Selene

Top 23 Supercomputer (December 2024)

Selene is NVIDIA's in-house supercomputer built on the modular DGX SuperPOD architecture, combining 560 DGX A100 systems to deliver exceptional performance for both AI and high-performance computing (HPC) workloads. (Angela Chen, 2021) Designed for rapid deployment and energy efficiency, Selene supports critical research in deep learning, autonomous systems, and scientific simulation, while serving as a blueprint for next-generation AI infrastructure.

## 1. Introduction

The **Selene** supercomputer, developed and operated by NVIDIA, marks a significant breakthrough in high-performance computing (HPC), with a specific focus on artificial intelligence (AI) and deep learning workloads. Unlike traditional supercomputers, which are typically built by government research laboratories for tasks such as physics simulations or weather forecasting, Selene was purpose-built to address the growing demand for AI-driven research and enterprise applications. (Stephanie Condon, 2020)

Selene is constructed using NVIDIA's DGX SuperPOD architecture. Upon its debut in 2020, it was ranked as the world's fifth-fastest supercomputer on the TOP500 list. (TOP500, 2020) It originally featured 280 NVIDIA DGX A100 nodes and was later expanded to 560, with each node housing eight A100 Tensor Core GPUs based on the Ampere architecture.

Selene achieved 63.4 petaflops on the High-Performance Linpack (HPL) benchmark, highlighting its exceptional capability for compute-intensive scientific tasks. More notably, it delivers nearly 2.8 exaflops of peak AI tensor-core performance. This illustrates Selene's dual strength in both traditional HPC and large-scale AI model development. (Brian Caulfield, 2020)


*Figure 1: NVIDIA Selene Supercomputer* (Rick Merritt, 2020)

## 1.1 Background

Selene was developed in 2020 during the early stages of the COVID-19 pandemic. The system was urgently built to support AI research and COVID-19-related scientific computing, including areas such as drug discovery and virus modeling. Remarkably, NVIDIA engineers assembled the initial 280-node system in under three weeks, while adhering to strict health protocols and using a rotating skeleton crew. This rapid deployment was made possible by NVIDIA's modular DGX SuperPOD design, which incorporated pre-configured infrastructure such as standardized cable bundles and 20-node units. Drawing from earlier systems like the 96-node "Circe" DGX-2 SuperPOD, Selene's architecture allows for fast scalability.(Mike Wheatley, 2020)

## 2. System Architecture

Selene's architecture is built on the modular design of NVIDIA DGX A100 systems, which serve as the supercomputer's core compute units. Each DGX A100 is a powerful, self-contained server that integrates high-performance GPUs, CPUs, memory, and networking into a unified platform. (Angela Chen, 2021) The following diagram illustrates the internal architecture of a single DGX A100 system, highlighting its major components and interconnections.

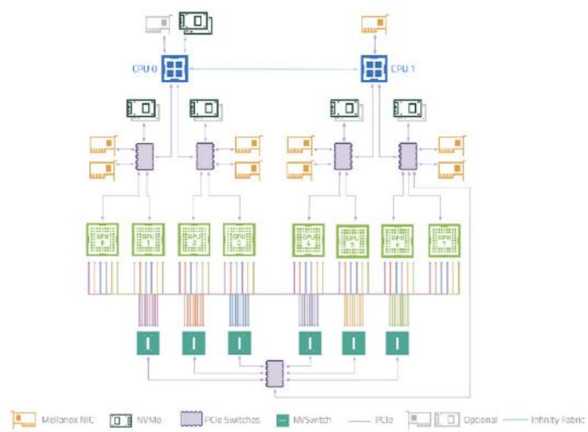


*Figure 2: Internal Architecture of a NVIDIA DGX A100 System (NVIDIA, 2024b)*

### 2.1 Compute Nodes

Selene consists of 560 NVIDIA DGX A100 systems (4 pods × 140 nodes each). Each DGX A100 contains 8× NVIDIA A100 (80 GB) GPUs and 2× AMD EPYC 7742 (64-core) CPUs. In total, Selene has 4,480 A100 GPUs (with 80 GB HBM2e each) and 555,520 CPU cores. Inside a DGX A100, GPUs are connected by NVLink/NVSwitch for 2.5 TB/s all-to-all GPU interconnect, and each CPU socket has its own PCIe NIC to avoid cross-node traffic. (NVIDIA, 2024b)

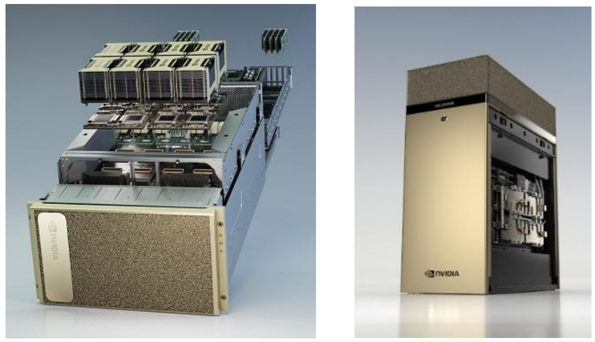

*Figure 3: A NVIDIA DGX A100 System (Nvidia, 2024)*

### 2.2 Connectivity

In terms of connectivity, each node is equipped with Mellanox HDR InfiniBand (200 Gb/s) adapters, typically eight 200 Gb/s HDR links for compute plus additional ports for storage/management. This 1:1 GPU-to-NIC ratio (two NICs per two GPUs) doubles earlier bandwidth, yielding a fat-tree network across all pods. (Patrick Kennedy, 2020) NVIDIA uses “thin-switch” modules to connect blocks of 20 nodes each, forming a large-scale fat-tree or DragonFly+ fabric for the full system. (NVIDIA, 2023)

### 2.3 Storage

Selene’s storage system is designed as a high-performance, multi-tiered hierarchy to support massive AI and HPC data workflows. At the node level, each DGX A100 includes 2× 1.92 TB NVMe drives for the OS and 30 TB of internal NVMe storage. Each SuperPOD features a Lustre-based parallel file system using 10 DDN AI400X units, offering 2.5 PB capacity and read/write speeds of 490/250 GB/s. This is supported by an Oracle ZFS5-2 shared file system (76.8 TB raw) for home directories and user data. Higher tiers include a 112 TB/s memory cache, 14 TB/s NVMe cache (8.4 PB), 1 TB/s network file system (5 PB), and a 100+ PB Ceph object store. (Patrick Kennedy, 2020)

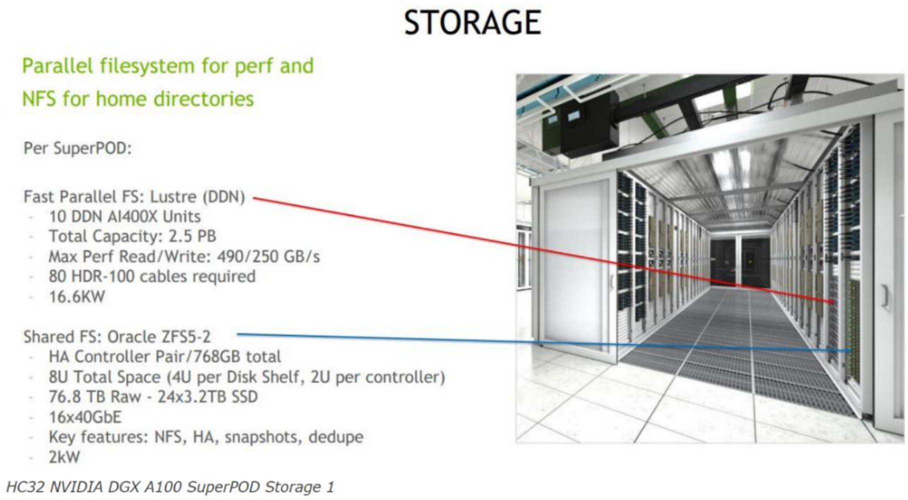


*Figure 4: NVIDIA DGX A100 SuperPOD Storage*

### 2.4 Cooling & Rack

For cooling, Selene runs in a standard datacenter with air-cooled, high-density racks. The design uses ~35 kW per rack (able to go beyond 50 kW), staying within ordinary facility cooling. No exotic cooling was needed, despite the high-power density. (Rick Merritt, 2020)

### 2.5 Modular Design

Lastly, the DGX SuperPOD design emphasizes modularity. For instance, racks are grouped by function (compute, storage, network, management) and assembled in scalable 20-node segments. This allows incremental expansion, and NVIDIA can easily add newpods for additional compute capacity. (Patrick Kennedy, 2020)

### 3. Manufactures & Institution

The Selene Supercomputer was developed and operated entirely by **NVIDIA Corporation**, a global leader in visual computing and accelerated computing technologies. Founded in 1993 and headquartered in Santa Clara, California, NVIDIA is best known for inventing the GPU in 1999, an innovation that revolutionized computer graphics and later enabled groundbreaking advances in artificial intelligence (AI) and high-performance computing (HPC). Over the years, NVIDIA has expanded its influence beyond graphics, becoming a major player in data center computing. (Wikipedia, 2025)

*Figure 5: Logo of NVIDIA* *Figure 6: NVIDIA Headquarter (Santa Clara, California)*

Selene was constructed in under three weeks in April 2020 and currently is located at NVIDIA's headquarters in Santa Clara, California. (Tiffany Trader, 2020) It plays a strategic role as both a research and development (R&D) resource and a reference platform for enterprise-scale AI computing. Today, Selene is actively used by NVIDIA engineers and researchers for developing technologies in autonomous driving, generative AI, large language models, and software development. Besides, Selene's design has influenced and been adopted by other large-scale institutions, such as Argonne National Laboratory for COVID-19 research and the University of Florida's HiPerGator AI supercomputer. (Mike Wheatley, 2020)

## 4. Unique Features

Selene is a powerful supercomputer for many reasons, with several unique features that distinguish it from the high-performance computing (HPC) landscape.

Most notably, it is the **first modular AI supercomputer based on NVIDIA's DGX SuperPOD architecture**. Built with 560 DGX A100 nodes, Selene exemplifies a scalable and flexible design, enabling rapid deployment and expansion from small-scale units to full supercomputer-class capability. This approach introduces a new paradigm in HPC system construction. (Angela Chen, 2021)

Another defining trait is Selene's record as **the fastest build of a supercomputer** in its class. Constructed in under one month during the COVID-19 pandemic using skeleton crews and remote tools, it demonstrated that modular, pre-configured architectures can achieve rapid deployment without compromising performance or reliability. (Rick Merritt, 2020) Selene also made history as the **first supercomputer to break the 26 GFLOPS/W efficiency barrier**. A single scalable unit achieved 26.2 GFLOPS/W, while the full system maintained an average of 24 GFLOPS/W, setting new standards for energy efficiency on the Green500 list. (Dion Harris, 2020)

Selene's versatility is further underscored by its **dual capability to handle both HPC and AI workloads** in a unified platform. It can seamlessly run scientific simulations (e.g., molecular dynamics) while powering large-scale AI tasks such as language model training and autonomous vehicle research, validated by its record-breaking MLPerf benchmark results. (Chris Porter, 2022)

As an **innovation testbed**, Selene is used internally by NVIDIA to test and refine emerging technologies at scale. These include MIG (Multi-Instance GPU) support, containerized environments using Pyxis + Enroot, DDN $A^3I$ storage, and HDR InfiniBand networking, enabling faster product iteration and optimization. (William Beaudin, 2021) Besides, Selene excels in **doubling its capacity** post-launch, showcasing the power of its modular architecture. For instance, NVIDIA expanded Selene from 280 DGX A100 nodes to 560 nodes, which effectively doubled its GPU count from 2,240 to 4,480. (Erik Deumens, 2021)

Lastly, Selene has played a critical role in **powering cutting-edge AI research**, advancing applications such as conversational AI, large language modeling, and self-driving simulation, all of which demand extreme computational performance and reliability. (Brian Caulfield, 2020)

## 5. Operating System

Selene operates in a Linux-based environment, primarily using Ubuntu (with a 5.3+ kernel) and CentOS, chosen for their stability, flexibility, and strong support for high-performance computing. The system runs NVIDIA's GPU software stack, including CUDA and cuDNN, enabling efficient acceleration of both scientific computing and deep learning tasks. This foundation allows seamless integration with popular AI frameworks like TensorFlow and PyTorch, unlocking the full potential of Selene's A100 GPUs. Combined with tools like Slurm for workload scheduling and Pyxis with Enroot for lightweight container deployment, Selene delivers a powerful, scalable, and customizable software environment for demanding AI and HPC workloads. (NVIDIA, 2024a)

## 6. Performance

Selene's computing performance has been thoroughly evaluated using industry-standard metrics. In the High-Performance Linpack (HPL) benchmark, which measures a supercomputer's ability to solve dense systems of linear equations, Selene achieved an impressive **Rmax** (maximum sustained performance) of **63.46 petaflops** (PFlop/s). This figure reflects the actual performance the system can sustain during real workloads. Its **Rpeak** (theoretical peak performance) stands at **79.22 PFlop/s**, indicating the maximum output the hardware could achieve under ideal conditions. (TOP500, 2024)

Selene also delivered strong results, **1,622.51 teraflops** (TFlop/s) on the **High-Performance Conjugate Gradients (HPCG)** benchmark, highlighting its capability to handle memory-bound and communication-intensive tasks. The system also achieved an **Nmax** value of **6,598,656**, representing the largest matrix size solved in the HPL benchmark. (TOP500, 2024) In terms of AI performance, Selene's GPU Tensor Cores provide a peak throughput of approximately **2,795 petaflops** (FP16/BF16 with sparsity). In July 2020, Selene broke records across all eight MLPerf training benchmarks, demonstrating its leadership in AI training workloads. (Brian Caulfield, 2020)

These results prove that Selene is one of the most powerful GPU-accelerated supercomputers, capable of handling both advanced scientific computing and enterprise-scale AI development.

## 7. Energy Efficiency

Energy efficiency is a central design principle of the Selene supercomputer. Selene draws approximately **2.646 megawatts (MW)** of power to deliver 63.46 petaflops on the LINPACK benchmark, resulting in an energy efficiency of approximately **24.0 GFLOPS/W**. This earned Selene the 2nd spot on the Green500 list in June 2020 and ranked 5th in November 2020. Notably, a single modular SuperPOD block (comprising 140 DGX A100 nodes) achieved an even higher efficiency of **26.2 GFLOPS/W.** (Brian Caulfield, 2020)

Selene's power efficiency is driven by its core, NVIDIA A100 Tensor Core GPUs. Each DGX A100 system can deliver up to 5 petaflops of AI performance while consuming just 6.5 kW. (Dion Harris, 2020)Mellanox 200 Gbit/s HDR InfiniBand networking ensures fast, low-latency data transfer with minimal energy loss, while the 1:1 ratio of NICs to GPUs optimizes data bandwidth without increasing power usage. Technologies like MIG, Slurm, Pyxis, and Enroot further optimize resource usage and reduce idle power consumption. (Angela Chen, 2021)

Moreover, Selene uses air cooling with cold aisle containment, a strategy that directs chilled air precisely to where it's needed, avoiding the energy costs associated with overcooling or liquid systems. Selene's modular DGX SuperPOD architecture further supports power efficiency by allowing unused nodes to be powered down when idle, ensuring energy is only consumed when necessary. (Angela Chen, 2021) Together, these choices make Selene a model for energy-efficient, scalable computing in the era of exascale systems.

## 8. Use Cases or Key Applications

Selene is designed not only as a showcase of NVIDIA's cutting-edge hardware and software integration, but also as a highly functional platform supporting real-world, mission-critical workloads. Below are three key areas where Selene has delivered significant impact:

### 1. AI Training and Autonomous Driving R&D

Selene leverages NVIDIA A100 GPUs to lead AI research and high-performance computing (HPC). It is used extensively for training large-scale models in natural language processing, computer vision, and autonomous vehicle simulations, accelerating innovation in cutting-edge AI development. (NADDOD, 2025)

### 2. COVID-19 Research

Originally built with 280 DGX A100 systems and Mellanox HDR InfiniBand networking, Selene showcased the scalability of NVIDIA's DGX SuperPOD architecture. One of its first major roles was to support COVID-19 research, performing complex protein docking and quantum chemistry simulations to advance understanding of the virus and aid in drug discovery. (Tiffany Trader, 2020)

### 3. Large-Scale Scientific Simulations

Selene supports massive simulations in areas like climate science, genomics, and physics by leveraging its vast parallel computing resources. It is also deployed across industries such as healthcare (drug discovery and medical imaging), energy (renewable source simulation), and manufacturing (design optimization) to demonstrate its versatility in scientific and industrial HPC applications. (Wikipedia, 2023)

## 9. Conclusion

Selene, NVIDIA's DGX A100-based supercomputer, is a major milestone in high-performance computing and artificial intelligence. Built using a modular DGX SuperPOD architecture, it achieved world-class performance and energy efficiency while rapidly deployed during the COVID-19 pandemic.

Through this research, I learned how modular design and commercially available components can lead to supercomputers that are easier to build, scale, and operate. I was especially surprised by how quickly Selene was constructed in under three weeks and how NVIDIA uses it daily for real-world tasks, not just as a showcase.

Selene is significant because it redefines how supercomputers can be designed, deployed, and applied in real-world scenarios. For science, it accelerates simulations and research. For industry, it drives innovation in AI, such as self-driving cars and language models. For society, it shows how computing power can be scaled efficiently and sustainably, pushing the boundaries of what is possible in healthcare, climate modeling, and more.

Looking ahead, Selene's architecture sets the stage for future exascale systems. Its role as a testbed for new technologies ensures that future supercomputers will be faster, greener, and better suited for the convergence of AI and scientific computing. Overall, Selene's influence will continue shaping the future of AI-driven research and infrastructure.

**References:**


Angela Chen. (2021, April). *Understanding Selene: A Modular NVIDIA Supercomputer*. https://www.nvidia.com/en-us/on-demand/session/gtcspring21-s31700/

Brian Caulfield. (2020, December 18). *Role of the New Machine: Amid Shutdown, NVIDIA's Selene Supercomputer Busier Than Ever*. https://blogs.nvidia.com/blog/nvidia-selene-busy/#:~:text=in%20academia%20when%20it%20goes,is%20also%20based%20on%20SuperPOD

Chris Porter. (2022, November 9). *Tuning AI Infrastructure Performance with MLPerf HPC v2.0 Benchmarks | NVIDIA Technical Blog*. https://developer.nvidia.com/blog/tuning-ai-infrastructure-performance-with-mlperf-hpc-v2-0-benchmarks/

Dion Harris. (2020, June 22). *Green Light! TOP500 Speeds Up, Saves Energy with NVIDIA*. https://blogs.nvidia.com/blog/top500-isc-supercomputing/

Erik Deumens. (2021, April). *NVIDIA Selene: Leadership-Class AI Supercomputing Infrastructure | GTC Digital April 2021 | NVIDIA On-Demand*. https://www.nvidia.com/en-us/on-demand/session/gtcspring21-s31844/

Mike Wheatley. (2020, August 14). *Nvidia built its Selene supercomputer for coronavirus research in just 3 weeks*. https://siliconangle.com/2020/08/14/nvidia-built-selene-supercomputer-coronavirus-research-just-3-weeks/#:~:text=Nvidia%20Corp,in%20just%20under%20three%20weeks

NADDOD. (2025, May 6). *Comparing NVIDIA's Top AI GPUs H100, A100, A6000, and L40S | by NADDOD | May, 2025 | Medium*. https://naddod.medium.com/comparing-nvidias-top-ai-gpus-h100-a100-a6000-and-l40s-03c2876e9982

NVIDIA. (2023, May 23). *Dragonfly+ Topology Validation - NVIDIA Docs*. https://docs.nvidia.com/networking/display/ibdiagnetusermanualv211/dragonfly++topology+validation

NVIDIA. (2024a, October 1). *Using Containers — NVIDIA DGX SuperPOD: User Guide*. https://docs.nvidia.com/dgx-superpod/user-guide-dgx-superpod/latest/containers.html

Nvidia. (2024, October 16). *Introduction to the NVIDIA DGX Station A100*. https://docs.nvidia.com/dgx/dgx-station-a100-user-guide/intro-to-station-a100.html

NVIDIA. (2024b, November). *Introduction to the NVIDIA DGX A100 System — NVIDIA DGX A100 User Guide*. https://docs.nvidia.com/dgx/dgxa100-user-guide/introduction-to-dgxa100.html

Patrick Kennedy. (2020, August 16). *NVIDIA DGX A100 SuperPod Detailed Look - ServeTheHome*. https://www.servethehome.com/nvidia-dgx-a100-superpod-detailed-look/

Rick Merritt. (2020, August 14). *Amid a Pandemic, Making Selene, an AI Supercomputer | NVIDIA Blogs*. https://blogs.nvidia.com/blog/making-selene-pandemic-ai/

Stephanie Condon. (2020, August 4). *How Nvidia built Selene, the world's seventh-fastest computer, in three weeks*. https://www.zdnet.com/article/how-nvidia-built-selene-the-worlds-seventh-fastest-computer-in-three-weeks/

Tiffany Trader. (2020, June 22). *Nvidia Nabs #7 Spot on Top500 with Selene, Launches A100 PCIe Cards*. https://www.hpcwire.com/2020/06/22/nvidia-nabs-7-spot-on-top500-with-selene-launches-a100-pcie-cards/

TOP500. (2020, November). *TOP 10 Supercomputer for November 2020*. https://top500.org/lists/green500/2020/11/

TOP500. (2024, November). *Selene - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband*. https://www.top500.org/system/179842/#:~:text=Linpack%20Performance%20%28Rmax%29%2063,22%20PFlop%2Fs%20Nmax%206%2C598%2C656

Wikipedia. (2023, September 27). *Nvidia Selene Supercomputer*. https://en.wikipedia.org/wiki/Selene_(supercomputer)#:~:text=Selene%20utilizing%201080%20AMD%20Epyc,single%20model%20to%20create%20art.

Wikipedia. (2025, June 15). *Nvidia* . Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Nvidia#:~:text=Nvidia%20Corporation%20(%2F%C9%9Bn%CB%88,California%2C%20and%20incorporated%20in%20Delaware.

William Beaudin. (2021, April). *Accelerating AI at-scale with Selene DGXA100 SuperPOD and Parallel Filesystem Storage | GTC Digital April 2021 | NVIDIA On-Demand*. https://www.nvidia.com/en-us/on-demand/session/gtcspring21-s31522/

**Acknowledgment by the student**
When Gen AI is allowed to be used, acknowledgement of how Gen AI is used can be included as follows:
(Place a tick '✓' in the checkbox where relevant)

| | Type | Acknowledgement |
|---|---|---|
| ✓ | Gen AI is allowed but students choose not to use it | I did not use any Gen AI tools for this assignment. |
| | Generate content which is modified in the final assessment submission | I acknowledge the use of …………………. to general materials such as …………… in this assessment. I have also provided the prompt used, the output generated by the Gen AI tools, how the output was used and the pages where the Gen AI-generated content can be found. |
| | Use Gen AI to polish language before further modification for final submission | I acknowledge the use of …………………. to improve the academic tone and accuracy of language, including grammatical structures, punctuation and vocabulary. I have also provided the prompt used, the output generated by the Gen AI tools, how the output was modified further to better represent my tone and style of writing. |
| | I attached Appendix in this assignment the chat history, including all the prompts that I used and the output from the Gen AI tool(s) for this assignment. | |

The use of GenAI has to be referenced appropriately according to the referencing style as required:

| **Referencing Style*** | **In-text citation** | **Reference** |
|---|---|---|
| APA | … (AI platform, year) | Company. (year). *Platform Used* (version) [model]. URL<br><br>***Example:***<br>OpenAI. (2023). *ChatGTP* (Mar 14 version) [Large language model]. https:chat.openai.com/chat |

*Please refer to the latest referencing style.

**Declaration**

I confirm that:
a) I understand what constitutes academic integrity violations, including plagiarism, collusion, fabrication, falsification, contract cheating, and improper use of Gen AI.
b) I confirm that my work or any part of this assessment has neither been previously and is not concurrently submitted for any other programme at Taylor's University or any other institution, save except when re-use of the same work is permitted by the module leader.

c) I acknowledge that using Gen AI or any external assistance without proper attribution constitutes academic misconduct and may be sanctioned accordingly.
d) I understand that if there are indications of academic integrity breaches, including improper Gen AI use, my work will be subject to investigation.
e) This assignment is my own work, and I have properly acknowledged all sources, tools, and external contributions, including the use of Gen AI where applicable.
f) I also understand that if there is any indication of inappropriate use of generative AI in my assignment, I may be required to, for instance, attend an oral presentation to justify my work as part of the review process.
g) I acknowledge and authorize the submission and/or storage of my work in a database for the purpose of verifying its originality and/or conducting tests using artificial intelligence software, and I hereby consent to this process.
h) I acknowledge that this submission is subject to Taylor's University/College Academic Integrity Procedure (THE-ACA-PROC-AINT) and all applicable university regulations.

| | | |
|---|---|---|
| **Name1 & Student ID** | | **Chloe Tee Rouyi (0354731)** |
| **Name2 & Student ID** | | **Hng Qi Yean (0364483)** |
| **Name2 & Student ID** | | **Rachel Yap Qi Qi (0362056)** |
| **Signature1:** ***Chloe*** | **Signature2:** ***Hng*** | **Signature3:** ***Rachel*** |
| **Date: 21 Jun 2025** | | |