SIML CUDA vs. nVidia Triton for Large Scale Cellular Simulation

Deep SIML Labs · 18 November 2026 · Version 1.0

SIML CUDA simulated 2,078,368 additional interacting cells under a common 6.0 GB GPU memory budget. On one NVIDIA GeForce RTX 3060, SIML CUDA advanced 8,100,736 cells through six seconds of model time. The comparison used the Triton GPU programming language on the same NVIDIA hardware. Triton advanced 6,022,368 cells under the same memory and accuracy requirements. The measured capacity difference was 34.51%.

At the largest population both systems could advance, SIML CUDA also used 24.93% less peak GPU memory. Its five advances after the first simulated second averaged 541.04 seconds of wall time, compared with 599.45 for Triton. This gives an observed throughput ratio of 1.108× in one paired run.

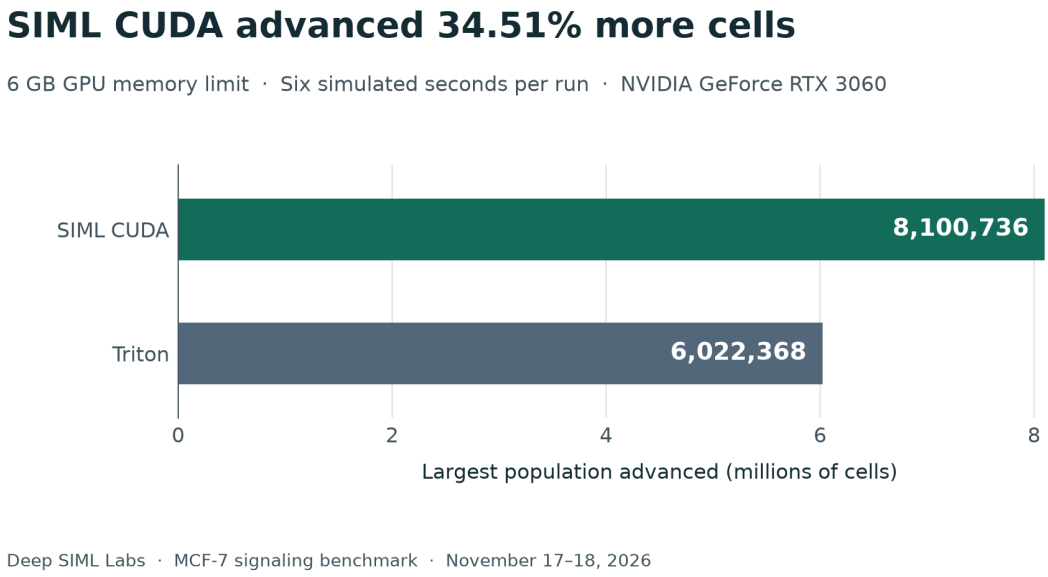


Figure 1. Each population completed the same six-second simulation and passed its accuracy checks. At each boundary, requesting 32 additional cells triggered the memory limit.

What a cell represents

Each cell represents a patch of MCF-7 culture with its own receptor abundances and intracellular signaling state. The benchmark arranges the patches in a line. Neighboring patches exchange epidermal growth factor and heregulin. The intracellular equations come from the ErbB signaling model of Birtwistle and colleagues [1, 2]. In this benchmark each patch has 116 evolving molecular state variables and 102 reaction entries. The counted population thus measures how many interacting copies of a defined signaling system the GPU can advance.

Memory and time at equal population

The two runs at 6,022,368 cells used the same initial conditions, geometry, reactions and accuracy criteria. SIML CUDA reached a measured process peak of 4.383 GB and Triton reached 5.838 GB. At equal population this freed 1.455 GB of the budget. Timing both systems at this population also gives a direct comparison of the same simulated work.

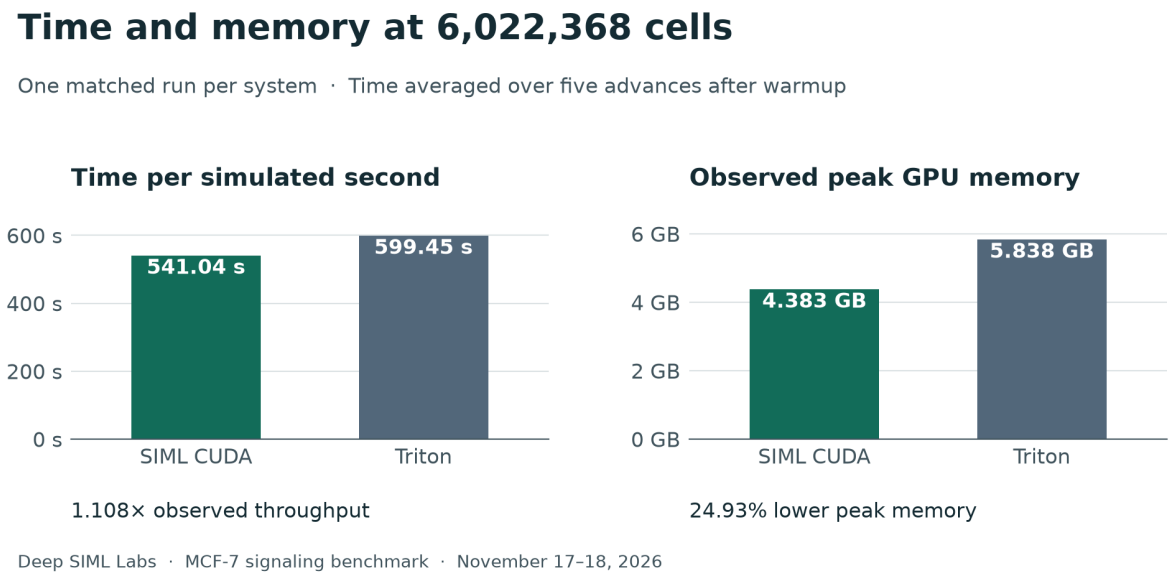


Figure 2. Each system ran once in a fresh process. The time measure averages five consecutive one-second advances after the first simulated second. Memory is the observed GPU process peak over the run.

The cost of the first simulated second

The first advance consumed most of the wall time. It took 3.72 hours with SIML CUDA and 4.10 hours with Triton at the shared population. Later advances became progressively faster. For instance, the second and sixth simulated seconds took 963 and 294 seconds with SIML CUDA, and 1,065 and 327 seconds with Triton. A researcher planning to start a field from these initial conditions should include the first advance in the time budget.

At 6,022,368 cells	SIML CUDA	Triton
First simulated second (wall hours)	3.72	4.10
Next five simulated seconds, total (wall hours)	0.75	0.83
Entire process including setup (wall hours)	4.47	4.94

The five advances follow a single trajectory in each process. Their mean describes this interval. Independent process pairs would be needed to estimate run-to-run uncertainty at this population. An earlier 4,096-cell comparison with 20 alternating process pairs measured 1.082× throughput, with a one-sided 95% lower bound of 1.079×. That result fell below the study's prespecified 1.2× criterion for a speed advantage of that magnitude.

How we tested the result

Common workload and memory limit

Both systems ran sequentially on one RTX 3060 with 12 GiB of physical memory at default driver clocks. A common limit of 6,000,000,000 bytes applied to each process, including a 128 MB reserve for overhead. We report process peaks in decimal GB. At their respective maximum populations, the observed peaks were 5.845 GB for SIML CUDA and 5.838 GB for Triton.

We used fresh processes to find the boundary to a resolution of 32 cells. We then ran three simulations. Triton ran at 6,022,368 cells, and SIML CUDA ran at both 6,022,368 and 8,100,736. All used the same heterogeneity seed. Every full advance included transport, intracellular reactions, error checks, synchronization and publication of the next field state. All 18 advances across the three runs passed.

Numerical agreement

Before the capacity runs, both systems passed eight screening cases and all 32 held-out cases on six-cell fields over 20 simulated seconds. Each molecular trajectory had to stay within 1% of its reference peak plus a small absolute allowance scaled to the published initial concentration. The reference used an independent double-precision integration. The checks on normalized ERK and Akt responses required agreement within one percentage point.

At large populations, every advance checked conserved quantities across the whole field and compared reaction values at nine sampled patches to a double-precision reference. The final states at the shared population also passed a direct comparison at 16 patches. These checks establish numerical agreement at the coverage described. Biological predictions at tissue scale require separate experimental validation.

Implications for research partners

A partner using this signaling model can place about 2.08 million more interacting patches on the same GPU within the tested budget. For a project holding population fixed, the matched run shows how much GPU memory each system used. Both results apply to this six-second interval, one-dimensional geometry, hardware and comparison settings. Longer runs, different biological models and additional spatial dimensions require their own capacity and accuracy measurements.

Sources and publication

[1] Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB and Kholodenko BN. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Molecular Systems Biology* (2007). doi:10.1038/msb4100188.

[2] BioModels. *Birtwistle2007 ErbB signalling model*. Accession BIOMD0000000175.

Deep SIML Labs conducted this benchmark. Technical whitepaper, November 2026. This publication has not undergone peer review. Research partnership inquiries.