**Enhancing the efficiency of animal-alternative in silico drug cardiotoxicity prediction through CUDA-based parallel processing**

Iga Narendra Pramawijaya1\*, Ariyadi1, Aroli Marcellinus1 , Ali Ikhsanul Qauli1 , Ki Moo Lim1

1: Department of IT Convergence Engineering, Kumoh National Institute of Technology, Korea

\*: [iga@kumoh.ac.kr](mailto:iga@kumoh.ac.kr)

**Abstract:**

This research focuses on enhancing in silico cardiotoxicity prediction by utilising GPU-based parallel computing. Traditional CPU-based simulations are computationally expensive, especially for large-scale studies. By leveraging CUDA programming, this research aims to optimise simulation efficiency while maintaining the accuracy of cellular electrophysiological models. The study employed three well-established cardiac cell models: ORd 2011, ORd 2017, and ToR-ORd. Simulations were conducted using GPU-based implementations of ordinary differential equation (ODE) solvers, with the Rush-Larsen method applied for ORd 2011 and a Forward Euler approach for ORd 2017 and ToR-ORd. The simulations were validated against CPU-based OpenCOR results, with performance evaluated in both drug-free and drug-induced conditions. GPU simulations demonstrated equivalent accuracy to CPU-based results, effectively replicating action potential dynamics and key biomarkers across all cell models. However, the Forward Euler solver required more computation time compared to the Rush-Larsen method. Computational performance analysis revealed significant efficiency improvements (up to 49 times in fastest configuration) in GPU-based simulations. This research successfully validates GPU-based parallel computing as a reliable and efficient approach for in silico cardiotoxicity prediction. The findings support its potential for accelerating drug discovery processes while reducing reliance on animal testing. Future work will focus on expanding model complexity and variabilities to further enhance the system’s applicability.

**Keywords:** In silico, cardiotoxicity, CUDA, GPU Parallel Programming

**1 Introduction**

Cardiovascular diseases are the leading global causes of death, which emphasizes the importance of effective methods for cardiac drug discovery. Traditionally, drug cardiotoxicity prediction is achieved using animal testing, which takes time due to ethical clearance and effortful. Modern *in silico* or computer-based methods for drug cardiotoxicity prediction show promising results as an animal-alternative solution. Nevertheless, at some point, this approach is computationally inefficient due to large amount of sample it needs to compute to mimic natural variations. As the sample size increases, the complexity of the calculations grows, resulting in longer processing times and reduced efficiency.

This efficiency limitation makes it difficult for traditional computational approaches to handle large-scale simulation (such that uses multi-sample scenario or inter-individual variations) within a reasonable timeframe. This research introduces an updated solution to address the computational inefficiencies of current in silico drug cardiotoxicity simulations. By implementing NVIDIA’s CUDA (Compute Unified Device Architecture)-based parallel programming on Graphics Processing Units (GPU) [1], this method can significantly accelerate overall computational process, enabling faster handling of large-scale simulations. By leveraging the power of parallel processing, computational time will be reduced and this can accelerate preclinical testing, potentially reducing drug development costs and reliance on animal testing.

Parallelisation has been a valuable concept in computational biology for some time. For example, the Cells in Silico (CiS) framework developed by Berghoff et al. [8] provides a modular and parallel design for simulating biological tissue growth and development. This flexibility allows researchers to configure various model assumptions for diverse research questions. A key application of CiS was the simulation of a 1000³ voxel-sized cancerous tissue at sub-cellular resolution, showcasing its ability to handle complex biological processes with high detail.

The use of GPUs in biological cell computing has also been explored extensively in prior research. Martinez et al. [9] introduced an adaptive parallel simulator to mitigate performance loss in massive parallel membrane computing devices, also known as P systems. By extending an existing simulator for Population Dynamics P systems, their approach improved performance by up to 2.5 times on both GPUs and multicore processors. Similarly, McIntosh-Smith et al. developed BUDE (Bristol University Docking Engine), a drug discovery tool for molecular docking, to work with OpenCL, a standard for parallel programming. Their adaptation enabled peak performance of 46%, or 1.43 TFLOP/s, on a single Nvidia GTX 680 [10]. Amar et al. extended parallelisation to biochemical simulations of metabolic pathways, while Barth et al. leveraged parallel computing in BUDE to handle more complex models with an increased number of chemicals and reactions, achieving more realistic and computationally efficient outcomes [11].

Recent work has increasingly focused on using in silico methods to address specific cardiac conditions and evaluate pharmacological interventions. For example, Whittaker et al. employed computational modelling to examine mutations linked to Short QT Syndrome and their effects on atrial arrhythmias. Their study demonstrated how in silico simulations could assess drug responses and inform strategies for treating genetic cardiac disorders [12].

Advances in computer hardware and parallel processing have further enhanced the speed and efficiency of in silico cardiac simulations, enabling the analysis of larger datasets and more intricate models. However, despite these advancements, optimising *in silico* simulations specifically for cardiotoxicity prediction remains underexplored. Therefore, this study main goal is to improve the efficiency and scalability of cardiotoxicity simulations by employing CUDA-based parallel processing techniques. The research aims to:

1. Address computational bottlenecks caused by increasing sample sizes and complex calculations in traditional methods.
2. Optimize GPU resources for faster, large-scale simulations without compromising accuracy.
3. Validate the accuracy and reliability of GPU-based simulations compared to CPU-based methods.
4. Develop a practical and cost-effective approach suitable for real-world drug discovery applications, reducing reliance on animal testing.

**2 Methodologies**

This chapter describes the development process of the GPU-based cardiac cell simulation software. The focus is on enabling multi-sample simulations, where each cardiac cell model is simulated in parallel. This chapter will guide readers through the process of modifying the generated code for GPU simulation, and implementing parallelisation techniques to handle multiple samples efficiently. Additionally, this chapter will explain how ordinary differential equations (ODEs) in the cell models are solved within this framework. The goal is to ensure that researchers or software engineers can follow the steps presented here to replicate the parallelisation process and build their own GPU-based multi-sample simulation platforms.

Once cell model file is acquired in C or C++ format, conversion in terms of GPU memory adjustment and offsetting is needed. CUDA-based application heavily depends on how data is transferred between the host (CPU) and device (GPU), as well as how it is organized within the GPU's memory. This research uses three types of GPU memory: global, shared, and constant memory.

In the GPU, global memory is the largest and most flexible, but with slower access speed. Shared memory is faster to access but has limited space (10 KB) and is restricted to threads within the same computing block. Constant memory is the fastest, with purpose to deliver execution commands to all threads. Constant memory’s size is more suitable to store constants during execution due to its 64 KB limit.

Proper adjustment of these memory types can significantly enhance computational efficiency. In this research, the global memory is used for storing variables, constant memory for orchestrating commands from CPU, and shared memory used to optimise GPU to CPU feedbacks.

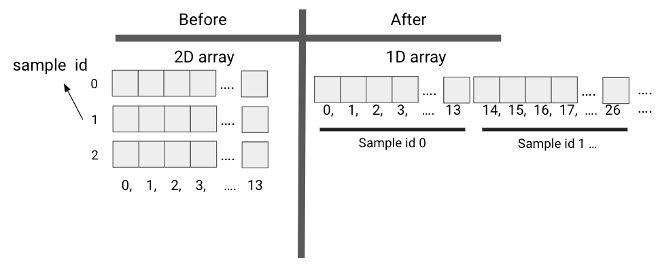


Figure 1. Main difference in values storing paradigm after CUDA-parallelisation, assuming column size is 13.

In this research, selecting an optimal number of threads per block is crucial to maximising utilisation of shared memory. Two factors need to be considered to optimise performance when selecting core per block value: thread grouping in CUDA, and number of samples. CUDA executes threads in groups called warps, which consist of 32 threads [16]. Using a block size that is a multiple of 32 ensures that all warps are fully utilised, minimising idle threads and maximising efficiency.

It is known that this configuration is not transferable across different GPUs. As the time writing this, NVIDIA 40xx series GPU supports 32 core per block, while older series like the 30xx only support 16 core per block. 30xx series also uses wraps, but it has less computing core. Therefore, for the 30xx series, 16 cores per block were selected, aligning well with the hardware’s limitations, as 16 × 2 = 32 matches the warp requirements.

The choice of cores per block was also influenced by the number of samples. Each simulation sample usually comes in the multiplication of 2000 (2000, 4000, etc.). To optimise warp utilisation (with warps consisting of 32 threads), To optimize warp utilization, the number of cores should be close to 32 and evenly divisible by 2000. Through trial and error, 20 cores per block were found to provide the most efficient configuration. This adjustment ensures that each sample is allocated its own computing core.

By default, the code is configured to use 32 cores per block. However, hardware limitations may cause errors with this configuration. In such cases, changing the number of cores per block to 20 ensures the parallel processing process remains stable. Since 32 does not divide evenly into 2000, the code automatically rounds up the number of blocks to ensure that at least one additional core is available for each sample.

Offsetting is a technique used to manage data indexing efficiently, ensuring that thread indices correspond to the correct memory addresses. This can reduce memory bank conflicts and improve overall performance. Proper offset calculation is also crucial when dividing large datasets across multiple threads and blocks, ensuring each thread processes its designated segment efficiently and correctly. In this research, offsetting also used to simplify any multi-dimensional input. In the previous iteration of the simulation based on CPU, it uses vector of struct to temporarily store simulation results.

In CUDA programming, there is no native multidimensional vector type like in higher-level programming languages. This research simplified all multi-dimensional vector used in the previous iteration into 1 dimensional (1D) array. Offsetting is mainly used for pointing the correct data in a flattened 1D array. Each sample will have their own identification number called the sample\_id, so each row correlates to one sample\_id. In the GPU version, instead of using rows to differentiate samples, it will utilise fact that each arrays have same number of columns. For example, array STATES has 43 columns, then STATES[0] up to STATES[42] is reserved for sample\_id = 0, STATES[43] up to STATES[84] is reserved for sample\_id = 1, and so on. Adapting with this approach, the current index can be determined by knowing row dimension, sample\_id and the number of specific columns selected.

The model relies on algebraic calculations and dynamic functions expressed in the form of ordinary differential equations (ODEs), which are essential for simulating the complex behaviours of biological systems. To efficiently solve these ODEs within a CUDA-based parallel processing framework, two distinct numerical methods were employed depending on the specific cell model: the Rush-Larsen method and a custom implementation of the forward Euler method. These methods were chosen to balance computational efficiency, numerical stability, and compatibility with the CUDA architecture. This research implements the ODE solver inside the cell model code as a function.

For the ORd 2011 model, the Rush-Larsen method was utilised due to its computational efficiency and computational stability in this context. This method effectively integrates stiff components of the equations, making it well-suited for the dynamic features of the ORd 2011 model. This method was implemented with a dynamic time-stepping mechanism by adjusting the time step during each iteration, while balancing acceptable numerical errors [17].

Initially, the Rush-Larsen method was meant to be used across all cell models. However, when applied to ORd 2017 and ToR-ORd models, this approach exhibited instability, failing to provide reliable results. To address this limitation, a simple forward Euler method was implemented for these two models.

Forward Euler is a simple method to solve ODE in particular time. The forward Euler method calculates the next value of a variable by taking its current value (STATES array) and adding the product of the rate (RATES array) of change and the time step (dt). This straightforward approach makes the method computationally simple and easy to implement. Mathematically, it is expressed as in Equation 1:

*xn+1=xn + rate(xn) ⋅ Δt* (1)

where *xn+1* is the current value, *rate(xn)* represents the rate of change at *xn*, and *Δt* is the time step. In the converted code, a function should be added to implement this calculation.

While the Euler method produced accurate and stable results, especially for the ORd 2017 and ToR-ORd models, it proved to be computationally intensive, significantly increasing the runtime. To optimise the parallelisation process, algorithm was simplified, enables parallel threads to process multiple samples rather than multiple equations simultaneously. Despite the trade-off between computational speed and stability, this combination of methods ensures that the CUDA-based framework effectively supports the diverse requirements of different cell models, while maintaining accuracy and to reduce numerical errors.

The simulation produces two distinct types of output files, a biomarker file and time-series data files, along with one intermediate cache file. The cache file is generated as the output of the `kernel\_DoDrugSim` function, which represents the initial phase of the simulation. During this phase, the function runs the simulation for thousands of cycles (referred to as paces) to amplify the drug effects within the model. After this initial phase, the `kernel\_DoDrugSim\_single` function is executed, which generates the biomarker file and the time-series data files. All output files are organised into a dedicated folder within the `result` directory for efficient storage and retrieval.

The biomarker file provides a summary of key features extracted from the simulation for each sample. It includes data such as sample numbers, qNet, qInward, and action potential shape analysis result. These biomarkers represent crucial physiological parameters simulated under drug influence, and they are instrumental for downstream

analyses, such as machine learning-based predictions. The time-series file offers a detailed temporal view of each sample’s behaviour. Each sample has its own individual time-series file; thus, a simulation involving 2000 samples will result in 2000 time-series files. These files capture parameters such as time, action potential, voltage gradient over time, Cai, INa, INaL, ICaL, IKs, IKr, IK1, and Ito. Using this detailed data, it is possible to plot the drug-induced cellular responses over a single cycle, facilitating visualisation and deeper analysis of dynamic behaviours.

**3 Results and Discussion**

This chapter presents the results of GPU-based simulations for three cardiac cell models: ORd 2011, ORd 2017, and ToR-ORd. It analyses simulation accuracy, drug-induced changes, and computational performance. Two types of simulations were conducted: control (no drug) and drug-induced using bepridil at concentrations of 33 mMol (cmax 1), 66 mMol (cmax 2), and 132 mMol (cmax 4). Each simulation was run for 1000 pacing cycles, with each cycle lasting 1000 milliseconds.

Table 1. Simulation time for different cell models using CPU and GPU

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cell Model | Mean Absolute Error | Single-Core CPU Time (s) | Multi-Core CPU Time (s) | GPU Time (Both Single and multi-core) (s) |
| ORd 2011 | 0.078 | 57 | 45600 | **928** |
| ORd 2017 | 0.004 | 390 | 312000 | **40089** |
| TOR-ORd | 0.023 | 455 | 364000 | **38542** |

**3.1 Simulation Validation**

Simulation validation is a crucial step in ensuring the accuracy and reliability of GPU-based cardiac electrophysiology simulations. The results of the GPU simulations were validated by comparing them with CPU-based simulations using OpenCOR as the ground truth. Across all three cell models—ORd 2011, ORd 2017, and ToR-ORd—the validation process involved calculating the mean absolute error (MAE) between the GPU and CPU results and evaluating key electrophysiological biomarkers, including action potential duration (APD) and calcium transient properties.

For the ORd 2011 model, the GPU simulation showed an MAE of 0.078 mV, demonstrating a high degree of accuracy. The Rush-Larsen method, used in this model for solving ordinary differential equations (ODEs), contributed to its computational stability and efficiency. The validation confirmed that the GPU accurately captured the physiological behaviours of the cardiac cells, including both control and drug-induced conditions, providing confidence in the simulation's robustness.

The ORd 2017 model achieved an exceptionally low MAE of 0.004 mV, reflecting near-perfect agreement with the CPU results. This model employed a forward Euler method for solving ODEs, which, while computationally more intensive than Rush-Larsen, delivered precise results. Validation under both no-drug and drug-induced conditions demonstrated the GPU’s capability to reproduce the intricate dynamics of the cardiac cells simulated in this model.

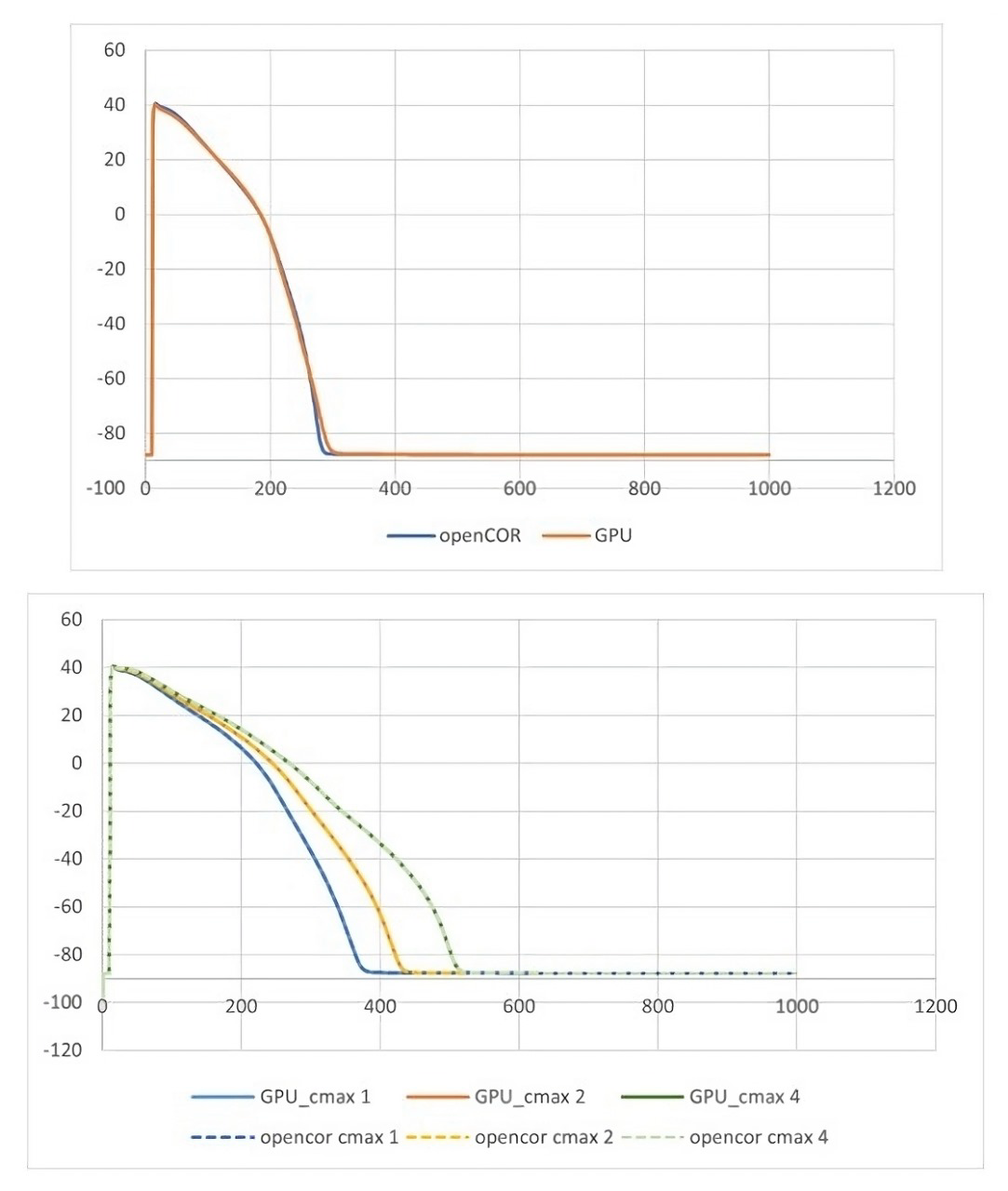


Fig. 2. Action Potential from GPU simulation in ORd 2011 model, compared to CPU simulation in non-drug and under-drug situation.

In the ToR-ORd model, the GPU simulation yielded an MAE of 0.023 mV, indicating strong alignment with the CPU reference. Despite the model's higher complexity and computational demands, the GPU efficiently handled both control and drug-induced scenarios. The simulation accurately replicated the cardiac cells' electrophysiological responses to bepridil at varying concentrations, further confirming the GPU's suitability for large-scale, high-fidelity simulations.

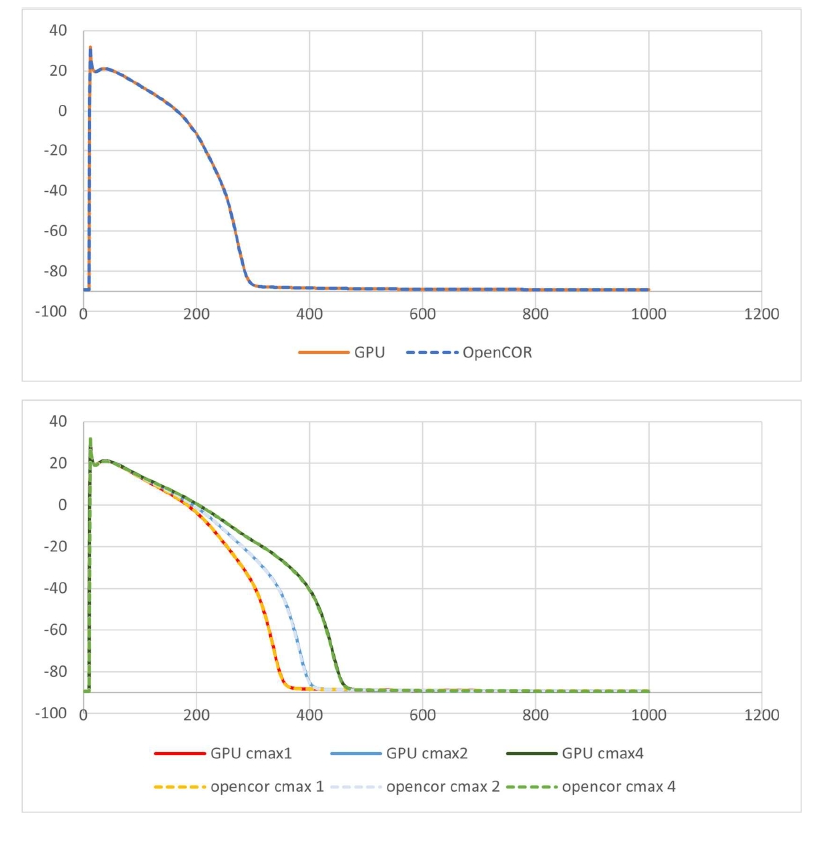


Fig. 3. Action Potential from GPU simulation in ORd 2017 model, compared to CPU simulation in non-drug and under-drug situation.

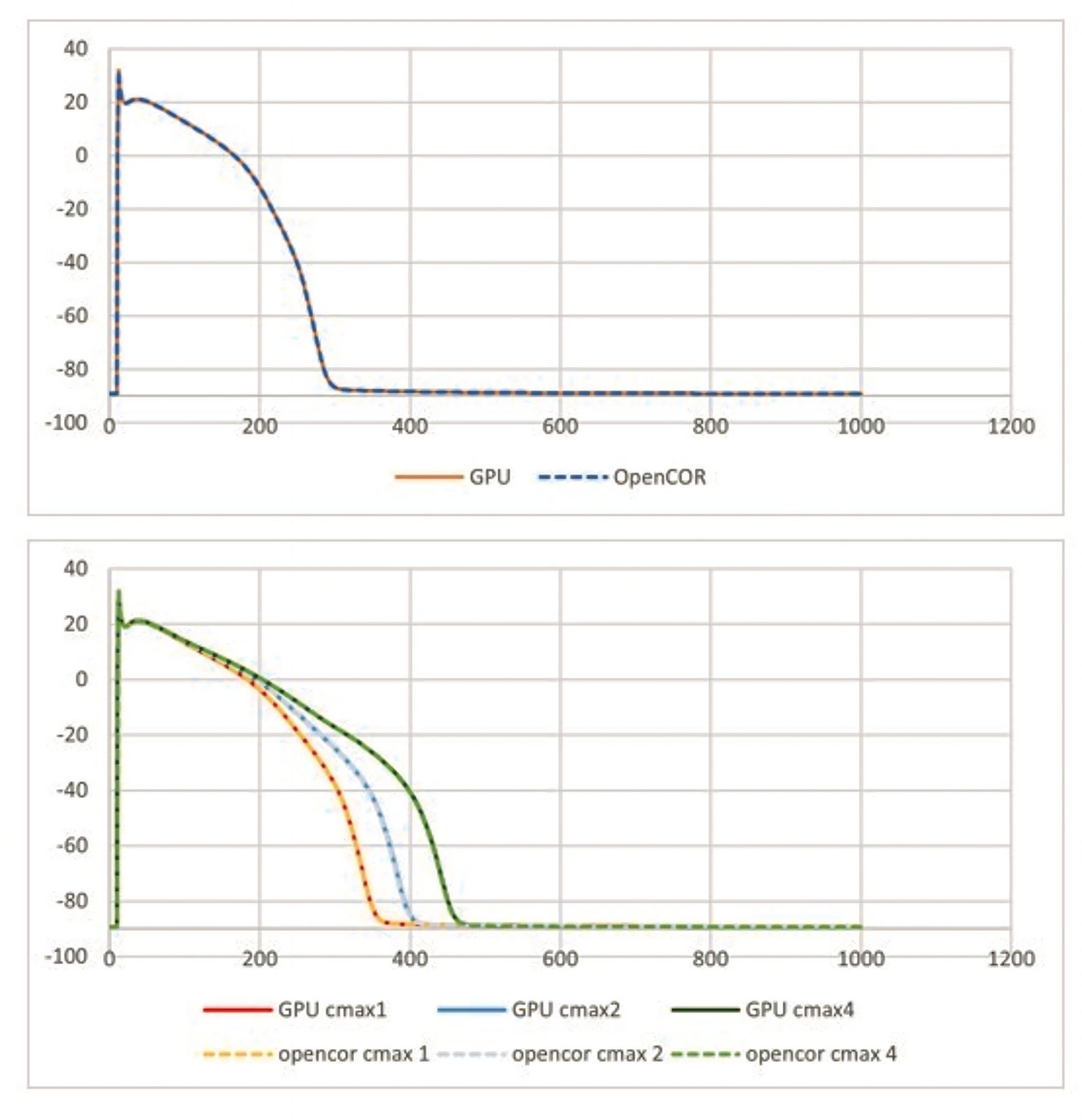


Fig. 4. Action Potential from GPU simulation in ToR-ORd model, compared to CPU simulation in non-drug and under-drug situation.

Overall, the validation results for all three cell models underscore the accuracy and reliability of the GPU-based approach. The low MAE values across models confirm that the parallel computing implementation does not compromise the simulation's fidelity, making it a viable alternative to traditional CPU-based methods. These findings support the broader application of GPU-based simulations in cardiac electrophysiology and drug discovery research.

**3.2 Computation Time Performance**

The computational time performance of GPU-based simulations was compared against both single-core and multi-core CPU implementations for the three cardiac cell models: ORd 2011, ORd 2017, and ToR-ORd. This comparison highlights the efficiency gains achieved through CUDA-based parallel processing on GPUs.

For the ORd 2011 model, the GPU achieved significant time savings compared to its CPU counterparts. A single-core CPU required 57 seconds to complete the simulation for one sample, while a 10-core multi-core configuration took 45,600 seconds for 8,000 samples. In contrast, the GPU completed the same workload in just 928 seconds, representing a substantial speedup. This efficiency is attributed to the Rush-Larsen method employed in ORd 2011, which is well-suited for GPU parallelisation and reduces computational overhead.

The ORd 2017 model, using a forward Euler ODE solver, was computationally more demanding. A single-core CPU required 390 seconds per sample, and the multi-core CPU setup took 312,000 seconds for 8,000 samples. Despite the increased complexity, the GPU outperformed the CPU significantly, completing the same task in 40,089 seconds. While the GPU’s advantage is slightly less pronounced for this model, it still demonstrated its capability to handle large-scale simulations more efficiently than traditional CPUs.

The ToR-ORd model, being the most complex of the three, posed the greatest computational challenges. Single-core CPU simulations required 455 seconds per sample, and the multi-core setup took 364,000 seconds for 8,000 samples. In comparison, the GPU completed the same workload in 38,542 seconds, achieving a notable speedup. This performance demonstrates the scalability of GPU-based simulations for handling high-fidelity cardiac models with greater computational demands.

Across all three cell models, the GPU consistently outperformed both single-core and multi-core CPU configurations, particularly as the sample size increased. The results underscore the GPU’s efficiency in large-scale simulations, making it a superior choice for computationally intensive applications. These findings highlight the GPU’s potential to accelerate cardiac electrophysiology research and reduce the time required for preclinical drug testing.

**4. Conclusions, Limitations and Suggestions**

This research proven its ability to aim a more computationally efficient *in silico* cardiotoxicity prediction by leveraging CUDA-based parallel processing. CUDA enables this research to do massive parallelisation with up to 40.91 times faster simulation speed for 8000 drug samples, compared to CPU-based 10-core simulation. Validation results confirmed the accuracy of GPU simulation. Little to no difference were found in both GPU and the well-used CPU simulation results. This result highlights not only the GPU capability to produce more samples in shorter amount of time, but also it is similarly reliable to common CPU cardiotoxicity simulation.

While this research demonstrates significant advancement in GPU-based parallel simulations *in silico* cardiotoxicity prediction, several limitations should be noted. First, this research restricts hardware use by requiring NVIDIA GPUs, as it relies on CUDA for its programming. This dependency makes the research incompatible with another brand’s GPU architecture. Another limitation is the reliance on simplified ODE solvers, such as the Rush-Larsen and forward Euler methods. Other methods should be heavily modified to be applied into this research. Furthermore, the fixed configurations for threads and blocks may not optimally utilise newer or varying GPU hardware, limiting future scalability.

This research focuses solely on drug-induced simulations using specific cell models and parameters, leaving the generalisation to other cell types or drugs unexplored. This point raises future suggestion to expand the research by incorporating another biological cell models and variables. More testing on diverse GPU hardware models would provide insights into performance variations across systems, ensuring broader applicability. Furthermore, assessing the economic viability of the GPU-based approach, including long-term operational costs and energy efficiency, could ensure its sustainability. Standardising protocols for simulation workflows would help maintain consistency and reproducibility, which are critical for scientific and industrial adoption.

**Acknowledgements**

This research was partially supported by the Ministry of Food and Drug Safety (22213MFDS3922 and (24212MFDS266), National Research Foundation of Korea (NRF) under the Basic Science Research Program (2022R1A2C2006326), and Ministry of Science and ICT Korea (MSIT) under the Grand Information Technology Research Center support program (RS-2020-II201612) supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP).

**Declarations**

The authors have no financial or any conflict of interest to declare during the research.

**References**

1. Jason Sanders and Edward Kandrot. 2010. CUDA by Example: An Introduction to General-Purpose GPU Programming (1st. ed.). Addison-Wesley Professional.

2. “What is OpenMP?” Accessed: Nov. 10, 2024. [Online]. Available: https://cvw.cac.cornell.edu/openmp/intro/what-is-openmp.

3. R. L. Graham, G. M. Shipman, B. W. Barrett, R. H. Castain, G. Bosilca and A. Lumsdaine, "Open MPI: A High-Performance, Heterogeneous MPI," 2006 IEEE International Conference on Cluster Computing, Barcelona, Spain, 2006, pp. 1-9, doi: 10.1109/CLUSTR.2006.311904.

4. A. Garny et al., “CellML and associated tools and techniques,” Sep. 13, 2008, Royal Society. doi: 10.1098/rsta.2008.0094.

5. M. Gómez, J. Carro, E. Pueyo, A. Pérez, A. Oliván, and V. Monasterio, “In Silico Modeling and Validation of the Effect of Calcium-Activated Potassium Current on Ventricular Repolarization in Failing Myocytes,” IEEE J Biomed Health Inform, pp. 1–9, 2024, doi: 10.1109/JBHI.2024.3495027.

6. C. M. Lloyd, J. R. Lawson, P. J. Hunter, and P. F. Nielsen, “The CellML Model Repository,” Bioinformatics, vol. 24, no. 18, pp. 2122–2123, 2008, doi: 10.1093/bioinformatics/btn390.

7. N. Le Novère et al., “BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.,” Nucleic Acids Res, vol. 34, no. Database issue, 2006, doi: 10.1093/nar/gkj092.

8. M. Berghoff, J. Rosenbauer, F. Hoffmann, and A. Schug, “Cells in Silico-introducing a high-performance framework for large-scale tissue modeling,” BMC Bioinformatics, vol. 21, no. 1, Oct. 2020, doi: 10.1186/s12859-020-03728-7.

9. M. Martínez-del-Amor, I. Pérez-Hurtado, D. Orellana-Martín, and M. J. Pérez-Jiménez, “Adaptative parallel simulators for bioinspired computing models,” Future Generation Computer Systems, vol. 107, pp. 469–484, Jun. 2020, doi: 10.1016/j.future.2020.02.012.

10. S. McIntosh-Smith, J. Price, R. B. Sessions, and A. A. Ibarra, “High performance in silico virtual drug screening on many-core processors,” International Journal of High Performance Computing Applications, vol. 29, no. 2, pp. 119–134, May 2015, doi: 10.1177/1094342014528252.

11. P. Amar, M. Baillieul, D. Barth, B. LeCun, F. Quessette, and S. Vial, “Parallel Biological In Silico Simulation,” Nov. 2014, doi: 10.1007/978-3-319-09465-6\_40.

12. D. G. Whittaker, J. C. Hancox, and H. Zhang, “In Silico Assestment of Pharmacotherapy for Human Atrial Patho-Electrophysiology Associated With hERG-Linked Short QT Syndrome” Jan. 2019, doi: 10.3389/fphys.2018.01888.

13. T. O'Hara, L. Virág, A. Varró, and Y. Rudy, “Simulation of the Undiseased Human Cardiac Ventricular Action Potential: Model Formulation and Experimental Validation.” 2011, PLOS Computational Biology 7(5): e1002061. Available: https://doi.org/10.1371/journal.pcbi.1002061.

14. S. Dutta, K.C. Chang, K.A. Beattie, J. Sheng , P.N. Tran, W.W. Wu, M. Wu, D.G. Strauss, T. Colatsky, and Z. Li. "Optimization of an In silico Cardiac Cell Model for Proarrhythmia Risk Assessment." Front Physiol. Aug 2017 doi: 10.3389/fphys.2017.00616.

15. Tomek, Jakub et al. “Development, calibration, and validation of a novel human ventricular myocyte model in health, disease, and drug block.” eLife vol. 8 e48890. Dec 2019, doi:10.7554/eLife.48890.

16. “Parallel Thread Execution ISA Version 8.5” Accessed: Nov. 24, 2024. [Online]. Available: https://docs.nvidia.com/cuda/parallel-thread-execution/index.html.

17. C., Yves, C. D. Lontsi, and C. Pierre. "Rush-Larsen time-stepping methods of high order for stiff problems in cardiac electrophysiology." 2017, arXiv preprint arXiv:1712.02260.

18. G.R. Mirams, Y. Cui, A. Sher, M. Fink, J. Cooper, B. M. Heath, et al. "Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk." 2011, Cardiovasc. Res. 91 (1), 53–61. doi:10.1093/cvr/cvr044.