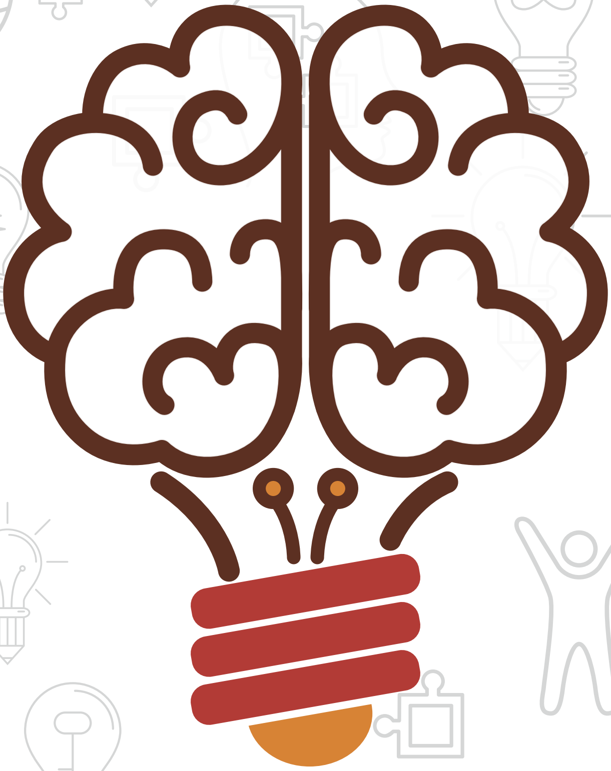


10th International BSC Severo Ochoa Doctoral Symposium 2023

9th - 10th May, 2023

Book of Abstracts

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA
SEVERO
OCHOA

Barcelona Supercomputing Center
Plaça Eusebi Güell, 1-3
08034 Barcelona (Spain)

education@bsc.es
www.bsc.es

@BSC_CNS
/BSCCNS
/BSC_CNS
/barcelona-supercomputing-center
/BSCCNS

Book of Abstracts
10th International BSC Severo Ochoa Doctoral Symposium 2023

Editor
Carolina Olmo

Cover
Design based on artwork created by macrovector.com

This book is available at
https://www.bsc.es/education/predoctoral-phd/doctoral-symposium

published by
Barcelona Supercomputing Center

Severo Ochoa Excellence Center (2023-2027) with reference CEX2021-001148-S (BSC)

10th Edition, May 2023

ACKNOWLEDGEMENTS

The BSC Education & Training team gratefully acknowledges all the PhD candidates, advisors, Postdoc researchers, experts and especially the Keynote Speaker Dr. Alba Cervera-Lierta, for contributing to this Book of Abstracts and participating in the 10th International BSC Severo Ochoa Doctoral Symposium 2023. We also wish to thank the volunteers that supported the organisation of the event: Albert Cañellas Sole, María Chavero Díez, Carlos García Guillamón, Othmane Hayoun Mya, Josep Pocurull, Iria Pose Lagoa, Guillermo Prol Castelo, Ignasi Puch Giner, Alba Santos Espeso, Ivan Vargas Valdivieso.

BSC Education & Training team
education@bsc.es

EDITORIAL COMMENT

We are proud to present the Book of Abstracts for the 10th International BSC Severo Ochoa Doctoral Symposium.

During more than fifteen years, the Barcelona Supercomputing Center has been receiving undergraduate, master and PhD students, and providing them training and skills to develop a successful career. Many of those students are now researchers and experts at BSC and in other international research institutions.

In fact, the number of students has never decreased. On the contrary, their number and research areas have grown and we noticed that these highly qualified students, especially the PhD candidates, needed a forum to present their findings and fruitfully exchange ideas. As a result, in 2014, the first BSC Doctoral Symposium was born.

In this 10th edition of the International BSC Severo Ochoa Doctoral Symposium we are offering a keynote talk titled “The quest for the quantum computer" by Dr. Alba Cervera-Lierta and a tutorial on Cross-Departmental Collaboration at BSC.

The talks will be held in five different sessions and have been distributed from an interdisciplinary approach. They will tackle the topics of:

- Genomics
- Simulations and Modelling
- Medical and Health Applications
- Tools Development
- Artificial intelligence and its Applications.

The posters will be exhibited and presented during three poster sessions that will give the authors the opportunity to explain their research and results.

This Book of Abstracts is the result of their contributions.

WELCOME ADDRESS

I am delighted to welcome all the PhD students, Postdoc researchers, advisors and experts to the 10th International BSC Severo Ochoa Doctoral Symposium.

Once again, in this 10th edition of the International BSC Severo Ochoa Doctoral Symposium, the goal of the occasion is to provide a framework to share research results of the projects developed by PhD thesis that use High Performance Computing in some degree. The symposium was conceived in the framework of the Severo Ochoa Program at BSC, following the project aims regarding the talent development and knowledge sharing and provides a forum for PhD students considering both the ones just beginning their research and others who have developed their research activities during several years.

As a consequence, I highly appreciate the support provided by BSC and the Severo Ochoa Center of Excellence Programme that make possible to celebrate this event.

I am very grateful to the BSC directors for supporting the symposium, to the group leaders and to the advisors for encouraging the participation of the students in the event. Moreover, I wish to specially thank the keynote speaker Alba Cervera-Lierta and for her willingness to share with us her knowledge and expertise.

I would also like to thank all PhD students and Postdoc researchers for their papers and presentations. I wish you all the best for your career and I really hope you enjoy this great opportunity to meet other colleagues in some cases for the first time and share your experiences.

Last but not least, I wish to thank the Education and Training Team who put great effort and enthusiasm on the event.

Dr. Maria Ribera Sancho
Manager of BSC Education & Training

KEYNOTE SPEAKER

Alba Cervera-Lierta
Senior Researcher at the Barcelona Supercomputing Center

The quest for the quantum computer

Multiple technological advances marked the second half of the 20th century. Among these, the invention of the transistor carried out the development of the computational era. It was also a golden century for fundamental physics, with the standard model and its experimental validation being one of its greatest exponents. However, advances in the study of the most fundamental properties of matter, described by quantum mechanics, highlighted the experimental limits to understanding it. Even the newest and most advanced computers were insufficient to describe large quantum systems due to a fundamental limitation: the space to perform the simulation grows exponentially with the number of particles we want to study. That's when, in the 1980s, some physicists suggested why not design a computer that obeys the same quantum laws of the systems we want to study, i.e., a quantum computer. Although quantum physics had already given rise to applications such as GPS, magnetic resonance, or the laser, the technology was not ready for this new and ambitious proposal at that time. Second-generation quantum technologies (communication, computing, and sensors) were developed mainly at the theoretical level, giving rise to unique potential applications not only in fundamental physics, but also in cryptography, mathematics, chemistry and material science.

A universal fault-tolerant quantum computer that can efficiently solve such challenging problems requires millions of quantum bits (qubits) with low error rates and long coherence times. While the experimental advancement toward realizing such devices will potentially take decades of research, noisy intermediate-scale quantum (NISQ) computers are now a reality. These computers are composed of hundreds of noisy qubits, i.e., not error-corrected qubits, and perform imperfect operations in a limited coherence time. In the search for quantum advantage with these devices, algorithms have been proposed for applications in various disciplines. Such algorithms aim to leverage the limited available resources to perform classically challenging tasks. HPC infrastructures will play a central role in developing these near-term applications of quantum computers. The quantum processing units (QPU) will be integrated into these infrastructures and become accelerators for some algorithms.

In this talk, I will explain the brief history of quantum technologies, with particular emphasis on near-term quantum computation, where we are, and the challenges and applications of this technological revolution.

Alba Cervera-Lierta is a Senior Researcher at the Barcelona Supercomputing Center. She earned her PhD in 2019 at the University of Barcelona, where she studied her physics degree and a Msc in particle physics. After her PhD, she moved to the University of Toronto as a postdoctoral fellow at the Alán Aspuru-Guizik group. She works on near-term quantum algorithms and their applications, high-dimensional quantum computation, and artificial intelligence strategies in quantum physics. Since October of 2021, she is the coordinator of the Quantum Spain project, an initiative to boost the quantum computing ecosystem that will acquire and operate a quantum computer at the BSC-CNS.

TUTORIAL

“Cross-Departmental Collaboration at BSC: Success Stories and a Workshop”

Cross-departmental collaboration is when a group of people with different job responsibilities or functions come together and work towards a common goal, project or solution. It can also be when people with similar job responsibilities or functions from different departments meet and work together. It leads to more ideas, shared workloads, significant process improvements and a culture of continuous learning.

The objective of this tutorial is to learn about the potential of cross-departamental collaborations and discover who you could colaborate with.

This tutorial will be divided in two parts:

14:00h - 15:00h **Alba Jené** (Bioinformatics Unit Coordinator, LS), **Carlos García y Guillermo Marin** (Data Pre&Post Processing Researchers, CASE) and **Daniele Lezzi** (Workflows and Distributed Computing Established Researcher, CS) will present several cross-departmental collaborations at BSC and share tips and advice on how to network and collaborate.

15:30h - 17:30h Hands-on workshop: Discover who you could collaborate with.
Following a speed-dating structure, participants will have 6’ to share one-on-one informal discussions with peer colleagues to ask questions and explore specific areas of their research and practices (several rounds of discussions). The aim is to find a person to collaborate with or to explore what others do to engage in a potential future collaboration.

PROGRAM

DAY 1 (May 9th)

Start time	Activity	Speaker/s	Chair

8.30h Registration

9.00h	Welcome and opening	**Josep Mª Martorell,** BSC Associate Director	**Maria Ribera Sancho**
9.20h	Keynote talk: The quest for the quantum computer	**Alba Cervera-Lierta,** Senior Researcher at the Barcelona Supercomputing Center and Quantum Spain project coordinator	
	Abstract:In this talk, I will explain the brief history of quantum technologies, with particular emphasis on near-term quantum computation, where we are, and the challenges and applications of this technological revolution.		

10.30h Event Photo

10.40h Coffee break & First Poster Session

Spatially-resolved multiscale models shed light into personalized drug treatments, **Alejandro Madrid Valiente** (LS)

Characterizing the Impact of Graph-Processing Workloads on Modern CPU's Cache Hierarchy, **Alexandre Valentin Jamet** (CS)

Representational Learning for the Study of Breast Cancer Progression through Pseudo-Time, **Guillermo Prol Castelo** (LS)

11.40h First Talk Session: Genomics

11.40h	Horizontal Gene Transfer in Asgard Archaea, **Saioa Manzano-Morales** (LS)	**Toni Gabaldón**
12.00h	Exhaustive Variant Interaction Analysis using Multifactor Dimensionality Reduction, **Gonzalo Gómez** (CS)	
12.20h	Antibody-Derived Tag normalization for ASAP andscCUT&TAG-PRO, **Xavier Soler-Sanchis** (LS)	
12.40h	Microbiome profiling from Fecal Immunochemical Test reveals microbial signatures with potential for Colorectal Cancer screening, **Olfat Khannous-Lleiffe** (LS)	

13.00h Lunch Break

14.00h Tutorial

"Cross-Departmental Collaboration at BSC: Success Stories and a Workshop"	**Alba Jené** (Bioinformatics Unit Coordinator, LS)**, Carlos García and Guillermo Marin** (Data Pre&Post Processing Researchers, CASE), **Daniele Lezzi** (Workflows and Distributed Computing Established Researcher, CS) and and **Albert Soret** (Earth System Services Group Leader, ES)

The objective of this tutorial is to learn about the potential of cross-departamental collaborations and discover who you could colaborate with.

14:00h - 15:00h Speakers will present several cross-departmental collaborations at BSC and share tips and advice on how to network and collaborate.

15:30h - 17:30h Hands-on workshop: Discover who you could collaborate with.
Following a speed-dating structure, participants will have 6' to share one-on-one informal discussions with peer colleagues to ask questions and explore specific areas of their research and practices (several rounds of discussions). The aim is to find a person to collaborate with or to explore what others do to engage in a potential future collaboration.

17.30h Adjourn

DAY 2 (May 10th)

Start time	Activity	Speaker/s	Chair
9.00h	Opening of the second day		
9.10h	**Second Talk Session: Simulations and Modelling**		
9.10h	Consensus Essential Dynamics Analysis: Application to Biomolecular Simulations, **Luis Jordà** (LS)		**Sara Royuela**
9.30h	Understanding North Atlantic deep-water formationdrivers in an eddy-resolving climate model, **Eneko Martin-Martinez** (ES)		
9.50h	Towards data driven reduced order models for the automotive industry, **Benet Eiximeno Franch** (CASE)		
10.10h	Climate Services Ecosystems: What Are They and Why Are They Are Important?, **Carmen Gonzalez Romero** (ES)		
10.30h	Fault-tolerant applications through OpenMP, **Adrián Munera** (CS)		
10.50h	**Coffee break & Second Poster Session**		
	Open-Source GEMM Hardware Kernels Generator: Toward Numerically-Tailored Computations, **Louis Ledoux** (CS)		
	Prediction of Bacterial Interactomes Based on Genome-Wide Coevolutionary Networks: an Updated Implementation of the ContextMirror Approach, **Miguel Fernández** (LS)		
	Parallelizing Recurrent Neural Networks and variantsusing OmpSs, **Robin Sharma** (CS)		
	Scaling RTL Simulations with Metro-MPI, **Guillem López Paradís** (CS)		
	Sorting Impact in Decision Support Benchmarks TPC-H and TPC-DS for Row-Oriented Relational Databases, **Iván Vargas Valdivieso** (CS)		
11.40h	**Third Talk Session: Medical and Health Applications**		
11.40h	Recapitulating experimental drug synergies in AGSthrough multiscale agent-based simulations, **Othmane Hayoun-Mya** (LS)		

12h	What is the Added Value of Climate Information for Predicting Infectious Disease Outbreaks, **Chloe Fletcher** (ES)	**Valentina del Olmo**
12.20	Machine Learning approaches for thecharacterization of COPD, **Iria Pose Lagoa** (LS)	
12.40	Ranking Allosteric Activators of AMPK by Absolute Binding Free Energy Calculations with Fun-metaD, **Rhys Evans** (LS)	
13.00h	**Lunch Break**	
14:00h	**Fourth Talk Session: Tools Development**	
14:00h	NTRU Cryptosystem: A solution to the quantum threat, **Miquel Guiot Cusidó** (CS)	**Petar Radojkovic**
14:20h	JLOH: Inferring Loss of Heterozygosity Blocks fromShort-Read Sequencing Data, **Matteo Schiavinato** (LS)	
14:40h	VAQUERO: A Scratchpad-based Vector Acceleratorfor Query Processing, **Julián Pavón Rivera** (CS)	
15:00h	TIGER: The gene expression regulatory variation landscape of human pancreatic islets, **Lorena Alonso** (LS)	
15:20h	A Stochastic Method for Solving Time-Fractional Differential Equations, **Nicolas L. Guidotti** (CS)	
15.40h	**Coffee break & Third Poster Session**	
	Variability of univariate and compound hot and dry events, **Alvise Aranyossy** (ES)	
	A Stroke Risk Assessment Tool for Edge-to-Cloud Platforms, **Fatemeh Baghdadi** (LS)	
	Numerical analysis of MHD flow of magnetically confined plasma in cylindrical geometry usingOpenFOAM, **Robert Benassai** (CASE)	
	In Silico Bioprospecting of Enzymatic PEF Synthesis and Degradation, **Rubén Muñoz Tafalla** (LS)	
16.40h	**Fifth Talk Session: Artificial intelligence and its Aplications**	
16:40h	Multi-dimensional Fourier series with quantum circuits, **Berta Casas Font** (CASE)	**Artur García**
17:00h	Tailored Molecular Modelling and Machine Learning solutions for small-molecule drug discovery, **Isaac Filella Merce** (LS)	
17:20h	New Tensor Network Structures in 1.5D, **Sergi Masot Llima** (CASE)	
17:40h	DNN Acceleration in the limits of Energy Efficiency, **Jordi Fornt** (CS)	
18.00h	**Final remarks and End of Doctoral Symposium**	
18:20h	**Cool off**	

Abstracts

TIGER: the Gene Expression Regulatory Variation Landscape of Human Pancreatic Islets

Alonso L.[1,*], The T2DSystems Consortium[+], Torrents D.[1,2,*]

[1] Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain

[1]lorena.alonso@bsc.es, [1,2]david.torrents@bsc.es

[+] Members of the T2DSystems Consortium are provided in Acknowledgement section

***Keywords*— Pancreatic islets, Functional interpretation, Public resource**

Extended Abstract

A. Introduction

One of the main focuses of computational genomics is to broaden the understanding of the genomic basis of complex diseases, such as Type 2 Diabetes (T2D). To that extent, first, the discovery of genomic changes (variants) associated with the disorder enhances the early detection and prevention of the disease development. Then, the comprehension of the molecular mechanisms underlying the associations facilitates the development of new drugs and treatments. Despite the broad use of Genome Wide Association Studies (GWAS) has facilitated the discovery of a large list of variants associated with complex disorders, the vast majority of these signals still remain without functional interpretation [1]. To approach this ongoing and still challenging problem, gene expression variation and regulatory regions analyses have enhanced the detection of numerous associations between variants and their change in gene expression, and the creation of wide catalogues of disease-related regulatory elements, such as enhancers and promoters. Remarkably, the transcriptomic and epigenetic study of disease-related tissues facilitates the understanding of the molecular mechanisms underlying GWAS disease-susceptibility loci. This is the case of pancreatic islets from which progressive failure and dysfunction is related to the development of T2D [2, 3]. Despite its interest to gain insights of the disease, the many complexities surrounding the accessibility and analysis of pancreatic islets has limited the advance on the study of the genetic and regulatory landscape of human islets and T2D [4].

B. Materials and Methods

Here, within the context of the T2DSystems, a Horizon2020 Project, we collected RNA-seq and genotyping data from 514 human islet samples and performed harmonisation, quality control, genotype phasing and imputation. We integrated a) GWAS associations from the summary statistics of multiple studies and large T2D meta-analyses, b) variant annotation, characterization and gene functional impact, c) epigenomic marks from islet DNA-methylation sites, chromatin accessibility and CHiP-seq profiles, d) gene, lncRNAs and islet regulome annotations, e) gene expression from normalised islet RNA-seq counts, microarrays and other tissues expression, and f) computed expression quantitative loci (eQTL) and combined allelic specific expression (cASE) and created the largest regulatory variation database from human pancreatic islets.

C. Results

We developed the Translational human Islet Genotype tissue-Expression Resource (TIGER), a large human islets regulatory expression database (http://tiger.bsc.es) [5]. This database contains information for more than 27 million variants and 59,625 genes and incorporates a genome browser to ensure the comprehensive data integration. It encloses tools for visualising, querying, and downloading human islet data enhancing the study of T2D and other islet-related diseases. TIGER facilitates follow-up by providing genetic and molecular findings related to T2D pathophysiology with a gene or a variant summary, eQTL and cASE results, associations with T2D and other related traits or diseases, genomic context information such as the islet chromatin landscape and direct access to other genomic databases.

D. Conclusions

The comprehensive collation in TIGER of genomic, transcriptomic and epigenetic human islet datasets, and the integration with T2D GWAS and regulatory variation, represents a formidable resource to interrogate the molecular aetiology of beta-cell failure.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program T2Dsystems under grant agreement no. 667191. Piron A.[3,4,*], Morán I.[1,*], Guindo-Martínez M.[1], Bonàs-Guarch S.[5,6], Atla G.[5,6], Miguel-Escalada I.[5,6], Royo R.[1], Puiggròs M.[1], Garcia-Hurtado X.[5,6], Suleiman M.[7], Marselli L.[7], Esguerra J.L.S.[8], Turatsinze J-V.[3], Torres J.M.[9,10], Nylander V.[11], Chen J.[12], Eliasson L.[8], Defrance M.[3], Amela R.[1], MAGIC[24], Mulder H.[13], Gloyn A.L.[10,11,14,15,16], Groop L.[8,13,17], Marchetti P.[7], Eizirik D.L.[3,18], Ferrer J.[5,6,19], Mercader J.M.[1,20,21,22,*] & Cnop M.[3,23,*] . [1] Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain. [2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. [3] ULB Center for Diabetes Research, Université Libre de Bruxelles, Belgium. [4] Interuniversity Institute of Bioinformatics in Brussels, Belgium. [5] Bioinformatics and Genomics Program, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Spain. [6] Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas, Spain. [7] Department of Clinical and Experimental Medicine and AOUP Cisanello University Hospital, University of Pisa, Italy. [8] Unit of Islet Cell Exocytosis, Lund University Diabetes Centre, Sweden. [9] Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, UK. [10] Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, UK. [11] Oxford Centre for Diabetes, Endocrinology, and Metabolism, Radcliffe Department of Medicine, University of Oxford, UK. [12] Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, UK. [13] Unit of Molecular Metabolism, Lund University Diabetes Centre, Sweden. [14] Division of Endocrinology, Department of Pediatrics, Stanford University School of Medicine, CA, USA. [15] NIHR Oxford Biomedical Research Centre, Churchill Hospital, UK. [16] Stanford Diabetes Research Centre, Stanford University, CA, USA. [17] Finnish Institute of Molecular Medicine Finland (FIMM), Helsinki University, Finland. [18] WELBIO, Université Libre de Bruxelles, Belgium. [19] Section of Epigenomics and Disease, Department of Medicine, Imperial College London, UK. [20] Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, MA, USA. [21] Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, MA, USA. [22] Department of Medicine, Harvard Medical School, MA, USA. [23] Division of Endocrinology, Erasmus Hospital, Université Libre de Bruxelles, Belgium. [24] Members of the MAGIC consortium. [*] These authors have contributed equally.

References

[1] Alonso L., Morán I., Salvoro C. & Torrents D. "In search of complex disease risk through genome wide association studies". Mathematics. 2021.

[2] Del Guerra S. et al. "Functional and molecular defects of pancreatic islets in human type 2 diabetes". Diabetes. 2005.

[3] Eizirik D.L. et al. "Pancreatic β-cells in type 1 and type 2 diabetes mellitus: different pathways to failure". Nature Reviews Endocrinology. Nature Research. 2020.

[4] Gloyn A.L. et al. "Every islet matters: improving the impact of human islet research". Nature Metabolism. 2022.

[5] Alonso L., Piron A., Morán I. et al. "TIGER: The gene expression regulatory variation landscape of human pancreatic islets". Cell Reports. 2021.

Author Biography

Lorena Alonso obtained her bachelor's degree in Mathematics at Universitat de Barcelona in 2011. From that point, and until finishing three Master Sciences studies in Applied Statistics, Bioinformatics and Biostatistics, she worked in Applied Mathematics. In 2016, she started to work as a research engineer at Dr. David Torrents' lab. In January 2023, she finished her PhD thesis in Biomedicine, which is entitled "From the discovery of epistatic events in Type 2 Diabetes Mellitus to the study of related gene expression regulatory variation". Now, Lorena is a postdoctoral researcher in Computational Genomics. Her research is centered in broadening the genomic understanding of complex diseases.

Variability of univariate and compound hot and dry events

Alvise Aranyossy*†, Markus Donat*‡, Paolo de Luca*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat de Barcelona, Barcelona, Spain
‡Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
E-mail: {alvise.aranyossy, markus.donat, paolo.deluca}@bsc.es

***Keywords*—Hot-dry compound events, drought indices.**

I. Extended Abstract

The vulnerability of societal and economical systems to future extreme events is an emerging topic, which is not limited anymore to the scientific environment but is also taken into consideration in decision-making processes. More specifically, a category of extreme events that have shown to create more damage is the so-called compound extremes, a combination of multiple drivers or hazards which can lead to a significant increase in risk and impacts, such as hot and dry summers or storm surges coinciding with severe precipitations or anomalous storm tracks [1] [2] [3]. This study is part of a bigger doctoral project which aims to study univariate and compound extremes in reanalysis data and evaluate the skill of extreme events both in real-forecast decadal hindcasts and in a perfect model setup. In addition, we aim to explore the underlying teleconnections between these events and the large-scale climate to increase the skill of predictions, and the application of new methods to extend the prediction of univariate and compound extreme events beyond the 10-year forecast covered by decadal predictions. This abstract focuses on the variability of univariate and compound extremes in observation and reanalysis from 1961 to 2016, thus representing the preliminary results of the project's first stage.

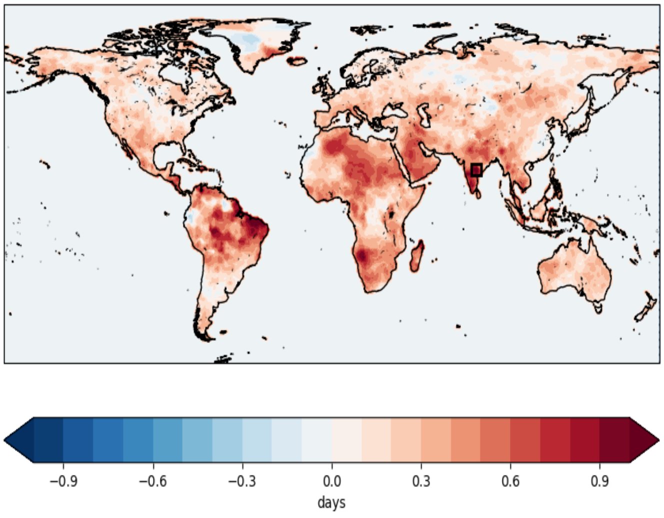


Fig. 1. Linear trend of yearly SPEI3_dry from 1961 to 2016. The black box in the Indian subcontinent represents the area for which the time series is shown in Figure 2.

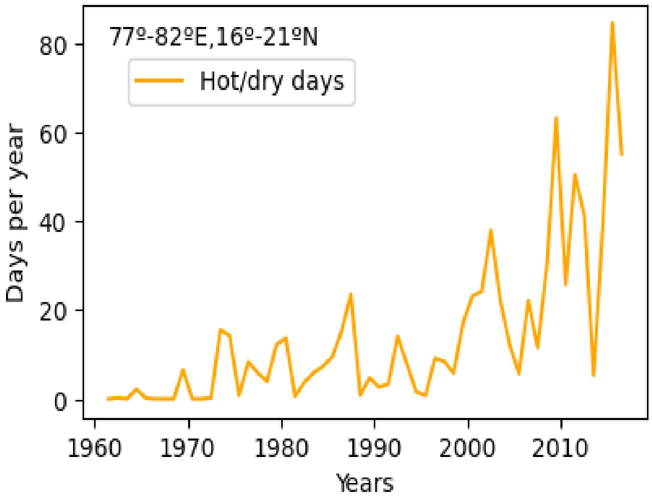


Fig. 2. Yearly SPEI3_dry time series for the hotspot area in the Indian subcontinent. Area correspond to the spatial mean of the black box in Figure 1.

A. Methodology

The initial part of the project focuses on understanding the variability of univariate and hot and dry compound extremes in observations. For drought indices, we use the Standardized Precipitation Index (SPI, [4]) and the Standardized Precipitation Evapotranspiration Index (SPEI, [5]), using the Hargreaves method to approximate Potential Evapotranspiration (PET, [6]). The SPI provides information regarding meteorological droughts uniquely in terms of precipitation, while SPEI represents the general water availability considering also atmospheric water demand. We calculate both SPI and SPEI for 3-, 6-, and 12-month accumulation periods. For the hot events, we select the days above the 90th percentile of the daily maximum temperature (*tx*). The percentile is calculated for each day of a year with a 5-day window. We run these indices on monthly values of accumulated precipitation from the Rainfall Estimates on a Gridded Network dataset (REGEN, [7]), while we take daily maximum and minimum temperatures from the Berkeley Earth Surface Temperatures dataset (BEST, [8]). To select hot/dry compound events, we select all the tx90p days which occur during dry conditions as measured by an

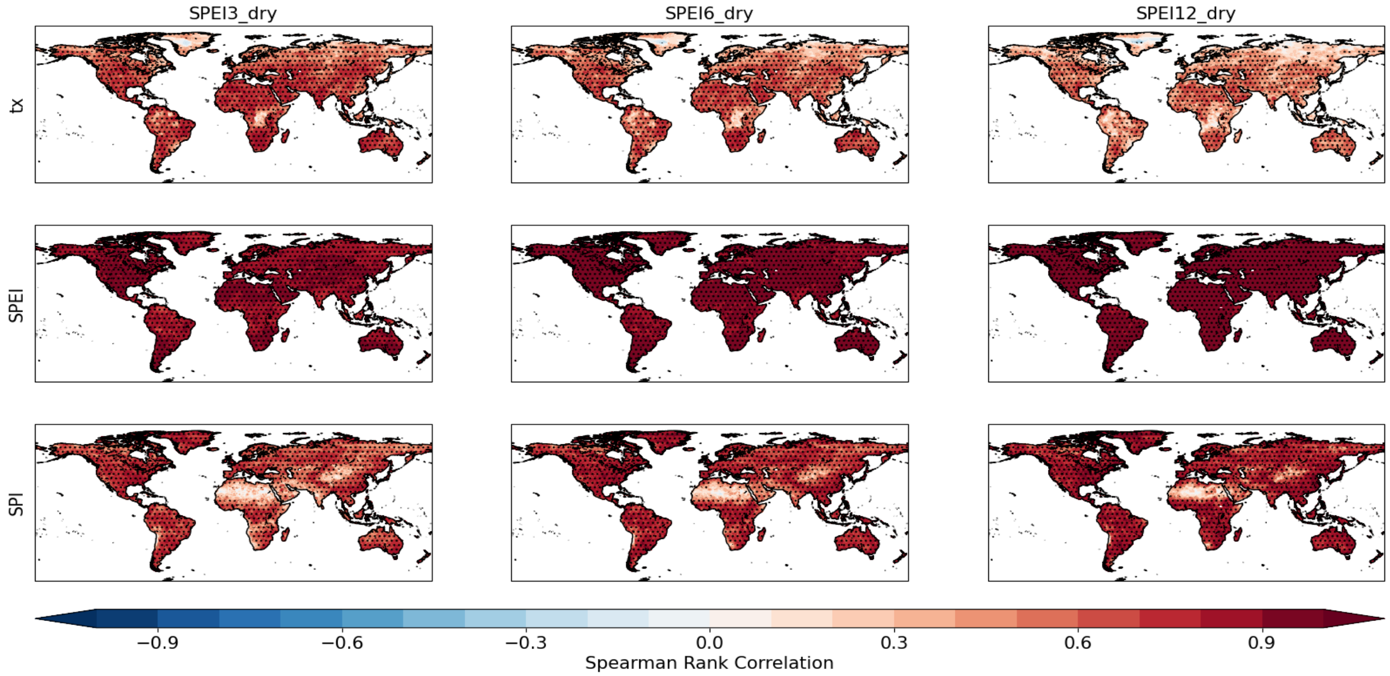


Fig. 3. Sperman Rank Correlation values between yearly SPEI3_dry, SPEI6_dry, SPEI12_dry and univariate extremes days in the period 1961-2016. SPIs and SPEIs (rows 2 and 3) correspond the the ones used to calculate the compound hot/dry extremes. Gridded areas show significant correlation.

SPEI value ≤ -1, following the method proposed by De Luca et al. [9]. To homogenize the monthly SPEI time series to the daily tx90p, we convert the SPEI data into daily by copying the value throughout the specific month. We apply this method with SPEI3, SPEI6 and SPEI12. Thus, we refer to this drought index as SPEI3_dry, SPEI6_dry and SPEI12_dry [9].

B. *Preliminary Results*

Preliminary results in the observations datasets show a general global increasing trend of the number of hot/dry compound events for the period 1961-2016 (Figure 1). Several hot spots are found in South America, Africa and the Indian subcontinent (Figure 2). In addition, we analyse the correlation between the yearly SPEI3_dry, SPEI6_dry, SPEI12_dry and the relative univariate events (*tx*, SPEI and SPI -3, -6 and -12), to gain some information regarding what is the main driver of the trend (Figure 3). We find that the correlation of the SPEI is the highest for all the accumulation months, indicating that the drivers have to be searched both in temperature and precipitation trends. However, we also find that *tx*s shows a decreasing correlation with longer accumulation months, while SPIs show an opposite trend. These results suggest the relative importance of temperature extremes and meteorological precipitation in short and long periods, respectively.

C. *Outlines*

The next steps of the research will focus on better understanding the physical mechanism behind these connections. In addition, we will perform the same analysis in decadal hindcasts from the WCRP Decadal Climate Prediction Project (DCPP) and evaluate them against the observational datasets.

References

[1] J. Zscheischler *et al.*, "Future climate risk from compound events," *Nature Climate Change*, vol. 8, no. 6, pp. 469–477, 2018.

[2] R. H. Grumm, "The central european and russian heat event of july–august 2010," *Bulletin of the American Meteorological Society*, vol. 92, no. 10, pp. 1285–1296, 2011.

[3] K. Emanuel, "Assessing the present and future probability of hurricane harvey's rainfall," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12 681–12 684, 2017.

[4] T. B. McKee *et al.*, "The relationship of drought frequency and duration to time scales," in *Proceedings of the 8th Conference on Applied Climatology*, vol. 17, no. 22. Boston, 1993, pp. 179–183.

[5] S. M. Vicente-Serrano *et al.*, "A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index," *Journal of climate*, vol. 23, no. 7, pp. 1696–1718, 2010.

[6] G. H. Hargreaves and Z. A. Samani, "Reference crop evapotranspiration from temperature," *Applied engineering in agriculture*, vol. 1, no. 2, pp. 96–99, 1985.

[7] S. Contractor *et al.*, "Rainfall estimates on a gridded network (regen)–a global land-based gridded dataset of daily precipitation from 1950 to 2016," *Hydrology and Earth System Sciences*, vol. 24, no. 2, pp. 919–943, 2020.

[8] R. A. Rohde and Z. Hausfather, "The berkeley earth land/ocean temperature record," *Earth System Science Data*, vol. 12, no. 4, pp. 3469–3479, 2020.

[9] P. De Luca and M. Donat, "Global warming increases hot, dry and compound hot-dry extremes over global land regions," *In preparation for GRL.*

Alvise Aranyossy received his BSc degree in Environmental Sciences from Ca' Foscari University, Italy in 2018. He completed his MSc degree in Earth and Climate System Sciences in 2021, studying in University Hohenheim and Hamburg University, Germany. The following year he worked in the Climate Modelling group at Hamburg University. Since November 2022, he has been with the Climate Variability and Change group of the Barcelona Supercomputing Center (BSC) as well as a PhD student at the department of meteorology of the University of Barcelona (UB), Spain.

A Stroke Risk Assessment Tool for Edge-to-Cloud Platforms

Fatemeh Baghdadi[#1], Davide Cirillo [*2]

[#1, *2] *Barcelona Supercomputing Center (BSC), Barcelona, Spain*

[1]fatemeh.baghdadi@bsc.es,

[#1] *Universitat de Barcelona, Barcelona, Spain*

[2]davide.cirillo@bsc.es

***Keywords*—– Stroke, Machine learning, AI SPRINT**

INTRODUCTION

Stroke is the second most common cause of death and a leading cause of adult physical disability in the European Union. The number of people living with stroke is estimated to increase by 27% between 2017 and 2047 in the European Union, mainly because of population ageing and improved survival rates. Stroke survivors can experience a wide range of outcomes that are long-lasting, including problems with mobility, vision, speech and memory; personality changes; fatigue; and depression [1]. The solution to reduce these statistics could be addressed by early detection and diagnosis using remote health care, virtual care, mobile health, or e-health which all essentially lead to the range of solutions that are enabled by wearable devices for continuous and remote monitoring to provide reliable clinical diagnosis by collecting physiological health data over long periods of time.

In this paper we introduce AI-SPRINT "*Artificial intelligence in Secure PRIvacy-preserving computing coNTinuum*" medical use case on personalized healthcare with its focus on stroke risk assessment and prevention. AI-Sprint project has received funding from the European Union Horizon 2020 research and innovation programme under Grant Agreement No. 101016577 and provides programming models, specialised building blocks, and mechanisms for automatic deployment and dynamic reconfiguration to seamlessly design, partition, and run AI applications in computing continuum environments. All this enables flexible and secure AI applications, benefiting developers, integrators, cloud providers, and end-users [2]. Personalized Healthcare use case concerns with the development of AI models for health monitoring using wearable and mobile devices. Specifically, it focuses on the assessment of stroke risk by combining quantitative data (sensor data) and qualitative data (lifestyle information) to create risk stratification models operating in the edge-cloud continuum in real-time.

OBJECTIVE

The AI-SPRINT medical use case conduct a pilot study to collect different type of information and integrating quantitative and qualitative data to design a personalized healthcare risk forecasting model using high-performance data analytics framework in Edge-to-Cloud platforms to manage distribution and parallelism across the resources by ensuring the protection of sensitive data through GDPR compliance and mechanisms to preserve privacy and security, including data anonymization and federated learning.

METHODOLOGY

Life science department of Barcelona Supercomputing Centre (BSC) conducts a pilot study with joint efforts of two subcontracted entities: the stroke awareness foundation Freno al Ictus and the company Smart Solutions Technologies S.L. (henceforth referred to as Nuubo) manufacturer of the adopted wearable device. The adopted wearable device consists of a vest with electrodes and a reusable recorder. This system is designed for ambulatory electrocardiogram (ECG) monitoring up to 30 days. The pilot study planning consisted in defining the strategy to organize and run the campaign for the recruitment of the first group of volunteers involved in the pilot study as well as detailing the dynamics of data collection and sharing. Freno al Ictus organized the recruitment campaign and led to the identification of 10 individuals (5 healthy subjects and 5 stroke survivors, out of >30 individuals planned to be monitored by the end of the project. Nuubo, in collaboration with Freno al Ictus and BSC, organized the training session and during the training session, the volunteers received an overview of the AI-SPRINT project and assets, instructed on the use of the Nuubo device and compiled a questionnaire on cerebrovascular risk factors designed in collaboration with the use case medical advisor Dr. María Alonso de Leciñana (Hospital Universitario La Paz, Madrid, Spain). After the monitoring of each group of volunteers ends and all wearable devices are sent back to the company, the recordings will be anonymized and pre-processed by Nuubo along the anonymized questionnaires which are digitized by Freno al Ictus to be delivered to BSC for data analysis and modelling. Figure 1 illustrates the general overview of pilot study.

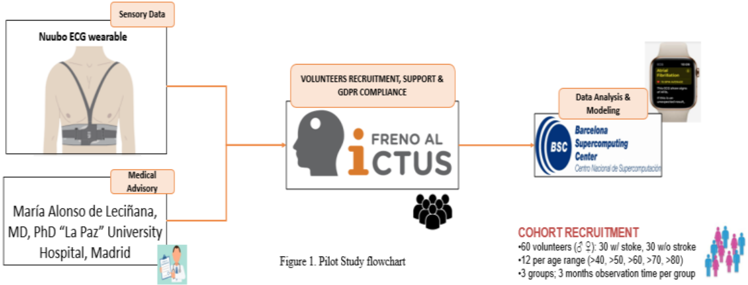

Sensory Data
Nuubo ECG wearable
Medical Advisory
María Alonso de Leciñana, MD, PhD "La Paz" University Hospital, Madrid
VOLUNTEERS RECRUITMENT, SUPPORT & GDPR COMPLIANCE
FRENO AL ICTUS
Data Analysis & Modeling
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
COHORT RECRUITMENT
•60 volunteers (♂ ♀): 30 w/ stoke, 30 w/o stroke
•12 per age range (>40, >50, >60, >70, >80)
•3 groups; 3 months observation time per group

Figure 1. Pilot Study flowchart

RESULTS

In the Demonstration phase of project, we trained several classifiers such as Cascaded support vector machine (CSVM), K-Nearest neighbour (KNN) and random forest using the Dislib library [3] developed by BSC Computer Science Department to classify Electrocardiogram signals (ECG) from the PhysioNet database and the random forest classifier scored 89.9% accuracy. The lack of clinical information about these signals represents an apparent limitation in the use of this dataset for the development of a stroke risk stratification model. Nevertheless, the differential analysis of these signals provided valuable insights into this particular arrhythmia (atrial fibrillation) that is strongly associated to stroke occurrence. In the next stage, Nuubo long-term ECG data and stroke risk factors questionnaires pre-processed and used to train a stroke risk stratification model to be used on non-invasive commercial fitness trackers, such as FitBit and Apple smartwatches.

CONCLUSION

The AI-SPRINT Personalized Healthcare use case for stroke monitoring will allow improving several aspects that are currently limited by the standard practice for stroke monitoring, specifically the immediacy of the responses to healthcare professionals and the engagement of patients in a healthier lifestyle and the use of new technologies. The inclusion of stroke patients and associations in the realization of the project as well as the related communication and dissemination activities increase trust and acceptance of the use of AI and advanced technologies in the healthcare domain by the civil society.

References

[1] Stroke Alliance for Europe,. (2017). The Burden of Stroke Report 2017. Brussels: Stroke Association House.

[2] https://www.ai-sprint-project.eu/

[3] J. Álvarez Cid-Fuentes, S. Solà, P. Álvarez, A. Castro-Ginard, and R. M. Badia, “dislib: Large Scale High Performance Machine Learning in Python,”.

Author biography

Fatemeh Baghdadi was born in Tehran, Iran, in 1991. She received the B.E. degree in Communication and Electronic engineering from UCSI University, Kuala Lumpur, Malaysia, in 2013, and the M.Sc.Eng. degree in Electrical-Digital electronics engineering from Sharif university of Technology, Tehran, Iran, in 2018. In July 2021, she has joined the Life science department of Barcelona supercomputing centre, Spain, where she started as research engineer, and a year later her PhD studies started in 2022. His current research interests include artificial intelligence in biomedical research, biomedical data fusion, Bio signal analysis.

Numerical analysis of MHD flow of magnetically confined plasma in cylindrical geometry using OpenFOAM

Robert Benassai Dalmau* and Shimpei Futatani†
*Universitat de Barcelona, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: rbenasda7@ub.edu*, shimpei.futatani@upc.edu†

***Keywords*—MHD, plasma, OpenFOAM, Magnetic confinement, numerical simulation**

I. Introduction

Climate change is one of the most challenging threats to human life. The need to find clean alternatives to greenhouse-emitting energy production mechanisms is urgent. Nuclear fusion, being sustainable and free from CO_2 emission, is considered as one of the promising future energy resources. In nuclear fusion reactors, plasma is confined using strong magnetic fields. Depending on the system's geometry and other factors, plasma will undergo several types of instabilities, which hamper long-lasting operations. In this work, numerical simulations of magnetohydrodinamical (MHD) flow of magnetically confined plasma in a cylindrical geometry have been performed. Furthermore, a preliminary study of the effects of a time-independent pressure profile in the velocity equation has been performed. The presence of the pressure-gradient term may make the plasma instability more complex.

II. OpenFOAM and MHD equations

Simulating the time evolution of MHD flow and its instabilities is crucial in the study of plasma physics. The dynamical equations for the plasma combine fluid dynamics and Maxwell's electromagnetism laws. This combination yields the visco-resistive MHD equations, which are the Navier-Stokes equation with the Lorentz force (1) and the induction equation (2), which combines Ohm's law, Faraday's equation and Ampère's law. These equations read (using normalized velocity by Alfven's velocity $C_a = \frac{B_0}{\sqrt{\rho\mu_0}}$ and a reference magnetic field B_0.):

$$\frac{\partial \mathbf{B}}{\partial t} = (\mathbf{B}\cdot\nabla)\mathbf{u} - (\mathbf{u}\cdot\nabla)\mathbf{B} + \lambda\nabla^2\mathbf{B} \tag{1}$$

$$\frac{\partial \mathbf{u}}{\partial t} = (\mathbf{B}\cdot\nabla)\mathbf{B} - \frac{1}{2}\nabla\mathbf{B}^2 - (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla^2(\nu\mathbf{u}) - \nabla P, \tag{2}$$

where $\mathbf{B}$ is the magnetic field and $\mathbf{u}$ is the velocity field. Also, ν corresponds to the dimensionless kinematic viscosity and λ is the dimensionless magnetic diffusivity respectively. The non-linear couplings of the magnetic and velocity fields create complex MHD dynamics. This process includes the self-organization of the plasma [1]. In order to solve the MHD equations, the OpenFOAM software [2] has been used. OpenFOAM is a free computational fluid dynamics (CFD) software. The source code is written in *C++* and the software is open source, which means that the existing solvers can be modified to fit the needs of any specific case. In this case, a modified version of the *mhdFoam* solver [3] has been used.

A. Magnetic fields for a cylindrical geometry

The magnetic field used in this work consists of an axial and an azimuthal component. In cylindrical coordinates, the magnetic field is described by $\mathbf{B} = (B_r, B_\theta, B_z) = (0, \frac{B_{wall}}{R}r, B_{axis})$, where R is the radius of the cylinder and B_{wall} and B_{axis} are the values of the azimuthal magnetic field at the cylinder wall and the constant magnetic field in the axial direction respectively. The parametric scan of the magnitude of B_{wall} and B_{axis} has been carried out.

B. Simulation set-up

At the side wall of the cylinder, a no-slip (and non penetration) boundary condition is applied. Periodic boundary conditions are applied at both ends. Initially, an uncorrelated gaussian solenoidal velocity field is imposed, with kinetic energy of the order 10^{-7}. The kinematic viscosity and magnetic resistivity are set to $\nu = \lambda = 0.02$ for all cases. The simulation geometry consists of five sub-domains. Each domain has $(150, 150, 150)$ grid points in the (x,y,z) directions respectively. The time step used in the time-integration is 0.001 (in the dimesionless unit).

III. Results

In the first part of the work, the simulations have been carried out with a constant and homogeneous pressure profile, i.e. $\nabla P = 0$. The simulations have been performed from time $t = 0$ to $t = 10$. The simulation of $B_{wall} = 7.0$ and $B_{axis} = 4.5$ has been performed. As shown in Figure 1, the cross-section of the velocity field obtained by OpenFOAM analysis shows a similar profile to one of the observations of Ref [4]. The simulation reproduced the helical structures of the velocity field as well. Similarities of Figure 1 with the corresponding figure of [4] only allow qualitative comparisons, since the used parameters of kinematic viscosity and magnetic diffusivity are different in this work.

The analysis of parametric scan has been carried out; $B_{axis} = 4.5$ is fixed, and B_{wall} is varied from $B_{wall} = 3.0$ to

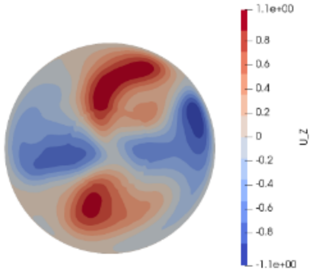


Fig. 1: Cross section of the cylinder for $B_{wall} = 7.0$ and $B_{axis} = 4.5$ with the red and blue areas being the positive and negative isosurfaces of u_z respectively. $\nu = \lambda = 0.02$.

$B_{wall} = 10.0$. Using the fluid parameters mentioned earlier, different cases have been analysed for $B_{wall} < B_{axis}$ and $B_{wall} > B_{axis}$. Figure 2 shows the time evolution of the kinetic energy and the total and magnetic energies of the system. For low values of B_{wall}, the plasma is stable in the time range studied, $t = 10$. On the other hand, when B_{wall} is increased, instabilities occur earlier and kinetic energies grow larger, as shown in Figure 2. In the bottom panel of Figure 2, an exchange between magnetic and kinetic energy is observed when an instability is triggered.

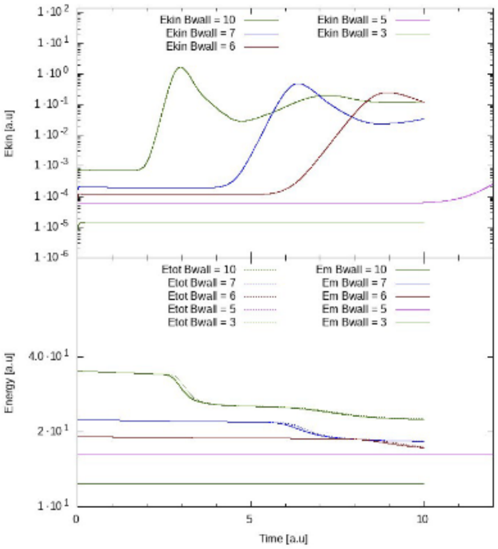


Fig. 2: Top panel: Evolution of kinetic energies with time for $B_{axis} = 4.5$ and $B_{wall} = 3$, 5, 6, 7 and 10. Bottom panel: Evolution of the total and magnetic energies with time for the above-mentioned B_{wall} and B_{axis} values. The vertical axis is logarithmic in both cases.

In the second part of the work, after studying the case with a homogeneous pressure field, a parabolic pressure profile has been applied as the initial condition. Note the pressure profile contributes to the term ∇P despite being time-independent. The inclusion of the term ∇P yields a different observation in the evolution of the dynamics as shown in Figure 3.

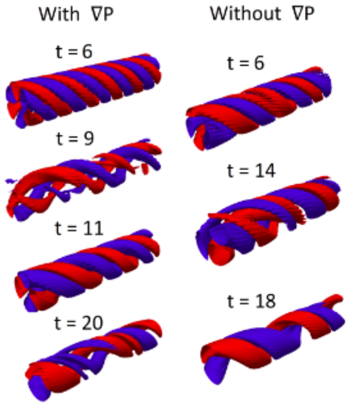


Fig. 3: Isosurfaces for the axial velocity $u_z = \pm 0.2$ where the red and blue colors correspond to the positive and negative directions of u_z respectively. Left: The case with pressure gradient. Right: The case without pressure gradient.

IV. Conclusions

In this work, the MHD dynamics of magnetically confined plasma in cylindrical geometry have been studied using a modified version of *mhdfoam* of OpenFOAM software. The simulation results show the parametric dependence of the plasma parameters on the MHD dynamics, i.e. when the poloidal magnetic field component at boundary (B_{wall}) is increased, an instability is triggered due to the coupling and the nonlinear nature of the MHD equations. Furthermore, a preliminary implementation of the pressure profile has been carried out, i.e. time-independent pressure profile has been implemented to the MHD system in OpenFOAM. During the non-linear regime, a helical structure of the velocity field is observed in both cases of $\nabla P = 0$ and $\nabla P \neq 0$. In the presence of ∇P, the evolution of the non-linear system shows more complexity.

References

[1] S. Futatani, J. A. Morales, and W. J. T. Bos, "Dynamic equilibria and magnetohydrodynamic instabilities in toroidal plasmas with non-uniform transport coefficients," *Physics of Plasmas*, vol. 22, no. 5, p. 052503, May 2015. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2015PhPl...22e2503F

[2] "Openfoam website," https://openfoam.org/.

[3] D. Garrido, D. Suarez, and S. Futatani, "Non-linear mhd simulations of magnetically confined plasma using openfoam," in *7th BSC Severo Ochoa Doctoral Symposium (2020).*

[4] M. Roberts, M. Leroy, J. Morales, W. Bos, and K. Schneider, "Self-organization of helically forced mhd flow in confined cylindrical geometries," *Fluid Dynamics Research*, vol. 46, no. 6, p. 061422, nov 2014. [Online]. Available: https://dx.doi.org/10.1088/0169-5983/46/6/061422

Robert Benassai Dalmau received his BSc degree in Physics from University of Barcelona (UB) in 2022. He is currently completing the MSc in Physics of Complex Systems and Biophysics from University of Barcelona while working at the Open University of Catalunya (UOC) as research assistant in urban mobility.

Multi-dimensional Fourier Series with Quantum Circuits

Berta Casas*, Alba Cervera-Lierta*
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {berta.casas, alba.cervera}@bsc.es

Keywords—Quantum machine learning, Function fitting, Fourier series

I. Extended Abstract

Quantum Machine Learning is the field that aims to integrate Machine Learning with quantum computation. Some works have proposed to implement supervised learning models in quantum computers with applications in problems such as function fitting or classification. Indeed, it has been shown that a single qubit (the quantum equivalent of a bit) can act as a universal approximant. In particular, function fitting models with a particular data encoding strategy output one-dimensional Fourier series. However, models used for multi-dimensional function fitting have not been explored with the same level of detail.

We study quantum strategies for fitting multi-dimensional functions with a general formalism for qudits (quantum information units of dimension d). Using different types of circuit ansatzes (a particular choice of the quantum circuit), we show that the outputs of the models are multi-dimensional Fourier series. We found that the degrees of freedom required for fitting such functions grow faster than the available degrees in the Hilbert space of the circuit. These results exhibit that, for these types of problems, the model does not have enough freedom to fit a general Fourier series. The goal of the work is to study the expressibility of these multivariate quantum models, in other words, the type of functions that can be generated. We contribute to the study of multi-feature quantum machine learning algorithms and conclude that new encoding strategies beyond Fourier series formalism can be more convenient.

A. Introduction

The development of computational paradigms such as quantum computation opens the path to explore the use of quantum devices to perform Machine Learning (ML) tasks, which raises the question of whether Quantum Machine Learning (QML) algorithms can offer an advantage compared to classical ones.

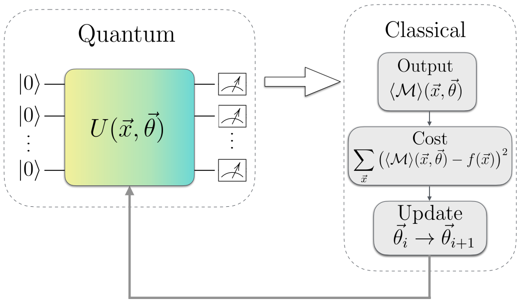

Quantum
$|0\rangle$
$|0\rangle$
$|0\rangle$
$U(\vec{x},\vec{\theta})$
Classical
Output
$\langle\mathcal{M}\rangle(\vec{x},\vec{\theta})$
Cost
$\sum_{\vec{x}}\left(\langle\mathcal{M}\rangle(\vec{x},\vec{\theta})-f(\vec{x})\right)^2$
Update
$\vec{\theta}_i \to \vec{\theta}_{i+1}$

Fig. 1. Representation of a quantum supervised learning model implemented via a Variational Quantum Algorithm (VQA). The quantum part is composed of a quantum circuit with a given unitary U that depends on the data point $\vec{x}$ and the trainable parameters $\vec{\theta}$. The classical part consists of optimizing a cost function(in our case, tailored for the classification of a target function $f(\vec{x})$ that measures the distance with the expectation value of some observable M) and optimize it with respect to the parameters via a classical subroutine.

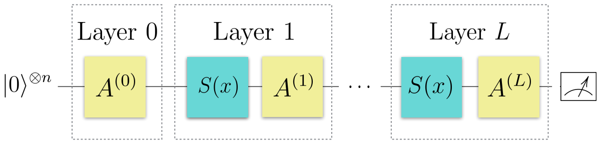


Fig. 2. General structure of a data re-uploading procedure. The 0th layer is used to generate an initial superposition. The other layers contain an encoding gate $S(x)$ and a processing gate $A(\vec{\theta}_l)$, containing parameters which are optimized.

We study a supervised QML algorithm that pertains to the family of Variational Quantum Algorithms [1] (hybrid algorithms that have some parameters in the quantum circuit, which are optimized classically), depicted in Fig. 1. The quantum circuit follows the data re-uploading strategy [2], which consists of defining a structure called layer that encodes and processes the data with some trainable gates $A^{(l)}(\vec{\theta}_l)$. Data is repeatedly introduced into the quantum circuit (see Fig. 2). With this re-uploading strategy and with an encoding gate $S(x) = e^{ixH}$, where H is an arbitrary encoding Hamiltonian and x the data encoded, the output of the model (the expectation value of an observable $\mathcal{M}$) is a one-dimensional Fourier series [3]. The frequencies of the Fourier series are related to the eigenvalues of the encoding Hamiltonian H and the coefficients are determined by the trainable gates $A^{(l)}(\vec{\theta}_l)$.

B. Multi-dimensional quantum models

In this section, we extend the one-dimensional quantum model to multi-dimensional, besides studying its expressivity. We propose two quantum circuit models, named the Line and Parallel ansatzes. They are depicted in Fig. 3. Both models generate multi-dimensional Fourier series as outputs:

$$\langle\mathcal{M}(\vec{x})\rangle = \sum_{\omega_1,\omega_2,\ldots,\omega_M=-D}^{D} c_{\vec{\omega}} e^{i\vec{x}\cdot\vec{\omega}} \tag{1}$$

We study in which cases the models can generate general Fourier series. For this, we compare the degrees of freedom

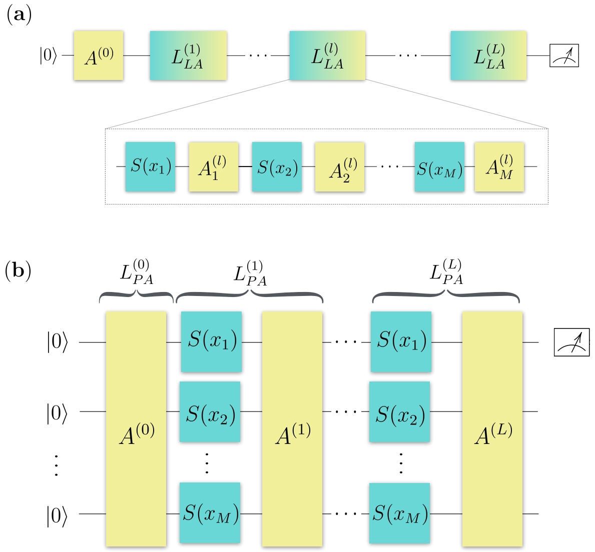


Fig. 3. Quantum circuit ansatzes of the models. The Line ansatz (**a**) encodes each data feature in a single qudit. Thus the circuit depth grows linearly with the total number of features M. The Parallel ansatz (**b**) encodes the M features in M qudits instead.

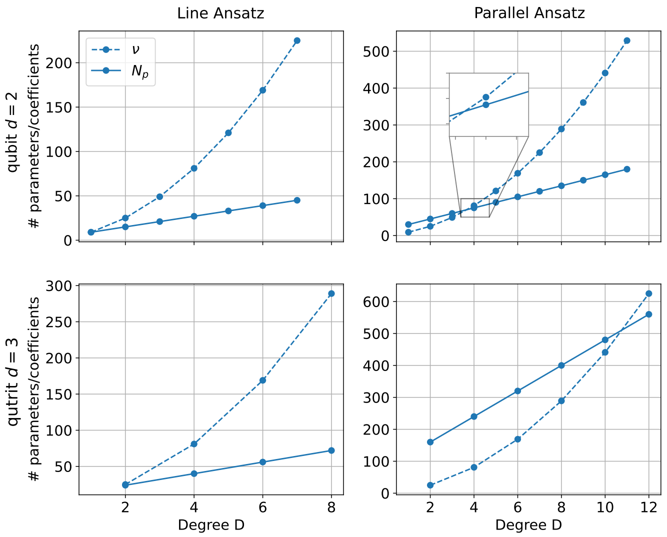


Fig. 4. Comparison of the degrees of freedom condition for the Line (first column) and Parallel (second column) ansatzes using qubits (first row) and qutrits (second row). The dashed line indicates the degrees of freedom ν of the Fourier series generated by the model. The solid line shows the number of trainable parameters N_p in the circuit. The Parallel Ansatz fulfils the degrees of freedom condition for higher-degree Fourier series compared to the Line Ansatz, being the gap wider when using qutrits.

(d.o.f.) ν of the output Fourier series of the models with the number of independent parameters N_p on the quantum circuit. The d.o.f. are the real and imaginary part of all the independent coefficients $c_{\vec{\omega}}$. We study if the models have enough d.o.f. to fit a general Fourier series, meaning $\nu \geq N_p$, which we name the d.o.f. condition.

C. Results

For a one-dimensional function the d.o.f. condition is always accomplished since $N_p > \nu$ for this case. However, for data of higher dimension the d.o.f. grow faster than the number of parameters in the circuit. We show an example of this for two-dimensional data in Fig. 4 for the Line and Parallel ansatzes. We represent the d.o.f. condition in front of the Fourier series degree $D = \max(\vec{\omega})$, which depend on the number of layers L and the dimension of the qudit d. We observe that the Line ansatz does not accomplish the d.o.f. condition and therefore, it does not have enough free parameters to fit a general Fourier series. On the other hand, the Parallel ansatz achieves the condition until a certain degree D, but the d.o.f. ν grow faster than the parameters available in the circuit. Besides this, by using higher information units (qutrits in this case, the quantum equivalent to a trit), we achieve to accomplish the condition up to higher-degree Fourier series.

D. Conclusion

Besides extending a well-known QML formalism to multi-dimensional datasets, this work aims to establish a trade-off between the number of qudits, circuit depth (measured with the number of layers of the circuit), data dimension, and local qudit dimension. The minimum number of parameters needed to characterize the multi-variable Fourier series grows exponentially with the data dimensions, resulting in poor scaling. Other methods, such as using non-commuting gates to embed the data features, should be explored, although some may depart from the Fourier series formalism. Besides this, with the proposed multi-qudit models, we can approximate functions of a higher degree, which can be used as a suitable approximation for some problems.

II. Acknowledgment

This work has been financially supported by the Ministry of Economic Affairs and Digital Transformation of the Spanish Government through the QUANTUM ENIA project call – Quantum Spain project, and by the European Union through the Recovery, Transformation and Resilience Plan – NextGenerationEU within the framework of the Digital Spain 2026 Agenda.

References

[1] M. Cerezo and *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, Aug. 2021. [Online]. Available: https://doi.org/10.1038/s42254-021-00348-9

[2] A. Pérez-Salinas and *et al.*, "Data re-uploading for a universal quantum classifier," *Quantum*, vol. 4, p. 226, Feb. 2020. [Online]. Available: https://doi.org/10.22331/q-2020-02-06-226

[3] M. Schuld and *et al*, "Effect of data encoding on the expressive power of variational quantum-machine-learning models," *Physical Review A*, vol. 103, no. 3, Mar. 2021. [Online]. Available: https://doi.org/10.1103/physreva.103.032430

Berta Casas received hes BSc degree in Physics from Universitat de Barcelona (UB), in 2021. The following year, she completed her MSc degree in Quantum Science and Technologies from Universitat de Barcelona (UB), Universitat Autonoma de Barcelona (UAB) and Universitat Politecnica de Catalunya (UPC) in Barcelona. Since 2022, he has been with the Quantic group of Barcelona Supercomputing Center (BSC) as well as a PhD student at Universitat de Barcelona (UB), Spain.

Multi-annual predictions of daily temperature and precipitation extremes

Carlos Delgado-Torres*, Markus G. Donat*†, and Albert Soret*
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
Corresponding author: Carlos Delgado-Torres (carlos.delgado@bsc.es)

***Keywords*—Decadal climate predictions, Extreme events, Forecast quality assessment, CMIP6 multi-model, Impact of initialization**

I. Introduction

Characteristics of climate extremes are changing in a warming climate, with in particular hot temperature and heavy precipitation extremes becoming more intense and frequent, thus increasing their potential impact on nature, economy and society. Besides, internal climate variability also modulates the occurrence of extreme events. Trustworthy predictions are essential to develop strategic planning to adapt, build more resilience to their risk and anticipate their impacts ahead of time. Predictions of extremes may be more relevant to users than predictions of average variables, as extremes typically break the resilience of a system and cause the heaviest impacts on society and environment.

Predictions of variations in the frequency and intensity of extremes in the forthcoming years can potentially be provided by decadal climate predictions. In addition to long-term changes due to external forcings (natural and anthropogenic), decadal predictions aim to also capture the internal variability of the climate system (slow, natural oscillations). For this reason, climate models are initialized with observation-based products. Decadal predictions have been shown to skillfully predict essential climate variables such as near-surface air temperature and, to a lesser extent, precipitation in many regions of the world. However, the predictability of mean quantities and extreme events might differ .

We perform a forecast quality assessment of multi-model decadal predictions of annual and seasonal extreme temperature and precipitation indices with all available hindcasts contributed to the Decadal Climate Prediction Project Component A (DCPP-A) of the Coupled Model Intercomparison Project Phase 6 (CMIP6). The evaluation is performed globally for predictions of the next five years. The skill for extreme predictions is compared to that for mean temperature and precipitation variations, and the impact of model initialization is assessed by comparing the skill of decadal predictions and historical forcing simulations.

II. Data and Methods

Daily minimum and maximum near-surface air temperature and precipitation have been used to compute the extreme indices. Besides, monthly means of near-surface air temperature (TAS) and precipitation (PR) have been used to compare the forecast quality for extreme indices and mean variables.

All the available CMIP6 decadal hindcasts (DCPP; 133 members) have been used. Besides, the CMIP6 historical simulations (HIST; 134 members) performed with the same forecast systems as DCPP have been used to estimate the impact of the model initialization on the forecast quality. The Berkeley Earth Surface Temperatures (BEST) and Rainfall Estimates on a Gridded Network (REGEN) datasets have been used as observational references for the extreme indices based on daily minimum and maximum temperatures and daily precipitation, respectively. The Global Historical Climatology Network v4 (GHCNv4) and Global Precipitation Climatology Centre (GPCC) datasets have been used as references for TAS and PR, respectively.

The Expert Team on Climate Change Detection and Indices (ETCCDI) defined a set of extreme climate indices to detect, characterize and monitor changes in the frequency and severity of extreme events, such as heat waves, cold spells, floods and droughts. We have selected six extreme indices: TN10p (percentage of days when minimum temperature is below the 10th daily percentile), TNn (minimum of daily minimum temperature), TX90p (percentage of days when maximum temperature is above the 90th daily percentile), TXx (maximum of daily maximum temperature), R95p (sum of precipitation in days where daily precipitation exceeds the 95th percentile of daily precipitation) and Rx5day (maximum 5-day consecutive precipitation). Thus, for each variable, we evaluate a measure of relatively moderate extremes, which occur on average several times per year, and a measure representing the most intense event of the year.

The Spearman Anomaly Correlation Coefficient (ACC), which estimates the linear relationship between the observed and predicted time series, has been used to evaluate the deterministic forecasts. The ACC ranges between -1 (worst forecast) and 1 (perfect forecast). The residual correlation has been used to assess whether DCPP captures more observed variability than the already captured by HIST, and also ranges between -1 and 1. Positive values indicate that DCPP captures more observed variability than HIST, meaning the opposite otherwise. The statistical significance of the ACC and residual correlation has been estimated with a one-sided and two-sided t-test, respectively, accounting for the time series autocorrelation and controlling for multiple testing by applying the False Detection Rate (FDR) procedure using $\alpha_{FDR} = 0.1$.

III. Results

We have evaluated the quality of the CMIP6 decadal forecast systems in predicting TAS, PR, and a set of extreme indices based on daily minimum and maximum temperatures and precipitation for predictions of the next five years. The forecast quality has been estimated and compared to that of the historical simulations to assess the impact of model initialization.

The skill of the DCPP multi-model (Fig. 1) is high in predicting mean and extreme temperature indices computed at annual frequency over most of the globe. The skill is lower and limited to some regions for mean and extreme precipitation. There is a generally higher skill in predicting the mean variables than the extreme indices. The skill for both extreme temperature and precipitation is higher for the moderate extremes (TN10p, TX90p and R95p; related to frequency) than for the most extreme extremes (TNn, TXx and Rx5day, related to intensity).

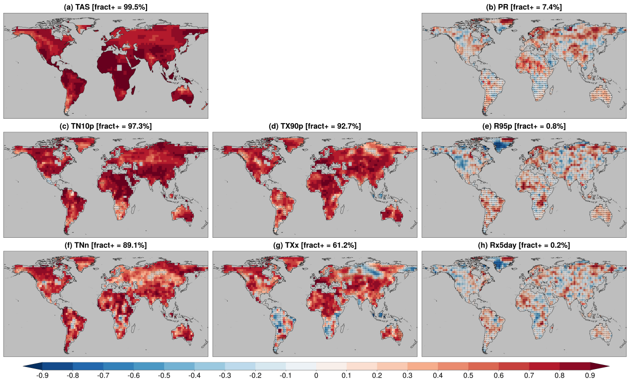


Fig. 1. Multi-annual forecast quality: ACC obtained with the DCPP multi-model ensemble for the forecast years 1-5 for the mean variables and extreme indices. The percentage of the global area with statistically significant positive or negative values is shown in the titles. Crosses indicate that the values are not statistically significant.

The comparison between DCPP and HIST (Fig. 2) shows a region-dependent impact of initialization on the skill. The added value due to initialization is higher for the mean variables than for the extreme indices. Besides, such skill differences differ between indices, especially those representing extreme temperature.

In conclusion, we find that the CMIP6 decadal forecast systems can skillfully predict characteristics of climate extremes, in particular extremely hot and cold temperatures. While the prediction skill for the extremes indices is mostly lower than for annual or seasonal means, these forecast systems still provide useful predictions for the more impact-relevant aspects of climate. However, to exploit all the potential usefulness of decadal predictions, user-oriented indicators could be explored to facilitate their applicability in climate-sensitive sectors, which might be based on variables other than temperature and precipitation, such as wind speed and solar radiation. Besides, the analogous forecast quality assessment should be performed for the particular region, index and forecast period for each user-specific need. This systematic evaluation of decadal hindcasts is essential when providing a climate service based on decadal predictions so that the user is informed on the trustworthiness of the forecasts. Also, comparing decadal hindcasts and historical simulations might help climate services providers to select the highest-quality climate information for each particular case.

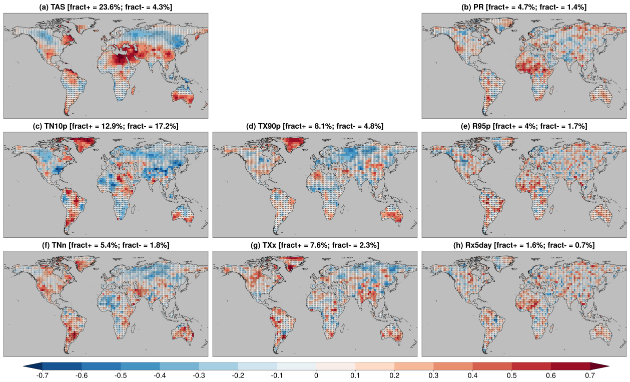

(a) TAS [fract+ = 23.6%; fract- = 4.3%]
(b) PR [fract+ = 4.7%; fract- = 1.4%]
(c) TN10p [fract+ = 12.9%; fract- = 17.2%]
(d) TX90p [fract+ = 8.1%; fract- = 4.8%]
(e) R95p [fract+ = 4%; fract- = 1.7%]
(f) TNn [fract+ = 5.4%; fract- = 1.8%]
(g) TXx [fract+ = 7.6%; fract- = 2.3%]
(h) Rx5day [fract+ = 1.6%; fract- = 0.7%]
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Fig. 2. Impact of model initialization: Residual correlation obtained with the DCPP multi-model ensemble with respect to the HIST multi-model ensemble for the forecast years 1-5 for the mean variables and extreme indices. The percentage of the global area with statistically significant positive or negative values is shown in the titles. Crosses indicate that the values are not statistically significant.

Publication

Delgado-Torres, C., Donat, M. G., Soret, A., Gonzalez-Reviriego, N., Bretonnière, P., Ho, A., Pérez-Zanón, N., Cabré, M. S., and Doblas-Reyes, F. J. (2023). Multi-annual predictions of the frequency and intensity of daily temperature and precipitation extremes. Environmental Research Letters. https://doi.org/10.1088/1748-9326/acbbe1

Acknowledgment

CDT acknowledges financial support from the Spanish Ministry for Science and Innovation (FPI PRE2019-509 08864 financed by MCIN/AEI/10.13039/501100011033 and by FSE invierte en tu futuro). We acknowledge the use of the climdex.pcic, startR, s2dv, CSTools, multiApply, and ClimProjDiags R-language-based software packages, all of them available on the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/). The authors thank all the modeling groups for making the simulations available through the Earth System Grid Federation system (ESGF; https://esgf-node.llnl.gov/search/cmip6/).

Carlos Delgado-Torres holds a BSc in Physics and an MSc in Meteorology and Geophysics from the Complutense University of Madrid. During his years at the university, he did internships at the Agencia Estatal de Meteorología (AEMET), Instituto Nacional de Técnica Aeroespacial (INTA) and eltiempo.es. He joined the BSC in June 2019 to develop his Master's thesis and, after finishing the MSc, he worked as a software developer for some months in the private sector before returning to the BSC. Carlos is developing his PhD thesis on decadal climate prediction and predictability for climate services at the Earth Sciences Department. During his PhD, he visited the International Research Institute for Climate and Society (IRI) of Columbia University for his PhD stay.

Towards Data Driven Reduced Order Models for the Automotive Industry

Benet Eiximeno*†, Ivette Rodríguez†, Oriol Lehmkuhl*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: benet.eiximeno@bsc.es, ivette.rodriguez@upc.edu, oriol.lehmkuhl@bsc.es

***Keywords*—Computational fluid dynamics, reduced order models, large eddy simulations.**

I. Extended Abstract

The aerodynamic performance of road vehicles is highly influenced by the wind incidence angle. Perturbations on the incident velocity, break the symmetry of the flow field as the hairpin vortex in the leeward side of the recirculation gains intensity, increases the suction, augments the drag force and with it the fuel consumption. This effect is even stronger in square-back cars, as they have an abrupt end that evokes a forced separation and recirculation area. The goal of the present work is to develop a reduced order model to predict the variations in the back pressure, and thus, the drag using the configuration without wheels of the Windsor body.

A. Numerical methodology

Wall-modelled large eddy simulations of the yawed Windsor body at $\delta = 2.5°$ have been used to generate the dataset for the reduced order model analysis. In order to have a consistent model with the experiments made by Varney et. al. [1], the Reynolds number has been set to $Re = UL_{ref}/\nu = 2.9 \times 10^6$, where L_{ref} is the length of the model. The domain is equivalent to the wind tunnel used in the experiments and at the bottom wall a non-slip boundary condition is given. The computational grid has a total amount of 41.42 million grid points with a maximum y^+ value in the car of $y^+ = 62.5$ and a maximum ratio between the mesh size and the Kolmogorov scales in the wake of $h/\eta = 119.9$. The Vreman model [2] is used to treat the turbulence and the near wall region is modelled with the Reichardt wall law with an exchange location method in the fourth node [3]. This results are satisfactory when compared with the experimental measurements by Varney et al. [1] (e.g. $C_D = 0.3400$ vs $C_D = 0.3298$ (exp)).

Proper orthogonal decomposition (POD) [4] is applied to the simulation data as a feature recognition technique and to find patterns in the dataset linked with the yaw angle. The database is formed by 569 snapshots which extend over a period of $t = 21.2$TU.

B. Results

Figure 1 shows an instantaneous snapshot of the back recirculation at a horizontal plane located at $z/L = 0.186$ (left) together the same state reconstructed using only the first 10 POD modes of the velocity field (right). Thanks to this technique, it is possible to filter the smaller scales and shed light to the most important structures in the flow field. It is clearly seen that the recirculation is dominated by the hairpin vortices (marked with a 1 and a 2 in the right picture). Eventhough the yaw angle of the flow is considerably small, the hairpin vortex from the leeward side (1) is bigger than the hairpin vortex in the windward side (2), which increases the suction in the leeward side of the back face.

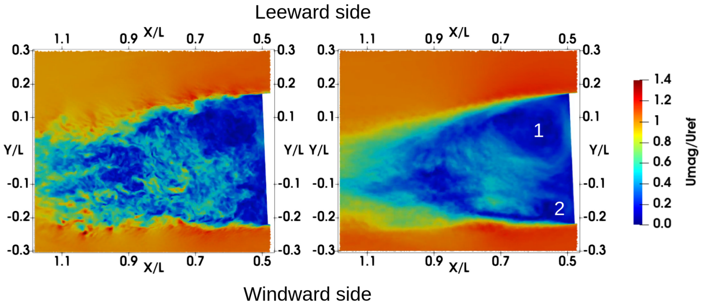


Fig. 1: Velocity magnitude at an horizontal plane located at $z/L = 0.186$. Left: all scales, right: reconstructed with the fist 10 POD modes

Figure 2 shows the two most dominant POD modes of pressure coefficient on the surface of the car. These modes contain a 8.37% of the total energy and their evoloution is related with the changes in intensity of the leeward hairpin vortex changes due to the variations in the yaw angle. Therefore, one may wonder if it would be possible to reconstruct the mean back-pressure at any yaw angle giving an arbitrary contribution to these distributions. Here we propose the following model:

$$P_{mean}[\delta] = P_{mean}[2.5°] + S_1[\delta]M_1 + S_2[\delta]M_2 \qquad (1)$$

where M_1 and M_2 are the spatial distributions coming from the POD of the car rotated 2.5° in the yaw axis. S_1 and S_2 are the contributions of each mode in the studied yaw angle. The contribution of S_1 and S_2 is null if the yaw angle is 2.5°. To compute them at the rest of yaw angles, first their values at $\delta = 5°$ and $\delta = 10°$ are obtained by minimizing the mean error of the back-pressure in the location of the experimental probes used by Varney et. al. [1]. Figure 3 illustrates, for the case of $\delta = 5°$, the mean error function and the relative error of the back-pressure in the experimental probes. The mean error in the back pressure obtained with the optimal S_1 and S_2 values is of $\varepsilon = 5.34\%$ for $\delta = 5°$ and of $\varepsilon = 14.64\%$ for

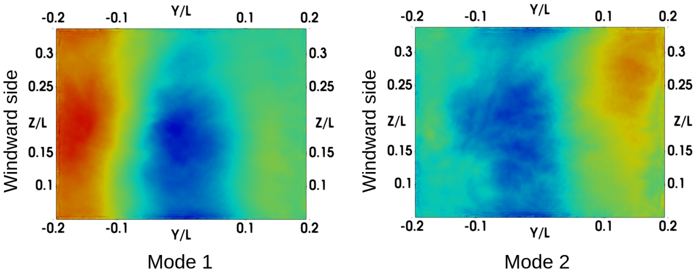


Fig. 2: Two most dominant surface pressure coefficient modes

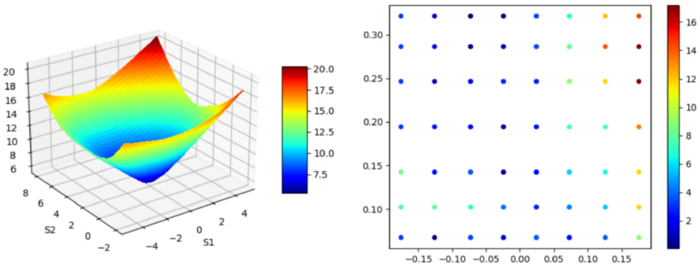


Fig. 3: Mean error function (left) and error in the probes from Varney et. al. [1] (right) for $\delta = 5°$

$\delta = 10°$. The results obtained with the reduced order model are rather good for $\delta = 5°$ (see Figure 4) although the error in the top leeward corner increases with the yaw angle. As the coefficients are null at $\delta = 2.5°$, it is possible to do a linear regression to find an expression for S_1 and S_2 at any yaw angle between $\delta = 2.5°$ and $\delta = 10°$.

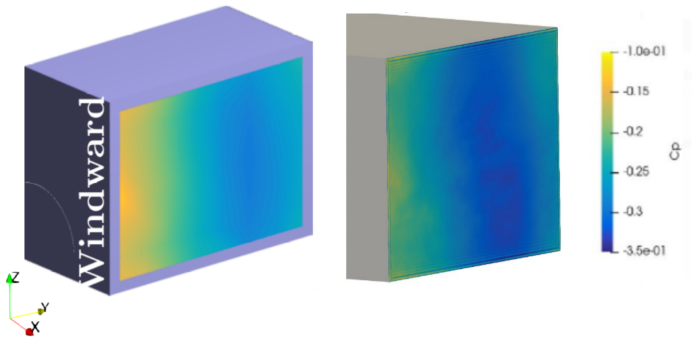


Fig. 4: Experimental back-pressure [1] (left) and model prediction (right) at $\delta = 5°$

The future work will be focused on performing new wall modelled LES simulations to add information at other yaw angles. It is expected then to capture better the changes of the physics in the hairpin vortices and also to find a general expression for any yaw angle without the need of relying on experimental data. Moreover, it will be discussed if adding lower rank modes can improve the accuracy of the model. Finally, the possibility using the dynamic mode decomposition (DMD) [5] to find a relationship between the yaw angles will also be discussed.

II. Acknowledgements

This work has been partially financially supported by the Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain (ref. PID2020-116937RB-C21 and PID2020-116937RB-C22).

This work has been submitted to the 14th European Turbulence Modelling and Measurements 2023 and the second Spanish Fluid Mechanics Conference (SFMC).

References

[1] Max Varney, Giancarlo Pavia, Martin Passmore, and Conor Crickmore. Windsor model experimental aerodynamic dataset, 2021.

[2] AW Vreman. An eddy-viscosity subgrid-scale model for turbulent shear flow: Algebraic theory and applications. *Physics of fluids*, 16(10):3670–3681, 2004.

[3] O Lehmkuhl, GI Park, ST Bose, and P Moin. Large-eddy simulation of practical aeronautical flows at stall conditions. *Proceedings of the 2018 Summer Program, Center for Turbulence Research, Stanford University*, pages 87–96, 2018.

[4] J. L. Lumley. Rational Approach to Relations between Motions of Differing Scales in Turbulent Flows. *The Physics of Fluids*, 10(7):1405, December 1981. Publisher: American Institute of PhysicsAIP.

[5] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

Benet Eiximeno received his BSc degree in Aerospace engineering from Universitat Politècnica de Catalunya, Spain in 2020. He then studied a master degree in Aeronautic engineering in the same university. During the Bachelor degree he did an internship in the TUrbulence and Aerodynamics REsearch Group (TUAREG) from UPC and he took part in the Summer of High Performance Computing program during the Summer of 2021 (between the first and second year of the MSc degree). In March 2022 he joined the Large-scale computational fluid dynamics group in BSC to do the master thesis on reduced order models applied to computational fluid dynamics. Finally, in October he enroled to the computational and applied physics PhD program of UPC.

Ranking Allosteric Activators of AMPK by Absolute Binding Free Energy Calculations with Fun-metaD

Rhys Evans#1, F. Javier Luque#2, Carolina Estarellas#3

#*Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus de l'Alimentació de Torribera,*
Av. Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain

[1]revans@ub.edu, [2]fjluque@ub.edu, [3]cestarellas@ub.edu

***Keywords*— computational biophysics, metadynamics, binding free energy, molecular dynamics, AMPK, CVDs**

Extended ABSTRACT

A. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year, and consequently, the global burden of CVD is a major public health crisis worldwide and the development of more effective treatments for CVD is an urgent challenge. Due to the great number of pathways it is involved in, and its tissue-specificity expression, AMPK has emerged in recent years as a crucial therapeutic target for treating CVDs. AMP-activated protein kinase (AMPK) is a Ser/Thr kinase known as a key regulator of cellular energy homeostasis. It functions as a fuel gauge of the cell by controlling the AMP/ADP/ATP ratio, acting via the switching of cellular metabolism from anabolic to catabolic pathways through the phosphorylation of approximately 30 targets. Hence, AMPK has been identified as a key therapeutic target to tackle metabolic disorders such as type 2 diabetes, and obesity and has more recently been involved in several cancer processes.

From a structural point of view, it comprises 12 different AMPK complexes expressed and differentially distributed along the body's tissues. This fact is essential to develop direct activators, located at the interface of the α- and β-subunits called ADaM site, that exhibit isoform selectivity. Following on from the first activator identified, A-769662[1], other alternatives have been proposed including PF-739[2], SC4[3], MT47-100[4], and MK-8722. Unfortunately, all efforts to develop a selective modulator towards β2-containing AMPK complexes have been unsuccessful. To understand the molecular factors that govern the structure-function-dynamics of these complexes, we have calculated the respective absolute binding free energy (ABFE) of, and to corroborate the isoform selectively presented by, these activators, in complex with α2β1 and α2β2 AMPK systems.

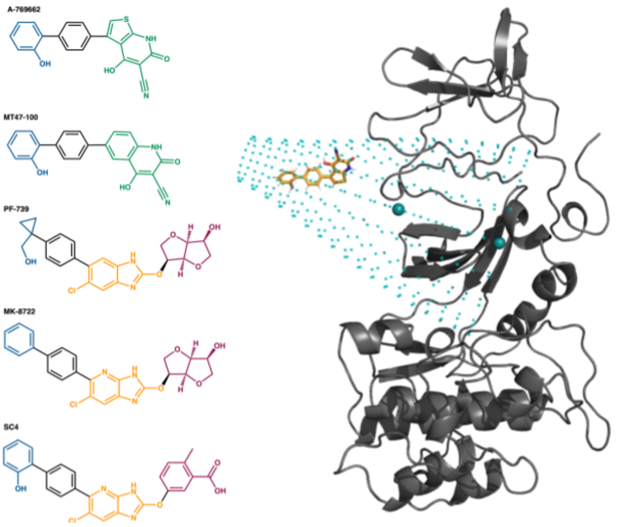


Fig. 1 LEFT: the structures of the five activators under investigation in this project. RIGHT: An example simulation system of AMPK α2β2 (shown here in grey cartoon) in complex with A-769 (shown with orange sticks), in the unbound state. The funnel-shaped restraints (cyan) are shown, defined from the axis created by the anchor points (spheres).

However, experimental data on the binding activity of these activators is limited and inconsistent, leading to difficulties in understanding the development of better drugs[2].

This project presents a computational ranking of absolute binding free energies using funnel-shaped restraint metadynamics, a method demonstrated to make such calculations accurately and efficiently in complex biological targets. We provide, for the first time, a binding affinity ranking in order to unify our understanding of, and design, more effective allosteric activators for, this crucial target.

B. Methodology

To measure the ABFE, we ran funnel-shaped restraint metadynamics (fun-metaD)[5]. This single replica CV-based metadynamics technique has previously proven to be effective at calculating ABFEs and uses a simple set of distance-based CVs. We ran a total of 24 500 ns simulations, comprising three replicas of each complex α2β1 and α2β2 with A-769662, PF-739, SC4, MT47-100, and MK-8722. These simulations were run on MareNostrumIV with GROMACS (v. 2021.4) using the Plumed plugin (v. 2.8.0). The analysis performed focused on establishing the efficacy of the technique to produce a proper ranking of the binding affinities. This was calculated from the reconstructed free-energy surfaces and was adjusted for the volume of the funnel that restricts the free movement of the ligand (1.63 kcal·mol^{-1}). Furthermore, the number of re-crossing events obtained during the simulation will also be considered, as this is a way to check the simulation time needed for the convergence of the systems[5]. One re-crossing is defined as the event that the ligand leaves from the pocket to the bulk water, and then returns to the original binding site, and for our simulations was measured using the values of the collective variables that establish the funnel-shaped restraints.

TABLE I
Average ABFE Results for AMPK Complexes

Activator	**Reported Literature Activity***		**ABFE / kcal·mol^{-1}**	
	α2β1	**α2β2**	**α2β1**	**α2β2**
A-769662 [1]	AF 14.3	AF n/a	-5.6 ± 2.2	-4.4 ± 1.0
PF-739 [2]	AF 5.2	AF 2.9	-7.5 ± 2.1	-7.8 ± 2.3
SC4 [3]	AF 2.3	AF 0.5	-4.8 ± 0.2	-8.4 ± 5.2
MT47-100 [4]	EC 5.23	EC 42.4	-8.7 ± 3.1	-9.2 ± 8.5
MK-8722	EC 2.4	EC 50	-7.0 ± 1.6	-4.4 ± 3.7

* AF = Activation Fold (no units), EC = EC_{50} (nM)

C. Results

Our results are summarised in Table 1, and show the ABFE values that were successfully calculated after 500 ns of simulation. These values are the mean values over all the replicas, with the standard deviations provided. In alignment with our findings when developing the protocol for fun-metaD, a minimum number of re-crossings was set for a simulation to qualify as converged. Therefore, any simulation that exhibited fewer than 2 recrossing within the 500 ns (of which there were 3) was considered to be insufficiently converged and excluded from the averages in Table 1.

D. Conclusion and Future Enhancement

In the future, such converged binding affinity results can be used to determine a distinctive pharmacophore between β1- and β2-containing AMPK complexes and will be taken into consideration in future drug design steps. By identifying the chemical groups that increase the selectivity to the α2β2 isoform, we will propose new candidates, which will be validated by metadynamics and alchemical means. The enzymatic activity of these novel compounds will also be experimentally assessed, through the host's collaboration with Prof. J. Oakhill (University of Melbourne).

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme; in particular, the author gratefully acknowledges the support of Dr Carolina Estarellas of the Department of Nutrition, Food Sciences and Gastronomy, University of Barcelona and the computer resources and technical support provided by Barcelona Supercomputing Center.

References

[1] B. Xiao et al., "Structural basis of AMPK regulation by small molecule activators," Nature Communications, vol. 4, no. 1, 2013.

[2] E. C. Cokorinos et al., "Activation of skeletal muscle AMPK promotes glucose disposal and glucose lowering in non-human primates and mice," Cell Metabolism, vol. 25, no. 5, 2017.

[3] K. R. W. Ngoei et al., "Structural determinants for small-molecule activation of skeletal muscle AMPK α2β2γ1 by the glucose Importagog SC4," Cell Chemical Biology, vol. 25, no. 6, 2018.

[4] J. W. Scott et al., "Inhibition of AMP-activated protein kinase at the allosteric drug-binding site promotes islet insulin release," Chemistry &amp; Biology, vol. 22, no. 6, pp. 705–711, 2015.

[5] R. Evans et al., "Combining machine learning and enhanced sampling techniques for efficient and accurate calculation of absolute binding free energies," Journal of Chemical Theory and Computation, vol. 16, no. 7, pp. 4641–4654, 2020.

Author biography

Rhys A. Evans studied an integrated Master's degree in Natural Sciences at University College London (UCL, 2013-2017). He transitioned to the field of Biophysics under the supervision of Prof. F. L. Gervasio, joining the group as a PhD student in September 2017. His thesis focussed on exploring the use of computational techniques and simulation-based methods to aid in the pharmaceutical development process. With Dr Carolina Estarellas he worked on the usage of enhanced sampling techniques and their application to various stages of the drug discovery pipeline.

As his PhD came to a close in 2022, he travelled to the University of Barcelona through the HPC-Europa3 Transnational Access Fellowship, to work with Dr Estarellas within the group of Prof. F. J. Luque. There he works as a post-doctoral researcher, applying the methods developed during his PhD to the study of drug discovery towards therapies for metabolic diseases.

Prediction of bacterial interactomes based on genome-wide coevolutionary networks: an updated implementation of the ContextMirror approach.

Fernández, Miguel[1,2], Pontes, Camila[1], Ruiz-Serra Victoria[1]

[1]Barcelona Supercomputing Center (BSC), Life Science, 08034, Barcelona, Spain.

[2]miguel.fernandez@bsc.es

Keywords**—** Coevolution, interactome, protein-protein interaction

Extended ABSTRACT

Interacting proteins undergo reciprocal evolutionary change to retain biological functions. Coevolution, quantified as the degree of similarity between the phylogenetic trees of protein families, has been shown to be a reliable indicator of protein-protein interactions [1]. However, previous methods treat every protein pair as an independent interaction partner and do not consider the influence of other proteins. The ContextMirror approach addresses this limitation by using the coevolutionary network of an organism (network of similarities between all pairs of proteins in a genome) to evaluate coadaptation by integrating the influence of every interactor on a given protein pair [2]. In this study, we present an updated version of the ContextMirror approach that is optimized in terms of time and resource utilisation by integreating the most optimal alignment (MAFFT) and tree-building softwares (FastTree) for large-scale analysis, tackling one of the main imitations of the original ContextMirror approach [3]. Due to the lack of original results to compare against, we verified the phylogenetic inference of the updated ContextMirror approach using the MirrorTree Webserver [4], an online tool that implements the widely used MirrorTree method for quantifying the correlation between protein pairs. A reduced set of 15 proteins, comprising two protein complexes and 5 additional non-interacting proteins as a control was inputed to the updated implementation of the ContextMirror approach and the MirrorTree Webserver (Figure 1).

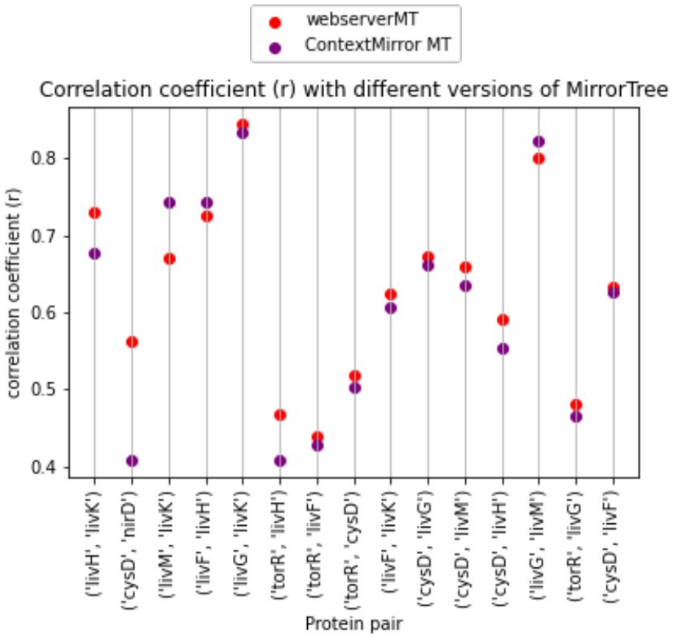


Fig. 1 Comparison between the results of phylogenetic inference on 15 pairs of proteins with the MirrorTree Webserver (red) and the updated implementation of the ContextMirror approach (purple).

Our results differ 0.032 on average from the Webserver's results, taken as ground truth, which indicates that the updated ContextMirror approach does indeed capture the evolutionary resemblance between different protein families, interpreted as an indicator of protein-protein interaction.

Further steps of the ContextMirror approach consider the whole evolutionary context to improve the predictions made by the MirrorTree method, successfully implemented as the first step of the ContextMirror updated approach.

We tested the complete updated implementation of the ContextMirror approach with the same reduced dataset that was used to validate the phylogenetic inference. The results vary considerably depending on the set of parameters used to make the prediction, but they suggest that the complete implementation of the updated ContextMirror approach is in fact able to detect protein-protein interactions based on coevolutionary information alone.

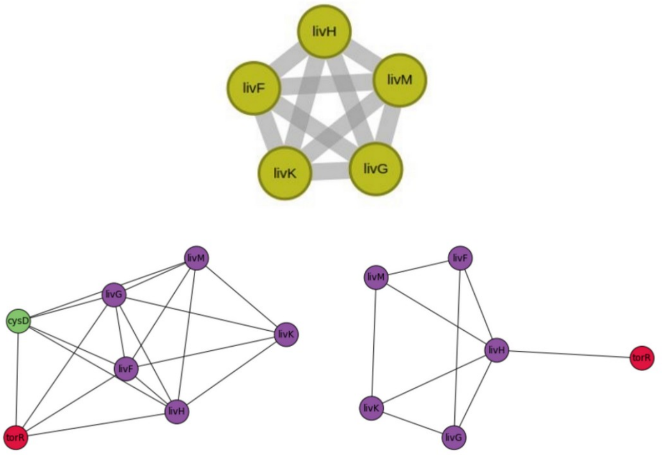


Fig. 2 Experimentally validated interactions in the LIV–I protein complex (top). Predictions of the updated ContextMirror approach for the LIV–I protein complex (purple) without (bottom left) and with filtering (bottom right).

Figure 2 presents the prediction of the LIV-I branched chain amino acid and phenylalanine ABC transport system (LivJHMGF) compared to the protein complex retrieved from experimentally validated databases. Different analytic strategies have been tested and, so far, filtering the absolute correlation values by third protein influence (≤3) and correlation coefficient ($r > 0.8$) has shown to improve the prediction by removing many false positives from the signal, recovering 8/10 real interactions and only one false positive (torR-livH).

Currently, we are still fine-tuning the parameters to increase accuracy and remove false positives. The implementation is being tested with the full proteome of Escherichia coli, and we will complete the parameter optimization to increase the performance comparing the results derived from the full proteome against experimentally validated databases. Later, we intend to use this optimized software to perform a comparative analysis of the predicted interactomes of different bacteria to spot differences in pathways that might be of importance in antibiotic resistance and other relevant functions.

ACKNOWLEDGEMENTS

I personally wish to acknowledge the Life Sciences Department at the Barcelona Supercomputing Center for their assistance in the production of this project and also Dr. Florencio Pazos from the Computational Systems Biology Group of the National Centre for Biotechnology (CNB-CSIC) for the creation and maintenance of the MirrorTree Webserver and his cooperation with us during the development of the updated implementation of the ContextMirror approach.

References

[1] Pazos, F., & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Engineering, Design And Selection, 14(9), 609-614. doi: 10.1093/protein/14.9.609

[2] Juan, D., Pazos, F., & Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proceedings Of The National Academy Of Sciences, 105(3), 934-939. doi: 10.1073/pnas.070967110

[3] Shatnawi, M. (2015). Review of Recent Protein-Protein Interaction Techniques. 99–121. doi: 10.1016/B978-0-12-802508-6.00006-5

[4] David Ochoa & Florencio Pazos. Studying the co-evolution of protein families with the Mirrortree web server. Bioinformatics (2010) vol. 26 (10) pp. 1370-1. doi: 10.1093/bioinformatics/btq137

Author biography

Miguel Fernández was born in Alcalá de Henares, Spain in 1999. He received a degree in Genetics from the Autonomous University of Barcelona (UAB), in 2021, and a Masters degree in Bioinformatics from the Autonomous University of Barcelona with a mention in protein structure and drug design, in 2022. For the last year he has been working in the Computational Biology group in the Life Sciences department at the Barcelona Supercomputing Center (BSC-CNS), conducting research on protein coevolution and its implication in protein-protein interaction and functionality. His current interests include the study of protein structure and how it is maintained throughout evolution.

Tailored Molecular Modelling and Machine Learning solutions for small-molecule drug discovery

I Filella-Merce[#1], J Vilalta-Mor[#], V Guallar[#*&]

[#]*Life Science Department, Barcelona Supercomputing Center (BSC), Plaça d'Eusebi Güell, 1-3, 08034, Barcelona, Spain*
[*]*Nostrum Biodiscovery S.L., Av. de Josep Tarradellas, 8-10, 3-2, 08029, Barcelona, Spain*
[&]*ICREA, Pg. Lluis Companys 23, 08010, Barcelona, Spain*
[1]isaac.filella1@bsc.es [2]victor.guallar@bsc.es

***Keywords*— Machine Learning, Generative Modelling Networks, Molecular Modelling, Active Learning**

Extended Abstract

A. Introduction

Computational techniques can help speed up the drug discovery process while reducing its associated costs [1]. Thanks to the breakthrough of vast biological and chemical data, Machine Learning approaches started to rise. Particularly, generative modelling networks (GMN) have benefited from using large chemical datasets of molecules [2]. These datasets are required to successfully train GMNs, as they use them to learn how to generate chemically viable molecules. To further refine the model and obtain target-specific molecules, GMNs can incorporate an additional specific set of molecules with demonstrated experimental activity with the target.

Additionally, an active learning phase can be integrated into a GMN through an iterative process to improve the model's predictability. At this stage, molecular modelling (MM) methods can take part. As determined by MM descriptors, favorable molecules are included in the specific set for the next generation round, while unfavourable ones are discarded.

B. Results and Workflow

This work aims to combine a GMN with cutting-edge MM techniques to generate novel target-specific molecules. We utilised a text-based GMN powered by an active learning phase of three steps: 1) molecules are accepted or discarded according to chemoinformatics descriptors, 2) the remaining molecules are screened using a fast rigid body docking method, 3) the affinity of the top molecules is estimated using PELE [3], a quick Monte Carlo simulation method which rescores the docking poses. This pipeline allowed us to identify and retain only the most promising molecules for further *in silico* and experimental validation.

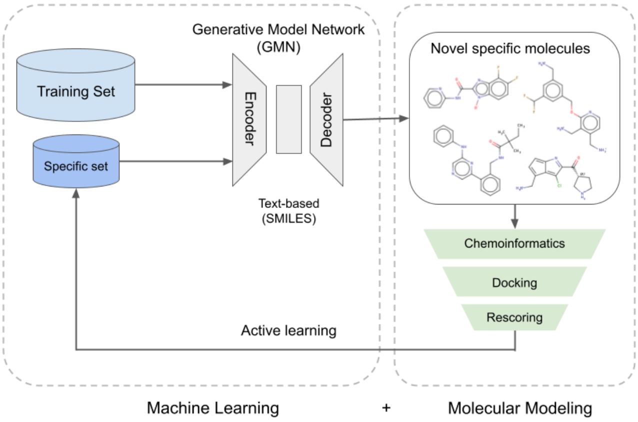


Fig. 1 Schematic Workflow of the coupling of a GMN with an active learning step powered by MM techniques.

References

1. Lin X, Li X, Lin X. Molecules. 2020 Mar 18;25(6):1375.
2. Bilodeau C, Jin W, Jaakkola T, Barzilay R, and Jensen K F. 2022. Comput. Mol. Sci. Vol. 12 Issue 5.
3. Borrelli K, Vitalis A, Alcantara R, Guallar V. Journal of Chemical Theory and Computation Int Ed. 2005, 6, 1304-1311.

Author biography

Isaac Filella Merce was born in Barcelone, Spain, in 1992. He received the B.E. degree in Mathematics and the B.E. degree in Physics from the University of Barcelone (UB), Spain, in 2016. In 2018 he obtained the Msc. degree in Bioinformatics from the University Pompeu Fabra (UPF) Barcelone, Spain.

In 2019 he began his PhD at the Structural Bioinformatics Unit at the Institut Pasteur in Paris, France, under the supervision of Michael Nilges and Riccardo Pellarin. In 2022 he obtained his PhD from Sorbonne University in Paris, France. Since Abril 2022, he work as a Post-Doc at the Electronic and Atomic Protein Modelling group at Barcelona Supercomputing Center (BSC), where he works under the supervision of Victor Guallar.

What is the Added Value of Climate Information for Predicting Infectious Disease Outbreaks?

Chloe Fletcher[1,2*], John Palmer[3], Rachel Lowe[1,4]

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[3]Department of Political and Social Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[4]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

[*]chloe.fletcher@bsc.es

***Keywords*— Infectious Diseases, Climate Change, Public Health, Bayesian Model, Reanalysis, Downscaling, Novel Data Streams**

EXTENDED ABSTRACT

Climate change is exacerbating the spread of infectious diseases. Models can identify drivers and predict outbreaks, providing insights to aid local epidemic preparation. Whilst models are valuable forecasting tools, derived climate-disease associations are impacted by data inputs, spatial coverage and downscaling methods. These factors propagate errors through model frameworks, affecting disease predictions downstream. There are few studies assessing these effects, signifying a gap in the literature. This research seeks to forecast epidemics in multi-scaled endemic settings and evaluate the added value of climate inputs, downscaling methods and novel data streams.

A. Introduction

Infectious diseases are increasingly emerging and re-emerging globally, amplified by rising temperatures, changes in precipitation and more frequent and intense weather events caused by climate change [1]. This disease transmission is further compounded by socioeconomic inequities, human mobility and the biology of vectors and pathogens [2].

Models can quantify climate-driven disease risk at multiple scales. However, predictions are impacted by data quality, spatiotemporal coverage, inputs and downscaling [3]. These factors can introduce errors into the model framework which affect disease predictions and decisions made downstream. Furthermore, novel data streams offer new means to enhance disease surveillance [4], including the identification of vectors via citizen science. Few studies have evaluated the added value and limitations of data inputs and downscaling methods in disease forecasts, highlighting a weakness in the literature.

This research will evaluate the added value of integrating downscaled climate inputs and novel data into early warning systems by: (i) developing impact-based forecast models for local decision makers; and (ii) investigating the effects of data inputs, spatial scales and downscaling on disease predictions.

B. Case Studies

Four case studies in endemic settings will be explored:

a) Using climate forecasts to inform disease risk in Caribbean small island developing states (SIDS);
b) Investigate the impacts of spatial scale and additional data inputs on dengue prediction in Vietnam;
c) Evaluate the added value of applying downscaled climate inputs and novel data streams into a multi-disease early warning system for Bangladesh; and
d) Assess the effects of statistical downscaling methods on climate-driven disease prediction.

C. Materials and Methods

a) Case Study A: Caribbean SIDS

Caribbean SIDS are highly vulnerable to climate change and, over the last decade, have experienced concurrent outbreaks of infectious diseases, prompting urgent calls for regional capacity building [1].

In this work, a temporal Bayesian model will be developed with monthly leptospirosis cases in Barbados as the response; climatic factors as covariates (including extreme events); and seasonal and interannual random effects. Distributed lag nonlinear models (DLNMs) will be used to capture nonlinear, lagged associations of 1-6 months. The best model will be selected via cross validation and goodness-of-fit metrics.

Once drivers and associations have been established, seasonal climate forecasts will be used to make probabilistic predictions of leptospirosis outbreaks with up to 3 months lead time. Hindcasts will be made to evaluate and calibrate forecasts, which will be downscaled to weather station data, bias corrected and quality assessed. The disease forecast will then be evaluated against out-of-sample posterior predictive distributions to assess the model's ability to produce probabilistic predictions of leptospirosis outbreaks.

Further work will explore the concurrent prediction of dengue and leptospirosis outbreaks using downscaled seasonal climate forecasts. This framework will be extended to model climatic interactions with dengue in Grenada and St Lucia.

This work supports the Caribbean Public Health Agency (CARPHA) early warning initiative and contributes to the Wellcome Trust funded modelling project, IDExtremes.

b) Case Study B: Vietnam

This work will draw upon research conducted under the D-MOSS project and will directly contribute to the Horizon Europe funded initiative, E4Warning.

To maintain consistency with previous iterations, spatio-temporal models will be fitted in a Bayesian framework using R-INLA. Associations between dengue cases and climate, environmental and socioeconomic covariates will be estimated at various spatial scales (province, district and city level). Seasonal climate forecasts and a bespoke hydrological forecast model will be used to predict dengue outbreaks up to 4 months in advance for each spatial scale, using a threshold-based approach to determine the likelihood of an outbreak. Model associations and prediction skills will be compared to establish optimal spatial scales for an operational dengue early warning system in Vietnam.

This research will also quantify the added value of the hydrological forecast model to predict dengue in Vietnam, and consider other novel data streams such as human mobility, vector abundance and seroprevalence data.

c) Case Study C: Bangladesh

This case study contributes to the Horizon Europe funded initiative, IDAlert. The research will establish key climatic, socioeconomic and environmental drivers for two diseases, selected by local stakeholders, using a Bayesian model framework. Previous methods employed during this PhD will be consolidated, including: (i) the prediction of multiple diseases using climate forecasts (Case Study A), and (ii) quantifying the added value of novel data streams, likely focusing on vector data from citizen science and smart traps, to predict disease outbreaks (Case Study B).

Furthermore, this work will assess how different climate data products impact downstream disease predictions. A forecasting model will be developed for each disease and climate data product, including reanalysis, satellite and weather station data. These models will be evaluated and compared to determine how the choice of climate data input affects subsequent disease predictions.

d) Case Study D: Downscaling

This study will utilise an existing prediction model which employs a similar Bayesian model framework to Case Studies A-C. This model will incorporate climate covariates from reanalysis products, ERA5 or ERA5-Land, and subseasonal-to-seasonal (S2S) climate forecasts to make probabilistic predictions of disease risk 1-3 months in advance. Different statistical downscaling techniques will be applied to resolve S2S forecasts at a consistent spatial scale to ERA5 or ERA5-Land. Associations between covariates and the response along with disease predictions will be evaluated for each downscaling technique and compared. This will help identify optimal spatial scales and downscaling methods to improve disease prediction accuracy in health impact models.

D. Purpose and Impact

Gridded climate products are solutions to overcome sparse weather station coverage and provide spatially continuous datasets. However, output accuracy can vary due to biases and disparities in data sources, model structures and assumptions [3]. These inaccuracies are rarely considered in health impact models, potentially affecting disease outcomes downstream. One study showed that 3 out of 4 climate products mis-characterised the relationship between rain, temperature and malaria in Machala, Ecuador [3]. This demonstrates how off-the-shelf usage of data products can undermine the reliability of climate-disease modelling and decision making. Further, spatial units and downscaling may fail to capture variability and dynamics, leading to inaccurate assessments.

There is therefore an urgent need to evaluate how different climate products and downscaling techniques individually impact disease outcomes across multiple spatial scales and settings. This will help build awareness of the associated methodological and scaling issues and provide a framework to strengthen health impact models and early warning systems.

Another challenge in climate-health modelling is finding a balance between complexity in dynamics and generalisability for effective prediction. This raises questions about how much value data inputs add to disease forecasts. Recent studies have shown strong associations between extreme weather and disease outbreaks [1, 5], which requires further exploration in other endemic settings. Additionally, probabilistic approaches rely on random effects to account for unknown characteristics, which could prove less significant when incorporating other relevant data inputs, potentially from novel data streams.

Explicitly evaluating the added value of tailored and novel data inputs may highlight new ways to enhance disease prediction. This could inform new data collection and policy interventions to reduce disease burden in at-risk populations.

ACKNOWLEDGEMENTS

This research will contribute to Wellcome Trust funded projects, HARMONIZE (224694/Z/21/Z) and IDExtremes (226069/Z/22/Z); Horizon Europe funded projects, IDAlert (Grant Agreement 101057554) and E4WARNING (Grant Agreement 101086640); and the CARPHA early warning system initiative, Strengthening Climate Resilient Health Systems in the Caribbean. We are grateful for all resources provided by project contributors to support this research.

References

[1] R. Lowe, A. Gaspirrini, C.J. Van Meerbeeck, C.A. Lippi, R. Mahon *et al.*, “Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study,” PLoS Med., Vol. 15, No. 7, Art. e1002613, Jul. 2018.

[2] J. Hess, L.G. Boodram, S. Paz, A.M. Stewart-Ibarra, J.N. Wasserheit, and R. Lowe, “Strengthening the global response to climate change and infectious disease threats,” BMJ, Vol. 371, Art. m3081, Oct. 2020.

[3] I.K. Fletcher, A.M. Stewart-Ibarra, M. García-Díez, J. Shumake-Guillemot and R. Lowe, “Climate services for health: From global observations to local interventions,” Med, Vol. 2, No. 4, Pp. 355-361, Apr. 2021.

[4] B.M. Althouse, S.V. Scarpino, L.A. Meyers, J.W. Ayers, M. Bargsten *et al.*, “Enhancing disease surveillance with novel data streams: challenges and opportunities,” EPJ Data Sci., Vol. 4, Art. 17, Oct. 2015.

[5] R. Lowe, S. Lee, K.M. O’Reilly, O.J. Brady, L. Bastos *et al.*, “Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: A spatiotemporal modelling study,” PLoS Med., Vol. 5, No. 4, Pp. e209-e219, Apr. 2021.

Author biography

Chloe Fletcher is a PhD candidate within the Global Health Resilience group at the Barcelona Supercomputing Center and the Biomedicine PhD in the Department of Medicine and Life Sciences at Universitat Pompeu Fabra. Chloe obtained an MSc in Environmental Modelling at UCL in 2018, where she achieved the Dean’s List award for outstanding academic performance. She also gained a BSc in Mathematics from UCL in 2015.

Since joining the BSC in March 2022, Chloe’s research has focused on the development of statistical models to forecast climate-sensitive infectious diseases outbreaks across endemic settings and evaluate how data impacts outbreak predictions.

DNN Acceleration in the Limits of Energy Efficiency

Jordi Fornt*†
Francesc Moll*†, Josep Altet †,
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: jordi.fornt@bsc.es, {francesc.moll, josep.altet}@upc.edu

***Keywords*—Hardware accelerators, energy efficiency, deep learning, neural networks, artificial intelligence**

I. Extended Abstract

Deep neural network (DNN) models dominate state-of-the-art artificial intelligence. Due to their high computational cost, the use of hardware accelerators is key in any system that trains or deploys deep learning models. GPUs are widely used for executing neural networks, but they are too power hungry for many applications, especially in edge computing. In the last years, a wide variety of deep learning-specific accelerators have been proposed that greatly outperform GPUs in terms of energy efficiency, but even those accelerators are far from the desirable efficiency for many applications. In this extended abstract, we present two accelerator architectures we have designed in order to experiment with different energy efficiency maximization ideas.

A. SAURIA: dataflow accelerator with on-the-fly im2col and approximate arithmetic

DNN-specific accelerators exploit the embarrassing parallelism of DNNs to achieve high throughput and energy efficiency. Instead of replicating general-purpose compute cores, like GPUs and vector processors, DNN accelerators use massive replication of simple multiply-add units [1], which are enough to perform most of the operations of these models. Dataflow accelerators, in particular, use parallelization to achieve a high degree of data reuse, a key feature of modern accelerators that allows them to amortize the energy of each memory access as much as possible.

SAURIA (Systolic Array tensor Unit for aRtificial Intelligence Acceleration) is a Convolutional Neural Network (CNN) accelerator built using a systolic array, which is a highly-pipelined dataflow architecture that allows for very high throughput. Its array (see Figure 1) features a mesh of Processing Elements (PEs) that jointly perform a matrix-matrix multiplication with a high degree of parallelism and data reuse over several clock cycles. The array dataflow is output stationary, meaning that the output values (or partial sums) stay in-place in the PEs during the computation of a context, while the input and weight values throughout the pipeline change every clock cycle.

The systolic array at the core of SAURIA is able to execute matrix-matrix multiplications efficiently, but in order to accelerate convolutional neural networks, the input tensors must be lowered using *im2col*. Even though it can be executed very efficiently and with minimal memory footprint, software-based *im2col* produces a large inflation of the data that must be transferred from main memory to the accelerator. This causes an increase in the total memory transactions (as much as x9 for 3x3 kernels) that compromises the energy efficiency of the system, since DRAM accesses are very energy-expensive. To solve this, SAURIA's IFmap Feeder supports performing on-the-fly *im2col* to the input tensors, so that they can be transferred in their original shape. This strategy saves up to 65% of memory traffic when executing YOLOv3 object detection network.

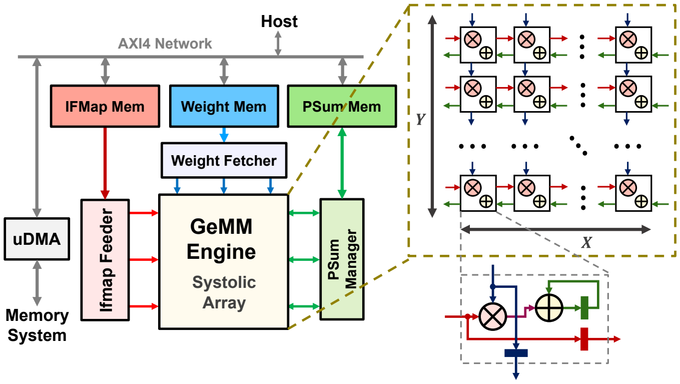


Fig. 1: Architecture of the SAURIA accelerator.

A version of SAURIA featuring an 8x16 array of FP16 PEs was integrated in the Kameleon SoC, the final chip of the DRAC project [2], which was fabricated as an ASIC in 22 nm technology. Figure 2 shows the layout of the accelerator after the full physical design flow, and Table I summarizes its main specifications. In order to maximize energy efficiency, the accelerator PEs were modified to use approximate multipliers

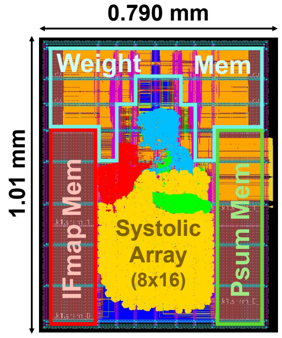


Fig. 2: Physical layout (Kameleon's SAURIA).

Technology Node	22 nm
Rows (Y)	8
Columns (X)	16
Memory sizes	32 kB
Arithmetic	FP16
Supply Voltage	0.8 V
Max. Frequency	500 MHz
Power	91 mW
Peak GFLOP/s	128
Energy Efficiency (GFLOP/sW)	1406

TABLE I: Specifications (Kameleon's SAURIA).

and adders in the mantissa computation. In particular, we use the ABM-M3 booth multiplier proposed in [3] and a Set-One Adder [4]. By tuning the parameters of these approximate units taking into account the tradeoff between power reduction and CNN accuracy degradation, we are able to improve energy efficiency by 30% with negligible accuracy losses. We use the YOLOv3 network to benchmark accuracy, by simulating its execution using the proposed approximate FP16 unit.

B. Bit-Serial Computing: exploiting flexible precision and selective undervolting

Dataflow accelerators, like SAURIA, typically rely on a fix arithmetic precision for the whole accelerator. This poses an inherent issue in terms of energy efficiency optimization, since the precision requirements of deep neural networks vary between layers of a particular model.

Bit-serial accelerators [5] tackle this issue by rearranging the computations so that the input operand bits become sequential loops that can be arbitrarily configured. In this setting, the multiplications in the PEs become of size 1xN or even 1x1 bits, which can be implemented by AND gates. By applying a shift to the multiplication result before the accumulation, the full result can be reconstructed bit by bit (see Listing 1). Thanks to this loop reordering, bit-serial accelerators are able to compute convolutions and matrix-matrix multiplications with fully-flexible precision.

From this concept, we have designed a parametrizable bit-serial accelerator architecture, depicted in Figure 3. This accelerator is a work in progress for which we still do not have definitive results, so we will focus on the main ideas of its architecture. The core component of the design is an array of parallel AND gates (1-bit multipliers) with adder trees to reduce the input channels dimension. The shift and accumulate operations are split in two levels, following a design idea similar to memory caching. In the first shift-accumulate stage, stored in the Array Accumulator (L0 cache), a small shift is performed for all partial results at every clock cycle. On the second stage, the PSum Accumulator (L1 cache) performs the remaining shift and accumulates the full partial sum values.

Listing 1: Bit-serial computation of a convolution.

```
for bi in range (i_bits):       # Bit-serial mult
 for bw in range (w_bits):      # Bit-serial mult
  for j in range(ky):           # Kernel size y
   for i in range(kx):          # Kernel size x
    for y in range(Y):          # Tensor size y
     for x in range(X):         # Tensor size x
      for k in range(K):        # Output Channels
       for c in range(C):       # Input Channels

        P[k,y,x] +=
        (I[c,y+j,x+i][bi] & W[k,c,j,i][bw])<<(bi+bw)
```

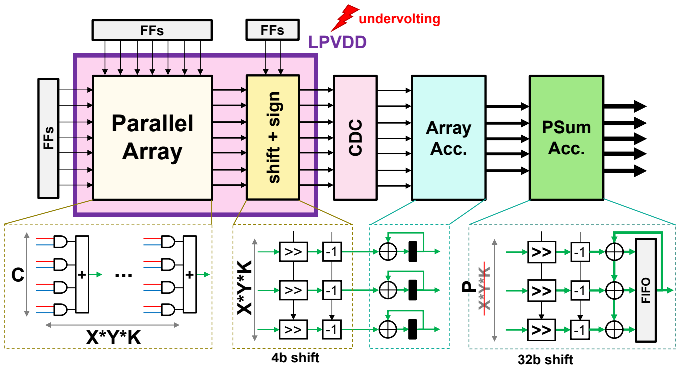


Fig. 3: Architecture of the bit-serial accelerator prototype.

The purple region highlighted in Figure 3, which contains the logic responsible for most of the power consumption, uses an independent power supply. By reducing the supply voltage of this region we can decrease the accelerator power without affecting the overall system, up to a certain limit in which we start experiencing setup timing violations due to excessive gate delays. Undervolting consists on going past this limit and letting timing violations happen due to an aggressively low voltage supply. This technique has a very good synergy with bit-serial computing, since we can apply aggressive undervolting when computing the LSBs of the operands, generating low-impact errors, and raise the voltage when computing the MSBs to avoid high-impact ones. To solve possible metastability issues during undervolting, we include a Clock-Domain Crossing (CDC) module to synchronize the outputs of the independent power domain.

C. Conclusion

We have presented two experimental accelerator architectures developed for optimizing the energy efficiency of deep learning workloads: SAURIA, a systolic array-based CNN accelerator with on-chip *im2col* and approximate arithmetic, which has been fully implemented in an ASIC; and a bit-serial accelerator that seeks to exploit mixed precision computing and undervolting to push energy efficiency further.

REFERENCES

[1] Y. Chen *et al.*, “Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 292–308, 6 2019.

[2] “Home — DRAC project,” https://drac.bsc.es/en.

[3] S. Venkatachalam *et al.*, “Design and Analysis of Area and Power Efficient Approximate Booth Multipliers,” *IEEE Transactions on Computers*, vol. 68, no. 11, pp. 1697–1703, 2019.

[4] W. Liu *et al.*, “Design and Evaluation of Approximate Logarithmic Multipliers for Low Power Error-Tolerant Applications,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 9, pp. 2856–2868, 2018.

[5] S. Ryu *et al.*, “BitBlade: Energy-Efficient Variable Bit-Precision Hardware Accelerator for Quantized Neural Networks,” *IEEE Journal of Solid-State Circuits*, vol. 57, no. 6, pp. 1924–1935, 2022.

Jordi Fornt received the B.S. degree in industrial electronics from the Universitat Politècnica de Catalunya (UPC) in 2019, and the M.S. degree in electrical engineering and information technology from the Swiss Federal Institute of Technology of Zürich (ETH Zürich) in 2021. He is currently pursuing the Ph.D. degree with the High-Performance Integrated Circuits (HIPICS) group from UPC. During 2020 he worked as research intern at IBM Research Zürich, with the In-Memory Computing group. In 2021, he joined the Barcelona Supercomputing Center (BSC), where he currently conducts research towards his Ph.D. degree in energy-efficient deep learning accelerator design.

Exhaustive Variant Interaction Analysis using Multifactor Dimensionality Reduction

Gonzalo Gómez Sánchez*†, Ignasi Morán†, Josep Ll. Berral*†,
*Universitat Politècnica de Catalunya, Barcelona, Spain
†Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {gonzalo.gomez, ignasi.moran, josep.ll.berral}@bsc.es

***Keywords*—genomics, variant interaction, High-performance computing.**

I. Extended Abstract

In human genetics, the goal is to create a map between genotype and phenotype. Typically, the effect of each variant with the complex disease is studied one at a time, and then, the sum of the effects is calculated to analyze the disease risk. This model, called additive, assumes that the effect of each variant is independent of the rest of them. In Epistatic studies [1], the question asked is different: each variant is studied to see if, with the interaction of at least a second variant, they can be related to the disease.

The challenge of finding epistatic interactions lies in the size of genomic data, where methods based on frequentist statistical inference that is common in the study of single variants are not viable. Here we apply Multifactor Dimensionality Reduction (MDR) [2], a statistical approach for detecting a combination of variants combined with a data filtering based on the Chi-Square. The pairwise combinations detected by the method show a clear enrichment, indicating that we are finding variant interactions that are related to the disease.

A. The Data

1) The dataset: The dataset used for this study is the Northwestern NuGENE project cohort [3]. It is composed of 11,297,253 variants, forming a dataset of 11,297,253 rows x 3,389 columns. In a variant interaction analysis, we are studying the association between each pair with the disease, which means that in the most simple case, we are going to process every possible pair. This means that the number of combinations ascends approximately to (11,297,253 x 11,297,252)/2 = 63,813,957,024,378. This volume of combinations is the major problem of all epistatic studies and the reason why we need HPC frameworks such as the one presented to process the datasets.

2) Filtering the data: since we are testing the interaction between pairs, if we want to reduce the search space we need to do it taking into account this interaction. A frequentist statistical method that gives a measure of how significant the distribution of cases/controls is, is the Chi-Square method. Using a contingency table, it performs a statistical test that examines the differences between the expected values and the observed ones.

Typically, it provides a value of significance that, corrected by multiple testing could select by itself which pair is significant and related to the disease. However, because of the size of the data, this correction will only find extremely significant cases, leaving out many possible interesting interactions. Here we use it as a filter for data selection, using an empirical significant value.

B. The Method

1) Multifactor Dimensionality Reduction: With the aim of improving the power to detect variant-variant combinations, MDR is a statistical method that, based on contingency tables, reduces the dimension of the problem. More specifically, it converts the counts obtained for the cases and controls into a simple binary variable by classifying all the possible allelic combinations for each pair in high-risk/low-risk, reducing the analysis to only one dimension. It follows a naive Bayes approach, building a probabilistic classifier from every variant-variant interaction and summarizing the best combinations for prediction. As illustrated in fig.1, the method can be divided into 5 different steps:

- The first step consists of dividing the data into a training set and a testing set for a k-fold cross-validation. This step is done to avoid potential overfitting and, therefore, to find pairs of variant-variant combinations that can have an effect on any dataset, not just the one being evaluated.
- In the second step, we will build a contingency table for each of the evaluated variant-variant pairs. We are working with a 2 factors table (variant-variant), each one with 3 classes corresponding to the variant genotypes (AA, Aa, aa), ending with a 9 cells table called multifactor classes. Then, each multifactor class is going to be filled with the number of cases and controls, building the 2-dimensional table.
- In step three, the tables are going to be transformed into a 1-dimension space using the original cases and control distribution of the dataset as a threshold, T. The cells where the ratio is bigger than T are going to be classified as 'high risk' and as 'low risk' if the ratio is lower.
- In step four, each of the variant-variant table is used to classify the testing set of the data and sort all the variant-variant tables using the prediction error.

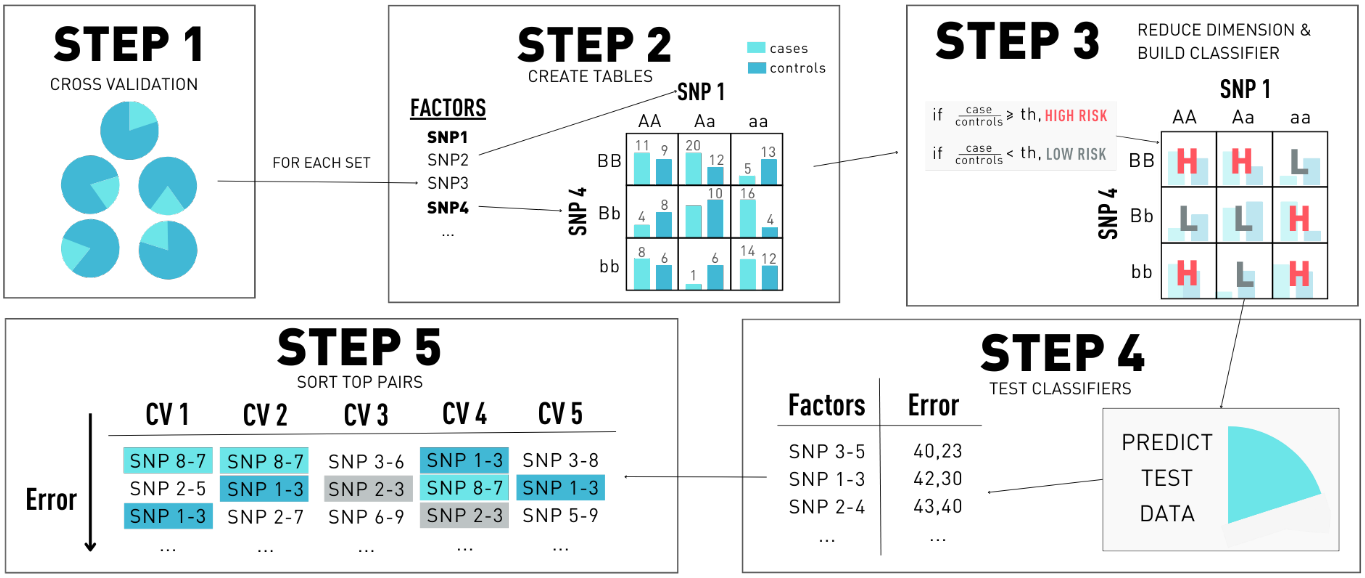


Fig. 1. Multifactor Dimensionality Reduction algorithm steps. Step 1 is the cross-validation division. Step 2 is the building of contingency tables. In step 3, the dimension of the contingency tables is reduced to 1. In step 4, each multifactor is tested. Steps 2-4 are repeated for each cross-validation set and in Step 5, the top pairs are selected.

Then, step two, three, and four are repeated for each of the possible cross-validation sets from step one.

- In step five, the best and most consistent variant-variant combinations are selected, which means picking the ones that appear in the most cross-validation sets as a top pair.

The method has been developed leveraging Apache Spark framework for parallel computation [4]. Apache Spark is an open-source distributed processing system for high-volume workloads. Thanks to its in-memory caching and a system of optimized queries, it can be used against data of any size. It has a master-slave architecture, combining a single master with multiple slaves. Despite using a parallel environment, the computation power needed to process 63 thousand billion pairs in a feasible time using MDR required huge computational resources. That is why we perform a data filtering using the Chi-Square described above.

C. The experiments

The first part of the experiments has been performed to choose the Chi-Square value for filtering the data. In order to do so, the first step has been performing a Chi-Square test on all the possible combinations of the NuGENE dataset. Then, we selected different percentages of the most significant pairs and apply the MDR to them. We detected that the MDR was not selecting as significant any pair that had a Chi-Square significance value above 10^{-7}. Therefore, using this value to filter the pairs, we ensure that no potentially significant pair is left outside the analysis. This step allows us to decrease the number of pairs to be processed by the MDR from thousand of billions to less than 2 million pairs (1,883,192). Then, from these pairs, MDR selected 104 pairs as significant in all the cross-validation sets performed.

D. Conclusion

While Chi-Square by itself lacks the power to detect significant pairs, it has the advantage of its simplicity, having a very low computational expense. On the other hand, MDR has the power but also a high computational expense. Combining both, we achieve an efficient method to perform an exhaustive study of the cohort NuGENE. From the 104 pairs obtained with our method, we found that 93% of these variants is a GWAS variant that has already been marked as significant individually, proving an enrichment of the method. The other 7% correspond with unmarked variants that may have also an effect on the disease. Furthermore, after a functional analysis, we detected that there where 2 pairs that had a direct relation with the paths that can affect a disease such as Type 2 Diabetes.

References

[1] e. a. Wood, "Another explanation for apparent epistasis," *Nature*, vol. 514, 2014.

[2] M. Ritchie *et al.*, "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity," *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, vol. 24, no. 2, pp. 150–157, 2003.

[3] e. a. Omri Gottesman, "The electronic medical records and genomics (emerge) network: past, present, and future," *Genetics in Medicine*, 2013.

[4] M. Zaharia *et al.*, "Apache spark: a unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.

Gonzalo Gómez received his BSc degree in Telecomunications Engineering from Universitat Politècnica de Catalunya (UPC), Barcelona in 2016. He completed his MSc degree in Artificial Intelligence from Universitat de Barcelona (UB), Universidat Rovira I Virgill (URV) and UPC, Catalunya in 2018. Since 2018, he has been a PhD student at the department of Computer Science in Barcelona Supercomputing Center (BSC) and Universitat Politècnica de Catalunya (UPC), Spain.

Climate Services Ecosystems: What are they and why they are important

Gonzalez Romero, Carmen[1,1], Muñoz, Ángel G. [1,2], Goddard, Lisa[3]

[1]Earth Science Department, Barcelona Supercomputing Center, C/Jordi Girona, 29, Barcelona, Spain

[1]carmen.gonzalezromero@bsc.es, [2]angel.g.munoz@bsc.es,

[3]deceased 2022

***Keywords*— Climate services ecosystem, resilience, bayes theorem, adaptation, climate change and natural climate variability**

EXTENDED ABSTRACT

Societies use climate services as part of their mitigation and adaptation strategies to a changing climate [2] Traditionally, climate services have been defined and framed around particular single applications or sectors, either agriculture, health, energy, water management or disaster risk management [3] -just to mention a few. Whilst this can bring potential benefits such as high specialization and adaptation [8], co-benefits of articulated climate services among different sectors have not been fully assessed in the broader societal system, where these are developed and implemented. Understanding and valuing the nexus between the sectors during the design, development and implementation of climate services might help project optimization, and eventually benefit the community, country, entire region or society.

Climate services ecosystems are defined -slightly modifying the business-perspective definition of Vargo and Akaka [13]- as relatively self-contained, self-adjusting systems of resource-integrating actors connected by shared institutional goals, and mutual-value creation through exchange of climate services [4]. In other words, a climate-services ecosystem involves interactions between different sectors sharing the same or similar climate services, which enhances resilience, and lends efficiency and value, by optimally orchestrating the available solutions. These ecosystems tend to be more robust to climate impacts than a collection of climate services focused on certain applications or just one sector, because shocks to one part of the ecosystem are redistributed and dampened through the entire network.

Since, by definition, these ecosystems take advantage of existing climate services in different society-relevant sectors, the overall benefit is directly dependent on the ecosystem configuration itself. The ability to scale high-quality climate services, not just to other locations but to other sectors, and the ability for these climate-service networks to organize into ecosystems is hypothesized to be a crucial ingredient to resilience in the face of climate variability and change, given that resources are finite.

A first step to address these issues would be to define rules of games and resources for the orchestration of potential and actual climate services interacting in a given space. Due to the similarities with the ecological definition, we propose here to define that space of interactions as climate services ecosystems.

Following the definition of climate services, as the development of climate data and climate knowledge, the translation, communication and use; a first attempt to define climate-services ecosystems was recently proposed [4] as slightly modifying the business- perspective definition of Vargo and Akaka- defined a couple of paragraphs above. Nonetheless, this definition is not as straight-forward as one might desire, as it includes concepts (e.g., self-contained, self-adjusting) that are not trivially understood without further definitions.

A. Definition of climate services Ecosystem

A climate-service ecosystem is a dynamic complex network of institutions, agents, information, knowledge, product and services functioning as a unit in any size scale with the objective of enhancing the resilience of the system to a changing climate. Similarly, to environmental ecosystems, the climate-services ecosystems are self-contained and self-adjusting, having the ability to adapt in response to changes in the network or system [10]. The elements of the ecosystems are interconnected and interdependent, in such a way that the more interconnected they are, the higher the value and resilience of the network is. Although this piece analyses just one particular size of climate-services ecosystems, the authors acknowledge that there are different sizes or scales of ecosystems and admit, that more likely than not, subsets of ecosystems can be found within bigger ones. Despite the size of the network, understanding the value and the dynamic of the interactions of the elements is essential to define the idea of climate services ecosystem at any size scale.

B. How to define the interactions within the ecosystem?

Certainly, a climate-services ecosystem involves interactions between different institutions, agents or sectors sharing the same or similar climate services. As per the definition of climate services, these interactions require climate data or climate knowledge share, as well as communication and feedback between users and providers of climate services within the system. These interactions within the network enhance its resilience, and lends efficiency and value, by optimally orchestrating the available resources [4] For example, climate-services ecosystems tend to be more robust to impacts than a collection of climate services focused on certain applications or just one sector, allowing the shocked received by the ecosystem to be redistributed and dampened through part or the entire. Some interactions might follow a functional redundancy (or repetition of climate services) helping to keep the equilibrium of the ecosystem when dealing with changes or shocks impacting the network. Some services or institutions might make a unique or singular contributions to the ecosystem functioning, while others might present a higher weight in the ecosystem due to the functional redundancy, or their role in the network - therefore their loss would be a great concern [10].

C. What is the value of the climate services ecosystem?

The concept of value has long been a philosophical question without a straightforward answer [11]. The utilitarian (anthropocentric) definition of value, widely used in many disciplines, frames the paradigm of value on the principal of humans' preference satisfaction [10], which is aligned with the user-centred perspective of climate services, widely accepted in the literature [1, 9]. The value definition of the climate service ecosystem is thus determined by the time and objective of the system itself, which should aim to respond to the particular needs of the users of the system in their decision-making process to adapt to a changing climate, at a specific point in time. This approach links the value of the climate service ecosystem to the concepts of risk, resilience and adaptive capacity [7], but it also allows to include budget constraints and resource optimization in the framework of climate services ecosystem- as it usually happens in the real world.

The term resilience refers to the state of the ecosystem and how it responds to different hazards, or crises under various probabilistic conditions [6]. It is subject to different state conditions established in a timely manner by the interactions and characteristic of the network in that specific point in time [5]. Therefore, it is natural to analyse these ecosystems through Dynamical Casual Network Theory which allows us to understand, characterize and foresee potential behaviour and changes in relationships between the elements of the networks, hence supporting decision-making processes for adaptation. The overall purpose is to provide an objective tool to define climate services ecosystems and the casual relationships between their elements, the dynamics of their interactions and the value of the ecosystem.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Lisa Goddard, former Director or the International Research Institute for Climate and Society at Columbia University, for her support as out tutor, colleague and friend through out her life. She inspired many of the work developed by the authors and, in particular, she co-authored the first publication on the topic of climate services ecosystem which led to the development of the thesis of the first author of this abstract. She will always be in our hearts as a great leader and inspiring our research.

References

[1] Buontempo, C. & Hewitt, C. & Doblas-Reyes, F. & Dessai, S., "Climate service development, delivery and use in Europe at monthly to inter-annual timescales. Climate Risk Management".6. 10.1016/j.crm.2014.10.002. 2014.

[2] Cortekar, J. & Bender, S. & Brune, M. & Groth, M., "Why climate change adaptation in cities needs customised and flexible climate services", Climate Services. 4. 10.1016/j.cliser.2016.11.002. 2016.

[3] Council, N.R. A, "Climate Services Vision: First Steps Toward the Future", https://doi.org/10.17226/101. 2001.

[4] Goddard L., Gonzalez Romero C., Muñoz A.G. et al. "Climate Services Ecosystems in times of Covid-19", World Meteorological Organization bulletin nº Vol 69 (2)- 2020.

[5] Haimes, Y.Y. "On the Definition of Resilience in Systems", Risk Analysis, 29: 498-501. https://doi-org.recursos.biblioteca.upc.edu/10.1111/j.1539-6924.2009.01216.x. 2009.

[6] Haimes, Y.Y. "On the Complex Quantification of Risk: Systems-Based Perspective on Terrorism". Risk Analysis, 31:1175-1186.https://doi-org.recursos.biblioteca.upc.edu/10.1111/j.1539-6924.2011.01603.x. 2011.

[7] Klein, R. J. T., Nicholls, R. J. and Thomalla, F., "Resilience to natural hazards: How useful is this concept?" Global Environmental Change Part B: Environmental Hazards, 5:1, 35-45, DOI: 10.1016/j.hazards.2004.02.001. 2003.

[8] Lemos, M. "Usable climate knowledge for adaptive and co-managed water governance", Current Opinion in Environment Sustainability, 12, 48–52. https://doi.org/10.1016/j.cosust.2014.09.005. 2016.

[9] Lu, J., Lemos, M. C., Koundinya, V. and Prokopy, L. S., "Scaling up co-produced climate-driven decision support tools for agriculture". Na1ture Sustainability, doi:10.1038/s41893-021-00825-0. 2021.

[10] Millennium Ecosystem Assessment "Ecosystems and Human Well-being: Synthesis". Island Press, Washington, DC.https://millenniumassessment.org/en/Framework.html. 2003.

[11] Perry, R. B., "The Definition of Value. The Journal of Philosophy, Psychology and Scientific Methods", 11(6), 141–162. https://doi.org/10.2307/2013053. 1914.

Author biography

Gonzalez Romero, Carmen, was born in Granada, Spain in 1987. She received two Master-degree level diplomas on Business Administration and Law by University of Granada, Spain, in 2011 and 2012, and a Master's in International Affairs by Columbia University, USA, in 2019. She is currently pursuing her PhD at University of Bologna.

In 2018 she started working at the International Research Institute for Climate and Society (IRI), at Columbia University, where she developed climate servicers specialized on agriculture and health in Latin America. In 2023, she joint BSC, in the Earth Science Department as part of the Knowledge Integration Team working on projects like Climateurope2, ASPECT or HARMONIZE. She also has 4 years of working experience in the FMCG private sector and has represented several of the institutions where she worked at international events like the COP.

A Stochastic Method for Solving Time-Fractional Differential Equations

Nicolas L. Guidotti [*], Juan A. Acebrón [*†], José Monteiro[*]
[*]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
[†]Department of Mathematics, Carlos III University of Madrid, Spain
E-mail: nicolas.guidotti@tecnico.ulisboa.pt, juan.acebron@tecnico.ulisboa.pt, jcm@inesc-id.pt

***Keywords*—Monte Carlo method, time-fractional differential equations, anomalous diffusion, Mittag-Leffler function, matrix functions, parallel algorithms, high performance computing.**

I. Extended Abstract

Among the many non-conventional statistical phenomena observed experimentally in the last few years, the anomalous diffusion stands out for its major impact in a variety of scientific disciplines [1]. In the anomalous diffusion, the mean squared displacement (MSD) vary nonlinearly in time, *i.e.*, $\langle x^2(t)\rangle \sim t^\alpha$ with $\alpha \neq 1$, during the entire process. Here $x(t)$ is the relative position at time t of a particle with respect to a given reference point. This contrasts with the classical diffusion problems, where it is well known that the MSD grows linearly in time. It has been proposed in the literature several pathways leading to anomalous diffusion. The most commonly found are the presence of long-range correlations, non-identical displacements, or non-finite mean or variance of the probability density function for the trajectory displacements. Within this context, fractional calculus has been proven to be extremely useful in order to provide realistic models for many real-life processes and phenomena [2], [3]. The main reason for this is due to fractional derivative operators being in practice non-local operators, and therefore they are specially suited for describing the long-time memory and spatial heterogeneity effects typically found in any anomalous diffusion problem.

The formal solution to the simplest dynamical problem defined by a system of linear fractional rate equations for an n-dimensional process $\mathbf{y}(t)$ subject to a given initial condition $\mathbf{y}(0) = \mathbf{y_0}$ is

$$\mathbf{y}(t) = E_\alpha(\mathbf{A}\, t^\alpha)\, \mathbf{y_0}, \tag{1}$$

where $\mathbf{A}$ is an $n \times n$ matrix of constant coefficients. When $\alpha = 1$, the Mittag-Leffler (ML) function $E_\alpha(\mathbf{A}\, t^\alpha)$ reduces to the matrix exponential $e^{\mathbf{A}\, t}$. The ML function also appears in other scientific domains [4], [5], [6], [7], [8], [9]. It is important to remark here that this system of linear rate equations serve as building-blocks for more complex systems. However, an efficient algorithm for solving large-scale problems is still missing. Therefore, our main contribution is the proposal of a probabilistic method for the efficient computation of the ML matrix function as well as an extensive analysis of its performance and convergence for a few relevant examples. Our second contribution was to parallelize and analyze the scalability of the stochastic algorithm for a large number of processors (up to $16,384$ CPU cores) in the Karolina Supercomputer located in IT4Innovations National Supercomputing Centre.

A. Description of the Method

Consider $\mathbf{A} = (a_{ij})$, a general n-by-n matrix with $a_{ii} < 0$ $\forall i$; $\mathbf{u}$, a given n-dimensional vector; and $\mathbf{y}$, an n-dimensional vector, such that $\mathbf{y} = E_{\alpha,\beta}(\mathbf{A}\, t^\alpha)\, \mathbf{u}$ with $0 < \alpha \leq 1$, $\beta > 0$ and $t \geq 0$. Here t denotes the value of time when the solution is computed. Let us define the following matrices: a diagonal matrix $\mathbf{D}$, represented hereafter as a vector $\mathbf{d}$, with entries $d_{ij} = 0$ for $\forall i \neq j$ and $d_{ii} = d_i = a_{ii}$ for $i = 1, \ldots, n$; $\mathbf{M}$, a matrix obtained as $\mathbf{M} = \mathbf{A} - \mathbf{D}$, and hence with zero diagonal entries; $\mathbf{Q}$, the matrix with entries $q_{ij} = |m_{ij}| / \sum_{j=1}^{n} |m_{ij}|$ for $i \neq j$, 0 otherwise; and finally a vector $\mathbf{r}$ such that $r_i = (\sum_{j=1}^{n} |m_{ij}|)/d_i$.

Theorem 1. *Let $\{X_t : t \geq 0\}$ be a stochastic process with finite state space $\Omega = \{1, 2, \ldots, n\}$ given by*

$$X_t = \sum_{m=1}^{\infty} Z_{m-1} \mathbb{1}_{[T_{m-1} \leq t \leq T_m]}. \tag{2}$$

Such process changes states according to a Markov chain $Z = (Z_m)_{m \in \mathbb{N}}$, which takes values in Ω and $\mathbf{Q}$ is the corresponding transition matrix. Here T_k is the time of the k-th event, and $\mathbb{1}_E$ denotes the indicator function, being 1 or 0 depending on whether the event E occurs. The sojourn times in the i-th state follows the Mittag-Leffler distribution $ML_\alpha(d_i)$. Then, we have that any entry i of the solution vector $\mathbf{y}$ can be represented probabilistically as

$$y_i = \mathbb{E}[u_{X_0} \mathbb{1}_{[T_0 > t]} + \omega\, \mathbb{1}_{[T_0 \leq t]}] \tag{3}$$

with

$$\omega = \left(\prod_{j=1}^{\eta} r_{X_{T_{j-1}}}\, g_{X_{T_{j-1}}, X_{T_j}} \right) u_{X_{T_\eta}} \tag{4}$$

where $X_0 = i$. Here $\mathbb{E}$ is the expected value with respect to the joint distribution function of the random variables T and η, where η is the number of events occurring between 0 and t.

Based on the Theorem 1, we can construct a stochastic algorithm that estimated the value of the i-th entry of the solution vector $\mathbf{y}$ through the simulation of the stochastic process X_t, which consists in generating random paths from the Markov chain Z and then computing the realization of a random variable ω over these paths. Each random path starts at state X_0 and time $\tau = 0$ and then jumps from one state to the next until it reaches the time $\tau = t$. The next state is

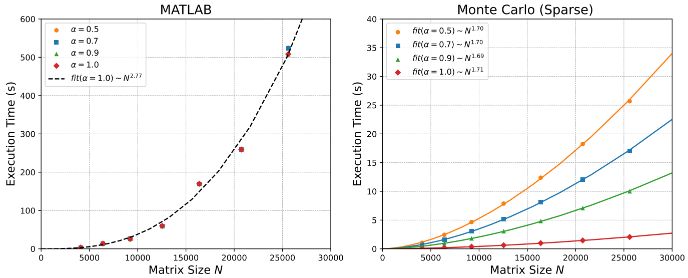


Fig. 1. The parallel execution times for solving the 2D diffusion equation as a function of the matrix size $N = m^2$ for $t = 0.1$ and 8 threads. For the Monte Carlo algorithms, the accuracy ϵ was kept fixed to be approximately 3×10^{-4}.

always selected at random based on the transition matrix $\mathbf{Q}$ and the sojourn time in each state follows an ML distribution.

B. *Experimental Results*

All simulations regarding the shared-memory architecture were executed on a commodity workstation with an AMD Ryzen 5800X 8C @4.7GHz and 32GB of RAM, running Arch Linux. The Monte Carlo Algorithm was implemented in C++ with OpenMP and uses `PCG64DXSM` [10] as its random number generator. The code was compiled with the Clang/LLVM v14.0.0 with the `-O3` and `-march=znver3` flags. We compare our implementation against the only freely available code capable of computing the matrix ML function [11]. It was written in MATLAB 2021a and is based on the Schur-Parlett algorithm [12], [13].

To compare the two implementations, we solved the 2D time-fractional diffusion equation:

$$D_t^{\alpha} u(\mathbf{x}, t) = \nabla^2 u(\mathbf{x}, t), \tag{5}$$

for $\Omega = [-1, 1]^2$, $t > 0$, $\mathbf{x} = (x, y) \in \mathbb{R}^2$, $u(\mathbf{x}, t)|_{\partial\Omega} = 0$ and $0 < \alpha < 1$. Considering a discrete mesh with m cells in each dimension, such that $\Delta x = \Delta y = 2\mu/m$, and the standard 5-point stencil finite difference approximation, the approximated solution $\hat{u}(\mathbf{x}, t)$ for (5) can be written as

$$\hat{u}(\mathbf{x}, t) = E_{\alpha}\left(\frac{m^2}{4\mu^2}\,\hat{\mathbf{L}}\,t^{\alpha}\right)\,\hat{u}_0(\mathbf{x}) \tag{6}$$

where $\hat{u}_0(\mathbf{x}) = c\,m^2\,\delta(\mathbf{x} - \mathbf{x}_c)$ with $\mathbf{x}_c = (m/2, m/2)$ and $c = 1/4096$. Fig. 1 shows the parallel execution time for different matrix sizes $N = m^2$ when using 8 threads. In terms of execution time, the Monte Carlo algorithm is several times faster than the `matlab` code, while maintaining a reasonable precision. When using 8 threads and considering a discrete mesh with $m = 128$, both `matlab` shows a speedup of 6.5, while `mc_sparse` have a speedup of 7.5.

C. *Conclusion*

We propose a novel stochastic method for solving time fractional partial differential equations. These equations are already being used for modeling natural phenomenon subject to memory effects, and microscopically are typically described by non-Markovian processes. Fractional equations are capable of capturing such effects due to the inherent non-locality of the operator, consequently, the classical numerical schemes often suffer from heavy memory storage requirements and high computational cost. Our stochastic method, on the other hand, compute the solution through an expected value of a functional of random processes which resembles the non-Markovian process found in the microscopic description of the phenomenon, and hence exploits somehow naturally the non-locality of the fractional operators.

II. Acknowledgment

This work was submitted to the Journal of Computational Physics. The preprint is available in ArXiv.

References

[1] R. Metzler and J. Klafter, "The random walk's guide to anomalous diffusion: A fractional dynamics approach," *Physics Reports*, vol. 339, no. 1, pp. 1–77, Dec. 2000.

[2] B. J. West, *Fractional Calculus View of Complexity: Tomorrow's Science*. Boca Raton, FL: CRC Press, Jan. 2016.

[3] A. Lischke, G. Pang, M. Gulian, F. Song, C. Glusa, X. Zheng, Z. Mao, W. Cai, M. M. Meerschaert, M. Ainsworth, and G. E. Karniadakis, "What is the fractional Laplacian? A comparative review with new results," *Journal of Computational Physics*, vol. 404, p. 109009, Mar. 2020.

[4] E. Estrada, "Fractional diffusion on the human proteome as an alternative to the multi-organ damage of SARS-CoV-2," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 8, p. 081104, Aug. 2020.

[5] R. Hilfer, *Applications of Fractional Calculus in Physics*. Singapore: World Scientific, Mar. 2000.

[6] F. Mainardi, *Fractional Calculus And Waves In Linear Viscoelasticity: An Introduction To Mathematical Models*. London: World Scientific, May 2010.

[7] J. Sabatier, O. P. Agrawal, and J. A. Tenreiro Machado, Eds., *Advances in Fractional Calculus: Theoretical Developments and Applications in Physics and Engineering*. Dordrecht: Springer, 2007.

[8] Y. Chen, I. Petras, and D. Xue, "Fractional order control - A tutorial," in *2009 American Control Conference*, Jun. 2009, pp. 1397–1411.

[9] F. Arrigo and F. Durastante, "Mittag–Leffler Functions and their Applications in Network Science," *SIAM Journal on Matrix Analysis and Applications*, vol. 42, no. 4, pp. 1581–1601, Jan. 2021.

[10] M. E. O'Neill, "PCG: A Family of Simple Fast Space-Efficient Statistically Good Algorithms for Random Number Generation," Harvey Mudd College, Claremont, CA, Tech. Rep. HMC-CS-2014-0905, Sep. 2014.

[11] R. Garrappa and M. Popolizio, "Computing the matrix Mittag-Leffler function with applications to fractional calculus," *Journal of Scientific Computing*, vol. 77, no. 1, pp. 129–153, Oct. 2018.

[12] P. I. Davies and N. J. Higham, "A Schur-Parlett Algorithm for Computing Matrix Functions," *SIAM Journal On Matrix Analysis and Applications*, vol. 25, no. 2, pp. 464–485, 2003.

[13] N. J. Higham, *Functions of Matrices*, ser. Other Titles in Applied Mathematics. Philadelphia, PA: Society for Industrial and Applied Mathematics, Jan. 2008.

Nicolas L. Guidotti received his B.Sc. in Electrical and Computer Engineering from Escola Politécnica da Universidade de São Paulo, Brazil. During his graduation, he participated in the Double Degree program, studying at Instituto Superior Técnico (IST), Portugal during 2 years as a Bologna Master student. At the end of the program, he also received M.Sc. in Computer Science and Engineering from IST. Since 2021, he is a Ph.D. student in Computer Science and Engineering at IST and an early stage researcher in INESC-ID, Portugal.

NTRU Cryptosystem
A solution to the quantum threat

Miquel Guiot Cusidó*†, Xavier Guitart Morales†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat de Barcelona, Barcelona, Spain
E-mail: miquel.guiot@bsc.es

***Keywords*—Cryptography, Lattice, NTRU, Quantum computing**

I. Extended Abstract

Most of the current public key cryptosystems' security is based on the integer factorization and the discrete logarithm, which are mathematical problems that are considered difficult to solve. However, the publication in 1994 of the Shor's algorithm [1] was a major setback for the world of cryptography, since it provided a quantum algorithm capable of solving efficiently these mathematical problems. Therefore, in recent years many governments and technology companies have put a lot of effort into developing what is known as post-quantum cryptography, i.e. the science that studies cryptosystems that are resistant to quantum attacks. Along these lines, in 2016 the National Institute of Standards & Technology launched a competition with the aim of establishing and standardising the best performing post-quantum cryptosystems, being the NTRU (*Nth Degree Truncated polynomial Ring Units*) cryptosystem one of the candidates still on the table.

The aim of this project is to study the quantum-resistant cryptosystem NTRU, both from a theoretical and practical point of view. Firstly, the mathematical results on which it is based are presented and the existing attacks against it are analyzed. Finally, a practical implementation of the cryptosystem and four of its attack is carried out, from which experimental results are extracted.

A. NTRU Cryptosystem

The origin of the NTRU public key cryptosystem dates back to 1998 when it was presented in society by Hoffstein, Pipher and Silverman [2]. As its name suggests, although it is conceived as a cryptosystem based on convolutional polynomial rings, the mathematical problem on which it is based can be formulated in terms of lattice theory, more specifically as the SVP (*Shortest Vector Problem*) [3].

Broadly speaking, the SVP simply consists in finding the vector of shortest length from a given lattice. A priori, this problem seems almost trivial, but it turns to be computationally demanding as the dimension of the lattice increases. In fact, from a complexity theory perspective, it is a well-known result that the SVP is NP-Hard under random reductions [4]. In addition, contrary to the case of integer factorization, so far there are no known quantum algorithm able to solve the SVP more efficiently than traditional classical algorithms. Therefore, this is the reason why NTRU is considered a post-quantum cryptosystem.

In that sense, despite existing several approaches, the most efficient known attacks against the NTRU cryptosystem are based on trying to solve the SVP. In particular, lattice-based attacks consist in exploiting the LLL (*Lenstra–Lenstra–Lovász*) algorithm [5], which solves an approximate version of the SVP in polynomial time. Thus, despite getting a solution in a reasonable amount of time, it may be incorrect due to the loss of accuracy. Other interesting attacks exploit some vulnerabilities when not choosing the public parameter sproperly, and the brute-force attack, which consists of simply trying all possible combinations of private keys that satisfy certain mathematical conditions [3]. Moreover, the brute-force attack can be slightly improved to what is known as meet-in-the-middle attack, which reduces the number of checks to be performed in exchange for increasing its memory demand. However, both versions have exponential time execution with respect to the degree of the polynomials used as keys. Another important aspect of the NTRU cryptosystem is its key lengths. As seen in Table I, for different security levels the key lengths remain in a reasonable size.

B. Practical Implementation

The entire code has been developed using the free software *SageMath* [6], a computational algebraic system written in Python language and built on multiple packages such as *NumPy*, *Maxima* or *R*. The triad is based on the fact that *SageMath* stands out for having a wide range of functionalities related to abstract algebra and, in particular, on the concepts related to the NTRU cryptosystem: the lattices and the convolutional polynomial rings.

The code corresponds to a script that instantiates the NTRU cryptosystem with choosen public parameters and generates text files with the derived public and private keys. It also encrypts text files using the generated public key and decrypts text files using the generated private key. Finally, it also implements the four previously discussed attacks: parameters vulnerability, brute-force, meet-in-the-middle, and lattice-based. It has been run in a Intel Core i5-7200U with 2,5 GHz, and the implementation is publicly available in GitHub: Criptosistema NTRU i atacs diversos.

C. Experimental Results

Table II shows the execution time for the decryption mechanism for different public parameters of the cryptosystem and the four attacks. The first thing to notice is that the data in the table correspond, in their entirety, to extremely small N values. This is because, as explained in Section I-A, with

TABLE I. NTRU SECURITY LEVELS & KEY LENGTHS

(N, p, q, d)	Security level (bits)	Private key length (bits)	Public key length (bits)
(163, 3, 1024, 54)	35	517	1630
(251, 3, 1024, 84)	66	796	2510
(347, 3, 2048, 116)	99	1100	3817
(439, 3, 2048, 146)	133	1392	4829
(593, 3, 2048, 247)	256	1880	6523
(745, 3, 2048, 204)	354	2362	8195

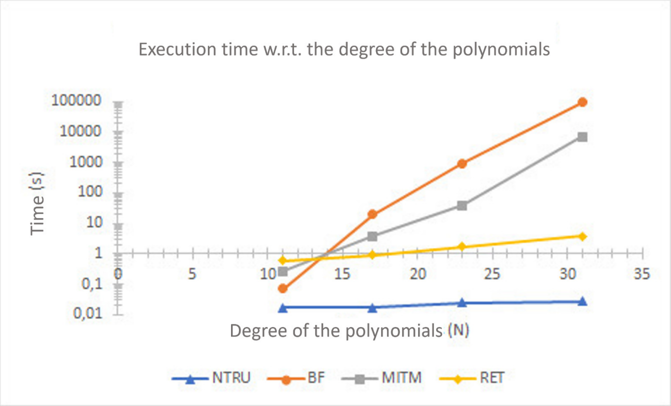


Fig. 1. Execution time w.r.t. the degree of the polyomials N.

the exception of the lattice-based attack, the time execution of the rest of the attacks grows exponentially with the size of the lattice. Therefore, the only way to be able to study their increase in execution time is starting from reduced values.

In the same direction, Figure 1 allows to visualise more clearly the growth of the execution time of each of the above decryption mechanisms. From there, analysing the values obtained it is corroborated that parameter vulnerability (when they can be carried out) and lattice-based attacks are much better in terms of performance than brute-force and meet-in-the-middle attacks. More specifically, the execution time of the parameter vulnerability attack is practically identical to that of the cryptosystem itself, regardless of the value of N. On the other hand, in the brute-force and meet-in-the-middle attacks, as expected, the decryption process increases exponentially as the degree of the polynomials increases.

Therefore, it is clear that the lattice-based attack has a much better performance than the other attacks for reduced N values. However, in practice, the NTRU cryptosystem implementations used usually have N values greater than 100, for which the lattice-based attack also fails because of the loss of accuracy mentioned in Section I-A.

D. Conclusion

From what has been seen throughout this work, it can be concluded that the NTRU cryptosystem is a viable post-quantum alternative to become the public key encryption system of reference in the coming decades. The size of its keys (no more than 1000 bytes), together with a considerable encryption and decryption speed, make it an efficient cryptosystem. Moreover, it is resistant in terms of security, since all attacks capable of breaking it, both classical and quantum, are of exponential complexity.

REFERENCES

[1] P. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, pp. 124–134.

[2] J. Hoffstein *et al.*, "Ntru: A ring-based public key cryptosystem," in *ANTS*, 1998.

[3] ——, *An introduction to Mathematical Cryptography*. Springer, 2008, ch. Lattices and Cryptography.

[4] D. Micciancio and S. Goldwasser, *Complexity of Lattice Problems: a cryptographic perspective*, ser. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, 2002, vol. 671.

[5] P. Q. Nguyen and B. Vallée, *The LLL Algorithm*, ser. Information Security and Cryptography. Springer Berlin, 2002, vol. 671.

[6] T. S. Developers, "Sagemath, the sage mathematics software system (version 9.2)," https://www.sagemath.org, 2021.

Miquel Guiot Cusidó received BSc degrees in Mathematics and Computer Science from the Universitat de Barcelona (UB). He is currently pursuing the MSc degree in Advanced Mathematics at the same university and working as a research engineer in Memory Technologies team at the Barcelona Supercomputing Center (BSC). He is collaborating with Micron Technologies US on novel memory architectures and analytical modeling.

TABLE II. EXECUTION TIME FOR DECRYPTION BY NTRU & DIFFERENT ATTACKS

(N, p, q, d)	NTRU (s)	VUL (s)	BF (s)	MITM (s)	RET (s)
(11, 8, 512, 4)	0,016	0,059	0,104	0,247	0,254
(17, 8, 512, 4)	0,018	0,034	15,571	3,494	0,712
(23, 8, 512, 8)	0,024	0,046	899,241	37,281	1,572
(31, 8, 512, 10)	0,025	0,058	90263,178	6413,558	3,331

Recapitulating experimental drug synergies in AGS through multiscale agent-based simulations

Othmane Hayoun-Mya*, Miguel Ponce-de-León*, Alfonso Valencia*[†]
*Barcelona Supercomputing Center, Barcelona, Spain
[†]ICREA, Pg. Lluís Companys, 23, 08010 Barcelona, Spain
E-mail: {ohayoun, miguel.ponce, alfonso.valencia}@bsc.es

***Keywords*—Multiscale modelling, Agent-based modelling, ABM, PhysiBoSS, Drug synergies, Gastric adenocarcinoma, Simulation-based optimization**

I. Extended Abstract

Research on the discovery of novel therapeutic strategies against tumor systems is often focused on combinatorial approaches and *in silico* testing. The former is a consequence of the well-known apparition of adaptive and acquired cancer resistance mechanisms. The latter aims to alleviate the bottleneck of using animal models or *in vitro* methodologies for drug research.

Our work aims to tackle both issues through the development of a multiscale, agent-based modelling simulation framework, that is, a computational model for high-throughput combinatorial drug assays for development of novel cancer therapies. Building on top of the work on novel drug synergies in Gastric Adenocarcinoma cells from [1], our aim was to employ PhysiBoSS [2] to develop a multiscale agent-based simulation model that recapitulates their experimental results. This will allow for further research on novel computational approaches for developing novel combinatorial therapeutic strategies for cancer.

A. Data integration into PhysiBoSS

In order to set up a computational template that reflects the experimental single and combined drug assays of [1], we developed a simulation setup that mimics the initial experimental cell disposition (a 2D monolayer of cells), and cell growth until reaching confluence (by including a contact-inhibition function) within the simulation domain. Total assay time (4200 min.) and drug injection time (1200 min.) were also replicated in our simulations. On top of this, we employed the AGS-specific Boolean Model (BM) from [1], integrated as a signalling pathway within each agent in our PhysiBoSS simulation. This model includes key regulatory elements of known cancer signalling pathways (PI3K, AKT, MEK and TAK1), which are the targets of the experimental drug synergy assays from [1]. The output of this Boolean Model is a complex combination of pro-survival and anti-survival nodes that affect growth rate and apoptosis rates of a given agent. Moreover, a Simple Diffusion transport model was employed to model the drug transport.

However, to add a more fine-grained control over the interface between microenvironment, agent and BM, we developed three transfer functions: 1) A Hill-like transfer function that computes the probability of deactivating a specific node of the Boolean network according to the internal drug concentration. This allows us to employ the exact same GI50 as in the experimental assays. 2) and 3) Hill-like transfer functions that affect growth and apoptosis rate according to the pro-survival and anti-survival readouts, respectively. On this aspect, a Hill-like function was also used to include contact-inhibition in our simulations. Based on the pressure of a given agent, its growth rate was affected. This allowed to us to reflect the confluence effect in the experimental assays of [1].

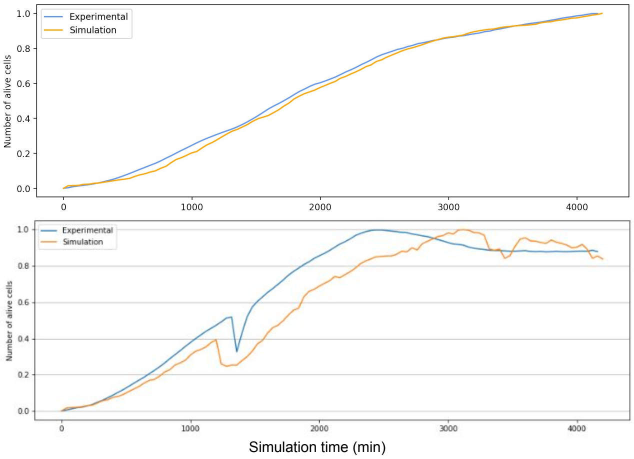


Fig. 1. **A.** Control curve fitting with PhysiBoSS, with the naïve Boolean Model from [1]. **B.** One of the best CMA results from an optimization run. Fitting of the experimental assay with the PI3K inhibitor. In both curves, the similarity was computed through the RMSE between both curves. Scaling of the curves between 0 and 1 was also needed to perform the comparison.

With the combination of these elements, we performed single and combined drug assays *in silico*, to perform a qualitative analysis of our simulations.

B. Simulation-based optimization of simulations

A step further from the qualitative results is to really reproduce the exact same experimental growth curves for the different single and combinatorial assays from [1]. Given the complexity of our simulation system shown in Section I-A, and the amount of free parameters (H value, and half-max values for all four Hill-like transfer functions and drug diffusion permeability), the most suitable approach for fitting our experimental data is by simulation-based optimization through heuristic methods. We do so with EMEWS [3], where we employ a Genetic Algorithm (GA) and Covariance Matrix Adaptation (CMA-ES) strategies in order to find the best sets

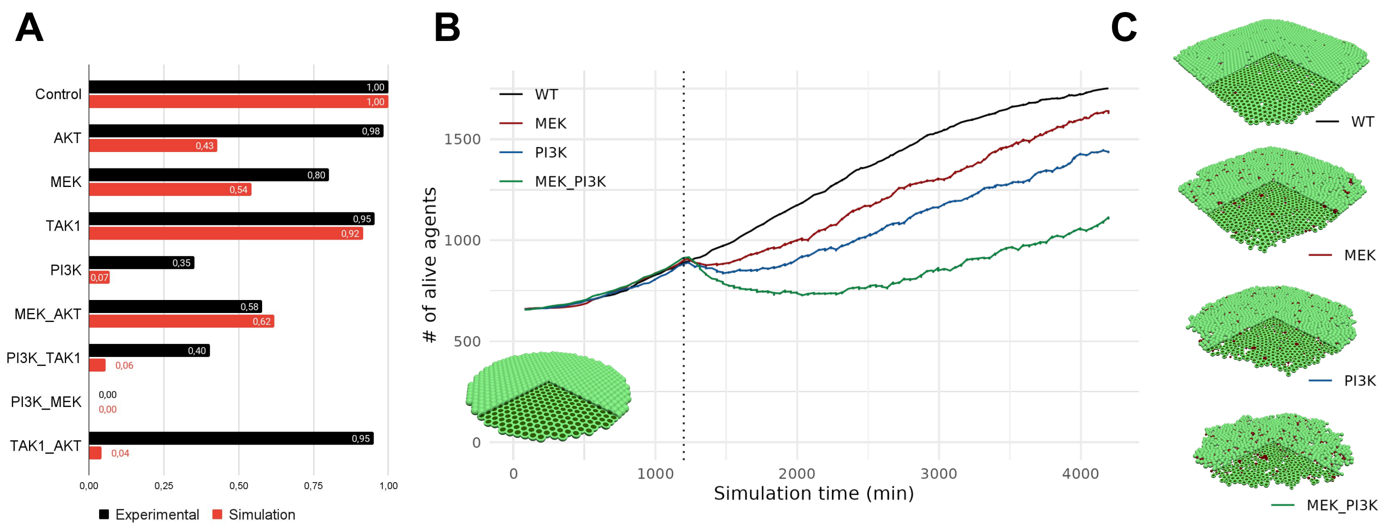


Fig. 2. **A.** Qualitative comparison of single and combined drug experiments by using the AUC of the normalized growth curves. **B.** PhysiBoSS drug synergy experiment with MEK and PI3K inhibitors. **C.** Simulation snapshot of the 2D monolayer of cells at 4200 min.

of parameters that replicate the experimental growth dynamics of [1]. Using the experimental Cell Index counts from the paper as our ground truth, our objective function is the RMSE between the simulated and experimental normalized curves (See Fig. 1).

C. Results

A first preliminary analysis of the Boolean Model already points to the correct dynamics, given that experimental drug combinations from [1] provide greater anti-survival readouts than those from single drugs. An exception was found for the TAK1-targeting drug, which presented no difference regarding the Control readouts, with a fully pro-survival steady state.

The fitting of the experimental control growth curve was performed without the need for a simulation-based optimization, and just by modification of the parameters of the contact-inhibition function manually (See Fig. 1). Once this fitting was accomplished, we further performed the same battery of single and combined drug assays as in [1]. The qualitative results from our simulations correctly reflect the experimental synergies observed (Fig. 2).

The simulation-based optimization workflow employed to fit the experimental drug curves provided few relevant results (Fig. 1). We are currently improving the parameters employed for optimization, along with the evaluation metric employed, and improving the parameter range to be explored by incorporating knowledge regarding the physicochemical properties of the drugs employed. However, the results obtained so far do indicate that this is the right direction.

D. Conclusion

The present work aims to push forward the development of *in silico* tools that allow for biologically realistic high-throughput development and testing of novel combinatorial therapeutic strategies against cancer in a personalized manner. By integrating many different sources of data at different biological and mechanistic levels, our model can qualitatively replicate experimental drug synergies within the AGS cell line. Furthermore, we are currently working on the fitting of different single drug experiments in order to obtain, in an emergent manner, the experimentally-shown drug synergies.

With this, it is a useful, personalizable template that can be used for testing different drugs in different cell-types, as well as setting an ideal tool for researching the apparition of acquired drug resistance through simulating heterogeneous cell populations. For this, we believe it will further advancing research on novel cancer therapies and the much-needed overcoming multi-drug resistance.

II. Acknowledgment

This work bas been supported by the PerMedCoE project (grant agreement Nº951773).

References

[1] Flobak *et al.*, "Discovery of drug synergies in gastric cancer cells predicted by logical modeling," *PLOS Computational Biology*, vol. 11, no. 8, pp. 1–20, 08 2015. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1004426

[2] G. Letort *et al.*, "PhysiBoSS: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling," *Bioinformatics*, vol. 35, no. 7, pp. 1188–1196, 08 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty766

[3] J. Ozik *et al.*, "From desktop to large-scale model exploration with swift/t," in *2016 Winter Simulation Conference (WSC)*, 2016, pp. 206–220.

Othmane Hayoun received his BSc degree in Biotechnology from Universitat de València, Spain in 2019 and his MSc in Bioinformatics for Health Sciences at Universitat Pompeu Fabra, Spain in 2022. His MSc Thesis was focused on the implementation of transport mechanisms within a multiscale agent-based modeling software, PhysiCell. He is now currently working on this same research line within the Computational Biology group at the Life Sciences Department of the Barcelona Supercomputing Center (BSC-CNS).

Characterizing the Impact of Graph-Processing Workloads on Modern CPU's Cache Hierarchy

Alexandre Valentin Jamet*†, Lluc Alvarez*†, Marc Casas*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {alexandre.jamet, lluc.alvarez, marc.casas}@bsc.es

***Keywords*—cache management, cache bypassing, big data, graph processing, workload evaluation, irregular workloads, micro-architecture**

I. Extended Abstract

In recent years, graph-processing has become an essential class of workloads with applications in a rapidly growing number of fields. Graph-processing typically uses large input sets, often in multi-gigabyte scale, and data-dependent graph traversal methods exhibiting irregular memory access patterns. Recent work [1] demonstrate that, due to the highly irregular memory access patterns of data-dependent graph traversals, state-of-the-art graph-processing workloads spend up to 80 % of the total execution time waiting for memory accesses to be served by the DRAM. The vast disparity between the Last Level Cache (LLC) and main memory latencies is a problem that has been addressed for years in computer architecture. One of the prevailing approaches when it comes to mitigating this performance gap between modern CPUs and DRAM is cache replacement policies.

In this work, we characterize the challenges drawn by graph-processing workloads and evaluate the most relevant cache replacement policies.

A. Graph-processing Workloads

Graph-processing is a class of emerging workloads that, nowadays, can be found in various applications. Graph-processing can be found in both industry and academia, from social network analytics to web search engines and biomedical applications.

Graph-processing typically uses sparse data formats such as *Compressed Sparse Row/Column (CSR/CSC)* to manage a large amount of data. The CSR/CSC format is used to encode the graph adjacency matrix using two data structures: 1) the *Offset Array (OA)* and; 2) the *Neighbours Array (NA)*. Finally, additional data structures contain numerical data corresponding to a graph's vertices called *Property Arrays (PA)*.

Manipulating these sparse data structures often produces irregular memory access patterns. For example, when computing the *Sparse Matrix-Vector (SpMV)* multiplication $y = A \cdot x$, accesses to vector x are indexed by the column indices of matrix A, which are non-contiguous and constitute an irregular access stream. Graph-processing workloads also display highly irregular memory access patterns driven by operations like graph traversals that require visiting all the vertices V of a graph, that is, scanning the adjacency matrix rows following the graph's connectivity.

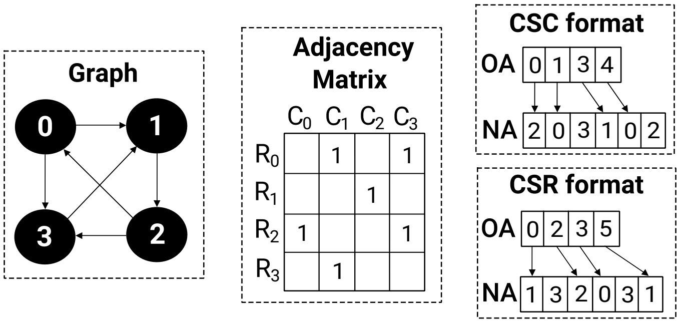


Fig. 1. Example of a graph representation in memory using the CSR/CSC formats.

Figure 1 shows an example graph along with its adjacency matrix and its representation using either the CSR or the CSC formats.

B. Cache Replacement Policies

For this work, we evaluate six of the most relevant replacement policies for the LLC. SRRIP, DRRIP [2] hahandleshe replacement process by predicting reuse distances. SHiP [3] leverages SRRIP and extends it with the addition of a PC feature. Hawkeye [4], Glider [5] a,nd MPPPB [6] further improve by making the addition of machine learning inspired techniques (*e.g.*: perceptron, SVM, etc.). Specifically, these advanced cache replacement policies leverage micro-architectural features based on bits extracted from the program counters and virtual/physical addresses to establish correlations and produce predictions.

C. Experimental Setup

Our evaluation considers ChampSim, a detailed trace-based simulator that models a Cascade Lake micro-architecture. The micro-architecture simulated has only one core, L1 instruction, and data cache of 32KB each, an L2 cache of 1MB, and an L3 cache of 1.375MB. The system also includes an 8GB main memory based on DDR4 SDRAM with a data rate of 2.933GT/s.

D. Experimental results

Figure 2 shows the MPKI rates in all three cache hierarchy levels for workloads of the GAP [7] benchmark suite. This figure shows that graph-processing workloads suffer from a large number of misses at all levels of the cache hierarchy.

The average MPKI rates of these workloads in the L1D, L2C, and LLC are 53.2, 44.2 and 41.8. We can further observe that a considerable portion (78.6%) of the accesses that trigger L1D misses also miss in the lower levels of the cache hierarchy and require a DRAM access.

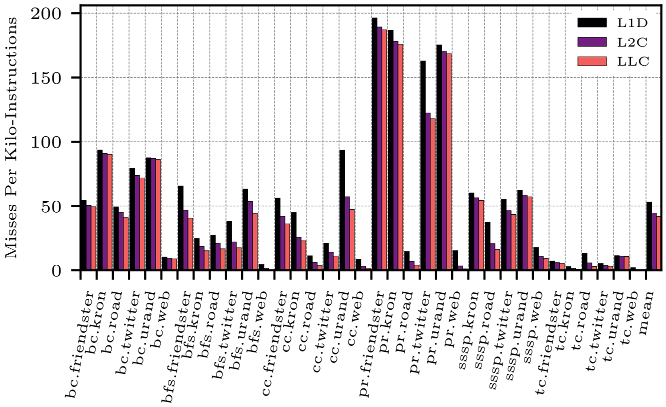


Fig. 2. Misses-Per-Kilo-Instruction (MPKI) across the different levels of the cache hierarchy triggered by graph-processing workloads.

Cache replacement policy is an obvious topic when it comes to improving the behavior of the cache hierarchy for certain types of workloads. As such, we evaluated the replacement policies presented in I-B to understand their impact on performance.

Figure 3 shows the geometric mean speed-up of the state-of-the-art cache replacement policies evaluated over the baseline LRU policy for various benchmark suites, including SPEC 2006 & 2017, along with the GAP workloads. The results show that the different policies can catch different kinds of access patterns and benefit different workloads.

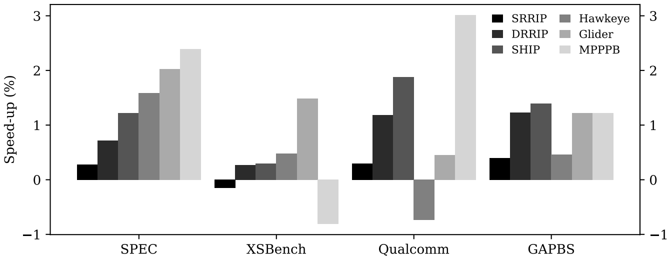


Fig. 3. Geometric mean speed-up over LRU of state-of-the-art LLC replacement policies for the different benchmark suites.

These results show that more complex replacement policies such as Hawkeye, Glider, and MPPPB have difficulties generalizing to benchmark suites beyond SPEC 2006 & 2017. This is due to the underlying assumptions on memory access patterns used to build these complex cache replacement policies. As shown in I-A, graph-processing is a prime example where the number of PC is very limited and where each PC maps to a very large number of addresses making correlations nearly impossible to establish.

E. Conclusion

Overall, this work highlights the poor ability of state-of-the-art cache replacement policies to leverage significant benefits against a baseline using an LRU policy for graph-processing workloads, despite the very high hardware complexity of such techniques. We show pieces of evidence that this bleak outlook stems from two factors: i) the very distinct nature of graph-processing workloads and; ii) the immense pressure these workloads create on the cache hierarchy.

II. Acknowledgment

This work has been published in Proceedings of the International Symposium on Workload Characterization (IISWC), 2020 [8].

References

[1] Y. Zhang, V. Kiriansky, C. Mendis, S. Amarasinghe, and M. Zaharia, "Making caches work for graph analytics," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 293–302.

[2] A. Jaleel, K. B. Theobald, S. C. S. Jr, and J. Emer, "High performance cache replacement using re-reference interval prediction (RRIP)," in *Proceedings of the 37th annual international symposium on Computer architecture*, vol. ISCA'10. IEEE, pp. 60–71. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1815961.1815971

[3] C.-J. Wu, A. Jaleel, W. Hasenplaugh, M. Martonosi, S. C. Steely, and J. Emer, "SHiP: signature-based hit predictor for high performance caching," in *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture - MICRO-44 '11*. ACM Press, p. 430. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2155620.2155671

[4] A. Jain and C. Lin, "Back to the future: Leveraging belady's algorithm for improved cache replacement," in *Proceedings of the 43rd International Symposium on Computer Architecture*, ser. ISCA '16. IEEE, pp. 78–89. [Online]. Available: https://dl.acm.org/citation.cfm?doid=3007787.3001146

[5] Z. Shi, X. Huang, A. Jain, and C. Lin, "Applying deep learning to the cache replacement problem," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, pp. 413–425. [Online]. Available: https://dl.acm.org/doi/10.1145/3352460.3358319

[6] D. A. Jiménez and E. Teran, "Multiperspective reuse prediction," in *Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture - MICRO-50 '17*, ser. MICRO '17. IEEE, pp. 436–448. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3123939.3123942

[7] S. Beamer, K. Asanović, and D. Patterson, "The GAP benchmark suite." [Online]. Available: http://arxiv.org/abs/1508.03619

[8] A. V. Jamet, L. Alvarez, D. A. Jiménez, and M. Casas, "Characterizing the impact of last-level cache replacement policies on big-data workloads," in *2020 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 134–144. [Online]. Available: https://upcommons.upc.edu/bitstream/handle/2117/343622/IISWC20-paper.pdf?sequence=1

Alexandre Valentin Jamet studied two years of Higher School Preparatory Classes with a Physics and Engineering Sciences major at LGT Baimbridge, Guadeloupe. In the following years, he pursued his MSc degree in parallel with an Engineer Diploma from TELECOM Nancy with a major in Embedded Computing. He concluded his studies in Nancy in 2018. Since 2018, he has been a Ph.D. candidate at the Computer Architecture departments of Barcelona Supercomputing Center (BSC) and Universitat Politècnica de Catalunya (UPC), Spain.

Consensus Essential Dynamics Analysis: Application to Biomolecular Simulations

Luis Jordà[#1], Josep Lluís Gelpí[#*2]

[#]Barcelona Supercomputing Center (BSC)

[*]University of Barcelona (UB), Department of Biochemistry and Molecular Biology, 08028-Barcelona, Spain

[1]luis.jorda@bsc.es, [2]gelpi@ub.edu

***Keywords* — Molecular dynamics, essential dynamics, proteins, pathogenicity prediction, pyruvate kinase.**

I. Extended Abstract

One of the main focuses of our current research revolves around the development of an analytic approach that can detect and characterize the similarities and differences in the dynamic behavior of conformational ensembles of macromolecules. We present here Consensus Essential Dynamics Analysis (CEDA), a protocol that integrates the information from Principal Components Analysis (PCA) applied independently to several equivalent molecular dynamics (MD) simulations of a system and derives a consensus set of vectors that enable trajectory comparison in a common framework. The protocol has been tested with a collection of MD simulations of the human erythrocyte pyruvate kinase (PKR) that covers multiple biological conditions. The obtained results are discussed in the light of the potential application of this protocol in functional studies of proteins in general, and with a particular perspective on pathogenicity prediction studies.

A. Introduction

Molecular dynamics (MD) simulations are commonly referred to as a veritable “computational microscope”, capable of valuably complementing many experimental methodologies and facilitating discovery in spatial and temporal domains that would otherwise be inaccessible [1]. With MD we can simulate the motion of a given atomic system and store such information in a collection of snapshots, called a trajectory. Nowadays, this methodology remains the prevalent choice among computational techniques when it comes to deciphering functional traits of biomacromolecules like proteins or nucleic acids [2]. MD simulations are especially useful to identify flexibility patterns and monitor conformational changes of macromolecules. These events are at the basis of virtually every biological process, such as metabolic reactions, nutrient transport, communication and signaling pathways, and immunologic response.

As computational power and infrastructures keep improving, we are increasingly able to generate longer MD simulations that allow capturing dynamic events at biologically relevant timescales. MD trajectories typically generate an overwhelming amount of data. Thus, we need to find suitable metrics to extract, quantify and present the relevant information depending on the target of the study. The scenario is even more challenging when we aim to analyze multiple trajectories and compare their similarity. Among the proposed strategies to explore the comparability between trajectories, essential dynamics analysis (EDA) approaches are a common framework, where Principal Components Analysis (PCA) or other dimensionality reduction techniques are applied to express the differential behavior between trajectories in terms of the underlying collective features of the ensemble. For instance, in the strategy known as combined-PCA, the involved trajectories are concatenated into a “multitrajectory”, enabling the extraction of the features of the average dynamic behavior of the ensemble. Even if this technique provides a single reference property space for the whole ensemble of trajectories, its interpretability is not straightforward and can lead to biased conclusions [3].

Overcoming these obstacles is bound to have a crucial impact on clinical assistance, personalized medicine and health research in general. Recently, Galano-Frutos *et al*. [4] provided insightful considerations for pursuing the goal of interpreting human genetic variations at large scale through dynamic data (and specifically, with MD simulations). By exploring mutations at the protein level in their structural context, it is possible to estimate their functional impact in terms of the expected structural alterations. However, structural and dynamic features have been largely neglected in the field of pathogenicity prediction, mainly due to the high computational cost of MD and the lack of robust analytic metrics. Still, through the years, a large volume of independent studies achieved meaningful results with MD-based approaches and dynamic data as the source for predicting the implications of mutations [5].

B. Protocol overview

The Consensus Essential Dynamics Analysis (CEDA) strategy consists in deriving a set of consensus eigenvectors by applying a clustering algorithm on the most relevant eigenvectors obtained from the traditional EDA of the available trajectories. Consensus eigenvectors: i) maintain the most relevant collective fluctuations (*i.e.*, the fraction of correlation that is shared between the members of the cluster), and ii) filter out the minor collective fluctuations (*i.e.*, the fraction of correlation that is present only in particular members of the cluster). The analysis of the projections of the involved trajectories to the single consensus set allows the comparison of the different simulation conditions in a consistent way. In summary, the full protocol consists of the following main steps:

1. Generate MD trajectories belonging to the reference and target conditions of the macromolecular system.
2. Get the average structure of the reference condition trajectories.
3. Perform (Cartesian) PCA of each reference condition trajectory.
4. Perform a clustering of the resulting eigenvectors. Use the cosine distance (expressed in absolute value) to build the pairwise distance matrix between vectors.
5. Get the consensus eigenvector (centroid) from each relevant cluster.
6. Perform a least-squares fitting of all trajectories to the structure of step 2 and center trajectory data around it.
7. Examine the projection of the individual trajectories obtained in step 6 onto the desired consensus eigenvectors. Explore the concerted motions captured in each consensus eigenvector by translating back to Cartesian coordinates the projection values.

8. Compare the projection data density distributions within and between conditions. Identify relevant values of the data distributions and characterize the distinctive structural conformations of each condition.

C. Preliminary results

The analysis of the projections of the individual trajectories to the single consensus set allows the comparison of the different simulation conditions in a consistent way. In this report we include an illustrative example of our results to show the effectiveness and usefulness of the method.

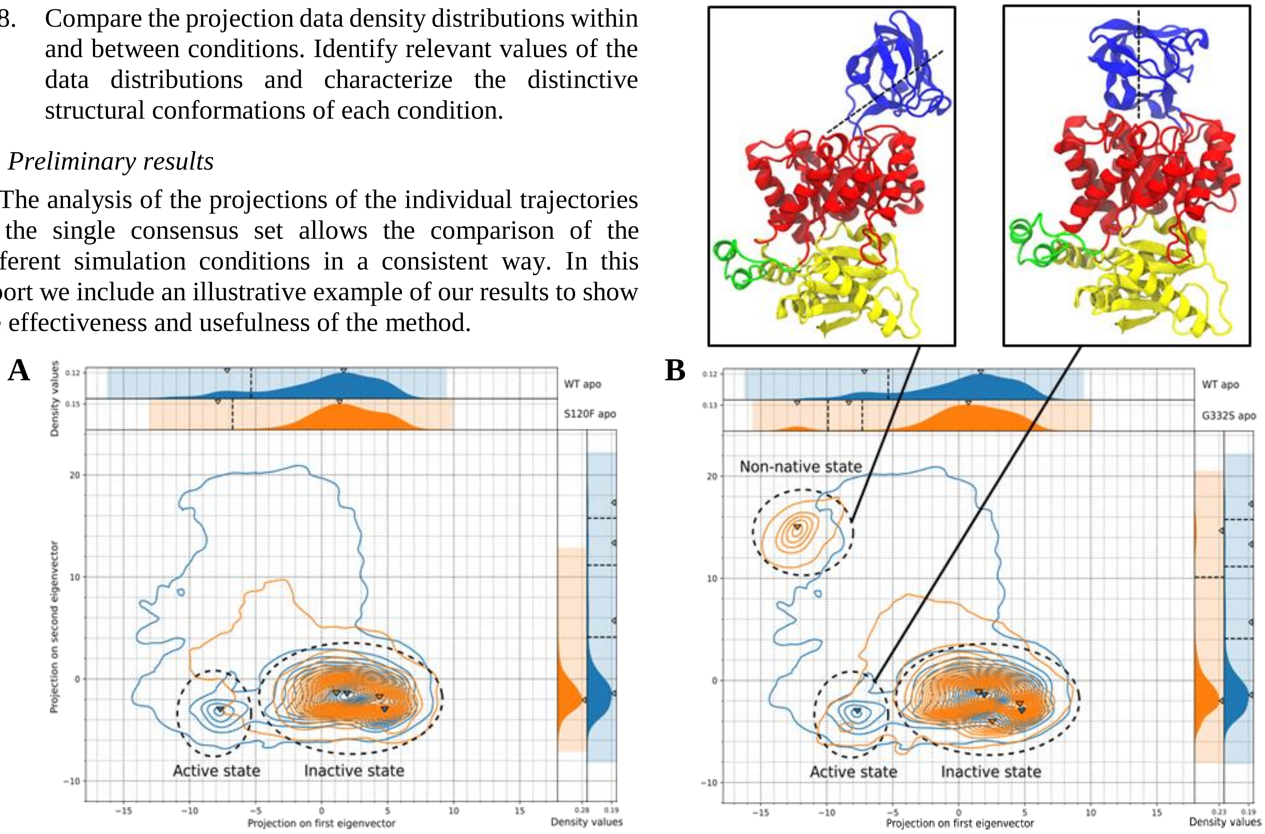


Fig. 1. Comparison of populations obtained from the trajectory projections onto two major consensus eigenvectors of the apo simulations of WT and selected mutants. A) WT vs. S120F, B) WT vs. G332S. The motion captured along the eigenvectors corresponds to the opening and the shaking of domain B of the protein. Even when in the apo form, the WT protein shows a significant population in the active conformation, supporting the concept of conformational selection. S120F is unable to explore such conformation, while G233S shows an additional non-native conformation. The protein figures were generated with the VMD software, in cartoon representation and with domains colored as follows: N-t (green), A (red), B (blue) and C (yellow).

The chosen example involves the analysis of domains A+B of the PKR monomer, where relevant native and differential dynamical behaviors are elucidated.

Figure 1 shows the comparison of system states as obtained from the analysis of projections, for the WT and two well-known pathological mutants (S120F and G332S). Our analysis reveals the inability of S120F (Figure 1A) to sample the active state conformation of the enzyme to the same extent as WT, thus providing initial evidence for the pathogenicity of the variant. The hydroxyl group of S120 is one of the chemical groups responsible for binding the cofactor K^+. Therefore, in addition to a possibly altered cofactor binding, also a significant alteration on the allosteric dynamics of the protein contributes to the deleterious effect. The analysis of G332S (Figure 1B) shows that, in addition to the fact that the active state is again barely sampled in comparison with the WT protein, a new non-native conformation is sampled.

II. Acknowledgments

We thankfully acknowledge the computer resources at MareNostrum and the technical support provided by Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS) and Red Española de Supercomputación (RES) to cope with the computational requirements of our project (RES-BCV-2019-1-0004, BCV-2019-3-0006, BCV-2020-3-0005, and BCV-2021-3-0002).

References

[1] R.O. Dror *et al*., "Biomolecular simulation: A computational microscope for molecular biology", Annual Review of Biophysics, 41(1), 429–452, 2012.

[2] A. Hospital *et al.*, "Surviving the deluge of biosimulation data", Wiley Interdiscip. Rev. Comput. Mol. Sci., 10(3), 2020.

[3] G. Pierdominici-Sottile *et al.*, "New insights into the meaning and usefulness of principal component analysis of concatenated trajectories", J. Comput. Chem., 36(7), 424–432, 2015.

[4] J.J. Galano-Frutos *et al*., "Molecular dynamics simulations for genetic interpretation in protein coding regions: where we are, where to go and when", Brief. Bioinform. 22(1), 3–19, 2021.

[5] V.E. Angarica *et al*., "Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: Linking SNPs with disease phenotypes in familial hypercholesterolemia", Hum. Mol. Genet. 25(6), 1233–1246, 2016.

Luis Jordà was born in Barcelona, Spain, in 1993. He received the BSc in Biochemistry from the University of Barcelona (UB), in 2016, and later the MSc in Bioinformatics for Health Sciences, from the University Pompeu Fabra (UPF), in 2018, in Barcelona, Spain.

He is currently pursuing his PhD in Biomedicine (UB) and performing his research at the facilities of the Barcelona Supercomputing Center (BSC), under the supervision of Prof. Josep Lluís Gelpí. He is also an active collaborator of the Molecular Modeling and Bioinformatics (MMB) group from the Institute for Research in Biomedicine (IRB) of Barcelona. His current research interests include structural bioinformatics, protein dynamics and modeling of biomolecules.

Microbiome profiling from Fecal Immunochemical Test reveals microbial signatures with potential for Colorectal Cancer screening

Olfat Khannous-Lleiffe[1,2], Toni Gabaldón[1-3*]

[1] *Life Sciences Department , Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain*

[2] *Mechanisms of Disease, Institute for Research in Biomedicine (IRB), Barcelona, Spain*

[3] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*

[1]olfat.khannous@bsc.es, [3]toni.gabaldon@bsc.es[*]

***Keywords*— Colorectal cancer; Microbiome; 16S rRNA sequencing; Screening; Diagnosis;**

Extended Abstract

Colorectal cancer (CRC) is a global healthcare challenge that involves both genetic and environmental factors. Early diagnosis of CRC, which saves lives and enables better outcomes, is generally implemented through a two-step population screening approach based on the use of Fecal Immunochemical Test (FIT) followed by colonoscopy if the test is positive. However, the FIT step has a high false positive rate, and there is a need for new predictive biomarkers to better prioritize cases for colonoscopy. Here we used 16S rRNA metabarcoding from FIT positive samples to uncover microbial taxa, taxon co-occurrence and metabolic features significantly associated with different colonoscopy outcomes, underscoring a predictive potential and revealing changes along the path from healthy tissue to carcinoma. Finally, we used machine learning to develop a two-phase classifier which reduces the current false positive rate while maximizing the inclusion of CRC and clinically relevant samples.

A. Introduction

Colorectal cancer (CRC) is the third most common cancer type and the second leading cause of cancer-related deaths worldwide. This malignant disease develops from the pathological transformation of normal colonic epithelium to adenomatous polyps, which ultimately leads to invasive cancer.

Several pieces of evidence suggest that alterations in the gut microbiota, microorganisms that are inhabiting the gastrointestinal tract, may influence colon tumorigenesis through chronic inflammation or the production of carcinogenic compounds [1].

Diagnosis of CRC is challenging and involves a complex process that usually starts with the detection of the first symptoms by the patient and it is followed by clinical diagnostic procedures, mainly based on colonoscopy. In Catalonia, as implemented in many regions, we have a pre-colonoscopy screening strategy which is based on Fecal Immunochemical test (FIT) that signals if there is hidden blood in the stool, followed by colonoscopy if the test is positive. However, this test has a high false positive rate (around 65%) and there is a need to speed the inclusion of clinically relevant cases to colonoscopy, to have early diagnosis and better prognosis.

B. Materials and Methods

A total of 2,889 FIT-positive (> 20 µg hemoglobin/g feces) samples recruited from asymptomatic participants from the Catalan CRC screening program were analysed. Metadata comprised six different clinical variables for each sample, including diagnosis after colonoscopy evaluation, the number of polyps, the FIT value (µg of hemoglobin/g of feces), the hospital, age and sex. The considered diagnosis groups were: Clinically relevant (CR) including Colorectal cancer (CRC), Carcinoma in situ (CIS), High risk lesion (HRL) and Intermediate risk lesion (IRL) and Non-Clinically relevant: Low risk lesion (LRL), Lesion not associated to risk (LNAR) and Negative (N) exploration.

The region V3-V4 was amplified from extracted DNA and sequenced by Illumina MiSeq. Dada2 pipeline was used to pre-process the data and taxonomy was assigned by using SILVA database. Both alpha diversity (within samples) and beta diversity (between samples) metrics were calculated. Normalization was performed by transforming counts to centered log-ratios (clr) and taxa that appeared in fewer than 10 samples and at low abundances (fewer than 100 reads) and samples with fewer than 1000 reads were filtered out.

We assessed the effect of different clinical variables on the overall microbiome composition by performing Permutational Multivariate Analysis of Variance (PERMANOVA) using the adonis function from the Vegan R package and performed differential abundance analysis using clr data considering the different taxonomic ranks across different clinical variables using linear mixed effects models, implemented in the R package lme4. We also studied co-occurrence networks specific per diagnosis and predicted the functional content in terms of functional orthologous groups.

We developed a predictive model based on a two-phase classification using a neural network algorithm implemented in the caret package. For each phase we trained a random 75% of the data with a 10-fold cross validation and tested with the remaining samples. The process was repeated 100 times to avoid “lucky” splits and to evaluate the variability in predictive performance. We performed a feature selection based on the differential abundance results including taxa found as having significantly different abundances in our study and incorporating FIT-value, age and sex variables. Samples with missing values for the considered metadata were removed. Taxa abundances were included as clr. The two-phase classifier proceeds as follows: in the first phase the method classifies CRC vs non-CRC samples. Samples that are classified as non-CRC in the first phase are subjected to a second model that classifies CR vs non-CR samples. At the end of the two-phase classification, the mean percentage of misclassified CRC and CR samples was calculated and the performance of the model was evaluated.

To validate our strategy we built a model training with all the CRIPREV samples and tested it in two independent datasets.

C. RESULTS AND DISCUSSION

We performed DNA extraction and 16S metabarcoding analysis of the V3-V4 region on the selected samples. A total of 2,889 FIT-positive samples passed all quality filters and were included in the study. We obtained a mean value of 56,219.03 filtered reads per sample, which comprised a total of 376 assigned taxa. Bacteroidetes and Firmicutes were the most represented phyla, and the ten most abundant genera were, in this order: *Bacteroides, Faecalibacterium, Prevotella, Blautia, F.Lachnospiraceae.UCG, Ruminococcus, Agathobacter, Bifidobacterium, Alistipes* and *Akkermansia* . These results are consistent with previous studies using stool samples, and with earlier analyses showing a high correspondence between stool and FIT samples from the same individuals. We compared our data with that of a recent Spanish population gut microbiome study. The two cohorts differ in several features such as the age range, but most notably our cohort was entirely formed by individuals with blood in stool, a factor shown to impact the gut microbiome, and hence differences are expected. Nevertheless, the two sample sets were largely similar in terms of dominating phyla and genera, reinforcing the validity of FIT sampling as a proxy of the gut microbiome.

We quantified the overall microbiome diversity by computing alpha and beta diversity metrics. We observed differences in both Observed index and Simpson index according to the Diagnosis. Representing the Aitchison distance in a multidimensional plot we did not observe clear clustering of the samples according to the diagnosis but we detected a significant but subtle effect of the variable by the adonis test (P=0.001) considering as covariates the sex and the age, and the sequencing run as a possible source of batch effect.

Our differential analysis detected 41 taxa as differentially abundant according to CRC vs others and 34 comparing CR vs Non-CR. Taking profit on the observed differences, we developed a two-phase machine learning classifier with high sensitivity for CRC (98.98%) and CR samples (97.78%) [2].

D. ACKNOWLEDGEMENTS

We are thankful to all CRC-screening participants that provided consent for the use of their samples in this study. O.K-L was funded by the “Formación de profesorado universitario (FPU)” program from the Spanish Ministerio de Universidades (FPU 2020-02907).

References

[1] Saus E et al. Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential. Mol Aspects Med. 2019;69:93–106.

[2] Khannous-Lleiffe O et al. on behalf of the CRIPREV Consortium. Microbiome Profiling from Fecal Immunochemical Test Reveals Microbial Signatures with Potential for Colorectal Cancer Screening. Cancers. 2023; 15(1):120. https://doi.org/10.3390/cancers15010120

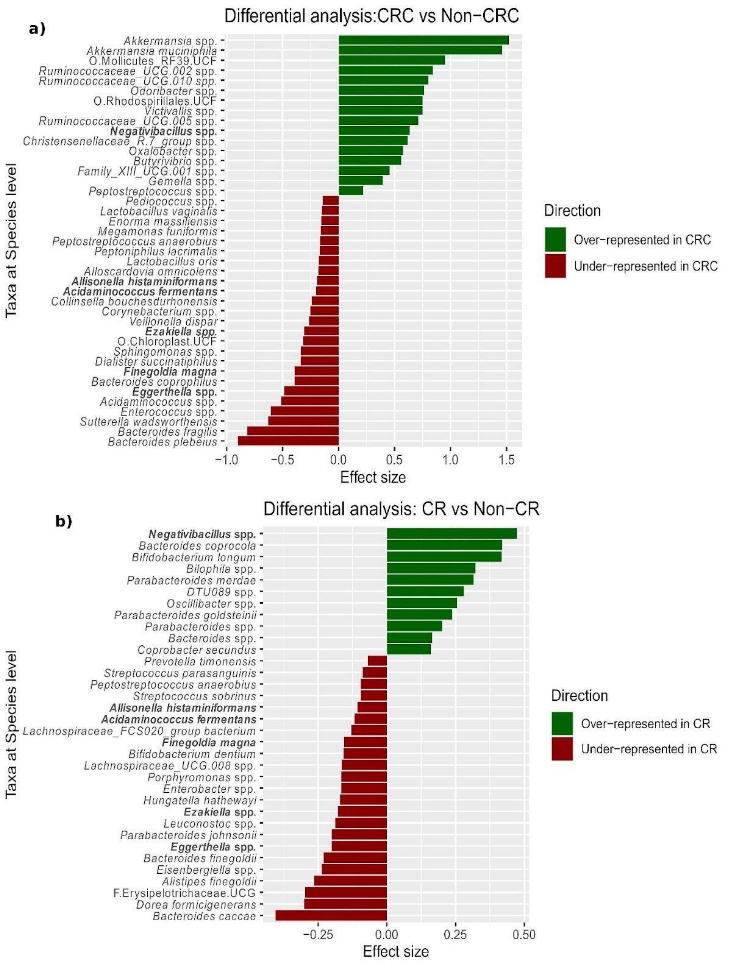


Fig. 1 Effect size of species found as significantly differentially abundant when comparing A) CRC vs Non-CRC samples and B) CR vs Non-CR samples. Bars are green for overrepresentation and red for underrepresentation. The bars are sorted according to the effect size. In bold are highlighted the taxa that appeared as differentially abundant in both comparisons.

Author biography

Olfat Khannous Lleiffe was born in Vic, Spain, in 1996. She received the BSc in Biotechnology from University of Vic (UVic) in 2018 and the MSc in Bioinformatics for Health Sciences from University Pompeu Fabra (UPF) in 2020.

She is currently doing a PhD in Biomedicine at Toni Gabaldon’s Comparative Genomics group with the “Formación de profesorado universitario (FPU)” fellowship from the Spanish ministry of universities. She is also a teaching assistant at the University of Barcelona (UB).

Open-Source GEMM Hardware Kernels Generator: Toward Numerically-Tailored Computations

Louis Ledoux*†, Marc Casas*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {louis.ledoux, marc.casas}@bsc.es

***Keywords*—GEMMs, matrix-matrix-multiply, full stack framework, automated pipeline, flopoco, OpenCAPI, OpenBLAS, High Performance Computing, approximate/trans/extended precision.**

I. Extended Abstract

Many scientific computing problems can be reduced to Matrix-Matrix Multiplications (MMM), making the General Matrix Multiply (GEMM) kernels in the Basic Linear Algebra Subroutine (BLAS) of interest to the high-performance computing community. However, these workloads have a wide range of numerical requirements. Ill-conditioned linear systems require high-precision arithmetic to ensure correct and reproducible results [1]. In contrast, emerging workloads such as deep neural networks, which can have millions up to billions of parameters, have shown resilience to arithmetic tinkering [2] and precision lowering [3].

General purpose arithmetic units and computer formats such as the IEEE754 standard naturally underperform in this vaste land of scenarios. We propose the generation of numerically tailored circuits where the necessary and sufficient internal precision is generated to target the computations requirements in terms of numerical quality while improving the energy cost.

A. *Open Source SW/HW co-designed framework for numerically tailored MMMs*

As depicted by Fig. 1, our framework is composed of two distinct phases, the prior Hardware generation flow and the runtime execution flow. Because MMMs are basically made of arbitrary long dot products, we design a custom Fused Dot Product (FDP) operator that is agnostic to the computer format and supports posit, IEEE754, and bfloat16 variations, while never rounding between two accumulations. The intermediate precision of the fixed-point accumulator used in the dot-product is a key aspect of this work, and is configurable through the length of the scratchpad delimited by the parameters MSB (Most Significant Bit) and LSB (Least Significant Bit). We leverage the automated pipeline feature of *flopoco* [4] which is an effective tool for efficiently exploring the wide range of functional specifications along with performance specifications to produce MMM kernels with the necessary basic elements (LUTs, FFs, Carry chains, DSPs) for a targeted $(chip, frequency)$ couple (see Fig. 1- Ⓑ).

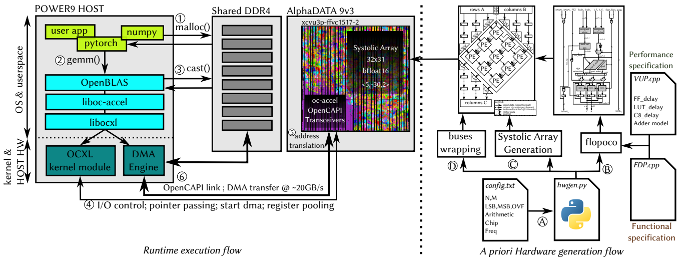


Fig. 1. Overview of the 2 phases framework. Left is Runtime execution flow and right is Hardware generation flow.

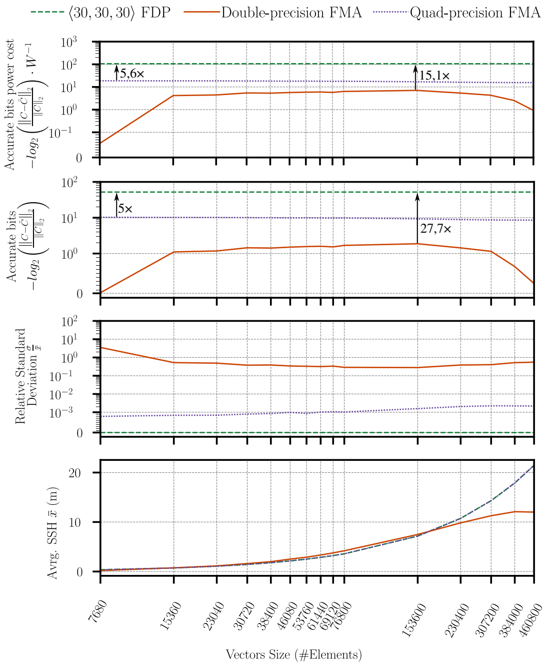


Fig. 2. Sea Surface Height computation comparing IEEE-754 double-, quad- pecision FMAs and a 91-bit FDP wrt numerical quality and power consumption.

The essence of this work is to make intermediate precision tweakings from the hardware accessible to high-end software code as transparent as possible. We achieve that by taking into account that many HPC codes rely on BLAS libraries to perform MMM operations. Such libraries receive the function call to perform a GEMM and dispatch adequately to the underlying hardware at their disposal.

B. *HPC workloads results*

We experiment with two families of real HPC workloads with contrasting numerical requirements, namely Artificial Intelligence (AI) and Sea Surface Height (SSH), whose respective results can be observed in Fig. 3 and Fig. 2.

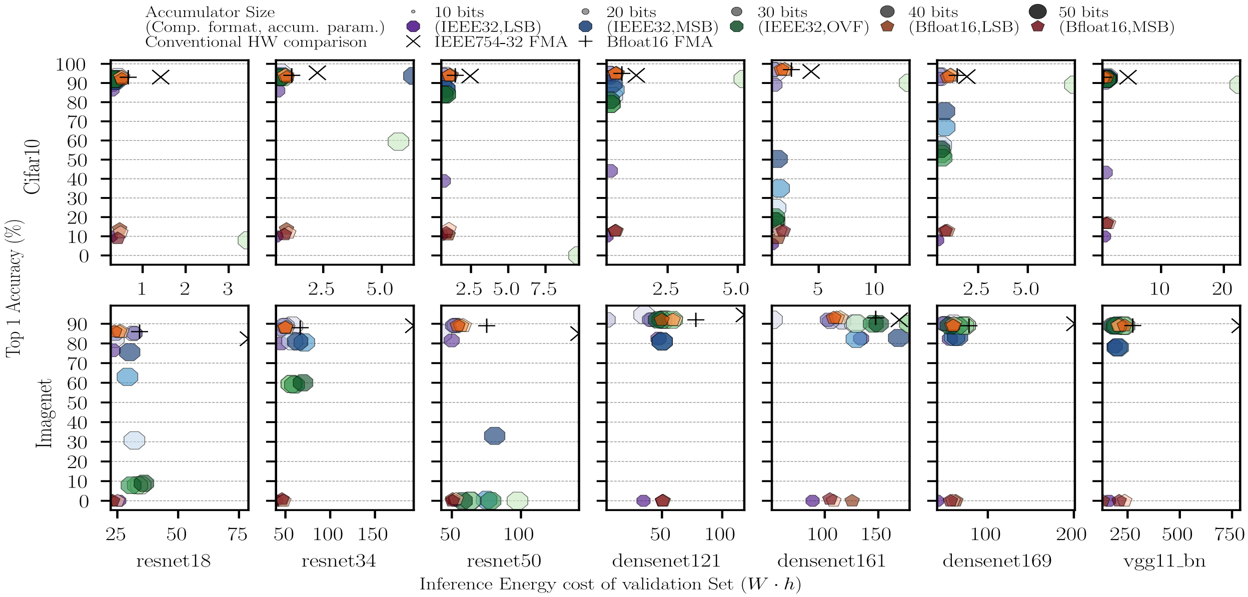


Fig. 3. Top1 Accuracy vs validation dataset inference Energy cost for various combinations of datasets,models,computer formats, and accumulators.

For the SSH computation, the results obtained with 64-bit and 128-bit FPUs exhibit decreasing reproducibility as the vector size increases. In contrast, our 91-bit $\langle ovf:30, msb:30, lsb:30\rangle$ FDP maintains reproducibility for all vector sizes without deviation. Our proposed FDP consistently exhibits 52 correct bits, which is at least 5× and 27.7× more than quad-precision and double-precision. Our measurements on VU3P-2 FPGA at $200MHz$ show that the units power consumption are 0.266, 0.549, and 0.491 watts for double-precision FMA, quad-precision FMA, and the 91-bit FDP, respectively. For all evaluated sizes, the 91-bit FDP yields at least 5.6× and 15.1× more correct bits for the same wattage as quad-precision and double-precision FMAs, respectively.

For AI workloads, we employ Pytorch as a base framework and link it to our modified OpenBLAS. We use popular neural network models such as ResNet18, ResNet34, ResNet50, DenseNet121, DenseNet161, DenseNet169, and VGG11 with batch normalization, and evaluate them on the CIFAR-10 and ImageNet datasets. To measure power consumption and accuracy, we use the BrainFloat16 and IEEE-754 32-bit formats for our computations with a large variety of accumulators varying their OVF, MSB, and LSB parameters. Fig. 3 shows the relationship between power consumption and accuracy for different accumulator and arithmetic combinations. For example, if 84% Top1 accuracy is satisfying for Imagenet with Resnet50, the most suited arithmetic/accumulator combination is IEEE-754 32-bit/$\langle ovf:9, msb:6, lsb:-20\rangle$ represented by a light purple hexagon as all other markers are either on the right or below.

C. Conclusion

Overall, our work highlights the importance of numerically tailored accumulators for reproducibility in scientific computing applications. Our results provide valuable insights into the trade-offs between power consumption and accuracy, and we believe that our results have the potential to inform the design of future AI and scientific computing systems, and we encourage other researchers to explore the possibilities of low precision accumulators using our open-source framework.

II. Acknowledgment

Marc Casas is supported by Grant RYC-2017-23269 funded by MCIN/AEI/ 10.13039/501100011033 and by “ESF Investing in your future”.

References

[1] D. Bailey and J. M. Borwein, “High-Precision Computation and Mathematical Physics,” in *Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research — PoS(ACAT08)*. Erice, Italy: Sissa Medialab, Oct. 2009, p. 014. [Online]. Available: https://pos.sissa.it/070/014

[2] J. Johnson, “Rethinking floating point for deep learning,” *arXiv:1811.01721 [cs]*, Nov. 2018, arXiv: 1811.01721. [Online]. Available: http://arxiv.org/abs/1811.01721

[3] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1,” *arXiv:1602.02830 [cs]*, Feb. 2016, arXiv: 1602.02830. [Online]. Available: http://arxiv.org/abs/1602.02830

[4] M. Istoan and F. de Dinechin, “Automating the pipeline of arithmetic datapaths,” in *Design, Automation & Test in Europe Conference & Exhibition (DATE 2017)*, Lausanne, Switzerland, Mar. 2017. [Online]. Available: https://hal.inria.fr/hal-01373937

Louis Ledoux received his BSc degree in 2016 in Computer Science from Université de Rennes1, France. The following years, he pursued his MSc degree in parallel with an Engineer diploma from École Supérieure d’Ingénieurs de Rennes (ESIR). He concluded in 2018 his studies in Rennes with a position of Hardware Engineer at b<>com, a national research laboratory. This position allowed him to experiment with the first FPGAs in the cloud and their virtualizations. Since 2018, he has been a PhD candidate at the Computer Architecture departments of Barcelona Supercomputing Center (BSC) and Universitat Politècnica de Catalunya (UPC), Spain.

Scaling RTL Simulations with Metro-MPI

Guillem López-Paradís *[†], Adrià Armejach*[†], Miquel Moretó*[†]
[*]Barcelona Supercomputing Center, Barcelona, Spain
[†]Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {guillem.lopez, adria.armejach, miquel.moreto}@bsc.es

***Keywords*—RTL Simulation, MPI, High-Performance Computing, Verilator.**

I. Extended Abstract

Current chips have tens to few hundreds of cores which translates to several billions of transistors per ASIC. The verification of these chips can be prohibitively expensive, usually only verifying small to medium portions of a chip at the same time. In addition, EDA tools have become a bottleneck in the typical hardware design process. The parallelization of these tools is very limited or non-existent and does not scale as the size of the design increases.

The main goal of this work is to apply High Performance Computing (HPC) techniques to RTL Simulations. We have selected MPI as the HPC technique to parallelize the simulations. Hardware designs that have replicated hardware blocks and are inter-connected by latency-insensitive interfaces like, e.g., networks-on-chip or AXI, can greatly benefit from the use of Metro-MPI. We have selected OpenPiton+Ariane as a representative design to evaluate Metro-MPI and Verilator as our baseline RTL simulator. In this setup, Metro-MPI achieves 2.7 MIPS of RTL simulation throughput for the first time on 1,024 Linux-capable cores. Additionally, compared to the sequential and multithreaded Verilator of smaller designs, Metro-MPI achieves up to 135.98× and 9.29× speedups.

A. Metro-MPI

Metro-MPI is a generic methodology to distribute and parallelize RTL simulations. We use parallel programming techniques from HPC to partition repeated hardware blocks in a given design into separated simulation processes that communicates through Message Passing Interface (MPI) [1] distributed computing runtime. Metro-MPI is very effective with designs that have replicated hardware blocks of a similar size, such as the tiles in manycores with NoCs. Figure 1 shows on the left the MPI division of a 2×2 manycore. Using Metro-MPI, every tile will be simulated in a separate MPI process and at every cycle communicates with every neighbour.

Every cycle, each MPI process simulates in parallel and then synchronises with its neighbours. We do it in this manner making use of the latency-insensitive interfaces [2] that can be found on the design in the form of NoCs, AXI, etc.. Figure 1 shows on the right, an example of this latency insensitive connection between two hardware blocks.

In this work, we use OpenPiton+Ariane tiled manycore [3], [4], which connects tiles via NoCs, while peripherals and accelerators use NoCs or AXI. In addition, OpenPiton also has a separate chipset with a bootrom, memory controller and many other periperhals. This can be simulated with Metro-MPI on a separate process. Additionally accelerators connected through the AXI interface to OpenPiton such as MIAOW GPGPU [5] have also been simulated with Metro-MPI.

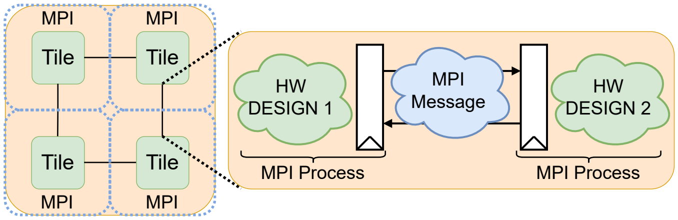


Fig. 1: Metro-MPI Connection between tiles.

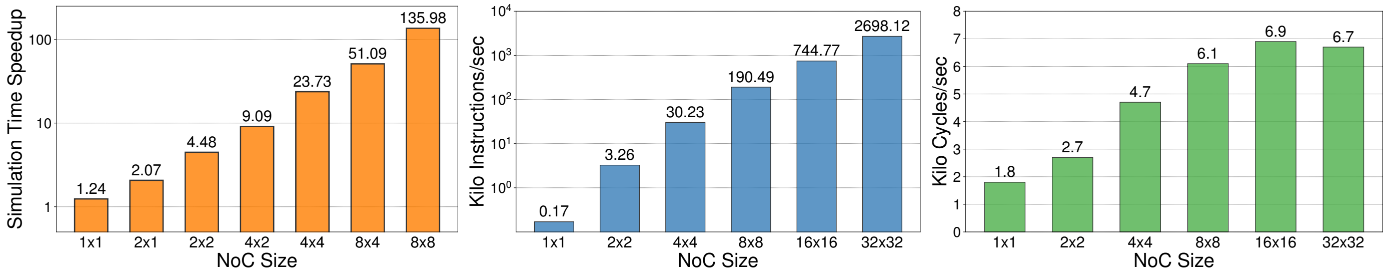
B. Experimental Environment

We use the Marenostrum IV Supercomputing HPC system with nodes that contain 2 sockets of Intel Xeon Platinum 8160 CPUs with 24 cores (48 in total) and 32MB LLC each, running at 2.10GHz. The nodes have 96GB DDR4-2667 of main memory (2GB per core) and are connected via a 100Gbit/s Intel Omni-Path HFI Silicon to other nodes. We evaluate simulations across SoC sizes using a 2D mesh NoC topology, from 1 to 1,024 cores: 1×1, 2×1, 2×2, 4×2, 4×4, 8×4, 8×8, 16×8, 16×16, 32×16, and 32×32.

C. Verilator Profiling

Table I shows a profiling analysis performed showing ICache and TLB (ITLB) misses per kilo instruction (MPKI) and instructions per cycle (IPC) for sequential and Metro-MPI simulations with Verilator. The sequential design has high ICache and ITLB MPKIs that increase significantly with manycore size, reaching 1.14 ITLB MPKI and 29.30 ICache MPKI for 64 tiles, while IPC decreases by 3× from 1.05 to 0.34. This indicates a clear bottleneck in the host's front-end. The latter can be explained by the code generation of Verilator, which has very long functions and many hard-to-predict branches. Also, the sequential binary size ranges from 2MB for 1×1 and grows to 89MB for 8×8.

On the contrary, Metro-MPI has lower ICache and ITLB MPKIs which *do not increase significantly with chip size*; ITLB MPKI starts to increase at 8×8 and ICache MPKI remains relatively stable. Our findings are confirmed by a recent paper [6] describing Verilator's conversion of RTL core logic into long C++ files with low code reusability. Clearly, Metro-MPI alleviates the ICache and ITLB problems that come with the RTL simulation of large designs, and is leaved as an open problem improving Verilator code generation.



(a) Simulation time speedup over sequential.

(b) Actual KIPS scalability.

(c) Actual cycles per second scalability.

Fig. 2: Metro-MPI performance results using Verilator: (a) simulation time speedup normalised to the 1×1 sequential design; (b) simulated KIPS and (c) simulated cycles per second with NoC designs of up to 1,024 tiles (32×32).

TABLE I: Verilator simulation profiling results.

		Chip Size				
		1x1	**2x2**	**4x4**	**8x4**	**8x8**
ITLB MPKI	Sequential	0.03	0.54	1.11	1.06	1.14
	Metro-MPI	0.01	0.01	0.01	0.12	0.39
ICache MPKI	Sequential	11.71	8.69	17.14	19.16	29.30
	Metro-MPI	7.99	9.44	10.56	9.03	9.52
IPC	Sequential	1.05	0.87	0.55	0.53	0.34
	Metro-MPI	1.31	1.31	1.37	1.18	0.96

D. Results

1) Simulation Time Speedup: Figure 2a shows the speedup of Metro-MPI from 1×1 to 8×8 (1 tile to 64 tiles) normalized to sequential Verilator. We have evaluated the sequential design only up to 64 tiles due to its prohibitively compilation times (69+ hours for 8×8). We observe a 1.2× speedup for 1×1, near-linear speedups up to 4×2, and super-linear speedups afterwards, reaching 135.9× with 64 tiles.

2) Throughput Scalability: Figure 2b shows throughput scalability in Kilo Instruction per second (KIPS) for Metro-MPI on the y-axis (log scale) and manycore dimension on the x-axis for up to 1,024 tiles (32×32). We obtain throughputs from 0.17 KIPS for 1×1 to 2.7 MIPS for 32×32. Figure 2c shows scalability in Cycles per second (CPS) on the y-axis and manycore dimensions on the x-axis. We achieve 1750 CPS for 1×1 to close to 7000 CPS for 16×16. In 32×32, we see a slight decrease in throughput, which we attribute to MPI communication cost at 22 nodes.

E. Future Work

We plan to extend this work by adding more support to other RTL Simulators both open-source and commercial. Also, we would like to use Metro-MPI on other multicores, such as Blackparrot [7] or test other available cores in OpenPiton with parallel benchmarks. Additionally, supporting the new Verilator v5 is promising because it supports the Verilog timing model and additional performance. Finally, the automatic support for Metro-MPI inside Verilator or another open-source RTL simulator could be another outcome of this work

F. Conclusions

With this work we have shown evidence of the value of HPC techniques for RTL simulations. We have obtained, for the first time, 2.7 MIPS on a 1,024 core design in RTL simulation. We have obtanied speedups in simulation time and simulation throughput, and energy reductions for a fixed amount of work compared to sequential and another parallelization technique. We invite the community to adopt Metro-MPI and support more tools and platforms.

II. Acknowledgment

This work has been accepted at Design, Automation and Test in Europe Conference (DATE), 2023. It is Open-source and can be found at https://github.com/metro-mpi.

References

[1] M. P. Forum, “MPI: A message-passing interface standard,” Tech. Rep., 1994.

[2] M. B. Taylor, “Basejump stl: Systemverilog needs a standard template library for hardware design,” in *DAC*, 2018.

[3] J. Balkind, M. McKeown, Y. Fu, T. Nguyen, Y. Zhou, A. Lavrov, M. Shahrad, A. Fuchs, S. Payne, X. Liang, M. Matl, and D. Wentzlaff, “OpenPiton: An open source manycore research framework,” in *ASPLOS*. ACM, 2016.

[4] J. Balkind, K. Lim, M. Schaffner, F. Gao, G. Chirkov, A. Li, A. Lavrov, T. M. Nguyen, Y. Fu, F. Zaruba, K. Gulati, L. Benini, and D. Wentzlaff, “BYOC: A ”Bring Your Own Core” framework for heterogeneous-ISA research,” in *ASPLOS*, 2020.

[5] R. Balasubramanian, V. Gangadhar, Z. Guo, C.-H. Ho, C. Joseph, J. Menon, M. P. Drumond, Paul, and et al, “Enabling gpgpu low-level hardware explorations with miaow: An open-source rtl implementation of a gpgpu,” *ACM Trans. Archit. Code Optim.*, jun 2015.

[6] S. Beamer, “A case for accelerating software RTL simulation,” *IEEE Micro*, 2020.

[7] D. Petrisko, F. Gilani, M. Wyse, D. C. Jung, S. Davidson, P. Gao, C. Zhao, Z. Azad, S. Canakci, B. Veluri, T. Guarino, A. Joshi, M. Oskin, and M. B. Taylor, “Blackparrot: An agile open-source risc-v multicore for accelerator socs,” *IEEE Micro*, 2020.

Guillem López Paradís is a second-year PhD student at BSC and UPC. He received his BSc degree in Computer Engineering from UPC in 2017. The following year, he worked at Xlabs in Xilinx, Dublin in Ireland. Later, he completed his MSc degree in Research and Innovation in Informatics at UPC. Since 2021, he has been with the High Performance Domain-Specific Architectures group at BSC. Lately, he has spent 4.5 months as a visiting PhD student at the University of California, Santa Barbara (UCSB), USA. His interests involved interconnecting accelerators with the coherence systems, accelerating RTL Simulations and accelerating the computation and communication of computers.

Spatially-resolved multiscale models shed light into personalized drug treatments

Alejandro Madrid[#1], Alfonso Valencia[#*2], Arnau Montagud[#1]

[#]*Barcelona Supercomputing Center (BSC), Plaça Eusebi Güell, 1-3, Barcelona, Spain*

[1]alejandro.madrid@bsc.es, [3]arnau.montagud@bsc.es

[*]*ICREA, Pg. Lluís Companys 23, Barcelona, Spain*

[2]alfonso.valencia@bsc.es

***Keywords*— agent-based simulations, spatial transcriptomics, cancer**

Extended ABSTRACT

Multiscale models have been very helpful in tissue biology by providing novel hypotheses to uncover mechanisms and novel treatments that tackle diseases of interest. PhysiBoSS is an open-source software which combines intracellular signalling using Boolean modelling (MaBoSS) and multicellular behaviour using agent-based modelling (PhysiCell) [1]. Since 2018, it has been successfully used in tissue simulation of diseases such as cancer and COVID. PhysiBoSS can use Boolean models personalized using bulk omics data and its simulation can be set-up using manually-designed 3D architectures defining an extracellular matrix (ECM). In spite of these advances, current simulations use rather simplistic 3D set-ups and much remains to be done to have a simulation that accurately produces results that can seem like the real tissue.

To have spatially-resolved personalized multiscale models, we have taken advantage of a novel technique: single cell spatial transcriptomics. This new technique allows to have the spatial information of unique cells combined with its transcriptomic state [2]. We hereby present a user-friendly workflow called "PhysiBoSS-spatial" for setting up PhysiBoSS simulations using spatial transcriptomics data. This workflow enables the translation of a slice of spatial transcriptomics in the initial disposition of the multiscale model. Users can modulate the number of cell types to be captured (and which), the resolution of the clustering and the cell-type annotation. To showcase the use of this workflow, we hereby present the set-up of the simulations using breast cancer spatial omics datasets and their simulation with different drugs treatments.

A. *Clustering and annotation of the spatial omics datasets*

First, we use the raw spatial omics dataset and re-analise them to know which subtype of cells are present in the data. For this, our pipeline processes and analyses the datasets using *Scanpy* [3]. For the clustering, the pipeline allows the user to choose from *Scanpy* internal metrics or the ESTIMATE score. ESTIMATE is a tool which provides scores for tumor purity, stromal cells presence and infiltration level of immune cells in tumor tissues [4]. Additionally, immune and endothelial cells are annotated using the Python package *CellTypist* to increase the accuracy identifying these cell types due to their low number in some samples [5].

B. *Boolean Models personalization*

PhysiBoSS embeds Boolean Models of signaling pathways in each agent to integrate genetic and environmental cues to their phenotype simulation. To have personalized Boolean models, we modified the tool PROFILE [6] with our breast

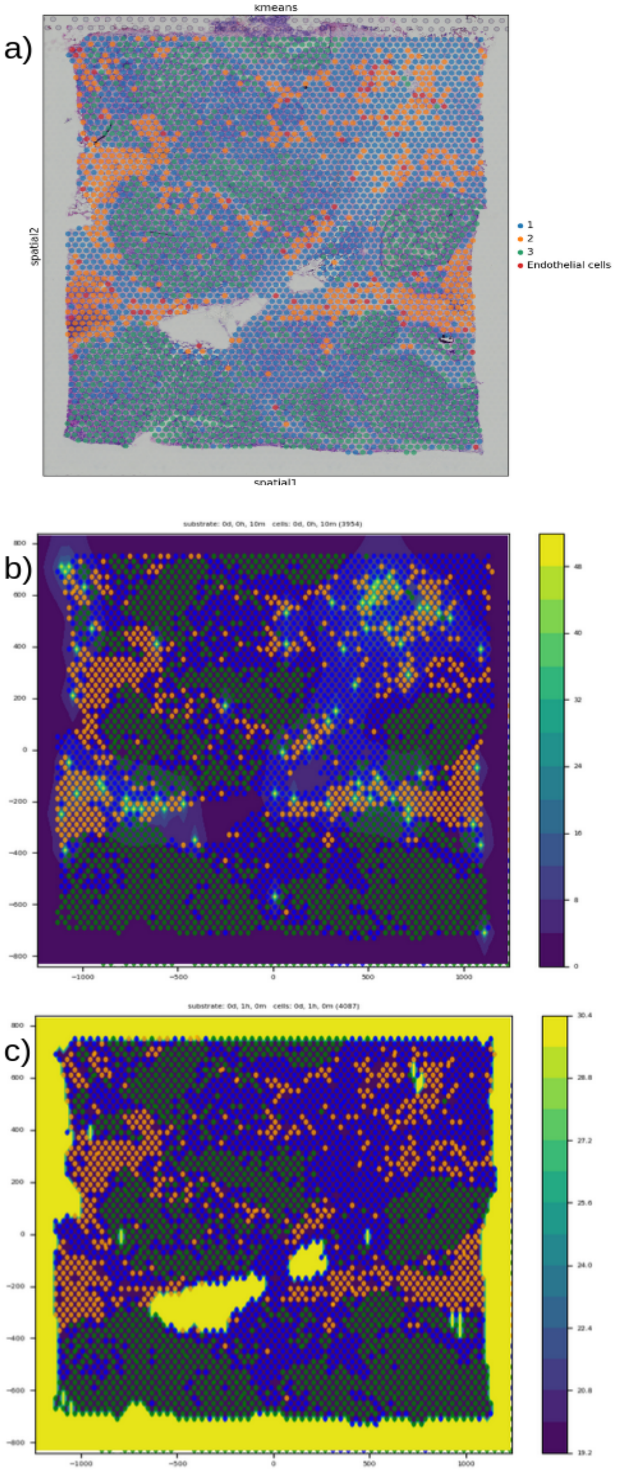


Fig. 1. Initial stage of the simulation from breast cancer spatial transcriptomics data (a) Clustering and annotation analysis of a breast cancer sample of spatial transcriptomics. Three cancer subtypes are shown with the endothelial cells. (b) Initial setup in PhysiBoSS of the tissue using our workflow, where the endothelial cells are displayed as a source of oxygen. (c) Same figure than b, but with the ECM depicted in yellow.

cancer spatial omics. As a first trial, we decided to personalize 3 different cancer subtypes and the endothelial cells. Thanks to the Boolean Model personalization our workflow recovers another layer of heterogeneity from the spatial transcriptomic data.

C. Creating the set up for PhysiBoSS

The pipeline then combines the spatial positions of each subtype of cells with the personalized Boolean Models to create all the files used for the initialization of the simulation. We have to highlight that in this first version we identify the endothelial cells as microvasculature where the oxygen, nutrients and drugs will appear in the simulation, each one with its own biophysical properties (Figure 1).

Our pipeline is also able to represent the extracellular matrix in our simulations as an obstacle that the cells cannot move through unless they produce metalloproteases to degrade it using an addon created by Ruscone et al. [7]. In this first version, we consider ECM to be present in the positions with the lowest (or no) level of RNA counts (see an unit test in Figure 2).

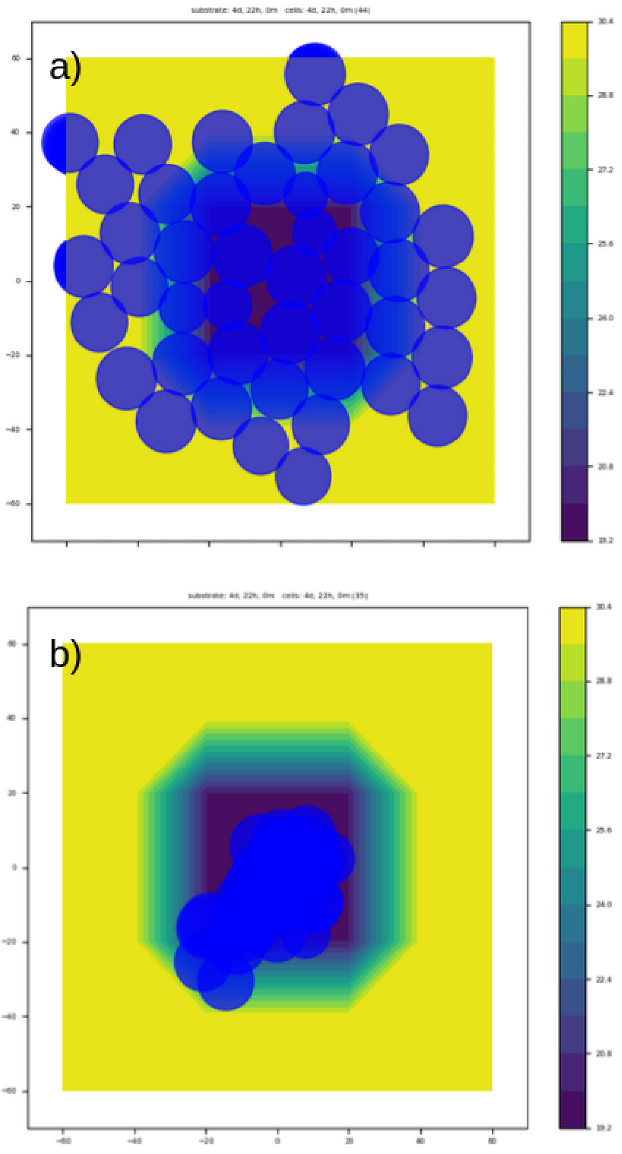


Fig. 2. Unit test for the interaction between ECM and cells. (a) Final stage of a simulation without cell-ECM interaction (b) Final stage of a simulation with cell-ECM interaction, where the cells cannot move through the ECM.

D. Discussion and future steps

PhysiBoSS-spatial is able to faithfully reconstruct a spatial omics 2D slice as an initial point for a multiscale simulation by integrating different information from the spatial transcriptomics data set: spatial position, cell type, and the transcriptomics to obtain personalized Boolean models. With this workflow users can create all the files needed to perform PhysiBoSS simulations in a straightforward way without having prior experience in spatial transcriptomic analysis or ESTIMATE or PROFILE tools. For the advanced users, our workflow allows the modification of the most important parameters in the scripts of the pipeline.

We will improve the PhysiBoSS-spatial workflow in the future and we will focus on the annotation of spatial omics to create more complex and heterogeneous simulations by including other cell types, like immune and stromal cells, and a better characterization of the cancer subpopulations.

ACKNOWLEDGEMENTS

This work has been partially supported by the European Commission under the PerMedCoE (H2020-ICT-951773).

References

[1] Letort, G., Montagud, A., Stoll, G., Heiland, R., Barillot, E., Macklin, P., ... & Calzone, L. (2019). PhysiBoSS: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling. *Bioinformatics*, *35*(7), 1188-1196.

[2] Rao, A., Barkley, D., França, G. S., & Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature*, *596*(7871), 211-220.

[3] Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, *19*, 1-5.

[4] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., ... & Verhaak, R. G. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, *4*(1), 2612.

[5] Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., ... & Teichmann, S. A. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, *376*(6594), eabl5197.

[6] Montagud, A., Béal, J., Tobalina, L., Traynard, P., Subramanian, V., Szalai, B., ... & Calzone, L. (2022). Patient-specific Boolean models of signalling networks guide personalised treatments. Elife, 11, e72626.

[7] Ruscone, M., Montagud, A., Chavrier, P., Destaing, O., Bonnet, I., Zinovyev, A., ... & Calzone, L. (2022). Multiscale model of the different modes of cancer cell invasion. bioRxiv, 2022-10.

Alejandro Madrid Valiente received his Bsc in Biochemistry and Biomedical Sciences from the Universitat de València, Spain in 2021. Now he is doing his MSc in Bioinformatics for Helath Sciences at the Universitat Pompeu Fabra, Spain. He is working in his MSc Thesis at the Computational Biology group of the Barcelona Supercomputing Center (BSC).

Horizontal Gene Transfer in Asgard Archaea

Saioa Manzano-Morales*†, Toni Gabaldón*†‡
*Barcelona Supercomputing Center, Barcelona, Spain
†Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain
‡Institución Catalana de Investigación y Estudios Avanzados, Barcelona, Spain
E-mail: saioa.manzano@bsc.es, toni.gabaldon@bsc.es

***Keywords*—Horizontal Gene Transfer, Asgard Archaea, Reticulate evolution, eukaryogenesis.**

I. Extended Abstract

Asgard archaea are considered to be the closest prokaryotic relative of eukaryotes [1]. They harbor many of what were previously thought to be eukaryote-exclusive proteins [1], including actin and actin-related proteins [2], and the presence of an actin cytoskeleton in particular has been proven in an isolated Lokiarchaeum [3]. As such, they are a key player in the debate surrounding the origin of eukaryotes (a process called eukaryogenesis) [4].

Being prokaryotes, the genome evolution of the Asgard Archaea is likely to have been shaped in no small part by Horizontal Gene Transfer (HGT), that is, the transfer of genetic material between organisms that are not bound by a parent-offspring relationship [5]. These transferred genes often encode for proteins that are beneficial for the cell and allow for adaptation to new niches [6].

In this work, we aim to unveil the fraction of the Asgard protein repertoire that stems from horizontal transfer events, by applying a HGT detection pipeline that combines homology-based and phylogeny-based methods. By analyzing the functional categories and putative donors of these genes, we hope to understand more about the evolution of Asgard archaeal genomes, so that we can employ this knowledge to shed light on the putative ecology and relationships of the archaeal partner of the symbiosis that would give rise to eukaryotic cells.

A. HGT detection pipeline

The genomic sequences and protein predictions for the cultured isolates *Candidatus* Prometheoarchaeum syntrophicum MK-D1 [7] (assembly accession GCF-008000775.1) and Candidatus Lokiarchaeum ossiferum/Lokiarchaeum sp. B-35 [3] (GenBank code CP104013.1) were downloaded from NCBI Assembly and NCBI Nucleotide/Protein, respectively.

We performed a similarity search with BLAST 2.11.0 [8] of the proteomes against a custom-made database comprised of all the species representatives of the Genome Taxonomy Database [9] species representatives and proteomes from a curated set of eukaryotes, to obtain a sufficiently representative sampling of protein sequences across the Tree of Life.

We parsed the BLAST results with HGTector [10], which systematically analyzes BLAST results looking for hit distribution patterns incongruent with a vertical evolution, given a series of hierarchically defined evolutionary categories. This step identified putative horizontally-transferred genes: for those, we retrieved the best 150 hits and reconstructed a gene tree following the algorithm implemented for PhylomeDB [11]. We further analyzed the resulting gene trees with Abaccus [12], which identifies taxonomical "jumps" in gene trees that do not follow the species tree and therefore further helps discern putative HGT events. Lastly, we performed a manual curation with an ete3-based in-house script [13] to further filter out false positives and to assess the acceptor and donor clades.

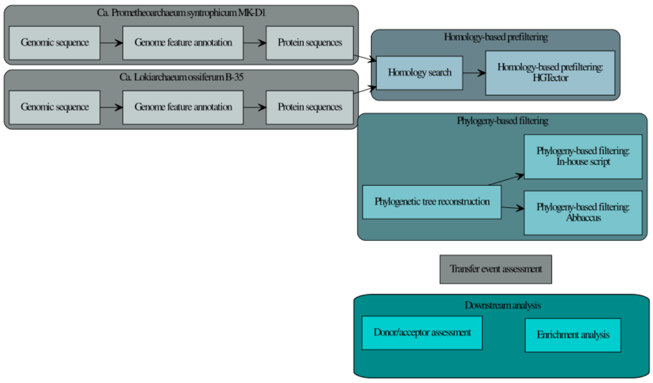

Ca. Prometheoarchaeum syntrophicum MK-D1
Genomic sequence
Genome feature annotation
Protein sequences
Ca. Lokiarchaeum ossiferum B-35
Genomic sequence
Genome feature annotation
Protein sequences
Homology-based prefiltering
Homology search
Homology-based prefiltering: HGTector
Phylogeny-based filtering
Phylogeny-based filtering: In-house script
Phylogenetic tree reconstruction
Phylogeny-based filtering: Abbaccus
Transfer event assessment
Downstream analysis
Donor/acceptor assessment
Enrichment analysis

Fig. 1. HGT detection pipeline

B. Results

Table I displays the number of putatively transferred genes per step in the pipeline and organism. 9.39% and 6.94% of the protein content of *Ca.* Lokiarchaeum ossiferum and *Ca.* Prometheoarchaeum syntrophicum, respectively, is of bacterial origin.

The transfer events have occurred over a series of time-points across the Asgard lineage 2: from genus-level to transfers that precede the diversification of the Loki lineage. Interestingly, there is a high degree of paraphyly, with many instances of the Asgard lineage forming two (or more) clades: one that branches close to Archaea (therefore, likely a copy of vertical inheritance) and one that branches closer to a bacterial clade (therefore, a likely transfer). This implies some degree

TABLE I. Number of Horizontally Transferred Genes)

Organism	Prot.	HGTector	Abaccus	HGTs
Ca. L. ossiferum	5119	717	513	481 (442)
Ca. P. syntrophicum	3890	432	359	270 (256)

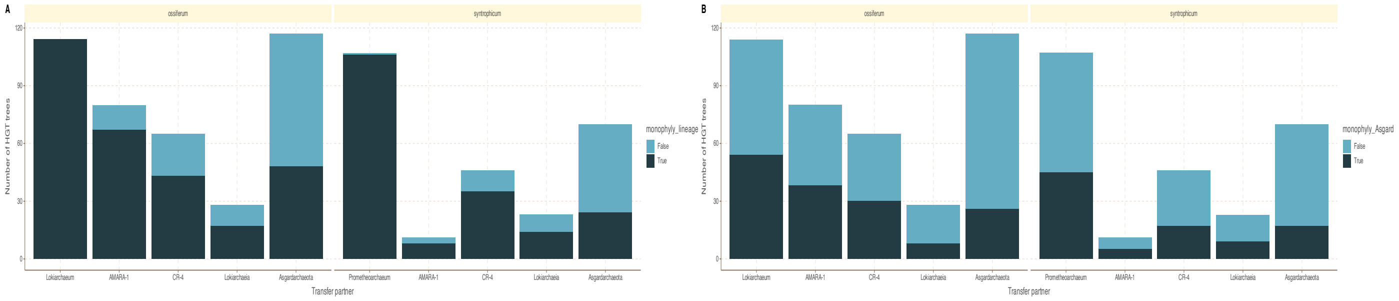


Fig. 2. Barplot displaying the number of trees per transfer acceptor. (A) Monophyly of the acceptor lineage. (B) Monophyly of the Asgard archaea.

of substitution of vertically-inherited copies by transferred ones, and a co-existence of both sources across the Asgard clade. Independent transfer events also cannot be ruled out. These transfers come from a wide arrange of donor phyla, with prominent donors being Firmicutes and Chloroflexota, followed by Proteobacteria, Spirochaeota, Desulfobacteriota and Bacteroidota. The contribution of Desulfobacterota is particularly interesting, as sulfate-reducing bacteria are known syntrophic partners of these Asgard archaea. The contribution of Anaerolineae within Chloroflexota is also non-trivial, since this lineage is known to inhabit marine sediments, a habitat where these Lokiarchaeia have been sampled.

We found instances of both Bacteria-to-Asgard and Asgard-to-Bacteria transfer, implying bidirectional flow between transfer partners.

Transferred genes seem to be enriched in metabolic functions, mainly related to lipid and amino acid metabolism, functions that seem central to the functions of the cell. They seem to mainly be components of the membrane (GO:0016021), and there is a high degree of overlap between both Lokiarchaeia.

C. Conclusion

In this study, we observe HGT events to be widespread across Asgard evolution, constituting a continuous flow of transferred genes at different points in the diversification of these archaea, and coming from a variety of donors, some of which can be linked by a metabolic or ecologic relationship.

II. Acknowledgment

This research was supported by Gordon and Betty Moore Foundation (Grant GBMF9742).

References

[1] K. Zaremba-Niedzwiedzka *et al.*, “Asgard archaea illuminate the origin of eukaryotic cellular complexity,” *Nature*, vol. 541, no. 7637, pp. 353–358, Jan. 2017.

[2] C. Akıl and R. C. Robinson, “Genomes of asgard archaea encode profilins that regulate actin,” *Nature*, vol. 562, no. 7727, pp. 439–443, Oct. 2018.

[3] T. Rodrigues-Oliveira *et al.*, “Actin cytoskeleton and complex cell architecture in an asgard archaeon,” *Nature*, vol. 613, no. 7943, pp. 332–339, Jan. 2023.

[4] E. V. Koonin, “The origin and early evolution of eukaryotes in the light of phylogenomics,” *Genome Biol.*, vol. 11, no. 5, p. 209, May 2010.

[5] W. F. Doolittle, “Lateral genomics,” *Trends Cell Biol.*, vol. 9, no. 12, pp. M5–8, Dec. 1999.

[6] J. J. Power *et al.*, “Adaptive evolution of hybrid bacteria by horizontal gene transfer,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 10, Mar. 2021.

[7] H. Imachi *et al.*, “Isolation of an archaeon at the prokaryote-eukaryote interface,” *Nature*, vol. 577, no. 7791, pp. 519–525, Jan. 2020.

[8] S. F. Altschul *et al.*, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990. [Online]. Available: https://doi.org/10.1016/s0022-2836(05)80360-2

[9] D. H. Parks *et al.*, “GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D785–D794, Jan. 2022.

[10] Q. Zhu *et al.*, “HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers,” *BMC Genomics*, vol. 15, p. 717, Aug. 2014.

[11] D. Fuentes *et al.*, “PhylomeDB v5: an expanding repository for genome-wide catalogues of annotated gene phylogenies,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D1062–D1068, Jan. 2022.

[12] M. A. Naranjo-Ortíz *et al.*, “Widespread inter- and intra-domain horizontal gene transfer of d-amino acid metabolism enzymes in eukaryotes,” *Front. Microbiol.*, vol. 7, p. 2001, Dec. 2016.

[13] J. Huerta-Cepas *et al.*, “ETE 3: Reconstruction, analysis, and visualization of phylogenomic data,” *Mol. Biol. Evol.*, vol. 33, no. 6, pp. 1635–1638, Jun. 2016.

Saioa Manzano-Morales received his BSc degree in Biochemistry and Molecular Biology from the University of the Basque Country (UPV-EHU), Spain in 2019. She then completed her MSc degree in Computational Biology from the Politechnical University of Madrid, Spain in 2021. After a brief internship in the CIB Margarita Salas (CSIC), she has been with the Comparative Genomics group of Barcelona Supercomputing Center (BSC), where she is developing her PhD.

Understanding North Atlantic deep-water formation drivers in an eddy-resolving climate model

Eneko Martin-Martinez*† (eneko.martin@bsc.es), Eduardo Moreno-Chamarro* (eduardo.moreno@bsc.es),
Pablo Ortega* (pablo.ortega@bsc.es)
*Barcelona Supercomputing Center (BSC)
†Departament de Dinàmica de la Terra i l'Oceà, Facultat de Ciències de la Terra, Universitat de Barcelona (UB)

***Keywords*—Ocean-dynamics, eddy-resolving, deep-water formation.**

I. Extended Abstract

Oceans play a significant role in regulating the Earth's climate. Oceanic deep mixing is a vital process in the oceans due to its contribution to the large-scale ocean circulation and to sustain marine life. This process usually happens after the surface's more saline and warm waters cool down, becoming denser and mixing with waters at deeper levels. The deep-water masses formed this way are essential for distributing heat and nutrients throughout the world's oceans. Moreover, deep mixing also plays a crucial role in maintaining the Earth's carbon cycle by storing carbon dioxide in the deep ocean and preventing it from entering the atmosphere.

Deep-water formation is crucial in the thermohaline circulation [1]; see Fig. 1. In the North Atlantic, the Gulf Stream current brings subtropical warmer and saltier surface waters poleward. Once these waters arrive to colder regions, i.e. subpolar areas, the relatively colder atmosphere acts to cool them down, while other processes like brine rejection can increase their content of salt. Then, the water masses become denser, forcing the mixing with the lower water masses. The newly formed deep-water masses move southward along the Western Boundary, bringing colder waters and closing the circulation with upwelling in the Southern Ocean. This process helps to balance the Earth's temperature by redistributing heat between the Equator and the poles.

A. Eddy-parametrising, eddy-permitting, and eddy-resolving models

Most models contributing to the Coupled Model Intercomparison Project phase 6 (CMIP6) parameterise the mesoscale processes due to their inability to resolve them explicitly. These processes have a length scale of $100\,km$ or smaller, and the ocean eddies are part of them. The ocean eddies are circular and horizontal spiral currents that usually last weeks to months and are vital for deep mixing and air-sea interactions. These models are known as eddy-parametrising models and usually use a horizontal resolution close to $100\,km$ in mid-latitudes.

However, the parametrisation in CMIP6 models cannot replace the interactions and feed-backs of mesoscale dynamics, which could lead to non-realistic convection in the North Atlantic Ocean [2] between other biases.

The eddy-permitting models are the next group of models when increasing the horizontal resolution in the ocean. These models usually have a horizontal resolution of $25\,km$, which allows resolving the Rossby deformation radius (and therefore the largest eddies) in low-latitudes (20ºS-20ºN). When comparing these models to eddy-parametrising ones, they tend to show deeper mixing in the Labrador Sea [3].

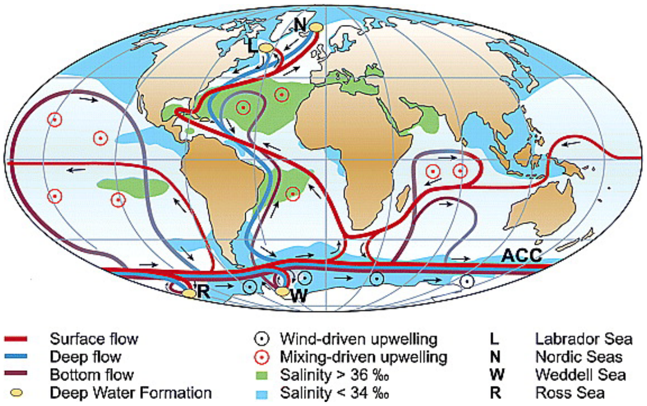


Fig. 1. Schematic of the global thermohaline circulation. Surface currents are in red, deep waters in light blue and bottom waters in dark blue. The main deep water formation sites are in orange. Source: [1]

Recent supercomputing power improvements enable us to take another step forward with coupled models that resolve the ocean mesoscale, its fine-scale interactions and feedbacks. These models are typically called eddy-resolving models and have a usual horizontal resolution close to $10\,km$ in the mid-latitudes, allowing us to resolve mesoscale eddies up to 50ºN. Several studies have shown that effectively resolving the mesoscale reduces model biases in the ocean [4] and improves air-sea interactions [5] when comparing to eddy-parametrised and eddy-permitting models. In particular, resolving the mesoscale is fundamental for the interior–boundary currents exchange in the Labrador Sea [6].

B. Methods

We use 76 years of a perpetual 1950 control simulation with the global climate model EC-Earth3P-VHR following the HighResMIP protocol [7] to investigate the role of fine-scale processes in the deep-water formation in the North Atlantic. The model EC-Earth3 is a version of EC-Earth3 [8] prepared for the PRIMAVERA project with a horizontal resolution of $1/12\,^\circ$ in mid-latitudes, approximately $10\,km$. The control configuration allows us to investigate the simulated internal variability exclusively, avoiding the trends and non-linear processes that are externally forced, e.g. the global warming produced by anthropogenic emissions of greenhouse gasses.

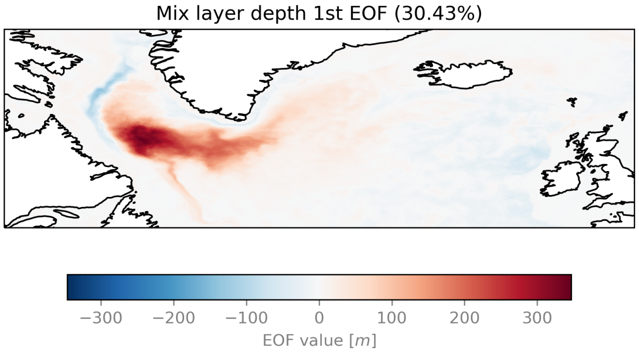


Fig. 2. First EOF of the mix layer depth in March in the North Atlantic Subpolar region. It represents the 30.43 % of the region variability.

This work mainly studies the major mode of climate variability of the mixed layer depth as described by a Principal Component analysis (PCA). The PCA allows decomposing n-dimensional climate variables in pairs of 1-dimensional time series Principal Component (PC) and (n-1)-dimensional Empirical Orthogonal function (EOF). The PC-EOFs decomposition is an orthogonal basis of the variable anomaly. Each component represents a percentage of the variability in the region that the analysis is applied.

C. Results

The first EOF of the mixed layer depth in the Subpolar North Atlantic represents the 30 % of the total variability, see Fig. 2. It captures the variability mainly in the interior of the Labrador Sea. This mode is related to the deep-water formation in that area, as it is related to density anomalies propagating down and southward (not shown).

The first PC of the mix layer depth is forced by the North Atlantic Oscillation (NAO) wind-stress pattern Fig. 3c. The same wind stress pattern enhances heat loss to the atmosphere in the Labrador Sea Fig. 3d, which cools surface temperature in the region Fig. 3a, as reported in previous studies with observations and reanalysis data [9].

There is also a significant positive surface salinity anomaly in the south of Greenland that has slowly propagated from the Eastern Subpolar North Atlantic Fig. 3b, which also contributes to increasing the local mixing. The associated time series (not shown) show that NAO-induced wind-stress changes drive the year-to-year variability, while westward propagating salinity anomalies drive the decadal variability in oceanic deep mixing.

D. Conclusions and future work

We have studied the deep-water formation in the North Atlantic with an eddy-resolving coupled climate model. We have found that positive NAOs and positive salinity anomalies can force a deeper mix layer depth in the Labrador Sea. This leads to a downward density anomaly propagation while leaving the Labrador Sea southward.

It remains to study these processes with eddy-parametrising and eddy-permitting configurations of the model, as well as to see the impact of the main modes of the mix layer depth on the intensity of the Atlantic Meridional Overturning Circulation (AMOC) and the climate of the neighbouring continents.

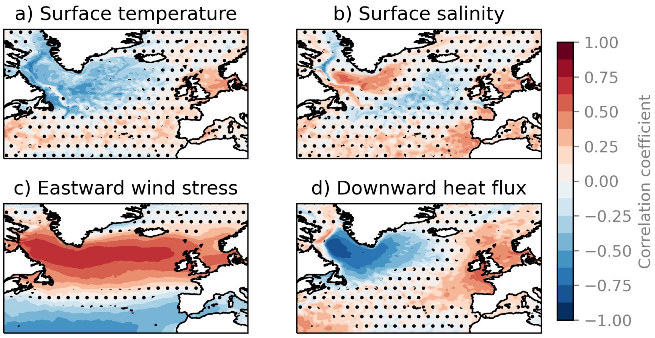


Fig. 3. Correlation coefficients of surface temperature (a), surface salinity (b), eastward wind stress (c), and downward heat flux at the surface (e) DJFM (winter) averages with the 1st PC of the mix layer depth in the North Atlantic. Black dots mask the area where the correlation is not statistically significantly different from 0 with a 95 % of confidence.

II. Acknowledgments

This work has been financed by the grant PRE2021-097163, funded by MCIN/AEI/10.13039/501100011033 and FSE+.

References

[1] T. Kuhlbrodt et al., "On the driving processes of the Atlantic meridional overturning circulation," *Reviews of Geophysics*, vol. 45, no. 2, 2007.

[2] C. Heuzé, "Antarctic Bottom Water and North Atlantic Deep Water in CMIP6 models," *Ocean Science*, vol. 17, no. 1, pp. 59–90, Jan. 2021.

[3] T. Koenigk et al, "Deep mixed ocean volume in the Labrador Sea in HighResMIP models," *Climate Dynamics*, vol. 57, no. 7, pp. 1895–1918, Oct. 2021.

[4] A. Marzocchi et al, "The North Atlantic subpolar circulation in an eddy-resolving global ocean model," *Journal of Marine Systems*, vol. 142, pp. 126–143, Feb. 2015.

[5] E. Moreno-Chamarro et al, "Can we trust CMIP5/6 future projections of European winter precipitation?" *Environmental Research Letters*, vol. 16, no. 5, p. 054063, May 2021.

[6] S. Georgiou et al, "Pathways of the water masses exiting the Labrador Sea: The importance of boundary–interior exchanges," *Ocean Modelling*, vol. 150, p. 101623, Jun. 2020.

[7] R. J. Haarsma et al., "High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6," *Geoscientific Model Development*, vol. 9, no. 11, pp. 4185–4208, Nov. 2016.

[8] R. Döscher et al., "The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6," *Geoscientific Model Development*, vol. 15, no. 7, pp. 2973–3020, Apr. 2022, publisher: Copernicus GmbH.

[9] L. Chafik et al., "Irminger sea is the center of action for subpolar AMOC variability," vol. 49, no. 17, p. e2022GL099133.

Eneko Martin-Martinez finished a double BSc in Physics and Mathematics at Universitat Autònoma de Barcelona (UAB) in 2019. During his last academic year of BSc, he also worked in the Climate Prediction group of Barcelona Supercomputing Center (BSC). He completed a MSc in Environmental Fluid Dynamics at Université Grenoble Alpes (UGA) in 2020. He made his MCs' thesis in the MEOM group of the Institut des Géosciences de l'Environnement (IGE) during a 6-month internship. He then worked as a Research Technician at Centre de Recerca Ecològica i Aplicacions Forestals (CREAF) for two years. In September 2022, he moved again to BSC's Climate Variability and Change group to start his PhD studies enrolled at the Universitat de Barcelona (UB).

New Tensor Network Structures in 1.5D

Sergi Masot Llima[#*1], Artur Garcia[#†2]

[#]CASE-Quantic, BSC, Barcelona, Spain [*]Universitat de Barcelona, Spain †Qilimanjaro, Barcelona, Spain

[1]sergi.masot@bsc.es [2]artur.garcia@bsc.es

***Keywords*— Tensor networks, Tree graphs, Quantum system simulation, Quantum computing, High-performance computing**

EXTENDED ABSTRACT

Tensor networks are a representation tool that allow simulation of large quantum systems. However, they only achieve an advantage for systems whose properties fit the correlation structure of the network, which must have some connectivity constraints to remain computationally feasible. Here we propose new structures that can adapt to the correlations of specific systems while keeping these constraints in mind. We find that for generic systems with no structure we can train structures that already achieve a small improvement. We also find ways to check that the dimension of the internal bonds allows a fair comparison.

A. Introduction

Quantum computing is steadily seeing advances in the quality and density of their devices, but is still subject to the limitations of the NISQ era with respect to what kind of algorithms they can perform [1]. The study of quantum systems has therefore been dominated by classical simulation for the last decades, both as a tool to check the quality of quantum devices as well as in terms of new findings.

Tensor networks are one of the main driving forces, beginning to see success a few decades ago with quantum inspired structures that are optimal to represent area-law quantum states, and has since grown to deal with many other quantum systems and even classical problems like machine learning [2]. In all these applications, however, it is crucial to choose the simplest tensor network structure that can encode the correlations of your system, since finding the path to contract your network quickly becomes hard, with the worst case of an arbitrary contraction being NP-complete [3]. There have been great improvements in the field regarding the contraction techniques of known structures [4], which work well along computational advances to increasingly push the feasibility bound together [5].

However, in the past there have been break-throughs that relied exclusively on how a new connectivity structure allowed, for example, to calculate critical systems [6]. For this reason, we think that there is an advantage to be gained by proposing network connectivities that adapt to the structure of the system that we want to represent while, at the same time, avoiding the properties that can make a tensor network hard to contract, such as large tree-width or lack of structure.

B. Methods

To test that the proposed structures have at least the same representation power as known structures, we prepare arbitrary quantum states of *n=12* qubits and train our proposed tensor networks to represent the same state with different bounds to the amount of data that they can store. Our arbitrary target system will be stored in dense form, which means that we will initialize at random each of the 2^n entries of a single tensor (representing the amplitude of each basis state) and then renormalize to ensure that it represents a physical state. The tensor networks that we want to train will also be initialized at random but, instead, will have *n* tensors with an open index (one for each qubit), that are connected by different structures of internal indices with dimension d≤χ, where χ is called the bond dimension. By using the following cost function *f* based on the fidelity (a quantum measure for similarity of states) between the state ψ and the target:

$$f(|\psi\rangle) = \left(\log\left(\langle\psi_{target}|\psi\rangle\right) - 1\right)^2$$

we can employ gradient descent methods until we converge on a tensor network that encodes the target state as well as possible. We then compare known structures to the proposed connectivities (fig. 1)

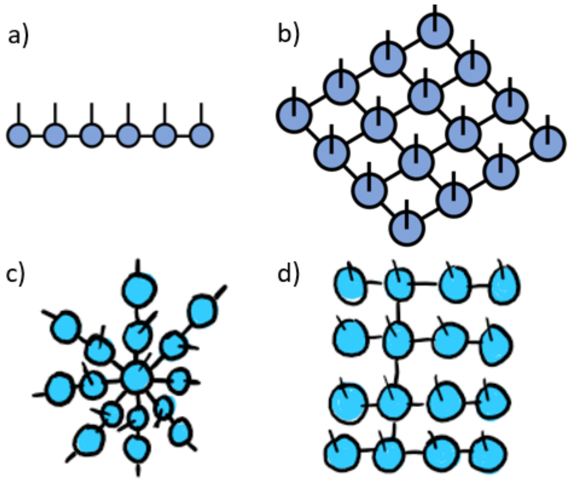


Fig. 1: Schematic of the used tensor networks structures in the training. MPS and PEPS are well-known structures shown in a) and b) respectively, whereas c) and d) are our proposals and are labeled "star" and "balanced tree".

We call these 1.5D structures because they are not linear but they avoid the loops that are characteristic in 2D and higher-dimensional structures such as PEPS.

For the second part of the work, we focus on how χ affects the representation power of the tensor networ. In particular, changing χ allows us to control the amount of information stored in these networks. Some quantum systems are able to be represented by tensor networks with low χ, allowing for the advantage over other types of simulation. We want to make the comparison in the case where the bond dimension is limited. This is because for unbounded χ, all structures are in principle able to represent the state perfectly, so we will check different, low-valued χ. However, with different types of connectivity, the same χ will affect tensor networks differently, and we want to ensure that the comparison between structures takes this into account.

The motivation behind the choice of *n* is to ensure states are big enough to present a training challenge for any χ other than extremely low ones (< 5). The performance is tracked with the infidelity (1-fidelity) between the trained state and the target. Finally, since we are getting our results with a training based on a cost function, we repeat the it 10 times and see how well the average training does.

C. Tools

To create and operate the tensor networks, we use python package *Quimb* [7], which we expand by implementing the new structures following its framework. Training is done with the L-BFGS-B method, which is a type of gradient descent. The corresponding gradients are calculated using automatic differentiation, specifically using the JAX package, which allows for great efficiency.

D. Results

The results of the training using the tools above for the case of *n=12* and *χ=8* is shown in Fig. 2, where all structures fare better than the lineal MPS, but worse than PEPS. This case was chosen because the trainings showed enough spread between structures, as well not reaching trivially low infidelities meaning that the representation power of the structure was too high.

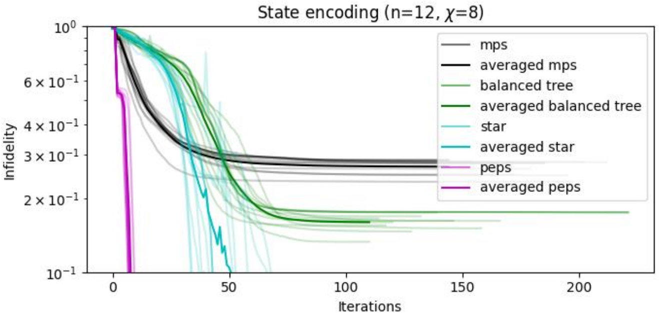


Fig. 2: Training results of different structures. Infidelity tracks improvement of the trained structured (lower is better) across the iterations of the training.

For the comparison between different bond dimensions, we can see in Fig. 3 that the advantage holds even when accounting for the different representation powers of the structures, even if the margin is less obvious.

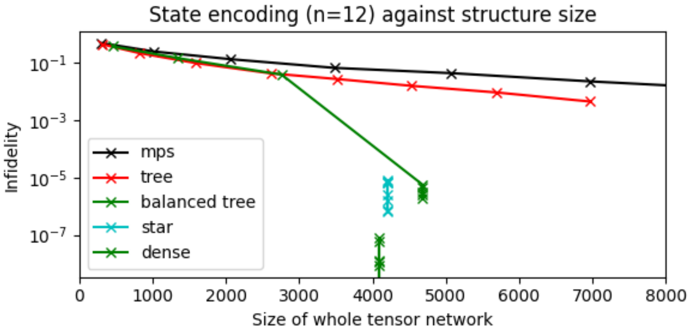


Fig.3: Performance (measured with infidelity, lower is better) of each structure for different representation powers, which we measure with the number of entries that the structure is able to hold (in the x axis). Some of the structures achieve extremely low infidelities because their representation passes the difficulty of the chosen case.

E. Conclusion and next steps

We have found that building tensor network structures strategically can bring a small advantage to the quality of simulations, even with arbitrary systems. This advantage is sustained even when accounting for the different representation powers that one finds by fixing *χ*. For this reason, the next steps of the project are testing the same structures with some system candidates that we suspect will benefit from the structure, and seeing if we can find that the advantage is greatly enhanced in a favorable situation. We also plan to generalize the ideas we used for our findings in this project to an algorithm that, given a system, can tailor the tensor connectivities to represent it to a good approximation.

F. ACKNOWLEDGEMENTS

For the progress in this work, still in progress, we want to acknowledge Sergio Sánchez and Jofre Vallès for their great work in supporting and improving our resources for tensor network simulations, in addition to their other duties, as well as the totality of the members of Quantic team for fruitful discussions on quantum algorithms and the field of tensor networks in general, without forgetting their part in building a welcoming and productive working environment.

References

[1] Noisy intermediate-scale quantum algorithms K. Bharti, A. Cervera-Lierta et al. Rev.Mod.Phys. 94, 015004 (2022)

[2] Orús, R. Tensor networks for complex quantum systems. Nat Rev Phys 1, 538–550 (2019)

[3] Faster identification of optimal contraction sequences for tensor networks R.N.C. Pfeifer, J. Haegeman, and F. Verstraete Phys. Rev. E 90, 033315 – (2014)

[4] Hyper-optimized tensor network contraction, J. Gray, S. Kourtis, Quantum 5, 410 (2021)

[5] Huang, C., Zhang, F., et al., Efficient parallelization of tensor network contraction for simulating quantum computation. Nat. Comput. Sci 1, 578–587 (2021)

[6] *Entanglement Renormalization*, G. Vidal, Phys. Rev. Lett. 99, 220405 – Published 28 November 2007

[7] quimb: A python package for quantum information and many-body calculations, The Journal of Open Source Software 3(29):819 (2018)

Author biography

Sergi Masot Llima is a physicist from Barcelona, born in 1995. He graduated from Universitat de Barcelona in 2019 with a Bachelor's degree in Physics and one in Mathematics. For his bachelor theses, he worked on a graphical representation of particle decay and the mathematical foundations of gravitational waves, respectively.

He later moved to Zürich, earning a master's degree in physics at ETHZ in 2021, where he focused on quantum computing technologies and quantum information. He worked on the realization of a quantum convolutional neural network in a real quantum device as part of his master's thesis, in the Quantum Devices group.

Having stayed in the field of Quantum computing, he is now a PhD student working at Quantic, in BSC, with a focus on large scale simulations with HPC using tensor networks, as well as the applications of this tool to quantum algorithm design. His interests include, in addition, understanding entanglement in complex quantum systems and the implementation of quantum algorithms both with real devices and using classical simulation.

Fault-tolerant applications through OpenMP

Adrian Munera*, Sara Royuela*, Eduardo Quinones*
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {adrian.munera, sara.royuela, eduardo.quinones}@bsc.es

***Keywords*—OpenMP, replication, fault-detection, CRTES, Multi/Many Core Architecture**

I. Introduction

Parallel and heterogeneous embedded platforms are commonly used to deploy complex critical real-time embedded systems (CRTES) from several domains, like automotive, due to the high-performance requirements of the most advanced functionalities, like the predictive cruise control. Additionally, the uncertainties derived from each scenario, such as randomness in the environment, errors in physical devices, and possible security attacks, make *dependability* a crucial aspect.

Highly parallel and heterogeneous processors are replacing the small microcontrollers traditionally used to deploy CRTES because they (1) boost performance through parallel processing, (2) reduce the number of components and hence costs, and (3) enable techniques like virtualization for the integration of real-time, general-purpose, and legacy software. There is however a common trend in processor design towards shrinking feature sizes, lower voltage levels, reduce noise margins, and increase clock rates for obtaining better performance at a lower power consumption. This makes systems more susceptible to faults that can affect correctness and safety, and so dependability.

Fault-detection is a common mechanism towards reliable, hence dependable, systems. One widespread technique in this scope is *redundancy*. Software faults are root causes in a high percentile of system failures in real time embedded computing (EC) systems [1], and hardware faults can also be detected by software mechanisms, as they produce errors in the software results. Therefore, this paper targets the detection of *transient software faults*, as they are one of the most common source of errors in many real time systems.

The overhead of replication can be unbearable for certain systems with tight end-to-end response time requirement, thus it becomes crucial to determine the most vulnerable parts of the program and establish a trade-off between performance and resilience. This paper introduces the following contributions:

1) A user-directed task-based replication technique exploiting parallelism through an extended version of OpenMP.
2) An implementation of the proposed technique in the LLVM compilation framework, including support in Clang, the LLVM compiler and the OpenMP runtime.
3) An evaluation of the productivity of the technique on a real-world use case from the railway domain, considering accuracy, overhead, and programmability.

II. Task-level replication with OpenMP

User-directed task-level redundancy is a flexible and convenient fault tolerance mechanism because it allows to easily define replication at different levels of granularity. The proposed mechanism maximizes flexibility by providing different ways to tune the replication with a type (spatial, temporal or both), the number of replicas, and the consolidation functions used to check the results.

The proposed syntax for OpenMP to expose replication is the following:

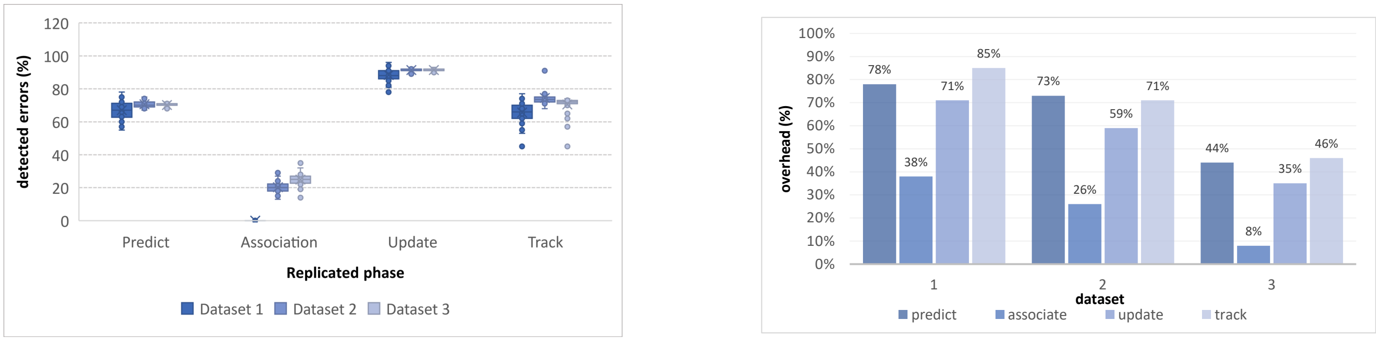
```
#pragma omp task replicated
  (n, (var:func [, var:func ...])
  [, spatial|temporal|spatial_temporal])
{/*functionality to replicate*/}
```

, where:

- *n* is the number of replicas to be created
- *var:func* is a tuple *variable:function* used to check the results by calling *func* with the original and the replicated values of *var* as arguments
- *spatial—temporal—spatial_temporal* defines the type of replication, where *spatial* forces each replica and the original task to be executed in a different core, allowing them to run in parallel; *temporal* forces each replica and the original task to be executed in mutual exclusion among them, so they have to be sequentialized, and *spatial_temporal* includes both cases.

The syntax and semantics of the proposed OpenMP mechanism have been implemented in the LLVM compilation framework [2], including support in Clang (the C/C++ frontend), LLVM (the compiler) and KMP (the OpenMP runtime).

In the extended LLVM, when a task with a `replicated` clause is found, *n+1* tasks are created and associated with the same task region. One of the tasks consumes the original data, while the rest consume copies of the written data to avoid race conditions. A synchronization task is inserted after the creation of the tasks, including as input dependencies all the tasks in the *replication set*, and inheriting as output dependencies those of the original task. Afterwards, for each *var:func* pair, one consolidation task per replica is generated; this function receives two parameters, i.e., one pointer to the original data and one pointer to the replicated data, and it implements the comparison between the original result and that from the replica, and returns a boolean expressing the correctness (equality) of the results.



(a) Accuracy of task replication

(b) Overhead of task replication

Fig. 1: Evaluation for the ODAS application depending on the dataset

III. Experimental results

Environment: The experiments are performed on a NVIDIA Jetson AGX Xavier, running a Linux OS, and featuring an 8-core NVIDIA Carmel ARM. The compilation toolchain used is the extended version of LLVM 15.0.0, supporting OpenMP 4.5 and the proposed extensions.

Application: The evaluation uses an obstacle detection and collision avoidance system (ODAS) provided by Thales, composed of three subsystems: (1) a set of *sensors*, combining radars, lidars and cameras; (2) a *data association and tracking module* that collects a huge mass of raw data from the sensors and elaborates it using Unscented Kalman Filters (UKF); and (3) a *collision checker module (CCM)* deciding whether a collision will occur to notify the driver. A collection of datasets has been created based on the analysis of real data collected on tram vehicles, including different types of objects moving around the train, with different speeds and trajectories: Scattered scene / quite zone (*dataset 1*), Crowded scene / city center) (*dataset 2*), and inflated scene / stress test (*dataset 3*).

Accuracy: Replication aims at detecting *erroneous results*. Figure 1a shows the number of errors correctly detected out of the total number of errors injected, for the three proposed datasets, when replicating different phases of the ODAS. The mechanism detects 70-80% of the injected errors in most of the cases, i.e., *predict*, *update* and *track*, while the *associate* phase, instead, shows poor results because the multitude of internal states are not easily visible to the variables checked during consolidation, hence causing *silent errors*.

Overhead: The overhead is computed as the percentage of execution time of the application with replication over the time without replication. As Figure 1b shows, the overhead highly depends on the functionality being replicated. When using 4 threads, the average overhead for the different datasets of the associate phase is a 24%, while for the predict, update and track phases is a 65%, 55% and 67%, respectively. The cost of replication is very low given that 3 replicas are executed together with the original, and the maximum overhead is 85%.

Programmability: Overall, the OpenMP task-based replication technique is a highly programmable mechanism to enhance the fault-tolerance of a system because (1) the intrusion in the code is minimal and it can easily be deactivated at compile-time by just ignoring the OpenMP directives, (2) most of the work (data copies and replication management) is performed by the compiler and the runtime, and (3) it requires very little knowledge about the application.

IV. Conclusions and future work

This paper tackles resilience and performance in CRTES by providing a novel mechanism for task-level replication based on OpenMP. The technique offers flexibility and programmability, by (1) exposing a simple interface to define different replication parameters, and (2) enriching the compilation and runtime systems with the processes needed to efficiently handle replication. The results show how the parallelism available in modern multi-core embedded systems can absorb the impact of the overhead of replication.

Compared to other mechanisms for task-level replication [3], which use a fixed attributes at runtime (i.e., 3x spatial replication), our technique presents better configurability and similar performance. In the future we will extend the replication with support for optimizations based on safety architectures, to cancel certain replicas if the results can be verified with less resources, and we will test the mechanism in further CRTES.

V. Acknowledgment

This work has been supported by the Generalitat de Catalunya project RESPECT under record no. 2021 PROD 00179 and the EU H2020 project AMPERE under the grant agreement no. 871669.

References

[1] M. Jain and R. Gupta, “Redundancy issues in software and hardware systems: an overview,” *International Journal of Reliability, Quality and Safety Engineering*, vol. 18, no. 01, pp. 61–98, 2011.

[2] C. Lattner and V. Adve, “LLVM: A compilation framework for lifelong program analysis and transformation,” 2004.

[3] O. Tahan and M. Shawky, “Using dynamic task level redundancy for OpenMP fault tolerance,” in *International Conference on Architecture of Computing Systems*. Springer, 2012, pp. 25–36.

Adrian Munera received his BSc in Computer Engineering (Computer Architecture specialization) from Universitat Politècnica of Valencia (UPV) in 2018. Then, he finished his MSc in Innovation and Research in Informatics at Universitat Politècnica de Catalunya (UPC), developing his thesis in the field of OpenMP and real-time systems. Currently, he is developing his PhD at UPC with an FPU grant and he is working in the Predictable Parallel Computing (PPC) group at BSC.

In Silico Bioprospecting of Enzymatic PEF Synthesis and Degradation

Rubén Muñoz-Tafalla[#1], Martin Floor[#2], Victor Guallar[#3]

[#]*Barcelona Supercomputing Center (BSC), Life Sciences Department, Barcelona, Spain*

[1]ruben.munoz@bsc.es, [2]martin.floor@bsc.es [3]victor.guallar@bsc.es

***Keywords*— Computational bioprospecting, Plastics, FDCA, PEF, Protein-ligand simulations, HMFO, PETase**

A. Introduction

Plastic waste accumulation is an urgent problem. Vast islands of accumulated plastic and microplastics spread endanger many ecosystems and are an imminent health threat to human populations [1]. Greener alternatives are being searched to fight against this environmental problem, including refining the recycling processes and searching for other options to fossil-based plastics.

Poly(ethylene-2,5-furandi-carboxylate) (PEF) is a biopolymer structurally similar to poly(ethylene terephthalate) (PET), a widely used petroleum-derived polymeric plastic. Compared to PET, PEF has shown improved mechanical properties, reduced oxygen permeability, and a higher glass transition temperature, among other properties [2]. However, its extensive use has yet to be adopted because of the cost associated with PEF synthesis and a still unclear degradation strategy. The former could be solved by improving the synthesis of 2,5-Furandicarboxylic acid (FDCA), the monomeric building block, while the latter by finding or improving enzymes for an effective depolymerization reaction.

We searched the available protein sequence and structural databases to find new enzymatic activities for the synthesis of FDCA and PEF degradation. We devised a vast computational bioprospecting strategy based on Monte Carlo simulations to uncover the interaction energy landscapes underlying the different enzyme and substrate combinations. This experiment revealed many uncharacterized enzymes showing good substrate affinity for reactive configurations and active site preorganization. Preliminary experimental results suggest that such a computational search can enormously narrow the experimental efforts when searching for enzymes able to act over novel chemical reactions.

B. Methodology

We searched for new enzymes that could show increased activity towards the synthesis of FDCA and the depolymerization of PEF, using a computational bioprospecting strategy. First, a PSI-BLAST search of sequences was done using as target enzymes described to perform these reactions: HMFO for the FDCA synthesis reaction, and, PETase and MEHTase, from *Ideonella sakaiensis*, and other cutinases capable of hydrolyzing PET, for PEF degradation. We filtered out sequences not bearing the required catalytic residues and then structural models were retrieved from the Alpha Fold database [3]. Enzymes were simulated using the PELE software [4] against the corresponding ligands, FFCA (5-Formyl-2-furancarboxylic Acid, the FDCA precursor substrate) and (Mono-(2-hydroxyethyl)furanic acid (MHEF) and Bis(2-hydroxyethyl)furanoate (BHEF), respectively.

PELE (Protein Energy Landscape Exploration) is an in-house all-atom Monte Carlo molecular modeling sampling software employed to sample the binding energy landscape between the ligands and the enzyme's active sites [4]. PELE results were used to rank the best-performing enzymes, taking into account substrate orientations and binding energies, catalytic distances, and oxyanion stabilization.

C. Results and Discussion

We identified thousands of non-redundant sequences using the PSI-BLAST tool. After filtering the structures, we obtained around two thousand proteins for each enzymatic target (synthesis and polymerization). We simulated those enzymes against the corresponding ligands using a fast screening approach with PELE, which allowed us to coarsely discern between strong catalytic candidates and enzymes that can not bind properly the studied ligands (Figure 1).

For both the FDCA synthesis and the PEF hydrolyzation, after selecting the best 60 enzymes from the PELE screening, we performed another round of simulations with increased sampling to further evaluate their performance. We ranked the proteins based on different catalytic parameters and selected the higher ranking models for *in vitro* evaluation (CSIC, Madrid). Preliminary experimental results showed several enzymes with comparable or better activity than the initial target enzyme.

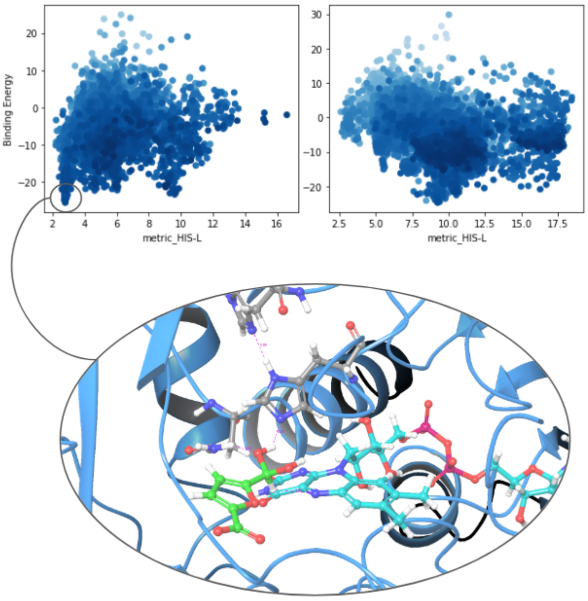


Figure 1. Comparison of the PELE simulations results for two different selected sequences regarding the synthesis of FDCA. Best pose of the left energy profile is shown at the bottom, including the value for the catalytic distances. The same study was done with PEF-like ligands for its hydrolysis.

D. Conclusions

Our computational bioprospecting strategy based on Monte Carlo simulations enabled us to identify several promising enzymes for FDCA synthesis and PEF degradation. This method can significantly narrow the experimental efforts when searching for enzymes capable of acting on novel chemical reactions. Further research is necessary to explore the potential of these enzymes for large-scale industrial applications, contributing to a more sustainable and eco-friendly future.

References

[1] E.S. Gruber, V. Stadlbauer, V. Pichler, K. Resch-Fauster, A. Todorovic, T.C. Meisel, S. Trawoeger, O. Hollóczki, S.D. Turner, W. Wadsak, A.D. Vethaak, L. Kenner, To Waste or Not to Waste: Questioning Potential Health Risks of Micro- and Nanoplastics with a Focus on Their Ingestion and Potential Carcinogenicity. Exposure and health, 2023, 15(1), 33–51.

[2] S.K. Burgess, J.E. Leisen, B.E. Kraftschik, C.R. Mubarak, R.M. Kriegel, W.J Koros, Chain mobility, thermal, and mechanical properties of poly(ethylene furanoate) compared to poly(ethylene terephthalate), Macromolecules, 2014, 47, 1383–91.

[3] M. Varadi, S. Anyango, M. Deshpande, et. al., AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D439–D444. (https://alphafold.ebi.ac.uk/)

[4] S. Acebes, E. Fernandez-Fueyo, E. Monza, M.F. Lucas, D. Almendral, F.J. Ruiz-Dueñas, H. Lund, A.T. Martinez, V. Guallar, Rational Enzyme Engineering Through Biophysical and Biochemical Modeling, ACS Catal. 2016, 6, 1624–1629.

Author biography

Rubén Muñoz is a PhD student at Barcelona Supercomputing Center, in the EAPM group led by Victor Guallar. Rubén holds a Bachelor's degree in Biochemistry and Molecular Biology from the Rovira I Virgili University in Tarragona, and a Master's degree in Biophysics from the Autonomous University of Madrid. In BSC, he designed two different pluriZymes, enzymes with more than one active centre. He is currently involved in a European project called FuturEnzymes and a Spanish national project called Furenpol, working with enzymes that synthesize and degrade plastic polymers.

VAQUERO: A Scratchpad-based Vector Accelerator for Query Processing

Julian Pavon*†, Osman Unsal*, Adrian Cristal*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {julian.pavon, osman.unsal, adrian.cristal}@bsc.es

***Keywords*—Vector Computing, Scratchpad Memory, Database Accelerator, PostgreSQL, MonetDB.**

I. Extended Abstract

Database Management Systems (DBMS) have become an essential tool for industry and research and are often a significant component of data centers [1]. There have been many efforts to accelerate DBMS application performance. One of the most explored techniques is the use of vector processing [2]. Unfortunately, DBMS have not been able to exploit the full potential of vector architectures.

In this paper, we present VAQUERO, our Scratchpad-based Vector Accelerator for QUEry pROcessing. VAQUERO improves the efficiency of vector architectures for DBMS operations that feature lookup tables, such as data aggregation and hash joins. VAQUERO introduces a novel Advanced Scratchpad Memory specifically designed with two mapping modes — direct- and associative-mode. These mapping modes enable VAQUERO to accelerate real-world databases with workload sizes that significantly exceed the scratchpad memory capacity. Additionally, the associative-mode allows to use VAQUERO with DBMS operators that use hashed keys, e.g. *hash-join* and *hash-aggregate*. VAQUERO has been designed considering general DBMS algorithm requirements instead of being based on a particular database organization. For this reason, VAQUERO is capable to accelerate DBMS operators for both row- and column-oriented databases.

A. Vectorizing DBMS Operators

DBMS applications expose high levels of DLP because the same operations (e.g. scanning, filtering and aggregating, among others) are applied to vast amounts of data. For this reason, vectorization is a recurrent technique to improve query processing performance and energy efficiency. There is special interest in linear access operators such as scans and compression which are an easy target for vectorization.

Other operators such as *join*, *aggregate* and *sort* also expose high levels of DLP. However, this DLP is irregular and turns vectorization into a difficult task to be efficiently accomplished with current commercial vector architectures. The main challenges can be summarized as *irregular memory access patterns* and *data dependencies* in vector operations.

B. VAQUERO Overview

VAQUERO is a scratchpad-based DBMS accelerator that is tightly coupled to the core pipeline that aims to tackle the previously mentioned challenges. VAQUERO is composed of two main building blocks: (1) an Advanced Scratchpad Memory (*ASPM*, Fig. 1.b) and (2) the VAQUERO control (*VAQ_CTRL*, Fig. 1.a). The *ASPM* works as a high-speed memory structure that stores irregularly accessed DBMS data, such as the aggregation state in the *group-by aggregation* operator. This scratchpad implements two mapping modes — direct- and associative-mode — to improve the performance of vector architectures over DBMS operators. In direct-mode, the *ASPM* works as a SRAM memory and directly uses an input vector register to generate as many parallel requests as the number of available ports in its memory. In associative-mode, the *ASPM* features a Content Addressable Memory to scale efficient hardware operations to large databases. To the best of our knowledge, we are the first to propose a scratchpad based design for a tightly coupled DBMS accelerator.

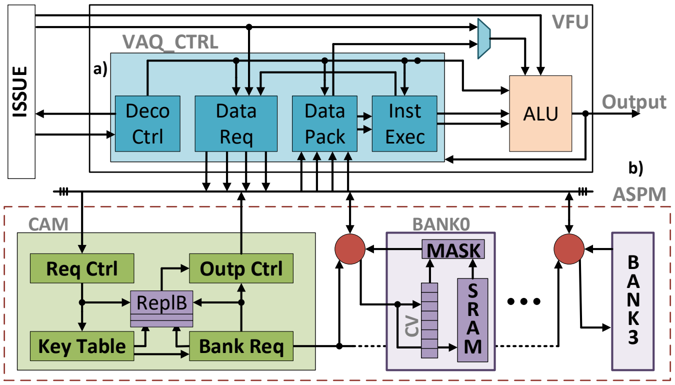


Fig. 1. Overall Architecture of VAQUERO. a) The control logic *VAQ_CTRL*. b) The Advanced Scratchpad Memory (ASPM).

The *VAQ_CTRL* works as the interface between the *ASPM* and the rest of the core micro-architecture components. It is composed of four components: (1) A decoder control (`Deco Ctrl`) that wakes up the other components when a VAQUERO instruction is issued to the Functional Unit. (2) A data request (`Data Req`) module that uses one of the input vector registers as the indices to access the *ASPM*. (3) A data packer (`Data Pack`) module that gathers in a single vector register all the elements read from the *ASPM*. And (4) an instruction control (`Inst Exec`) that coordinates the ALU in the Functional Unit to execute with data from the *ASPM*.

Several state-of-the-art vectorized software and hardware solutions use column-oriented databases as their baseline and do not evaluate the implications that row-oriented databases have for vector computing. However, row-oriented databases are highly utilized in industrial data warehouses [3] and present new challenges that are not properly addressed by recent proposals. VAQUERO was designed using the insights and requirements from different DBMS operators independently of the database organization. For this reason, VAQUERO is a general solution targeting to improve performance over several DBMS operations for both row- and column-oriented DBMS implementations. VAQUERO is the first on-chip DBMS accelerator that is tightly coupled to a standard vector processor pipeline and demonstrate its efficiency on both row- and column-oriented DBMS with minimal area and power impact.

C. Experimental Environment

Simulator: To evaluate VAQUERO, we use the gem5 simulator [4]. The simulated system is an X86 full-system running an Ubuntu 20.04 OS with a 5.15.0-48 Linux Kernel. We modelled a Xeon Gold 5318N-like architecture [5] which is a server class processor based on the Intel Ice Lake; and added to gem5 with our structures to simulate the *ASPM* and *VAQ_CTRL* functionality.

Benchmarks: We evaluate VAQUERO compared to PostgreSQL [6] and MonetDB [7], two highly optimized C/C++ DBMS for row- and column-oriented databases respectively. We use the Q01, Q03, Q09 and Q18 queries from TPC-H Benchmark [8] and include two extra queries referred as Q'01 and Q'06. The two last queries perform the same operations than Q01 and Q06 from the TPC-H benchmark. However, Q'01 and Q'06 generate larger output lookup tables that significantly exceed the capacity of the *ASPM*.

Datasets: The TPC-H database generator allows generating databases in various scales. Using this feature, we generated our input databases using a SF=1 (Scale Factor = 1GB). Additionally, to evaluate the performance impact from data distribution in real world databases, we use the *TPC-H Data Generation with Skew* tool from Microsoft [9] to create two additional database distributions. (1) A database with an hhiter distribution (`hhiter`, 50% of the data is a single heavy hitting value) and (2) a pseudo random distribution with a zipfian probability (`zipfian`).

D. Results

Figure 2 depicts the performance results obtained when comparing VAQUERO with PostgreSQL and MonetDB. On average, VAQUERO outperforms PostgreSQL and MonetDB by 2.01× and 2.82× respectively for all the executed queries. We make three observations: (1) Q01 is dominated by the *group-by aggregation* operator and its aggregation state fits in the *ASPM* in *direct-mode*. For this reason, it presents constant performance results across all the database distributions. (2) The remaining queries are either dominated by the *join* operator (Q03, Q09 and Q18) or its output aggregation state significantly surpass the size of the *ASPM* (Q'01 and Q'06). The efficiency of VAQUERO over these queries is directly related to the data distribution. For a zipfian distribution, VAQUERO outperforms by 2.45× and 3.32× to PostgreSQL and MonetDB respectively. (3) In real databases, keys can be heavily skewed (Similar to `hhiter` and `zipfian`). Thus, VAQUERO efficiently improves the performance of vector architectures over real queries and databases.

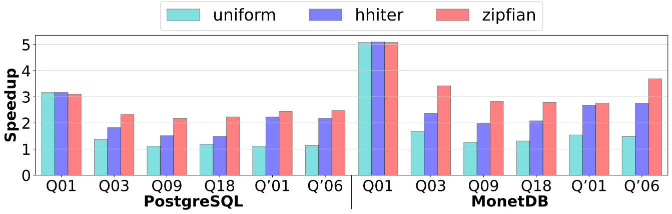


Fig. 2. Performance results of the studied queries. Results are normalized to PostgreSQL and MonetDB naive implementations.

E. Conclusion

In this paper, we introduce VAQUERO, a novel Scratchpad-based Vector accelerator that (1) eases vectorizing DBMS operators that work with lookup tables (e.g., *data aggregation* and *hash join*) and (2) significantly improves performance over DBMS operators independently of the database organization.

II. Acknowledgment

This work has been published in the Proceedings of The 29th IEEE International Symposium on High-Performance Computer Architecture (HPCA-29), 2023 [10].

References

[1] Z. Azad et al., "Hardware acceleration for DBMS machine learning scoring: Is it worth the overheads?" in *ISPASS*. IEEE, 2021, pp. 243–253.

[2] S. Jah et al., "Improving main memory hash joins on intel xeon phi processors: An experimental approach," *Proc. VLDB Endow.*, vol. 8, no. 6, pp. 642–653, 2015. [Online]. Available: http://www.vldb.org/pvldb/vol8/p642-Jha.pdf

[3] S. IT, "Db-engines ranking," https://db-engines.com/en/ranking, accessed: 2022-05-21.

[4] N. Binkert et al., "The gem5 simulator," *SIGARCH Computer Architecture News*, vol. 39, pp. 1–7, 08 2011.

[5] I. E. Papazian, "New 3rd gen intel® xeon® scalable processor (codename: Ice lake-sp)." in *Hot Chips Symposium*, 2020, pp. 1–22.

[6] B. Momjian, *PostgreSQL: introduction and concepts*. Addison-Wesley New York, 2001, vol. 192.

[7] monetdb, "monetdb," https://monetdb.org/home, [Web, accessed January 15, 2022].

[8] TPCH, "Tpc-h homepage," https://www.tpc.org/tpch/, accessed: 2022-04-20.

[9] Avatar Srikanth Kandula - Microsoft Corporation, "A parallel zipf-skewed data generator for TPC-H benchmark," https://github.com/SrikanthKandula/tpch_dbgen_zipf_skew, accessed: 2022-04-22.

[10] J. Pavón et al., "Vaquero: A scratchpad-based vector accelerator for query processing," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 1289–1302.

Julian Pavon was born in Panuco, Mexico, in 1992. He received the B.E degree in Electronic Engineering from the Panuco's Institute of technology, Mexico, in 2015, and the MIRI degree in Research of Informatics from the Universitat Politecnica de Catalunya (UPC) Barcelona, Spain in 2018.

Since March 2018 he has been with the *Computer Architecture for Parallel Paradigms* group, Barcelona Supercomputing Center, where he was a research engineering, and became a PhD student in 2019. His current research topics include vector architectures, hardware-software codesign, Embedded Systems, RTL design and RISCV SoC design.

Machine Learning approaches for the characterization of COPD

Iria Pose-Lagoa [*†], Alfonso Valencia [*‡]
[*]Barcelona Supercomputing Center, Barcelona, Spain
[†]Universitat Pompeu Fabra, Barcelona, Spain
[‡]ICREA, Barcelona, Spain
E-mail: iria.poselagoa@bsc.es, alfonso.valencia@bsc.es

Keywords—Chronic Obstructive Pulmonary Disease; Machine Learning, ; gene expression

I. Extended Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous and underdiagnosed disease characterized by the level of airflow limitation. It can be classified into four levels of severity, taking into account the ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC), FEV1/FVC < 0.7, (after bronchodilator), and FEV1 (mild $\geq 80\%$, moderate $50 - 80\%$, severe $30 - 49\%$, very severe $< 80\%$ predicted).

In recent years, Machine Learning techniques have demonstrated their ability to outperform traditional statistical approaches in a variety of biomedicine tasks. This major breakthrough, specially provided by deep neural networks, lies in their ability to find complex interaction patterns in the input data. Consequently, they can process raw data and recover relevant features that remain undetectable to less complex models.

Here, we use gene expression data to characterize the molecular basis of COPD and its clinical subtypes. Therefore, we will identify the genes and biological processes that may be involved in the disease risk and specific subtypes.

A. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a major chronic disorder with smoking exposure as the first risk factor. It constitutes the third leading cause of death worldwide, causing more than 3 million deaths in 2019, as the World Health Organization reported.

This inflammatory lung disease is characterized by its complexity and heterogeneity, comprising a wide range of non-identical patient profiles. It has been hypothesized that such heterogeneity results from the interplay between lung disease, the systemic effects of the disease, and its comorbidities, each with its particular dynamic [1].

In recent years, Machine Learning (ML) has provided tools for preventing, diagnosing, treating, and handling pathologies, improving patient care and health management, and leading to more accurate, effective, and personalized treatments. As a matter of fact, several studies have inspected machine-based learning algorithms and penalized regression models for the analysis of COPD and possible candidate therapeutic genes, achieving accuracies below 82% [2]. Therefore, understanding the diversity of the disease is important for diagnosing and treating COPD, enabling the implementation of more individualized therapies. Nonetheless, the previously described molecular subgroups have little overlap among studies.

Here, we developed innovative ML techniques of increasing complexities that will help diagnose the disease and acquire novel insights into the underlying molecular mechanisms of the different patient subgroups (endotypes) of COPD.

B. Methods

First, we collected gene expression data from the Lung Tissue Research Consortium (https://ltrcpublic.com). The data was extracted from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and belonged to two different microarray platforms containing both clinical (spirometry values, age, sex, smoking status) and molecular information for COPD and control patients. Then, we developed a pipeline using the limma package to compute an individual pre-processing for each platform and combine the different cohort studies.

Next, using different training methodologies – simple cross-validation, repeated cross-validation, Bayes optimization –, we trained classical ML models – Random Forest (RF), Support Vector Machines (SVM) with polynomial and radial kernels, K-Nearest Neighbours (KNN), XGBoost (XGB) and Penalized Regression Models (PenReg).

At the same time, we evaluated the effect of the number of genes in the classification task. To do so, we performed a differential expression analysis of our original data giving input transcriptomic profiles formed only by the top deferentially expressed genes. As an alternative approach to using the means as the differentiation factor, we applied an iterative method that selects features with maximum relevance concerning the target variable and minimum redundancy concerning the previously selected features (Maximum Relevance and Minimum Redundancy [3]). Moreover, we extracted relevant COPD genes from well-established databases of disease-gene associations and Genome-Wide Association Studies. Since it is also possible to examine underlying relationships using signaling, another approach was to generate new matrices in the context of signaling pathways for interpreting changes in gene expression. This last approach can help reduce the noise associated with analyzing large sets of genes and provide insights into the underlying mechanisms of complex diseases.

C. Results and discussion

Our preliminary results show that most classifiers perform similarly across different input gene sets, achieving accuracies of at most 0.903 (KNN) for a set of 45 genes using cross-validation. KNN achieved the best results, while XGB had the worst performance. We also observed a *”plateau phase”* between 50 and 200 genes for most classifiers, where accuracies remained higher than 0.85 for KNN (Figure: 1(a)).

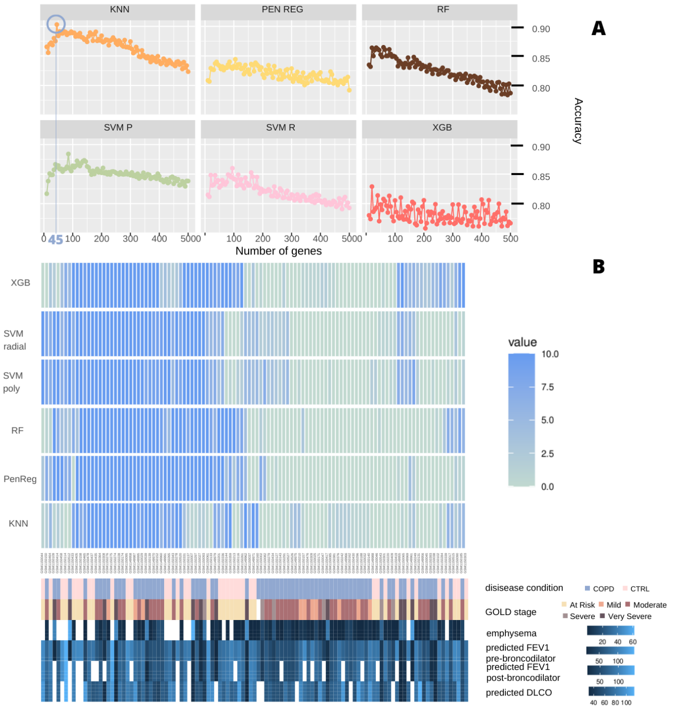


Fig. 1. Classification of COPD patients. A. These curves show how the prediction performance behaves while adding more genes as input to the ML methods. The curves are based on the accuracy achieved on repeated cross-validation tunning methodology. The blue circle marks the optimal accuracy across all methods. B. The histogram represents only the samples incorrectly classified at some point across different algorithms based on the set of 68 differentially expressed genes. The intensity value indicates the number of times (from 0-10) each sample has been misclassified. Both figures show the results obtained using a repeated 10-fold cross-validation methodology for tunning and training the classifiers. Only the clinical variables that are significantly enriched are shown as sidebars.

Interestingly, we still have many misclassified samples that remain consistent across different algorithms. Specifically, when we examine the misclassified samples across different ML methods using deferentially expressed genes as input, we find that the misclassified samples are enriched for controls and mild COPD, as well as some spirometry variables, such as emphysema or predicted FEV1 (Figure: 1(b)).
Overall, these findings suggest that a small set of fewer than 100 genes is sufficient to achieve good performance. Furthermore, while KNN performs well on this data set, there is still room for improvement in accurately classifying samples with specific characteristics.

D. Future directions

COPD is a highly heterogeneous and complex disease, comprising an extensive range of nonidentical patient profiles associated with various comorbidities such as cardiovascular diseases, depression, or lung cancer. However, since comorbidity data is not accessible, we have set a new objective, focused on studying the impact of age on FEV1 deterioration (a determinant of COPD severity), given the relevance of pulmonary function trajectories throughout life for the development of COPD [4]. Therefore, we will focus on values of FEV1 $\geq 80\%$(or $< 80\%$) of the predicted value in elderly ≥ 65 and young < 65 people for the prediction of COPD [5]. We will also complete the systems biology approach with aggregated biological features such as signaling pathways or gene regulatory networks. Moreover, we will implement more complex Deep Learning algorithms such as DeepType [6] – a framework that performs jointly supervised classification, unsupervised clustering, and dimensionality reduction to learn cancer (breast and bladder) data representation with a cluster structure – or autoencoders. We also want to explain and validate our model to understand the disease’s crucial genes and biological processes.

II. ACKNOWLEDGMENT

This work has received funding from HPC Technology Innovation Lab, Barcelona Supercomputing Center and Huawei research cooperation agreement (2020). We recognize Jon Sánchez Valle, Beatriz Urda García and Jose Carbonell Caballero as authors of the present study.

REFERENCES

[1] J. Roca, C. Vargas, I. Cano, V. Selivanov, E. Barreiro, D. Maier, F. Falciani, P. Wagner, M. Cascante, J. Garcia-Aymerich *et al.*, “Chronic obstructive pulmonary disease heterogeneity: challenges for health risk assessment, stratification and management,” *Journal of translational medicine*, vol. 12, no. 2, pp. 1–11, 2014.

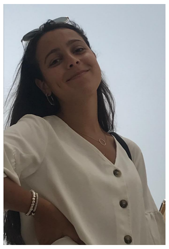
[2] S. Mostafaei, A. Kazemnejad, S. Azimzadeh Jamalkandi, S. Amirhashchi, S. C. Donnelly, M. E. Armstrong, and M. Doroudian, “Identification of novel genes in human airway epithelial cells associated with chronic obstructive pulmonary disease (copd) using machine-based learning algorithms,” *Scientific reports*, vol. 8, no. 1, pp. 1–20, 2018.

[3] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[4] A. Agusti and R. Faner, “Lung function trajectories in health and disease,” *The Lancet Respiratory Medicine*, vol. 7, no. 4, pp. 358–364, 2019.

[5] P. Lange, B. Celli, A. Agustí, G. Boje Jensen, M. Divo, R. Faner, S. Guerra, J. L. Marott, F. D. Martinez, P. Martinez-Camblor *et al.*, “Lung-function trajectories leading to chronic obstructive pulmonary disease,” *New England Journal of Medicine*, vol. 373, no. 2, pp. 111–122, 2015.

[6] R. Chen, L. Yang, S. Goodison, and Y. Sun, “Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data,” *Bioinformatics*, vol. 36, no. 5, pp. 1476–1483, 2020.



Iria Pose Lagoa was born in Galicia, Spain, in 1999. She received the BSc in Mathematics from the Universidade de Santiago de Compostela in 2021. She started her MSc in Bioinformatics for the Health Sciences from the Universitat Pompeu Fabra in September 2021. She joined Alfonso Valencia’s Computational Biology group as a master’s student the last September 2022, and is now doing her Master’s Thesis.

Representational Learning for the Study of Breast Cancer Progression through Pseudo-Time

Guillermo Prol-Castelo*, Davide Cirillo*, Alfonso Valencia*
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
E-mail: {guillermo.prolcastelo, davide.cirillo}@bsc.es

Keywords—Representational Learning, Cancer, Virtual Patients.

I. Extended Abstract

Cancer is the second most common cause of death worldwide, and its incidence is only increasing [1]. Cancer is not one disease, but many, and each cancer type has its own subtypes[2]. Some methods, such as PAM50 for breast cancer, have helped categorize cancer subtypes and how these may molecularly evolve to another subtype with a worse prognosis[2]. The rapid growth of sequenced biological data, or omics, has allowed for a more accurate picture of these subtypes. Even so, there is a scarcity of molecular data on intermediate tumor stages—these characterize a patient's clinical state and are time-dependent. In order to study tremendous amounts of information, researchers have taken to deep learning tools that can handle and learn from big data and gain new insights[3]. However, the curse of dimensionality [4] has held back the application of deep learning to omics.

For this reason, many researchers are applying representational learning[3], a set of techniques that reduces data dimensionality. In this paper, we discuss how we can learn from static data, such as that provided by The Cancer Genome Atlas (TCGA), to conduct experiments on cancer evolution in pseudo-time and *in silico*. To do so, we use the Variational Autoencoder (VAE), a type of representational learning, and we apply techniques of data oversampling and interpolations between patients at different stages. We highlight a recent literature review we performed, concluding that current applications of representational learning for cancer subtyping do not add much to the existing knowledge. Furthermore, we found no studies on cancer stages using deep learning, a missing chance in the literature to study cancer progression through time.

A. *Systematic Literature Review (SLR)*

The use of deep learning for biomedical research and clinical applications is now well-established and widespread[5]. However, these applications are usually intended for specific medical areas, such as clinical diagnosis, treatment, and disease monitoring using cardiovascular, lung, neurological, and breast imaging. Thus, we performed a SLR (following PRISMA guidelines [6]) in order to gain insight into the current state of the art of deep learning, with particular emphasis on representation learning, when studying human cancer progression through time or pseudo-time to forecast critical events. Specifically, we intend to improve our molecular understanding of cancer progression, a complex event, through time or pseudo-time using omics data and representation learning methods such as Autoencoders (AEs).

B. *Cancer Data*

Due to the absence of longitudinal human cancer data in time, we turn to TCGA to obtain a dataset of breast cancer (BRCA) from 1,064 female patients. Within the metadata of these patients, we find the specific stage at the time of data collection. The different stages are the most critical information to us. Patients may be diagnosed with one of four stages at a given time:

- Stage I involves a small tumor (<2 cm in diameter) confined to the breast tissue.
- Stage II, the tumor may be larger (<5 cm in diameter) and may have spread to nearby lymph nodes.
- Stage III, the tumor is even larger and may have invaded nearby tissues, such as the chest wall or skin, and may have spread to multiple lymph nodes.
- Stage IV represents the most advanced stage of breast cancer, where the tumor has spread to distant body parts, such as the bones, liver, or lungs.

Hence, the stage of a patient may change as the tumor progresses with time if treatment is ineffective.

C. *Variational Autoencoder*

To study these datasets, we turn to AEs, since they can embed a representation of the input data into a smaller subspace. Hence, AEs make classification tasks easier and faster and learn relevant characteristics from the generated subspace due to its reduced dimensionality. Specifically, the Variational AE (VAE) is a particular type of AE that relies on a Bayesian probabilistic generative model.

Combining VAE and bulk expression data reveals a more rounded-off picture of the subtypes than non-omics data but potentially less than single-cell expression data.

D. *Results*

As for the SLR, the initial search yielded a total of 300 records from three databases: 162 (Google Scholar), 67 (Web of Science), and 71 (Scopus), see (Figure 1). Of the 300 records, 46 duplicates were removed, and the remaining 254 were screened based on their abstract. We proceeded to remove 189 records that did not meet our criteria. From the remaining 65 records, we were unable to retrieve 2 of them, leaving a total of 63 texts for a full-text assessment. After this assessment, 51 records did not meet the initial criteria, leaving 12 records

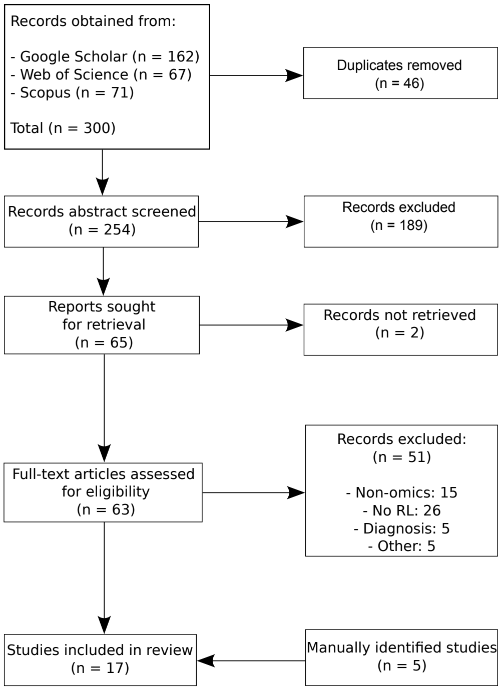


Fig. 1. Flow diagram of the systematic literature review process, following PRISMA guidelines. RL: representational learning

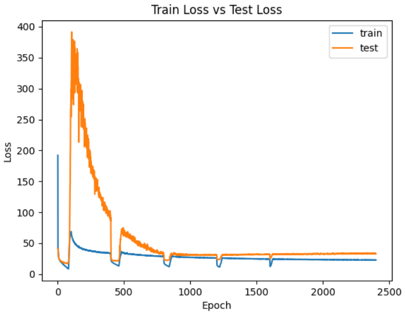


Fig. 2. Total loss obtained when training the VAE on BRCA data. Blue line represents training part, while orange represents testing.

for data extraction. Besides the database search, a reference search of the full-text-assessed records was performed, from which another 5 records were found to be relevant. All in all, 17 articles were included in the study.

Even though our keyword-based search did not return any cancer progression studies, our reference-identified studies did return two relevant studies [7], [8] in BRCA using TCGA and METABRIC data. However, this progression is in terms of the BRCA Intrinsic Subtypes, i.e., how the tumor evolves in a molecular sense. Even though these studies confirm a current interest in finding cancer progression, cancer stages, which indicate the time progression of cancer, have not yet been explicitly taken into account.

As for the BRCA data analysis, although still ongoing, we have preliminary results from training a VAE on this data with minimal loss (Figure 2). Future work will imply the study of optimally interpolating new virtual patients that may be used to perform BRCA progression analyses in pseudo-time—not real time since the data is newly generated and does not form a continuum but a discrete set.

E. *Conclusion*

AEs, or a modification of them, are the most commonly used representational learning algorithms in cancer subtyping. AEs are the most commonly used primarily because they can partly circumvent the curse of dimensionality, typically present in omics data. Another reason is that the smaller generated dataset implies computationally cheaper operations for classifying the subtypes.

To our knowledge, the study of cancer prognosis can be assessed from a molecular point of view with representational learning techniques. Thus, we believe that, in the same way cancer subtypes are identified through representational learning techniques, cancer stages could be identified following an analog procedure.

II. Acknowledgment

This project is funded by the European Union under Horizon Europe agreement No 101070430.

References

[1] M. C. Hulvat, "Cancer Incidence and Trends," *Surgical Clinics of North America*, vol. 100, no. 3, pp. 469–481, Jun. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0039610920300025

[2] S. K. Yeo and J.-L. Guan, "Breast Cancer: Multiple Subtypes within a Tumor?" *Trends in Cancer*, vol. 3, no. 11, pp. 753–760, Nov. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405803317301759

[3] A. Shmatko, N. Ghaffari Laleh, M. Gerstung, and J. N. Kather, "Artificial intelligence in histopathology: enhancing cancer research and clinical oncology," *Nature Cancer*, vol. 3, no. 9, pp. 1026–1038, Sep. 2022, number: 9 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s43018-022-00436-4

[4] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, "Digital medicine and the curse of dimensionality," *npj Digital Medicine*, vol. 4, no. 1, p. 153, Oct. 2021. [Online]. Available: https://www.nature.com/articles/s41746-021-00521-5

[5] M. Bakator and D. Radosav, "Deep Learning and Medical Diagnosis: A Review of Literature," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 47, Aug. 2018. [Online]. Available: http://www.mdpi.com/2414-4088/2/3/47

[6] M. J. e. a. Page, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, p. n71, Mar. 2021. [Online]. Available: https://www.bmj.com/lookup/doi/10.1136/bmj.n71

[7] G. Caravagna, Y. Giarratano, D. Ramazzotti, I. Tomlinson, T. A. Graham, G. Sanguinetti, and A. Sottoriva, "Detecting repeated cancer evolution from multi-region tumor sequencing data," *Nature Methods*, vol. 15, no. 9, pp. 707–714, Sep. 2018. [Online]. Available: http://www.nature.com/articles/s41592-018-0108-x

[8] Y. Sun, J. Yao, L. Yang, R. Chen, N. J. Nowak, and S. Goodison, "Computational approach for deriving cancer progression roadmaps from static sample data," *Nucleic Acids Research*, p. gkx003, Jan. 2017. [Online]. Available: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx003

Guillermo Prol Castelo received his BSc degree in Engineering Physics from Universitat Politècnica de Catalunya (UPC), Spain in 2019. He completed his MSc degree in Multidisciplinary Research in Experimental Sciences from Universitat Pompeu Fabra, Spain in 2021. Since 2022, he has been with the Machine Learning for Biomedical Research unit of the Barcelona Supercomputing Center (BSC) as well as a PhD student at the Department of Medicine and Life Sciences of Universitat Pompeu Fabra, Spain.

JLOH: Inferring Loss of Heterozygosity Blocks from Short-Read Sequencing Data

Matteo Schiavinato[#*1], Valentina del Olmo[#*2] and Toni Gabaldón[#*3]

[#]*Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain.*

[*]*Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain*

[1]matteo.schiavinato@bsc.es, [2]valentina.delolmo@bsc.es, [3]toni.gabaldon@bsc.es

***Keywords*— software, genomics, hybrids, LOH**

Extended Abstract

Introduction

Heterozygosity is a genetic condition that exists when two or more alleles are present at a specific genomic locus. It is widespread and pervades all non-haploid organisms, but is particularly high and relevant for hybrid organisms [1]. Hybrids are either the offspring of two diverged organisms from different species (inter-species hybrids) or from genetically distant populations of the same species (intra-species hybrids). Their genomes are highly heterozygous, as chromosome loci can present different alleles inherited from each of the parents. During their evolution and adaptation to a specific niche, hybrids often lose some of these heterozygous sites, a process known as loss of heterozygosity (LOH), which results in particular genomic patterns in which homozygous and heterozygous blocks alternate along a chromosome [2]. Generally, LOH information is extracted from genomic variation data using short-read sequencing. However, these data are often analysed semi-manually, using ad-hoc pipelines and parameters which are hardly reproducible in other environments. A general, easy-to-use tool that requires only a few common input data types and can handle hybrid genomes is missing from the landscape. Here, we present a software named “JLOH” that allows users to infer, analyse and visualize LOH blocks making use of parallel computing and commonly-available input data. Due to its easy implementation, JLOH can be easily added to existing pipelines, adding a layer of information to the corresponding analysis (Figure 1).

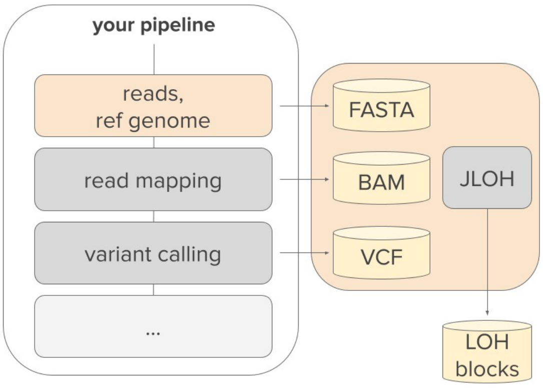


Figure 1: Implementation of JLOH within an existing variant calling pipeline.

Code modules

JLOH is subdivided in 9 modules which cover the important steps in LOH inference. The “stats” module analyses the single-nucleotide polymorphism (SNP) distribution from the input files and returns to the user the SNP density distribution quantiles. These are needed to set thresholds in the inference step. The “g2g” module identifies diverging regions between the parental genomes of a hybrid, to exclude false positives deriving from identical regions. The “extract” module is the core of the toolkit. Using the optional information from “stats” and “g2g”, it clusters homozygous and heterozygous SNPs separately to identify regions of heterozygosity and regions lacking it. Based on the SNP density quantiles identified by “stats”, it filters out clusters that have too many heterozygous SNPs within them, leaving the others as candidate LOH blocks. The candidates are then screened for coverage per position, to infer their zygosity (hemi- or homozygous). When using data from hybrids, homozygous SNPs are used to infer the allele of each block. The “filter” and the “intersect” modules allow the user to manipulate the output to their needs and filtering criteria. The “chimeric” and “junctions” modules identify genes with blocks from two different haplotypes, and LOH haplotype junctions along the genome sequence. The “plot” module provides a handy representation of LOH blocks from the output files of “extract”. Finally, the “sim” module simulates a genome with adjustable levels of divergence and LOH from the input reference genome.

Sensitivity and specificity

We tested the ability of the “extract” module to correctly infer true positive LOH blocks without missing blocks or including false positives. To do so, we used the “sim” module to simulate 24 genomes at different levels of divergence (1%, 3%, 5%, 10%, 15%, 20%) and LOH rate (10%, 20%, 30%, 40%) using the genome of *Saccharomyces cerevisiae* (*Sc*) as template. We refer to the divergent genomes as *Sd*. The “sim” module provides a list of all the generated haplotypes within the *Sd* genomes (be them LOH or not), which served as true positives and true negatives. We simulated hybrid reads from these genomes concatenated with the original one (*Sc*); hence, these reads contained both the *Sc* and the *Sd* genotype. We then mapped these reads against their corresponding *Sc* and *Sd* genomes independently and recovered SNPs from the mapping. We then used these 24 sets of SNPs and read mapping information to perform 960 “jloh extract” runs with random combinations of program parameters such as the minimum LOH block length (500, 1000, 2000, 5000, 10000 bp), the SNP density quantile as computed with “jloh stats” (Q5, Q10, Q15, Q50), and the simulated sequencing coverage (10X and 30X). In each run we evaluated the number of inferred true positives, true negatives, false positives and false negatives (Figure 2).

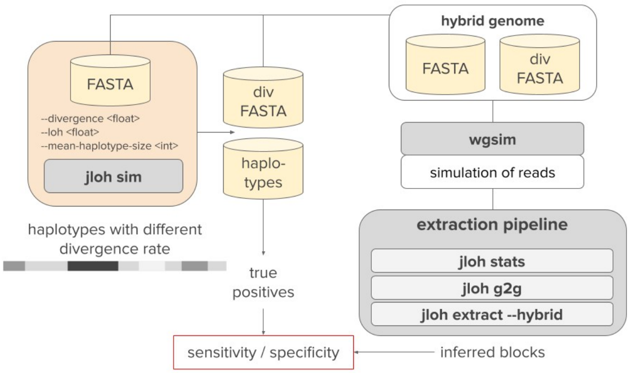


Figure 2: Schematic of the "sensitivity and specificity" workflow.

All 960 runs had sensitivity ≥ 0.90, of which a notable 93.3% reached a sensitivity of 0.99, while 68.3% of the run had specificity ≥ 0.90 (Figure 3). Results show that “extract” correctly infers LOH blocks at low and mid divergence (1-10%) between genotypes represented in the reads (from which heterozygosity arises) which a large amount of true positives are included at higher divergences (15-20%).

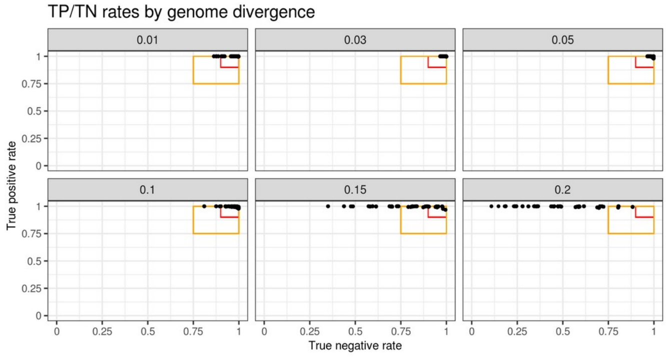


Figure 3: Sensitivity (TP rate) and specificity (TN rate) grouped by input data sequence divergence.

Re-analysis of public data

With the provided framework, it is possible to extract LOH blocks from pre-existing variation SNP data that was used for other purposes. To demonstrate the advantage of JLOH in this context, we processed five genome sequencing samples from a dataset of publicly available hybrid yeasts taken from a collection of 204 publicly available yeast hybrid samples [3]**.** which the authors characterized in terms of their environment (e.g. beer, wine, olive). Within that study, the authors extract LOH from five *S. cerevisiae x S. paradoxus* hybrids only due to software limitations. When looking at the LOH propensity of chromosome 12 of *S. cerevisiae*, which they show to have strong LOH towards *S. cerevisiae*, four out of the five strains show the expected *S. cerevisiae* predominance also when using JLOH to extract the LOH blocks. However, one (CBS7002) shows an LOH profile that is almost entirely favouring *S. paradoxus* alleles (Figure 4A). Moreover, when mapping the reads onto chromosome 12 of *S. paradoxus*, the same strain shows LOH towards an alternative allele in its second half of the sequence. This suggests the presence of a third genotype which is neither *S. paradoxus* nor *S. paradoxus*. Hence, these results demonstrate the usefulness of JLOH in the analysis of LOH blocks stratified by sample, which may reveal strikingly diverging results in different samples.

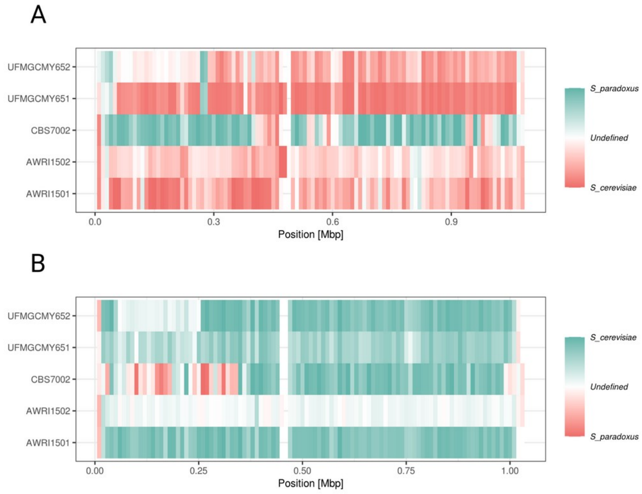


Figure 4: LOH propensity towards *S. cerevisiae* (Sc) or *S. paradoxus* (Sp) in five publicly available Sc x Sp hybrids. Labels refer to strain names, while color refers to the allele (see legend).

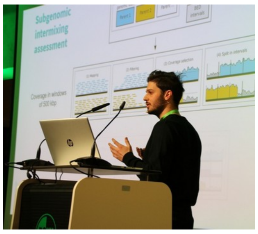
References

[1] Gabaldon T. “Hybridization and the origin of new yeast lineages”. FEMS Yeast Research. 2020 Aug;20(5):foaa040.

[2] Smukowski Heil CS, DeSevo CG, Pai DA, Tucker CM, Hoang ML, Dunham MJ. “Loss of heterozygosity drives adaptation in hybrid yeast”. Molecular biology and evolution. 2017 Jul 1;34(7):1596-612.

[3] Bendixsen DP, Peris D, Stelkens R. “Patterns of genomic instability in interspecific yeast hybrids with diverse ancestries”. Frontiers in Fungal Biology. 2021:52.

Author biography



Matteo Schiavinato was born in Italy in 1990. He received both the Bachelor’s and the Master’s degree in Molecular Biology at the University of Padua (Veneto, Italy) completing his studies in 2015.

Between 2016 and 2020 he completed a PhD in Bioinformatics and Genomics at the University of Natural Resources and Life Sciences (BOKU, Vienna, Austria).

After the PhD, he stayed in the same institute until mid 2021 first as postdoc and then as the Operative Head of the Bioinformatics Core Facility.

Since October 2021, he has arrived in Barcelona to work as a postdoc in the group of prof. Toni Gabaldón at the Barcelona Supercomputing Center (BSC-CNS). His research involves the investigation of the genomic changes underlying hybrid genome evolution.

Parallelizing Recurrent Neural Networks and variants using OmpSs

Robin Sharma*†, Marc Casas*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {robin.sharma, marc.casas}@bsc.es

***Keywords*—Deep neural network (DNN), Wavefront Parallelization, Task Parallelism, Recurrent Neural Networks (RNNs), Bidirectional recurrent neural networks (BRNNs), Long-short term memory (LSTM), Gated Recurrent Units (GRU).**

I. Extended Abstract

Recurrent neural networks (RNNs) are widely used for natural language processing, time-series prediction, or text analysis tasks [1]. RNNs models have been widely used in combination with convolutional neural networks (CNNs). RNNs contain memory units that display dynamic and temporal connections between past and future data. The outstanding text and signal analysis properties of RNNs and other recurrent models like Long-Short Term Memories (LSTMs) [2] and Gated Recurrent Units (GRUs) [3] make them the prevalent choice to analyze sequential and unsegmented data like text or speech signals.

RNNs have two widely used variants; one is uni-directional RNNs [1], which only preserves the information of the past because the only inputs it has seen are from the past, and the second is bi-directional RNNs (BRNNs) [4] which preserves both past and future information. The internal structure of RNNs and its variants inference and training in terms of data or control dependencies across their fundamental numerical kernels complicate the exploitation of model parallelism, which is why just data-parallelism has been traditionally applied to accelerate RNNs [1]. Model parallelism has not been fully exploited to accelerate the forward and backward propagation of RNNs on multi-core CPUs.

We present two model parallelism-based approaches: W-Par (Wavefront-Parallelization), a comprehensive approach for uni-directional RNNs, and B-Par (Bidirectional-Parallelization) for bi-directional RNNs inference and training on CPUs that relies on applying model parallelism into RNNs models. We use fine-grained pipeline parallelism in terms of tasks to accelerate multi-layer RNNs running on multi-core CPUs.

A. W-PAR Approach

We propose W-Par (Wavefront-Parallelization), an approach to parallelize multi-layer RNNs [5]. W-par conceives RNN's forward and backward propagation as a computational graph where nodes represent computation and edges identify data and control dependencies across them. W-Par exploits model parallelism on multi-layer RNNs network by creating multiple parallel tasks and specifying at the source code level their data or control dependencies. A run-time system - OmpSs [6] software orchestrates the parallel execution by considering dependencies across different computing routines and scheduling them across multi-core CPU devices. W-Par relies on the basic structure of multi-layer RNNs, where a cell on a particular layer depends on the previous cell of the same layer and its counterpart cell of the previous layer.

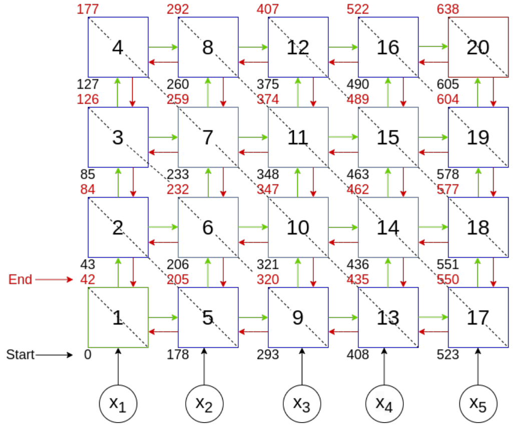


Fig. 1. 4 layer with 5 sequence length unrolled deep LSTMs network with initial and final computational graph indexes per each cell

Figure 1 represents an unrolled LSTMs network with 4 layers and a sequence length of 5 with the corresponding initial and final state numbers per each cell. State-of-the-art RNN forward propagation compute cell outputs in the sequential order displayed in Figure 2, which implies processing first Cell 1, then Cell 2, until Cell 20. W-Par maps each cell's computations in a single sequential task and orchestrates the parallel run taking into account dependencies across tasks, which defines a parallel wavefront scheme where Cell 1 is processed first, which produces the input dependencies consumed by Cells 2 and 5. For the case of LSTM cells, the 43 states of each cell and its corresponding updates are encapsulated within a single sequential task. Similarly, backward propagation can be parallelized by performing the update of Cell 20 as the starting sequential task. For a 4-layer RNN with a sequence length of 5, the maximum number of tasks running in parallel is 4. In general, for a N-layer RNN model with a sequence length of M, the maximum parallelism degree is minimal (N , M).

B. B-PAR Approach

B-Par (Bidirectional-Parallelization) is a parallel execution model for deep BRNNs [7]. B-Par conceives BRNN forward

TABLE I. LSTMS SPEED-UP OF W-PAR AND B-PAR ON CPU AGAINST SEQUENTIAL(SEQ), KERAS(K), PYTORCH(P) ON CPU AND GPU SYSTEMS.

Model Parameters				Speed-up of W-Par-CPU wrt			Speed-up of B-Par-CPU wrt				
Input	**Hidden Size**	**Batch Size**	**Seq Len**	**Seq-CPU**	**K-CPU**	**K-GPU**	**B-Seq-CPU**	**K-CPU**	**K-GPU**	**P-CPU**	**P-GPU**
64	256	128	100	3.67	4.5	1.47	2.39	1.79	0.13	3.25	0.60
256	256	128	100	3.61	4.3	1.43	2.59	1.90	0.14	4.24	0.63
1024	256	128	100	2.96	3.3	1.00	2.37	1.58	0.17	3.19	0.52
256	256	1	2	1.21	1.5	3.46	1.35	1.17	1.64	1.37	1.61
256	256	1	10	2.01	3.1	6.72	2.45	1.50	1.18	2.21	2.61
256	256	1	100	3.56	4.7	9.09	3.07	1.93	0.56	3.22	3.60
64	256	256	100	3.39	2.4	0.79	2.72	1.76	0.11	3.35	0.36
64	1024	256	100	3.56	3.2	0.10	4.09	1.64	0.07	8.51	0.00
256	256	256	100	4.03	4.3	0.92	2.75	1.75	0.13	3.42	0.35
256	1024	256	100	4.67	3.1	0.10	4.59	1.83	0.08	9.16	0.00
1024	256	256	100	3.34	3.4	0.72	2.48	1.58	0.17	3.12	0.31
1024	1024	256	100	3.50	3.1	0.10	4.43	1.78	0.09	7.31	0.00

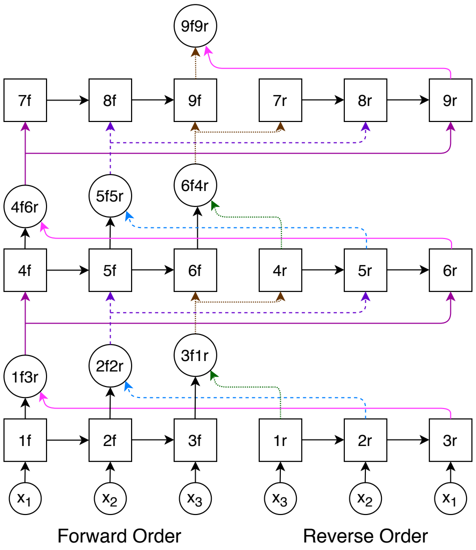


Fig. 2. Many-To-One BRNN Model

and backward propagation routines as graphs where nodes represent computation and edges identify data and control dependencies between them. B-Par executes two uni-directional RNN models simultaneously, one model processes input data in forward order and another in reverse order. B-Par exploits model parallelism on BRNN models by conceiving forward and reverse-order input computations in terms of multiple sequential pieces of code, which we denote as tasks. B-Par relies on the basic structure of deep BRNNs, where a cell on a particular layer depends on a previous cell of the same layer, its counterpart cell of the previous layer, and a cell of the previous layer in reverse order.

Figure 2 shows a 3-layer deep BRNN model with a sequence length of three. BRNNs are composed of a uni-directional RNN model for forward order input processing and another uni-directional RNN model for reverse order input processing, as shown in Fig 2.

C. Performance on Speech Recognition Task

Table I shows W-Par and B-Par against TensorFlow-Keras and PyTorch running on CPUs and GPUs, and Sequential implementation on CPUs, considering a speech recognition task on the TIDIGITS data set covering a wide range of RNN model sizes. W-Par and B-Par are always faster than TensorFlow-Keras, and PyTorch on CPUs for all configurations. Training time includes the forward and backward propagation plus the gradient update time per batch.

D. Conclusion

W-Par and B-Par deliver good scalability for the training and inference phases of uni and bi-directional RNN models on multi-core CPU devices and outperforms state-of-the-art TensorFlow-Keras and PyTorch frameworks. Future work is centered around removing connections in-between cells of RNNs layers.

II. ACKNOWLEDGMENT

Marc Casas has been supported by the Spanish Ministry of Economy, Industry and Competitiveness under the Ramon y Cajal fellowship No. RYC-2017-23269.

REFERENCES

[1] S. Mittal and Umesh, "A survey on hardware accelerators and optimization techniques for rnns." Elsevier, 2020, p. 101839.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," 1997.

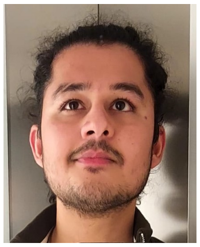
[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 12 2014.

[4] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," vol. 45, no. 11. Ieee, 1997, pp. 2673–2681.

[5] R. K. Sharma and M. Casas, "Wavefront parallelization of recurrent neural networks on multi-core architectures," in *Proceedings of the 34th ACM International Conference on Supercomputing*, ser. ICS '20, New York, NY, USA, 2020.

[6] A. Duran, E. Ayguadé, R. M. Badia, and Jesús, "Ompss: a proposal for programming heterogeneous multi-core architectures." vol. 21, 06 2011.

[7] R. K. Sharma and M. Casas, "Task-based acceleration of bidirectional recurrent neural networks on multi-core architectures," in *IPDPS*, 2022.



Robin Kumar Sharma has been a Ph.D. candidate at the Computer Architecture departments of Barcelona Supercomputing Center (BSC) and Universitat Politècnica de Catalunya (UPC), Spain, since 2019. In 2018, he pursued his MSc in parallel computing systems from the University of Amsterdam and Vrije University, Amsterdam. He worked as Software Developer in SAP, India, from 2014 to 2016.

Antibody-Derived Tag normalization for ASAP and scCUT&TAG-PRO

Xavier Soler-Sanchsis[#1,2], R. Gonzalo Parra[*1], François Serra[#2,]

[#]*Life Sciences Department, Computational Biology Group, Barcelona Supercomputing Center*
[1]xsoler@bsc.es, [2]gonzalo.parra@bsc.es

[*]*3D Chromatin Organization Department, Biola Javierre's Laboratory, Institut Josep Carreras*
[2]fserra@carrerasresearch.org

***Keywords*— single-cell, epigenomics, proteomics, normalization, scCUT&TAG-PRO**

EXTENDED ABSTRACT

A. Introduction

Recent advancements in single-cell sequencing technology make it possible to identify membrane proteins on the same cell on which we perform sequencing. This capability has been applied in scRNA-seq (CITE-seq), ATAC-seq (ASAP) and more recently in the successor of ChIP-seq, scCUT&TAG (scCUT&TAG-PRO) [1].

Current strategies to normalize Antibody-Derived Tag (ADT) counts in scCUT&TAG-PRO or ASAP experiments are based on the assumption that all cells should have a similar number of surface proteins, independently from their cell type. These normalizations sometimes overcorrect some sources of meaningful biological variation. We believe that this is the case of the state-of-the-art ADT normalization Centered Log Ratio (CLR) transformation [2] when it is applied cell-wise. Since the number of features in ADTs is usually relatively low (<100 membrane protein types vs >10.000 peak coordinates or distinct transcripts) the assumption of equal abundances should lead to even more unbalanced variances. Moreover, these corrections are agnostic to some sources of bias, leading to differences in counts which are completely artifacts of the technology, such as amplification bias.

In this work, we propose a normalization method for ASAP and scCUT&TAG-PRO data that helps remove the amplification bias while maintaining other, potentially more meaningful, differences. Our method is based on the assumption that each sequencing droplet is subject to a specific sequencing bias and hence we can correct ADT data using information from the corresponding genomic library. Concretely we use the number of background counts (sequenced fragments outside the defined peaks) in the cell. Our general assumption is validated by the currently undescribed correlation between ADT counts and genomic counts.

B. Materials and Methods

In order to ensure the reproducibility of our normalization strategy we included it into an ASAP/scCUT&TAG-PRO processing pipeline. This pipeline, the first specific to scCUT&TAG-PRO, uses the output from the cellranger utility (proprietary 10xGenomics) and covers all the steps of processing and basic analysis up to the embedded clustering of the data from both genomic and ADT libraries. It also takes into account different combinations of parameters by evaluating the quality of the resulting clusters. This evaluation is based on objective clustering quality metrics. For instance, we use the normalized mutual information score to assess how congruent the clusterings are based on data from only the genomic library or only the ADT library. We also have implemented a version of the Residual Average Gini Index (RAGI) measure [3]. This metric measures how specific genes have unequal signal counts between the different clusters defined by the integration of both, ADT and genomic libraries. This measure gives higher values when the marks for the gene are more unequaly distributed between clusters (that is when they are specific of one or a few clusters). The RAGI is measured in two sets of genes, marker genes specific to the cell-types in the sample (automatically picked from Cell Marker 2.0 [4]), and housekeeping genes (picked from HRT Atlas va.0 [5]). If the clustering correctly separates cell-types, we expect marker genes to be differentially represented in clusters while housekeeping genes not. We thus define the best clustering as the one that maximizes the median RAGI ratio between marker genes and house-keeping genes.

C. Results

When we run our pipeline with multiple combinations of parameters we can compute the RAGI ratio for all of them. The first 6 samples (Tonsil H3K27ac, Cord Blood H3K27ac, PBMC H3K27ac, PBMC H3K27ac, PBMC H3K4me3 and PBMC H3K4me2) correspond to the detection of histone marks that are characteristic of active promoters. The RAGI ratio is significantly higher in 5 of these samples when background normalization is used. H3k4me2 also marks active promoters and has indeed a very high ratio but the difference with and without our normalization is not statistically significant. Next H3K4me1 is a mark associated with active enhancers, and we see a significant improvement only in one of the two replicates. As the RAGI score is centered in the TSS of genes we do expect to see less signal for marks in enhacers and therefore the RAGI ratios are very close to one. The last two marks H3K9me3 and H3K27me3 are repressive marks and their RAGI ratio is as expected not higher than one, their relation to marker genes and housekeeping genes is probably weak (Figure 1).

D. Conclusion and Future Directions

These results show that background normalization represents a powerful tool for scCUT&TAG-PRO data processing that can lead to notorious improvements. The main advantage of this normalization strategy is that it specifically targets technical biases keeping unchanged biological differences between samples or cells. Currently, we are working in the benchmarking of this tool for ASAP and we aim to apply a similar approach to also correct CITE-seq data.

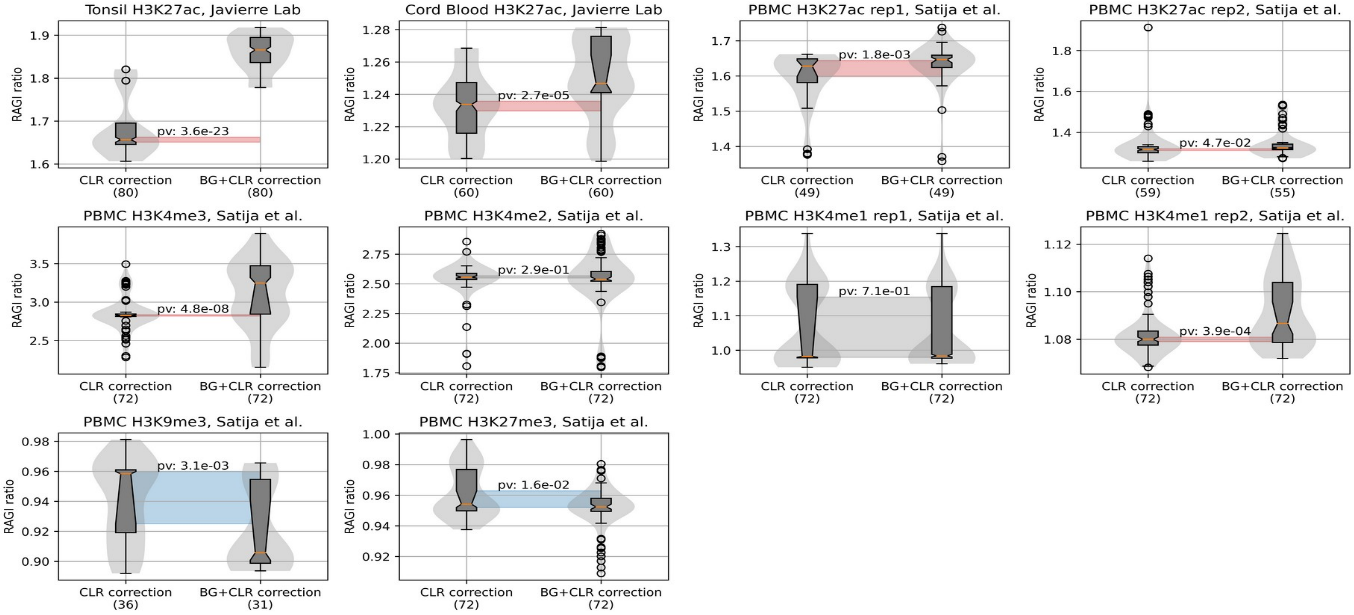


Fig. 1. RAGI ratio distributions for the different samples and different combinations of parameters in each run. In each sample the distribution is plotted for the combinations that include background normalization and for those that does not include it. The band shows 95% confidence of the median after 10,000 bootstraps in the CLR correction results; red colouring means significance of the difference towards higher RAGI ratio upon the corrected samples, blue colouring means significance towards a lower median in the CLR corrected samples and grey is applied in the non-significant difference cases. Numbers bellow X tick labels are sample sizes. P-values are from a Mann-Whitney U test.

References

[1] S. Ramesh Babu, V. S. Senthil Kumar, L. Karunamoorthy, and G. Madhusudhan Reddy, "Investigation on the effect of friction stir processing on the superplastic forming of AZ31B alloy," Mater. Des., Vol. 53, Pp. 338–348, Jan. 2014.

[2] W. Woo, H. Choo, M. B. Prime, Z. Feng, and B. Clausen, "Microstructure, texture and residual stress in a friction-stir-processed AZ31B magnesium alloy," Acta Mater., Vol. 56, No. 8, Pp. 1701–1711, May 2008.

[3] A. Dhanapal, S. R. Boopathy, and V. Balasubramanian, "Developing an empirical relationship to predict the corrosion rate of friction stir welded AZ61A magnesium alloy under salt fog environment," Mater. Des., vol. 32, no. 10, Pp. 5066–5072, Dec. 2011

[4] R. Zeng, W. Dietzel, R. Zettler, J. Chen, and K. U. Kainer, "Microstructure evolution and tensile properties of friction-stir-welded AM50 magnesium alloy," Trans. Nonferrous Met. Soc. China, Vol. 18, Pp. s76–s80, Dec. 2008.

Author biography

Xavier Soler Sanchis received his BSc degree in Biochemistry and Biomedical Sciences from Universitat de València (UV), València in 2021. The following year he joined the Master in Bioinformats for Health Sciences of the Universitat Pompeu Fabra (UPF). He is doing his internship in Biola Javierre's Laboratory at the Institut Josep Carreras (IJC) and in the Alfonso Valencia's Group at the Barcelona Supercomputing Center (BSC) working mainly in the epigenetics of the B-cell differentiation process at single-cell level.

Sorting Impact in Decision Support Benchmarks TPC-H and TPC-DS for Row-Oriented Relational Databases

Ivan Vargas Valdivieso*†, Adrian Cristal*, Osman Unsal*
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {ivan.vargas, adrian.cristal, osman.unsal}@bsc.es

***Keywords*—DBMS, Sorting, Decision Support, Relational Databases, TPC-H, TPC-DS.**

I. Extended Abstract

Sorting is an important database operation because: ❶It improves the performance of certain queries [1]. For example, when performing a GROUP BY operation or calculating aggregates such as MAX or MIN, the data needs to be sorted to ensure accurate results. ❷When data is sorted, it can be retrieved much more quickly than if it is unsorted. This is because when a query is executed, the database can use an index to find the data much more efficiently when it is sorted. ❸Sorted data is also easier to analyze, as it allows for better visualizations and easier comparisons. ❹Sorting can also simplifies database maintenance tasks, such as data backups and avoid replications. When data is sorted, it reduces the risk of data loss or corruption. Eliminating data replication simplify the overhead of extra analysis.

In summary, sorting is important in databases because it improves performance, facilitates data analysis, and simplifies database maintenance tasks.

The main objective of this work is to analyze and profile the TPC-H and TPC-DS Decision Support Benchmarks, understand the impact of sorting, and propose a novel hardware and software solution.

A. Decision Support Benchmarks

A decision support benchmark is a set of queries developed to evaluate the performance of database management systems when performing decision-related workloads. Such workloads are widely used in business intelligence, data maning, forecasting, customer segmentation and supply chain optimizations. These queries are designed to test real problems that servers are exposed to and include the most common scenarios.

In this paper, we use the decision support benchmark designed by the Transaction Processing Performance Council (TPC). The TPC benchmarks are widely used in the industry to evaluate the performance of different database systems and other computing platforms[2]. The TPC decision support benchmarks are the TPC-H and TPC-DS [3], [4].

1) TPC-H: TPC-H is extensively used in research[3], [4], [5] and industry[2]. It evaluates and compares the performance of different database systems over different technologies. The benchmark measures the performance of the database system in terms of response time and throughput for each query.

This benchmark consists of 22 queries executed against a database that includes 8 tables and a variety of data types, including integers, floating-point numbers and character strings. TPC-H includes a generator that allows to create databases from 1 GB to 100 TB, depending on the capacity of the hardware system under test.

In this work, we profile a TPC-H-generated database with scale factor (SF) 10 and 100. Results for both SF behave similarly, in the Fig. 1 we show the results obtained for SF 10 database.

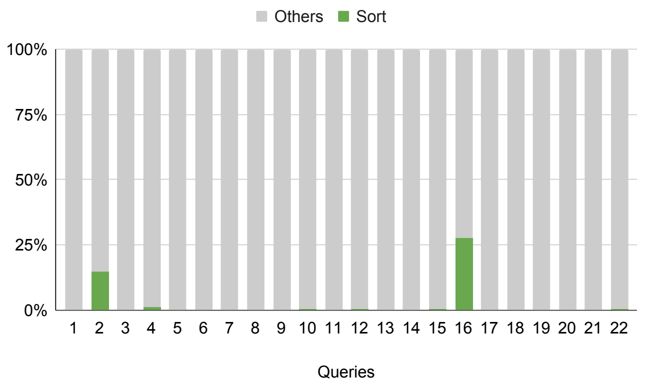


Fig. 1. Sorting impact in all TPC-H queries

2) TPC-DS: TPC-DS (Transaction Processing Performance Council - Decision Support) is the successor to the TPC-H benchmark. The TPC-H is a benchmark based on aggregations and joins. TPC-DS is a benchmark that evaluates a wider range of operations to simulate more realistic workloads. However, TPC-DS is not as popular as TPC-H. This is because it is easier to analyse the 22 queries of the TPC-H benchmark than the 99 queries from the TPC-DS benchmark.

TPC-DS not only includes a wider range of queries, but also executes over a larger data schema than TPC-H. The database generated by the TPC-DS generator is 1.9× larger than the generated for TPC-H, this is comparing the databases with the same SF. TPC-DS includes 17 tables that target more realistic scenarios, with larger intra and sub table relations.

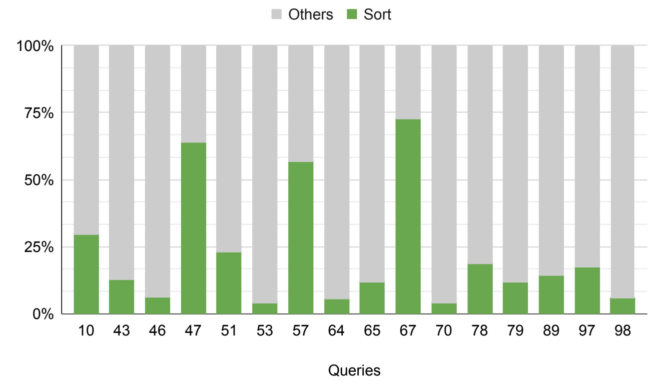


Fig. 2. Sorting impact in TPC-DS queries

In the same way that TPC-H, TPC-DS is used in the industry to compare the performance of different database systems.[2]. In this work, we profile a TPC-DS-generated database with the SF 10. The results behave similarly, in the Fig. 2 we show the results obtain for the SF 10 database. Since TPC-DS explore a wider range of operations, we found that more queries present the sorting bottlenecks, compared with the results of TPC-H.

B. Relational Database

There are two types of databases. When the data is organized in tables, it is called a relational database and is organized through the industry standard Structured Query Language(SQL). When the data is not organized in tables but typically in key-value pairs, the database is known as non-relational database or NoSQL (no SQL). In this proposal, we are going to focus on the SQL database which is the most widely used in the industry. In the SQL world, there are two ways to organize relational databases: row-oriented and column-oriented.

Row and Column-oriented databases are used for different use cases; it is more efficient to use one or the other, depending on the use case. On the one hand, for applications that focus on online transactions named OLTP (Online Transaction Processing), row-oriented databases are more efficient. On the other hand, for applications that require intensive data analysis or OLAP(Online Analytical Processing), the column-oriented databases are the best choice.

In row-oriented databases, also known as *traditional databases*, data is stored by row, such that the columns of a single row are next each other.

For instance, let's consider the next table:

Name	City	Age
Oriol	Sevilla	27
Xavi	Valencia	30
Jordi	Barcelona	33

In a row-oriented database, this data would be stored on a disk row by row, as follows:

Oriol	Sev...	27	Xavi	Val...	30	Jordi	Bar...	33

C. Methodology

We chose PostgreSQL as our reference DBMS(Database Management System), because PostgreSQL is one of the most widely used DBMSs in the industry. For a long time PostgreSQL has proven to be an optimized and stable tool. PostgreSQL is also an open source tool, this feature allows us to navigate deep into the code to analyze, propose and implement optimizations of the algorithms.

D. Conclusion

In this study, we analyzed the impact of sorting in TPC-H and TPC-DS benchmarks to measure the performance degradation of the sorting execution. The results shown comes from the direct execution of the sorting operation. However, we still take into account that sometimes sorting is used to perform other operations such as aggregation. We conclude that sorting is a problem shared among different queries. In the queries where sorting is presented it takes a big part of the execution time. We believe this study will justify the importance of improving the execution of sorting in DBMS.

References

[1] J. Mitrovski, L. Djinevski, M. Gusev, and S. Arsenovski, "TPC-H benchmark Q3, Q6 and Q12 sequential, openmp parallel and CUDA parallel implementation," in *44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021, Opatija, Croatia, September 27 - Oct. 1, 2021*, M. Koricic, K. Skala, Z. Car, M. Cicin-Sain, S. Babic, V. Sruk, D. Skvorc, S. Ribaric, B. Jerbic, S. Gros, B. Vrdoljak, M. Mauher, E. Tijan, T. Katulic, J. Petrovic, T. G. Grbac, N. F. Fijan, and V. Gradisnik, Eds. IEEE, 2021, pp. 938–943. [Online]. Available: https://doi.org/10.23919/MIPRO52101.2021.9597197

[2] TPC, "Tpc homepage," https://www.tpc.org/, accessed: 2023-02-04.

[3] M. Dreseler, M. Boissier, T. Rabl, and M. Uflacker, "Quantifying TPC-H choke points and their optimizations," *Proc. VLDB Endow.*, vol. 13, no. 8, pp. 1206–1220, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol13/p1206-dreseler.pdf

[4] M. Stufi, B. Bacic, and L. Stoimenov, "Big data architecture in czech republic healthcare service: Requirements, TPC-H benchmarks and vertica," *CoRR*, vol. abs/2001.01192, 2020. [Online]. Available: http://arxiv.org/abs/2001.01192

[5] J. Pavón, I. V. Valdivieso, J. Marimon, R. Figueras, F. Moll, O. S. Unsal, M. Valero, and A. Cristal, "VAQUERO: A scratchpad-based vector accelerator for query processing," in *IEEE International Symposium on High-Performance Computer Architecture, HPCA 2023, Montreal, QC, Canada, February 25 - March 1, 2023*. IEEE, 2023, pp. 1289–1302. [Online]. Available: https://doi.org/10.1109/HPCA56546.2023.10070958

Ivan Vargas Valdivieso received his BSc degree in Mechatronic Engineering from Mixteca Technological University (UTM), Mexico, in 2010. He completed his MSc degree in Computer Science from National Polytechnic Institute, Mexico in 2019. Since 2020, he has been with the Computer Architecture for Parallel Paradigms group of Barcelona Supercomputing Center (BSC). He started a PhD at the department of computer architecture of Universitat Politècnica de Catalunya (UPC), Spain in 2022.

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Plaça Eusebi Güell, 1-3
08034 Barcelona (Spain)

education@bsc.es
www.bsc.es

@BSC_CNS
/BSCCNS
/BSC_CNS
/barcelona-supercomputing-center
/BSCCNS