|  |
| --- |
|  |
| Thesis for Master of Engineering |
|  |
| Enhancing the efficiency of animal-alternative in-silico drug cardiotoxicity prediction through CUDA-based parallel processing |
| … 2024 |
|  |
| Graduate School  Kumoh National Institute of Technology |
|  |
| Department of IT Convergence Engineering |
|  |
| Iga Narendra Pramawijaya |
|  |

|  |
| --- |
| Thesis for Master of Engineering |
|  |
| Enhancing the efficiency of animal-alternative in-silico drug cardiotoxicity prediction through CUDA-based parallel processing |
| …. 2024 |
|  |
| Graduate School  Kumoh National Institute of Technology |
|  |
| Department of IT Convergence Engineering |
|  |
| Iga Narendra Pramawijaya |
|  |

|  |
| --- |
| Enhancing the efficiency of animal-alternative in-silico drug cardiotoxicity prediction through CUDA-based parallel processing |
|  |
| Supervisor Ki Moo Lim |
|  |
| This Thesis Presented for the Master of Engineering |
| … 2024 |
|  |
| Graduate School  Kumoh National Institute of Technology |
|  |
| Department of IT Convergence Engineering |
|  |
| Iga Narendra Pramawijaya |

|  |
| --- |
| Approval of the Thesis for the Master of Engineering Submitted by Ali Ikhsanul Qauli |
|  |
| December 2020 |
|  |
|  |
|  |
|  |
| C:\Users\Med IT\Pictures\Other\thesis\sign2.pngChair of Committee Kwang Soup Song (seal) |
|  |
| C:\Users\Med IT\Pictures\Other\thesis\sign5.pngCommittee Ki Moo Lim (seal) |
|  |
| Committee Hojong Choi (seal) |
|  |
| Graduate School  Kumoh National Institute of Technology |
|  |

|  |
| --- |
|  |
| Enhancing the efficiency of animal-alternative in-silico drug cardiotoxicity prediction through CUDA-based parallel processing |
|  |
| Iga Narendra Pramawijaya |
|  |
| Department of IT Convergence Engineering,  Graduate School  Kumoh National Institute of Technology |
|  |
| Abstract |
|  |

Introduction: The comprehensive *in vitro* proarrhythmia assay (CiPA) has opened broad opportunities to incorporate *in silico* experiment as an integral part of drug assessment. Following that, the initial assessment of the Hill coefficient and IC50, which are two important parameters describing drug cellular dynamics, can be included in the in silico experiment as well. The initial assessment consists of Hill fitting and parameters bootstrapping. In this work, I propose a new approach to deploy simple Monte Carlo (MC) simulation with Gaussian scatter incorporated into the calculation of bootstrap samples.

Method: The experimental dose-response data is processed into Hill fitting with the least square method (LSM) to obtain best-fit parameters. These fitted parameters combined with Gaussian scatter of blocking data will act as a data generator for MC simulation that will result in bootstrap samples of Hill coefficient and IC50. Finally, several aspects of those bootstrap samples are assessed such as the 95% confidence interval, the parameter distribution (in the form of the histogram), and finally sensitivity analysis by predicting blocking response from various drug concentrations.

Results: It is found that for insufficient dose-response data, both the proposed MC algorithm and the existing Markov chain Monte Carlo (MCMC) method can jump to the quite high value of uncertainties faster especially with observed mean blocking percentage below 60%. However, both algorithms can yield a similar maximum uncertainty profile except within a region of 15-25% minimum observed mean blocking where the MC method generated higher uncertainty than MCMC. In addition, I was able to show that for sufficient observed experimental data where allowed drug dosage is within this range, the proposed MC method match quite well with existing MCMC in term of distribution of fitted parameters.

Conclusion: The proposed MC method for bootstrapping Hill coefficient and IC50 is quite simpler to implement than the existing MCMC method as it adds only Gaussian sampling into the main Hill fitting algorithm. In addition to its simplicity, the proposed MC method has comparable results with the existing MCMC method in terms of fitted parameter distribution and uncertainty quantification.

|  |
| --- |
|  |

|  |
| --- |
| CUDA기반 병렬처리를 통한 동물대체 인실리코 약물 심독성 예측 효율성 증대 |
|  |
| Iga Narendra Pramawijaya |
|  |
| 금오공과대학교 대학원 IT융복합공학과 |
|  |
| 요 약 |
|  |

소개 : 종합 체외 부정맥 분석(CiPA)은 약물 평가의 필수 부분으로 In silico 컴퓨터 시뮬레이션을 통합할 수 있는 광범위한 기회를 열었다. 그에 따라 약물에 의한 세포 역학을 설명하는 두 가지 중요한 매개 변수인 Hill계수와 IC50의 초기 평가는 In silico 실험에 포함될 수 있다. 이 초기 평가는 Hill 피팅과 매개 변수의 부트스트랩과정으로 구성된다. 본 연구에서는 부트스트랩 샘플의 계산을 위해 가우스 산포가 포함된 몬테카를로(MC) 시뮬레이션을 사용하는 새로운 접근방식을 제안하였다.

방법 : 최소제곱법(LSM)을 사용한 힐 피팅을 통해 실험 복용량-반응 데이터로부터 최적의 매개변수를 생성하였다. 약물에 의한 차단 데이터의 가우스 산포와 결합된 피팅 된 매개변수는 MC 시뮬레이션을 위한 데이터 생성기로 사용되어 Hill계수와 IC50의 부트스트랩 샘플을 생성한다. 부트스트랩 된 샘플들은 95%의 신뢰구간에 분포하는 정도, 매개 변수의 분포(히스토그램의 형태) 및 다양한 약물 농도에 따른 차단 반응의 예측을 통한 민감도 분석을 통해 평가되었다.

결과 : 복용량-반응 데이터가 충분하지 않은 경우 MC 알고리즘과 Markov Chain Monte Carlo (MCMC) 알고리즘 모두에서 매우 높은 불확실성 수치를 생성하였으며, 이는 특히 측정된 평균 차단 비율이 60% 미만인 경우에서 두드러지게 관찰되었다. 하지만, 두 알고리즘 모두 유사한 최대 불확실성 프로파일을 생성할 수 있었으며 최소 측정된 평균 차단 정도가 15~25%인 영역에서는 MC 방법이 MCMC 방법보다 더 높은 불확실성을 생성하였다. 또한, 허용된 약물 복용량이 이 범위 내에서 측정된 실험 데이터가 충분한 경우 연구에서 제안한 MC 방법이 피팅 된 매개 변수의 분포 측면에서 기존의 MCMC 방법과 매우 잘 일치 하는 것을 보여주었다.

결론 : Hill 계수와 IC50을 부트스트랩하기 위해 연구에서 제안한 MC 방법은 주 Hill 피팅 알고리즘에 가우시안 샘플링 과정이 추가되기 때문에 기존의 MCMC 방법에 비해 구현이 쉽다. 이러한 단순성 외에도 제안된 MC방법은 적합 매개 변수의 분포 및 불확실성의 정량화 측면에서 MCMC 방법과 유사한 결과를 낼 수 있었다.

**Contents**

[[List of Figures] ii](#_Toc182812268)

[[List of Tables] iii](#_Toc182812269)

[[Glossary] iv](#_Toc182812270)

[[Acknowledgement] v](#_Toc182812271)

[Chapter 1. Introduction 1](#_Toc182812272)

[1.1 In-Silico Electrophysiology Simulation 1](#_Toc182812273)

[1.2 Parallel Computing 2](#_Toc182812274)

[1.2.1 Central Processing Unit (CPU) for Parallel Computing 2](#_Toc182812275)

[1.2.2 Graphics Processing Unit (GPU) for Parallel Computing 3](#_Toc182812276)

[1.2.3 CUDA 3](#_Toc182812277)

[1.2.4 CellML 3](#_Toc182812278)

[1.3 Previous Study 4](#_Toc182812279)

[1.4 Objectives 5](#_Toc182812280)

[Chapter 2. Methodologies 5](#_Toc182812281)

[2.1 Generating C Code from CellML 5](#_Toc182812282)

[2.2 GPU Memory Adjustment and Offseting 6](#_Toc182812283)

[2.3 Solving Ordinary Differential Equations 8](#_Toc182812284)

[2.4 Output Format 8](#_Toc182812285)

[2.5 Compilation, Input Files, and Testing 8](#_Toc182812286)

[Chapter 3. Results and Discussion 9](#_Toc182812287)

[3.1 Best-fit parameters and their confidence intervals 9](#_Toc182812288)

[3.2 Uncertainty of blocking response 13](#_Toc182812289)

[Chapter 4. Conclusion and Limitation 20](#_Toc182812290)

[[References] 21](#_Toc182812291)

[Appendix 23](#_Toc182812292)

[A. Code for Data processing 23](#_Toc182812293)

[B. Code for linear and nonlinear LSM 25](#_Toc182812294)

[C. Code for Monte Carlo simulation 28](#_Toc182812295)

[D. Complete code for Hill fitting with LSM and Monte Carlo simulation 29](#_Toc182812296)

# [List of Figures]

**No table of figures entries found.**

**No table of figures entries found.**

[[Figure 2.1] OpenCOR interface when selecting Tomek model and converting it to C codes on MacOS. 6](#_Toc182814237)

[[Figure 3. 1] Comparison of the results from scheme 1 (left) and scheme 2 (right) on Dofetilide hERG channel 12](#_Toc182814238)

[[Figure 3. 2] Comparison of the results from scheme 1 (left) and scheme 2 (right) on Dofetilide INa channel 13](#_Toc182814239)

[[Figure 3. 3] The comparison of Monte Carlo simulation of scheme 1 (left) and scheme 2 (right). Note that the grey shaded area is the result of MCMC as a reference. Also note that means the blocking response of 2.5 to 97.5 quantiles of the marginal distribution, while for 5 to 95 quantiles. 14](#_Toc182814240)

[[Figure 3. 4] Uncertainties of blocking response from scheme 1 as a function of observed mean blocking percentage at the highest experimental drug concentration (). Please note that the dashed vertical line means 60% of observed mean blocking at . 15](#_Toc182814241)

[[Figure 3. 5] Uncertainties of blocking response from scheme 2 as a function of observed mean blocking percentage at the highest experimental drug concentration (). The vertical dashed line represents the same meaning as in [Figure 3. 4]. 16](#_Toc182814242)

[[Figure 3. 6] The maximum uncertainty as a function of the minimum observed average blocking response. Left panel from scheme 1 and the right one from scheme 2. Please note that the dashed line represents the result of MCMC. 17](#_Toc182814243)

[[Figure 3. 7] The blocking response distribution at a drug concentration of from scheme 1 (left) and scheme 2 (right). axis is the relative frequency and axis is the blocking percentage. Please note that the middle and lower panels have a secondary vertical axis for the results of MCMC. 18](#_Toc182814244)

# [List of Tables]

[[Table 2. 1] 12 CiPA training drugs and its TdP risk classification 11](#_Toc57331937)

[[Table 3. 1] Hill equation parameters and confidence intervals (square bracket) for the 12 CiPA training drugs with scheme 1 approach from [Figure 2. 1] 18](#_Toc57331935)

[[Table 3. 2] Hill equation parameters and confidence intervals (square bracket) for the 12 CiPA training drugs with scheme 2 approach from [Figure 2. 1] 19](#_Toc57331936)

# [Glossary]

|  |  |
| --- | --- |
| TdP | Torsade de pointes |
| CiPA | Comprehensive *in vitro* Proarrhythmia Assay |
| ECG | Electrocardiogram |
| UQ | Uncertainty quantification |
| MC | Monte Carlo |
| MCMC | Markov chain Monte Carlo |
| DRAM | Delaying rejection adaptive Metropolis |
| AARJ | Adaptive automatic reversible jump |
| LSM | Least square method |
| LM | Lavenberg-Marquardt |
| GSL | GNU Scientific library |
| N/A | Not applicable |
| hERG | Human Ether-à-go-go-Related Gene |
| ICaL | L-type calcium current |
| IK1 | Inward rectifier potassium current |
| IKs | Slow delayed rectifier potassium current |
| Ito | Transient outward current |
| INaL | Late sodium current |
| INa | Sodium current |
|  |  |

# [Acknowledgement]

In the name of God, The Most Gracious and The Most Merciful. Alhamdulillah, all praises to God. This thesis would not have been completed without God’s will and guidance. Also, may blessing and peace always be upon Prophet Muhammad, his family, his companion, and his followers all over the world.

First of all, I would like to express my sincere gratitude to my supervisor Professor Ki Moo Lim. Thank you for allowing me to be part of the research member of the Computational Medicine Lab (CML). Your valuable guides and support really help me excel in many aspects I am good at during my study and research in CML. Also, I would like to thank the referee Professor Kwang Soup Song, and Professor Ho Jong Choi for their time, support, and constructive comments.

I would also like to extend my gratitude to all lab mates that I have ever met and collaborated with; Aulia, Getu, Abebe, 다운,하영, Marcell,준효, and 에담. Thank you for all support and advice that had been given to me. I do not think I can survive during my study and research without your help, guys.

For all Indonesian students that I have ever met and collaborated with during my study in Korea, especially those who stayed in Shinpyeong; Aulia, Rahma, Salwa, Naufal, other students from KIT, also from Indonesian Students Association in South Korea (Perpika); Anafi, CQ, and other students from various campuses and cities in Korea that I cannot include here one by one, thank you to be important parts of my life. You really help me excel to become a better human being as a whole. Take care and stay healthy, guys.

Lastly, I wish to express my gratitude to my family, my dad, mom, brother, and sisters. They have always unconditionally supported me to be as best as I can be. Thank you for all that support. May The Almighty God grants you the best life in this small, temporary world and in the eternal afterlife.

# Introduction

Cardiovascular diseases are the leading global causes of death, which emphasizes the importance of effective methods for drug discovery. Traditionally, drug cardiotoxicity prediction is achieved using animal testing, which is controversial and has several drawbacks. Modern in-silico or computer-based methods for drug cardiotoxicity prediction show promising results as an animal-alternative alternative. Nevertheless, some of them are 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 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], our method significantly accelerates overall computational process, enabling faster handling of large-scale simulations. By leveraging the power of parallel processing, this approach not only enhances the in-silico simulation but also ensures that drug toxicity evaluations are both more practical and accurate, paving the way for broader and more ethical applications in real-world drug testing.

## In-Silico Electrophysiology Simulation

Biological *In-Silico* simulation is a field in computational biology in means of using the aid of computing devices to conduct mathematical calculations that accurately simulates cardiac responses within different conditions. Electrophysiology is a study of electrical activity in the heart, and it can be explained using mathematical model in form of ordinary differential equations (ODE). By studying how cardiac cell responses through its electrical activity, we can explain various phenomenon in detail such as cardiovasicular disesases, how effective a drug is, and how toxic a drug that is not meant for the heart cell is, to the heart cell. Furthermore the toxicity of a drug to the heart cell will be mentioned as cardiotoxicity. Simpy, *in-silico* electrophysiology simulation is a powerful computational tool used to model and study the electrical activity of the heart. These simulations provide valuable insights into cardiac function, disease mechanisms, potential treatments and what might harmful to cardiac function with a minimum invasive approach to collect data.

## Parallel Computing

Over recent decades, parallel computation has promised to accelerate overall computation speed. Parallel computing, from a technical standpoint, means performing many calculations simultaneously, based on the principle that large problems can often be divided into smaller tasks that can be processed at the same time. For programmers, the main challenge is how to allocate these concurrent tasks across multiple computing resources such as cores or even computers.

Parallel computing has both hardware and software requirements that are deeply interlinked. Computer hardware architechture must ensure it has more than one computing core, while parallel programming designs code utilizing more than one computing core. The hardware aspect of computer architecture supports parallelism by providing an infrastructure that can handle multiple, simultaneous processes or threads. Meanwhile, parallel programming focuses on efficiently using this hardware to perform tasks concurrently. This programming paradigm involves mapping tasks to these available cores to achieve simultaneous execution, ensure every core runs in harmony, and arranging output from each cores. When writing non-parallel programs, understanding the computer architecture is less crucial. However, in multi-core programming, a solid understanding of multicore architectures becomes essential for developing efficient and correct parallel programs.

### Central Processing Unit (CPU) for Parallel Computing

For decades, one key method of improving consumer computing performance was to increase the CPU’s clock speed. However, due to power, heat, and physical limitations in transistor miniaturization, this approach has reached its limits. As a result, manufacturers have shifted their focus from boosting clock speed to increasing the number of cores per CPU, inspired by supercomputers that achieve high performance by using large numbers of processors. Rather than relying solely on single-core performance, adding multiple cores also allows personal computers to improve processing power without clock speed increases. Widely-known standards to do parallel processing with a GPU is to use Open Multi-Processing (OpenMP) or the Message Passing Interface (MPI).

OpenMP (Open Multi-Processing) is a programming model designed for parallel computing on shared-memory architectures, typically used to exploit multicore CPUs. It enables developers to add parallelism to existing C, C++, and Fortran code using simple compiler directives (pragmas). OpenMP divides tasks across multiple threads that share the same memory space, allowing for straightforward parallelization of loops and sections of code with minimal modifications. OpenMP is relatively easy to implement, making it an excellent choice for applications requiring high performance on a single multicore processor, where shared memory among threads simplifies data access.

MPI (Message Passing Interface), on the other hand, is a standard used for parallel computing on distributed-memory systems, such as clusters or supercomputers. MPI allows multiple processes running on separate memory spaces to communicate by sending and receiving messages, making it suitable for applications where tasks need to run on different nodes in a network. This model is more complex than OpenMP, as it requires explicit data sharing, but it offers greater flexibility and scalability, allowing parallelization across a large number of processors. MPI is ideal for high-performance computing tasks that involve significant data exchange across nodes, making it the go-to solution for distributed systems.

### Graphics Processing Unit (GPU) for Parallel Computing

Graphics processing unit initially designed to compute graphical calcucations that is repetitive, relatively more simple compared to what CPU calculates, but quantitatively much more calculations compared to CPU. Unlike CPUs, which are optimized for a wide variety of tasks and tend to have a smaller number of powerful cores, GPUs have thousands of smaller cores designed for high-throughput parallelism. This architecture allows GPUs to perform many calculations simultaneously, making them highly efficient for tasks that can be broken down into smaller, identical operations.

Beside of graphics computing, GPUs are also able to accelerate other computing purposes such as scientific simulation and machine learning. This demand creates new sub-field in computer programming, called GP-GPU programming, that stands for general-purposed graphics processing unit programming. Its high-throughput parallelism makes GPU suitable for scientific calculation that has massive datasets or extensive matrix calculation. With frameworks like CUDA and OpenCL, programmers can leverage GPU architectures to perform GP-GPU.

### CUDA

CUDA, short for Compute Unified Device Architecture, is a parallel computing platform and programming model developed by NVIDIA. It enables developers to harness the immense computational power of NVIDIA GPUs for general-purpose processing tasks. CUDA is built on the foundation of extending standard C/C++ programming with GPU-specific features, making it accessible for developers already familiar with these languages. It provides APIs and tools that allow fine-grained control over GPU resources, enabling efficient parallel execution of computationally intensive tasks across thousands of GPU cores.

The computational model of CUDA is structured around a hierarchical organization of cores, blocks, and threads. At the highest level, the GPU consists of multiple streaming multiprocessors (SMs), each containing numerous CUDA cores. These cores execute the smallest unit of work in CUDA, referred to as a thread. Threads are grouped into blocks, which can contain hundreds or thousands of threads, depending on the GPU's architecture. Blocks are further organized into a grid, creating a hierarchy that allows the distribution of computational tasks across the entire GPU.

This hierarchy is essential for mapping complex problems onto the GPU efficiently. Developers can define the number of threads per block and the number of blocks per grid based on the problem's computational requirements and the GPU's hardware limitations. Each thread has a unique thread ID within its block, and each block has a unique block ID within the grid. Using these IDs, CUDA programs can assign specific tasks or data portions to individual threads, ensuring even workload distribution.

The flexibility of CUDA's core-block-thread hierarchy enables scalability and efficiency. By tuning the grid and block dimensions, developers can optimize memory usage and parallelism for diverse applications, from image processing to numerical simulations. This design, combined with CUDA's extensive library ecosystem and advanced debugging tools, makes it a powerful platform for unlocking the full potential of GPU-based computing.

### CellML

CellML is an XML-based language created to represent mathematical models in a platform-independent format, facilitating model sharing between researchers and secure archival in repositories. This standardization in a machine-readable form is essential in bioinformatics, as it enhances scientific accuracy, accelerates model development, and enables the integration of multiple models into complex, combined systems. CellML supports collaboration by allowing models to be easily exchanged and archived. Several public databases host extensive collections of CellML models, with the CellML Model Repository being one of the most prominent. Additionally, the BioModels database includes models converted from the Systems Biology Markup Language (SBML) into CellML, broadening accessibility and compatibility for researchers using these bioinformatics resources. [[cite](https://doi.org/10.1098%2Frsta.2008.0094)] [[cite](https://doi.org/10.1093%2Fbioinformatics%2Fbtn390)] [[cite](http://scholar.google.com/scholar_lookup?&title=BioModels%20Database%3A%20a%20free%2C%20centralized%20database%20of%20curated%2C%20published%2C%20quantitative%20kinetic%20models%20of%20biochemical%20and%20cellular%20systems&publication_year=2006&author=Le%20Novere%2CN&author=Bornstein%2CB&author=Broicher%2CA&author=Courtot%2CM&author=Donizelli%2CM&author=Dharuri%2CH&author=Li%2CL&author=Sauro%2CH&author=Schilstra%2CM&author=Shapiro%2CB)]

## Previous Study

Parallelisation in computational biology is not an entirely new concept. The Cells in Silico (CiS) framework presented by Berghoff et al. (2020) [[cite](https://link.springer.com/article/10.1186/s12859-020-03728-7)] offers a tool for simulating the growth and development of biological tissues. The modular and parallel design of CiS allows for flexible configuration of different model assumptions, making it applicable to a wide range of research questions. As demonstrated by the example of a 10003 voxel-sized cancerous tissue simulation at sub-cellular resolution, CiS can be used to explore complex biological processes at a high level of detail.

Utilisation of GPU in biological cell computing has been explored in previous researches. One of them is from Miguel, et al [[cite](https://www.sciencedirect.com/science/article/abs/pii/S0167739X19308817)] in 2020. Miguel, et al. explored an adaptive parallel simulator to solve performance loss in massive parallel membrane computing devices known as membrane systems or P systems. The paper demonstrates the effectiveness of this approach by extending an existing simulator for Population Dynamics P systems. Experimental results show that this adaptive simulation can significantly improve performance, up to 2.5x on both GPUs and multicore processors.

Related to drug toxicity and discovery, other researchers tried to approach and optimise drug development process using parallel computing approach as well. Previously, McIntosh-Smith et, al. developed a in-silico drug screening method on multiple core processors. McIntosh-Smith et, al. developed BUDE (Bristol University Docking Engine), a drug discovery tool, simulating molecular docking. To speed up calculations on powerful processors with multiple cores, BUDE has been adapted to work with OpenCL, a common language for parallel programming [[cite](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4425459/)]. As a result, McIntosh-Smith et, al. achieved of 46% at peak, or 1.43 TFLOP/s on a single Nvidia GTX 680.

Barth et, al. developed a parallelisation on biochemical simulation of metabolic pathways in their high level computational simulation. This method allows Barth et, al. to run simulations with more complex models, featuring a greater number of chemicals and reactions. Hence, Barth et, al. can achieve more realistic, lifelike outcomes while using less computing time [[cite](https://www.researchgate.net/publication/281886386_Parallel_Biological_In_Silico_Simulation)].

## Objectives

The research group of Computational Medicine Laboratory (CML) in Kumoh National Institute of Technology (KIT) has proposed computational simulation for drug assessment as follow:

# Methodologies

Throughout this chapter, the author explains how to create sample-based parallelisation cardiac cell simulation platform on GPU, and parallelise CellML based *in-silico­* cell model. Each subchapter will explain in detail development and code conversion process in this research. At the end, a software engineer or other researcher can replicate the parallelisation concept and utilise this base to more multi-sample *in-silico* simulation.

## Generating C Code from CellML

## 

In order to generate codes in various programming language based on CellML’s XML, we can use various third party libraries or application. One of the most popular application is OpenCOR. OpenCOR is a versatile software tool designed for modeling and simulation of biological processes, including those described in CellML. It provides functionality to parse CellML models and convert them into executable code in various programming languages, including C. In addition to generating C code, OpenCOR supports model editing, running simulation, and analysis, making it a comprehensive platform for working with CellML models. Its extensibility and integration with other tools, such as Python scripting, further enhance its utility for researchers. By leveraging OpenCOR, users can streamline the process of implementing CellML models into broader computational workflows, such as those involving high-performance computing or in-silico drug testing. Results from this GPU modification will also validated using OpenCOR’s result.

This research involves three different cell models. O’Hara-Rudy 2011, 2017 and Tomek. After installing OpenCOR, open the application and search for these three cell models in the search bar on top left corner (PMR). Select file with .cellml extension. Select tools, and export the CellML file to C code. Figure 2.1 shows the OpenCOR GUI on MacOS.



[Figure 2.1] OpenCOR interface when selecting Tomek model and converting it to C codes on MacOS.

Having the C code of the cell model is important because CUDA programming uses .cu format, that is similar to C. CUDA is built upon C/C++, extending it for GPU programming. Both use C syntax as their base language, many fundamental syntax elements like loops, conditionals, functions, and function definitions are similar in both CUDA and C. CUDA programs include host code that runs on the CPU, which is essentially standard C/C++ code. For programmers familiar with C/C++, CUDA maintains a relatively low learning curve by building upon familiar concepts, only adding memory allocation and parallel processing paradigms.

## GPU Memory Adjustment and Offseting

Efficient memory management is critical in GPU programming, as the performance of a 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. The GPU has several memory types, including global, shared, constant, and local memory, each with unique characteristics and access speeds. Proper adjustment and offsetting of these memory types can significantly enhance computational efficiency.

Memory adjustment involves allocating and aligning data in a way that minimizes access latency and maximizes memory throughput. For instance, global memory, though large, has higher access latency and benefits from coalesced memory access patterns, where threads access memory in a sequential and aligned manner. In contrast, shared memory, which resides on-chip, is much faster but limited in size, making it ideal for frequently accessed data within thread blocks.

Offsetting refers to managing data indexing to optimize memory access patterns. For example, ensuring that thread indices correspond to contiguous memory addresses 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, offseting implemented to simplify any multi-dimentional input. In order to enhance the efficiency of drug cardiotoxicity prediction, utilising multi-sample simulation is crucial. Cell model code from OpenCOR is only designed for single sample simulation, then conversion should be applied. On the other side, simplicity from the OpenCOR should be maintained. Also in the previous iteration of drug toxicity *in-silico* simulation based on CPU, it uses 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 dimentional vector used in the previous iteration into 1 dimentional (1D) array. Offsetting is mainly used for pointing the correct data in a flattened 1D array.

Since CUDA natively operates on linear memory, storing data in a flattened 1D format aligns well with the GPU's memory architecture, allowing for optimal performance while maintaining simplicity. In this method, a multidimensional array is represented as a single contiguous block of memory, and elements are accessed using calculated indices. For example, a 2D array with dimensions (rows, columns) can be indexed as index = (current row \* row size) + columns. This linearization simplifies memory allocation and transfer between the host and device, ensuring compatibility with CUDA's memory management functions. Figure 2.2 explains graphically the flattening process and how multi-dimentional vector data differentiate samples (called as sample id) in the previous iteration of drug toxicity *in-silico* simulation.



[Figure 2.2] Main difference in values storing paradigm after CUDA-parallelisation, assuming column size is 13.

The previous iteration of the code uses row to indicates different samples. 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 has 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 dimention, sample\_id and the number of specific column we want to select.

In the code from OpenCOR, most of the constants declarations, states calculations, rates calculations and algebraic formulations will be delivered in this form:

CONSTANTS[5] = 8314;

Or, for algebraic formulation

ALGEBRAIC[3] = 1.00000/(1.00000+exp((STATES[0]+87.6100)/7.48800));

At the top of the code from OpenCOR, we can see how many CONSTANTS, STATES (similar to number of RATES), and ALGEBRAIC are there in the code. It looks like:

There are a total of 223 entries in the algebraic variable array.

There are a total of 43 entries in each of the rate and state variable arrays.

There are a total of 163 entries in the constant variable array.

These are going to be the row size, while number of sample will be the current row. Number of sample will be declared later in this chapter, we will name it ‘sample\_id’.

Apply offsetting in every CONSTANTS, STATES (similar to RATES), and ALGEBRAIC array occur in the code by applying this to every array index:

new\_index = (sample\_id \* row size) + columns.

Hence, all of the declarations and calculations in the code should look like:

CONSTANTS[(sample\_id \* 163) + 5] = 8314;

Or, for algebraic formulation

ALGEBRAIC[(sample\_id \* 223) + 3] = 1.00000/(1.00000+exp((STATES[(sample\_id \* 43) + 0]+87.6100)/7.48800));

Notice that the calculation for new index of each arrays are calculated within the declaration of each elements, and all arrays are treated the similar way.

## Solving Ordinary Differential Equations

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 behaviors 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.

For the ORd 2011 model, the Rush-Larsen method was utilized 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. Initially, we aimed to generalize the use of the Rush-Larsen method across all cell models. However, when applied to ORd 2017 and Tomek models, this approach exhibited instability, failing to provide reliable results. To address this limitation, a simple forward Euler method was implemented for these models. While the Euler method produced accurate and stable results, particularly for the ORd 2017 and Tomek models, it proved to be computationally intensive, significantly increasing the runtime.

To optimize the parallelization process, we focused on simplifying the algorithm by enabling parallel threads to process multiple samples rather than multiple equations simultaneously. The Forward Euler solver also can be upgraded with a dynamic time-stepping mechanism, by adjusting the time step during each iteration. 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.

This research implements the ODE solver inside the cellmodel code as a function. 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:

*xn+1=xn+rate(xn)⋅Δt*

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 we are working now, add a function to implement this calculation. This can be achieved by implementing for loop such as:

void solveEuler( double \*STATES, double \*RATES, double dt, int sample\_id)

{

for(int i=0;i<43;i++){

STATES[(43 \* sample\_id) + i] = STATES[(43 \* sample\_id) + i] + RATES[(43 \* sample\_id) + i] \* dt;

}

}

The function uses a for loop to iterate over the 43 state variables associated with a single sample in Tomek cell model. For each state variable, it calculates the new value using the forward Euler method formula. The formula is applied to the corresponding state variable and rate of the selected sample, determined by the sample\_id. By multiplying sample\_id by 43 (the number of state variables per sample in Tomek cell model), the function accesses the correct block of memory in the flattened 1D array for both STATES and RATES. This ensures that updates are sample-specific and do not interfere with other samples.

## Repository Structure and Code Organisation

## Output Format

## Compilation, Input Files, and Testing

# Results and Discussion

## Best-fit parameters and their confidence intervals

The outcomes from LSM fitting will be the optimum or best-fit parameters of and . It is found that the best-fit parameters from Hill fitting with LSM match exactly with previous work by [15]. Meanwhile, the 95% confidence interval of the parameters from two schemes shows different results among them that can be seen in [Table 3. 1] and [Table 3. 2]. In each table, the N/A or “not applicable” means the fitted is not well defined as no detectable blocking appeared during experiment as stated by Li et al [22].

[Table 3. ] Hill equation parameters and confidence intervals (square bracket) for the 12 CiPA training drugs with scheme 1 approach from [Figure 2. 1]

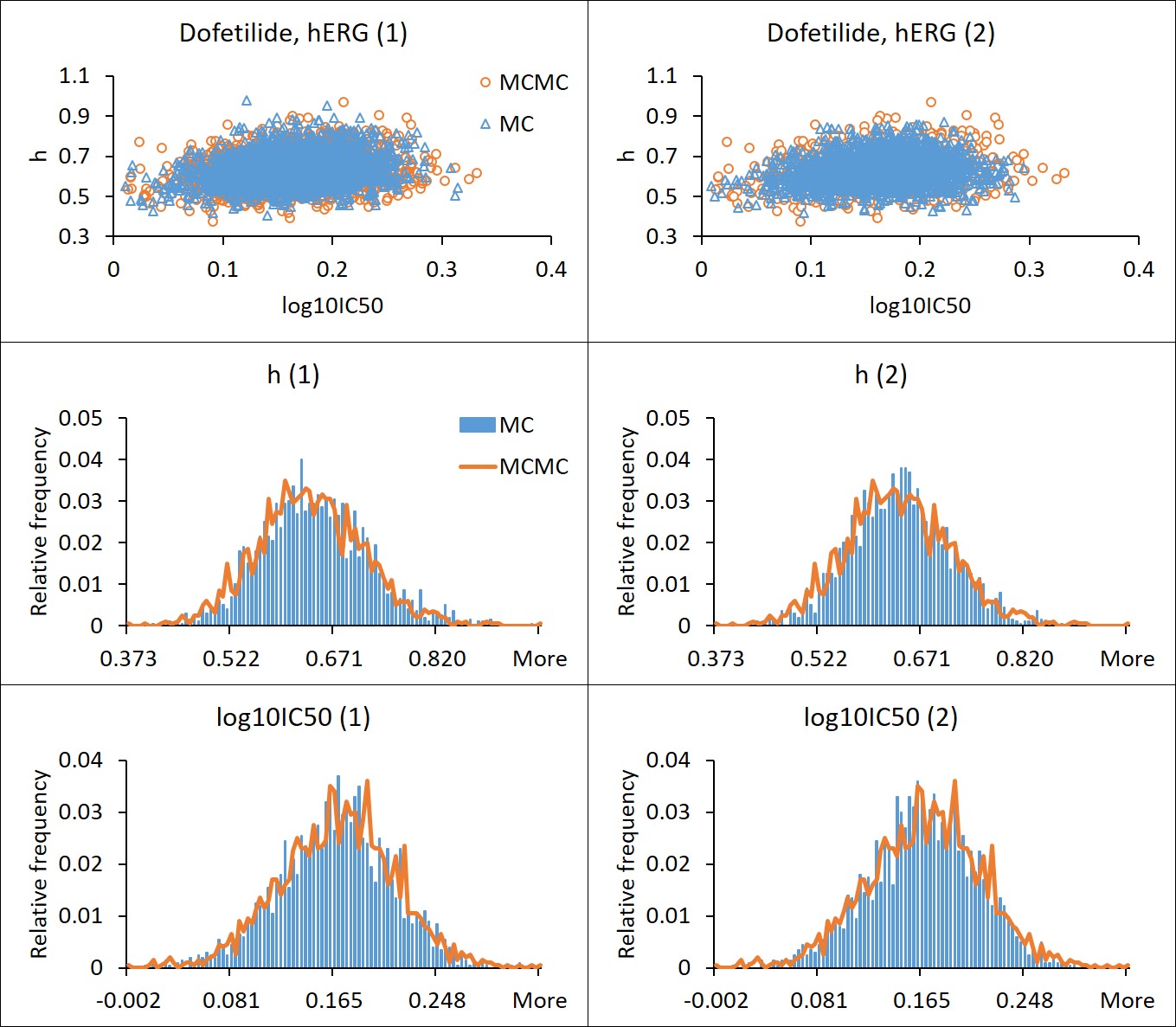
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dofetilide | | Cisapride | |
|  | log10IC50 | h | log10IC50 | h |
| ICaL | 2.42[1.80,10.0] | 1.16[0.21,1.70] | 6.97[4.69,10.0] | 0.43[0.24,0.70] |
| IK1 | 2.60[1.47,10.0] | 0.77[0.15,1.69] | 4.47[3.33,10.0] | 0.51[0.15,0.81] |
| IKs | N/A | N/A | 7.91[4.63,10.0] | 0.29[0.20,0.60] |
| Ito | 1.27[1.04,1.79] | 0.77[0.47,1.11] | 5.34[3.55,10.0] | 0.24[0.11,0.46] |
| INaL | 5.88[1.81,10.0] | 0.26[0.13,1.05] | N/A | N/A |
| INa | 2.58[1.63,10.0] | 0.90[0.19,1.76] | N/A | N/A |
| hERG | 0.16[0.09,0.23] | 0.63[0.52,0.77] | 1.07[1.02,1.12] | 1.34[1.17,1.55] |
|  | **Bepridil** | | **Verapamil** | |
|  | log10IC50 | h | log10IC50 | h |
| ICaL | 3.45[3.30,3.67] | 0.65[0.47,0.85] | 2.30[2.23,2.38] | 1.10[0.88,1.35] |
| IK1 | N/A | N/A | 8.54[4.55,10.0] | 0.27[0.19,0.80] |
| IKs | 4.46[4.09,5.47] | 0.71[0.39,1.04] | N/A | N/A |
| Ito | 3.93[3.63,10.0] | 3.54[0.26,10.0] | 4.13[3.58,5.83] | 0.82[0.37,1.48] |
| INaL | 3.26[3.20,3.31] | 1.42[1.15,1.70] | 3.85[3.26,10.0] | 1.03[0.15,2.73] |
| INa | 3.47[3.39,3.58] | 1.16[0.84,1.51] | N/A | N/A |
| hERG | 2.17[1.96,2.36] | 0.88[0.61,1.33] | 2.70[2.64,2.75] | 1.10[0.93,1.28] |
|  | **Terfenadine** | | **Ranolazine** | |
|  | log10IC50 | h | log10IC50 | h |
| ICaL | 2.85[2.79,2.90] | 0.66[0.60,0.72] | N/A | N/A |
| IK1 | N/A | N/A | N/A | N/A |
| IKs | 5.60[4.27,10.0] | 0.54[0.21,1.03] | 7.56[5.96,10.0] | 0.52[0.29,0.96] |
| Ito | 5.38[4.30,9.01] | 0.26[0.12,0.39] | N/A | N/A |
| INaL | 4.30[3.83,5.50] | 0.60[0.35,0.89] | 3.90[3.80,3.99] | 0.94[0.76,1.16] |
| INa | 3.68[3.26,5.06] | 1.02[0.39,2.14] | 4.84[4.62,5.67] | 1.42[0.60,2.40] |
| hERG | 1.27[1.08,1.45] | 0.59[0.45,0.75] | 3.81[3.71,3.92] | 0.84[0.66,1.02] |
|  | **Sotalol** | | **Mexiletine** | |
|  | log10IC50 | h | log10IC50 | h |
| ICaL | 6.85[6.68,7.17] | 0.87[0.61,1.14] | 4.58[4.37,5.00] | 1.03[0.67,1.48] |
| IK1 | 6.48[6.41,6.59] | 1.20[0.93,1.53] | N/A | N/A |
| IKs | 6.63[6.51,6.84] | 1.17[0.80,1.58] | N/A | N/A |
| Ito | 7.63[6.95,9.98] | 0.66[0.28,1.29] | N/A | N/A |
| INaL | N/A | N/A | 3.95[3.89,4.03] | 1.41[1.06,1.80] |
| INa | 9.06[7.52,10.0] | 0.51[0.35,1.11] | N/A | N/A |
| hERG | 4.94[4.78,5.08] | 0.94[0.73,1.23] | 4.39[4.20,5.05] | 2.59[1.03,4.94] |
|  | **Quinidine** | | **Ondansetron** | |
|  | log10IC50 | h | log10IC50 | h |
| ICaL | 4.71[4.48,5.04] | 0.59[0.47,0.73] | 4.35[4.23,4.54] | 0.75[0.57,0.95] |
| IK1 | 7.60[4.88,10.0] | 0.35[0.19,0.91] | N/A | N/A |
| IKs | 3.69[3.63,3.78] | 1.36[1.00,1.76] | 5.76[5.29,6.83] | 0.65[0.40,0.93] |
| Ito | 3.54[3.50,3.59] | 1.28[1.09,1.51] | 6.01[5.19,10.0] | 0.99[0.27,1.67] |
| INaL | 3.97[3.91,4,08] | 1.34[1.06,1.63] | 4.28[4.21,4.37] | 1.03[0.82,1.25] |
| INa | 4.09[3.97,4.33] | 1.49[0.99,2.08] | 4.76[4.51,6.35] | 1.02[0.31,1.77] |
| hERG | 2.54[2.40,2.63] | 1.03[0.78,1.33] | 3.17[3.03,3.31] | 0.96[0.71,1.29] |
|  | **Diltiazem** | | **Chlorpromazine** | |
|  | log10IC50 | h | log10IC50 | h |
| ICaL | 2.05[1.91,2.19] | 0.71[0.55,0.92] | 3.91[3.80,4.04] | 0.84[0.63,1.05] |
| IK1 | N/A | N/A | 3.97[3.84,4.14] | 0.69[0.53,0.85] |
| IKs | N/A | N/A | N/A | N/A |
| Ito | 9.45[6.89,10.0] | 0.17[0.14,0.29] | 7.25[5.55,10.0] | 0.37[0.20,0.68] |
| INaL | 4.34[4.24,4.46] | 0.68[0.56,0.81] | 3.66[3.59,3.73] | 0.94[0.79,1.09] |
| INa | 5.04[4.76,6.07] | 0.70[0.37,0.94] | 3.66[3.58,3.73] | 2.00[1.65,2.46] |
| hERG | 3.82[3.73,3.92] | 0.79[0.67,0.91] | 3.05[2.94,3.16] | 0.89[0.71,1.09] |

[Table 3. ] Hill equation parameters and confidence intervals (square bracket) for the 12 CiPA training drugs with scheme 2 approach from [Figure 2. 1]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dofetilide | | Cisapride | |
|  | log10IC50 | H | log10IC50 | h |
| ICaL | 2.42[1.90,10.0] | 1.16[0.21,1.52] | 6.97[5.37,10.0] | 0.43[0.23,0.56] |
| IK1 | 2.60[1.72,10.0] | 0.77[0.14,1.19] | 4.47[3.69,10.0] | 0.51[0.15,0.62] |
| IKs | N/A | N/A | 7.91[5.30,10.0] | 0.29[0.19,0.47] |
| Ito | 1.27[1.06,1.73] | 0.77[0.49,1.03] | 5.34[3.78,10.0] | 0.24[0.11,0.40] |
| INaL | 5.88[1.96,10.0] | 0.26[0.13,0.90] | N/A | N/A |
| INa | 2.58[1.78,10.0] | 0.89[0.19,1.47] | N/A | N/A |
| hERG | 0.16[0.09,0.23] | 0.63[0.52,0.76] | 1.07[1.02,1.12] | 1.34[1.17,1.54] |
|  | **Bepridil** | | **Verapamil** | |
|  | log10IC50 | H | log10IC50 | h |
| ICaL | 3.45[3.31,3.67] | 0.65[0.47,0.81] | 2.30[2.22,2.39] | 1.10[0.88,1.37] |
| IK1 | N/A | N/A | 8.54[4.60,10.0] | 0.27[0.18,0.75] |
| IKs | 4.46[4.20,5.34] | 0.71[0.41,0.88] | N/A | N/A |
| Ito | 3.93[3.63,10.0] | 3.54[0.26,10.0] | 4.13[3.80,5.52] | 0.82[0.40,1.08] |
| INaL | 3.29[3.20,3.31] | 1.42[1.14,1.68] | 3.85[3.27,10.0] | 1.03[0.15,2.70] |
| INa | 3.47[3.39,3.59] | 1.16[0.78,1.63] | N/A | N/A |
| hERG | 2.17[1.97,2.36] | 0.88[0.61,1.29] | 2.70[2.64,2.75] | 1.10[0.90,1.34] |
|  | **Terfenadine** | | **Ranolazine** | |
|  | log10IC50 | H | log10IC50 | h |
| ICaL | 2.85[2.80,2.91] | 0.66[0.60,0.72] | N/A | N/A |
| IK1 | N/A | N/A | N/A | N/A |
| IKs | 5.60[4.60,10.0] | 0.54[0.21,0.83] | 7.56[6.37,10.0] | 0.52[0.29,0.76] |
| Ito | 5.38[3.96,10.0] | 0.26[0.11,0.49] | NA | NA |
| INaL | 4.30[3.91,5.86] | 0.60[0.31,0.79] | 3.90[3.80,4.00] | 0.94[0.76,1.14] |
| INa | 3.68[3.36,5.13] | 1.02[0.38,1.63] | 4.84[4.62,5.68] | 1.42[0.60,2.32] |
| hERG | 1.27[1.08,1.45] | 0.59[0.45,0.81] | 3.81[3.70,3.92] | 0.84[0.68,0.98] |
|  | **Sotalol** | | **Mexiletine** | |
|  | log10IC50 | H | log10IC50 | h |
| ICaL | 6.85[6.68,7.18] | 0.87[0.60,1.16] | 4.58[4.42,4.98] | 1.03[0.68,1.31] |
| IK1 | 6.48[6.42,6.59] | 1.20[0.93,1.45] | N/A | N/A |
| IKs | 6.63[6.51,6.84] | 1.17[0.82,1.52] | N/A | N/A |
| Ito | 7.63[7.29,9.63] | 0.66[0.30,0.82] | N/A | N/A |
| INaL | N/A | N/A | 3.95[3.89,4.03] | 1.41[1.08,1.73] |
| INa | 9.06[7.61,10.0] | 0.51[0.35,0.97] | N/A | N/A |
| hERG | 4.94[4.78,5.08] | 0.94[0.73,1.24] | 4.39[4.22,5.11] | 2.59[0.98,4.47] |
|  | **Quinidine** | | **Ondansetron** | |
|  | log10IC50 | H | log10IC50 | h |
| ICaL | 4.71[4.53,5.02] | 0.59[0.47,0.69] | 4.35[4.23,4.54] | 0.75[0.56,0.92] |
| IK1 | 7.60[4.51,10.0] | 0.35[0.18,1.29] | N/A | N/A |
| IKs | 3.69[3.63,3.78] | 1.36[0.96,1.99] | 5.76[5.45,6.84] | 0.65[0.40,0.78] |
| Ito | 3.54[3.50,3.59] | 1.28[1.07,1.52] | 6.01[5.17,10.0] | 0.99[0.27,1.74] |
| INaL | 3.97[3.91,4.09] | 1.34[1.04,1.58] | 4.28[4.21,4.38] | 1.03[0.80,1.32] |
| INa | 4.09[3.99,4.32] | 1.49[1.02,1.90] | 4.76[4.50,6.17] | 1.02[0.33,1.81] |
| hERG | 2.54[2.41,2.63] | 1.02[0.77,1.39] | 3.17[3.02,3.31] | 0.96[0.72,1.24] |
|  | **Diltiazem** | | **Chlorpromazine** | |
|  | log10IC50 | H | log10IC50 | h |
| ICaL | 2.05[1.90,2.19] | 0.71[0.55,0.92] | 3.91[3.80,4.06] | 0.84[0.64,1.08] |
| IK1 | N/A | N/A | 3.97[3.84,4.14] | 0.69[0.53,0.82] |
| IKs | N/A | N/A | N/A | N/A |
| Ito | 9.45[7.11,10.0] | 0.17[0.14,0.27] | 7.25[6.11,10.0] | 0.37[0.19,0.50] |
| INaL | 4.34[4.25,4.48] | 0.68[0.55,0.79] | 3.66[3.59,3.73] | 0.94[0.79,1.09] |
| INa | 5.04[4.81,6.01] | 0.70[0.38,0.88] | 3.66[3.58,3.73] | 2.00[1.64,2.45] |
| hERG | 3.82[3.72,3.90] | 0.79[0.66,0.88] | 3.05[2.94,3.15] | 0.89[0.70,1.12] |

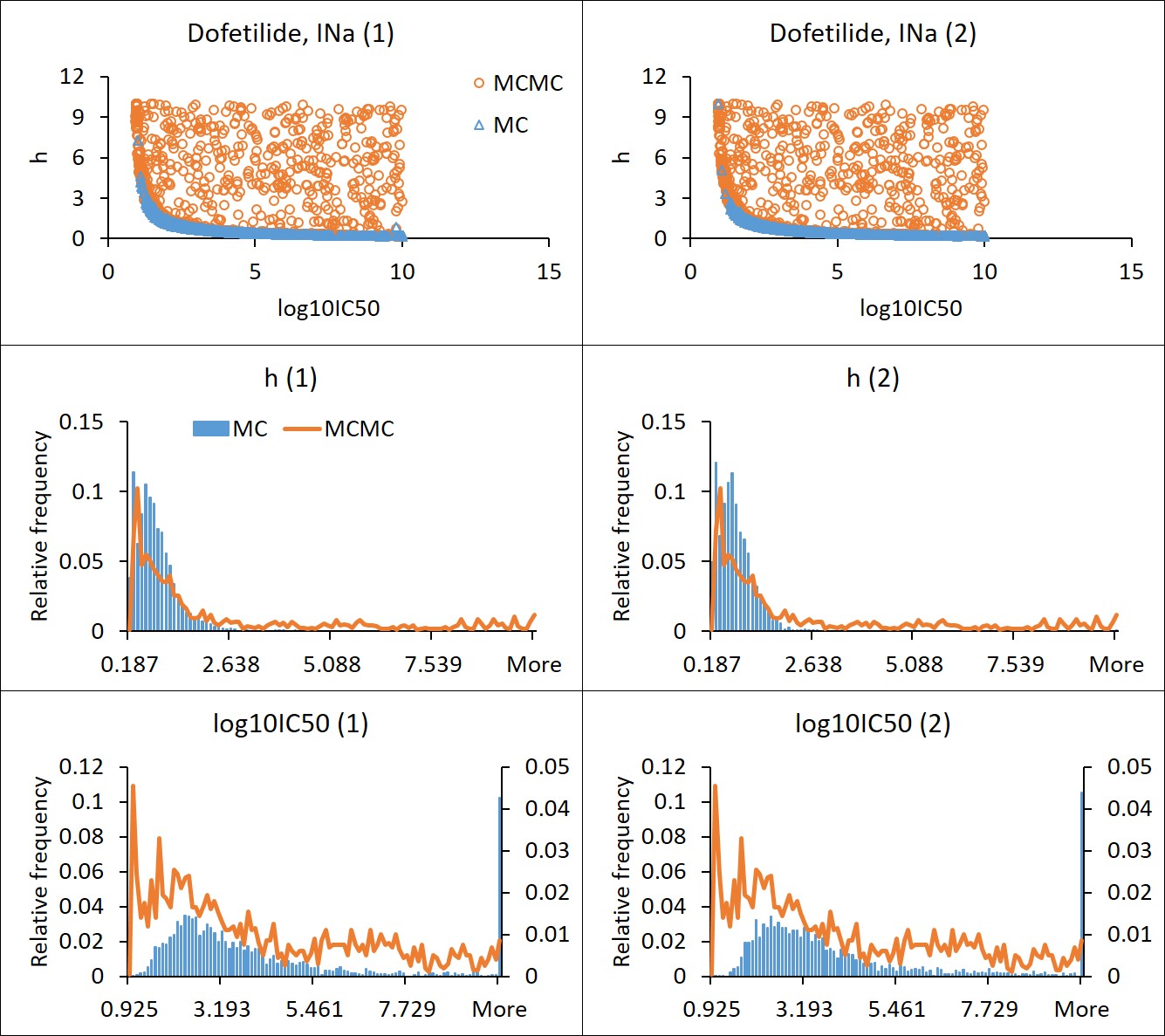
Further analysis can be carried out with the plot of the relationship among fitted parameters ( and ) as well as its histogram as shown in some sample results on [Figure 3. 1] for Dofetilide hERG channel and [Figure 3. 2] for Dofetilide INa channel. For cases with sufficient experimental data like the Dofetilide hERG channel on [Figure 3. 1], both MCMC and the proposed simulation yield similar spread of fitted parameters where there is no strong relation between and for both schemes (see top panels). The histogram plots also show that the posterior distribution of and predicted by MCMC are similar to the MC’s results which resemble Gaussian distribution (see middle and bottom panels). On the other hand, the Dofetilide INa channel on [Figure 3. 2] yields inverse proportionality of two fit parameters (see on top panels). Furthermore, the histogram plots show a completely non-Gaussian shape of parameter distribution. Both MCMC and proposed method result mostly grouped near its lower bounds while for also grouped near its lower bound except that proposed algorithm yield quite high relative frequency on upper bound of .

Moreover, from the histogram in the middle and bottom panel of [Figure 3. 1] and [Figure 3. 2], one can clearly predict and compare the confidence intervals of fit parameters between the proposed MC algorithm and MCMC. For cases like the Dofetilide hERG channel, where the histogram of MCMC and MC algorithm overlap quite well, the confidence interval of fit parameters would be similar. However, for cases with insufficient drug response like the Dofetilide INa channel where the histogram peak is not well defined and its shape does not resemble Gaussian distribution, the confidence intervals from MCMC and MC algorithm can be completely different and hardly meaningful. The term “insufficient dose-response data” will be discussed in the next subsection.



[Figure 3. 1] Comparison of the results from scheme 1 (left) and scheme 2 (right) on Dofetilide hERG channel

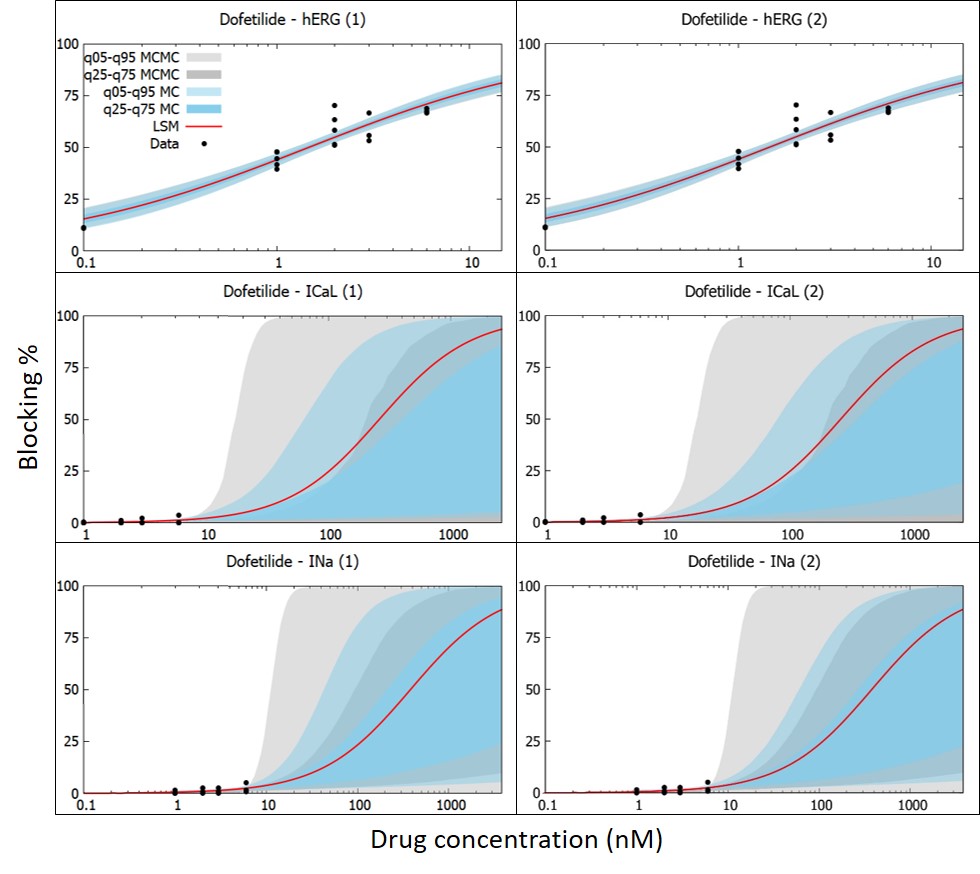
Note that some histogram plots in [Figure 3. 2] (the bottom panels) have a secondary vertical axis on the right side of the panel that shows the results from MCMC. The secondary vertical axis is necessary for some cases because the results from our algorithm grouped massively near either lower or upper bounds of parameters that could yield a relatively huge number of relative frequencies compared to MCMC. In the other words, the secondary vertical will not be used where the histogram vertical value of the proposed algorithm is comparable to MCMC. More detailed figures and comparisons for all drugs can be seen in supplementary figures (https://intip.in/NA8X).



[Figure 3. 2] Comparison of the results from scheme 1 (left) and scheme 2 (right) on Dofetilide INa channel

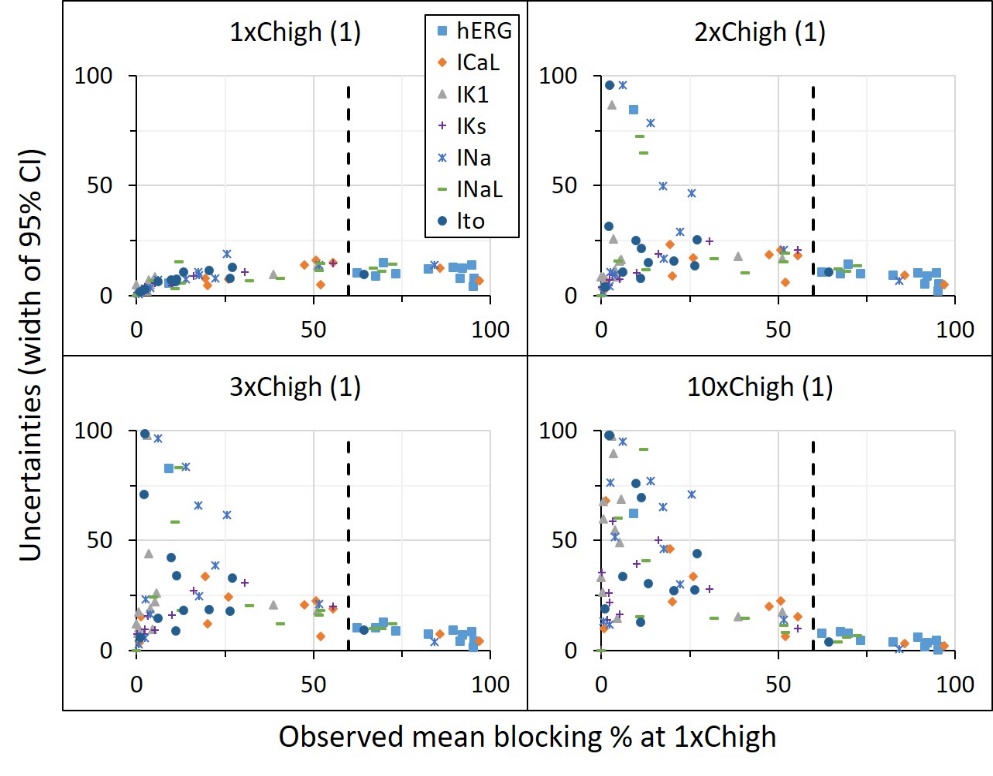
## Uncertainty of blocking response

In the previous subsection, it is obtained that while the results from proposed nonlinear LSM fitting match exactly with the one from the previous study by [15], the confidence intervals can show quite different results compared to MCMC for cases with insufficient dose-response data. The term “insufficient” can be seen from blocking response in sample results of [Figure 3. 3] that maximum blocking response obtained from experiment are very small that some cases have even lower than 20% (Dofetilide ICaL and INa channel).



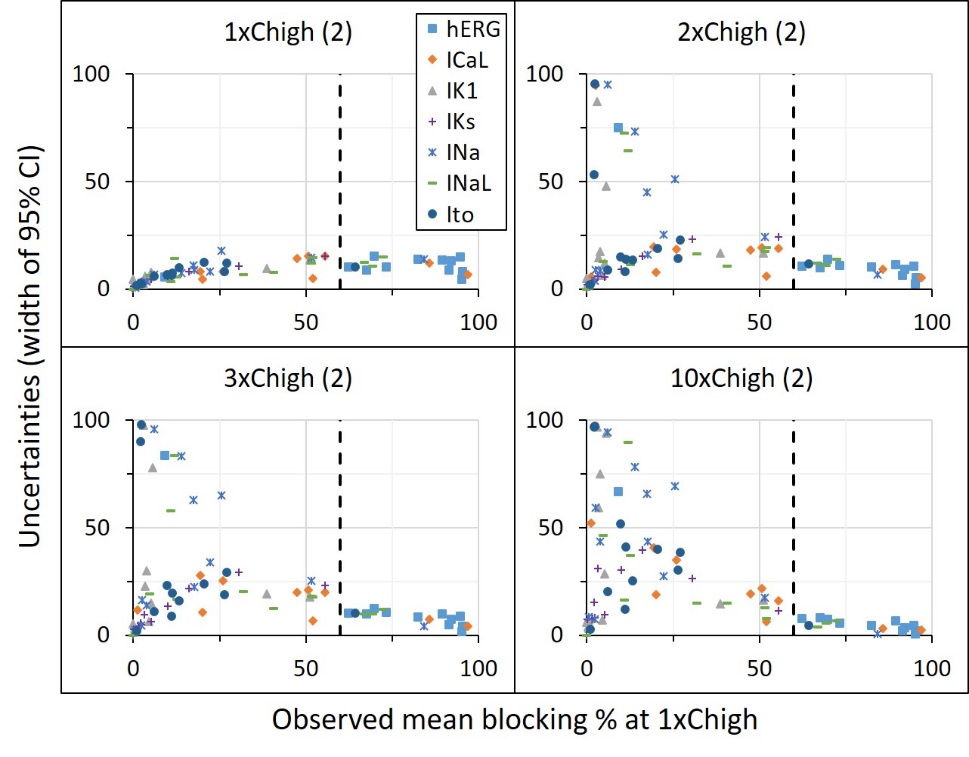
[Figure 3. 3] The comparison of Monte Carlo simulation of scheme 1 (left) and scheme 2 (right). Note that the grey shaded area is the result of MCMC as a reference. Also note that means the blocking response of 2.5 to 97.5 quantiles of the marginal distribution, while for 5 to 95 quantiles.

In [Figure 3. 3], the bootstrap samples of fit parameters are deployed to obtain a prediction of drug blocking response of the channel. Each bootstrap sample will be used for generating a prediction of blocking response with 100 different drug concentrations. Scheme 1 generally yields a broader range of blocking response compared to scheme 2. However, both schemes can produce a quite similar prediction region only on the Dofetilide hERG channel where at the same time MCMC yields a similar response. This can happen because the sufficient observed blocking responses of the drug, which are about 60-70%, is achieved. Also, from the sample figure on [Figure 3. 3], one can clearly observe that for sufficient experimental data, like the one from the hERG channel, both schemes 1 and 2 generate overlapping regions to the one from MCMC. However, when the blocking response is insufficient, like the one on ICaL and INa channels, the prediction of MCMC yields broader blocking intervals compared to both schemes 1 and 2. This also happens on most of the non-hERG channels. More detailed figures and comparisons for all drugs can be seen in supplementary figures (https://intip.in/NA8X).



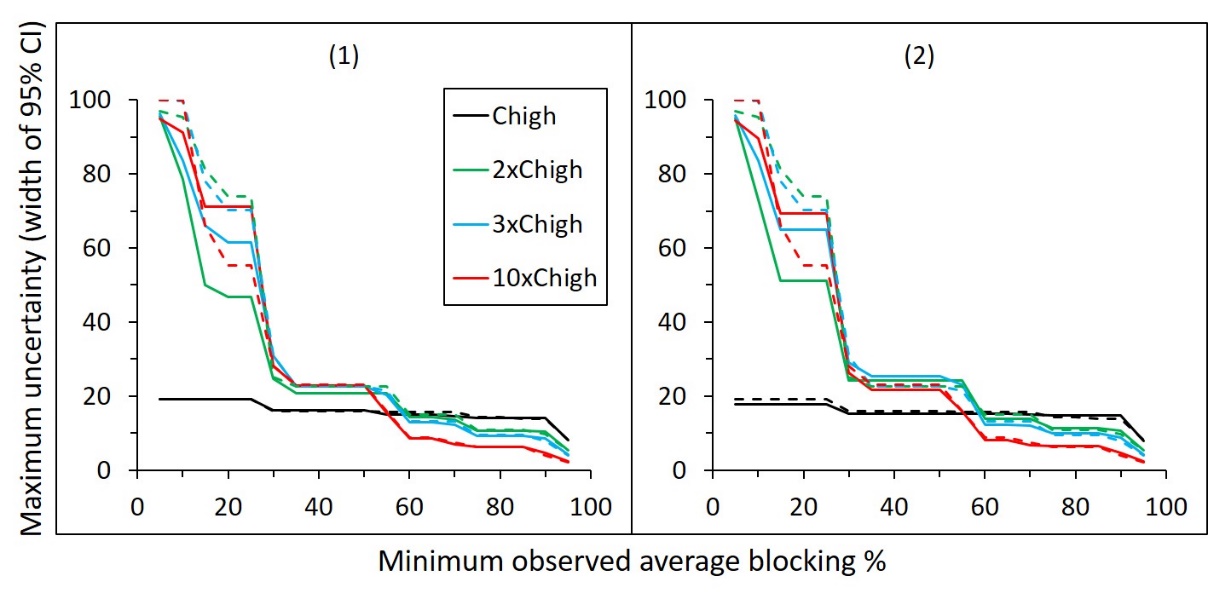
[Figure 3. 4] Uncertainties of blocking response from scheme 1 as a function of observed mean blocking percentage at the highest experimental drug concentration (). Please note that the dashed vertical line means 60% of observed mean blocking at .

Furthermore, the insufficient observed blocking response can lead to high uncertainty in predicting the drug response especially during extrapolation with bootstrap parameters. In this work, I defined the uncertainty as to the width of 95% confidence interval of blocking the response of the corresponding ionic channel. [Figure 3. 4] and [Figure 3. 5] show that both schemes yield high uncertainty (higher than 16%) when the observed mean blocking is less than 60%, the same as results obtained by a previous study [15] in [Figure 1. 3]. However, while the minimum threshold for observed mean blocking is the same as previous work, the proposed MC algorithm yields a different spread of uncertainty as shown in [Figure 3. 4] and [Figure 3. 5] compared to MCMC in [Figure 1. 3] especially for cases with low observed mean blocking (below 60%). MCMC generates significant jumps in the spread of uncertainties for to while both scheme 1 and 2 show a less significant jump in the sense that fewer cases reach high uncertainties with the same drug concentration and observed mean blocking response.



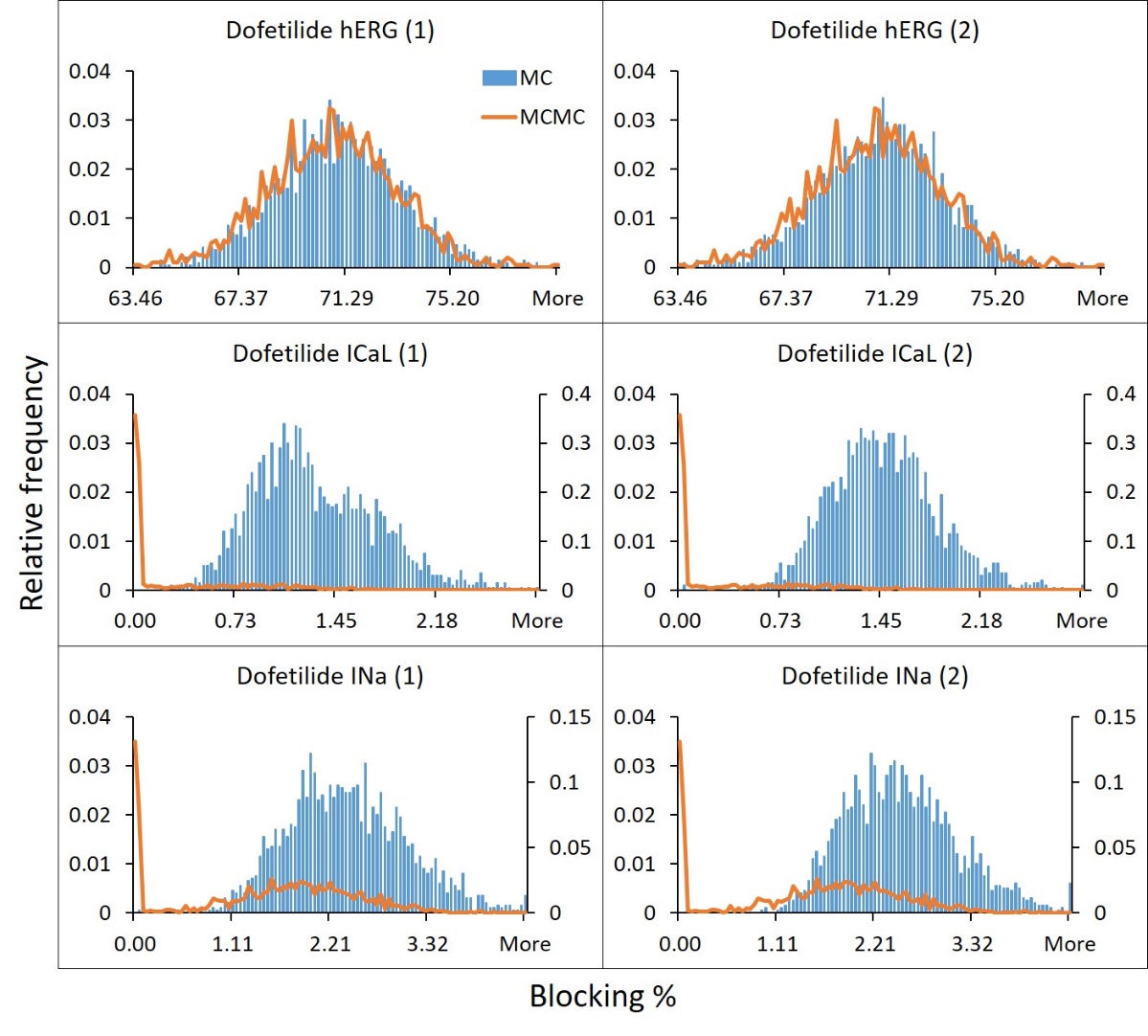
[Figure 3. 5] Uncertainties of blocking response from scheme 2 as a function of observed mean blocking percentage at the highest experimental drug concentration (). The vertical dashed line represents the same meaning as in [Figure 3. 4].

Further analysis from [Figure 3. 6] can reveal that both MCMC and the proposed MC algorithm show a similar spread of maximum uncertainty when the minimum observed blocking is applied to the blocking response prediction in [Figure 3. 4] and [Figure 3. 5]. One can see that within the range of 15-25% of minimum observed average blocking response, MCMC clearly yields higher uncertainty compared to the proposed method for most drug concentration except (red lines). This happens because at the blocking response prediction from MCMC can reach the highest possible blocking response quicker than the proposed algorithm. Therefore, while the MCMC already reaches saturation (the blocking responses are close to 100%), the proposed MC algorithm still in the process to reach it, thus gives higher uncertainty.



[Figure 3. 6] The maximum uncertainty as a function of the minimum observed average blocking response. Left panel from scheme 1 and the right one from scheme 2. Please note that the dashed line represents the result of MCMC.

In addition, from [Figure 3. 6] we can notice that below 60% of the minimum observed average blocking response, the prediction will have uncertainty higher than 16% for both MCMC and proposed MC algorithm. However, for drug concentration at , the uncertainty is 20% or less for all minimum observed blocking response. This indicates that within the range of experimental data, both MCMC and the proposed MC algorithm can provide a good prediction of the blocking response. In addition, the 60% minimum threshold obtained from [Figure 3. 4], [Figure 3. 5], and [Figure 3. 6] emphasize the results from [Figure 3. 3] that the experimental design needs to be changed if the researchers need more reliable prediction. The change that can be made is at least the minimum observed average blocking response has to be 60%. Thus, from that modification, researchers can expect the uncertainty of blocking responses to be less than 16% for all range of drug concentrations.



[Figure 3. 7] The blocking response distribution at a drug concentration of from scheme 1 (left) and scheme 2 (right). axis is the relative frequency and axis is the blocking percentage. Please note that the middle and lower panels have a secondary vertical axis for the results of MCMC.

One important assumption in the proposed MC simulation is that the scatter of data around the curves follows Gaussian distribution. One can clearly observe in [Figure 3. 7] that for all cases (including the cases with insufficient dose-response data), the blocking response follows the Gaussian distribution. Meanwhile, MCMC generates Gaussian scatter of data only for the Dofetilide hERG channel (top panels) which has sufficient dose-response data. On the other hand, for insufficient cases, as shown in the middle and bottom panels, MCMC yields quite irregular distribution as a relatively huge number of blocking responses are grouped close to its minimum value. This indicates that my assumption for the scattering of data is appropriate whenever sufficient data is achieved and MCMC results emphasize that assumption.

# Conclusion and Limitation

Hill fitting and bootstrapping can play an important role as integral parts of drug assessment in the CiPA project. Providing simple yet reliable parameter fitting and predicting capabilities, I showed that Hill fitting with LSM and Monte Carlo simulation could help researchers during their first step of computational drug assessment. In addition, the proposed MC method generates quite similar results compared to the rather complicated existing MCMC method, both in producing best-fit parameters and predicting the uncertainties. These advantages indicate that the proposed methodology, especially MC bootstrapping samples, can be easily applied in Hill fitting and it also can be extended to other problems incorporating bootstrapping samples in general.

Nevertheless, despite the promising results from the proposed methodology especially for bootstrapping, one can observe that this method has one prominent limitation. During the generation of blocking data by adding Gaussian scatters, I assumed that the scatters follow Gaussian distribution. In Hill fitting, the Gaussian distribution of scatter data produces quite consistent results especially with sufficient dose-response data compared to MCMC. However, for other types of problems, the errors may follow other of distribution, thus one may need to change with different distribution sampling during MC simulation.

# [References]

1. Hill, A.V., *The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves.* The Journal of Physiology, 1910. **40**: p. i--vii.

2. Gesztelyi, R., et al., *The Hill equation and the origin of quantitative pharmacology.* Archive for History of Exact Sciences, 2012. **66**(4): p. 427-438.

3. Langmuir, I., *THE ADSORPTION OF GASES ON PLANE SURFACES OF GLASS, MICA AND PLATINUM.* Journal of the American Chemical Society, 1918. **40**(9): p. 1361-1403.

4. Neubig, R.R., et al., *International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology.* Pharmacol Rev, 2003. **55**(4): p. 597-606.

5. Kenakin, T., *Data-driven analysis in drug discovery.* J Recept Signal Transduct Res, 2006. **26**(4): p. 299-327.

6. Motulsky, H. and A. Christopoulos. *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*. 2004.

7. Black, J.W. and P. Leff, *Operational Models of Pharmacological Agonism.* Proceedings of the Royal Society of London. Series B, Biological Sciences, 1983. **220**(1219): p. 141-162.

8. Keller, F., et al., *PK-PD curve-fitting problems with the Hill equation? Try one of the 1-exp functions derived from Hodgkin, Douglas or Gompertz.* Int J Clin Pharmacol Ther, 2002. **40**(1): p. 23-9.

9. Giraldo, J., et al., *Assessing the (a)symmetry of concentration-effect curves: empirical versus mechanistic models.* Pharmacol Ther, 2002. **95**(1): p. 21-45.

10. Van der Graaf, P.H. and W.B. Stam, *Analysis of receptor inactivation experiments with the operational model of agonism yields correlated estimates of agonist affinity and efficacy.* Journal of Pharmacological and Toxicological Methods, 1999. **41**(2): p. 117-125.

11. Gintant, G.A., *Preclinical Torsades-de-Pointes screens: advantages and limitations of surrogate and direct approaches in evaluating proarrhythmic risk.* Pharmacol Ther, 2008. **119**(2): p. 199-209.

12. Sager, P.T., et al., *Rechanneling the cardiac proarrhythmia safety paradigm: a meeting report from the Cardiac Safety Research Consortium.* Am Heart J, 2014. **167**(3): p. 292-300.

13. Vicente, J., et al., *Mechanistic Model-Informed Proarrhythmic Risk Assessment of Drugs: Review of the "CiPA" Initiative and Design of a Prospective Clinical Validation Study.* Clin Pharmacol Ther, 2018. **103**(1): p. 54-66.

14. Strauss, D.G., et al., *Comprehensive In Vitro Proarrhythmia Assay (CiPA) Update from a Cardiac Safety Research Consortium / Health and Environmental Sciences Institute / FDA Meeting.* Ther Innov Regul Sci, 2019. **53**(4): p. 519-525.

15. Chang, K.C., et al., *Uncertainty Quantification Reveals the Importance of Data Variability and Experimental Design Considerations for in Silico Proarrhythmia Risk Assessment.* Front Physiol, 2017. **8**: p. 917.

16. Haario, H., et al., *DRAM: Efficient adaptive MCMC.* Statistics and Computing, 2006. **16**(4): p. 339-354.

17. Laine, M. and J. Tamminen, *Aerosol model selection and uncertainty modelling by adaptive MCMC technique.* Atmos. Chem. Phys., 2008. **8**(24): p. 7697-7707.

18. Soetaert, K. and T. Petzoldt, *Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME.* Journal of statistical software, 2010. **33**: p. 1.

19. Crumb, W.J., Jr., et al., *An evaluation of 30 clinical drugs against the comprehensive in vitro proarrhythmia assay (CiPA) proposed ion channel panel.* J Pharmacol Toxicol Methods, 2016. **81**: p. 251-62.

20. Levenberg, K., *A METHOD FOR THE SOLUTION OF CERTAIN NON-LINEAR PROBLEMS IN LEAST SQUARES.* Quarterly of Applied Mathematics, 1944. **2**(2): p. 164-168.

21. Marquardt, D.W., *An Algorithm for Least-Squares Estimation of Nonlinear Parameters.* Journal of the Society for Industrial and Applied Mathematics, 1963. **11**(2): p. 431-441.

22. Li, Z., et al., *Improving the In Silico Assessment of Proarrhythmia Risk by Combining hERG (Human Ether-a-go-go-Related Gene) Channel-Drug Binding Kinetics and Multichannel Pharmacology.* Circ Arrhythm Electrophysiol, 2017. **10**(2): p. e004628.

# Appendix

## Code for Data processing

Prior to fitting procedure, there will be data processing steps that read experimental data, store it into arrays, sort it with respect to channel name, and average blocking data on each channel. These steps are described by several C++ functions and steps as follow:

* Read experimental data and find amount of data per channel

int readDataPerChannel(int numberOfData, std::string drugName, std::string channelName, std::string\*\* dataStr){

int iter;

iter = 0;

for(int i=0;i<numberOfData;i++){

if(dataStr[i][0]==drugName && dataStr[i][2]==channelName){

iter = iter + 1;

}

}

return iter;

}

This function will return an integer value of the number of data on each channel of the drug.

* Store the data per channel into array

void storeDataPerChannel(int numberOfData, std::string drugName, std::string channelName, std::string\*\* dataStr,double \*\*dataPerChannel, double \*\*data){

int iter;

iter = 0;

for(int i=0;i<numberOfData;i++){

if(dataStr[i][0]==drugName && dataStr[i][2]==channelName){//change channelList index for other channel

dataPerChannel[iter][0] = data[i][0];//conc

dataPerChannel[iter][1] = data[i][1];//block

iter = iter + 1;

}

}

return;

}

The output of this function is the dataPerChannel array that contains all experimental data per channel.

* Reading data per channel to obtain number of data having same amount of drug concentration

int readConcPerChannel(int numberOfDrugDataPerChannel, double \*\*dataPerChannel){

int iter;

double conc;

iter = 0;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

if(i==0){

conc = dataPerChannel[i][0];//conc variable

iter = iter + 1;

}

else{

if(dataPerChannel[i][0]!=conc){

conc = dataPerChannel[i][0];//conc variable

iter = iter + 1;

}

}

}

return iter;

}

It will return an integer value of the number of different concentrations within the data per channel.

* Calculating the averaged data for each drug concentration as described in [Figure 2. 2]

void findAveragedData(int numberOfDrugDataPerChannel, double \*\*dataPerChannel, double \*\*sortedDataPerChannel){

int iter,iter2;

double block,conc;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

if(i==0){//first data of dataPerChannel

conc = dataPerChannel[i][0];//conc variable

block = dataPerChannel[i][1];//block variable

iter = 1;//counting number of same conc values

iter2 = 0;//row of sortedDataPerChannel

}

else{

if(dataPerChannel[i][0]==conc){

if(i<numberOfDrugDataPerChannel-1){

block = block + dataPerChannel[i][1];//adding block variables

iter = iter + 1;

}

else{//last data of dataPerChannel

block = block + dataPerChannel[i][1];//adding block variables

iter = iter + 1;

block = block/iter;

sortedDataPerChannel[iter2][0] = conc;

sortedDataPerChannel[iter2][1] = block;

//std::cout << i << " Sorted " << iter2 << " has value of conc = " << conc;

//std::cout << " and average block " << block << std::endl;

}

}

else{

block = block/iter;//average block

sortedDataPerChannel[iter2][0] = conc;

sortedDataPerChannel[iter2][1] = block;

//std::cout << i << " Sorted " << iter2 << " has value of conc = " << conc;

//std::cout << " and average block " << block << std::endl;

conc = dataPerChannel[i][0];//set new conc variable

block = dataPerChannel[i][1];//set new block variable

iter = 1;//restart counting same conc values

iter2 = iter2 + 1;//next row of sortedDataPerChannel

}

}

}

return;

}

The sorted data will be the stored in sortedDataPerChannel array.

## Code for linear and nonlinear LSM

As stated in Section 2.1.1, especially (Eq. 2.1.2) to (Eq. 2.1.4) that express the linear equation for linear LSM fitting, the implementation of LSM relies on ALGLIB C++where the main calculation is as follow:

* The minimum requirement to conduct the linear LSM fitting is declared first.

The algorithm requires that the experimental data have at least 4 different drug concentrations for each channel. It is implemented as:

if(numberOfDataLinFit<4){

std::cout << "Too few number of distinct conc values to fit Hill equation " << std::endl;

continue;

}

* If the minimum requirement is fulfilled, the next step is to calculate the actual initial guess of and . If the data is either dominated by very low or very high blocking, the and will be set as follow:

//criteria for linear fitting

if(flag1==numberOfDataLinFit){//all of the sortedDataPerChannell are very small blocking

//c0 = -log10(gsl\_vector\_max(dataConc));//log10(IC50)

c0 = -log10(\*std::max\_element(dataConc,dataConc+numberOfDataLinFit));//log10(IC50)

c1 = 1;//Hill coefficient

}

else if(flag2==numberOfDataLinFit){//all of the sortedDataPerChannell are very big blocking

c0 = -log10(\*std::min\_element(dataConc,dataConc+numberOfDataLinFit));//log10(IC50)

c1 = 1.0;//Hill coefficient

}

* If neither very low nor very high blocking dominate the data or ,in other words, we have a regular data, we can securely conduct linear LSM as follow:

//ignore data with dataFlag!=0

numberOfDataLinFit = numberOfDataLinFit - flag1 - flag2;

//std::cout << "This is the number of data to fit: " << numberOfDataLinFit << std::endl;

if(2 <= numberOfDataLinFit){

dataX = new double [numberOfDataLinFit];

dataY = new double [numberOfDataLinFit];

ignoreSomeData(numberOfDataLinFit,flag1,flag2,dataFlag,dataX,dataY,sortedDataPerChannel);

//initialize linear lsm environment

real\_1d\_array ylin;

real\_2d\_array fmatrix;

ylin.setlength(numberOfDataLinFit);

fmatrix.setlength(numberOfDataLinFit,2);

for(int i=0;i<numberOfDataLinFit;i++){

ylin[i] = dataY[i];

fmatrix[i][0] = 1;

fmatrix[i][1] = dataX[i];

}

int\_t infolin;

real\_1d\_array clin;

lsfitreport replin;

// Fitting without weights

lsfitlinear(ylin, fmatrix, infolin, clin, replin);

//printf("# best fit: Y = %g + %g X\n",clin[0],clin[1]);

//printf("# h = %g, log10IC50 = %g, IC50 = %g\n",clin[1],-clin[0]/clin[1],pow(10.0,-clin[0]/clin[1]));

c0 = clin[0];

c1 = clin[1];

delete [] dataX;

delete [] dataY;

Once the linear LSM finish, the initial and are further processed into nonlinear LSM fitting by following the procedure explained in Section 2.1.2. The implementation of nonlinear LSM fitting is done by using the following function:

void nonLinearFit(int numberOfDrugDataPerChannel, double c0, double c1, double \*log10IC50, double \*h, double \*\*dataPerChannel){

double conc;

real\_2d\_array x;

real\_1d\_array y;

x.setlength(numberOfDrugDataPerChannel,1);

y.setlength(numberOfDrugDataPerChannel);

for(int i=0;i<numberOfDrugDataPerChannel;i++){

x[i][0] = dataPerChannel[i][0];//drug concentration

y[i] = dataPerChannel[i][1];//blocking percentage

}

real\_1d\_array c;//parameters to fit

c.setlength(2);

//check the linear fitting results

//lower and upper bounds for h

if(c1 <= 0 || 10 <= c1){

//set new h parameter

c1 = 0.9;

}

//lower and upper bounds for log10IC50

if(-c0/c1 <= -10 || 10 <= -c0/c1){

//find average conc value

conc = 0.0;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

conc = conc + dataPerChannel[i][0];//conc

}

conc = conc/numberOfDrugDataPerChannel;

//new log10IC50 = -c0/c1

c0 = -log10(conc)\*c1;

}

c[0] = -c0/c1;//log10IC50

c[1] = c1;//h

real\_1d\_array bndl;

bndl.setlength(2);

bndl[0] = -10.0;

bndl[1] = 0;

real\_1d\_array bndu;

bndu.setlength(2);

bndu[0] = 10.0;

bndu[1] = 10.0;

double epsx = 0.000001;

ae\_int\_t maxits = 0;

ae\_int\_t info;

lsfitstate state;

lsfitreport rep;

double diffstep = 0.0001;

lsfitcreatef(x, y, c, diffstep, state);

lsfitsetbc(state, bndl, bndu);

lsfitsetcond(state, epsx, maxits);

alglib::lsfitfit(state, function\_cx\_1\_func);

lsfitresults(state, info, c, rep);

//store results

\*log10IC50 = c[0];

\*h = c[1];

return;

}

This function will result the optimum log10IC50 and h.

## Code for Monte Carlo simulation

The implementation of Monte Carlo simulation is basically the same as the previous Appendix B except that it repeats the LSM fitting as many bootstrap samples as we need and it has different set of data on every iteration. There are two available schemes as in [Figure 2. 1]. In the first scheme, the data is generated by using optimal or best-fit parameters from LSM fitting with Gaussian noise. The implementation of data generator for this scheme is as follow:

for(int j=0;j<numberOfDrugDataPerChannel;j++){

conc = dataPerChannel[j][0];//data of drug concentration

//block = dataPerChannel[j][1];//obtaining blocking from actual data

block = blocking(log10IC50\_Best,conc,h\_Best);//obtaining blocking from previous best-fit value

noise = gsl\_ran\_gaussian(r,SD);

dataPerChannelRand[j][0] = conc;//conc

if(block+noise<0){

dataPerChannelRand[j][1] = 0;//block

}

else{

dataPerChannelRand[j][1] = block+noise;//block

}

//printf("Conc: %f Blocking: %f Noise: %f \n",conc,block,noise);

}

The noise = gsl\_ran\_gaussian(r,SD) is given to the blocking data in every iteration. Each iteration will result different noise thus generates a different set of data. Meanwhile, the second scheme generates data from actual experimental data with Gaussian noise simply by changing the following expression:

block = blocking(log10IC50\_Best,conc,h\_Best);//obtaining blocking from previous best-fit value

to

block = dataPerChannel[j][1];//obtaining blocking from actual data

## Complete code for Hill fitting with LSM and Monte Carlo simulation

#include "stdafx.h"

#include <stdlib.h>

#include <stdio.h>

#include <math.h>

#include "interpolation.h"

#include <iostream>

#include <iomanip>

#include <fstream>

#include <sstream>

#include <cstring>

#include <string>

#include <algorithm>

#include <math.h>

#include <gsl/gsl\_rng.h>

#include <gsl/gsl\_randist.h>

using namespace alglib;

//function for linear fitting

double Y(double block){

return log10(1.0/(1.0-block/100.0)-1.0);

}

double X(double conc){

return log10(conc);

}

double blocking(double log10IC50, double conc, double h){

return 100.0\*(1.0-1.0/(1.0+pow((conc/pow(10.0,log10IC50)),h)));

}

void function\_cx\_1\_func(const real\_1d\_array &c, const real\_1d\_array &x, double &func, void \*ptr);

void readData(std::string fileName, double \*\*data, std::string \*\*dataStr);

int readChannels(int numberOfData, std::string drugName, std::string\*\* dataStr);

int readDataPerChannel(int numberOfData, std::string drugName, std::string channelName, std::string\*\* dataStr);

void storeDataPerChannel(int numberOfData, std::string drugName, std::string channelName, std::string\*\* dataStr,double \*\*dataPerChannel, double \*\*data);

int readConcPerChannel(int numberOfDrugDataPerChannel, double \*\*dataPerChannel);

void storeChannelList(int numberOfData, std::string drugName, std::string\*\* dataStr, std::string\* channelList);

void findAveragedData(int numberOfDrugDataPerChannel, double \*\*dataPerChannel, double \*\*sortedDataPerChannel);

void ignoreSomeData(int numberOfDataLinFit, int flag1, int flag2, int\* dataFlag, double\* dataX, double\* dataY, double\*\* sortedDataPerChannel);

void nonLinearFit(int numberOfDrugDataPerChannel, double c0, double c1, double \*log10IC50, double \*h, double \*\*dataPerChannel);

double stdev(int numberOfDrugDataPerChannel, double log10IC50, double h, double \*\* dataPerChannel){

double result = 0.0, block, conc;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

conc = dataPerChannel[i][0];

block = dataPerChannel[i][1];

result = result + pow(blocking(log10IC50,conc,h)-block,2.0);

}

result = sqrt(result/(numberOfDrugDataPerChannel-2));

return result;

}

int main(int argc, char \*\*argv)

{

//

// In this example we demonstrate Hill fitting by

// f(c,x)=100\*(1-1/(1+pow(x/pow(10,c0),c1)))

// subject to bound constraints

// -10.0 <= c0 <= 10.0

// 0.0 <= c1 <= 10.0

// using function value only.

// c0 = log10IC50 and c1 = h.

//

// Gradient is estimated using combination of numerical differences

// and secant updates. diffstep variable stores differentiation step

// (we have to tell algorithm what step to use).

//

//declaration for linear fitting

int numberOfData,numberOfChannels,numberOfDrugDataPerChannel;

int \*dataFlag,numberOfDataLinFit;

double conc,block,noise,eps=0.01;

double c0,c1,cov00,cov01,cov11,chisq;

double \*weight,SD,log10IC50,h;

double \*\*dataRaw,\*\*dataPerChannel,\*\*dataPerChannelRand,\*\*sortedDataPerChannel,\*\*finalSamples;

std::string \*\*dataStr,\*channelList,tempString,line,fileName,oFileName,bFileName,sFileName,headerSFile;

std::string drugName,channelName,drug,channel,units,pacing;

//initialization

//==============

if (argc != 2){

fprintf (stderr,"usage: ./lsm.out drugName\n");

exit (-1);

}

drugName = argv[1];

bFileName = "data/bootstrap\_"+drugName+".csv";

sFileName = "data/IC50\_samples\_"+drugName+".csv";

//open output file to store results

std::ofstream bFile(bFileName);

std::ofstream sFile(sFileName);

bFile << "channel,log10IC50,h" << std::endl;

//storing data from drug\_block.csv

//================================

fileName = "data/drug\_block\_validation.csv";

std::ifstream inputFile(fileName);

if(!inputFile.is\_open()){

std::cout << "Input drug\_block.csv file failed to open" << std::endl;

std::cout << "File name:" << fileName << std::endl;

std::cout << "==============================" << std::endl;

}

//skip file header

getline(inputFile,line);

//calculate number of drug data

numberOfData = 0;

while(getline(inputFile,line)){

numberOfData = numberOfData + 1;

}

inputFile.clear();

inputFile.seekg(0, inputFile.beg);

//preparing data array for conc and block

dataRaw = new double \* [numberOfData];

for(int i=0;i<numberOfData;i++){

dataRaw[i] = new double [2];

}

//preparing an array for drug,channel,units and pacing

dataStr = new std::string\* [numberOfData];

for(int i=0;i<numberOfData;i++){

dataStr[i] = new std::string [4];

}

//store inputFile into data and dataStr matrix

readData(fileName,dataRaw,dataStr);

//collecting name of channels

numberOfChannels = readChannels(numberOfData,drugName,dataStr);

//preparing an array for channel list

channelList = new std::string [numberOfChannels];

//storing channel names into channelList

storeChannelList(numberOfData,drugName,dataStr,channelList);

//preparing to store the finalSamples for sFile (drug IC50 bootstrap samples)

finalSamples = new double \*[2000];

for(int i=0;i<2000;i++){

finalSamples[i] = new double [2\*numberOfChannels];

}

//Start calculation value for all channels

//========================================

for(int channelNum=0;channelNum<numberOfChannels;channelNum++){

//collecting drug data per channel

channelName = channelList[channelNum];

headerSFile = headerSFile+channelName+"\_IC50,"+channelName+"\_h,";//Header for IC50 2,000 samples file

numberOfDrugDataPerChannel = readDataPerChannel(numberOfData,drugName,channelName,dataStr);

//storing dataPerChannel

dataPerChannel = new double \* [numberOfDrugDataPerChannel];

dataPerChannelRand = new double \* [numberOfDrugDataPerChannel];//For storing dataPerChannel with random noise

for(int i=0;i<numberOfDrugDataPerChannel;i++){

dataPerChannel[i] = new double [2];

dataPerChannelRand[i] = new double [2];

}

storeDataPerChannel(numberOfData,drugName,channelName,dataStr,dataPerChannel,dataRaw);

//find number of different conc values

numberOfDataLinFit = readConcPerChannel(numberOfDrugDataPerChannel,dataPerChannel);

std::cout << "Drug " << drugName << " with";

std::cout << " channel "<< channelName << " has ";

std::cout << numberOfDrugDataPerChannel << " number of data per channel";

std::cout << " and " << numberOfDataLinFit << " number of different conc values" << std::endl;

sortedDataPerChannel = new double \* [numberOfDataLinFit];

for(int i=0;i<numberOfDataLinFit;i++){

sortedDataPerChannel[i] = new double [2];

}

//finding average for sortedDataPerChannel

findAveragedData(numberOfDrugDataPerChannel,dataPerChannel,sortedDataPerChannel);

//linear fitting procedure

//========================

int flag1,flag2;

double \*dataX,\*dataY,\*dataConc;

//store initial dataFlag

dataFlag = new int [numberOfDataLinFit];

for(int i=0;i<numberOfDataLinFit;i++){

dataFlag[i] = 0;

}

//check the total number of sortedDataPerChannel

if(numberOfDataLinFit<4){

std::cout << "Too few number of distinct conc values to fit Hill equation " << std::endl;

continue;

}

else{

//counting flag1 for too many very small blocking

std::cout << "Count many very small blocking " << std::endl;

flag1 = 0;

for(int i=0;i<numberOfDataLinFit;i++){

if(sortedDataPerChannel[i][1]<eps){

flag1 = flag1 + 1;

dataFlag[i] = 1;

}

}

std::cout << "Number of too small blocking: " << flag2 << std::endl;

//counting flag2 for too many very big blocking

std::cout << "Count many very big blocking " << std::endl;

flag2 = 0;

for(int i=0;i<numberOfDataLinFit;i++){

if(sortedDataPerChannel[i][1]>100.0-eps){

flag2 = flag2 + 1;

dataFlag[i] = 2;

}

}

std::cout << "Number of too big blocking: " << flag2 << std::endl;

//copying conc data from sortedDataPerChannel

std::cout << "check" << std::endl;

dataConc = new double [numberOfDataLinFit];

for(int i=0;i<numberOfDataLinFit;i++){

dataConc[i] = sortedDataPerChannel[i][1];

}

//criteria for linear fitting

if(flag1==numberOfDataLinFit){//all of the sortedDataPerChannell are very small blocking

c0 = -log10(\*std::max\_element(dataConc,dataConc+numberOfDataLinFit));//log10(IC50)

c1 = 1;//Hill coefficient

std::cout << "check2" << std::endl;

}

else if(flag2==numberOfDataLinFit){//all of the sortedDataPerChannell are very big blocking

c0 = -log10(\*std::min\_element(dataConc,dataConc+numberOfDataLinFit));//log10(IC50)

c1 = 1.0;//Hill coefficient

std::cout << "check3" << std::endl;

}

else{

//ignore data with dataFlag!=0

numberOfDataLinFit = numberOfDataLinFit - flag1 - flag2;

std::cout << "This is the number of data to fit: " << numberOfDataLinFit << std::endl;

if(2 <= numberOfDataLinFit){

dataX = new double [numberOfDataLinFit];

dataY = new double [numberOfDataLinFit];

ignoreSomeData(numberOfDataLinFit,flag1,flag2,dataFlag,dataX,dataY,sortedDataPerChannel);

//initialize linear lsm environment

real\_1d\_array ylin;

real\_2d\_array fmatrix;

ylin.setlength(numberOfDataLinFit);

fmatrix.setlength(numberOfDataLinFit,2);

for(int i=0;i<numberOfDataLinFit;i++){

ylin[i] = dataY[i];

fmatrix[i][0] = 1;

fmatrix[i][1] = dataX[i];

}

ae\_int\_t infolin;

real\_1d\_array clin;

lsfitreport replin;

// Fitting without weights

lsfitlinear(ylin, fmatrix, infolin, clin, replin);

printf("# best fit: Y = %g + %g X\n",clin[0],clin[1]);

printf("# h = %g, log10IC50 = %g, IC50 = %g\n",clin[1],-clin[0]/clin[1],pow(10.0,-clin[0]/clin[1]));

c0 = clin[0];

c1 = clin[1];

delete [] dataX;

delete [] dataY;

}

else{

std::cout << "check4" << std::endl;

c1 = 0.9;//Hill coeff

//log10IC50 = -log10(conc)\*h

c0 = -log10(0.5\*(\*std::max\_element(dataConc,dataConc+numberOfDataLinFit)-\*std::min\_element(dataConc,dataConc+numberOfDataLinFit)))\*c1;

}

}

delete [] dataConc;

}//else

//Nonlinear lsm procedure

//=======================

double log10IC50\_Best, h\_Best,SD;

nonLinearFit(numberOfDrugDataPerChannel,c0,c1,&log10IC50\_Best,&h\_Best,dataPerChannel);

SD = stdev(numberOfDrugDataPerChannel,log10IC50\_Best,h\_Best,dataPerChannel);

printf("Best fit parameters from LSM:\n");

printf("log10IC50 = %f, h = %f \n",log10IC50\_Best,h\_Best);

printf("SD of residuals = %f\n",SD);

printf("=============================\n");

//Start Monte Carlo simulation to generate bootstrap

printf("Starting Monte Carlo simulation ...\n");

double \*MClog10IC50Samples,\*MChSamples;

MClog10IC50Samples = new double [2000];

MChSamples = new double [2000];

for(int MC=0;MC<2000;MC++){

//Modifying the block percentage on dataPerChannelRand with Gaussian noise

const gsl\_rng\_type \* M;

gsl\_rng \* r;

gsl\_rng\_env\_setup();

M = gsl\_rng\_default;

r = gsl\_rng\_alloc(M);

gsl\_rng\_set(r,clock());

for(int j=0;j<numberOfDrugDataPerChannel;j++){

conc = dataPerChannel[j][0];//data of drug concentration

//block = dataPerChannel[j][1];//obtaining blocking from actual data

block = blocking(log10IC50\_Best,conc,h\_Best);//obtaining blocking from previous best-fit value

noise = gsl\_ran\_gaussian(r,SD);

dataPerChannelRand[j][0] = conc;//conc

if(block+noise<0){

dataPerChannelRand[j][1] = 0;//block

}

else{

dataPerChannelRand[j][1] = block+noise;//block

}

//printf("Conc: %f Blocking: %f Noise: %f \n",conc,block,noise);

}

//find number of different conc values

numberOfDataLinFit = readConcPerChannel(numberOfDrugDataPerChannel,dataPerChannelRand);

//finding average for sortedDataPerChannel

findAveragedData(numberOfDrugDataPerChannel,dataPerChannelRand,sortedDataPerChannel);

//linear fitting procedure Monte Carlo

//====================================

int flag1,flag2;

double \*dataX,\*dataY,\*dataConc;

//store initial dataFlag

dataFlag = new int [numberOfDataLinFit];

for(int i=0;i<numberOfDataLinFit;i++){

dataFlag[i] = 0;

}

//check the total number of sortedDataPerChannel

if(numberOfDataLinFit<4){

std::cout << "Too few number of distinct conc values to fit Hill equation " << std::endl;

continue;

}

else{

//counting flag1 for too many very small blocking

flag1 = 0;

for(int i=0;i<numberOfDataLinFit;i++){

if(sortedDataPerChannel[i][1]<eps){

flag1 = flag1 + 1;

dataFlag[i] = 1;

}

}

//counting flag2 for too many very big blocking

flag2 = 0;

for(int i=0;i<numberOfDataLinFit;i++){

if(sortedDataPerChannel[i][1]>100.0-eps){

flag2 = flag2 + 1;

dataFlag[i] = 2;

}

}

//copying conc data from sortedDataPerChannel

dataConc = new double [numberOfDataLinFit];

for(int i=0;i<numberOfDataLinFit;i++){

dataConc[i] = sortedDataPerChannel[i][1];

}

//criteria for linear fitting

if(flag1==numberOfDataLinFit){//all of the sortedDataPerChannell are very small blocking

c0 = -log10(\*std::max\_element(dataConc,dataConc+numberOfDataLinFit));//log10(IC50)

c1 = 1;//Hill coefficient

}

else if(flag2==numberOfDataLinFit){//all of the sortedDataPerChannell are very big blocking

c0 = -log10(\*std::min\_element(dataConc,dataConc+numberOfDataLinFit));//log10(IC50)

c1 = 1.0;//Hill coefficient

}

else{

//ignore data with dataFlag!=0

numberOfDataLinFit = numberOfDataLinFit - flag1 - flag2;

if(2 <= numberOfDataLinFit){

dataX = new double [numberOfDataLinFit];

dataY = new double [numberOfDataLinFit];

ignoreSomeData(numberOfDataLinFit,flag1,flag2,dataFlag,dataX,dataY,sortedDataPerChannel);

//initialize linear lsm environment

real\_1d\_array ylin;

real\_2d\_array fmatrix;

ylin.setlength(numberOfDataLinFit);

fmatrix.setlength(numberOfDataLinFit,2);

for(int i=0;i<numberOfDataLinFit;i++){

ylin[i] = dataY[i];

fmatrix[i][0] = 1;

fmatrix[i][1] = dataX[i];

}

ae\_int\_t infolin;

real\_1d\_array clin;

lsfitreport replin;

// Fitting without weights

lsfitlinear(ylin, fmatrix, infolin, clin, replin);

c0 = clin[0];

c1 = clin[1];

delete [] dataX;

delete [] dataY;

}

else{

c1 = 0.9;//Hill coeff

//log10IC50 = -log10(conc)\*h

c0 = -log10(0.5\*(\*std::max\_element(dataConc,dataConc+numberOfDataLinFit)-\*std::min\_element(dataConc,dataConc+numberOfDataLinFit)))\*c1;

}

}

delete [] dataConc;

}//else

//nonLinear fitting procedure Monte Carlo

//=======================================

double log10IC50,h;

nonLinearFit(numberOfDrugDataPerChannel,c0,c1,&log10IC50,&h,dataPerChannelRand);

bFile << channelName << "," << log10IC50 << "," << h << std::endl;

finalSamples[MC][channelNum\*2] = pow(10,log10IC50);

finalSamples[MC][channelNum\*2+1] = h;

MClog10IC50Samples[MC] = log10IC50;

MChSamples[MC] = h;

gsl\_rng\_free(r);

}

//Sort the MC samples to obtain uncertainties (95% confidence intervals)

std::sort(MClog10IC50Samples,MClog10IC50Samples+2000);

std::cout << "Upper bound of 95% CI interval for log10IC50: " << MClog10IC50Samples[1899] << std::endl;

std::cout << "Lower bound of 95% CI interval for log10IC50: " << MClog10IC50Samples[99] << std::endl;

std::sort(MChSamples,MChSamples+2000);

std::cout << "Upper bound of 95% CI interval for h: " << MChSamples[1899] << std::endl;

std::cout << "Lower bound of 95% CI interval for h: " << MChSamples[99] << std::endl;

//Finish Monte Carlo simulation to generate bootstrap

printf("Monte Carlo simulation is finished...\n");

printf("=====================================\n");

delete [] MClog10IC50Samples;

delete [] MChSamples;

delete [] dataPerChannelRand;

delete [] dataPerChannel;

delete [] sortedDataPerChannel;

}//channel

//print sFile

sFile << headerSFile << std::endl;

for(int i=0;i<2000;i++){

for(int j=0;j<numberOfChannels;j++){

sFile << finalSamples[i][j\*2] << "," << finalSamples[i][j\*2+1] << ",";

}

sFile << std::endl;

}

delete [] finalSamples;

delete [] dataRaw;

delete [] dataStr;

delete [] channelList;

sFile.close();

bFile.close();

inputFile.close();

//outputFile.close();

return 0;

}

void function\_cx\_1\_func(const real\_1d\_array &c, const real\_1d\_array &x, double &func, void \*ptr)

{

// this callback calculates f(c,x)=100\*(1-1/(1+pow(x/pow(10,c0),c1)))

// where x is a drug concentration on X-axis and c is adjustable parameter

// c0 = log10IC50, c1 = h

func = 100\*(1-1/(1+pow(x[0]/pow(10,c[0]),c[1])));

}

void readData(std::string fileName, double \*\*data, std::string \*\*dataStr){

int iter;

double block,conc;

std::string line,tempString,drug,channel,units,pacing;

std::ifstream inputFile(fileName);

//skip file header

getline(inputFile,line);

iter = 0;

while(getline(inputFile,line)){

std::stringstream ss(line);

getline(ss,tempString,',');//drug

drug = tempString;

getline(ss,tempString,',');//conc

conc = atof(tempString.c\_str());

getline(ss,tempString,',');//units

units = tempString;

getline(ss,tempString,',');//channel

channel = tempString;

getline(ss,tempString,',');//block

block = atof(tempString.c\_str());

getline(ss,tempString,',');//pacing

pacing = tempString;

//store string and numerical data into arrays

data[iter][0] = conc;//store conc data

data[iter][1] = block;//store block data

dataStr[iter][0] = drug;//store dataStr

dataStr[iter][1] = units;//store dataStr

dataStr[iter][2] = channel;//store dataStr

dataStr[iter][3] = pacing;//store dataStr

iter = iter + 1;

}

inputFile.close();

return;

}

int readChannels(int numberOfData, std::string drugName, std::string\*\* dataStr){

int numberOfChannels;

std::string channelName;

numberOfChannels = 0;

for(int i=0;i<numberOfData;i++){

if(dataStr[i][0]==drugName && numberOfChannels==0){

channelName = dataStr[i][2];//first channel

numberOfChannels = numberOfChannels + 1;

}

else if (dataStr[i][0]==drugName && numberOfChannels>0){

if(dataStr[i][2]!=channelName){

channelName = dataStr[i][2];//next channel

numberOfChannels = numberOfChannels + 1;

}

}

}

return numberOfChannels;

}

int readDataPerChannel(int numberOfData, std::string drugName, std::string channelName, std::string\*\* dataStr){

int iter;

iter = 0;

for(int i=0;i<numberOfData;i++){

if(dataStr[i][0]==drugName && dataStr[i][2]==channelName){

iter = iter + 1;

}

}

return iter;

}

void storeDataPerChannel(int numberOfData, std::string drugName, std::string channelName, std::string\*\* dataStr,double \*\*dataPerChannel, double \*\*data){

int iter;

iter = 0;

for(int i=0;i<numberOfData;i++){

if(dataStr[i][0]==drugName && dataStr[i][2]==channelName){//change channelList index for other channel

dataPerChannel[iter][0] = data[i][0];//conc

dataPerChannel[iter][1] = data[i][1];//block

iter = iter + 1;

}

}

return;

}

int readConcPerChannel(int numberOfDrugDataPerChannel, double \*\*dataPerChannel){

int iter;

double conc;

iter = 0;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

if(i==0){

conc = dataPerChannel[i][0];//conc variable

iter = iter + 1;

}

else{

if(dataPerChannel[i][0]!=conc){

conc = dataPerChannel[i][0];//conc variable

iter = iter + 1;

}

}

}

return iter;

}

void storeChannelList(int numberOfData, std::string drugName, std::string\*\* dataStr, std::string\* channelList){

int iter;

std::string channelName;

iter = 0;

for(int i=0;i<numberOfData;i++){

if(dataStr[i][0]==drugName && iter==0){

channelName = dataStr[i][2];//first channel

channelList[iter] = channelName;

iter = iter + 1;

}

else if (dataStr[i][0]==drugName && iter>0){

if(dataStr[i][2]!=channelName){

channelName = dataStr[i][2];//next channel

channelList[iter] = channelName;

iter = iter + 1;

}

}

}

return;

}

void findAveragedData(int numberOfDrugDataPerChannel, double \*\*dataPerChannel, double \*\*sortedDataPerChannel){

int iter,iter2;

double block,conc;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

if(i==0){//first data of dataPerChannel

conc = dataPerChannel[i][0];//conc variable

block = dataPerChannel[i][1];//block variable

iter = 1;//counting number of same conc values

iter2 = 0;//row of sortedDataPerChannel

}

else{

if(dataPerChannel[i][0]==conc){

if(i<numberOfDrugDataPerChannel-1){

block = block + dataPerChannel[i][1];//adding block variables

iter = iter + 1;

}

else{//last data of dataPerChannel

block = block + dataPerChannel[i][1];//adding block variables

iter = iter + 1;

block = block/iter;

sortedDataPerChannel[iter2][0] = conc;

sortedDataPerChannel[iter2][1] = block;

}

}

else{

block = block/iter;//average block

sortedDataPerChannel[iter2][0] = conc;

sortedDataPerChannel[iter2][1] = block;

conc = dataPerChannel[i][0];//set new conc varibale

block = dataPerChannel[i][1];//set new block variable

iter = 1;//restart counting same conc values

iter2 = iter2 + 1;//next row of sortedDataPerChannel

}

}

}

return;

}

void ignoreSomeData(int numberOfDataLinFit, int flag1, int flag2, int\* dataFlag, double\* dataX, double\* dataY, double\*\* sortedDataPerChannel){

int iter;

iter = 0;

for(int i=0; i<numberOfDataLinFit+flag1+flag2;i++){

if(dataFlag[i]==0){

dataX[iter] = X(sortedDataPerChannel[i][0]);//conc

dataY[iter] = Y(sortedDataPerChannel[i][1]);//block

iter = iter + 1;

}

}

return;

}

void nonLinearFit(int numberOfDrugDataPerChannel, double c0, double c1, double \*log10IC50, double \*h, double \*\*dataPerChannel){

double conc;

real\_2d\_array x;

real\_1d\_array y;

x.setlength(numberOfDrugDataPerChannel,1);

y.setlength(numberOfDrugDataPerChannel);

for(int i=0;i<numberOfDrugDataPerChannel;i++){

x[i][0] = dataPerChannel[i][0];//drug concentration

y[i] = dataPerChannel[i][1];//blocking percentage

}

real\_1d\_array c;//parameters to fit

c.setlength(2);

//check the linear fitting results

//lower and upper bounds for h

if(c1 <= 0 || 10 <= c1){

//set new h parameter

c1 = 0.9;

}

//lower and upper bounds for log10IC50

if(-c0/c1 <= -10 || 10 <= -c0/c1){

//find average conc value

conc = 0.0;

for(int i=0;i<numberOfDrugDataPerChannel;i++){

conc = conc + dataPerChannel[i][0];//conc

}

conc = conc/numberOfDrugDataPerChannel;

//new log10IC50 = -c0/c1

c0 = -log10(conc)\*c1;

}

c[0] = -c0/c1;//log10IC50

c[1] = c1;//h

real\_1d\_array bndl;

bndl.setlength(2);

bndl[0] = -10.0;

bndl[1] = 0;

real\_1d\_array bndu;

bndu.setlength(2);

bndu[0] = 10.0;

bndu[1] = 10.0;

double epsx = 0.000001;

ae\_int\_t maxits = 0;

ae\_int\_t info;

lsfitstate state;

lsfitreport rep;

double diffstep = 0.0001;

lsfitcreatef(x, y, c, diffstep, state);

lsfitsetbc(state, bndl, bndu);

lsfitsetcond(state, epsx, maxits);

alglib::lsfitfit(state, function\_cx\_1\_func);

lsfitresults(state, info, c, rep);

//store results

\*log10IC50 = c[0];

\*h = c[1];

return;

}

|  |
| --- |
| Uncertainty Quantification through the Monte Carlo method for Parameter Estimation of In silico Model used for Drug Toxicity Assessment |
|  |
| Thesis for Master of Engineering |
| December 2020 |
| Ali Ikhsanul Qauli |