SIXTH ISC WORKSHOP ON HPC APPLICATIONS IN PRECISION MEDICINE
May 25, 2023
Hamburg, Germany
9:00 am - 1:00 pm Central European Time
Program

9:00 am – 9:10 am **Welcome—Sixth ISC Workshop on HPC Applications in Precision Medicine (HAPM23)**
Eric Stahlberg, *Frederick National Laboratory for Cancer Research, USA*

9:10 am – 9:30 am **Accelerating Drug Discovery by Combining Machine-Learning and Physics-Based Methods**
Agastya P. Bhati and Shunzhou Wan, *University College London*, United Kingdom; Jens Glaser, *Oak Ridge National Laboratory*, United States; Sean Black, *Frederick National Laboratory for Cancer Research*, United States; Marco Klahn and Hannes Loffler, *AstraZeneca*, Sweden; Matteo Turilli and Andre Merkzy, Rutgers University, United States; Mikhail Titov, *Brookhaven National Laboratory*, United States; Ola Engkvist, *AstraZeneca*, Sweden; Shantenu Jha, *Rutgers University*, United States; Eric Stalhberg, *Frederick National Laboratory for Cancer Research*, United States; Peter V. Coveney, *University College London*, United Kingdom

Presenter: Peter V. Coveney, *University College London*, United Kingdom
Moderator: Charles Gillan, *Queen's University Belfast*

Abstract: The drug discovery process currently employed in the pharmaceutical industry typically requires about 10 years and $2–3 billion to deliver one new drug. This is both too expensive and too slow, especially in emergencies like the COVID-19 pandemic. In silico methodologies need to be improved both to select better lead compounds, so as to improve the efficiency of later stages in the drug discovery protocol, and to identify those lead compounds more quickly. Traditional methods of drug discovery usually begin with virtual screening of a large library of ligands followed by in vitro and finally in vivo processes on the selected few compounds with filtering at each step. In silico methods involved, irrespective of their level of accuracy, rely heavily on human intelligence for applying chemical knowledge to filter out or suggest structural features that should be incorporated into ligand molecule to improve its binding interaction with the target protein along with optimising other ligand properties. This makes the process quite slow and is a major bottleneck in the drug discovery process currently. Machine learning techniques are increasingly being used as a substitute for physics-based in silico methods to overcome this bottleneck. However, they come with their own set of constraints.

Here, we describe an Integrated Modeling PipEline for COVID Cure by Assessing Better LEads (IMPECCABLE) that couples machine learning and physics-based methods, collating the speed of ML-based surrogates and the reliability of physics-based models, each compensating for the limitations of the other. IMPECCABLE employs multiple methodological innovations to accelerate the drug discovery process. We have developed ML models that can substantially speed up various steps in drug discovery including traversing the huge chemical space (both real as well as virtual), docking surrogate for high-throughput virtual screening and ligand pose optimisation. The usual scarcity of training data is overcome by generating large amount of relevant data from PB simulations, both structural and energetics. There is extensive flow of information upstream as well as downstream across the various components of IMPECCABLE. It is an iterative workflow that generates new data in each cycle making the predictions better each time. During the first few iterations, our focus is on ensuring diversity in order to cover a wide extent of the chemical space allowing us to identify potential regions of interest. Thereafter, we focus on a more localised search of molecules. We also developed the computational framework to support these innovations at scale and characterized the performance of this framework in terms of throughput, peak performance, and scientific results. We exhibit how augmenting human intelligence with artificial intelligence can substantially reduce the throughput time for exploring a huge chemical space thereby accelerating drug discovery.

9:30 am – 9:50 am **AIM-AHEAD: Artificial Intelligence and Machine Learning to Address Health Disparities and Achieve Health Equity**

Jamboor Vishwanatha and Lavanya Vishwanatha, *AIM-AHEAD Coordinating Center*, United States

Presenter: Lavanya Vishwanatha, *AIM-AHEAD Coordinating Center*
Moderator: Charles Gillan, *Queen's University Belfast*

Abstract: The rapid increase in the volume of data generated through electronic health records (EHR) and other biomedical research presents exciting

opportunities for developing data science approaches (e.g., AI/ML methods) for biomedical research and improving healthcare. Many challenges hinder more widespread use of AI/ML technologies, such as the cost, capability for widespread application, and access to appropriate infrastructure, resources, and training. Additionally, lack of diversity of both data and researchers in the AI/ML field runs the risk of creating and perpetuating harmful biases in its practice, algorithms, and outcomes, thus fostering continued health disparities and inequities. Many underrepresented and underserved communities, which are often disproportionately affected by diseases and health conditions, have the potential to contribute expertise, data, diverse recruitment strategies, and cutting-edge science, and to inform the field on the most urgent research questions, but may lack financial, infrastructural, and data science training capacity to apply AI/ML approaches to research questions of interest to them. This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

The AIM-AHEAD Coordinating Center (A-CC) is a consortium of institutions and organizations that have a core mission to serve under-represented or underserved groups (minority populations, low socioeconomic, rural, sexual gender minorities) impacted by health disparities (e.g., Historically Black Colleges and Universities, Tribally Controlled Colleges and Universities, etc.). The A-CC focused initially on coordination, assessment, planning, and capacity building to enhance the use of artificial intelligence (AI) and machine learning (ML) in research among the consortium institutions and organizations; and to build and sustain trusted relationships between the consortium and groups impacted by health disparities. Through the AIM-AHEAD CONNECT, the networking platform of the A-CC, training, mentoring and networking resources available to the stakeholders will be demonstrated.

9:50 am – 10:10 am **High Performance Computing in Radiotherapy and PET Medical Imaging**

Othmane Bouhali, Maya Abi Aki, and Zakaria Ait Elcadi, *Texas A&M University at Qatar*

Presenter: Othmane Bouhali, *Texas A&M University at Qatar*
Moderator: Charles Gillan, *Queen's University Belfast*

Abstract: The high energy and medical physics group at Texas A&M University in Qatar is conducting research in areas of radiotherapy simulation and medical imaging. In the field of radiation therapy, Monte Carlo simulations are used to reconstruct the energy deposition due to radiation therapy dose. Advanced anatomically correct 4D computational phantom are used as an input to reconstruct realistic RT treatment planning and to measure the voxel-wise dose for different organs. In order to simulate the different tissues, specific phantoms were used to represent the density of each tissue. Experimental measurement at a hospital were compared to the simulation data.
In the field of medical imaging our group is actively involved in designs of new positron emission tomography (PET) scanners. PET is a non-invasive imaging

modality used mainly in oncology, radiology and neurology to assess functional changes in different areas of the body. A radiotracer consisting of a tracer coupled to a radionuclide (positron emitter) is administered to the patient to examine physiological functions of their body. Pairs of 511 keV photons will be emitted following the positron-electron annihilation and detected by the scintillation crystals arranged in rectangular blocks and positioned in a cylindrical gantry. Considered as the main building blocks of a PET scanner, the scintillation crystals absorb the high energy photons and convert them into optical photons. In this work, we use Monte Carlo methods to simulate a clinical PET scanner investigate and optimize the performances of new scanners. The use of a high performance computing machine (4200 cores) was crucial to simulate the different scenarios. We will present an overview of the research activities describe above and will discuss the crucial role of the HPC tools in this research.

10:10 am – 10:30 am **Performance Analysis Tools Are Used as Guidelines to Build Efficient, HPC-Scalable Human Digital Twins**

Jose Luis Estragués Muñoz, *Barcelona Supercomputing Center*, Spain; Alfonso Valencia, *Barcelona Supercomputing Center & ICREA - Institució Catalana de Recerca i Estudis Avançats*, Spain; Arnau Montagud, Thaleia Ntiniakou, Miguel Ponce-de-León, Jose Carbonell Caballero, and Davide Cirillo, *Barcelona Supercomputing Center*, Spain

Presenter: Jose Luis Estragués Muñoz, Barcelona Supercomputing Center, Spain
Moderator: Andrea Townsend-Nicholson, *University College London*

Abstract: Precision medicine requires high performance computing (HPC) platforms to model complex and massive volumes of biomedical data. In particular, multiscale cell simulators have proven useful in several precision medicine applications thanks to their ability to help uncover and explain disease mechanisms (Ponce-de-Leon et al., 2022). Nevertheless, these simulators usually consider thousands of cells, rarely reaching the cm3 size.

PhysiCell is an open-source cell simulator that renders multicellular systems as many interacting cells that respond to and influence their microenvironment (Ghaffarizadeh et al., 2018). The simulation size scale of these tools is an open problem whose solution is attached to the efficient usage of HPC resources. While the state of the art has reached the tissue level, that is, simulations up to 109 cells (Montagud et al., 2021), the ultimate goal in this area is represented by larger and more realistic simulations, collectively called digital twins.

We have reached an efficient, HPC-scalable multiscale simulation tool by working on two fronts. On the one hand, we have analyzed PhysiCell's performance and scalability efficiency. We detected the limiting factors of the scalability of PhysiCell in the Marenostrum 4 supercomputer. We examined in-depth the execution behavior of each process within the simulation using detailed metrics provided by the Barcelona Supercomputing Center's performance tools (Munera et al., 2020). We discovered that the instructions per cycle (IPC) scalability, the number of instructions and the load balance are critical for the efficiency of PhysiCell's scalability.

On the other hand, we have built an MPI implementation of PhysiCell to distribute the simulation across multiple computation nodes and named it PhysiCell-X (https://gitlab.bsc.es/gsaxena/physicell_x). The distribution of the domain allows larger simulations to be modeled across compute nodes, potentially enabling real-sized tumor simulations. This effectively changes the limiting factor of the simulations from the memory size of one node to the number of compute nodes of the HPC cluster.

Despite this change in limiting factor, we wanted to compare the efficiencies of the distributed PhysiCell-X with the shared memory PhysiCell. Specifically, we are currently applying the same efficiency methodology to PhysiCell-X to reach an optimized scalability efficiency, obtaining promising preliminary results.

Modeling is among the most promising techniques in precision medicine, but it is also a very demanding task in terms of energy resources. Thus, the field would benefit from examples of tools properly adapted and optimized to high-performance clusters. This work has set the first steps in the biomedical field to optimize the use of multiscale cell simulations in HPC platforms leading the way to the actionable realization of digital twins.

10:30 am – 10:55 am **HPC for Accelerated Precision Medicine on AWS**

Aniket Deshpande, *Amazon Web Services*, United States, and Brian Skjerven, *Amazon Web Services,* United Kingdom

Presenter: Brian Skjerven, *Amazon Web Services,* United Kingdom

Moderator: Andrea Townsend-Nicholson, *University College London*

Abstract: Biological data is increasingly getting more massive, complex, and data-dense. The primary reasons are due to the release of next-generation, higher-throughput laboratory instruments in genomics, cryo-EM, digital pathology, etc., and population-scale, multi-modal data commons like the NIH AllOfUs, UKBiobank, Genomics England, NCI Genomic Data Commons, etc. Biomedical researchers are now looking to leverage the latest in cloud-based HPC and applied ML technologies to transfer, store, process, and analyze data at scale and securely share those findings with other collaborators. In this presentation, we will cover research and clinical use cases on how customers are using AWS HPC and ML services to accelerate drug discover workloads and enable precision medicine in the clinic.

Some examples in genomics include how Baylor College of Medicine has analyzed 5000 Whole Genome Sequencing (WGS) samples/month for NIH's AllOfUs clinical genomics workloads using FPGA instances, and how AstraZeneca runs 51 billion tests in 1 day using AWS Batch and AWS Lambda. In structure based drug discovery, Vertex and Eli Lilly are using AWS ParallelCluster along with GPU and HPC instances to reduce Cryo-EM data processing costs by 50%, while researchers at Columbia university optimizing Protein folding costs with OpenFold using AWS Batch. For imaging, digital pathology startups like Paige.ai using GPUs, and FSx for Lustre filesystem to deliver high performance for ML training and HPC applications in the cloud. Finally, discover how Allen Institute building multi-modal brain datasets as part of the NIH's BICAN initiative using transcriptomics data processing on AWS. Researchers can expect to learn how AWS' customers are building 'fit-for-purpose' workloads for genomics, cryo-EM, digital pathology, and applied AIML along with other resources such as price/performance benchmarks, rightsizing instance recommendations, reference architectures, available workshops, AWS Quickstarts, etc. Finally, we will briefly cover AWS HealthAI's newly released managed services for genomics and medical imaging; Amazon Omics and Amazon HealthLake Imaging.

10:55 am – 11:30 am **Break**

11:30 am – 12:15 pm **Panel on "A FAIR HPC Ecosystem for Innovation in Medicine - How Do We Work Together?"**

Moderator: Eric Stahlberg, *Frederick National Laboratory for Cancer Research*, United States
Othmane Bouhali, *Texas A&M University at Qatar,* Qatar
Jamboor Vishwanatha, *AIM-AHEAD Coordinating Center,* United States
Peter V. Coveney, *University College London*, United Kingdom

12:15 pm – 12:55 pm **Open Discussion: Creating a Medical Digital Twin Ecosystem**
Moderator: Eric Stahlberg*, Frederick National Laboratory for Cancer Research*

12:55 pm – 1:00 pm **Wrap Up**

Organizing Committee
Eric Stahlberg – Frederick National Laboratory for Cancer Research (United States)
Charles Gillan – Queen's University Belfast (UK)
Arnau Montagud – Barcelona Supercomputer Center (BSC) (Spain)
Jan Nygard – Cancer Registry of Norway (Norway)
Thomas Steinke – Zuse Institute Berlin (Germany)
Andrea Townsend-Nicholson – University College London (UK)
Lynn Borkon – Frederick National Laboratory for Cancer Research (United States)
Petrina Hollingsworth – Frederick National Laboratory for Cancer Research (United States)

Presenter Bios

Othmane Bouhali, Texas A&M University at Qatar
Dr. Bouhali received his PhD in Science from the Universite Libre de Bruxelles in 1999. Since 1994, he has been participating to the Large Hadron Collider Project (LHC) at the European Organization for Nuclear and Particle Physics (CERN). He served as the head of the computing group at the high energy physics institute in Brussels. He is Director of Research Computing and Research Professor at Texas A&M University at Qatar. His field of expertise includes high performance computing, artificial intelligence, radiation detectors and medical physics. He is the founder of the TAMU-Q Advanced Scientific Computing (TASC)center. He is affiliated with the Qatar Computing Research Institute (QCRI) and visiting professor at Georgetown University at Qatar. He is leading the high energy and medical physics group at TAMUQ. He is the recipient of Dean’s Distinguished Achievement Award (twice), the teaching excellence award (twice) and the received the best UREP research project with his students from the Qatar National Research Fund.

Peter V. Coveney, University College London
Peter Coveney is a professor of physical chemistry, honorary professor of computer science, and Director of the Centre for Computational Science (CCS) and Associate Director of the Advanced Research Computing Centre at University College London (UCL). He is also professor of applied high performance computing at the University of Amsterdam (UvA) and professor adjunct at the Yale School of Medicine, Yale University. He is a fellow of the Royal Academy of Engineering and Member of Academia.

Jose Luis Estragués Muñoz, Barcelona Supercomputing Center
Jose L. Estragués is a research engineer at the Barcelona Supercomputer Center - Spanish National Center of Computation. His expertise lies in hardware architecture and high performance computation, with a background as a computer engineer from the Polytechnic University of Catalonia (UPC). He also completed a master’s program in Innovation and Research in Informatics at UPC. For the past two years, Estragués has been part of the computational biology group led by Alfonso Valencia. His work focuses on evaluating multi-scale cell simulators on different HPC platforms and addressing the challenges of scaling these tools.

Charles Gillan, Queen’s University Belfast
Dr. Gillan studied applied mathematics and physics at Queen's University Belfast, completing his PhD, 1988, in the application of high performance computing to study low energy electron molecule scattering. He continued his research in computational chemistry as a post doc at the IBM Almaden Research Centre in San Jose.

After a career in embedded software development in the telecommunications industry, he returned to work at the ECIT Institute at QUB in 2004. He continues his research in high performance computing applied to physics, chemistry, and more recently physiology and has published over 50 research papers. His current research applies machine learning and artificial intelligence (using HPC to train complex models) to predictive analysis of streams of physiological parameters received from patients in intensive care.

Brian Skjerven, Amazon Web Services
Brian Skjerven is a Solutions Architect at AWS specializing in high performance computing, particularly for customers in the healthcare and life sciences space. His background is in applied math and parallel computing, and prior to joining AWS he was at the Pawsey Supercomputing Centre (Australia) and Argonne National Labs.

Eric Stahlberg, Frederick National Laboratory for Cancer Research
Eric Stahlberg serves as the director of Cancer Data Science Initiatives at the Frederick National Laboratory for Cancer Research (FNLCR). Joining the team at Frederick in 2011 to establish and lead the bioinformatics core supporting the NCI Center for Cancer Research, Dr. Stahlberg shifted his attention in 2014 to lead a new NCI CBIIT initiative to accelerate cancer research through applications of high-performance computing. Working collaboratively with NCI leadership Dr. Stahlberg helped established the NCI-US Department of Energy Collaboration as well as Accelerating Therapeutics for Opportunities in Medicine (ATOM), a public-private collaboration to dramatically increase the pace and success of new treatments. Driven to drive advances at the intersection of leading-edge science and computing, Dr. Stahlberg continues to build the cross-disciplinary community through efforts with the Computational Approaches for Cancer and HPC Applications of Precision Medicine workshops. In 2017, he was recognized as one of FCW’s Federal 100. Stahlberg holds a PhD in computational chemistry from the Ohio State University and bachelor’s degrees in computer science, chemistry, and mathematics.

Andrea Townsend-Nicholson, University College London
Dr. Townsend-Nicholson obtained her Bachelor of Science Specialist degree in molecular genetics and molecular biology, with a major in zoology and a minor in religion, from the University of Toronto in 1986. She moved to the Laboratoire de Génétique Moléculaire des Eucaryotes in Strasbourg France, investigating the establishment of the dorsoventral polarity axis in Drosophila melanogaster and obtained her doctorate in cellular and molecular biology from the Université Louis Pasteur in 1990. From 1991 to 1996, she switched from transcriptional studies to cell signalling, studying mammalian G protein-coupled receptors as a postdoctoral fellow in the Neurobiology Division of the Garvan Institute of Medical Research (Sydney, Australia). During this time, she cloned and characterised several adenosine receptor subtypes and learned about the benefits of wide-brimmed hats and factor 50 sunscreen. Having started her research career at University College (University of Toronto) in Canada, she is now at University College London, where she was appointed as a member of academic staff in 2001, following three and a half years of postdoctoral study and 18 months as a British Heart Foundation Intermediate Research Fellow.

Jamboor Vishwanatha, AIM-AHEAD Coordinating Center
Dr. Vishwanatha is a Regents Professor, Vice President, and Founding Director of the Texas Center for Health Disparities at the University of Texas Health Science Center at Fort Worth. He is a principal investigator of the National Research Mentoring Network, a NIH Common Fund initiative to provide mentorship, networking and professional development for a diversified biomedical and behavioral workforce. He is also a principal investigator of the NIH Specialized Center of Excellence in Health Disparities, AIM-AHEAD Coordinating Center, as well as the Texas CEAL Consortium. Dr. Vishwanatha received the Presidential Award for Excellence in Science, Mathematics and Engineering Mentoring (PAESMEM) from US White House in October 2019.

Lavanya (Lavi) Vishwanatha, AIM-AHEAD Coordinating Center
Lavi Vishwanatha is the Research Enterprise Solutions Director for the Artificial Intelligence and Machine Learning to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program at the University of North Texas Health Science Center at Fort Worth. In her role as the director, Lavi coordinates all the project activities of the program and the administrative core. She oversees the programs, integration, accomplishments, and reporting of the functional cores as well as the Leadership core of the AIM-AHEAD Coordinating Center.