to use, without sacrificing its ability to simulate important genetic processes like drift and selection. This approach demonstrates that scientific tools don't always need to be the most complex to be the most impactful; sometimes, simplicity is the key to clarity and understanding.

87: *De Novo* Origins of Promoters 2

Citation: Usman, S, Chua, JW, Ardhanari-Shanmugam, KD, Thong-Ek C, B, V, Shahrukh, K, Woo, JH, Kwek, BZN, Ling, MHT. 2019. *Pseudomonas balearica* DSM 6083T promoters can potentially originate from random sequences. MOJ Proteomics & Bioinformatics 8(2): 66–70.

Abstract: Recent studies and researches have proposed that many genes are plausibly emerged from previously non-coding genomic regions. However, how a promoter can emerge and function properly from *de novo* genes remain debatable as this has not been shown in large numbers of organisms. Therefore, this study aims to explore the possibility of *de novo* evolution of a promoter from random sequences by using *Pseudomonas balearica* DSM 6083T as the model organism. Our result shows that 39.3% of the generated random sequences have 68.6% probability to be a functional promoter. Evolution simulation was carried out to observe the effect of evolution in the putative *P. balearica* promoter over generations. The simulation result proves that selection enhances the functionality of the generated random sequences overtime. Therefore, it is plausible that *P. balearica* promoter could emerge from random sequences, which is consistent with findings from previous studies.

Context: The idea that genes may arise from previously non-coding sequences has become a powerful narrative in molecular evolution, but the question of how these newly formed genes become transcriptionally active remains underexplored. Central to this process is the emergence of functional promoters – genomic sequences that enable the initiation of transcription. In this study, we investigated whether *de novo* promoters could originate in *Pseudomonas balearica* DSM 6083T, a relatively understudied bacterium known for its metabolic diversity and resilience in extreme environments. By generating 10,000 random sequences and analyzing their alignment to known promoter motifs, we discovered that nearly 40% exhibited a moderate probability (~68.6%) of functioning as promoters. Furthermore, simulated evolutionary selection significantly increased their functional potential over generations. This work complements our earlier study in *Bacillus subtilis* (Chapter 84) and extends the hypothesis that promoter origination from random sequences is not an isolated phenomenon but may be a generalizable feature across diverse bacterial lineages.

Reflection: This paper was a continuation of my inquiry into the plausibility of evolutionary mechanisms giving rise to regulatory elements from scratch. The intellectual excitement came from watching randomness yield to order, even if only probabilistically. Collaborating with Sharlene Usman and the rest of the team added a layer of mentorship to the project. Sharlene, now a business owner in Batam, was then an aspiring researcher. The fact that she and her peers could independently carry forward a concept from Bacillus to Pseudomonas was a small but meaningful testament to the reproducibility of our approach and the maturity of the students involved. While the publication was modest in its reach, it contributed to a deeper understanding of genome evolvability and furthered my interest in digital evolution as a valid lens for studying molecular biology. This chapter is not just about promoter emergence; it's also a record of emergence of ideas, of students, and of possibilities.

Sidebar: *Pseudomonas balearica* – An Emerging Environmental Model

Originally isolated from marine environments, *Pseudomonas balearica* DSM 6083T is gaining recognition as a versatile environmental bacterium with notable tolerance to pollutants, heavy metals, and hydrocarbon-rich substrates. Its adaptability makes it a promising candidate for bioremediation studies. Unlike laboratory workhorses such as *E. coli* or *Bacillus subtilis, P. balearica* brings ecological authenticity to synthetic biology and evolutionary modeling. By testing *de novo* promoter origination in this organism, our study implicitly acknowledges the need to validate evolutionary hypotheses not just in idealized lab strains, but in real-world microbes facing real-world challenges. The species' under-characterization also made it a fertile ground for novelty – a microbial "blank slate" on which we could paint genomic possibilities.

88: SeqProperties

Citation: Ling, MHT. 2020. SeqProperties: A Python Command-Line Tool for Basic Sequence Analysis. Acta Scientific Microbiology 3(6): 103-106.

Abstract: A Python Command-Line Tool for Basic Sequence Analysis.

Context: In 2020, I published “SeqProperties: A Python Command-Line Tool for Basic Sequence Analysis” in Acta Scientific Microbiology. The primary motivation behind developing SeqProperties was to create a user-friendly, accessible tool for students and researchers working in fields like biotechnology and biomedical science, particularly those without a deep background in computational biology. During the honours projects of Jung Kwan and Argho, I observed that tools like BioPython, while powerful, were often too complex for students who were more familiar with laboratory work than computational programming. The complexity of these libraries posed a significant barrier to entry, especially for those new to sequence analysis.

SeqProperties was designed to address this gap by providing a simplified command-line tool that could perform essential sequence analysis tasks without the steep learning curve associated with more advanced libraries. As I worked closely with the students on their projects, I continuously added functionality to SeqProperties based on the specific needs that arose, making it a dynamic tool that could evolve with the requirements of the projects. This iterative approach ensured that the tool was both practical and tailored to the real-world needs of the students.

The focus of SeqProperties was on basic sequence analysis tasks, such as sequence alignment, searching, and basic statistical analysis. The simplicity of the tool allowed students to focus more on the biological implications of their results rather than getting bogged down by the complexities of software libraries. By enabling students to analyze biological sequences quickly and efficiently, SeqProperties became an essential part of their toolkit for conducting research and learning about bioinformatics.

Reflection: Reflecting on the development of SeqProperties, I realize how important it was to bridge the gap between computational biology and students who might not have a strong programming background. One of the most gratifying aspects of this project was seeing the immediate impact it had on the honours students I was working with. By providing them with a tool that was tailored to their needs, I was able to make their research projects more manageable and less intimidating. For students like Jung Kwan and Argho, who were primarily focused on their lab work, having an easy-to-use tool for sequence analysis meant they could dive into the biological questions without being overwhelmed by the technical complexities of bioinformatics.

The process of adding functions to SeqProperties based on student requirements was a learning experience for me as well. It made me more aware of the types of analyses that were most commonly needed in their research, and it also helped me appreciate the value of adaptability in software development. What started as a simple tool for basic sequence analysis evolved into a more comprehensive resource that could accommodate the diverse needs of students across various projects. The flexibility of SeqProperties is something I value deeply, as it demonstrates how a tool can evolve over time to better serve its users.

One of the key reasons I published SeqProperties was to make it easier to refer to and reference in academic and research projects. By making the tool publicly available and officially published, I ensured that it would be easier for others, especially students and researchers, to integrate it into their own work and cite it in their projects. This publication not only provided greater visibility for the tool but also allowed me to contribute to the field by offering a solution that could be used by others in a straightforward and accessible way.

Looking back, I am proud of the fact that SeqProperties was able to provide such a straightforward solution to an often complex problem. It also reinforced my belief in the importance of developing practical, accessible tools for researchers at all levels. By focusing on the core tasks and making the tool easy to use, I was able to help students focus more on their research questions and less on the intricacies of software programming. This experience reaffirmed my commitment to creating tools that empower users and simplify the complexities of scientific research.

Sidebar: The Challenge of Balancing Simplicity and Functionality

One of the key challenges in developing SeqProperties was finding the right balance between simplicity and functionality. While the tool had to be easy to use and accessible for students with limited computational experience, it also needed to be powerful enough to handle the common tasks in sequence analysis. This required careful consideration of the most essential features to include, as well as a focus on keeping the interface simple and intuitive. The process of continually adding functions based on the needs of the students highlighted how important it is to remain flexible and responsive when designing tools for research. In many ways, SeqProperties represents a perfect example of how a simple, focused tool can be just as effective as a more complex software suite, provided it meets the needs of its users.

89: Variability in Controls

Citation: Teng, RSY, Kwang, JCY, Chin, ASQ, Sander, CJ, Ang, IZL, Foong, JH, Cheong, KC, Hon, RYH, Ling, MHT. 2020. Correlation Analysis on Transcriptomes from Published Human Skin Studies Show Variations between Control Samples. EC Clinical and Medical Case Reports 3(6): 143-146.

Abstract: Reproducibility has been shown to be a problem in many areas of science, leading to a "reproducibility crisis". Many studies had examined factors limiting experimental reproducibility and one of the factors suggested is the stability of control samples underpinning all experimental findings. This study examines the transcriptomes of the control samples from three published human skin studies using correlation analysis to evaluate the stability of clinical control samples. Our results show significant differences (t-test p-value < 5.4E-5, Mann-Whitney U p-value < 0.00001) between within data set correlations and between data set correlations, suggesting significant differences between control samples from different data sets. This may have potential implications on the interpretation of clinically important results.

Context: In this project, conducted under the Differential Research Programme (DRP) at the School of Applied Science, Temasek Polytechnic, we addressed a subtle but potentially consequential issue in biomedical research – the assumed stability of control samples. Controls are the backbone of experimental design but rarely are they questioned with the same rigour as experimental treatments. In this study, I guided a group of DRP students through an investigation into transcriptome data from three publicly available human skin studies. Our question was simple: Are control samples as comparable and reproducible as we think they are across different studies?

Using basic correlation analysis techniques, the students compared expression profiles within and across datasets. The findings were both surprising and sobering – control samples from different studies showed statistically significant variation, suggesting that they may not be interchangeable or as "neutral" as commonly assumed. This has deep implications, especially in clinical research, where results often hinge on subtle expression differences between control and case samples.

Reflection: This DRP project is a testament to the power of asking foundational questions with curiosity and rigour, even in a short-term undergraduate research setting. The students were not working with novel algorithms or cutting-edge experimental data. Instead, they explored existing datasets with a critical eye – an approach that is often undervalued in science education. What they discovered was both real and important: control samples are not universally comparable, and this realization deserves more attention in experimental design, data interpretation, and meta-analyses.

For me, this study reinforced the need to teach students that science is not just about discovery, but also about questioning assumptions; even those that seem fundamental or settled. That control samples can differ significantly across datasets is a stark reminder that the “baseline” in an experiment is not always fixed. It can drift subtly depending on context, sample source, or even technical handling, which in turn can lead to vastly different conclusions if not carefully accounted for.

This experience also underscored how short-term research programmes like DRP can produce meaningful results, not only in terms of scientific content but in shaping how young researchers think. The paper, though short and focused, is a demonstration that critical thinking applied to publicly available data can yield insights with implications well beyond the scope of the classroom. As a supervisor, I found it deeply rewarding to see the students grapple with real-world variability and come away with a healthy sense of scientific skepticism.

Sidebar: The Control Sample Fallacy

In many research settings, controls are treated as interchangeable standards as if one control sample is as good as any other. This paper challenges that assumption. By demonstrating statistically significant differences between transcriptomic profiles of controls from different studies, it invites a rethink of how we interpret “normal” in biomedical data. Especially in clinical and genomic studies, where minute differences matter, this variability could spell the difference between a genuine discovery and a misleading artefact.

90: Chemistry to Biochemistry

Citation: Sim, BKY, Ling, MHT. 2020. Possibility of Abiotic Genesis of Biochemistry. EC Microbiology 16(6): 104-109.

Abstract: One of the first and primary life origin questions is how life can originate from primordial Earth chemistry. More than 60 years ago, Stanley Miller and Harold Urey conducted the famous Miller-Urey experiment where heated mixture of water, methane, ammonia, and hydrogen; representing early compounds on Earth; produces several amino acids when passed through an electrical discharge representing lightning. This gave rise to the possibility of abiotic genesis of biochemistry. Over the next six decades, evidence supporting various primordial macrobiomolecules emerged; leading to the concepts of RNA, DNA, and peptides being the first primordial macrobiomolecules. In this short review, possibility of each world originating separately, and coevolving were examined. Current evidence suggests that RNA world and peptide/amyloid world may originate independently and substantial possibility of interplay between these three worlds. Hence, RNA world, DNA world, and peptide/amyloid world may coevolve regardless of whether they originate independently. Thus, this calls for a reconciliation into a peptide-nucleic acid world.

Context: The question of how life began, how simple molecules on the primordial Earth gave rise to the complex biochemistry of living organisms, has fascinated scientists for decades. This short review was an intellectual collaboration between myself and Benjamin Sim, whom I met during my time as a research fellow at Nanyang Technological University (NTU). Benjamin, then an undergraduate and a valedictorian of his diploma cohort at Republic Polytechnic, showed a keen interest in origin-of-life studies. We began drafting this paper during our overlapping time at NTU, though it would take several years before the work matured into publishable form.

The paper revisits one of the most compelling questions in science: Can life originate abiotically from simple chemistry? Beginning from the foundational Miller-Urey experiment, we explored the evolving landscape of ideas that have given rise to competing and complementary hypotheses; namely, the RNA world, DNA world, and the peptide or amyloid world. Drawing from recent evidence, we proposed a reconciliatory model in which these biochemical domains might not have emerged in isolation but coevolved, interacting and shaping one another in the earliest stages of life's development. Our suggestion of a peptide–nucleic acid world attempts to bridge these once-competing paradigms into a unified conceptual framework.

Reflection: What made this work especially meaningful was not just the subject matter but the long arc of its creation. From casual conversations in NTU's research corridors to a collaborative draft that followed us through multiple phases of our careers, this paper is a testament to intellectual persistence across time and changing contexts. By the time it was published, Benjamin had transitioned into a career as a crime scene investigator, and eventually became a full-time commissioned police officer. Yet, the work we began as scientists endured.

This chapter marks one of the few times where I took a philosophical and integrative stance in the natural sciences. Instead of focusing on a specific molecule or pathway, we explored conceptual reconciliation – a theme that resonates with my broader scientific journey. Whether working with digital organisms, gene expression data, or educational simulations, my approach has always leaned toward building bridges; between ideas, disciplines, and people. This paper encapsulates that spirit.

It also stands as a quiet celebration of non-linear progress. The draft sat dormant for years, not for lack of interest, but because both of us were navigating different paths in life. And yet, when the time came, we dusted it off, refined it, and released it into the world. To me, this is a reminder that science can have a long gestation, and meaningful contributions sometimes come from revisiting past curiosities with a renewed perspective.

Sidebar: A Life of Parallel Tracks

Benjamin's journey from a bioscience student to a commissioned officer in the Singapore Police Force illustrates the multi-faceted nature of scientific curiosity. The fact that he could contribute to theoretical discussions about abiogenesis while preparing for a life in law enforcement speaks to the universal relevance of scientific thinking – its emphasis on evidence, logic, and model-building. The duality of his trajectory is not unusual but rather reflective of how diverse scientific interests can be carried across very different professional lives. This paper is both a scientific exploration and a marker of such interdisciplinary continuity.

91: *De Novo* Origins of Coding Sequences

Citation: Neo, CY, Ling, MHT. 2020. Prevalence and Length of Open Reading Frames Vary Across Randomly Generated Sequences of Different Nucleotide Compositions. EC Microbiology 16(7): 72-78.

Abstract: The emergence of open reading frames is an important step in the origination of *de novo* genes. However, the conditions leading to the origination of *de novo* genes is not well-understood. This study aims to determine the effect of nucleotide composition on the length and occurrence of ORFs by examining various ORF parameters using randomly generated sequences from 85 different nucleotide compositions. Our results suggest that various ORF parameters are significant across different nucleotide compositions (p-value < 1E-120). The average length, standard error of the average length, average maximum length, and standard error of the average maximum length of ORFs can be moderately predictable (0.43 < r^2 < 0.59) by nucleotide compositions. These results suggest that the prevalence and length of ORFs may be function of the underlying nucleotide composition.

Context: The formation of open reading frames (ORFs), continuous stretches of nucleotides that could encode proteins, is a critical precursor to the emergence of *de novo* genes. While *de novo* gene birth has gained increasing attention in molecular evolution, the precise conditions enabling such events remain elusive. This study arose from an exploratory project with Chuan Yang, an intern I supervised at Temasek Polytechnic. The work aimed to investigate whether nucleotide composition alone could influence the likelihood and properties of ORFs arising from random sequences.

We generated synthetic DNA sequences representing 85 different nucleotide compositions, simulating a wide spectrum of GC and AT biases. Using computational analysis, we examined key ORF metrics; average length, maximum length, and their respective variabilities; cross these compositions. Our findings were striking: nucleotide composition had a statistically significant influence on ORF characteristics, with moderate predictability in ORF parameters based on composition (r^2 values ranging from 0.43 to 0.59). These results hint that genomic "grammar" alone, independent of selection, may shape the raw potential for gene emergence.

Reflection: This paper sits at the intersection of randomness and biological possibility. It grew from a simple idea – what happens if we shuffle the genetic alphabet in different proportions? – into a data-rich investigation of sequence potential. The idea that the mere statistical properties of DNA can modulate the probability of ORF formation is both compelling and humbling. Evolution, it seems, doesn't just work on functional sequences; it works on a canvas of probabilistic constraints laid down by base composition.

Working with Chuan Yang was a joy. He approached this abstract question with a level of clarity and technical diligence that belied his age. It’s gratifying to know that he completed his undergraduate studies in 2024, likely carrying forward the same blend of curiosity and discipline that shaped this project.

In my broader scientific arc, this paper connects with several themes: the role of sequence-level forces in shaping higher-order biological outcomes, the emergence of novelty from stochastic systems, and the power of computational simulation as a means of theoretical exploration. It also reinforces a foundational idea that runs through my work, meaning can emerge from structure, even before function arrives.

Sidebar: Building Biology from Statistics

This project is a reminder that biology doesn’t begin with function, it begins with form. The nucleotide composition of a genome may seem like a background feature, but it lays the groundwork for what forms can feasibly emerge. ORFs, the precursors of genes, don’t appear at random; their statistical likelihood is shaped by GC content, codon usage tendencies, and reading frame constraints. In this way, the genome is both code and constraint – an encoding of biological potential shaped by the language it speaks. This study makes the case that randomness is not chaotic; it’s patterned and those patterns matter.

92: Predicting Growth from Culture Media

Citation: Cheong, KC, Hon, RYH, Sander, CJ, Ang, IZL, Foong, JH, Ling, MHT. 2020. A Simulation Study on the Effects of Media Composition on the Growth Rate of *Escherichia* coli MG1655 using iAF1260 Model. Acta Scientific Microbiology 3(8): 40-44.

Abstract: Media compositions are important determinants of growth rate and genome-scale models (GSMs) had been used for optimizing media for metabolite production and growth. Recently, iAF1260, a GSM based on *Escherichia* coli MG1655, was used to study the effects varying glucose concentration in media on growth rate and metabolic fluxes. In this study, the effects of other media components in the presence of varying glucose concentrations on the predicted growth rate of *E. coli* MG1655 were examined. Our results show that 10 media components (ammonium, calcium, chloride, copper, glucose, manganese, magnesium, molybdate, phosphate, and potassium) demonstrate substantial impact on the predicted growth rate of *E. coli* MG1655. Of which, 4 components (glucose, ammonium, magnesium, and phosphate) have the most impact. However, our results also demonstrate the limitations of iAF1260 as media components that had been shown to affect *E. coli* growth rate were not reflected by the model.

Context: The nutrient environment in which bacteria grow exerts a profound influence on their physiology, metabolism, and gene expression. This Differential Research Programme (DRP) project builds on an earlier initiative by my student, Ega, where we began dissecting Monod's growth equation, traditionally focused on carbon limitation, into its underlying metabolic dependencies. Rather than treat media composition as a black box, we asked: How do individual nutrients interact with each other in shaping bacterial growth potential?

To address this, we used the iAF1260 genome-scale model of *Escherichia coli* MG1655, a well-characterized metabolic reconstruction, and systematically varied glucose concentrations alongside ten other key media components. The simulation revealed that while multiple nutrients influence growth rate predictions, four components (glucose, ammonium, magnesium, and phosphate) consistently exert the greatest effect. Yet, the study also highlighted critical limitations in the model's predictive fidelity, as known empirical effects of certain nutrients were not reflected.

Reflection: This paper sits at the intersection of systems biology and biochemical intuition. On one level, it was a data-driven exploration, playing out nutrient permutations *in silico*. On another, it was a conceptual continuation of the idea that growth is not just about carbon, but a multi-nutrient orchestration involving cations, cofactors, and metabolic balance.

This project exemplifies a pedagogical loop I often aim for, where students are not merely learning tools, but are actively interrogating their assumptions. The fact that the iAF1260 model failed to capture known effects reminds us that computational models are not oracles, but reflections of curated knowledge and that missing knowledge can lead to blind spots.

There is an enduring satisfaction in turning Ega’s initial biochemical insight into a computational simulation study carried out by another cohort of students. It reinforces my long-held belief: mentorship isn’t about transmission, it’s about continuity. This paper extends that thread and shows that a single idea, nurtured over time, can bloom in multiple forms.

Sidebar: Beyond Carbon – Redefining Growth-Limiting Factors

The Monod equation has long dominated microbiological thinking: growth as a function of a single limiting nutrient. But real microbial ecosystems are far more complex. Magnesium stabilizes ribosomes, phosphate is essential for ATP and nucleic acids, ammonium provides nitrogen for biosynthesis. In this simulation study, *E. coli*'s growth rate was shown to hinge on a constellation of such metabolites. Importantly, models like iAF1260, though powerful, cannot replace empirical insight; especially when their assumptions under-represent essential cofactors. Future models must strive for biochemical completeness, not just computational elegance.

93: New Functions from Existing Coding Sequences

Citation: Gunalan, K, Wong, CQL, Neo, MPY, Ling, MHT. 2020. One Percent of *Escherichia* coli O157:H7 Peptides May Contain Putative Beta-Lactamase Activity. EC Microbiology 16(8): 73-79.

Abstract: Beta-lactamases are enzymes conferring resistance to beta-lactam antibiotics, which has become a global challenge. Studies had suggested that beta-lactamases are primitive enzymes that existed before the antibiotic era, leading to the question on potential sources and emergence of beta-lactamases. This study examines the possibility of putative beta-lactamases in *Escherichia* coli O157:H7 by sequence comparison to known extended-spectrum beta-lactamases (ESBLs) from *E. coli*. Our results suggest that 57 peptides out of 5021 (1.14%) *E. coli* O157:H7 peptides have 64.7% probability of beta-lactamase activity. Phylogenetic analysis clustered the top 10 (by sequence similarity score) of these 58 peptides within known ESBLs. This suggests that these peptides may contain putative beta-lactamases activity and potentially be a source of putative beta-lactamase.

Context: Antibiotic resistance has become one of the most pressing challenges in global health. At the core of resistance to beta-lactam antibiotics lies a molecular culprit, beta-lactamases, enzymes that dismantle the antibiotic before it can act. What's fascinating is the hypothesis that beta-lactamases predate the antibiotic era, possibly existing in microbial populations long before humans discovered penicillin. This paper arose from an honours project that extended Brenda Kwek's earlier work, which investigated the genomic landscape of *E. coli* for these elusive enzymes.

In this study, we focused on *Escherichia* coli O157:H7, a well-studied pathogenic strain, to assess the prevalence of peptides that might harbour beta-lactamase-like activity. Using sequence similarity analysis against known extended-spectrum beta-lactamases (ESBLs), we found that approximately 1% of its peptides could potentially possess such activity, supported by clustering in phylogenetic analyses. These findings hint at a cryptic reservoir of resistance potential, lying dormant within bacterial genomes.

Reflection: There's something sobering about the idea that antibiotic resistance is not just an outcome of clinical misuse, but an ancient evolutionary inheritance. This study reminded me that bacteria, far from being passive targets, have an evolutionary head start. We're merely discovering tools they've evolved for eons.

This project stood out because of its layered continuity. It began with Brenda, and then passed on to this group. Each student added a brick to the wall of understanding, working across different years but toward the same goal. That sense of scientific relay, where insights accumulate over time, mirrors how knowledge is truly built.

The result, while modest in scale, demonstrates the power of computational screening. With minimal resources and clever algorithmic comparison, we were able to spotlight 57 candidates that might someday become clinically relevant. This was not only an honours project; it was a quiet warning – resistance doesn’t just spread, it emerges.

Sidebar: Ancient Weapons in Modern Wars

Beta-lactamases are not a recent invention. Long before antibiotics entered human medicine, microbes were already evolving enzymes to protect themselves from each other’s biochemical weapons. This evolutionary arms race has left molecular fossils in modern genomes – peptides that look, fold, and potentially function like beta-lactamases. Our study hints that these dormant or low-activity proteins may act as evolutionary precursors, ready to adapt under selective pressure. If true, resistance doesn't just spread, it awakens. Understanding these precursors may be key to staying ahead of the next antibiotic resistance wave.

94: *Pseudomonas* Core Genome

Citation: Tan, XT, Ramesh, A, Wang, VCC, Kamarudin, NJ, Chew, SSM, Murthy, MV, Yablochkin, NV, Mathivanan, K, Ling, MHT. 2020. Core *Pseudomonas* Genome From 10 *Pseudomonas* Species. MOJ Proteomics & Bioinformatics 9(3): 68–71.

Abstract: Core genome of a set of organisms represents the set of homologous genes shared between the set of organisms with many applications. The *Pseudomonas* genus is highly diverse with both plant and animal pathogens. Hence, the core genome of *Pseudomonas* genus can be useful. Current studies presented contradictory results with the core genome of *Pseudomonas* genus marginally larger than that of *Pseudomonas aeruginosa*. In this study, we attempt to identify a core *Pseudomonas* genome from 10 publicly available annotated genomes by intersecting homologous coding sequences using BLAST. Our results suggest a 218-gene core genome, which is 3.46% of the coding sequences of *P. aeruginosa*. 136 of 218 genes were mapped to official gene symbols and were enriched in 8 clusters in Gene Ontology biological processes related to central metabolism.

Context: The *Pseudomonas* genus is a fascinating tapestry of ecological versatility and home to both benign soil dwellers and notorious pathogens like *P. aeruginosa*. One of the central questions in comparative genomics is what genes are universally conserved across a genus – a “core genome” that underlies shared biological identity.

This honours project took on the challenge of resolving contradictory findings in the literature regarding the *Pseudomonas* core genome. By analyzing ten annotated *Pseudomonas* genomes, we identified 218 homologous genes shared across species; just 3.46% of the total coding sequences in *P. aeruginosa*. Interestingly, 136 genes could be linked to Gene Ontology terms, forming eight enriched clusters, mostly tied to central metabolic functions.

Reflection: There’s a deep satisfaction in finding commonality in diversity, especially in a genus as functionally rich as *Pseudomonas*. Each genome is a narrative of ecological adaptation, but the core genome is the shared grammar beneath these stories.

What struck me was how a small number of students, using simple tools like BLAST and careful filtering, could build a robust comparative insight from public data. The project may seem technical on the surface, but it actually wrestled with a philosophical question in biology: what is essential?

The project also highlighted how accessible meaningful research has become. Ten genomes, freely available; standard annotation; a good pipeline and suddenly you’re holding the genomic common ground of an entire genus. It was a proud moment to see this group of honours students, from different backgrounds, collectively assemble a foundational map of *Pseudomonas* biology.

Sidebar: Why Core Genomes Matter

A core genome isn't just a list of conserved genes, it is the genomic fingerprint of identity. In microbial taxonomy, biotechnology, and even drug target discovery, knowing the core genome can provide stable reference points. For *Pseudomonas*, this means understanding what makes these bacteria *Pseudomonas* regardless of their niche. This project’s finding that the core comprises mostly genes for central metabolism, reinforces the idea that life holds onto what is most necessary, and sheds the rest as the environment allows. Core genomes are the genomic bones beneath evolutionary flesh.

95: Loss of Genetic Variability in Captivity

Citation: Kamarudin, NJ, Wang, VCC, Tan, XT, Ramesh, A, Chew, SSM, Murthy, MV, Yablochkin, NV, Mathivanan, K, Ling, MHT. 2020. A Simulation Study on the Effects of Founding Population Size and Number of Alleles Per Locus on the Observed Population Genetic Profile: Implications to Broodstock Management. EC Veterinary Science 5(8): 176-180.

Abstract: Loss of genetic variability in small population, known as founder effect, is commonly seen in aquaculture, where broodstocks are not routinely supplemented from the wild, leading to detrimental effects. Yet, the relationship between founding population size and observed population genetic profile is not clear. Here, the effects of founding population size and number of alleles per locus on the observed population genetic profile across multiple generations were examined using simulation. Our results suggest that the number of alleles per locus (p-value = 1.2E-102) and generation counts (p-value < 1E-240) are significant factors in genetic drift but not founding population size (p-value = 0.12). This suggests that genetic drift occurs regardless of population sizes, which may have implications in broodstock management to constantly minimize the impact of genetic drift regardless of broodstock population.

Context: In aquaculture, maintaining genetic diversity within a broodstock is vital to ensuring healthy, resilient populations. However, a common oversight is the belief that simply increasing the founding population size will safeguard against genetic drift. This project, an honours study conducted at MDIS, questioned that assumption directly.

Using a custom simulation framework, our first application of the Island software, we explored how founding population size, number of alleles per locus, and generation count affect the observable genetic profile of a population. Surprisingly, our findings revealed that founding population size had no statistically significant effect (p = 0.12). Instead, it was allelic diversity and the number of generations that overwhelmingly shaped the trajectory of genetic drift.

Reflection: This project disrupted one of the most comfortable myths in population management: that a large starting population insulates against diversity loss. The simulations showed that genetic erosion is a quiet, relentless force, independent of how many individuals you begin with, especially when new genetic material isn’t periodically introduced.

The project had both technical and philosophical significance. Technically, it marked the first operational use of Island, a simulation platform designed to model genetic systems flexibly across generations. Philosophically, it underscored a hard truth in biology: entropy doesn’t care about headcount. Without deliberate genetic replenishment, drift is inevitable, even in numbers.

For the students involved, many of whom had little prior exposure to genetics or simulation, it was eye-opening. They witnessed how models can make the invisible, like allele frequency shifts, suddenly observable. I was proud not just of the findings, but of how this group internalized the responsibility their results placed on aquaculture practices.

Sidebar: Managing Against the Drift

In conservation and aquaculture, genetic drift acts like a slow leak; subtle, silent, and dangerous over time. This study adds urgency to what should already be a standard practice: routine supplementation from wild populations. Maintaining genetic diversity isn’t about how many broodstock you have but how often you refresh their gene pool. The Island software used here made that clear, one simulated generation at a time.

96: How Many Genes Required for Universal Phylogeny?

Citation: Wang, VCC, Kamarudin, NJ, Tan, XT, Ramesh, A, Chew, SSM, Murthy, MV, Yablochkin, NV, Mathivanan, K, Ling, MHT. 2020. A Case Study using Mitochondrial Genomes of the Order Diprotodontia (Australasian Marsupials) Suggests that Single Ortholog is Not Sufficient for Phylogeny. EC Clinical and Medical Case Reports 3(9): 93-114.

Abstract: All organisms exist today descended from a common ancestor and phylogenetic tree is a common means to analyze such evolutionary histories. Currently, orthologs are routinely used to construct phylogenetic trees. However, the number of orthologs required to determine the evolutionary history of a set of organisms is not clear. In this case study, we compare the generated phylogenetic trees from one ortholog against that of the complete set of orthologs using 13 mitochondrial genes of the 24 species from the Order Diprotodontia. Using the phylogenetic tree generated from the complete set of orthologs as benchmark, our results suggest that using single ortholog may result in distinctly different phylogenies as compared to benchmark and the average number of branch points from multiple single orthologs is significantly different (paired t-statistic = 8.01, p-value = 3.27e-14) from benchmark. This suggests that phylogenetic analysis from single ortholog or multiple single orthologs is not likely to reflect actual evolutionary history and the complete set of orthologs is required.

Context: In evolutionary biology, reconstructing phylogenetic relationships is a delicate balance between computational tractability and biological truth. Often, due to convenience or data limitations, researchers rely on single orthologs to infer phylogenetic trees. This approach is particularly tempting in teaching or rapid analyses. However, during the COVID-19 lockdown, a group of honours students used this constraint as a springboard to ask a deeper question: Can a single ortholog truly represent the evolutionary history of a lineage?

Using 13 mitochondrial genes from 24 species in the order Diprotodontia, a diverse group of Australasian marsupials, they compared phylogenetic trees derived from individual orthologs with a consensus tree built from the complete ortholog set. The results were clear: single-gene trees routinely contradicted the benchmark tree, both in topology and branching patterns. The average divergence (paired t-statistic = 8.01, p = 3.27e-14) was not trivial, it was systematic.

Reflection: This project, developed entirely during lockdown, was a testament to resilience under restriction. The students couldn’t access labs or campus resources, so they dove deep into computational biology from their homes. That isolation, paradoxically, gave them space to reflect more critically on basic assumptions in phylogenetics.

What emerged was a clear cautionary tale. The use of a single ortholog is not just a computational shortcut; it is a distortion of evolutionary history. The differences weren’t cosmetic; they were structural. For an order as genetically diverse as Diprotodontia, relying on one gene to tell the full story is like reading a novel through a keyhole.

For me, this paper underscored the value of systems thinking in biology. Evolution does not happen one gene at a time, and so our analytical methods must evolve as well. This case study became a perfect blend of pedagogy, computational insight, and biological nuance.

Sidebar: Lockdown Science and the Rise of Home Bioinformatics

COVID-19 shuttered classrooms and laboratories, but it also democratized research in an unexpected way. Students who might never have explored bioinformatics suddenly found themselves analyzing genomes from their bedrooms. This project not only challenged a central assumption in phylogenetics but also illustrated a broader truth: with the right questions and mindset, meaningful science can emerge from anywhere, even during a global lockdown.

97: UniKin 1

Citation: Murthy, MV, Balan, D, Kamarudin, NJ, Wang, VCC, Tan, XT, Ramesh, A, Chew, SSM, Yablochkin, NV, Mathivanan, K, Ling, MHT. 2020. UniKin1: A Universal, Non-Species-Specific Whole Cell Kinetic Model. Acta Scientific Microbiology 3(10): 04-08.

Abstract: Mathematical models of metabolism can be a useful tool for metabolic engineering. Genome-scale models (GSMs) and kinetic models (KMs) are the two main types of models. GSMs provide steady-state fluxes while KMs provide time-course profile of metabolites, which has more advantage in identifying metabolic bottlenecks. However, KMs require greater degree of accuracy for parameters than GSMs resulting in fewer large-scale KMs than GSMs. Recently, large-scale KMs have been developed but are not based on standard enzymatic rate equations resulting in difficulty in interpreting results in terms of enzyme kinetics. Here, we construct a universal, non-species-specific KM of core metabolism, based on Michaelis-Menten Equation, from glucose to the 20 amino acids and 5 nucleotides based on reactions listed in Kyoto Encyclopaedia of Genes and Genomes (KEGG). Non-species specificity is achieved by using the same Michaelis-Menten constant (Km), turnover number (Vmax), and concentration for each metabolite and enzyme for each equation. This forms a base model for developing species-specific whole cell KMs. The resulting model consists of 566 reactions, 306 metabolites, and 310 enzymes, involving in 1284 metabolite productions, and 1249 metabolite usages. Sensitivity analysis shows that 85% of the metabolite concentration changes with the change of one enzyme kinetic parameter. This forms a base model for developing species-specific whole cell KMs.

Context: Genome-scale metabolic models (GSMs) have dominated systems biology for over a decade, offering comprehensive static snapshots of organismal metabolism. However, these models are inherently limited by their steady-state assumptions. To truly understand the dynamics of metabolism, how systems shift and respond over time, kinetic models (KMs) are essential. Unfortunately, the field has been slow to scale kinetic models beyond single pathways due to the sheer complexity of obtaining accurate kinetic parameters.

The UniKin1 model was conceived as a workaround to this bottleneck. Developed during the COVID-19 lockdown as an honours project, it constructs a universal kinetic model of core metabolism, agnostic to species, by enforcing uniform parameters across all reactions. Drawing from the Kyoto Encyclopedia of Genes and Genomes (KEGG), it maps reactions from glucose through to all 20 amino acids and five nucleotides, resulting in a comprehensive model of 566 reactions, 306 metabolites, and 310 enzymes.

This radical simplification; fixing values for Km, Vmax, and metabolite concentrations; wasn't meant to capture species-specific realities but to establish a baseline kinetic scaffold from which more tailored models could evolve.

Reflection: UniKin1 was my first serious foray into whole-cell kinetic modeling, and it marked a deliberate parallel to the established world of GSMs. While others had made impressive strides in large-scale metabolic maps, very few had attempted to encode time-resolved behavior across the entire metabolic network. We weren't trying to compete with GSMs, but to complement them with a dynamic layer.

What surprised me most was how instructive even a universal model could be. The sensitivity analysis revealed that 85% of metabolites were responsive to just one kinetic parameter change, highlighting potential bottlenecks and leverage points in metabolic control. It offered a systems-level intuition that static models could not.

For the students, it was a chance to learn both computational modeling and philosophical humility. In science, sometimes a model is useful not because it is precise, but because it gives you a place to stand.

Sidebar: The First Stone in the Temple of Kinetics
UniKin1 was never meant to be perfect. It was meant to be foundational, a template upon which better, species-specific models could be built. Like early genome browsers, which started with incomplete annotations, UniKin1 provided a framework for further elaboration. In that sense, it was the kinetic analog to the early days of systems biology: coarse,but catalytic.

The vision was clear – one day, we would have whole-cell kinetic models for many organisms, allowing us to simulate not just metabolism, but metabolism over time. UniKin1 was the first stone laid in that temple.

98: AdvanceSyn Toolkit

Citation: Ling, MHT. 2020. AdvanceSyn Toolkit: An Open-Source Suite for Model Development and Analysis in Biological Engineering. MOJ Proteomics & Bioinformatics 9(4):83–86.

Abstract: Modelling and simulations are useful means to screen potential experimental designs for metabolic engineering. Genome-scale models of metabolism (GSM) and kinetic models (KMs) are the two main approaches for modelling, which resulted in largely disjoint computational tools for GSMs and KMs. Existing tools for GSMs require knowledge of the underlying programming languages while the development and merger of two or more KMs is difficult. In this work, AdvanceSyn Toolkit is an open-sourced high-level command-line tool to develop KMs, and to analyse GSMs and KMs; licensed under the Apache License, Version 2.0, for academic and not-for-profit use. It elevates the need to know the underlying programming language for GSM analysis. AdvanceSyn Model (ASM) specification is a simple and modular format for model development and AdvanceSyn Toolkit provides a method to merge two or more model files for simulation and sensitivity analysis.

Context: Metabolic engineering relies heavily on modelling to explore experimental space before benchwork begins. Yet the tools for genome-scale models (GSMs) and kinetic models (KMs) have long been fragmented, both conceptually and technically. GSMs typically demand programming proficiency in MATLAB or Python, while KMs often lack modularity or the ability to scale, making it difficult to merge or extend them.

In response, I developed the AdvanceSyn Toolkit – a high-level, command-line suite that unified GSM and KM development into a single workflow. It was the first official release from AdvanceSyn Private Limited, a company I co-founded in 2014, and this release in 2020 marked a philosophical milestone: the decision to make the software open-source, under the Apache License Version 2.0, for academic and non-profit use.

The toolkit introduced AdvanceSyn Model (ASM) format, a modular, human-readable specification for building models without requiring programming knowledge. Most importantly, it provided tools to merge multiple KMs, perform simulations, and sensitivity analysis, bridging a critical gap in the modeling ecosystem.

Reflection: AdvanceSyn Toolkit was not just a technical output, it was a manifestation of my values. I believed that accessibility should not be a barrier to biological discovery, and I wanted to remove the gatekeeping inherent in most modeling tools. Many scientists, especially in developing regions or at teaching institutions, had neither the time nor resources to master multiple programming languages just to run simulations. AdvanceSyn Toolkit allowed them to participate fully in metabolic design without that burden.

The decision to open-source the toolkit, despite being a product of a private company, was a deliberate act of scientific generosity. I had always envisioned AdvanceSyn Private Limited not as a profit-maximizing venture, but as a conduit for creating and disseminating ethical tools. This was my first chance to walk that talk.

Creating the ASM format was itself an act of abstraction – a distillation of years of modeling practice into something intuitive and modular. In doing so, I hoped to lay a foundation that others could easily build upon, just as I had once relied on others' open frameworks.

Sidebar: An Engineer's Declaration of Independence

AdvanceSyn Toolkit was my technical declaration of independence. It represented a turn from merely consuming models and tools to producing them. More than that, it reified my belief that scientific tools should be public goods, not private advantages.

By releasing the toolkit and its specification openly, I was inviting collaboration, replication, and critique. It was a statement that power in science should be distributed, not concentrated. That modelling, as a form of thought, belongs to everyone, not just the coders.

Sidebar: From Academic Scientist to Startup Co-Founder

The transition from academic scientist to startup co-founder is often portrayed as a leap from theory to practice, from ivory tower to marketplace. But for me, co-founding AdvanceSyn Private Limited was not a departure from science; it was an extension of it. I didn't pivot away from research, I operationalized its values.

In academia, I was driven by curiosity and the desire to democratize access to knowledge. As a co-founder, those same motivations shaped the company. AdvanceSyn wasn't conceived to chase venture capital or dominate a market. It was born from the frustration I saw in the lab: brilliant minds hamstrung by inaccessible, over-engineered tools. I wanted to create something better, something built on scientific empathy.

Startups and academic labs may seem worlds apart, but both demand intellectual honesty, resilience in the face of failure, and the ability to solve problems with limited resources. What changes is the interface: instead of publishing a paper, you publish a toolkit. Instead of being cited, you're downloaded. Instead of peer review, you get user feedback – immediate, sometimes harsh, but always instructive.

AdvanceSyn Toolkit’s release was a turning point. It transformed years of research into a practical, sharable asset. It validated that software rooted in academic rigor could also be productized, not for profit’s sake, but to propagate an idea more widely and sustainably.

For scientists considering the same path, know this: starting a company doesn’t mean you stop being a scientist. It means you learn to translate your values into code, into policy, into practice; and in doing so, bring your science closer to the people it was always meant to serve.

99: BactClass

Citation: Liu, TT, Ling, MHT. 2020. BactClass: Simplifying the Use of Machine Learning in Biology and Medicine. Acta Scientific Medical Sciences 4(11): 43-47.

Abstract: Machine learning has many applications in biology and medicine. However, most existing tools require substantial programming skills, which can be a challenge to many biologists. Here, we present BactClass as a command-line tool for machine learning algorithms on formatted data, aiming to reduce the challenges faced by biologists who are interested to use machine learning approaches. BactClass is part of the Bactome project (https://github.com/mauriceling/bactome) and is licensed under GNU General Public Licence version 3 for academic and non-commercial purposes only.

Context: Machine learning has rapidly permeated the life sciences, offering powerful tools for classification, prediction, and feature selection. Yet its practical uptake by many biologists and clinicians remains hindered by a steep learning curve, particularly around coding proficiency. Most ML frameworks are embedded in programming environments like Python or R, which creates a barrier for those trained outside of computational disciplines.

BactClass was developed precisely to bridge that gap. As a command-line tool under the Bactome project, BactClass allowed users to apply a suite of machine learning algorithms to well-formatted datasets without writing any code. Released under the GNU General Public License v3, it was positioned as an academic and non-commercial resource, designed to democratize access to ML techniques in biology and medicine.

The project also had a deeply human origin. Tiantong Liu, a young Chinese student seeking guidance on postgraduate options, reached out to me remotely; much like Amani and Jitesh had before him. Although we never met in person, we collaborated on BactClass to give him a meaningful technical experience. He later chose to pursue graduate studies at the University of Florida.

Reflection: BactClass was not my most technically ambitious project but it was one of my most intentionally inclusive. I have long held the conviction that science must serve the many, not just the technologically fluent few. BactClass arose from that philosophy: to enable those who were willing to learn, but lacked the coding background, to use machine learning meaningfully.

Working with Tiantong reminded me of the unique magic of global scientific mentorship. We were from different continents, never shared a physical workspace, and had no formal affiliation, just a shared curiosity and a willingness to build. In a way, BactClass is a digital artifact of trust, forged across borders during a time when physical travel was impossible.

By including BactClass in the larger Bactome project, I was also reinforcing my belief in ecosystemic thinking. Tools are more useful when they are part of a larger, interoperable framework. BactClass wasn’t just a standalone utility, it was a building block in an architecture that encouraged iterative learning and contribution.

Sidebar: Mentorship Without Borders

Tiantong was one of several students I mentored whom I never met in person. These relationships, forged through a simple message or email, reaffirmed my belief that mentorship can transcend geography. The digital age has redefined the boundaries of supervision and collaboration.

BactClass was a way to give Tiantong agency, a chance to shape something real before his formal studies even began. In helping him build, I was also building something larger: a culture of inclusion and empowerment in scientific computing.

100: Accumulated Potential

Citation: Chew, SSM, Murthy, MV, Kamarudin, NJ, Wang, VCC, Tan, XT, Ramesh, A, Yablochkin, NV, Mathivanan, K, Ling, MHT. 2020. Rapid Genetic Diversity with Variability between Replicated Digital Organism Simulations and its Implications on Cambrian Explosion. EC Clinical and Medical Case Reports 3(11): 64-68.

Abstract: Cambrian Explosion resulted in substantial increase in biodiversity, which may be attributed to both environmental and biological factors. Although increased genetic evolution rate had been shown during this period, the role of genetic evolution in increased biodiversity is unclear. Re-creating Cambrian Explosion experimentally is not feasible. In this study, we used digital organisms (DOs) at high rate of random point mutations in the absence of selective pressure to examine the extent genetic evolution possible during Cambrian Explosion. Our simulation results suggest rapid and significant genetic divergence in the absence of selective pressure can occur at a species level and at local population level with significant differences between each local population ($F \geq 15.97$, p-value $\leq$ 1.4E-79). Hence, the emergence of biodiversity in Cambrian Explosion may be due to the release of accumulated adaptive potential.

Context: The Cambrian Explosion remains one of the most intriguing evolutionary phenomena – a geologically brief period when life underwent a dramatic diversification. While many hypotheses attempt to explain this burst of biodiversity, experimentally re-creating such an event is unfeasible due to both ethical and biological constraints. That is where digital organisms (DOs) offer a powerful alternative.

This project used high-throughput simulations of DOs; simple, evolvable digital entities; to model the kind of rapid, mutation-driven divergence that may have occurred during the Cambrian period. Notably, the simulations excluded selective pressures, allowing for the unfettered expression of underlying genetic variation. The outcome was profound: significant divergence emerged not just between populations, but within them; suggesting that the Cambrian biodiversity might have been a release of latent adaptive potential rather than purely an outcome of selection.

This work was conducted entirely under COVID lockdown conditions, making it not just a scientific exploration but a logistical feat as well.

Reflection: There is a poetic symmetry in using digital organisms to understand the origin of biological complexity. What struck me most about this study was not just the biological plausibility of the results, but how much evolution could unfold in the absence of a guiding hand. The emergence of diversity from noise alone pointed to the creative potential embedded in randomness, a concept that aligns well with both evolutionary theory and systems thinking.

This was also a deeply personal project. The world was in chaos during the pandemic, and yet, in a small corner of digital space, we were replicating the birth of life’s complexity. There was a metaphor in that too – finding emergence, novelty, and meaning in constraint.

That we achieved statistically significant divergence across independent simulations ($F \geq 15.97$, $p \leq 1.4E-79$) underscored a key message: evolutionary outcomes are not deterministic but bounded by statistical structure. The Cambrian Explosion may have been less about radical new inputs and more about conditions allowing expression of what was already possible.

Sidebar: The Ethics of the Digital Lab

One critical advantage of this study was ethical. Recreating high mutation rates in biological organisms, especially under no selective pressure, is neither feasible nor responsible in a wet lab setting. But with digital organisms, we bypass that constraint.

This study reinforced my belief that digital evolution platforms are not toys – they are ethical, powerful, and essential alternatives for studying biological phenomena that would be impossible or unethical to explore *in vivo*. The Cambrian Explosion, one of life’s greatest mysteries, was thus reframed not through paleontology, but through lines of code.

101: Singapore COVID-19 Stories

Citation: Wang, VCC, Ling, MHT. 2020. Science/Education Portraits VI: Anecdotes of Life in Singapore During COVID-19 (February 2020 to September 2020). EC Clinical and Medical Case Reports 3(11): 98-111.

Abstract: No doubt that the world in 2020 looks very different from that of 2019, due to COVID-19 pandemic, which is the single most impactful event since World War II. This triggered a call to document this event for future ethnographical studies. Many authors from various countries had responded to this call. In this article, we respond with collection of 25 ad-hoc anecdotal portraits of life in Singapore between February to September 2020, to add to the collective documentation for this momentous time in this century. Several of our experiences echo that of other reports from other countries. We iterate the need to document this period. However, we also see anecdotal evidence of benefits from this crisis.

Context: The year 2020 was a global rupture. The COVID-19 pandemic altered lives in ways that were intimate, structural, and deeply human. As much as it was a public health crisis, it was also an existential event – prompting reflections on fragility, interconnectedness, and resilience. Recognizing the need to document the lived experience of this time, national libraries and international scholars issued a call for memory, before it was overwritten by the sheer force of time.

This paper, part of the Science/Education Portraits series, answered that call. Co-written with Victor Wang, a former honours student, we curated 25 anecdotal portraits drawn from daily life in Singapore between February and September 2020. These vignettes captured the microcosms of disruption: from queuing for masks to coping with isolation, from panic buying to the silence of once-busy streets. Though anecdotal, these stories mirrored global patterns while remaining uniquely local.

This project also marked my first deliberate step into ethnographic work – an attempt to bear witness rather than explain.

Reflection: At the time, writing this chapter felt more like a civic duty than a scientific contribution. There was no dataset to analyze, no model to validate; only human experience to hold and preserve. It reminded me that science does not always proceed in equations or hypotheses; sometimes, it whispers through stories.

Victor's involvement was particularly meaningful. Having previously worked with him on mitochondrial phylogenies, it was moving to shift from genetics to generational memory. Our portraits were honest, raw, and sometimes mundane but therein lay their power. They offered a snapshot of ordinary people navigating extraordinary times.

This chapter also shifted something within me. It taught me that my role as a scientist could include that of archivist and storyteller, not just theorist and builder. Ethnography, so often seen as outside the domain of computational science, revealed itself as deeply complementary, especially when the goal is to understand the world in its full complexity.

Sidebar: The Science of Remembering

In times of crisis, scientific memory tends to focus on data – case counts, transmission curves, mortality rates. But societal memory is built on anecdotes, rituals, and the small details of daily life. This project was a humble act of scientific remembrance – rooted not in metrics but in meaning.

We were inspired by Singapore’s National Library Board's COVID-19 oral history initiative, but our contribution stood apart in its informality and immediacy. It reminded me that science, too, needs witnesses; those willing to record not just what happened, but what it felt like to live through history.

102: PubMed and Google Scholar

Citation: Teo, YH, Ling, MHT. 2020. A Systematic Review on the Sufficiency of PubMed and Google Scholar for Biosciences. Acta Scientific Medical Sciences 4(12): 03-08.

Abstract: NCBI PubMed is the de facto bibliographic database for biosciences but has been shown to be insufficient for the purpose of systematic review and meta-analysis, which requires comprehensiveness. Among bibliographic databases, Google Scholar is most comprehensive. With arguments that PubMed, supplemented with Google Scholar, may be sufficient for a systematic review in biosciences; we reviewed 18 studies to determine whether a combination of PubMed and Google Scholar is sufficient. Current literature shows that the combined coverage of Google Scholar and PubMed is between 85% to 98% of the universe of bioscience articles, which may be sufficient. However, Google Scholar alone is not sufficient as the concordance between PubMed and Google Scholar is 30.3% with 20.3% of the articles unique to PubMed.

Context: In the domain of biosciences, literature review is foundational; yet, the tools we rely on for these reviews are often assumed rather than examined. PubMed has long held its place as the gold standard for biomedical indexing, maintained by the U.S. National Library of Medicine. However, its sufficiency for systematic reviews, which demand exhaustive coverage, has been repeatedly questioned.

Google Scholar, by contrast, is an omnivorous search engine indexing a broader and less curated collection of academic material. Some researchers argue that the combination of both might offer a pragmatic balance between precision and comprehensiveness. In this paper, we examined 18 prior studies to evaluate whether PubMed and Google Scholar, used in tandem, are sufficient for systematic reviews in the biosciences.

Our findings suggest that combined coverage ranges from 85% to 98%, a level that might be considered sufficient for most systematic reviews. However, neither platform alone is adequate: Google Scholar's low concordance with PubMed (30.3%) highlights the importance of PubMed's curated content.

Reflection: This work stands out not just for its content but for its context of creation. It was born during the early disruptions of COVID-19, when my former student from Temasek Polytechnic, Yun Hwee, had her internship delayed by three weeks. Rather than let the time slip by, I invited her to work on a short research project. Over WhatsApp and Microsoft Teams, without a single face-to-face meeting, we turned a logistical delay into a published paper.

The speed and clarity of our collaboration were invigorating. It reminded me that scholarship is as much about responsiveness as it is about rigour. This wasn’t a grand theoretical contribution, but it answered a very practical question in a moment when the world itself was being reorganized. And for Yun Hwee, this experience offered a fast-tracked immersion into real academic publishing under conditions that were anything but ideal.

This chapter reaffirmed something I had long believed: research doesn’t need perfect conditions, only intentional momentum.

Sidebar: Scholarship in Crisis

This project is an example of adaptive scholarship – born from disruption, not despite it. The COVID-19 pandemic created countless interruptions in academic life, especially for students whose internships or lab work were stalled. Rather than viewing those disruptions as wasted time, I saw them as opportunities for agile mentorship.

In many ways, the constraints sharpened our focus. No lab access, no meetings, no travel; just a clear question, accessible data, and a commitment to get something done. It proved that even during a pandemic, academic possibility can flourish in the margins.

103: Systematic Quirks in Machine Learning Classifiers

Citation: Ling, MHT. 2020. Low Classification Accuracy by Logistic Regression, Support Vector Classifier, and Multi-Layer Perceptron, but Not Decision Tree, on Random Attributes from Hadamard Matrix. EC Clinical and Medical Case Reports 3(12): 07-10.

Abstract: The use of machine learning classifiers is increasing with evidence of overtaking human judgement. This can be risky if workings and implications of machine learning classifiers remain a black box. Here, a case where a balanced and algorithmically generated data set, Hadamard matrix, classifies poorer than random using logistic regression (accuracy < 17.4%), support vector classifier (accuracy < 23.4%) and in most cases of multi-layer perceptron (accuracy < 27.9%) but not in decision tree (accuracy > 77.3%); despite perfect (100%) internal classification accuracy for both support vector classifier and multi-layer perceptron; is reported. This suggests a systematic and yet currently unexplained source of error.

Context: Machine learning, for all its promise, often remains a black box; particularly in biological and clinical domains, where interpretability can be more important than raw performance. In this paper, I wanted to explore not how well machine learning works, but how and when it breaks. The Hadamard matrix, with its orthogonal rows and perfect balance between 1s and -1s, provides a unique test bed – algorithmically generated, noise-free, and class-balanced.

I subjected these matrices to a suite of standard classifiers: logistic regression, support vector classifiers (SVCs), multi-layer perceptrons (MLPs), and decision trees. The results were jarring. While SVCs and MLPs boasted perfect internal accuracy during training, their external classification performance dropped below random; suggesting that they were not just overfitting, but doing so in a systematically misleading way. Only decision trees performed robustly and consistently, achieving classification accuracies above 77%.

The implication is sobering. Even in ideal, noise-free data scenarios, many popular machine learning algorithms can fail spectacularly, and in ways that go unnoticed if we rely solely on training metrics.

Reflection: This was a short, almost casual exploration but it hit a deeper nerve. For all the talk about the power of AI and machine learning, their adoption in medicine and biology has outpaced our understanding of their vulnerabilities. The Hadamard matrix, as mathematical as it is, became a metaphor for misplaced confidence in models that look good on paper but stumble in reality.

The most surprising result wasn’t the poor performance itself, it was the disconnect between perfect internal accuracy and catastrophic external results. This paper, modest in scope, reminded me how deceptively easy it is to believe in models just because they produce good numbers.

In hindsight, this work quietly re-anchored my approach to machine learning. It reinforced my preference for models like decision trees not because they’re always more accurate, but because they are more transparent, and in certain contexts, reliability is more useful than complexity.

Sidebar: Simplicity vs. Sophistication

Machine learning’s appeal often lies in its sophistication – deep layers, complex kernels, ensemble methods. But this study reminds us that sophistication doesn’t guarantee stability. Sometimes, the simplest tools, like decision trees, are the most trustworthy, especially when interpretability and reproducibility matter.

The Hadamard matrix isn't biological or medical in nature, but what it reveals is universally applicable: always test beyond the training data, and never assume that elegance in mathematics equals robustness in application.

Sidebar: The Risk of Blind Trust in AI/ML Outputs

One of the most dangerous habits in data science is assuming that a model’s high performance during training guarantees real-world success. This paper exposed that fallacy in stark terms. It showed that even on perfectly balanced, noise-free data, some of the most popular machine learning algorithms; logistic regression, support vector classifiers, and neural networks; can fail catastrophically outside their training scope, often without warning.

What makes this especially concerning is that these failures don’t announce themselves. On paper, everything looks fine. The training accuracy is perfect. The model converges. The metrics suggest success. But beneath that sheen lies a fundamental disconnect between what the model has learned and what it actually understands. If that distinction sounds subtle, its consequences are not; especially in high-stakes domains like medicine, finance, or criminal justice, where decisions affect lives.

Blind trust in AI and ML is not just a technical oversight; it's a cognitive trap. The more impressive the model, the more likely we are to defer to its outputs without questioning its behavior. This study, modest in its dataset and scope, offered a quiet but crucial warning: sophistication is not a substitute for scrutiny. In the end, transparency may prove to be the most vital metric of all.

Returns on Grants

Warning: This chapter may offend some readers; hence, you have been warned. Do not proceed if you think you can be, may be will be offended.

In the citation for Barbara McClintock's Nobel Prize, Professor Ringertz of Karolinska Institute said the following to Dr. McClintock, “your work is encouraging because it shows that great discoveries can still be made with simple tools.” I am a proponent of good stewardship of R&D grants as a large part of it are public funds. As an academic, I will want my students to leave me with a strong concept that R&D dollars is a privilege. However, there is no real measure and benchmark of R&D output. Hence, I am adapting the concept of “Returns on Equity” and “Returns on Investment” to measure my own efficiency in using grants, naming it “Returns on Grants” or ROG.

Grants comprises of the total amount of money I am given for research and development activities. ROG will then be the quotient of grants less allowable deductions (accessible grant) and the number of manuscripts produced.

The allowable deductions are:

- Manpower. Each peer-reviewed manuscript can only be used for 1 deduction of $25000 as manpower cost provided that the peer-reviewed manuscript is accepted within 2 years of the funding period.
- Equipment. 90% of equipment shared by different R&D groups (common equipment) and 80% of other equipment can be deducted.
- Patents. 30% of cost, up to 5000, incurred in the filing process of each patent may be deducted. The rest should be offset by revenue derived from the patent.
- Publication cost. Any cost related to publications can be deducted.
- Bringing forward of unused deductions. 50% of unused deductions can be brought forward to the following years. This is in concordance to how company losses can be brought forward to negate profits in subsequent years for tax computation.

Number of manuscripts equivalents are computed as follow: Each peer-reviewed manuscript is considered as 1 manuscript. Each book (monograph or collection) is 0.6 manuscript. Each book chapter (both peer-reviewed and un-reviewed) is 0.3 manuscript. Each technical report, R&D-based un-reviewed manuscript, abstract, poster or talk is 0.1 manuscript. Hence, my Return on Grants is calculated to be $887 per manuscript equivalent. Details are as follows:

Grants (calculated in Singapore Dollars)	**Grant**	**Deductions**	**Accessible Grant**	**Cumulative Accessible Grant**
2004-2008: Own Ph.D grant ($15000 for research, $100000 for stipend)	$115000	$100000 (3 manuscripts in 2007, 2008, 2009, 2010)	$15000	$15000
2008-2009: Internship for CyNote 1	$3200	$25000 (1 manuscript in 2010)	($10900)	$4100
2009: MOE SMP (1.5 projects)	$750		$750	$4850
2009: 1 DBT Final Year Project	$4660	$840 (cost of micropipetters = $1050); $25000 (1 manuscript in 2010)	($10590)	($5740)
2009: 1 SNHP Project	$350		$350	($5390)
2010: MOE SMP (1 project)	$500		$500	($4890)
2010: 2 DBT Final Year Projects	$8000	$25000 (1 manuscript in 2012) Note: $3500 allocated for general use.	($10250)	($15100)
2010: 2 SNHP Projects	$700		$700	($14400)
2011: MOE SMP (1 project)	$500		$500	($13900)
2011: EBD Grant (paying half of my salary in Life Technologies)	$37700	$25000 (1 manuscript in 2011])	$12700	($1200)
2012: NIH Grant (87.5% of salary)	$45500	$25000 (1 manuscript in 2012)	$20500	$19300
2013: Institutional Grant (SCBE, NTU) (100% of salary)	$61100	$25000 (1 manuscript in 2013)	$36100	$55400
2014: MOE Grant (100% of salary)	$61100	$25000 (1 manuscript in 2014)	$36100	$108205
2015: NEWRI Grant (100% of salary)	$55900	$25000 (1 manuscript in 2015)	$30900	$86300
2015: iJAM Grant (to AdvanceSyn Pte Ltd;	$10000	$25000 (1 manuscript in 2016)	($7500)	$78800

20% ownership of $50000 grant)				
2016: NEWRI Grant (100% of salary)	$55900	$25000 (1 manuscript in 2016)	$30900	$109700
2016: SPRING Proof-of-Concept Grant (to AdvanceSyn Pte Ltd; 10 of 12 months; 20% ownership of $208300 grant)	$41600	$25000 (1 manuscript in 2016)	$16600	$126300
2017: SPRING Proof-of-Concept Grant (to AdvanceSyn Pte Ltd; 02 of 12 months; 20% ownership of $41600 grant)	$8320	$25000 (1 manuscript in 2016)	($8340)	$117960
2017: NEWRI Grant (100% of salary)	$4800	$25000 (1 manuscript in 2017)	($10100)	$107860
2018: 11 Honours Year Projects, Northumbria University (UK)	$9900	$25000 (1 manuscript in 2019)	($7550)	$100310
2019: 13 Honours Year Projects, Northumbria University (UK)	$11700	$25000 (1 manuscript in 2018)	($6650)	$93660
2020: 21 Honours Year Projects, Northumbria University (UK)	$18900	$25000 (1 manuscript in 2020)	($3050)	$90610

Number of manuscripts equivalents = **102.2** comprising of

Type of Publication	**Count**	**Manuscript Equivalents**
Peer-reviewed manuscripts	91	91
Books (monographs and collections)	6	3.6
Book chapters (peer-reviewed and un-reviewed)	11	3.3
Technical reports, Un-reviewed manuscripts, Abstracts, Posters and Talks	43	4.3
Total:	151	102.2

My gross grant amount of $556080 (before any of my listed deductions) divided by 102.2 manuscript equivalents still put me at $5442 per manuscript equivalent.

Overall Reflection: An Evolution of a Journey

Looking back on the 103 chapters of this book, I find myself struck by the recurring themes of transformation, adaptation, and the subtle power of incremental change. The work is a mosaic of intellectual evolution, a trajectory that has spanned decades, from the early days of grappling with the intricacies of artificial life to the quieter, more contemplative moments where I reflect on the broader significance of science and its role in shaping human experience.

At its core, this journey has been one of **discovery** – not only in terms of technical achievements but also in the discovery of the kind of scientist, teacher, and individual I have become. These chapters chronicle the unfolding of my intellectual and personal life, often in parallel, as I navigated the challenges of academia, scientific inquiry, and the search for meaning within a rapidly changing world. They reflect my growth from a knowledge consumer to an enabler, from a scientist following in others' footsteps to one who sought to chart new paths for others to follow.

A recurring theme that emerges throughout these chapters is the importance of **infrastructure** – the unseen, often unnoticed work that supports the larger scientific enterprise. From the development of bioinformatics tools like MapMan for tobacco gene expression to the creation of Ragaraja, the esoteric genome interpreter, much of my work has revolved around enabling others to explore, innovate, and build upon the foundation I helped lay. In many ways, this resonates with the quiet ethos of **enabling progress**, whether through writing code or designing processes that make complex biological phenomena more accessible.

Underlying this is a **philosophical tension** between the quest for novelty and the value of pragmatic, often invisible contributions. Early on, I sought novelty; new ideas, new methods, groundbreaking discoveries. But with time, I began to appreciate the quiet, foundational work that makes it possible for others to create and discover. It’s easy to chase the loudness of novelty, but the true value often lies in the unsung moments of creation; the scripts, the tools, the frameworks that allow the next generation of thinkers to build on what came before.

Another recurring theme is **adaptation** – both in the scientific sense and the personal one. Through projects like the halophilization of *E. coli* and my engagement with synthetic biology, I explored how systems; biological, computational, and personal; can adapt to constraints, evolve under pressure, and thrive in unexpected environments. These chapters reflect my own process of adaptation within the academic world. I grappled with questions of identity, purpose, and impact, and I learned that growth often comes not in leaps and bounds, but through steady, deliberate changes. Just as *E. coli* adapted to 11% NaCl, I too learned to navigate the challenges of academia, personal relationships, and evolving scientific landscapes, one incremental step at a time.

In parallel, these chapters also delve into the complexities of **legacy** – what we leave behind as we move forward in our lives. While much of the work presented here is inherently technical, it's the reflections—on the people, the questions, and the moments of intellectual curiosity that truly define my legacy. I've come to see that impact doesn't always need to be loud or widely recognized. It can be quiet, humbling, and deeply personal. As much as my contributions to bioinformatics and artificial life are important, the more enduring legacy may be in the act of encouraging others to ask better questions, to explore deeper, and to see the beauty in the simplicity of the process.

The chapters also reflect a constant **reframing of self**, a continuous journey of reflection, questioning, and understanding my place in the broader scientific and personal context. I have learned to embrace the idea that, as much as we seek certainty in science, there is often beauty in uncertainty and complexity. This reflects the philosophical journey I have undertaken, moving from the certainty of structured academia to the more fluid and nuanced path of self-reflection and personal growth.

Ultimately, this collection is a celebration of **curiosity**, the relentless pursuit of understanding, both of the world around us and ourselves. It celebrates the process of creation, of building something meaningful, and of allowing oneself to evolve in the face of challenges. It is a record of how science and life are inextricably linked, each shaping and informing the other.

In reflecting on this journey, I realize that the chapters I've written are not merely about the accomplishments, the experiments, or the published papers. They are about the quiet, profound shifts that occur when we look back and ask ourselves: What was the purpose? What did I contribute? And most importantly, how did it shape who I've become?

As I close the book on this chapter of my life, I'm reminded of the wisdom of Ragaraja – one of transformation and purification. In the same way that my digital organisms evolve, so too have I transformed. I have come to see that the value of life's work is not only in the discoveries made but in the way we learn to adapt, to evolve, and to contribute meaningfully to the world around us, one small step at a time.

Postface: Closing the Circle

As I reach the conclusion of this journey, I find myself reflecting on the themes that have defined both this book and my life's work. The narratives woven through these 103 chapters are not merely technical achievements or personal milestones; they are a collection of moments, each one representing a shift in understanding; both of the world of science and of myself. And though this book marks an endpoint, in many ways, it feels more like a reflection at a turning point, the closing of one chapter to make way for the next.

Throughout this work, I have grappled with the challenges of translating biological principles into computational models and the philosophical implications of artificial life. In doing so, I have learned that the journey of scientific inquiry is as much about **asking the right questions** as it is about finding the answers. The work of adapting organisms to extreme conditions, designing new languages for digital genomes, and creating frameworks for understanding evolutionary processes has taught me that the path to knowledge is never linear. It is a winding road of discovery, failure, and adaptation, where even small insights can ripple out to profound consequences.

This book is not simply a collection of projects or papers, it is a narrative of transformation. From the first flickers of curiosity that sparked my interest in artificial life to the philosophical inquiries that now inform my work, the chapters are a testament to the **power of reflection**. What began as a scientific pursuit has transformed into a deeply personal journey of self-exploration and growth. I have come to understand that science is not just a series of experiments or theories, but a way of engaging with the world. It is a lens through which we can explore the intricacies of life, our place within it, and the stories we tell about who we are and why we matter.

As I move forward, I carry with me the knowledge that **legacy is a living thing**. It is not defined by the number of papers or the prestige of accolades, but by the quiet moments of impact—the way we shape others, inspire curiosity, and ask questions that push the boundaries of what we know. In the end, the greatest contribution we can make is not in the work we do, but in the way we live and the way we engage with others in their own journeys.

While this book may seem to mark the end of a chapter, it is, in reality, just another moment in the ongoing dialogue between the past and the future. The work I have done, the lessons I have learned, and the questions I have asked will continue to shape the way I approach both science and life. As the Buddhism's Wisdom King Ragaraja transmuted lust into wisdom, so too have I sought to transform my own experiences; through reflection, inquiry, and adaptation; into something of lasting value.

As I close this book, I know that the story is far from over. In many ways, this is just a beginning; it is a continuation of the narrative that has guided my work, my teaching, and my reflections on life. I look forward to the future with curiosity, knowing that, like the organisms I have studied, we are all capable of evolving in unexpected ways. What lies ahead will undoubtedly present new challenges, but it is these very challenges that make the journey worthwhile.

Thank you for joining me on this exploration of science, self, and story. It has been an honour to share these reflections, and I hope they inspire others to ask their own questions, seek their own truths, and continue the search for wisdom in the endless dance of life and knowledge. See you soon.