

**11<sup>th</sup> International BSC  
Severo Ochoa Doctoral Symposium 2024**

**7<sup>th</sup> - 8<sup>th</sup> May 2024**

**Book of Abstracts**



*Book of Abstracts*

11th International BSC Severo Ochoa Doctoral Symposium

*Editors*

Mary K. Chessey, Claudia Cutiño, and Sajjan Singh Mehta

*Cover*

Design based on artwork created by macrovector.com

*This is an open access book registered at UPC Commons*  
([upcommons.upc.edu](https://upcommons.upc.edu)) under a Creative Commons license to protect its  
contents and increase its visibility.

*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)

11th Edition, May 2024

# ACKNOWLEDGEMENTS

The BSC Education & Training team gratefully acknowledges the PhD students, early career researchers, advisors, postdocs, experts and especially the Keynote Speaker Dr. Mercè Crosas Navarro, for contributing to this Book of Abstracts and for participating in the 11th International BSC Severo Ochoa Doctoral Symposium. We also wish to thank the volunteers that supported the organisation of the event: Jose Miguel Ramirez, Manuel Giménez de Castro Marciani, Anastasia Sukhorukova, Miriam Poley Gil, Iria Pose Lagoa, Olfat Khannous, Maria Sopena Rios, Winona Oliveros, Sergi Masot Llima, Maria Paola Ferri, Eva Martin del Pico, Audrey Mendez-Pratt, Zahra Noori, and Ghazal Rahimi.

BSC Education & Training team  
[education@bsc.es](mailto:education@bsc.es)

# EDITORIAL COMMENT

We are proud to present the Book of Abstracts for the 11th International BSC Severo Ochoa Doctoral Symposium.

With the recent inauguration of the MareNostrum5 supercomputer in December 2023, the BSC and European HPC communities are now able to make use of the 8th most powerful supercomputer in the world and the 6th greenest. For 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 11th edition of the International BSC Severo Ochoa Doctoral Symposium we are offering a keynote talk titled "New Horizons for HPC Applied to Social Sciences and Humanities " by Mercè Crosas Navarro.

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

- Human Health
- Applications of Computational Methods
- Simulations and Modelling
- Genomics
- Earth Sciences & Transcriptomics
- Computer Architecture and Performance.

The posters will be exhibited and presented during four 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 students, researchers, advisors and experts to the 11th International BSC Severo Ochoa Doctoral Symposium.

Once again, in this 11th 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 theses that use High Performance Computing in some way. The Symposium was conceived in the framework of the Severo Ochoa Programme at BSC, following the project aims regarding talent development and knowledge sharing, and provides a forum for early career researchers to communicate their findings.

Consequently, I appreciate the support provided by BSC and the Severo Ochoa Center of Excellence Programme that make this event possible.

I am very grateful to the BSC directors for supporting the Symposium, and to group leaders and advisors for encouraging the participation of students in the event. Moreover, I wish to especially thank the keynote speaker Mercè Crosas for her willingness to share her knowledge and expertise with us.

I would also like to thank all early career researchers for their papers and presentations. I wish you all the best for your career and I hope you enjoy this great opportunity to meet other colleagues and share your experiences.

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

Dr. María-Ribera Sancho  
Education & Training Group Leader  
Barcelona Supercomputing Center

# KEYNOTE SPEAKER

**Mercè Crosas Navarro**

Head of Computational Social Sciences Programme  
Barcelona Supercomputing Center



## **New Horizons for HPC Applied to Social Sciences and Humanities**

The social sciences are more needed than ever to understand and address today's world's challenges. The vast amounts of data about human behaviors and interactions, and the second-to-second digital trace we leave behind, together with the advances of applied computational methods and compute resources, have given rise to the emergence of the new field of computational social sciences. Democratic quality and media consumption, changing demographics and living arrangements, social and ecological values in the digital world, social innovation to improve social mobility, understand what works in education and in science, and what doesn't, explore the volumes and volumes of medieval text and document our cultural heritage, are social science and humanities problems that we are aiming to shed light at the new Computational Social Science program at the Barcelona Supercomputing Center, with the use of a wide variety of data, from text, images, and massive statistical datasets, a wide variety of methods including NLP/LLMs for text analysis, Social Network Analysis, Agent Based Models for social simulations, multilevel statistical models, among other computational-intensive methods. This talk will introduce these initial projects, and describe the vision, goals, and structure of the Computational Social Science program, as well as the importance of responsible access and use of FAIR data and of conducting this computational research with open science in mind.

Mercè Crosas is a scientist and technologist at the Barcelona Supercomputing Center (BSC) focused on computational and data science, data sharing, and open and FAIR data (Findable, Accessible, Interoperable, Reusable). Since the beginning of 2023, she has been the Head of the Computational Social Sciences Program at the BSC, a new program that aims to facilitate the use of data and computing in the social sciences and humanities and advance new computational research in these domains. Crosas is also the President of CODATA, the Committee on Data of the International Science Council, and is an affiliate of the Institute for Quantitative Social Science at Harvard University.

Crosas has spent most of her professional life at Harvard University, first as an astrophysicist and a scientific software engineer, and recently as the Chief Data Science and Technology Officer at the Institute for Quantitative Social Sciences and the University Research Data Management Officer. Prior to her current position, from 2021 to 2022, Crosas was Secretary of Open Government at the Generalitat de Catalunya (Government of Catalonia), where she was responsible for open data, transparency, and citizen participation in democracy. She holds a doctorate in Astrophysics from Rice University, a degree in Physics from the University of Barcelona, and was a pre-doctoral and post-doctoral fellow at Harvard University.

## AGENDA

### 7 May 2024

08:30-09:00	Registration
09:00-09:20	Welcome Address
09:20-10:20	Keynote Presentation
10:20-10:40	Meet Attendees and Group Photo
10:40-11:40	POSTER SESSION I (with light refreshments)
11:40-12:40	TALK SESSION I - Human Health
12:40-13:40	Lunch
13:40-15:00	TALK SESSION II - Applications of Computational Methods
15:00-16:00	POSTER SESSION II (with tea & coffee)
16:00-17:00	TALK SESSION III - Simulations and Modelling
17:00	End Day 1

### 8 May 2024

08:50-09:10	Registration
09:10-10:50	TALK SESSION IV - Genomics
10:50-11:50	POSTER SESSION III (with light refreshments)
11:50-13:10	TALK SESSION V - Earth Science and Transcriptomics
13:10-14:10	Lunch
14:10-15:10	POSTER SESSION IV
15:10-16:30	TALK SESSION VI - Computer Architecture and Performance
16:30-17:30	Closing Session and Awards
17:30	Cool Off on the BSC terrace (with light refreshments)

## DAY 1 Presenters and Titles

### 7 May, 10:40-11:40, POSTER SESSION I

- Zahra Noori - *Adjusting UV-Vis Spectrum of Alizarin by Insertion of Auxochromes*
- Fatemeh Baghdadi - *Computational Methods for the Integration of Multimodal Cardiovascular Data*
- Giacomo Mutti - *Detecting non-vertical inheritance across eukaryotes*
- Sergi Palomas - *Evaluating computational performance metrics in Climate modelling: Insights from CMIP6*
- Maria Sopena Rios - *Single-cell atlas of the aging immune system*
- Marc Solé i Bonet - *The METASAT Hardware Platform v1.1: Identifying the Challenges for its RISC-V CPU and GPU Update*

### 7 May, 11:40-12:40, TALK SESSION I - Human Health

- Guillermo Prol Castelo - *A Benchmark of Synthetic Transcriptomic Cancer Data Reconstruction*
- Sara Peregrina Cabredo - *Implications of the human oral microbiome in Alzheimer's disease prognosis*
- Alejandro Navarro Martínez - *Evaluating the Impact of Recurrent Mobility in Air Pollution Exposure in Catalonia*

## DAY 1 *continued*

### 7 May, 13:40-15:00, TALK SESSION II - Applications of Computational Methods

- Miriam Poley Gil - *Exploring the biophysical boundaries of protein families with deep learning methods*
- Roc Farriol-Duran - *Design of a surface-accessible epitope panel using Brewpitopes to empower early lung cancer detection*
- Lidia Neves - *Big data and diversity: the specificities of analyzing discourses about refugees in Brazil*
- Varbina Ivanova - *Multiple-Copies Association Studies for Computational Binding Mode Elucidation of Fbw7 E3 Ligase Fragment Hits*

### 7 May, 15:00-16:00, POSTER SESSION II

- Jeremy Jens Giesen León - *ASCOM: Affordable Sequence-aware COntention Modeling in Crossbar-based MPSoCs*
- Ivan Vargas Valdivieso - *Design and Analysis of a Processing-in-Memory Sort Algorithm using UPMEM*
- Júlia Vilalta Mor - *Generative Strategies for Multi-target Drug Design: Generating Mpro Pan-inhibitors*
- Louis Ledoux - *LLMMMM: Large Language Models Matrix-Matrix Multiplications Characterization on Open Silicon*
- Iria Pose Lagoa - *Machine Learning Approaches for the Characterization of COPD*
- Hernán Domingo Ramos - *MAXWEL: Simulation of EM wave propagation in plasma using FEM*

### 7 May, 16:00-17:00, TALK SESSION III - Simulations and Modelling

- Patricia Blanco Gabella - *Development of Novel Non-Peptidic VHL Binders Using a Fragment-Based Approach*
- Miguel Luengo - *Designing a new generation of industrial proteases*
- Fernando Vázquez - *Methodologies for the Design and Development of Digital Twins*

## DAY 2

### 8 May, 09:10-10:50, TALK SESSION IV - Genomics

- Alejandro Alonso Marín - *BIMSA: Accelerating Long Sequence Alignment Using Processing-In-Memory*
- Winona Oliveros - *Inter-individual and Inter-tissue variation of DNA methylation*
- Saioa Manzano-Morales - *Reconstructing prokaryotic metabolic contributions to the Last Eukaryotic Common Ancestor*
- Moisés Bernabeu - *Horizontal gene transfer help in the process of organellogenesis*



DAY 2 *continued*

8 May, 10:50-11:50, POSTER SESSION III

- Júlia Orteu Aubach - *A Framework and Methodology for Performance Prediction of HPC Workloads*
- Carlos Rojas - *Agile and accurate microarchitecture modeling using Python and Salabim*
- Laura Ventura San Pedro - *Deep-learning-enhanced transcriptomic and histopathology analysis of the role of aging in female tissues*
- Ignasi Puch Giner - *Drug Repurposing for Mammalian Heart Regeneration: The Inhibition Mechanism of Neomycin and Paromomycin on Meis1-Hoxb13-DNA Trimer*
- Álvaro Redondo del Río - *Hybridisation in Emerging Fungal Pathogens*
- Pol Garcia Recasens - *Towards Pareto Optimal Throughput in Small Language Model Serving*

8 May, 11:50-13:10, TALK SESSION V - Earth Science and Transcriptomics

- Alba Santos Espeso - *Cooling Effect of Aerosols on Past Arctic Climate (1950-2014)*
- Pep Cos - *Atmospheric circulation leading to subtropical air intrusions in the Western Mediterranean*
- Rubén Chazarra Gil - *Alternative Splicing variability between human populations at single-cell resolution*
- Pau Clavell Revelles - *A pipeline to preprocess long reads Oxford Nanopore sequencing data to reveal the transcriptomic diversity of human populations*

8 May, 14:10-15:10, POSTER SESSION IV

- Victor Xirau Guardans - *Active Compute Memory: Enhancing Memory and Processing in Near-Memory Architectures for Vector Classification*
- Xavier Benedicto Molina - *Analysing Metabolic Changes in a Gastric Adenocarcinoma Cell Line under Different Drug Treatments*
- Lluís Frontera Perelló - *Disentangling inter-individual transcriptome variability at single-cell and pseudobulk resolution*
- Marko Ludaic - *Emergence of Energetic Constraints over the Evolution of Protein Families*
- Pau Manyer Fuertes - *EQUILI module in ALYA: a free-boundary Grad-Shafranov equation solver using CutFEM*
- Othmane Hayoun - *An AGS cell line digital twin for studying novel treatment strategies*

8 May, 15:10-16:30, TALK SESSION VI - Computer Architecture and Performance

- Alejandro Cano Cos - *Buffer architecture for Dragonfly topologies*
- Josep Pocerull Serra - *Performance analysis of an OpenFOAM HPC Grand Challenge using BSC performance tools*
- Sergi Laut Turon - *Architecture-aware Patterns for the Factorized Sparse Approximate Inverse Preconditioner*
- Elias Perdomo - *HBM performance on FPGAs*

The background of the page is filled with a variety of light gray line-art icons. These icons include lightbulbs (some with radiating lines to indicate they are lit), puzzle pieces, globes, and stylized human figures with their arms raised in a gesture of triumph or excitement. The icons are scattered across the entire page, creating a dense, conceptual pattern.

# Abstracts

# BIMSA: Accelerating Long Sequence Alignment Using Processing-In-Memory

Alejandro Alonso-Marín\*, Ivan Fernandez\* Santiago Marco-Sola\*

\*Barcelona Supercomputing Center, Barcelona, Spain

E-mail: {alejandro.alonso1, ivan.fernandez, santiago.marco}@bsc.es

**Keywords**—*Sequence Alignment, Wavefront Alignment Algorithm (WFA), Performance Analysis, Memory Bound, Hardware Acceleration, Processing-In-Memory (PIM)*

## I. EXTENDED ABSTRACT

The alignment of DNA, RNA, and protein sequences is fundamental for understanding multiple biological processes, including identifying genetic variations that can lead to diseases and health conditions. Recent advances in sequencing technologies have allowed the production of larger and more accurate sequences, reaching up to reads of 100K bases in some experiments.

Sequence alignment algorithms based on Dynamic Programming (DP), such as Needleman-Wunsch [1] and Smith-Waterman [2], [3], are widely used and effective for aligning relatively short sequences. However, this approach presents performance challenges when dealing with ultra-long sequences and large-scale datasets. In particular, DP algorithms present a time and space complexity that grows quadratically with the length of the sequences being aligned. Hence, these algorithms are not able to scale when aligning large-scale sequence datasets generated by modern sequencing technologies.

To overcome these challenges, WFA [4] is a novel state-of-the-art algorithm that improves over DP-based solutions by taking advantage of homologous regions between sequences to accelerate the alignment’s computation. In essence, WFA computes optimal alignments in  $O(ns)$  time and  $O(s^2)$  memory, where  $n$  is the sequence length and  $s$  is the optimal alignment score. Recently, BiWFA [5] improved the original WFA algorithm by maintaining compute time at  $O(ns)$  complexity but reducing memory space to linear  $O(s)$  complexity. However, BiWFA’s memory requirements still prevent memory-restricted architectures from performing ultra-long sequence alignment. We thoroughly characterize BiWFA in terms of performance and find that it remains a memory-bound algorithm, especially when aligning long and noisy sequences.

In this context, Processing-In-Memory (PIM) represents an emerging paradigm that seeks to alleviate the data movement bottleneck in conventional platforms. The key idea behind PIM is to place compute units as close as possible to the actual location where data resides (i.e., memory). There are already available PIM devices in the market that implement the PIM paradigm, such as the general-purpose UPMEM [6] platform.

Our **goal** in this work is to *enable fast and energy-efficient alignment of sequences of different lengths*, ranging from short sequences to long sequences. To this end, we present BIMSA, the first PIM-enabled memory-efficient alignment implementation that exploits the linear memory features of

BiWFA to efficiently perform sequence alignment in the PIM units of the UPMEM platform.

## A. BIMSA Implementation

BIMSA follows a coarse-grain parallelization scheme by assigning several sequence pairs to each thread inside each UPMEM compute unit. This is feasible since applications that require sequence alignment usually involve comparing thousands or millions of sequence pairs. Each of these alignments can be computed independently, using their own working memory space. We discard a collaborative parallelization scheme since 1) thread synchronization and communication operations are costly (1000s of cycles), 2) communication between compute units has to be done through the host CPU, and 3) BiWFA’s parallelism increases progressively during the alignment of a pair of sequences. Therefore, at the beginning of every alignment computation, there is minimal parallelism, which may underutilize the platform. Based on that, BIMSA’s coarse-grain parallelization approach is the best fit, removing the need for threads to synchronize or share data across compute units.

Adapting BiWFA to the UPMEM programming paradigm poses specific challenges, primarily due to its recursive nature. To tackle these challenges, we opt for a BiWFA-based application that splits sequences using breakpoints, generating additional BiWFA tasks through iteration until a threshold is reached, removing recursion completely. Finally, a WFA task is generated to finalize the alignment process.

BIMSA leverages two key properties of BiWFA. First, the breakpoint iteration identifies a coordinate along the optimal path where sequences can be effectively split, separating them into two sections with balanced error. Second, each breakpoint provides the distance for the iteration alignment.

We employ the BiWFA algorithm to iterate through sequences, breaking them down into sub-problems until their distances generate a WFA task that fits in the UPMEM’s compute unit scratchpad. These subproblems, (aka *base cases*), employ the regular WFA algorithm in which memory space is dictated by the error rate ( $O(s^2)$ ). Thus, we can precisely determine the memory required for a regular WFA within a sub-problem. Once we identify a suitable error distance, we execute the WFA, yielding a partial alignment.

## B. Experimental Environment

To evaluate and compare the performance of BIMSA, we select the state-of-the-art CPU application `WFA2lib` and the PIM pairwise alignment application (`AIM` library [7]), which implements a gap-affine optimized Needleman-Wunch (NW)

algorithm and the classical WFA algorithm for PIM. All the applications are executed in a UPMEM node equipped with PIM-enabled DIMM modules, which have up to 2556 operative compute units (DPUs) running at 350 MHz. Each DPU can handle up to 24 threads and is equipped with an MRAM (64 MB), a WRAM (64 KB), and an IRAM (24 KB). The node has two Intel Xeon Silver 4215 CPUs with a total of 16 hardware threads. We use simulated datasets in our evaluation.

### C. Results

Figure 1 illustrates the alignments per second achieved by the state-of-the-art and BIMSA across different balanced workload datasets. To compare performance, each bar displays the speedup needed to match the fastest application.

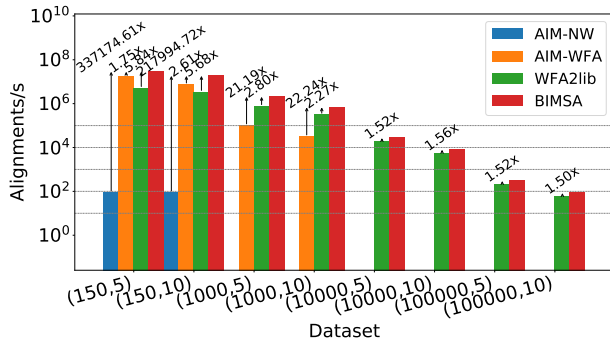


Fig. 1. Alignments per second achieved by BIMSA, WFA2lib (CPU), and AIM when aligning balanced workload datasets. The datasets are indicated by (length,error%). Note that the y axis follows a log scale.

First, we observe that BIMSA outperforms the state-of-the-art PIM (AIM-WFA) across all datasets, achieving up to 22.24× speedup in the best-case scenario. Additionally, while BIMSA handles sequences of 10000 bases or longer, the state-of-the-art PIM (AIM-WFA) is unable to handle them due to insufficient memory allocation space. Second, BIMSA outperforms the state-of-the-art CPU (WFA2lib) by up to 5.84× across all datasets. However, we note that as the sequence length and error percentage increase, BIMSA’s performance degrades, achieving a moderated 1.50× speedup in the worst-case scenario.

**Key Result I:** BIMSA (1) is able to handle long sequence pairs where state-of-the-art PIM fails and (2) provides higher performance than CPU for all datasets.

In Figure 2, we examine the scalability of the same datasets as in Figure 1 across varying numbers of DPUs. A dotted grey line denotes ideal scalability. We observe that BIMSA’s scalability with the number of DPUs is nearly linear, while CPU implementations do not scale well when increasing the number of cores.

**Key Result II:** The linear scalability of BIMSA will translate into significant performance improvements in the next UPMEM systems with a higher number of DPUs.

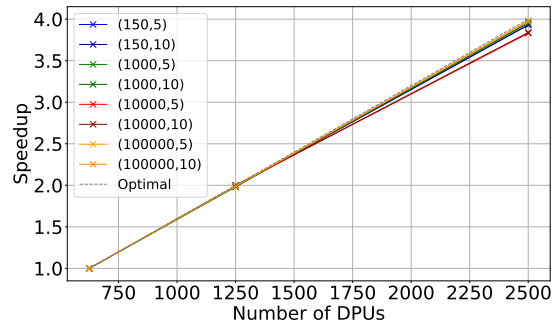


Fig. 2. Scalability with the number of DPUs for BIMSA with balanced workload datasets. The datasets are indicated by (length, error %).

### D. Conclusion

In this work, we introduced a PIM-based implementation of the BiWFA sequence alignment algorithm. BIMSA leverages the real-world PIM architecture provided by UPMEM, and it is designed to address the data-penalty challenges that often hinder the scalability of other sequence alignment implementations. Overall, BIMSA exhibits superior performance than other state-of-the-art sequence alignment implementations aligning datasets of different sequence lengths. Moreover, we observe ideal scalability with the number of compute units. Considering the early stages of technological development of PIM technologies, these results are very promising with plenty of room for improvements in the coming years.

## II. ACKNOWLEDGMENT

We thank UPMEM for both the infrastructure and technical support and BSC for the financing support.

## REFERENCES

- [1] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of molecular biology*, 1970.
- [2] M. S. Waterman, T. F. Smith, and W. A. Beyer, “Some biological sequence metrics,” *Advances in Mathematics*, 1976.
- [3] O. Gotoh, “An improved algorithm for matching biological sequences,” *Journal of molecular biology*, 1982.
- [4] S. Marco-Sola, J. C. Moure, M. Moreto, and A. Espinosa, “Fast gap-affine pairwise alignment using the wavefront algorithm,” *Bioinformatics*, 2021.
- [5] S. Marco-Sola, J. M. Eizenga, A. Guarracino, B. Paten, E. Garrison, and M. Moreto, “Optimal gap-affine alignment in  $o(s)$  space,” *Bioinformatics*, 2023.
- [6] UPMEM, “UPMEM Website,” <https://www.upmem.com>, 2024.
- [7] S. Diab, A. Nassereldine, M. Alser, J. Gómez Luna, O. Mutlu, and I. El Hajj, “A framework for high-throughput sequence alignment using real processing-in-memory systems,” *Bioinformatics*, 2023.



**Alejandro Alonso-Marín** earned his BSc in Computer Engineering from the Universitat Autònoma de Barcelona (UAB) in 2019. Over the next two years, he worked at the ALBA Synchrotron in Barcelona. He completed his MSc in High-Performance Computing at the Universitat Politècnica de Catalunya (UPC) in 2023. Since then, he has been pursuing his PhD in the Computer Science Department at the Barcelona Supercomputing Center (BSC).

# Computational Methods for the Integration of Multimodal Cardiovascular Data

Fatemeh Baghdadi<sup>#1</sup>, Davide Cirillo<sup>\*2</sup>

<sup>#1, \*2</sup> *Barcelona Supercomputing Center (BSC), Barcelona, Spain*

<sup>1</sup>*fatemeh.baghdadi@bsc.es,*

<sup>#1</sup> *Universitat de Barcelona, Barcelona, Spain*

<sup>2</sup>*davide.cirillo@bsc.es*

**Keywords**— **Data fusion, Cardiovascular disease (CVD), Precision medicine, Remote health care**

## INTRODUCTION

The exponential growth in clinical and biological data has significantly advanced healthcare for cardiovascular disease (CVD), which stands as a primary global cause of mortality [1]. The solution to improve CVD healthcare 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. This progress addresses two crucial contemporary needs within the field. Firstly, there's been a shift from merely collecting data towards effectively analyzing and interpreting it. Advanced analytical techniques now enable healthcare professionals to derive actionable insights from complex datasets, enhancing diagnoses, treatment decisions, and patient management strategies. Secondly, precision medicine, which tailors therapies to individual patients based on their genotype and phenotype, has become pivotal [2]. In order to enhance the incorporation of information from multiple sources, the concepts of 'data fusion' and 'data integration' have emerged with the main difference being a focus on combining information in the case of data integration and reducing information in the case of data fusion. In both cases, the combination of data from different modalities that provide separate views on a common phenomenon (e.g. a human disease) promises to solve prediction and classification problems with fewer errors than unimodal approaches [3]. The goal of such strategies is to effectively exploit complementary and interdependence of various modalities and to fully exploit these perspectives, as well as to harness artificial intelligence (AI) and machine learning (ML) methods that can combine structured and unstructured data with different statistical properties and granularities [4].

Data fusion also presents a promising solution to address sex representation and enhance population diversity concerns, including those involving minority populations, in health modelling. By integrating datasets where one type may predominantly represent one sex while another leans towards the other, data fusion can generate a more balanced and representative dataset. This reciprocal compensation capability extends to racial or ethnic diversities as well. By leveraging various datasets, data fusion enables a more comprehensive understanding of health disparities and facilitates the development of more inclusive and equitable healthcare models that cater to the diverse needs of populations.

## OBJECTIVE

The ultimate goal of this research is to develop AI and ML models for effective multimodal data combination, with special emphasis on the temporal dimension that is reflected in the data, specifically observations at a single point in time, such as cross-sectional data, and multiple observations over time, such as longitudinal data. The research focuses on human diseases with high mortality rates, including CVD, cerebrovascular diseases such as stroke, among others.

## METHODOLOGY

The research methodology for cardiovascular disease detection employs a multi-faceted approach integrating multimodal biomedical data and AI fusion models. Initially, in data collection phase the clinical data and recorded electrocardiogram (ECG) collected in the context of the Horizon Europe project AI-SPRINT (Grant agreement ID: 101016577) has been utilized [5]. Life science department of Barcelona Supercomputing Centre (BSC) conducted 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 ECG monitoring up to 30 days. Additionally, large-scale biomedical in-depth genetic and health information database containing diverse sources of clinical records, ECG signals, genetic information, and patient demographics will be utilized in the next phase for the fusion models. Subsequently, in the data pre-processing phase data is being pre-processed and standardized to ensure consistency and compatibility across modalities. The AI-based algorithms, including machine learning and deep learning techniques, are then trained on this enriched dataset to extract relevant features and patterns indicative of cardiovascular disease. Additionally, we employ fusion models to integrate information from multiple modalities, leveraging the complementary nature of diverse data sources to enhance detection accuracy. The fusion process involves combining outputs from individual AI models or developing novel fusion architectures tailored to the specific characteristics of cardiovascular disease data. To evaluate the performance of our methodology, we employ rigorous validation techniques, including cross-validation and independent testing on unseen data. Finally, we interpret the results and assess the clinical utility of our approach, aiming to contribute to the early detection and improved management of cardiovascular disease through advanced AI-driven methodologies and fusion techniques.

### PRELIMINARY RESULT

The initial model has been trained to classify ECGs from the PhysioNet database [6]. Classifiers such as the cascaded support vector machine (CSVM), k-nearest neighbor (KNN), random forest, and convolutional neural network (CNN), using the Dislib library [7] developed by the BSC Computer Science Department, were employed to classify ECGs from the PhysioNet database. The CNN achieved above state-of-the-art 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 (AF)—that is strongly associated with stroke occurrence. The final model detects AF in the ECGs and, if present, uses a standard AF-based stroke risk calculator called CHA2DS2-VASc (congestive heart failure, hypertension, age  $\geq 75$  (doubled), diabetes, stroke (doubled), vascular disease, age 65 to 74, and sex category (female)) [8]. Specifically, the model integrates the AF detection results with the risk scores to create a stroke risk stratification model suitable for use with non-invasive commercial fitness trackers, such as FitBit and Apple smartwatches. Fig. 1 describes the flowchart of the final model.

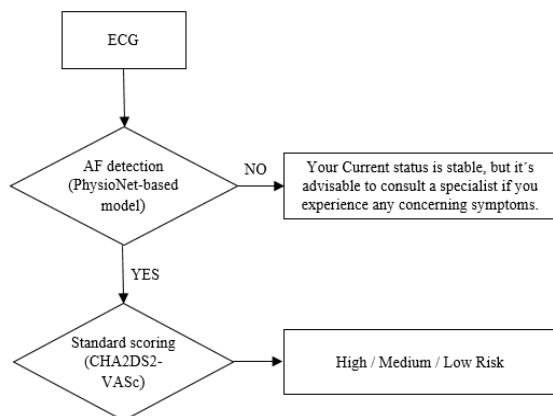


Fig. 1: Final model flowchart

### Future Direction

In future directions, the research aims to integrate genomic analysis into the existing model, enhancing its predictive capabilities and furthering our understanding of cardiovascular health. By fusing information from ECGs, clinical data, and genomic profiles, the model will provide a more comprehensive assessment of cardiovascular disease risk and prognosis. This multidimensional approach holds promise for uncovering novel biomarkers, elucidating disease mechanisms, and ultimately advancing personalized medicine interventions tailored to individual patient profiles. By leveraging the synergy of diverse data modalities, future iterations of the model seek to revolutionize cardiovascular risk assessment and treatment strategies, ultimately improving

patient outcomes and promoting proactive cardiovascular health management.

### References

- [1] Ahmad FB, Cisewski JA, Xu J, Anderson RN. Provisional Mortality Data — United States, 2022. *MMWR Morb Mortal Wkly Rep* 2023;72:488–492. DOI: <http://dx.doi.org/10.15585/mmwr.mm7218a3>
- [2] Leopold, J. A., Maron, B. A., & Loscalzo, J. (2020, January 2). The application of big data to cardiovascular disease: Paths to precision medicine. *The Journal of Clinical Investigation*. Retrieved February 15, 2023, from <https://www.jci.org/articles/view/129203>
- [3] Sören Richard Stahlschmidt, Benjamin Ulfenborg, Jane Synnergren, Multimodal deep learning for biomedical data fusion: a review, *Briefings in Bioinformatics*, Volume 23, Issue 2, March 2022, bbab569, <https://doi.org/10.1093/bib/bbab569>
- [4] Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;19(2):325–40.
- [5] Baghdadi, F., Cirillo, D., Lezzi, D., Lordan, F., Vazquez, F., Lomurno, E., ... & Matteucci, M. (2024). Harnessing the Computing Continuum across Personalized Healthcare, Maintenance and Inspection, and Farming 4.0. arXiv preprint arXiv:2403.14650.
- [6] Clifford GD, Liu C, Moody B, Li-wei HL, Silva I, Li Q, Johnson AE, Mark RG. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC) 2017 Sep 24 (pp. 1-4)*. IEEE. <https://doi.org/10.22489/CinC.2017.065-469>
- [7] J. Álvarez Cid-Fuentes, S. Solà, P. Álvarez, A. Castro-Ginard, and R. M. Badiá, “dislib: Large Scale High Performance Machine Learning in Python,”
- [8] Lip, G. Y., Nieuwlaat, R., Pisters, R., Lane, D. A., & Crijns, H. J. (2010). Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2), 263–272. <https://doi.org/10.1378/chest.09-1584>

### 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. Her current research interests include artificial intelligence in biomedical research, biomedical data fusion, Bio signal analysis.

# Analysing Metabolic Changes in a Gastric Adenocarcinoma Cell Line under Different Drug Treatments: A Constraint-Based Modeling Approach

Xavier Benedicto Molina\*<sup>†</sup>, Miguel Ponce-de-León\*

\*Barcelona Supercomputing Center (BSC), Barcelona, Spain

<sup>†</sup>Universidad Autónoma de Madrid, Madrid, Spain

E-mail: {xavier.benedicto, miguel.ponce}@bsc.es

**Keywords**—*Gastric carcinoma, AGS, drug synergies, metabolic tasks, genome-scale metabolic models.*

## I. EXTENDED ABSTRACT

Gastric carcinoma (GC) is one of the leading causes of cancer death globally. Managing advanced stages of GC necessitates a multifaceted approach, encompassing surgical interventions and comprehensive multidisciplinary strategies such as combinatorial drug treatments, recently emerging as promising avenues. Recent advances in computational biology have facilitated the development of genome-scale metabolic models (GEMs), which offer a comprehensive, mathematical framework for understanding the intricate metabolic dynamics. Integrating these GEMs with high-throughput omics data, GEMs provide a representations of cellular metabolism through metabolic tasks, allowing for the elucidation of complex metabolic phenotypes associated with the effects of combinatorial treatments. Herein, an approach to take advantage of metabolic tasks and high-throughput omics data using the previously described Tasks Inferred from Differential Expression (TIDEs) approach in conjunction with vital genes to metabolic tasks (anchor genes) is presented as ag-TIDEs.

The aim of this work is to explore aberrant metabolic phenotypes using different approaches to generate hypotheses on the mechanisms underlying the observed synergies in the combinatorial drug treatments of TAK1, MEK, and PI3K inhibitors on cells from the AGS cell line. Preliminary results propose distinct metabolic alterations induced by the drug treatments as a hypothesis to be further explored.

### A. Background

To this date, there are two main approaches to elucidate enriched functions with expression data: (i) gene-centric approaches, which use expression data directly from genes to estimate significant changes in already defined biological pathways, gene sets, or ontology terms (Subramanian et al., 2005; Huang et al., 2009; Kolberg et al., 2023); and (ii) reaction-centric approaches, which are based in genome-scale metabolic models of organism that integrate expression data directly onto metabolic reactions to infer significance (Richelle et al., 2021; Dougherty et al., 2021). Genome-scale metabolic models (GEMs) are valuable tools to study metabolism with many applications such as to elucidate malignant transformation (Yizhak et al., 2015) or to infer drug perturbations (Yizhak et al., 2014). Other kinds of models have been proposed,

although at a smaller scale than a GEM and for specific human cell lines, designed to elucidate drug synergies from a logical decision network (Flobak et al., 2015). These models rise from the need to evaluate the combination of hundreds of different drug treatments in order to discover new synergistic treatments (Robinson & Nielsen, 2017).

In this research project, the aim is to explore the metabolic changes that may occur in the gastric adenocarcinoma cell line AGS when exposed to different drugs (Flobak et al., 2015). Specifically, constraint-based modeling methods will be employed with the integration of gene expression data from various experimental conditions (control, individual drug and combinatorial drug treatments) to reconstruct metabolic phenotypes corresponding to each of the specific experimental conditions (Robinson et al., 2020; Richelle et al., 2021; Ponce-de León et al., 2019).

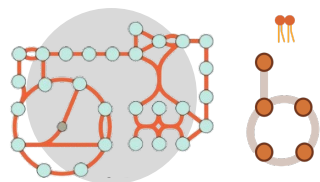
### B. Methods

Using constraint-based metabolic modeling methods, essential or anchor genes can be deduced from mathematical representations of metabolic functions (i.e., metabolic tasks). Thus, an anchor gene is defined as essential to its metabolic task and, when they are knocked-out using *in silico* methods, the task becomes mathematically infeasible. In this research project, 187 metabolic tasks were obtained from Richelle et al., 2021 and curated to the latest available human metabolic model Human1 (Richelle et al., 2021; Robinson et al., 2020). Anchor genes were computed for each metabolic task, and using an approach based on the already published TIDE method (Dougherty et al., 2021) in conjunction with the anchor genes (ag-TIDE), the metabolic phenotype of the different experimental conditions was inferred.

### C. Conclusion

PD (MEK inhibition) perturbations appeared as an impairment in nucleotide synthesis and recycling, whereas PI (PI3K inhibition) displayed alterations in the metabolism of some amino acids that could be linked to an imbalance in oxidative stress protection. Finally, a combination of alterations within different metabolic systems is proposed as a possible hypothesis that could explain the observed synergistic effects in AGS cell growth.

Metabolic tasks are identified and validated, along with their anchor genes associated



Log-FC from anchor genes are recovered to calculate task score

	Log-FC	Random		Log-FC	Random
Gene I	1.57	0.04	Gene IV	1.34	0.04
Gene II	-0.93	0.11	Gene V	2.87	0.67
Gene III	-1.20	0.31	Gene VI	0.87	-0.12

○ Gene    ● Anchor gene

**Task score: 1.77    Random score: 0.19**

Statistical significance for each task compared to randomized data

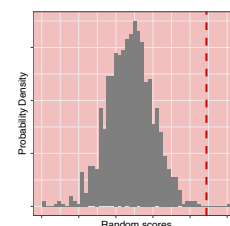


Fig. 1. Pipeline summary of ag-TIDEs.

## II. ACKNOWLEDGMENTS

This project could not have been finished without the valuable help of Åsmund Flobak, who kindly sent the data for the AGS experiments, and Miguel Ponce-de-León, who provided with the experience and guidance necessary to complete project.

## REFERENCES

- Blais, E. M., Rawls, K. D., Dougherty, B. V., Li, Z. I., Kolling, G. L., Ye, P., ... Papin, J. A. (2017). Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. *Nature Communications*, 8. doi: 10.1038/ncomms14250
- Dougherty, B. V., Rawls, K. D., Kolling, G. L., Vinakota, K. C., Wallqvist, A., & Papin, J. A. (2021). Identifying functional metabolic shifts in heart failure with the integration of omics data and a heart-specific, genome-scale model. *Cell Reports*, 34. doi: 10.1016/j.celrep.2021.108836
- Flobak, Å., Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., & Lægreid, A. (2015). Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS Computational Biology*, 11. doi: 10.1371/journal.pcbi.1004426
- Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1), 44-57. doi: 10.1038/nprot.2008.211
- Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., & Peterson, H. (2023). g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1), W207-W212. doi: 10.1093/nar/gkad347
- Ponce-de León, M., Apaolaza, I., Valencia, A., & Planes, F. J. (2019). On the inconsistent treatment of gene-protein-reaction rules in context-specific metabolic models. *Bioinformatics*, 36(6), 1986-1988. doi: 10.1093/bioinformatics/btz832
- Richelle, A., Kellman, B. P., Wenzel, A. T., Chiang, A. W., Reagan, T., Gutierrez, J. M., ... Lewis, N. E. (2021). Model-based assessment of mammalian cell metabolic functionalities using omics data. *Cell Reports Methods*, 1. doi: 10.1016/j.crmeth.2021.100040
- Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P.-E., Cook, D., Nilsson, A., ... Nielsen, J. (2020). An atlas of human metabolism. *Sci. Signal*, 13, 1482.
- Robinson, J. L., & Nielsen, J. (2017). Anticancer drug discovery through genome-scale metabolic modeling. *Current Opinion in Systems Biology*, 4, 1-8. (Big data acquisition and analysis • Pharmacology and drug discovery) doi: 10.1016/j.coisb.2017.05.007
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*.
- Swainston, N., Smallbone, K., Hefzi, H., Dobson, P. D., Brewer, J., Hanscho, M., ... Mendes, P. (2016). Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, 12. doi: 10.1007/s11306-016-1051-4
- Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., ... Palsson, B. O. (2013). A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, 31, 419-425. doi: 10.1038/nbt.2488
- Yizhak, K., Chaneton, B., Gottlieb, E., & Ruppin, E. (2015). Modeling cancer metabolism on a genome scale. *Molecular Systems Biology*, 11(6), 817. doi: 10.15252/msb.20145307
- Yizhak, K., Gaude, E., Le Dévédec, S., Waldman, Y. Y., Stein, G. Y., van de Water, B., ... Ruppin, E. (2014). Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife*, 3(Nov 21), e03641. doi: 10.7554/eLife.03641



**Xavier Benedicto** received his BSc in Biomedical Sciences from the Autonomos University of Barcelona (UAB) in 2022. Next year, he moved to Madrid to pursue his MSc in Bioinformatics and Computational Biology from the Autonomous University of Madrid (UAM). Since September 2023, he had been undergoing his master's internship at the Computational Biology group of Barcelona Supercomputing Center (BSC). Recently, he finished his master's thesis and started working as a Junior Research Engineer in the same group.



# Horizontal gene transfer help in the process of organellogenesis

Moisès Bernabeu\*<sup>†</sup>, Toni Gabaldón\*<sup>†‡§</sup>

\*Barcelona Supercomputing Center (BSC)

<sup>†</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona

<sup>‡</sup>Catalan Institution for Research and Advanced Studies (ICREA), Barcelona

<sup>§</sup>Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona

E-mail: {moises.bernabeu, toni.gabaldon}@bsc.es

**Keywords**—*Horizontal Gene Transfer, Organellogenesis, Paulinella micropora, HGT relative dating.*

## I. EXTENDED ABSTRACT

Symbiosis is an evolutionary strategy that has helped increase the complexity of life. It was the mechanism by which mitochondria and plastids originated and the one that generated the origin of eukaryotes. Despite the role of symbiosis in the origin of cellular organelles (organellogenesis), how an endosymbiont becomes an organelle is poorly understood. *Paulinella* species are unicellular eukaryotes that harbour a photosynthetic organelle named chromatophore. It was acquired independently and much more recently than the chloroplast, the photosynthetic organelle of plants. Here, we aim to understand how other bacterial symbionts may have helped during the chromatophore origin by analysing the history of horizontal gene transfers (HGT) in the genome of this organism.

### A. The origin of endosymbiotic organelles and the role of HGT

The origin of complex cells (eukaryotes) had a crucial step, the origin of mitochondria. This organelle originated when an alpha-proteobacteria entered and established as an endosymbiont in the proto-eukaryotic host [1]. In the initial stages, the endosymbiont was more bacterial-like, through the loss and the horizontal gene transfer (HGT) of some genes to the host nucleus [2], it lost its individuality and became an organelle dependent on the host cell, this process is known as organellogenesis. We observe similarities in the origin of photosynthetic eukaryotes. For a long time, we thought that all the photosynthetic organelles (named plastids) originated from a single event of primary endosymbiosis of a cyanobacterium within a eukaryotic host. Posteriorly, by subsequent secondary endosymbioses, many lineages of the eukaryotic tree of life became photosynthetic [3]. However, the molecular characterisation of the photosynthetic endosymbiont of *Paulinella chromatophora* [4] revealed that there were two primary endosymbioses, and the one in *Paulinella* happened much more recently [5]. These findings posited *Paulinella* species as one of the major models for studying organellogenesis.

Although the bacterial ancestor of the organelles had many genes, current organellar genomes have a low number of genes, 30-91 for mitochondria [6] and 120-130 in the case of

chloroplasts [7], as they were lost or transferred to the nucleus. However, the chromatophore of *Paulinella* still contains several genes, 800. This supports the idea that organellogenesis is still going on in this organism.

An important step during organellogenesis is the HGT from the endosymbiont to the host. HGT is the exchange of genes, apart from sexual recombination, between organisms [8]. HGT is important in driving the genome composition and adaptation to the environment in many bacteria [9]. HGTs from diverse bacteria were previously characterised in *Paulinella* species [10]; however, their impact on organellogenesis remains poorly understood. Here we assess the bacterial origins and functions of some HGTs to understand their role before, during and after the establishment of the chromatophore of *Paulinella*.

### B. Methods

To detect HGT genes, we took all the proteins encoded in the *Paulinella micropora* genome, which we submitted to a homology search against an in-house database comprising eukaryotic, bacterial, archaeal and viral proteins using BLAST. From the homology results, we selected some precandidates based on high proportions of non-eukaryotic homologs.

We performed a phylogenetic tree for each of these pre-candidates to seek the origin of the *P. micropora* sequence. We assumed that a *P. micropora* sequence originated from an HGT event when the sequences next to it were mostly non-eukaryotic. Specifically, to identify the donor of these genes, we required that the donor's proportion in the first sister group to the *P. micropora* sequence to be higher than 80%, to be present in the first and the second sister, and finally, the node to have a bootstrap support higher than 85%. Finally, we obtained the relative ages for these gene acquisitions using the stem length, which separates the *P. micropora* sequence from the donor.

### C. Results

The genome of *P. micropora* shows many HGT events from diverse donors. We detected 119 robust HGT events for which we can identify a specific donor. We observe significant contributions from Gammaproteobacteria, Cyanobacteria (the clade from which the chromatophore derives), Bacteroidota, Actinobacteria and Alphaproteobacteria, which are all diverse groups of bacteria (Figure 1a).

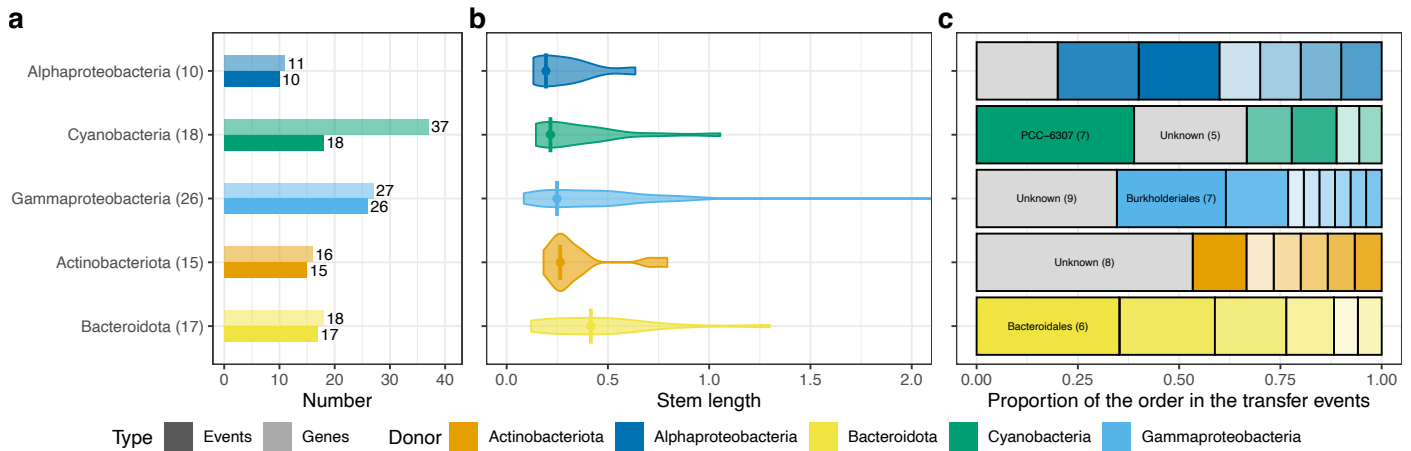


Fig. 1. HGT in *Paulinella micropora*. a) Count of HGT events (darker) and genes (lighter) for each bacterial phylum. b) Distribution of the relative ages of the gene transfer waves, the vertical bar indicates the position of the mode. c) The distribution of the transfer events that could be assigned to a specific order within the bacterial phylum for which we detected a transfer wave.

We sought the functions of the transferred genes, and we did not find a clear association between the function and the bacterial donor. However, we detected genes involved in carbohydrate metabolism from Alphaproteobacteria, signalling genes from Alphaproteobacteria and Bacteroidota, and nucleotide and amino acid metabolism genes transferred from Gammaproteobacteria. This would suggest that the HGT we found may have helped the interconnection between the host and the symbiont.

We observe that Alphaproteobacteria transferred some genes after Cyanobacteria, the clade from which the chromatophore arose. Moreover, we found transfers to *P. micropora* that predated the Cyanobacterial acquisition (Figure 1b). Specifically, Actinobacteria shows considerable transfer waves, and Gammaproteobacteria and Bacteroidota show even more specific donors (Burkholderiales and Bacteroidales, respectively, – Figure 1c). These results reveal that the genome of *P. micropora* is dynamic, and HGT occurred continuously and is probably still occurring.

## D. Conclusion

The results show that *P. micropora* is a dynamic genome that has had and still has a huge amount of HGT. Considering the habitats where *Paulinella* species live and the co-occurrence with many other bacteria, these transfers may help the organism to adapt to the environment and to communicate with the photosynthetic organelle properly. These results also suggest that the genome has adapted ancestral proteins and proteins originated from many other bacteria to the new endosymbiont, facilitating its incorporation as a functional organelle. Thus, HGTs from many bacterial donors help in organellogenesis.

## II. ACKNOWLEDGEMENTS

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

## REFERENCES

- [1] J. Martijn, J. Vosseberg, L. Guy, P. Offre, and T. J. G. Ettema, “Deep mitochondrial origin outside the sampled alphaproteobacteria,”

*Nature*, vol. 557, no. 7703, p. 101–105, May 2018, citation Key: Martijn2018ISBN: 4158601800.

- [2] A. Butenko, J. Lukeš, D. Speijer, and J. G. Wideman, “Mitochondrial genomes revisited: why do different lineages retain different genes?” *BMC Biology*, vol. 22, no. 1, p. 15, Jan 2024.
- [3] P. J. Keeling, “The endosymbiotic origin, diversification and fate of plastids,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1541, p. 729–748, Mar 2010.
- [4] B. Marin, E. C. M. Nowack, and M. Melkonian, “A plastid in the making: Evidence for a second primary endosymbiosis,” *Protist*, vol. 156, no. 4, p. 425–432, Dec 2005.
- [5] L. Delaye, C. Valadez-Cano, and B. Pérez-Zamorano, “How really ancient is paulinella chromatophora?” *PLOS Currents Tree of Life*, Mar 2016. [Online]. Available: <http://currents.plos.org/treeoflife/article/how-really-ancient-is-paulinella-chromatophora/>
- [6] D. Moreira, J. Blaz, E. Kim, and L. Eme, *A gene-rich mitochondrion with a unique ancestral protein transport system*, Feb 2024. [Online]. Available: <http://biorxiv.org/lookup/doi/10.1101/2024.01.30.577968>
- [7] H. Daniell, C.-S. Lin, M. Yu, and W.-J. Chang, “Chloroplast genomes: diversity, evolution, and applications in genetic engineering,” *Genome Biology*, vol. 17, no. 1, p. 134, Jun 2016.
- [8] P. J. Keeling, “Horizontal gene transfer in eukaryotes: aligning theory with data,” *Nature Reviews Genetics*, Jan 2024. [Online]. Available: <https://www.nature.com/articles/s41576-023-00688-5>
- [9] M. Dmitrijeva, J. Tackmann, J. F. Matias Rodrigues, J. Huerta-Cepas, L. P. Coelho, and C. Von Mering, “A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer,” *Nature Ecology Evolution*, Mar 2024. [Online]. Available: <https://www.nature.com/articles/s41559-024-02357-0>
- [10] E. C. M. Nowack, D. C. Price, D. Bhattacharya, A. Singer, M. Melkonian, and A. R. Grossman, “Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of paulinella chromatophora,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 43, p. 12214–12219, Oct 2016.



**Moisés Bernabeu** received his BSc degree in Biology from Universitat de València (UV), in 2020 specialising in phylogenetics. During the following two years, he studied an MSc degree in Biostatistics at the same university and moved to the Barcelona Supercomputing Center (BSC) to finish his MSc thesis. He is currently a PhD student at BSC and Universitat de Barcelona (UB) working on deep-time phylogenomics.

# Development of Novel Non-Peptidic VHL Binders Using a Fragment-Based Approach

Patricia Blanco-Gabella<sup>#1</sup>, Carles Galdeano<sup>#2</sup>, Jordi Juárez-Jiménez<sup>#3</sup>

<sup>#</sup> Department of Physical Chemistry, Faculty of Pharmacy, University of Barcelona, Joan XXIII s/n 08028, Barcelona, Spain

<sup>1</sup>patricia.blanco@ub.edu, <sup>2</sup>cgaldeano@ub.edu, <sup>3</sup>jordi.juarez@ub.edu

Targeted Protein Degradation, Fragment-Based Screening, Multiple-Ligand Association Studies

## EXTENDED ABSTRACT

Targeted Protein Degradation has emerged in recent years as a revolutionary strategy in drug discovery. The von Hippel–Lindau protein (VHL) is a well-validated E3 ligase that is recruited by many efficacious PROteolysis TARgeting Chimera molecules (PROTACs). However, all existing VHL warheads have been developed around a central hydroxyproline unit and, owing to their peptidic nature, exploiting them pharmacologically is challenging due to their poor absorption, distribution, and metabolism.

Here, we apply a fragment-based approach combining computational techniques and ligand-observed NMR studies to discover new chemotypes for VHL ligands that could be developed into more effective drugs. A virtual screening of all the accessible fragment space followed by paramagnetic relaxation enhancement assays allowed us to identify two novel fragment hits for the hydroxyproline binding site. Subsequently, we performed  $\mu$ s-long molecular dynamics simulations of VHL and multiple copies of each ligand to refine the predicted poses of the fragment hits. Finally, these compounds will be used as a starting point for fragment-growing strategies to obtain more potent ligands.

### A. Introduction

The von Hippel–Lindau protein (VHL) is a ubiquitously expressed E3 ligase that recognizes the hydroxylation of a proline of the Hypoxia Inducible Factor 1 alpha (HIF1 $\alpha$ ). VHL has a huge relevance in the Targeted Protein Degradation (TPD) field [1] as it is constitutively active, and multiple potent warheads (low nM affinity) have been developed for this target [2]. Many efficacious PROTAC molecules have been developed using VHL ligands as warheads. However, all of them present limitations because they are created around the central hydroxyproline unit, which confers the molecule a large polar surface and makes it susceptible to rapid (phase II) metabolism [3,4]. Consequently, few VHL PROTACs have reached clinical stages so far [5]. Furthermore, generally VHL-based PROTACs must be administered intravenously, whereas the CRBN-derived molecules can be administered orally. Thus, developing new warheads that improve the properties of VHL-based PROTACs would be of great interest for the field.

### B. Methods

We performed a virtual screening of all the available fragment space with rDock. We used all molecules containing up to 14 heavy atoms from ZINC and Enamine REAL. In addition, we included a pharmacophoric restraint at the hydroxyl position of hydroxyproline. After that, we run Dynamic Undocking simulations (DUck) on the top 10000 molecules applying a work to reach the quasi-bound state (Wqb) threshold of 10kcal/mol. After visual inspection, 19 molecules were selected for experimental evaluation.

The compounds were tested using Paramagnetic Relaxation Enhancement NMR. The signals of the fragment's protons

were compared in the presence and absence of the VHL ligand VH032 to detect the binders to the hydroxyproline site.

The hits were further studied *in silico* following a Multiple-Ligand Association Studies (MAS) approach to determine the most probable binding site. We performed 5 replicas of 1 $\mu$ s-long molecular dynamics simulations with 20 copies of the fragment. A DBSCAN clustering and a lifetime analysis were executed on the ligands to obtain the long-lived poses.

To search for analogues and bigger molecules containing the fragment hits substructure in the Enamine REAL space we used the SpaceLight and SpaceMACS programs. We performed tethered docking of these analogues using rDock.

### C. Results and Future Work

A virtual screening followed by Dynamic Undocking allowed us to select 19 fragments for experimental testing. We were able to identify two fragment hits, **1** and **5**, that showed a modest affinity for VHL in a Paramagnetic Relaxation Enhancement NMR assay. These fragments are novel chemotypes that displayed a stronger response than the L-hydroxyproline core on its own.

We used MAS to predict the most probable binding sites, and the most populated cluster for each fragment was located at the VH032 binding site. In the case of **5**, the predicted pose corresponds with the original docking pose, whereas **1** seems to bind on the right-hand side of the pocket (Fig.1). These predictions agree with the NMR data we obtained.

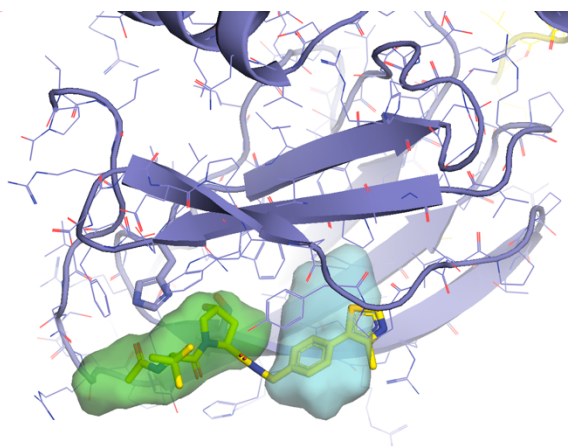


Fig. 1 Superposition of VH032 (PDB: 4w9h) and the surface of the predicted poses for **1** (blue) and **5** (green).

Finally, in order to further develop these promising fragments into better VHL warheads, we did a substructural search on Enamine REAL space to grow them to small molecules of up to 30 heavy atoms. A second virtual screening tethering the atoms of these compounds to the predicted poses was run to select the next round of molecules that will be tested experimentally.

## Acknowledgements

We thank Xavier Barril, who co-supervises this work. We are also thankful to Alessio Ciulli's group for the experimental validation, especially Kevin Haubrich, who has performed the NMR assays mentioned here. This work was funded by the research project PID2021-127693OB-I00 financed by the Ministry of Science and Innovation and the National Research Agency and by a FI-SDUR grant (2021 FISDU 00367). Access to HPC facilities was granted by the RES (project BCV-2023-1-0010).

## References

- [1] M. Békés, D.R. Langley and C.M. Crews, "PROTAC targeted protein degraders: the past is prologue". *Nat Rev Drug Discov*, vol. 21, pp. 181–200, Jan 2022.
- [2] J. Frost et al., "Potent and selective chemical probe of hypoxic signalling downstream of HIF- $\alpha$  hydroxylation via VHL inhibition", *Nat Commun*, vol. 7, pp. 13312, Nov 2016.
- [3] A. Pike, B. Williamson, S. Harlfinger, S. Martin, and D.F. McGinnity, "Optimising proteolysis-targeting chimeras (PROTACs) for oral drug delivery: a drug metabolism and pharmacokinetics perspective", *Drug Discov Today*, vol. 25, pp. 1793–1800, Oct 2020.
- [4] B. Castellani et al., "VHL-Modified PROteolysis TArgeting Chimeras (PROTACs) as a Strategy to Evade Metabolic Degradation in In Vitro Applications", *J Med Chem*, vol. 66, pp. 13148–13171, Sep 2023.
- [5] S. Chirmomas, K.R. Hornberger, C.M. Crews. "Protein degraders enter the clinic — a new approach to cancer therapy" *Nat Rev Clin Oncol*, vol.20, pp. 265–278, Feb 2023.

## Author biography



**Patricia Blanco Gabella** is a PhD student co-supervised by Dr. Jordi Juárez-Jiménez and Prof. Xavier Barril. She completed her bachelor's degree in Biotechnology at the Polytechnic University of Madrid in 2018 and afterwards, did a master's degree in Computational Biology at the same university. When she finished her master's, she carried out a 6-month Erasmus+ traineeship in Thierry Langer's group at the University of Vienna and, after that, she worked at Roche as Linux System Administrator for 1.5 years. Finally, in 2022 she decided to join Xavier Barril and Jordi Juárez's research group at the University of Barcelona and enrol in the Biomedicine PhD program. Her PhD project aims to develop a computational workflow for the rational design of molecular glues using molecular dynamics simulations, Markov State Models analysis and virtual screenings. During her PhD, she also did a research stay at Alessio Ciulli's lab (Dundee, UK) to perform biophysical assays on the projects.

# Buffer architecture for Dragonfly topologies

Alejandro Cano\*, Cristóbal Camarero\*, Carmen Martínez\*, Ramón Bevide \*

\* Universidad de Cantabria, Santander, Spain

E-mail: {canoca, cristobal.camarero, carmen.martinez, ramon.bevide}@unican.es

**Keywords**—Interconnection networks, buffer architecture, routing, Dragonfly topology

## I. EXTENDED ABSTRACT

An interconnection network comprises routers, links, and servers, and the arrangement of these elements is known as network topology. Various network topologies have been employed for supercomputers, including Dragonfly [1], Dragonfly+, Fat-tree, etc. Each topology presents advantages and disadvantages in terms of performance, cost, and scalability, with different routing mechanisms employed for each one.

The Dragonfly topology is composed of a *global* complete graph connecting super-nodes or groups of switches, which, in turn, are connected by means of *local* complete graphs. Thus, it is a two-level hierarchical network with diameter three, or the distance between any pair of switches is three. A small instance of a Dragonfly topology is showed in Figure 1. The Frontier supercomputer, currently number one in the TOP500 list, employs a Dragonfly of this type.

Apart from the topology, another crucial part in the design of interconnection networks is the routing mechanism. Usually, adaptive routing is employed so a packet can choose between following a minimal route, minimizing the number of hops of a packet over the network, or a non-minimal route, which usually follows the Valiant routing scheme. Specifically, in Dragonfly, minimal paths are of type *local-global-local* (hereinafter referred to as *lg1*). These paths start with a local link inside the source group, then a global link to achieve the destination group, and finally a local link to arrive to the destination switch; with any of the links being possibly omitted depending on the relative locations of the source and destination switches.

For non-minimal routes, initially, a Valiant routing variant was proposed. This variant selects an intermediate group of the Dragonfly to pass through before approaching the destination switch, making *lg-1g1* routes. However, it was discovered that Valiant routes should select a switch of the network instead of a group, due to over-subscription of the local links in some patterns, and therefore *lg1-1g1* routes should be used, composing Valiant routes of two equal phases [2].

Also, the routing mechanism should be designed to not to introduce cyclic buffer dependencies in the network, which could lead to packet-deadlock. Therefore, a buffer architecture should be employed to avoid this phenomenon, which is responsible for assigning a buffer (or Virtual Channel, VC) to each packet at each hop across the network.

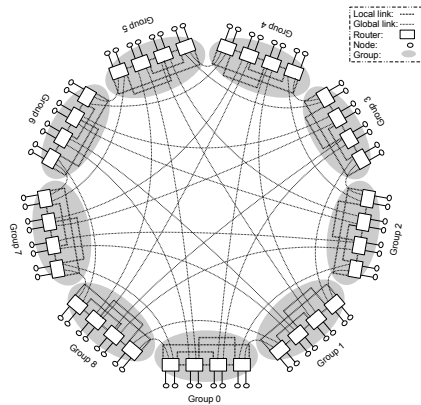


Fig. 1. Small instance of a Dragonfly topology.

## A. Motivation

In a high-loaded lossless interconnection network, a highly demanded portion of the network can propagate its congestion to the entire system, leading to instabilities in throughput, and poor performance.

Network congestion appears due to existing unfairness between network flows, and HoLB (Head of Line Blocking). This unfairness, leaves some servers in starvation, while others get a big portion of the capacity of the network.

Some mechanisms, act when these scenarios appear by typically notify sources to reduce injection rates, thereby leading to a decrease in load at bottlenecks. However, both unfairness and HoLB are directly related to the buffer architecture employed in the network, so buffer architecture can be studied to avoid the appearance of network congestion without the need of throttling the injection rate of the sources.

## B. Buffer Architecture

In Dragonfly networks, minimal paths (*lg1*) necessitate 2 VCs on local ports, visited in order, and 1 VC on global ports to prevent deadlock. To support non-minimal Valiant routing in Dragonfly (DF) topology, two mechanisms could be distinguished: Two Phases and Ladder.

A Valiant routed packet firstly go to a random intermediate switch of the network, and then from the intermediate switch to the destination switch, composing its route of two segments of minimal routes: *lg1 - 1g1*. Therefore, 2 local VCs and 1 global VC

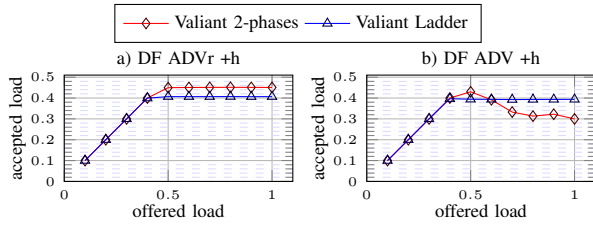


Fig. 2. Performance evaluation of Dragonfly (DF) topology, under ADV +h and ADVr +h traffic patterns. It shows the average accepted load for an offered load, both measured in  $\frac{\text{phits}}{\text{cycle}}$ . Results are averaged over ten runs.

are employed per segment, making a total of 4 and 2 VCs, for local and global ports, respectively.

The ladder is an old method which order VCs to break deadlock, and the strategy is to increase the order of the visited VC with each link traversed by the packet [3]. With the Ladder scheme, independently of the routing, the  $i$ th-hop of a packet goes through the  $VC_i$  of a certain port of a certain switch in the network. In case of a DF network with Valiant routing, a packet could take up to six hops, requiring six VCs in the Ladder.

### C. Methodology and Results

The presented buffer architectures are analyzed in a Dragonfly network. To conduct experiments, CAMINOS [4] network simulator is employed. The most typical synthetic traffic patterns utilized for DF networks are the ADV +i patterns, where each server at group  $g$  sends traffic to the server located at the same relative position at the  $g + i$  group. Also, the ADVr +i pattern, that do the same but the specific server in the destination group is randomized. Specifically, the ADVr +h, and ADV +h<sup>1</sup> patterns are used in this work, because they have been object of study in previous works [2]. In the simulations included, the accepted load (throughput) per server is showed for a specific offered load or injection rate per server. Both accepted and injected load are measured in  $\frac{\text{phits}}{\text{cycle}}$ . The performance results for the traffic patterns introduced are depicted in Figure I-C.

In the case of the ADVr +h pattern, it can be observed that 2-phases maintains a stable performance above 0.4 of accepted load the entire simulation. Also, the Ladder shows a stable performance with an accepted load around 0.4, but slightly below 2-phases performance. For the ADV +h traffic pattern, it can be observed that 2-phases mechanism experience congestion problems when injecting traffic above saturation ( $\geq 0.5$  of offered load), and the accepted load becomes unstable. In contrast, the Ladder exhibits an accepted load of 0.4 the entire simulation, without instabilities.

<sup>1</sup>h is a topological parameter of the DF network, which in the DF tested is h=6. Therefore, the ADV +6 and ADVr +6 patterns are simulated

### D. Conclusions and future work

In conclusion, the evaluation of buffer architectures in the DF network, when dealing with the ADVr +h and ADV +h traffic patterns, revealed distinct performance characteristics. While 2-phases mechanism proved stable performance in ADVr +h, its performance dropped for the ADV +h pattern. In contrast, the Ladder mechanism exhibited stable and high performance under both traffic patterns. These findings suggest that buffer architecture have significant impact on performance, and should be considered when designing interconnection networks.

Further study showed us that the impact of buffer architecture depend on a lot of factors, as the routing algorithm, traffic, topology, etc. Future work focus on analyzing its impact in other realistic scenarios, and how it can prevent network congestion or low performance.

For instance, an analysis could be carried out in the presence of more complex traffic patterns based on real applications. Aspects like job allocation, task placement and job scheduling of several applications in the network should be taken into account. Furthermore, other routing algorithms should be considered in the evaluation, like a source adaptive routing such as UGAL.

Lastly, with a general view, there are possibilities of proposing a generalized or specific buffer architecture which improves traditional mechanisms like the ones presented here, in a wide range of realistic scenarios.

## II. ACKNOWLEDGMENT

This work has been published in proceedings of the SBAC-PAD 2023 [5].

## REFERENCES

- [1] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proceedings of the 35th Annual International Symposium on Computer Architecture*. IEEE Computer Society, 2008, pp. 77–88.
- [2] M. García, "Routing mechanisms for dragonfly interconnection networks," Ph.D. dissertation, University of Cantabria, 2014.
- [3] K. D. Günther, "Prevention of deadlocks in packet-switched data transport systems," *IEEE Transactions on Communications*, vol. 29, no. 4, pp. 512–524, 1981.
- [4] "CAMINOS," <https://crates.io/crates/caminos>.
- [5] A. Cano, C. Camarero, C. Martínez, and R. Bevide, "Analysing mechanisms for virtual channel management in low-diameter networks," in *2023 IEEE 35th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, 2023, pp. 12–22.



**Alejandro Cano** received his BSc and MSc in Computer Science at the University of Cantabria, in Spain. Now, he is a PhD student at the Computer Architecture group in the same university.

# Alternative Splicing variability between human populations at single-cell resolution

Rubén Chazarra-Gil<sup>1</sup>, Martin Hemberg<sup>2</sup>, Marta Melé<sup>1</sup>

<sup>1</sup>Barcelona Supercomputing Center, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona, Spain

[ruben.chazarra@bsc.es](mailto:ruben.chazarra@bsc.es)

<sup>2</sup>Harvard Medical School, Boston, Massachusetts, USA

**Keywords**— single-cell RNA-seq, alternative splicing, differential transcript usage

Transcriptional response to immune challenges varies between individuals of different genetic ancestries which impacts susceptibility to infectious disease (1, 2). While humans of different populations differ in their expression of inflammatory genes upon immune stimulation (1), population variation in the alternative splicing landscape due to infection has been less explored, being restricted to isolated cell types or heterogeneous tissues. With a growing consensus in the contribution of alternative splicing to immunity and common disease risk, characterizing immune related alternative splicing differences between individuals becomes a priority. Single-cell RNA technologies can emerge as a valuable tool to understand population differences in the response to infection across a broad range of cell types. Until recently, the limited number of samples available and the sparsity of single-cell data made the study of alternative splicing at single-cell resolution extremely challenging. With the advent of large cohort single-cell RNA-seq (scRNA-seq) datasets available including hundreds of donors and thousands of cells, we can now for the first time uncover alternative splicing interindividual differences at single-cell resolution.

Methods for splicing analysis specifically designed for droplet-based scRNA-seq data only started to emerge recently. Strategies to detect splicing in 3' biased single-cell data are diverse, and vary between: i) splice junction quantification, ii) calculation of exon-inclusion levels with splice junction reads, or iii) calculation of differential exon-usage (3). Nonetheless, the mentioned strategies focus on individual elements of mRNA transcripts, such as exons or splice junctions, lacking the consideration of the entire transcript as a biological entity. Measuring differences in the expression ratios of transcripts, also known as differential transcript usage (DTU), provides a holistic view on the mRNA transcript. Furthermore, it enables the detection of isoform switches between conditions and provides a functional point of view to alternative splicing characterization. DTU methods for droplet-based scRNA-seq data often focus on cell type comparisons, and don't allow the incorporation of additional covariates when modelling transcript ratios. To assess transcript usage differences between individuals, the incorporation of confounding effects becomes mandatory. To fill this gap, we have developed a computational framework which allows for transcript ratio modeling in a multivariate fashion using 3' scRNA-seq data.

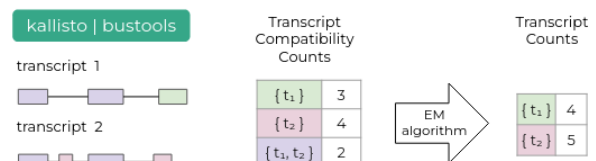
In this study, we present a computational strategy to study alternative splicing inter-individual variation at single-cell resolution. We apply this method to a 3' biased scRNA-seq dataset of peripheral blood mononuclear cells from individuals of different genetic ancestries. We detect genes changing their transcript usage between different human populations across all immune cell types. Our characterization of population differences in alternative splicing in blood immune cell types,

contribute to the understand inter-individual transcriptional response to immune stimuli.

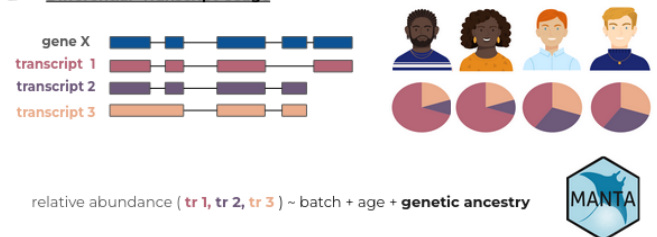
## A computational framework for inter-individual differential transcript usage using 3' single-cell RNA-seq data

To address population differences in transcript usage, we developed a computational framework to identify alternative splicing differences in 3' scRNA-seq datasets. This framework is composed of 2 steps: transcript quantification and DTU analysis. To quantify transcripts for each single cell, we first perform pseudo-alignment of 3' biased scRNA-seq reads to a reference transcriptome generating transcript compatibility counts (TCCs) (Fig 1. A). TCCs are counts for each of the equivalence classes defined by the pseudo-aligner. Transcript Counts are then estimated from the TCCs by running Expectation Maximization algorithm. We use kallisto-bustools suite (4) for the steps involving transcript quantification. Next, we conduct DTU analysis. We perform differential transcript usage by modeling transcript ratios of each gene between conditions for each cell type (Fig 1. B). To do so, we use the MANTA (5), which fits particularly well to our research context. On the one hand, by being non-parametric MANTA does not assume a particular distribution of transcript ratios, which may not fit an established trend in 3' scRNAseq data. On the other hand, by being multivariate MANTA allows for the incorporation of many covariates in our transcript ratio modelling, a required feature to test inter-individual differences.

### A Transcript Quantification



### B Differential Transcript Usage



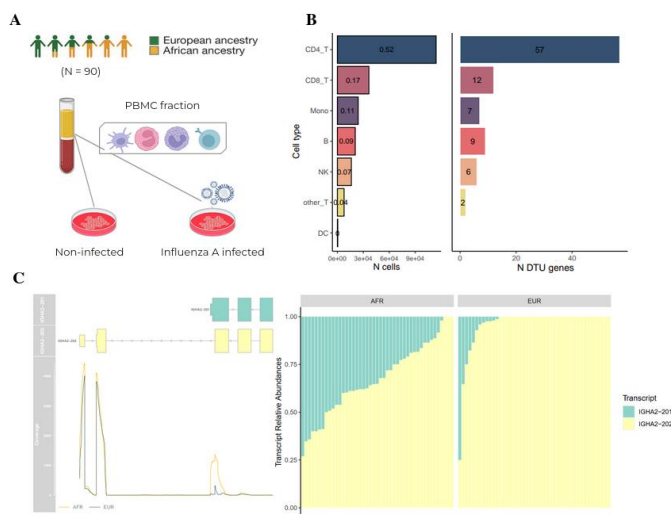
**Fig 1. A computational framework for differential transcript usage using 3' single-cell RNA-seq data.** A) Transcript quantification at single-cell level with kallisto-bustools (4). First, transcript compatibility counts (TCCs) are obtained by pseudo-aligning 3' scRNAseq reads to the reference transcriptome. Next, transcript counts are estimated from the TCCs by running the Expectation Maximization algorithm. B) Inter-individual differential transcript usage. Transcript relative abundances are obtained from the per cell type-donor pseudo bulked transcript counts. DTU between human populations are tested for each gene with > 1 transcript expressed in each cell type by modeling its

transcript relative abundances as a function of the genetic ancestry of each individual and other covariates. For the test, we use MANTA (5) which implements a multivariate nonparametric approach.

## Differences in transcript usage between human populations are present across all blood immune cell types

We apply our framework to study alternative splicing differences between 45 European and 45 African descent individuals in a scRNA-seq dataset of immune blood cells before and after influenza infection (Fig 2 A). First, we evaluate differences in transcript usage between populations at baseline, finding changes in transcript ratios in all immune cell types except Dendritic Cells (Fig 2 B). We find the number of populations DTU genes is highly correlated with the number of cells per cell type ( $\rho = 0.744$ ,  $p=0.034$ ). This is likely explained by the high sparsity of 3' single cell data, which causes cell types with lower number of cells to quantify less transcripts, and thus test less genes for DTU.

As an example of population DTU, we highlight the IGHA2 gene which encodes for the constant region 2 of the heavy chain of immunoglobulin A. We observe how African descent individuals express both transcripts of the gene, whereas European descent individuals mostly express a single transcript (Fig 2 B). Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) has previously been reported, highlighting the potential specificity of IGHA2 transcript expression.



**Fig 2. Ancestry associated DTU across the blood immune cell types.**

**A)** scRNA-seq PBMC dataset from individuals of mixed European and African ancestry before and after influenza infection. **B)** N of genes exhibiting differential transcript usage between human populations across blood immune cell types is highly correlated with N of cells per cell type. Left: N of cells per cell type. Right: N of population DTU genes per cell type at baseline. **C)** Illustrative example of a population DTU gene. The IGHA2 gene shows different transcript usage between European and African descent individuals in B cells. Left panel: the 2 transcripts of the IGHA2 gene (top) and the sequencing coverage split by population (bottom). The 3' bias of droplet-based scRNAseq data is appreciated as coverage only spans the 3' end of transcripts. Right panel: per-donor relative abundances of the IGHA2 transcripts quantification. African descent individuals express the 2 transcripts, whereas European descent almost exclusively express IGHA2-202.

## Discussion

Characterizing inter-individual alternative splicing variation is crucial to understand differences in transcriptional response to immune challenges between human populations which impacts susceptibility to infectious disease. To characterize population alternative splicing differences, we have developed a computational strategy to perform differential transcript usage with 3' scRNAseq data. Applying this pipeline to the circulating immune system between individuals of different genetic background, we detected genes with population specific alternative splicing patterns across all blood immune cell types. As a validation of our results, we find significant overlaps of our population differentially spliced genes with bulk RNA-seq data from blood immune cells (not shown). Our approach can be readily scalable to associate differences in transcript ratios with other individual traits or environmental factors in future studies.

## References

- 1) Randolph, Haley E., et al. "Genetic ancestry effects on the response to viral infection are pervasive but cell type specific." *Science* 374.6571 (2021): 1127-1133 <https://doi.org/10.1126/science.abg0928>
- 2) Aquino, Y., Bisiaux, A., Li, Z. *et al.* Dissecting human population variation in single-cell responses to SARS-CoV-2. *Nature* 621, 120–128 (2023). <https://doi.org/10.1038/s41586-023-06422-9>
- 3) Hu, Yu, Kai Wang, Mingyao Li. Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers. *PLoS computational biology* 16.6 (2020). <https://doi.org/10.1371/journal.pcbi.1007925>
- 4) Melsted, P., Boeshaghi, A.S., Liu, L. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* 39, 813–818 (2021). <https://doi.org/10.1038/s41587-021-00870-2>
- 5) Garrido-Martín, D., Calvo, M., Reverter, F. *et al.* A fast non-parametric test of association for multiple traits. *Genome Biol* 24, 230 (2023). <https://doi.org/10.1186/s13059-023-03076-8>



My name is **Ruben Chazarra Gil** and I am a 3rd year PhD pre-doc at the Transcriptomics and Functional Genomics in the Life Sciences Department at BSC. I studied a BSc in Biotechnology at the Universitat Politècnica de València (UPV) where I became interested in bioinformatics. Following this interest, I moved to Cambridge (UK) where I performed internships in the Sanger Institute and at the European Bioinformatics Institute (EBI-EMBL). Here, I developed my bioinformatic skills and became passionate about the field of single-cell transcriptomics. Next, I enrolled as a bioinformatician at the University of Cambridge where I worked in close collaboration with experimental teams performing extensive single-cell data analysis. Now I study inter-individual transcriptomics differences at single-cell level. On my personal side, I am a fan of road cycling and everything involving music, from playing to producing.



# A pipeline to preprocess long reads Oxford Nanopore sequencing data to reveal the transcriptomic diversity of human populations

Pau Clavell-Revelles<sup>#1</sup>, Fairlie Reese<sup>#2</sup>, Marta Melé<sup>#3</sup>

<sup>#</sup>Life Sciences, Barcelona Supercomputing Center (BSC), Spain

<sup>1</sup>pau.clavell@bsc.es, <sup>3</sup>marta.mele@bsc.es

<sup>2</sup>fairlie.reese@bsc.es

**Keywords**— Long-read sequencing, Human genetic diversity, Transcriptomic

## EXTENDED ABSTRACT

More than 20 years after the release of the human reference genome, the main human gene annotations (GENCODE and RefSeq) continue to grow in number of genes and transcripts. In parallel, there has been a surge of projects to functionally annotate genetic variants as well as to associate them to specific phenotypes. However, human genomics and transcriptomics studies present a sheer bias towards individuals of European ancestries. This can be detrimental in many aspects considering that despite most of the human genetic variants in an individual can be found in populations from other continents, projects like the 1000 Genomes revealed that most human genetic diversity is population-private [1]. This is particularly abundant in the African superpopulation which did not undergo the bottle-necks and founder effects that the Out-of-Africa populations suffered.

Recently, several efforts to reduce this bias in genomics have resulted in the sequencing of thousands of novel genomes from ancestry-diverse cohorts and the production of population-specific genomes, T2T genomes from non-European individuals and the human pangenome draft graph [2].

In a slower pace and with much lower sample sizes, RNA sequencing experiments have extended to more non-European populations but always with a strong focus on the genetic basis of transcriptomic traits (gene expression and alternative splicing). These investigations have shown that while genetic differences between ancestries can distinguish populations in a principal component analysis, transcriptomic differences do not [3]. Interestingly, human populations share a genetic regulatory architecture and therefore the identified transcriptomic differences arise basically from differences in allele frequencies [4]. As a consequence, genetic variants found in non-European populations but rare or nonexistent in Europeans might be affecting the alternative splicing of many genes leading to the formation of novel transcripts.

The identification of currently unannotated transcripts through pipelines of transcriptome assembly could enhance the current annotations in a population-aware manner. Their potential benefits include improved mappings and gene quantifications of specific genes in non-European populations.

Regardless, all publicly available RNA-seq data including individuals of non-European descent are based on microarrays and short-read sequencing technologies, which do not have or miss, respectively, information about the full-length transcript. Hence we designed a project to generate full-length RNA-seq data from a population-diverse cohort to enhance the current GENCODE gene annotation.

## A. Data overview

We have selected lymphoblastoid cell lines from 45 unrelated individuals belonging to eight different populations in four continents from the Coriell Institute for Medical Research. All the samples have been previously sequenced in at least one DNA-sequencing study. Most of the samples are included in the 1000G project but some others also belong to projects such as the HapMap, the Genome In a Bottle (GIAB) or the Human Genetic Diversity Panel (HGDP). Some selected cell lines were RNA-sequenced in the GEUVADIS and the Multi-Ancestry Gene Expression (MAGE) resource.

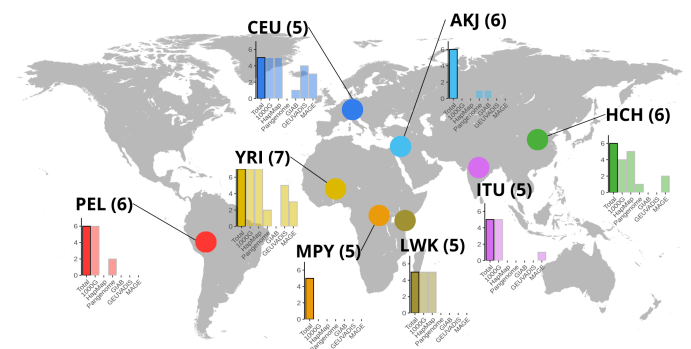


Fig. 1 Geographical representation of selected populations in our study. The barplots indicate the total number of individuals and the number of individuals belonging to different DNA-seq and RNA-Seq publications.

The selected populations are: Peruvians from Lima (Perú), Central and Northern Europeans from Utah (USA), Yoruba in Ibadan (Nigeria), Luhya in Webuye (Kenya), Mbuti Pygmies, Ashkenazi Jewish, Indian Telugu in the UK, Han Chinese in Beijing (China). These populations were chosen following a series of criteria meant to represent the maximum of populations-private genetic variation from different continents and inclusion in the human pangenome reference draft.

### B. ONT sequencing and preprocessing pipeline

In this project we use Oxford Nanopore Technologies to sequence full-length cDNA obtained by CapTrap protocol. CapTrap enriches in full-length mRNAs thanks to a dual capture based on the presence of a poly-A tail and a 5' cap. Then a retrotranscription to cDNA followed by PCR amplification provides enough cDNA material for Oxford Nanopore sequencing library preparation.

These experimental procedures lead to different issues that can be partially assessed *in silico*. Firstly, PCR duplications must be quantified and deduplicated. Secondly, reads originated from sequencing both strands of cDNA and therefore containing redundant information must be eliminated. Finally, chimeric reads generated by the concatenation of very closely sequenced molecules must be splitted. Here, we develop a snakemake pipeline to deal with these problems after the sequence basecalling with dorado.

Firstly, PCR duplication was assessed by leveraging the internal sixteen nucleotides long unique molecular identifiers (UMI-16) added during the CapTrap protocol. To do so, all reads are mapped to a blast database of the 69 nucleotides long linker containing the UMI-16. Those reads whose linker is identified are subsequently used to extract the UMI-16 sequenced thanks to the CIGAR. Then, the UMI-16 is appended to each read name and mapped to the reference genome (hg38) with minimap2. Afterwards, the mapped reads are processed using the `umi_tools` library, which removes those reads mapped to the same gene and sharing the same UMI sequence, under an edit threshold compatible with ONT error rate.

Secondly, the removal of complementary reads coming from the same cDNA molecule are removed thanks to the bam tag included in the unmapped bam file generated by dorado basecaller in duplex mode.

Finally, chimeric reads are splitted in two rounds of splitting based on the identification of different combinations of terminal adapters (ONT adapters and CapTrap linkers) thanks to the software `duplex_tools` contained in Guppy.

### C. Future directions

After running our preprocessing pipeline we will run LyRic and other transcriptome assembly tools to produce transcript models that will be assessed by the HAVANA team at the EBI. The accepted transcript models will complement the current GENCODE annotation to build an enhanced annotation meant to quantify the transcriptomic differences between human populations through Differential Gene and Transcript expression and usage, allele-specific analyses, etc.

### D. Conclusions

Our project will reveal the transcriptomic differences between human populations through long-reads sequencing of cDNA from a selection of individuals from different human populations. Here we have developed a snakemake pipeline to preprocess our CapTrap-seq data to remove to the maximum extent the information redundancy and sequencing errors. This will improve mapping, transcript identification and gene quantification to ultimately better understand the transcriptomic diversity between human populations.

### E. ACKNOWLEDGEMENTS

I would like to thank Carme Arnan, Silvia Carbonell-Sala, Zighereda Ogbah and Winona Oliveros their help in the data generation, Roderic Guigó, Tamara Perteghella, Gazaldeep Kaur, Emilio Palumbo and Silvia Carbonell-Sala their help in the data management and pre-processing.

### References

- [1] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation." *Nature* vol. 526,7571 (2015): 68-74. doi:10.1038/nature15393
- [2] Liao, Wen-Wei et al. "A draft human pangenome reference." *Nature* vol. 617,7960 (2023): 312-324. doi:10.1038/s41586-023-05896-x
- [3] Martin, Alicia R et al. "Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture." *PLoS genetics* vol. 10,8 e1004549. 14 Aug. 2014, doi:10.1371/journal.pgen.1004549
- [4] Taylor, Dylan J et al. "Sources of gene expression variation in a globally diverse human cohort." *bioRxiv* : the preprint server for biology 2023.11.04.565639. 8 Nov. 2023, doi:10.1101/2023.11.04.565639. Preprint.

### Author biography



**Pau Clavell-Revelles** was born in Vallgorguina, Barcelona, in 1999. He received the bachelor's degree in Biochemistry from the University of Barcelona, Barcelona in 2021, and masters degree in Omics Data Analysis from the Universitat de Vic, Barcelona, in 2022.

Since October 2022, he has been with the Transcriptomics and Functional Genomics Lab in the Life Sciences Department at the Barcelona Supercomputing Center (BSC), where he was a Junior Research Engineer until the start of his PhD in Biomedicine in 2023. His current research interests include human populations genomics and transcriptomics, recent human evolution and functional genomics. Apart from research, he is also a science disseminator in the digital magazine of science outreach *Ciència Oberta* and member of the platform *Neurones Fregides*.

# Circulation types leading to subtropical air intrusions in the Western Mediterranean

Pep Cos<sup>\*‡</sup>, Francisco Doblas-Reyes<sup>\*†</sup>, Raül Marcos<sup>‡</sup>, Matías Olmo<sup>\*</sup>, Ángel Muñoz<sup>\*</sup>, Lluís Palma<sup>\*</sup>, Diego Campos<sup>\*</sup>,  
<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain  
<sup>†</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain  
<sup>‡</sup>Universitat de Barcelona, Barcelona, Spain  
E-mail: {josep.cos}@bsc.es

**Keywords**—saharian intrusions, observed climate, clustering, weather types.

## I. EXTENDED ABSTRACT

One of the many relevant mechanisms that drive extreme temperatures at daily time scales in the Mediterranean region, especially during Summer, is the intrusions of Saharian or subtropical continental air. These warm and stable air masses that penetrate northward, heavily influence the Mediterranean region's temperature and have been linked to severe heat waves in the past [1]. This phenomenon hasn't been extensively studied in the literature, and therefore, we think it is of utmost importance to understand the intrusions and the mechanisms that drive them, especially as the amount of observed events has increased in the present climatology (1991-2022) with respect to the end of last century (1959-1990). An analysis was conducted to categorize air masses originating from low latitude subtropical desert areas in the historical period using data from the ERA5 observational dataset [2]. This analysis relied on basic thermodynamic air characteristics: the geopotential thickness of the 1,000-500 hPa layer [3] and the average potential temperature of the 925-700 hPa layer (). Utilizing the climatological mean values of these variables during summer (June-August) from 1959 to 2022 for each grid point, specific criteria were obtained to recognize air masses of subtropical origin. Our approach consists on identifying the days where these air masses move northward and reach regions in the Western Mediterranean basin. We explore the frequency, spatial distribution and persistence of these events and the potential trends in the historical period. The amount of intrusion days increase in the historical period, as seen in Figure 1, where the mean of summer intrusion days is shown for the period 1959-1990 and the period 1991-2022.

Our results show how the subtropical air intrusions can reach many regions in the Mediterranean and how there is an increase in the amount of observed events. The effects of these events are beyond the region where the intrusion reaches, and they can generate temperature anomalies in many other points of the Mediterranean basin. Therefore, the impacts are affect not only locally.

We want to understand which large-scale weather types tend to generate these intrusions and where. Therefore, we employ a powerful tool to cluster the most relevant circulation patterns that can lead to intrusions. To do so we explore different applications of the k-means [4]. This method involves

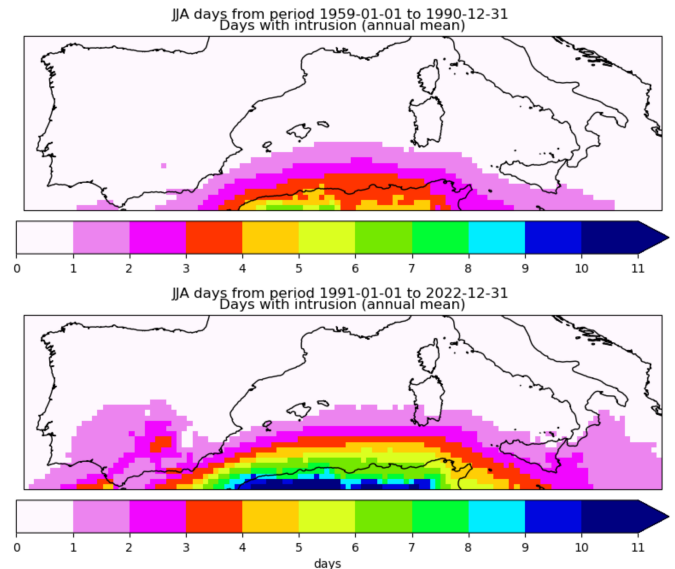


Fig. 1. mean JJA intrusion days in period 1959-1990 (top) and 1991-2022 (bottom)

partitioning days into a predetermined number of clusters, aiming to minimize the total squared Euclidean distances within each cluster set. The daily geopotential data is then projected onto its primary empirical orthogonal functions, capturing 95% of the variability [5].

In the future we would like to analyse if the models from the last phase of the Coupled Model Intercomparison Project (CMIP6) can reproduce satisfactorily the intrusions in the historical period.

## REFERENCES

- [1] P. M. Sousa *et al.*, "Saharan air intrusions as a relevant mechanism for Iberian heatwaves: The record breaking events of August 2018 and June 2019," *Weather and Climate Extremes*, vol. 26, p. 100224, Dec. 2019. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2212094719300349>
- [2] H. Hersbach *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020. [Online]. Available: <https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803>
- [3] J. Galvin, *An Introduction to the Meteorology and Climate of the Tropics*. John Wiley & Sons, 2015.

- [4] A. W. Robertson and M. Ghil, "Large-scale weather regimes and local climate over the western united states," *J. Climate*, vol. 12, 1999.
- [5] G. Muñoz *et al.*, "Cross-time scale interactions and rainfall extreme events in southeastern south america for the austral summer. part i: Potential predictors," *Journal of Climate*, vol. 28, no. 19, pp. 7894 – 7913, 2015. [Online]. Available: <https://journals.ametsoc.org/view/journals/clim/28/19/jcli-d-14-00693.1.xml>



**Pep Cos** received his BSc degree in Aerospace Engineering from Universitat Politècnica de Catalunya (UPC), Terrassa in 2019. He completed his MSc degree in Meteorology from Universitat de Barcelona (UB), in 2021. Since 2020, he has been with the Earth System Services group of Barcelona Supercomputing Center (BSC) first as a MSc intern and then as a PhD student at the department of physics of Universitat de Barcelona (UB).

# MAXWEL: Simulation of EM wave propagation in plasma using FEM

Hernán Domingo Ramos<sup>\*†</sup>, Alejandro Soba Pascual<sup>\*</sup>, Daniel Gallart Escolà<sup>\*</sup> Mervi Johanna Mantsinen<sup>\*‡</sup>

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Ghent University, Ghent, Belgium

<sup>‡</sup>ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

E-mail: hernan.domingo@bsc.es

**Keywords**—*Finite Element Method, Electromagnetic simulation, Nuclear fusion.*

## I. EXTENDED ABSTRACT

Finite Elements Method (FEM) is a numerical tool that allows to solve differential equations, subjected to a series of boundary conditions, up to a good approximation. Its applicability to a broad spectrum of fields, together with its good adaptability to different geometries and physical domains have allowed this method to extend and expand its uses since the 1940's [1], [2]. To be able to solve these Boundary Value Problems opens up a world of numerical simulations, reachable throughout any conventional computing device, and with the interesting potential of scalability to High Performance Computing environments. This would ultimately allow to reduce the required wall time of these simulations and even more importantly, to increase the domain's complexity and the results' accuracy.

The main focus of the present work is to develop a numerical tool able to solve a general three-dimensional electromagnetic wave equation in plasma environments, within the context of nuclear fusion and tokamak reactors. For example, this could stand for an injected wave coming from resonant heating antennas. This numerical tool is meant to be implemented in Alya's framework [3], which serves as a work environment to implement different physics solution using FEM.

The starting point is to work with a 2D circular domain and then moving towards a tokamak-like cross section and finally scaling to a final 3D solution of the full toroidal reactor shape (fig. 1). The ability to solve a general wave equation in the whole domain of the reactor, plays a crucial role when studying the magnetohydrodynamic equilibrium, transport phenomena and plasma heating.

### A. Alya Environment

Alya is a software developed by the BSC CASE department, and it is been improved and updated since 2004. Apart from a powerful FEM code, it is also a framework on which different physics can be treated depending on the interest of the researcher. Alya was designed from the very first line of code, optimised to get the maximum performance from a supercomputer machine.

Alya is computationally adapted to new architectures, being able to run in CPUs and GPUs, also with the future expectation to run in exascale computers. The long-term goal within the

FUSION group and in relation to this plan, is to develop a series of modules that will work together in different problems that occur in magnetically confined plasma reactors. Some modules are already found in Alya's environment, i.e. those related to thermohydraulics and thermomechanics, neutron transport and superconductor magnetic coils. The future perspective is focused on developing three new modules:

- 1) EQUILI. Able to solve the plasma equilibrium
- 2) MHDNOL. In order to describe nonlinear magneto-hydrodynamics phenomena
- 3) MAXWEL. The module described in this publication, responsible of analysing the electromagnetic behaviour of the plasma and its surrounding environment.

### B. Formulation

By using the time harmonic and monochromatic expressions of Maxwell-Faraday equation and Ampère's law, they can be combined in order to obtain the generalised expression of a 3D vector wave equation (eqs. (1) and (2)).

$$\nabla \times (\boldsymbol{\mu}^{-1} \cdot \nabla \times \mathbf{E}) - \omega^2 \boldsymbol{\varepsilon} \cdot \mathbf{E} = -j\omega \mathbf{J} \quad (1)$$

$$\nabla \times (\boldsymbol{\varepsilon}^{-1} \cdot \nabla \times \mathbf{H}) - \omega^2 \boldsymbol{\mu} \cdot \mathbf{H} = \nabla \times (\boldsymbol{\varepsilon}^{-1} \cdot \mathbf{J}) \quad (2)$$

It can be appreciated that this two expression have a common form, which considering an invariant geometry along a given coordinate axis, can be reduced to the 2D case:

$$-\nabla \cdot (\mathbf{p} \cdot \nabla u) + qu = f \quad (3)$$

being  $u$  our field of interest ( $E_z$  or  $H_z$ ) and  $\mathbf{p}$ ,  $q$  and  $f$  independent terms and coefficients related to the relative permeability  $\mu_r$ , the relative permittivity  $\varepsilon_r$  and current density  $\mathbf{J}$ .

Up to now, the work has been focused in reproducing a standard case, extracted from [4], in which they solve the *generalised homogeneous Helmholtz equation* (eqs. (4) and (5)), for a perfect electric conductor and for a dielectric object, both isotropic and anisotropic, in a circular domain and in absence of external sources. Helmholtz equation describes a time independent wave propagation phenomena, such as plane waves or, in this case, TM and TE propagation modes.

$$\nabla \cdot (\boldsymbol{\Lambda}_\mu \cdot \nabla E_z) + k_0^2 \varepsilon_r^{zz} E_z = 0 \quad (4)$$

$$\nabla \cdot (\boldsymbol{\Lambda}_\varepsilon \cdot \nabla H_z) + k_0^2 \mu_r^{zz} H_z = 0 \quad (5)$$

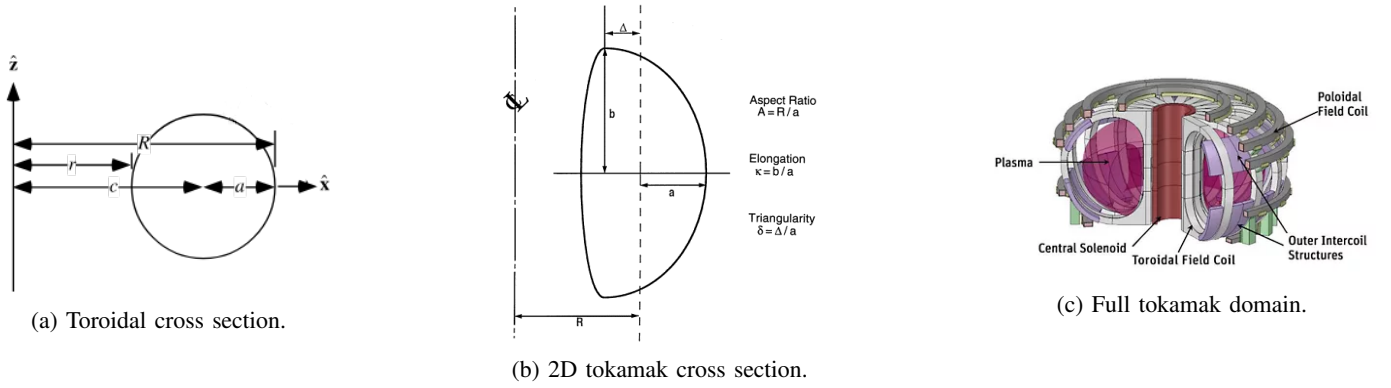


Fig. 1: Integration domain at different steps or degrees of approximation. Images are extracted from [5], [6], [7]

In eqs. (4) and (5), parameters are defined as  $k_0 = \omega\sqrt{\varepsilon_0\mu_0}$ ,

$$\Lambda_\mu = \frac{(\boldsymbol{\mu}_r)_{\text{sub}}^T}{|(\boldsymbol{\mu}_r)_{\text{sub}}^T|} \quad \text{and} \quad \Lambda_\varepsilon = \frac{(\boldsymbol{\varepsilon}_{rc})_{\text{sub}}^T}{|(\boldsymbol{\varepsilon}_{rc})_{\text{sub}}^T|}.$$

Matrices  $\boldsymbol{\mu}_r$  and  $\boldsymbol{\varepsilon}_{rc} = \boldsymbol{\varepsilon}_r - j\frac{\boldsymbol{\sigma}}{\omega\varepsilon_0}$  are the relative permeability and the complex relative permittivity tensors, whose sub index corresponds to the  $xy$   $2 \times 2$  submatrix.  $\varepsilon_{rc}^{zz}$  and  $\mu_r^{zz}$  are, consequently, the complex permittivity and permeability in the  $z$  direction.

The anisotropic dielectric case is of most importance since it can be adapted to a real plasma permittivity tensor, which would allow to simulate the wave propagation inside a 2D nuclear fusion reactor cross section. Some preliminary results are shown in fig. 2. In this case there is an incident wave coming from the left side, and it scatters in presence of a dielectric object. Since we are solving eq. (4), in fig. 2a it is shown the scattered field ( $E_z^{\text{scat}}$ ) from the incident field ( $E_z^{\text{inc}}$ ), while in fig. 2b the total field  $E_z = E_z^{\text{inc}} + E_z^{\text{scat}}$  is presented.

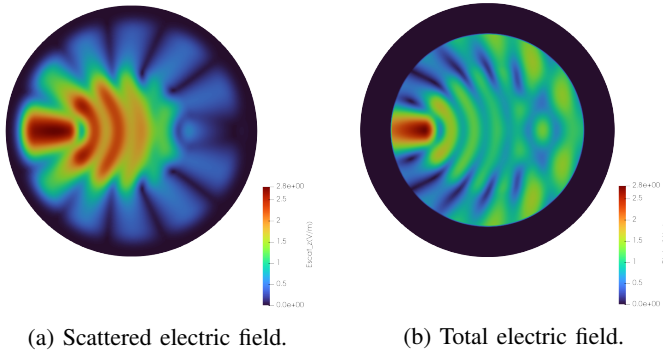


Fig. 2: TM wave propagation under a circular anisotropic dielectric object.

### C. Conclusion

We have presented some preliminary results derived from [4] as a first step towards a full implementation of the Helmholtz equation solution in Alya's environment. The final goal of this work is to develop a complete module to solve

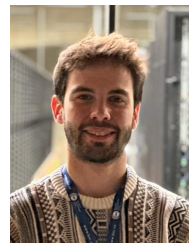
eqs. (1) and (2), in order to be able to retrieve a full characterisation or simulation of an electromagnetic wave in a three dimensional domain of a tokamak nuclear fusion reactor.

## II. ACKNOWLEDGMENT

This work has been carried out within the FuseNet Association mobility program and the EUROfusion Consortium framework, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

## REFERENCES

- [1] A. Hrennikoff, "Solution of Problems of Elasticity by the Framework Method," *Journal of Applied Mechanics*, vol. 8, no. 4, pp. A169–A175, Mar. 2021. [Online]. Available: <https://doi.org/10.1115/1.4009129>
- [2] R. Courant, "Variational methods for the solution of problems of equilibrium and vibrations," *Bulletin of the American Mathematical Society*, vol. 49, no. 1, pp. 1–23, 1943. [Online]. Available: <https://doi.org/10.1090/S0002-9904-1943-07818-4>
- [3] BSC CASE department, "bsc-alya · GitLab," Nov. 2022. [Online]. Available: <https://gitlab.com/bsc-alya>
- [4] Ö. Özgün and M. Kuzuoğlu, *MATLAB-based Finite Element Programming in Electromagnetic Modeling*. Boca Raton: CRC Press, Oct. 2018.
- [5] E. W. Weisstein, "Torus," publisher: Wolfram Research, Inc. [Online]. Available: <https://mathworld.wolfram.com/Torus.html>
- [6] C. C. Baker, R. W. Conn, F. Najmabadi, and M. S. Tillack, "Status and prospects for fusion energy from magnetically confined plasmas," *Energy*, vol. 23, no. 7, pp. 649–694, Jul. 1998. [Online]. Available: [https://doi.org/10.1016/S0360-5442\(97\)00068-6](https://doi.org/10.1016/S0360-5442(97)00068-6)
- [7] "Designing Nuclear Fusion Reactors with Simulation | Ansys." [Online]. Available: <https://www.ansys.com/en-gb/blog/designing-nuclear-fusion-reactors-simulation>



**Hernán Domingo Ramos** received his BSc degree in Physics from Universitat de Barcelona (UB), in 2021. The following year, he worked as Data Analyst and Data Engineer in SDG Group in Barcelona. He is in his final year of the European Master of Science in Nuclear Fusion and Engineering Physics. Currently, he is doing his final master thesis in the FUSION group of Barcelona Supercomputing Center (BSC), with the intention of becoming a PhD student in the present project.

# Design of a surface-accessible epitope panel using Brewpitopes to empower early lung cancer detection

Roc Farriol-Duran<sup>1,2,3,5</sup> Evelyn Fitzsimons<sup>3,4,5</sup>, Anna Maria Díaz-Rovira<sup>1</sup>, Víctor Montal<sup>1</sup>, Richard Lee<sup>6</sup>, Víctor Guallar<sup>1</sup>, Eduard Porta-Pardo<sup>1,2</sup> & Kevin Litchfield<sup>3,5</sup>

*1*Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain. *2*Josep Carreras Leukaemia Research Institute (IJC), Badalona, Spain *3*The Tumour Immunogenomics and Immunosurveillance (TIGI) Lab, UCL Cancer Institute, London, WC1E 6DD, United Kingdom. *4*Pre-Cancer Immunology Laboratory, UCL Cancer Institute, London, WC1E 6DD, United Kingdom. *5*Cancer Research UK Lung Cancer Centre of Excellence, London, UK. *6*The Royal Marsden Hospital Early Diagnosis and Detection Centre, London, UK

roc.farriol@bsc.es, k.litchfield@ucl.ac.uk

**Keywords—** Immunoinformatics, Antibody, Early Lung Cancer Detection

## EXTENDED ABSTRACT

Up to 50% of cancer patients are diagnosed at a late stage with tumours that are often unresectable, leading to intensive treatments and a preventable loss of life. Whilst multiple detection screenings have been developed for advanced tumours, many have shown limited sensitivity and specificity for early-stage malignancies. Hence, underscoring an urgent need for novel early detection strategies.

Peptide screening to capture antibody signatures in blood has been extensively used in infectious diseases' diagnosis. However, cancer proteins are less foreign to the immune system thus epitope prediction approaches for tumour detection need to prioritise specificity. To this end, we used the SERA discovery platform (Serimmune) to interrogate plasma samples from 60 stage I LUAD patients and analyzed the obtained dataset with IMUNE algorithm to identify an enriched epitope motif shared across the cohort. To generate a customized peptide screening panel and to ensure the surface accessibility of the candidate epitopes, we implemented the Brewpitopes pipeline on the proteins that contain the motif. Brewpitopes works upon protein sequences for linear epitope prediction and crystal structures or AlphaFold2 models for conformational epitopes. The pipeline leverages a compendium of state-of-the-art B-cell epitope predictors and a series of bioinformatic tools to map the candidate peptides to extracellular protein regions, to avoid glycosylation sites and to locate them in the 3D surface of the protein to select accessible regions.

The target epitope motif mapped to 24 human proteins (251 candidate peptides (11-mers)). The use of Brewpitopes led to an optimized panel comprised of 7 target proteins and 42 extracellular, non-glycosylated and surface-accessible candidates. The resulting panel will be validated in the NIMBLE early lung cancer detection study (>360 patients recruited to date). This study reports for the first time the implementation of Brewpitopes in cancer and displays its capacity to prioritize tumoral antigens for diagnostic purposes.

## A. Conclusion and Future Enhancement

In this work, Brewpitopes has been adapted from a pathogen context to a cancer setting.

The epitope pannel obtained will be validated in 300 plasma samples from patients with lung nodules susceptible to developing lung adenocarcinoma.

The antibody immune signature discovered by this work could be used to detect early lung cancers.

## B. ACKNOWLEDGEMENTS

Greatly acknowledge the work of Evie Fitzsimmons on coordinating the NIMBLE clinical trial under the supervision of Prof Richard Lee and Dr Kevin Litchfield. Thanks to Anna Díaz-Rovia and Víctor Montal for the support in developing Brewpitopes. Lastly, thanks to Prof Víctor Guallar and Dr Eduard Porta for their supervision during Brewpitopes development.


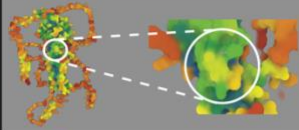

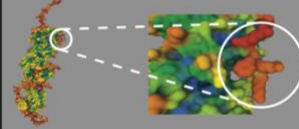
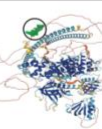
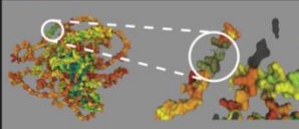
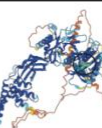
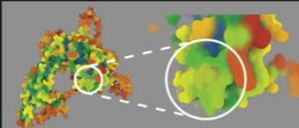

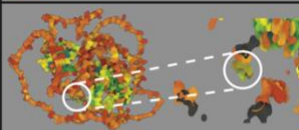

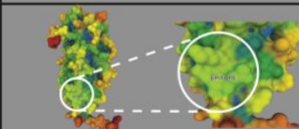


Protein	Prediction Efficiency	Epitope Location
<b>GRBT</b> Mediation of neuronal inhibition		
<b>GRIK4</b> Glutamate receptor ionotropic, kainate 4		
<b>PKHH2</b> Structural constituent of cytoskeleton		
<b>ANFK1</b> Ankyrin repeat and fibronectin type-III domain-containing protein 1		
<b>AJM1 (prev. C1172)</b> Apical junction component 1 homolog		
<b>FRIH</b> Ferritin heavy chain, N-terminally processed		
<b>KIF2B</b> Kinesin-like protein KIF2B		

Fig. 1 List of tumor-specific curated targets to design an epitope pannel for early lung cancer detection via antibody signatures in blood. In display, AlphaFold 2 prediction efficiency, epitope location and coloured the surface accessibility metric (RSA).

## References

[1] Farriol-Duran, R., López-Aladid, R., Porta-Pardo, E., Torres, A. & Fernández-Barat, L. Brewpitopes: a pipeline to refine B-cell epitope predictions during public health emergencies. *Frontiers in Immunology* 14, (2023).

## Author biography



Roc Farriol-Duran was born in Sabadell, Catalunya, in 1995. He received the B.E. degree in Biomedical Sciences from the Autonomous University of Barcelona, Barcelona, Catalunya, in 2018, the Msc. degree in Translational Biomedical Research for the Vall d'Hebron Hospital Campus (VH) Barcelona, Catalunya, in 2019 and the Msc. Degree in Omics Data

Analysis for the Vic University, Vic, Catalunya.

During 2016-17 he stayed at the Proteomics Group at the Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands. Followingly, he joined the Cellular Immunology Laboratory at the Institute for Biotechnology and Biomedicine, Autonomous University of Barcelona and collaborated closely with the Proteomics Unit at the CSIC-UAB. Then, he joined the lab of Tumor Immunology and Immunotherapy at the Vall d'Hebron Oncology Institute (VHIO). Since May 2020 he has been at the Life Sciences Department of the Barcelona Supercomputing Center, first at the Computational Biology group and later at the Electronic and Atomic Protein Modelling group. Currently, he is pursuing a PhD degree in Computational Modelling of T-cell and B-cell immunogenicity. The past 6 months Roc has undergone a placement at the UCL Cancer Institute under the mentorship of Dr Kevin Litchfield. Part of his work in the Tumor Immunogenomics and Immunosurveillance group is presented herein.



# Disentangling inter-individual transcriptome variability at single-cell and pseudobulk resolution

Lluís Frontera-Perello<sup>\*†</sup>, Aida Ripoll-Cladellas<sup>\*</sup>, Maria Sopena-Rios<sup>\*</sup>, Marta Melé<sup>\*</sup>

<sup>\*</sup>Department of Life Sciences, Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, 08034, Spain

<sup>†</sup>Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, 08005, Spain

E-mail: lluis.frontera@bsc.es, aida.ripoll@bsc.es, msopena@bsc.es, marta.mele@bsc.es

**Keywords**—*Transcriptomics, single-cell RNA-seq, aging, cell-to-cell variability, donor-to-donor variability, differential variability analysis.*

## I. EXTENDED ABSTRACT

Over the past 200 years, the average lifespan of humans has grown at a remarkable rate. As a result, age-related diseases are increasingly common in society, contributing to the burden of global health [1]. Despite being a difficult term to characterize, several hallmarks have been established regarding its cellular and molecular mechanisms [2]. Among them, one of the most documented ones is a gradual decline in immune function, leading to a higher risk of immune-related diseases or a permanent inflammation status (inflammaging) [3]. Previous studies using peripheral blood mononuclear cells (PBMCs) hypothesized that a combination of cell function decline, alterations in the extracellular matrix’s characteristics, and tissue disarray primarily caused by apoptosis, cellular communication, and epigenetic modifications would likely result in this immune dysregulation [4]. However, despite being highly accessible cell types usually used in large-scale population genomics studies, a full comprehension of the immune cellular and transcriptional changes accompanying aging is still lacking.

Aging has also been associated with changes in cell-to-cell heterogeneity [5] [6], defined as the difference in the measured level of variation in gene expression among cells supposed to be identical [5]. Single-cell technologies have been used to study the relationship between aging and transcriptional noise in a wide range of tissues and cell types [7] [8]. However, current studies show inconsistencies in age-related cell-to-cell gene expression variability [9] [10]. The use of different computational methods along with the heterogeneity in the datasets limits our ability to conduct meaningful comparisons [5]. Therefore, developing a standardized strategy to study how age affects transcriptional heterogeneity at cellular resolution is necessary to resolve these inconsistencies.

To fill this gap, we aim to use the OneK1K cohort [11], the largest available scRNA-seq dataset, to characterize the role of aging in gene expression variability. This comprehensive dataset, which encompasses more than one million human peripheral blood mononuclear cells (PBMCs) from 982 individuals spanning a wide range of ages, allows us to profile transcriptional noise at an unprecedented cellular resolution. Our results will offer a comprehensive characterization of how age influences transcriptional noise across immune cell types both at the cellular and donor level, providing valuable insights into the molecular and cellular alterations underlying aging in the immune system.

## A. Methodology

To pursue our objective, we use the OneK1K dataset. It consists of scRNA-seq data from 1.27 million PBMCs collected from 982 donors with Northern European ancestry, categorized according to their transcriptional profiles into 14 distinct immune cell types spanning the myeloid and lymphoid lineages.

To model cell-to-cell variability we use Scran [11], benchmarked as the best-performing method. The method assumes cell-to-cell variability (CCV) can be decomposed into biological and technical components by analyzing the total expression variance against log-normalized data. This fitted value for each gene is used as a proxy for the technical component of variation. The biological component of the CCV (CCV biological) is then defined as the residual from this trend. The resulting biological component is used to compute the difference in CCV biological ( $\delta$ ) between two conditions:  $\delta = \text{CCV biological (Old)} - \text{CCV biological (Young)}$ . Finally,  $\delta$  is normalized as a z-score and further converted to a p-value under a normal distribution.

To model donor-to-donor variability, we use an approach developed by Pique-Regi et al. [12]. This method uses a pseudobulk approach by aggregating gene counts of each individual to effectively account for zero inflation. Once aggregated, a linear regression statistical model is fitted for each individual following two steps: (1) Use a negative binomial model to obtain the estimation of the mean and dispersion per gene and remove the dispersion that can be predicted by mean expression, obtaining the residual dispersion. (2) Fit a linear model between mean and residual dispersion across genes to detect differential variable genes (DVGs) by applying a log2 transformation. DVG are identified based on a false discovery rate (FDR) threshold of 0.05.

## B. Results

Cell-to-cell variability results using Scran identify 186 differentially variable genes across cell types. Surprisingly, all of them showed a higher variability in young individuals. In order to validate our results, we performed permutations 10 times and obtained a higher number of DVGs in many of them, indicating the approach was not successful. We hypothesize this is most likely caused by the high dataset’s technical variability due to its sample size.

Given the poor performance of a single-cell approach, we decided to explore variability changes at the donor level.

We performed a donor-to-donor variability analysis and found 4294 differential variable genes between young and old individuals. In particular, 1666 increased variability with age whereas 2628 were less variable. We are currently validating those results to show whether gene expression variability with age could be studied in large-scale datasets using a pseudobulk approach.

### C. Conclusion

In this project, we study expression variability with aging both at the cellular and the donor levels and show that the currently available methodologies are not suitable for large scRNA-seq datasets. We believe that novel methodologies must be developed to provide a meaningful contribution to elucidating the transcriptional variation with aging at the single-cell resolution.

### REFERENCES

- [1] D. Melzer *et al.*, “The genetics of human ageing,” *Nature Reviews Genetics*, vol. 21, pp. 88–101, 2 2020.
- [2] D. Aw *et al.*, “Immunosenescence: Emerging challenges for an ageing population,” *Immunology*, vol. 120, pp. 435–446, 4 2007.
- [3] C. P. Martinez-Jimenez *et al.*, “Aging increases cell-to-cell transcriptional variability upon immune stimulation.” [Online]. Available: <https://www.science.org>
- [4] B. M. Owen *et al.*, “Evaluation of quantitative biomarkers of aging in human pbmcs,” *Frontiers in Aging*, vol. 4, 2023.
- [5] O. Ibáñez-Solé *et al.*, “Lack of evidence for increased transcriptional noise in aged tissues,” *eLife*, vol. 11, 12 2022.
- [6] H. Zheng *et al.*, “Measuring cell-to-cell expression variability in single-cell rna-sequencing data: a comparative analysis and applications to b cell aging,” *Genome Biology*, vol. 24, 12 2023.
- [7] M. Enge *et al.*, “Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns,” *Cell*, vol. 171, pp. 321–330.e14, 10 2017.
- [8] M. C. Salzer *et al.*, “Identity noise and adipogenic traits characterize dermal fibroblast aging,” *Cell*, vol. 175, pp. 1575–1590.e22, 11 2018.
- [9] R. Bahar *et al.*, “Increased cell-to-cell variation in gene expression in ageing mouse heart,” *Nature*, vol. 441, pp. 1011–1014, 6 2006.
- [10] M. Ximerakis *et al.*, “Single-cell transcriptomic profiling of the aging mouse brain,” *Nature Neuroscience*, vol. 22, pp. 1696–1708, 10 2019.
- [11] S. Yazar *et al.*, “Single-cell eqtl mapping identifies cell type-specific genetic control of autoimmune disease,” *Science*, vol. 376, 4 2022.
- [12] J. A. Resztak *et al.*, “Genetic control of the dynamic transcriptional response to immune stimuli and glucocorticoids at single-cell resolution,” *Genome Research*, vol. 33, pp. 839–857, 6 2023.



**Lluís Frontera** graduated with a Bachelor’s degree in Genetics from Universitat Autònoma de Barcelona (UAB) in 2018. The following year, he pursued an MC in Bioinformatics for Health Sciences at the Universitat Pompeu Fabra (UPF) in the same city. He is currently a researcher trainee at the Barcelona Supercomputing Center (BSC) in Spain, engaged on his master’s thesis in the Transcriptomics and Functional Genomics lab (TFGL).

# Towards Pareto Optimal Throughput in Small Language Model Serving

Pol G. Recasens<sup>#1</sup>, Yue Zhu<sup>\*2</sup>, Chen Wang<sup>\*3</sup>, Eun Kyung Lee<sup>\*4</sup>, Olivier Tardieu<sup>\*5</sup>,  
Alaa Youssef<sup>\*6</sup>, Jordi Torres<sup>#7</sup>, Josep Ll. Berral<sup>#8</sup>

<sup>#</sup>*Barcelona Supercomputing Center (BSC), Plaça d'Eusebi Güell, 1-3, 08034, Barcelona, Spain*

<sup>1</sup>pol.garcia@bsc.es, <sup>7</sup>jordi.torres@bsc.es, <sup>8</sup>josep.berral@bsc.es

<sup>\*</sup>*IBM Research, Thomas J. Watson Research Center, Yorktown Heights, New York, U.S*

<sup>2</sup>Yue.Zhu@ibm.com, <sup>3</sup>Chen.Wang@ibm.com, <sup>4</sup>eunkyoung.kee@us.ibm.com,

<sup>5</sup>tardieu@us.ibm.com, <sup>6</sup>asyoussef@us.ibm.com

**Keywords**— Language Models, Inference Optimization.

## EXTENDED ABSTRACT

Large language models (LLMs) have revolutionized the state-of-the-art of many different natural language processing tasks. Although serving LLMs is computationally and memory demanding, the rise of Small Language Models (SLMs) offers new opportunities for resource-constrained users, who now are able to serve small models with cutting-edge performance. In this paper, we present a set of experiments designed to benchmark SLM inference at performance levels. Our analysis provides a new perspective in serving, highlighting that the small memory footprint of SLMs allows for reaching the Pareto-optimal throughput within the resource capacity of a single accelerator. In this regard, we present an initial set of findings demonstrating how model replication can effectively improve resource utilization for serving SLMs.

### A. Introduction

Although the success of LLMs was traditionally attributed to scale, recent research suggests that a curated dataset might play an important role in training high-performance models. This paradigm shift, coupled with new serving optimization strategies, holds a substantial impact for a resource-constrained user, that is now able to serve SOTA small models. This rise of Small Language Models (SLMs) represents a significant step forward in making AI more accessible.

Despite the smaller size of SLMs, the incremental decoding of autoregressive language models limits the serving performance. Due to data dependencies in the self-attention layer, we process a single token per iteration, leading to matrix vector operations. This, coupled with the large cost of loading the model weights from memory, leads to very low arithmetic intensity during single-batch inference. One way to increase the arithmetic intensity, defined as the ratio between arithmetic operations and bytes accessed, is to batch requests and compute multiple tokens for the same transfer of weights. How large batches affect the serving performance of the less memory-demanding SLMs has yet to be explored.

However, batching techniques demand memory to store key-value pairs of previously processed tokens. The space in memory dedicated to store the intermediate results of previous tokens is known as KV cache, and handling it naively leads to memory fragmentation. PagedAttention algorithm identified this challenge and effectively reduced memory waste by dividing the KV cache in blocks, allowing to store KV pairs in non-contiguous memory space. In our experiments, we leverage vLLM [1], a high-throughput online serving engine

based on PagedAttention [1], to guarantee achieving the maximum batch size from our computational resources.

In this work, we benchmark SLM inference at performance level. In this regard, we serve OPT models ranging from 125M to 13B parameters in various online scenarios, sending requests generated from the ShareGPT dataset. We characterize the throughput and latency trade-off when the small memory footprint allows for large batches of requests. To the best of our knowledge, this provides a novel perspective, as previous inference benchmarking works are limited and primarily focused on large-scale serving. From our results, we observe that the Pareto-optimal throughput with small models is reached within the resource capacity of a single accelerator. This paves the way to new optimizations, such as partitioning of GPU resources in multi-model serving. In this context, we present an initial set of findings demonstrating how model replication can improve resource utilization for serving SLMs.

### B. Background

Performance of an inference step on a given processor can be memory-IO bound, limited by the time spent accessing memory, or compute bound, limited by the time spent computing operations. The metric used to measure the limiting factor is the arithmetic intensity, defined as the ratio ops:byte between compute operations and bytes transferred from HBM memory. Due to the low arithmetic intensity in the autoregressive generation phase, the performance of single-batch inference is commonly classified as memory-IO bound. This scenario is frequently found in memory limited scenarios, with large models and small batch sizes. Therefore, as long as memory-IO time overlaps compute time, we can theoretically increase the arithmetic intensity and improve the serving throughput without affecting end-to-end latency.

However, as we increase the inner size of the matrix-matrix operations with larger batches, compute might become important. For instance, a linear layer with a large batch is usually limited by arithmetic. This might also be influenced by how continuous batching sequentially processes attention operations of different requests. With large batches compute time grows to be larger the memory-IO time, reaching a Pareto-optimal throughput frontier. Beyond that point, further increasing of the batch size does not improve the serving performance.

### C. Experimental setup

The goal of our experiments is to characterise the performance of serving SLMs. Since these small models require significantly less memory than larger LLMs, we expect to batch enough requests to reach the throughput

frontier within the memory available in a high-end accelerator. In our experiments, we increasingly allocate more memory to the serving system via distributed inference to find the Pareto optimality point.

We employ models from the OPT family, with sizes ranging from 125M to 13B parameters. Introduced by MetaAI in May 2022, OPT belongs to decoder-only architectures such as GPT-3. The weights of the models are provided by the Transformers library from HuggingFace. We use an internal IBM cluster of machines composed of 4 NVIDIA A100 GPU's interconnected with NVLink. Each GPU has 40GB of HBM and a GPU memory bandwidth of 1555GB/s. In distributed inference, vLLM leverages the Megatron-LM's tensor parallel algorithm.

We generate 500 requests from the ShareGPT dataset, a collection of real conversations with ChatGPT. The prompt of each request is composed by 512 input tokens, and we limit the generation to 256 output tokens. This synthetic workload is intentionally restricted to study the effect of larger batches in a straightforward and consistent manner, providing an estimate of the amount of memory required to reach the optimal throughput. Each request is tokenized and sent to the vLLM engine through an http request, simulating a real-world deployment scenario. Although we include experiments with different arrival rates, we primarily focus on sending all the requests concurrently for various batch settings, helping to evaluate different batch configurations. It is worth noting that models larger than OPT-125M cannot process the 500 requests concurrently with a single GPU.

#### D. Results

We are interested in observing how large batches affect to the inference performance. While it proves challenging to batch even a small number of requests in LLM serving, SLMs introduce a unique scenario where a single accelerator can manage the memory requirements for storing a larger number. The performance implications of this aspect have yet to be explored. In this first analysis, we benchmark the system's performance when serving small models of increasing size for different batch sizes. If the batch size cannot be achieved with a single accelerator, we distribute the serving across multiple GPUs to increase available memory.

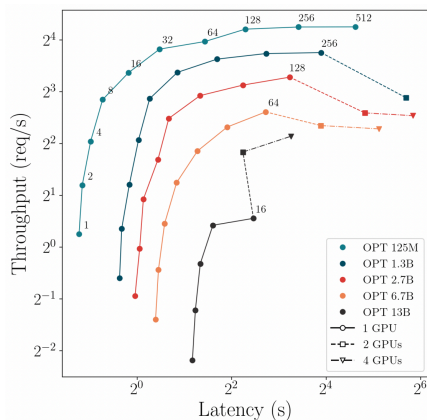


Fig. 1 Throughput and latency trade-off for models of increasing size and batch sizes in powers of two. With small models we observe that throughput reaches a Pareto-optimal frontier within the resource capacity of a single accelerator.

GPU resources are scarce, and our previous results show that over provisioning memory to SLMs, therefore increasing the batch size, does not necessarily correlate to a performance improvement. With this knowledge in hand, we can limit the memory allocated to each model, and run different models in the same accelerator, or replicate the same model with multiple instances. We provide a first set of findings on how we can leverage model replication to improve end-to-end serving performance.

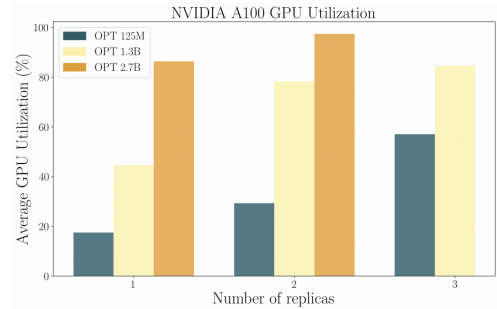


Fig. 2 Average GPU utilization when serving OPT 125M, OPT 1.3B and OPT 2.7B in an NVIDIA A100 GPU. The model is replicated up to three times, if possible.

#### E. Conclusion

This paper characterizes the serving performance of SLMs, highlighting the implications of memory allocation on inference throughput. Our analysis shows that for small models a single high-end accelerator has enough memory to reach a Pareto-optimal throughput frontier given a large batch of requests. Beyond that point, allocating more memory results in minimal or no improvements. In light of our results, we pave the way for new optimizations in model serving, presenting an initial set of findings that show how model replication on a single device improves overall inference performance. Further analysis should consider a more realistic serving scenario with heterogeneous requests and devices, and explore model replication with more suitable techniques.

#### ACKNOWLEDGEMENTS

This work has been partially financed by grant agreement EUHORIZON GA.101095717 and by the EU-HORIZON MSCA programme under grant agreement EU-HORIZON MSCA GA.101086248. Also, it has been partially financed by Generalitat de Catalunya (AGAUR) under grant agreement 2021-SGR-00478, and by the Spanish Ministry of Science (MICINN), the Research State Agency (AEI) and European Regional Development Funds (ERDF/FEDER) under grant agreement PID2021-126248OB-I00, MCIN/AEI/ 10.13039/501100011033/FEDER, UE.

#### References

- [1] Kwon, Woosuk, et al. "Efficient memory management for large language model serving with pagedattention." Proceedings of the 29th Symposium on Operating Systems Principles. 2023.

# ASCOM: Affordable Sequence-aware COntention Modeling in Crossbar-based MPSoCs

Jeremy Giesen<sup>\*†</sup>, Enrico Mezzetti<sup>\*</sup>, Jaume Abella<sup>\*</sup>, Francisco J. Cazorla<sup>\*</sup>

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {jeremy.giesen, enrico.mezzetti, jaume.abella, francisco.cazorla}@bsc.es

**Keywords**—Crossbar, multicore, contention, software timing analysis.

## I. EXTENDED ABSTRACT

Multicore interference that arises when several accesses contend for the same shared hardware resources poses a challenge to the already demanding consolidated verification and validation practice. The Sequence-Aware Pairing (SeAP) model approach exploits the parallelism granted by crossbars to derive tighter contention bounds. We show that SeAP suffers from scalability issues that hinders its applicability to more complex contention scenarios. We address SeAP limitations in terms of scalability by identifying two complementary techniques to reduce SeAP execution time requirements. We assess the proposed approaches to show how they effectively enable the application of SeAP to large sequences of accesses to the crossbar with limited impact on tightness, and scaling gracefully with the number of co-running cores.

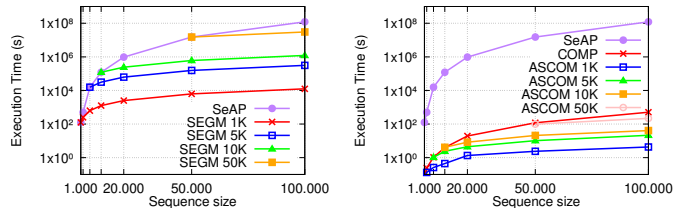
### A. Introduction and Motivation

Accurately estimating the Worst-Case Execution Time (WCET) [1] in Multicore Processor Systems on Chip (MP-SoCs) is a formidable task, especially considering the extensive hardware resource sharing inherent in such systems. Effective contention analysis is crucial before MPSoCs can be confidently deployed in high-integrity systems.

Contending accesses to Hardware Shared Resources (HSRs) in MPSoCs demand sophisticated modeling approaches. One such approach, the Sequence-Aware Pairing (SeAP) [2], exploits the precedence relation between accesses to enhance contention modeling. However, SeAP’s computational complexity, inherited from the pattern-matching methods it relies on, limits its scalability. Real-world multicore scenarios with numerous cores and extensive access sequences exacerbate this challenge.

### B. SeAP

SeAP’s reliance on precedence relations between accesses marks an improvement over models based solely on access counts [2], [3], [4], [5]. However, its computational demands hinder its practical application in large-scale scenarios. While SeAP provides tighter contention bounds, its scalability issues become evident as the number of cores and access sequences increase.



(a) SeAP and SEGM

(b) SeAP, COMP, and ASCOM

Fig. 1: Execut. time of SeAP, COMP, SEGM & ASCOM.

### C. ASCOM

SeAP’s scalability challenges prompt the development of ASCOM, which comprises two key techniques: Compositionality (COMP) and Segmentation (SEGM). COMP leverages the additive nature of contention timing interference, simplifying SeAP’s complexity through a divide-and-conquer approach. SEGM partitions access sequences into fixed-size segments, reducing computational demands while maintaining accuracy.

**COMP** exploits the additive nature of contention timing interference, significantly reducing SeAP’s computational complexity. By independently analyzing sequences of contending accesses, COMP achieves scalability without compromising accuracy.

**SEGM** addresses SeAP’s scalability limitations by segmenting access sequences into smaller, manageable units. By applying SeAP to these segments, SEGM achieves a balance between computational efficiency and accuracy.

### D. Experimental Evaluation

We conduct detailed experiments to assess the performance of COMP and SEGM across various factors, such as sequence size, number, dictionaries, and distribution patterns, focusing on scalability, accuracy, and practicality of ASCOM in multicore systems.

Figure 1b illustrates the time efficiency of COMP compared to SeAP in handling three sequences with sizes ranging from 1K to 100K elements, using a fixed sequence dictionary and shape. COMP remains efficient (under 10 minutes) even for larger sequences, while SeAP struggles, taking about a day for 10K elements.

For accuracy, Figure 2 indicates COMP’s relative overestimation compared to SeAP, with details on the Maximum Theoretical Overestimation (MTO) from Table I. Overestimation varies from 4% to 17%, averaging at 9.5%.

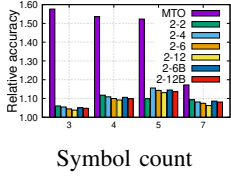
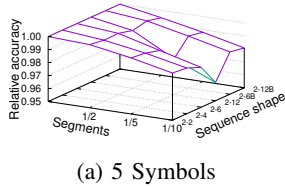
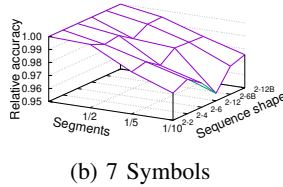


Fig. 2: COMP accuracy.

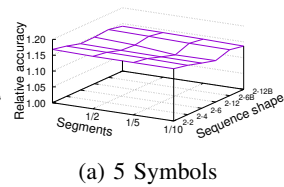


(a) 5 Symbols

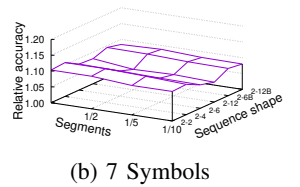


(b) 7 Symbols

Fig. 3: SEGM relative accuracy for three sequences of 10k elements.



(a) 5 Symbols



(b) 7 Symbols

Fig. 4: ASCOM relative accuracy for three sequences of 10k elements.

TABLE I: Real ( $W$ ) and forced ( $W^+$ ) linearity weight function.

Slowdown	1 Request				2 Requests					
	R		W		R+R		R+W		W+W	
	$W$	$W^+$	$W$	$W^+$	$W$	$W^+$	$W$	$W^+$	$W$	$W^+$
LMU Read	1	3	3	4	4	4	6	6	8	8
LMU Write	1	4	3	5	5	5	7	7	9	9
PFlash Read	4	6	n/a	n/a	11	11	n/a	n/a	n/a	n/a
DFlash Read	34	35	n/a	n/a	69	69	n/a	n/a	n/a	n/a

### 1) COMP:

*Execution time:* Figure 1b compares the time required by SeAP on 3 sequences or by leveraging compositionality with COMP, for a fixed sequence dictionary and shape, by only varying the sequence sizes from 1K to 100K elements. COMP timing requirements remain affordable (<10 minutes) even for larger sequences, whereas SeAP shows limitations with average-size sequences ( $\sim 1$  day for 10K elements).

*Accuracy:* Figure 2 displays the relative increase (overestimation) incurred by COMP compared to SeAP. Each plot also presents the Maximum Theoretical Overestimation (MTO), representing an upper bound to the overestimation incurred by enforcing linearity in the specific weight function  $W$  in Table I. The observed overestimation ranges between 4% and 17% in the worst case, with an average overestimation of 9.5%.

### 2) SEGM:

*Execution time:* Figure 1a shows SeAP’s execution times on a logarithmic scale for a 3-sequence scenario with varying sequence sizes, comparing them to SEGM performance across segment sizes from 1K to 50K elements. The execution time reduction is proportional to the segment ratio; halving the sequence size results in halving the execution time. The choice of segment size should consider the impact on accuracy.

*Accuracy:* As we cannot extrapolate, to obtain a reference SeAP result over full sequences, we focus on sequences with 10K elements. Surface plots in Figure 3 show accuracy across varying numbers of segments (1, 2, 5, and 10) and different sequence shapes/distributions, using various symbol/device dictionaries. SEGM consistently provides accurate results, with an average underestimation of 0.67%, peaking at 4.42% in the 7-symbol setup divided into 2-12, 1/10 segments.

### 3) ASCOM:

*Execution time:* Figure 1b shows that combining SEGM and COMP significantly improves execution time compared to using either alone for variable-sized sequences in a 3-sequence scenario. This advantage over SeAP increases with the segment ratio, as shown in Figure 1a.

*Accuracy:* Results for three sequences with 10,000 elements show that the accuracy impact of segmenting is minor and less reliant on the sequence shape/distribution. Figure 4 highlights COMP’s dominance, as shown in the left wall of each plot and detailed for COMP alone in Figure 2. When combined with COMP, the accuracy loss from segmenting (SEGM) is negligible (about 1% less), reducing the problem effectively to a two-sequence problem where SEGM’s effect is minimal.

### E. Conclusions

In conclusion, our study presents ASCOM as a comprehensive framework for scalable contention analysis in multicore systems. By leveraging innovative techniques like COMP and SEGM, ASCOM offers practical solutions to longstanding challenges, paving the way for enhanced verification and validation practices in modern computing environments.

## II. ACKNOWLEDGMENT

This work has been published in proceedings of the 38th ACM/SIGAPP Symposium On Applied Computing (SAC), 2023 [6].

## REFERENCES

- [1] Wilhelm R. et al., “The worst-case execution-time problem: overview of methods and survey of tools,” *ACM Trans. on Embedded Comp. Systems*, 2008.
- [2] J. Giesen *et al.*, “Modeling contention interference in crossbar-based systems via sequence-aware pairing (SeAP),” in *RTAS*. IEEE, 2020.
- [3] D. S. Hirschberg, “A linear space algorithm for computing maximal common subsequences,” *Commun. ACM*, 1975.
- [4] G. Jacobson and K.-P. Vo, “Heaviest increasing/common subsequence problems,” in *Combinatorial Pattern Matching*. Springer Berlin Heidelberg, 1992.
- [5] R. Li, “A linear space algorithm for the heaviest common subsequence problem,” *Utilitas Mathematica*, vol. 75, 03 2008.
- [6] J. J. Giesen Leon, E. Mezzetti, J. Abella, and F. J. Cazorla, “Ascom: Affordable sequence-aware contention modeling in crossbar-based mpsoCs,” in *38th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC ’23. Association for Computing Machinery, 2023, p. 471–474.



**Jeremy Giesen** received his BSc degree in Computer Engineering from Universidad de Las Palmas de Gran Canaria (ULPGC), Spain in 2018. The same year, he joined the Barcelona Supercomputing Center as a research student while studying a MSc degree in Innovation and Research in Informatics in the Universitat Politècnica de Catalunya (UPC). He completed his MSc degree in 2020. Since 2020, he has been a PhD student in the department of computer architecture of Universitat Politècnica de Catalunya (UPC), Spain.

# An AGS cell line digital twin for studying novel treatment strategies

Othmane Hayoun-Mya\*, Miguel Ponce-de-León\*, Arnau Montagud\*, Alfonso Valencia\*†

\*Barcelona Supercomputing Center, Barcelona, Spain

†ICREA, Pg. Lluís Companys, 23, 08010 Barcelona, Spain

E-mail: {ohayoun}@bsc.es

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

## I. EXTENDED ABSTRACT

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 implementation of a digital twin of the AGS cell line for exploring multidrug resistance in this cancer and novel therapeutic strategies. Building off of the computational and experimental work on novel drug synergies in Gastric Adenocarcinoma cells from [1], we implement a PhysiBoSS [2] multiscale agent-based simulation 3D model calibrated on experimental data that replicates their found drug synergies. With this calibrated model, we set to explore the efficacy of said synergies in light of a heterogeneous resistant population.

### A. Implementing a multiscale model of AGS

In order to set up a computational template that reflects the experimental single and combined drug assays of [1] on PI3K, MEK, AKT and TAK1. For this, 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. 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], embedded 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 simultaneous combination of pro-survival and anti-survival nodes that affect growth rate and apoptosis rates of a given agent, respectively. 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 Boolean model, we developed four Hill-shaped transfer functions: 1) A transfer function for the probability of deactivating a specific node of

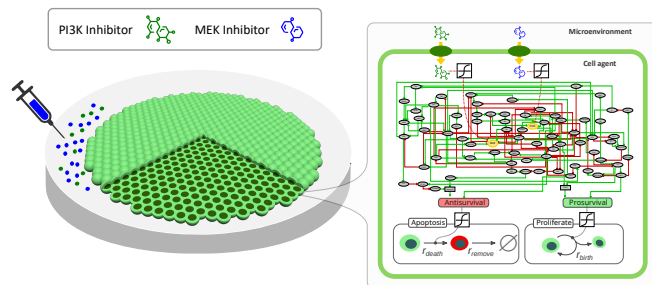


Fig. 1. Diagram showcasing the fully-implement PhysiBoSS AGS model. The disk of cells is the initial setup of the simulation. Zooming into a single one of the agents, on the right, there is an embedded Boolean model representing major signaling pathways of gastric cancer. Here, external inputs such as drug density imported are mapped to changes in specific nodes through a Hill function. Similarly, readouts from the Boolean model are mapped to agent phenotype rules.

the Boolean network according to the internal drug concentration. 2) and 3) Transfer functions that affect growth and apoptosis rate according to the pro-survival and anti-survival readouts, respectively. 4) A transfer function was also used to include contact-inhibition in our simulations that maps growth rate to pressure, reflecting cell confluence in the simulation.

After building the PhysiBoSS model, we set to calibrate its many parameters in order to faithfully replicate the experimental growth curves from [1].

### B. Simulation-based model calibration

Given the complexity of our simulation system shown in Section I-A, and the amount of free parameters, mostly related to Transfer function parameters, along with drug-specific ones such as drug 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 the Genetic Algorithm (GA) and Covariance Matrix Adaptation (CMA-ES) strategies in order to find the best sets of parameters that replicate the experimental growth dynamics of single-drug experiments from [1]. Using the experimental dimensionless Cell Index counts as our ground truth, our objective function is to minimize the RMSE between the simulated and experimental normalized curves (See Fig. 2). We perform this calibration on the single-drug experimental results for the PI3K, MEK and AKT inhibitors. We first perform a general calibration of all 14 parameters, from which we obtain a subset of common parameters that show good fittings for all three drugs. Then, we perform a parameter sweep for fine-tuning the three remaining drug-specific parameters.

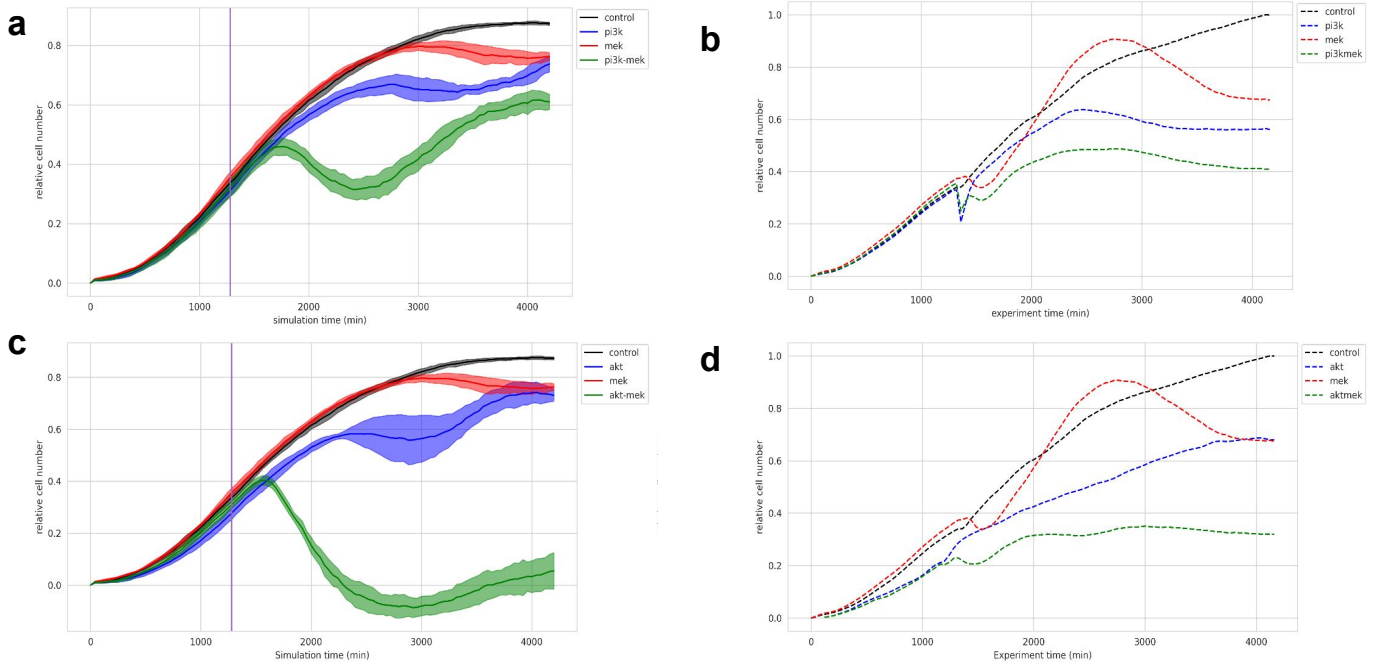


Fig. 2. Synergy comparison between experimental and simulation results. **a.** Simulated results from a combinatorial assay with PI3K and MEK inhibitors. **b.** Experimental results from [1] of this same experiment. **c.** Simulated results from a combinatorial assay with AKT and MEK inhibitors. **d.** Experimental results from this same experiment. Horizontal line in plots **a** and **c** indicates time of drug addition. In simulation results, standard deviation was also included.

### C. Results

From the calibration with the full set of parameters (14) within our model, we were able to correctly fit the experimental curves. Moreover, we found a subset of AGS-specific parameters, identical for all three calibrated drugs, that provided a good fitting (See Fig. 2). The subsequent fine-tuning of these parameters showed not only an improved calibration for each single-drug experiment, but it also allowed us to find sets of drug-specific parameter values that would correctly fit the experimental data and also showcase the same synergies observed in [1] for the PI3K-MEK and AKT-MEK combinatorial inhibitor experiments (Fig. 2).

The calibrated model is currently being employed for implementing a heterogeneously resistant population, in order to explore the critical effectiveness of state-of-the-art drug synergy-based treatments in light of a more realistic tumoral setting, with innate and acquired resistance to the injected drugs. We see a correct implementation of our mutational and heterogeneity addition, and results regarding simulations with this settings are currently being performed.

### D. Conclusion

The present work aims to push forward the development of *in silico* tools that allow for biologically realistic high-throughput hypothesis testing in the context of drug resistance emergence and overcoming. By integrating many different sources of data at different biological and mechanistic levels, our model can replicate experimental drug synergies within the AGS cell line.

We argue that this is a modular *in silico* 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 advance research on novel cancer therapies and the much-needed overcoming multi-drug resistance.

## II. ACKNOWLEDGMENT

This work has been supported by the PerMedCoE project (grant agreement N°951773).

## 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 doing his PhD on this same research line within the Computational Biology group at the Life Sciences Department of the Barcelona Supercomputing Center (BSC-CNS).



# Multiple-Copies Association Studies for Computational Binding Mode Elucidation of Fbw7 E3 Ligase Fragment Hits

Varbina Ivanova<sup>1,2#</sup>, Roger Castaño<sup>1,3</sup>, Carles Galdeano<sup>1,3</sup>, Jordi Juárez-Jiménez<sup>1,2</sup>, Xavier Barril<sup>1,2,3,4</sup>

<sup>1</sup>*Department of Pharmacy and Pharmaceutical Technology, and Physicochemistry, Faculty of Pharmacy and Food Sciences, University of Barcelona, Spain*

<sup>2</sup>*Institute of Theoretic and Computational Chemistry (IQTC), University of Barcelona, Spain*

<sup>3</sup>*Biomedicine Institute of University of Barcelona (IBUB), Spain*

<sup>4</sup>*Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain*

#varbina.ivanova@ub.edu

**Keywords**— E3 ligases, fragment binding mode, Fbw7, HPC, MAS

## EXTENDED ABSTRACT

Targeted Protein Degradation is a promising therapeutic approach for the regulation of proteins involved in cancer, neuropsychiatric diseases and other disorders. [1] [2] At the core of this approach, lies the development of small molecules able to engage E3 ubiquitin ligases. To date, however, many E3 ligases have been deemed "undruggable". In this context, the combination of fragment-based drug discovery and allosteric regulation provides an appealing path toward targeting undruggable proteins. [3]

In this work, we develop a structure-based computational approach to elucidate fragment binding modes, aimed at the rational design of novel ligands targeting the Fbw7 E3 ligase. Fbw7 acts as a crucial tumor suppressor by facilitating the ubiquitination of key oncogenes such as Cyclin E, C-Myc, and Notch1, along with other vital proteins like DISC1. [4] [5]

Previously, our research group had identified 10 potent Fbw7 fragment hits, but crystallography efforts to determine the complex structure have not been successful. Here we present a computational workflow for the elucidation of fragment binding modes through Multiple-copies Association Studies (MAS).

MAS combines classical molecular dynamics (MD) simulations with elevated ligand concentrations and employs a post-MD clustering algorithm to rank the probable binding modes and determine the correct one. We have developed a novel LJ potentials adjustment approach to mitigate small molecule aggregation typically observed in water simulations with high concentrations of organic molecules, enabling MAS to be conducted at high concentrations of ligand.

To further enhance the efficiency and scalability of our approach, we harness the power of High-Performance Computing (HPC), which enables us to conduct MAS at higher scales. By leveraging HPC resources, we accelerate the production and processing of the MAS molecular dynamics simulations, facilitating the rapid identification and validation of ligand binding modes.

With the help of HPC, the Multiple-copies Association Studies approach holds promise for uncovering unknown binding modes for known small-molecule binders. In this study, we first focus on the optimization and validation of the MAS by utilizing a diverse benchmarking set with protein-ligand systems with crystallographic and binding affinity data. Subsequently, we present the results of employing the Multiple-copies Association Studies (MAS) approach to elucidate the binding modes of Fbw7 fragment hits previously identified using various biophysical techniques.

## References

- [1] Saravanan KM, Kannan M, Meera P, Bharathkumar N, Anand T. E3 ligases: a potential multi-drug target for different types of cancers and neurological disorders. *Future Med Chem.* 2022;14(3):187-201.
- [2] Békés M, Langley DR, Crews CM. PROTAC targeted protein degraders: the past is prologue. *Nat Rev Drug Discov.* 2022;21(3):181-200.
- [3] Xie X, Yu T, Li X, et al. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal Transduct Target Ther.* 2023;8(1):335.
- [4] Welcker M, Clurman BE. FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. *Nat Rev Cancer.* 2008;8(2):83-93.
- [5] Yalla K, et al. FBXW7 regulates DISC1 stability via the ubiquitin-proteasome system." *Molecular Psychiatry* 23.5 (2018): 1278-1286.

## Author biography



**Varbina Ivanova** was born in Sofia, Bulgaria, in 1997. She received the B.E. degree in chemical engineering from the Sofia University, Sofia, Bulgaria, in 2020, and the M.S. degree in computational chemistry from the Sofia University, Sofia, Bulgaria, in 2021.

Since October 2021, she has been with the Department of Pharmacy and Pharmaceutical Technology, and Physicochemistry, University of Barcelona, where she was a Visiting Researcher and later became a Marie Curie PhD fellow and part of ITN ALLODD as an early stage researcher (ESR). Her current research interests include allostery in drug discovery, computer-aided drug discovery, fragment-based screening, and targeted protein degradation.

# Architecture-aware Patterns for the Factorized Sparse Approximate Inverse Preconditioner

Sergi Laut\*<sup>†</sup>, Ricard Borrell, Marc Casas\*<sup>†</sup>

\*Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {sergi.lautturon, marc.casas}@bsc.es, rickborrell@gmail.com

*Keywords*—Conjugate Gradient, FSAI, SpMV, CPU, GPU

## I. EXTENDED ABSTRACT

The Conjugate Gradient (CG) method is an iterative solver targeting linear systems of equations  $Ax = b$  where  $A$  is a Symmetric and Positive Definite (SPD) matrix. CG convergence properties improve when preconditioning is applied to reduce the condition number of matrix  $A$ . The Factorized Sparse Approximate Inverse (FSAI) preconditioner constitutes a highly parallel option based on approximating  $A^{-1}$ . FSAI is applied through two Sparse Matrix-Vector (SpMV) products in each iteration of the preconditioned CG. The SpMV kernel is memory-bound and is heavily influenced by the irregular memory access patterns on  $x$ , which are driven by the locations of the sparse matrix non-zero coefficients. A very important aspect of FSAI is the definition of its corresponding sparse pattern. While state-of-the-art solutions define this pattern by exclusively taking into account numerical considerations, we consider that low-level architecture-aware concepts should also be taken into account.

In this work, we propose and evaluate an approach to extend FSAI sparse patterns based on two fundamental concepts: First, an algorithm to extend sparse patterns that aim to reduce the CG iteration count while keeping the cost per iteration low. This optimization relies on low-level aspects of the cache hierarchy architecture, like indexing mechanisms or virtual memory management approaches for CPUs and the promotion of coalesced data accesses on GPUs. Second, an approach to filter out the smallest entries of the FSAI pattern extension without degrading its convergence properties.

### A. Introduction

The FSAI preconditioner is typically applied when solving SPD systems [1], [2]. FSAI approximates  $A^{-1}$  considering a factorization  $G^T G$ . Therefore, it requires computing two SpMV products in each iteration of the preconditioned CG,  $y = G^T Gx$ .  $G$  is a sparse lower triangular matrix approximating the inverse of the Cholesky factor  $L$  of  $A$ . Given a generic lower triangular sparse pattern  $\mathcal{S}$ , FSAI obtains  $G$  via the minimization problem,  $\min_{G \in \mathcal{S}} \|I - GL\|_F^2$ , where  $\|\cdot\|_F$  is the Frobenius norm. This problem can be solved independently for each row  $i$  of  $G$  by considering the local system  $A_{\mathcal{S}_i \mathcal{S}_i} g_i = e_i$ , where  $A_{\mathcal{S}_i \mathcal{S}_i}$  is the restriction of  $A$  to the coefficients of the  $i$ th row of the sparse pattern  $\mathcal{S}$ , and  $e_i$  is the  $i$ th column of the identity matrix restricted to the same space [2], [3].

Pattern  $\mathcal{S}$  is usually defined following numerical considerations. In Algorithm 1, we propose a redefinition of the process

---

**Algorithm 1** FSAI,  $G^T G \approx A^{-1}$  with pattern extension

---

- 1: Threshold  $A$  to produce  $\tilde{A}$ .
  - 2: Compute the pattern  $\tilde{A}^N$ , and let the pattern,  $\mathcal{S}$ , of  $G$  be the lower triangular part of the pattern of  $\tilde{A}^N$ .
  - 3: **Compute architecture-aware extension of the pattern of  $G$ ,  $\mathcal{S}_{ext}$ .**
  - 4: **Calculate an approximation  $\tilde{G}$  of the preconditioner, and filter out entries of  $\mathcal{S}_{ext}$  according to its values.**
  - 5: Calculate  $G$  on the sparse pattern obtained from the previous step.
- 

to compute FSAI [3]. We add Step 3, which consists of adding entries to the pattern  $\mathcal{S}$  of  $G$  following a defined strategy:

*Shared memory CPU context:* The extension must not increase the cache misses on accesses to the multiplying vector,  $x$ , in the SpMV product. This is achieved by adding entries to the pattern that require coefficients of  $x$  that are already loaded by the initial pattern but not used. Ultimately, this means that if an L1 cache line is loaded due to access on  $x$  and only one of the brought elements is used in the initial pattern, the extended pattern must ensure all of them are used. The cache line size is the main parameter for this extension. The added entries improve  $G$  spatial locality and  $G^T$  temporal locality. If, after the extension of  $G$ , the same process is followed on  $G^T$ , the spatial and temporal locality are improved in both FSAI SpMV products. We call this method the *Factorized Sparse Approximate Inverse with Pattern Extension (FSAIE)*.

*GPU context:* The extension must promote coalesced memory accesses and spatial locality. The SpMV product on GPUs is typically distributed by rows, where each thread in a warp computes one of them. In GPUs, performance is improved when all threads in a warp access the same or contiguous data. We obtain this by creating blocks of entries around the initial ones of the warp size. With these blocks, all threads in a warp require accessing the same entry in  $x$  at each step of the SpMV product. Spatial locality is improved as the following  $x$  accesses are contiguous for all threads. We call this method the *GPU-aware Factorized Sparse Approximate Inverse (GFSAI)*.

*Distributed memory context:* The extension must not increase communication costs. Entries added to the pattern of  $G$  must follow the CPU or GPU criteria and, in addition, not generate any new communications either in  $G$  or  $G^T$ . We call these methods *FSAIE-Comm* and *GFSAI-Comm*.

The extended patterns shall be used to compute  $G$ . Although the extended FSAI patterns deliver computational benefits in the SpMV product, some of the added  $G$  coefficients display small absolute values that have a negligible contribution to the PCG convergence while incurring useless memory accesses. For this reason, it is necessary to filter out the entries with small numerical contributions to the inverse approximation. This is Step 4 in Algorithm 1. The filtered  $G$  matrix is not the best approximation to the inverse of the filtered pattern, as  $\min_{G \in \mathcal{S}} \|I - GL\|_F^2$  does not hold anymore. For this reason, in Step 5,  $G$  is computed again.

## B. Methodology

We evaluate our methods by applying them to the CG solver and comparing them against FSAI. For all tests, the initial residual norm is reduced by eight orders of magnitude. We test different filtering-out values and report the averages for all matrices with their best filtering option.

We evaluate FSAIE on a single node in three machines based on Skylake, Power9, and A64FX architectures. We use a dataset comprising 72 matrices from different domains with a non-zero coefficient count between 48k and 4.8M. The cache line size in Skylake and Power9 is 64 bytes, while for A64FX, it is 256 bytes.

FSAIE-Comm is evaluated on three clusters based on Skylake, A64FX, and Zen 2 architectures. We use a dataset comprising 47 matrices from different domains with a non-zero coefficient count between 1M and 40M. We also test a larger dataset consisting of 8 matrices with non-zero entries between 40M and 320M on the Zen 2 cluster. The cache line size in Skylake and Zen 2 is 64 bytes, while for A64FX, it is 256 bytes. We select a hybrid configuration for each matrix of 8 CPU threads/cores per MPI process and a number of processes that allow for good scalability of the results.

GFSAI is evaluated on a single node with two different GPU architectures: NVIDIA V100 (Volta) and AMD MI50 (Vega 20). We use a dataset comprising 47 matrices from different domains with a non-zero coefficient count between 700k and 115M. The warp size for NVIDIA and AMD is 32 and 64, respectively.

TABLE I. AVERAGE ITERATION DECREASE, AVERAGE EXECUTION TIME DECREASE, HIGHEST TIME IMPROVEMENT, AND LARGEST PERFORMANCE DEGRADATION OF FSAIE AGAINST FSAI ON SKYLAKE, POWER9, AND A64FX. NUMBERS ARE PERCENTAGES.

Architecture	Avg. iter. decrease	Avg. time. decrease	Highest. time improvement	Highest. time degradation
Skylake	16.60	15.02	56.72	-2.06
Power9	15.15	12.94	56.72	-12.35
A64FX	24.91	22.85	76.99	-0.96

## C. Evaluation

Table I displays the results obtained on the shared memory CPU context. FSAIE obtains significant time-to-solution improvements compared to FSAI, with a very small chance of performance degradation. Skylake and Power9 64-byte cache lines make the results of the two architectures similar. A64FX has better results due to the larger 256-byte cache lines.

Table II displays the results obtained on the distributed memory CPU context. As in the shared memory context, the larger cache lines of A64FX lead to the best improvements.

TABLE II. AVERAGE ITERATION DECREASE, AVERAGE EXECUTION TIME DECREASE, HIGHEST TIME IMPROVEMENT, AND LARGEST PERFORMANCE DEGRADATION OF FSAIE-COMM AGAINST FSAI ON SKYLAKE, A64FX, AND ZEN 2. NUMBERS ARE PERCENTAGES.

Small Set				
Architecture	Avg. iter. decrease	Avg. time. decrease	Highest. time improvement	Highest. time degradation
Skylake	21.33	17.90	55.09	-0.34
A64FX	31.32	26.44	62.63	0.49
Zen 2	20.64	16.74	57.52	-1.05
Large Set				
Architecture	Avg. iter. decrease	Avg. time. decrease	Highest. time improvement	Highest. time degradation
Zen 2	13.89	12.59	19.09	3.94

TABLE III. AVERAGE ITERATION DECREASE, AVERAGE EXECUTION TIME DECREASE, HIGHEST TIME IMPROVEMENT, AND LARGEST PERFORMANCE DEGRADATION OF GFSAI AGAINST FSAI ON NVIDIA AND AMD. NUMBERS ARE PERCENTAGES.

Architecture	Avg. iter. decrease	Avg. time. decrease	Highest. time improvement	Highest. time degradation
NVIDIA	27.48	23.83	63.58	-1.00
AMD	31.25	26.07	68.09	-0.45

Table III displays the results obtained in the GPU context. The GPU extended  $G$  blocks, 32x32 for NVIDIA and 64x64 for AMD, allow for much larger pattern extensions than CPUs. For this reason, the averages in time-to-solution are much larger. The larger AMD warp size compared to NVIDIA improves the performance of GFSAI.

## II. ACKNOWLEDGMENTS

Marc Casas has been partially supported by the Grant RYC-2017-23269 funded by MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future. This research was supported by grant PID2019-107255GB-C21 funded by MCIN/AEI/10.13039/501100011033. The authors thank the support of Departament de Recerca i Universitats de la Generalitat de Catalunya to the Research Group "Performance understanding, analysis, and simulation/emulation of novel architectures" (Code: 2021 SGR 00865).

## REFERENCES

- [1] L. Y. Kolotilina, A. A. Nikishin, and A. Y. Yeremin, "Factorized sparse approximate inverse preconditionings. iv: Simple approaches to rising efficiency," *Numerical Linear Algebra With Applications - NUMERICAL LINEAR ALGEBRA APPL*, vol. 6, pp. 515–531, 10 1999.
- [2] L. Y. Kolotilina and A. Y. Yeremin, "Factorized sparse approximate inverse preconditionings i. theory," *SIAM Journal on Matrix Analysis and Applications*, vol. 14, no. 1, pp. 45–58, 1993. [Online]. Available: <https://doi.org/10.1137/0614004>
- [3] E. Chow, "Parallel implementation and practical use of sparse approximate inverse preconditioners with a priori sparsity patterns," *International Journal of High Performance Computing Applications*, vol. 15, 05 2001.



**Sergi Laut** received his BSc degree in Electronic Engineering from Universitat Politècnica de Catalunya (UPC) in 2016. After, he worked at the Superconductivity Group at Universitat Autònoma de Barcelona (UAB). He completed his MSc degree in Modelling for Science and Engineering from UAB in 2019. Since then, he has been with the Software research and development vehicles for New ARchitectures (SONAR) group of Barcelona Supercomputing Center (BSC) and is a Ph.D. student at the Department of Computer Architecture of UPC.

# LLMMM: Large Language Models Matrix-Matrix Multiplications Characterization on Open Silicon

Louis Ledoux\*<sup>†</sup>, Marc Casas\*<sup>†</sup>

\*Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {louis.ledoux,marc.casas}@bsc.es

**Keywords**—Large Language Models (LLM), Transformers, Generative Pre-Trained (GPT), Matrix-Matrix Multiplications, Floating-Points, arithmetic, ASIC, Open-Source Silicon (OSS)

## I. EXTENDED ABSTRACT

### A. Introduction

GPT transformers are useful for various applications, offering significant advancements in natural language processing tasks. However, their operational costs are substantial as shown in prior work which highlights the financial implications of deploying these models [1].

Essentially, matrix-matrix multiplications (MMM), with their intensive data movement and manipulation of arithmetic weights, underscore the computational demands of these architectures. Naturally, these observations are also found in recent efforts within the research community, which have concentrated on devising specialized formats and algorithms aimed at mitigating these costs. These innovations include reducing bit-width exemplified by Machine Learning eXchange (MLX) formats (essentially small floats), specialized hardware such as TPUs’ systolic arrays, model pruning of up to 40%, and more recently, ternary and binary LLMs (see BitNets [2]).

We introduce a generator of ASIC kernels agnostic to the PDK of MMM units for emerging and small floating-point formats, followed by the evaluation of such units. Concretely, our contributions include the automated generation of circuits for any floating-point format with automated pipelining, a systolic array architecture proposal—these two combined form the foundation of MMM units, a framework to automate the translation from high-level language (Python) to silicon for such matrices, the generation of 4 arithmetic formats  $\times$  2 accumulator configurations  $\times$  4 PDKs = 32 chips, and their performance and efficiency evaluation, all provided as open source.

### B. Asynchronous and Parallel Compilation

The necessity for rapid generation of specialized circuits primarily arises due to the challenges posed to the established laws of computer science, including Dennard scaling, Moore’s law, and the emergence of issues like power walls and the dark silicon era. An advanced and emerging solution that keeps pace is Open Source EDA, supported by its community [3]. In adopting this approach, we build our own open-source tool, accessible online<sup>1</sup>. Figure 1 illustrates this framework, which

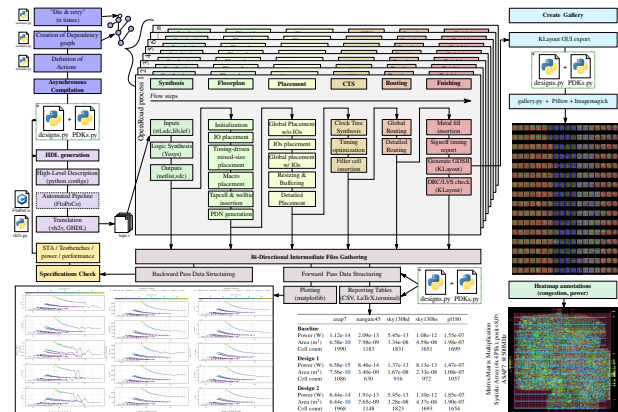


Fig. 1. Schematic Overview SUF: Centralized Management of Asynchronous OpenROAD Forks, Derived from Dependency Task Graphs. This illustration also encapsulates the extended capabilities ranging from Code Generation without manual RTL Writing to Advanced Plotting and Visualization Features.

facilitates the creation of multiple independent design entries, transitioning from high-level Python descriptions to silicon GDS outputs.

### C. Functional and Performance Specifications

We define and assess four computational formats distinguished by their compactness and mathematical attributes (dynamic range and precision). These include Nvidia’s e4m3 and e5m2, and the tapered formats, posit4 (es=0), and posit8(es=2) [4], [5]. Another significant aspect of our work is the proposal of two variations of internal paths for each of these formats. Internally, we execute the dot product as a fused operation (without rounding) in a fixed accumulator with varying boundaries (bit weights for lsb/msb/ovf). These variations, named  $\alpha$  and  $\beta$ , are configured as follows: (ovf = 2, msb = 3, lsb = -2) for an aggregate of 8 bits, and (ovf = 5, msb = 5, lsb = -5) for the 16-bit model. The

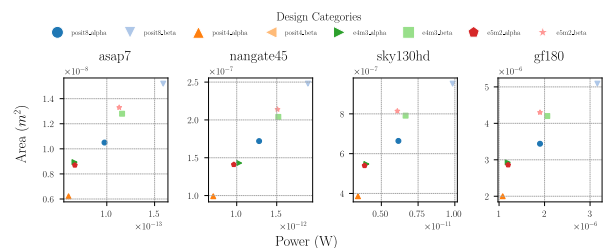


Fig. 2. Area vs. Power for 28 different MMM units combining different computer format, accumulator sizes, and Process Development Kits.

<sup>1</sup><https://github.com/Bynaryman/SUF>

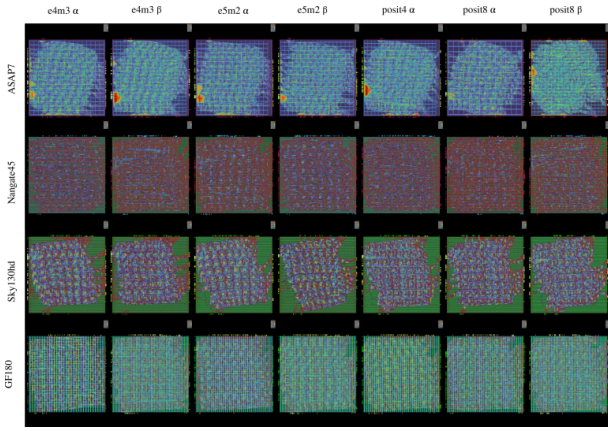


Fig. 3. Overview of the GDS layout of the 28 generated MMM units categorized by Arithmetics vs. PDKs. Each layout has a congestion heatmap which helps in the visualization of Processing Elements.

weights distribution of the embedding layers in the Llama-2-7b model dictates these boundaries. All Systolic Arrays of this work are set to  $8 \cdot 8 = 64$  PEs, which ends the definition of the *Functional* specifications set.

We augment this set with *Performance* specifications across four Process Development Kits (PDKs), specifically GF180, Sky130hd, nangate45, and ASAP7, all of which remain open source. The assessment of multiple PDKs facilitates the validation of design scalability, obviating the need for often imprecise manual scaling techniques.

#### D. Results

All configurations successfully “taped-out” within an hour, with the exception of  $\text{posit}4\beta$ , culminating in a total of 28 produced chips. Figure 3 illustrates the 28 systolic arrays, each a 2D mesh, arranged across arithmetic units versus PDK dimensions. Figure 2 details the performance metrics for the MMM units, indicating that beta costs exceed those of alpha, as expected. Posit arithmetic units incur higher costs relative to their counterparts of equivalent size and accumulation capabilities. Nonetheless, this observation warrants further investigation into accuracy (designated as future work). The configurations e5m2 and e4m3 exhibit minor differences and are positioned in closely situated clusters, reflecting their similar hardware characteristics.

Figure 4 zooms in one of the chip, specifically the e4m3 arith with beta accumulator for the Sky130 nanometer high density PDK. The picture allows to see the PEs thanks to routed congestion heatmap.

#### E. Conclusions

Overall, by the mean of a custom open source framework, we are able to generate MMM units for several arithmetic specifications and technology nodes. We show the performance metrics of 28 distinct chips that have been generated within an hour, which is possible thanks to open source EDA tools.

As a future perspectives, we need to correlate the performance metrics measured with accuracy metrics in order to find the best entry in the vast accuracy/energy efficiency design space exploration. In light of these promising results, we encourage researchers to interact with our tool.

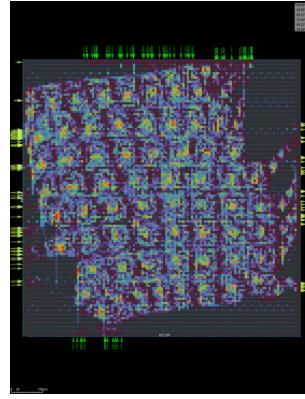


Fig. 4. Zoom-in view of the e4m3 beta chip with fine-tunes congestion heatmap allowing to clearly distinguish the  $8 \times 8 = 64$  PEs.

#### F. Biography



Louis Ledoux, originating from a comprehensive computer science background in Rennes (Bretagne, France), has transitioned towards a hardware focus. His journey began with a Bachelor’s degree, followed by a Master’s in Computer Science, culminating in a one-year internship in 2017, where he explored FPGA virtualization in the cloud. Since 2018, Louis has been engaged in a PhD in computer arithmetic at the Universitat Politècnica de Catalunya and the Barcelona Supercomputing Center, in Barcelona, Spain. His main focus are hardware implementations to address numerical requirements sparsity in HPC workloads.

## II. ACKNOWLEDGMENT

Marc Casas has been partially supported by the Grant RYC-2017-23269 funded by MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future. This research was supported by grant PID2019-107255GB-C21 funded by MCIN/AEI/ 10.13039/501100011033. Els autors agraeixen el suport del Departament de Recerca i Universitats de la Generalitat de Catalunya al Grup de Recerca ”Performance understanding, analysis, and simulation/emulation of novel architectures” (Codi: 2021 SGR 00865).

#### REFERENCES

- [1] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, “Estimating the carbon footprint of bloom, a 176b parameter language model,” 2022.
- [2] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, “The era of 1-bit llms: All large language models are in 1.58 bits,” 2024.
- [3] T. Ajayi, V. A. Chhabria, M. Fogaça, S. Hashemi, A. Hosny, A. B. Kahng, M. Kim, J. Lee, U. Mallappa, M. Neseem *et al.*, “Toward an open-source digital flow: First learnings from the openroad project,” in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–4.
- [4] P. Micikevicius, D. Stolic, N. Burgess, M. Cornea, P. Dubey, R. Grisenthwaite, S. Ha, A. Heinecke, P. Judd, J. Kamalu, N. Mellempudi, S. Oberman, M. Shoeybi, M. Siu, and H. Wu, “FP8 Formats for Deep Learning,” arXiv, Tech. Rep. arXiv:2209.05433, Sep. 2022, arXiv:2209.05433 [cs]. [Online]. Available: <http://arxiv.org/abs/2209.05433>
- [5] J. L. Gustafson and I. T. Yonemoto, “Beating Floating Point at its Own Game: Posit Arithmetic,” *Supercomputing Frontiers and Innovations*, vol. 4, no. 2, pp. 71–86, Apr. 2017, number: 2. [Online]. Available: <https://superfri.org/index.php/superfri/article/view/137>

# Emergence of Energetic Constraints over the Evolution of Protein Families

Marko Ludaic<sup>1</sup>, RG Parra<sup>1</sup>, A. Valencia<sup>1</sup>

<sup>1</sup>Computational Biology Team, Life Sciences Department, Barcelona Supercomputing Center (BSC-CNS)

E-mail: {marko.ludaic, gonzalo.parra, alfonso.valencia}@bsc.es

**Keywords**— protein evolution, energetic frustration, hemoglobin, lactamases

## EXTENDED ABSTRACT

Natural proteins fold by minimizing their internal energetic conflicts at their native states, satisfying the minimum frustration principle. However, 10-15% of the interactions between residues remain in energetic conflict with their local environment and are known as frustrated interactions. If a given energetic feature, energetically minimized or in conflict, is conserved across members of the same protein family it may hint to an evolutionary requirement of the family. In this study, we investigate how local frustration patterns have emerged over evolutionary timescales in two specific protein families, namely the globin and lactamase families.

Local energetic frustration quantifies how effectively a residue-residue interaction's energy is optimized for folding compared to random interactions within the polypeptide chain's non-native conformations. Utilizing the Z-score to assess the native energy against a distribution of decoy energies, interactions can be categorized into three groups - highly frustrated, minimally frustrated and neutral. Three methods of generating decoys yield three distinct frustration indices (FIs): mutational and configurational FIs for pairwise contacts, and single residue frustration index (SRFI). [1]

We have inferred sequences of ancestral proteins in the aforementioned families by performing Ancestral Sequence Reconstruction (ASR) at different timepoints based on the data from the family's phylogenetic tree. We predicted structures of ancestral and extant proteins using Deep learning systems such as AlphaFold2 and ESMFold, and calculated energetic frustration patterns with the Frustratometer tool.

We generated pseudotime dependent energetic profiles for each residue in the family's Multiple Sequence Alignments that describes how energetic constraints changed through evolution, fulfilling functional requirements of the family. We identified particular amino acid positions as important for functional divergence of extant groups of proteins. (Fig. 1)

Fig. 1 Energetic frustration profile through time of the superfamily of globins for K40 residue. K40 residue is found with highly conserved energetic frustration in  $\alpha$  globins and all the main common ancestors, while in  $\beta$  globins it differentiates.

In the lactamase family, we found some of them to be of high importance for the antibiotic resistance mechanism while in the globin family changes obey differential protein-protein interaction needs. We were able to trace such amino acid substitutions back in time to create a time dependent profile of events that led to the extant state. Additionally, we found that single-residue energetic frustration does not necessarily get maximized to advance the functionality trait associated with particular residue, but instead shifts to the neutral zone, which has been described as a relaxed state, allowing for more functional opportunity and flexibility for the protein. (Fig. 2)

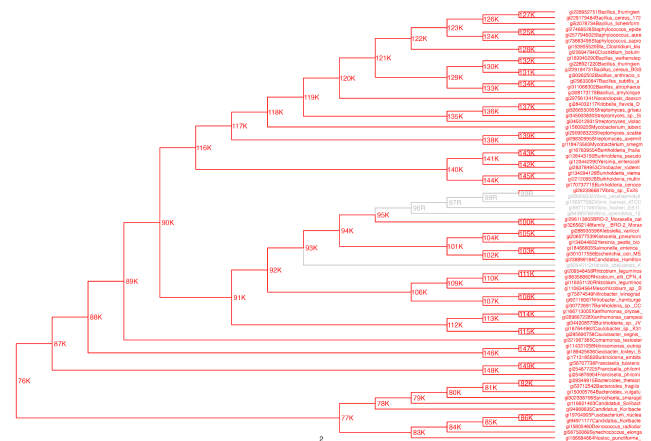
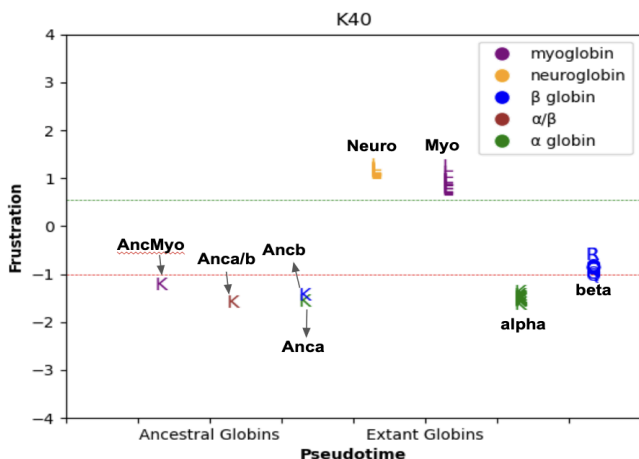


Fig. 2 Phylogenetic tree of lactamase family showing change of energetic frustration through evolution for the residue K234. In the clade of *Vibrio* sp. there was a change from highly frustrated to neutral.



Based on the previous studies in the family of lactamases, the catalytic residue K234 was found to play an important role in avibactam-mediated inhibition of the enzyme when mutated to arginine (K234R). This particular mutation was noted to shift energetic frustration to neutral level and was recorded in *Vibrio* clade. [2] In addition to K234R, another residue R220M has been associated with the same behavior. Surprisingly, both of these amino acid substitutions correspond to the *Vibrio* clade in the phylogenetic tree. Substitution R220M was found in *Vibrio splendidus* particularly. [3] These results, along with the previous experimental findings, suggest that *Vibrio* bacteria is more resistant to ampicillin-avibactam inhibition compared to other bacteria due to having these particular substitutions along with a change of energetic frustration at the same sites.

To represent mutational frustration residue-residue contacts of particular globin groups ( $\alpha$ ,  $\beta$ , myoglobin and neuroglobin) we made networks of residues that were in contact using mutational frustration. Mutational frustration indicates how the frustration changes as a function of the amino acid identities for a given pair of positions. Contact residue networks of selected  $\alpha$  and  $\beta$  globin residues suggest that  $\alpha$  globins are the group which frustration is the highest as they have many highly frustrated contacts and the overall frustration profile differs from the one found in  $\beta$ , and especially myoglobin and neuroglobin groups. (Fig. 3)

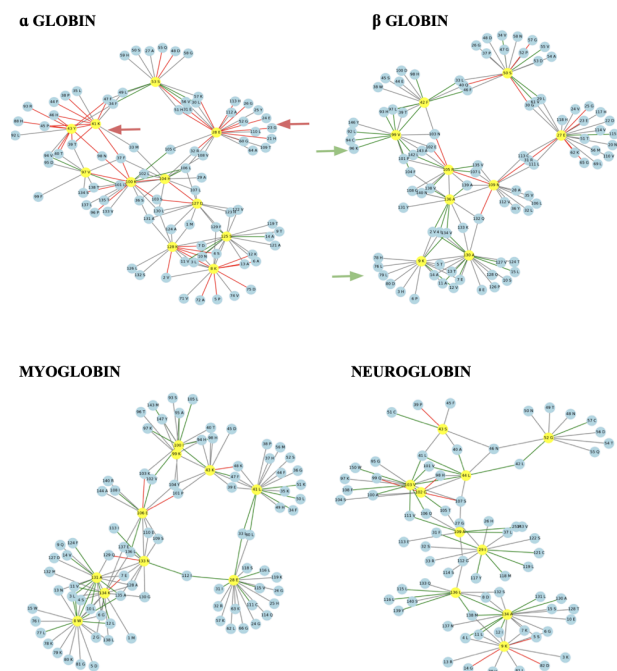


Fig. 3 Residue-residue contact networks for each group in the superfamily of globins.

To summarize the information from the contact networks we conducted a statistical analysis. We concluded that out of the four extant groups of globins,  $\alpha$  globins have the highest frequency of maximally frustrated contacts. The same pattern was observed for the common ancestors which suggests that  $\alpha$  globins are closely related to globin ancestors, and even more closely related to the ancestors than other extant groups of globin. This furthermore suggests that the functions of  $\alpha$  globin alone are more similar to those of the ancestors, identifying  $\alpha$  globins as evolutionary oldest members of the superfamily.

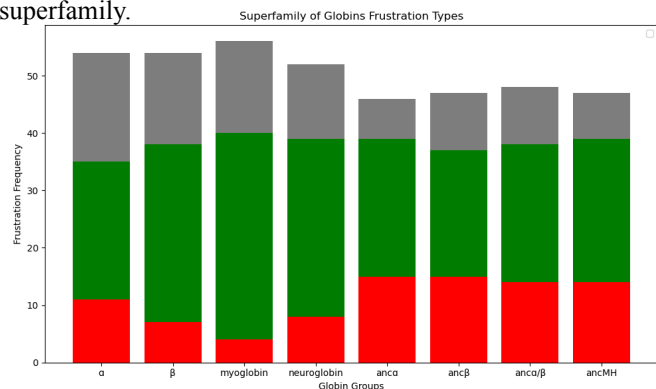


Fig. 4 Frequency histogram of energetic frustration in each extant globin

group and main common ancestors. Different colors represent different frustration states (red - highly frustrated, green - minimally frustrated, gray - neutral).

While exact mechanisms of how protein families evolved through evolution and developed their functional characteristics remain unexplored, our approach aims to bring new insights and suggests how energetic constraints have shaped evolution. We believe that this methodology can be generally applied to any other type of protein, globular, non globular or even disordered ones.

#### ACKNOWLEDGEMENTS

I wish to acknowledge the Life Sciences Department at the Barcelona Supercomputing Center for their assistance in the production of this project, especially the supervisors Dr. R Gonzalo Parra and Dr. Alfonso Valencia. Furthermore, I would like to thank collaborator Valeria Risso from the University of Granada (Departamento de Química Física, Universidad de Granada) for providing this research with the ancestral lactamase dataset.

#### References

- [1] Freiburger, M. I., Ruiz-Serra, V., Pontes, C., Romero-Durana, M., Galaz-Davison, P., Ramírez-Sarmiento, C. A., Schuster, C. D., Marti, M. A., Wolynes, P. G., Ferreira, D. U., Parra, R. G., & Valencia, A. (2023). Local energetic frustration conservation in protein families and superfamilies. *Nature communications*, 14(1), 8379.
- [2] Mangat, C. S., Vadlamani, G., Holicek, V., Chu, M., Larmour, V. L. C., Vocadlo, D. J., Mulvey, M. R., & Mark, B. L. (2019). Molecular Basis for the Potent Inhibition of the Emerging Carbapenemase VCC-1 by Avibactam. *Antimicrobial agents and chemotherapy*, 63(4), e02112-18.
- [3] Papp-Wallace, K. M., Winkler, M. L., Taracila, M. A., & Bonomo, R. A. (2015). Variants of  $\beta$ -lactamase KPC-2 that are resistant to inhibition by avibactam. *Antimicrobial agents and chemotherapy*, 59(7), 3710–3717.

#### Author biography



**Marko Ludaic** was born in Novi Knezevac, Serbia. In 2022 he graduated from the University of Belgrade, Faculty of Biology in the field of molecular biology and physiology. Afterwards, he commenced his master degree studies in Bioinformatics for health sciences at

Universitat Pompeu Fabra (UPF) in Barcelona, Spain and joined a Computational Biology team, Life science department at Barcelona Supercomputing center where he is currently developing his master thesis in the field of structural bioinformatics and molecular evolution.

# Designing a new generation of industrial proteases

Miguel Luengo<sup>#1</sup>, Martin Floor<sup>#2</sup>, Victor Guallar<sup>#3</sup>

<sup>#</sup>Barcelona SuperComputing Center, Life Sciences Department, Plaça d'Eusebi Güell, 1-3, 08034 Barcelona, Spain

<sup>1</sup>[miguel.luengo@bsc.es](mailto:miguel.luengo@bsc.es) <sup>2</sup>[martin.floor@bsc.es](mailto:martin.floor@bsc.es) <sup>3</sup>[victor.guallar@bsc.es](mailto:victor.guallar@bsc.es)

**Keywords**— Enzyme Engineering, Proteases, Detergents, Molecular Modelling, Structural Bioinformatics

## A. Introduction

Enzymes, especially proteases, play a critical role in numerous industrial processes, including detergents, cosmetics, and food. Proteases account for 60% of the enzyme market [1] and are a significant ingredient in detergent formulations, representing approximately 30% of worldwide enzyme sales [2].

However, the high cost of production remains a major obstacle to their industrial use as an alternative to chemical formulations. To address this challenge, it is crucial to improve enzyme activity and promiscuity. Increasing enzyme activity makes the reaction more efficient, reducing the amount of enzyme required while enhancing promiscuity allowing a single enzyme variant to eliminate a wider range of organic stains.

To enhance an enzyme's activity and promiscuity, a promising strategy is to create secondary active sites with substrate specificity that complements the primary site [3]. However, industrial subtilisin detergent proteases have limited cavities, making it challenging to add new catalytic pockets. To address this issue, we introduce a novel enzyme design approach that searches the protein surface for locations to insert extra catalytic triads while simultaneously enhancing substrate binding.

## B. Methodology

First, a Monte Carlo simulation is run to find alternative binding sites with our in-house PELE Monte Carlo algorithm. PELE evaluates the binding energy of the ligand around the surface of the protein while exploring the conformation of the protein-ligand complex. For each newfound site, an exhaustive search is performed to insert a new catalytic triad. This is done by first selecting all possible residues within a distance of the ligand that could be mutated to serine, the residue that performs the nucleophilic attack during catalysis. For each serine, we then find all neighboring residues that could be mutated to histidine, and the same is done to find possible aspartic or glutamic acid residues to obtain the rest of the necessary amino acids to form a catalytic triad. All the possible combinations are then evaluated in the design phase.

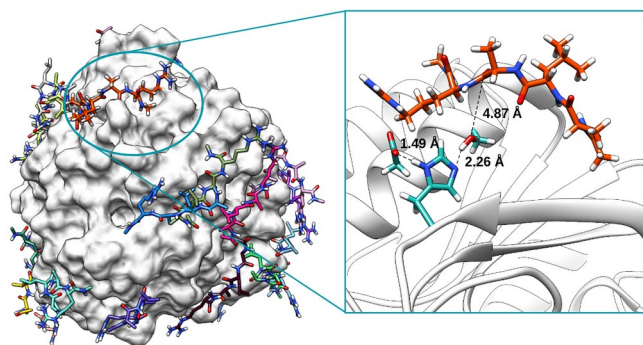


Fig. 1 Representation of the surface exploration of additional binding sites and example of putative inserted catalytic triad.

The design protocol utilizes a Monte Carlo search to perform conformational sampling of the protein-ligand complex while proposing mutations that stabilize the configuration. Only residues defined inside a sphere within the closest contacts of the inserted triad are mutated, while a second sphere designates residues to repack. To maintain triad stability while minimizing energy, a distance constraint is applied during the process which is lifted after each step to record unconstrained energy changes. These mutations result in a binding complementary surface at the protein crevices, allowing for catalytically competent ligand binding.

## C. Results

The protocol was applied to several industrial proteases used in the detergent industry, using peptides based on casein as the ligand. To ensure peptide binding stability and proper triad formation, the resulting enzymes were validated using both molecular dynamics and Monte Carlo PELE simulations.

The validation results demonstrate that the protocol successfully achieves its initial goal of designing additional stable triads that provide an effective binding surface to the ligand. Some of the enzymes designed showed low binding energy conformations based on the PELE simulations and maintained catalytic distances between the triad and the ligand throughout the molecular dynamics simulations.

The best designs in terms of ligand binding and triad preorganization were evaluated experimentally. Initial protease activity assays show an increase in specific activity of up to 5 fold compared to the native enzyme. Additionally, some designs also presented higher values of expression which may be attributed to the additional mutations stabilizing the overall protein.

## D. Conclusions and Further Work

Despite the initial challenge of working with enzymes with few alternative pockets, such as subtilisin proteases, the new protocol successfully obtained designs with high binding energies and preformed catalytic triads in computational simulations. Furthermore, initial experimental activity assays show promising results with designs that show improvement in both activity and expression. Additional experimental research is required to better characterize the newly designed active sites as well as evaluate their capacity in a detergent industrial context.

Another possibility that arises given the positive results is combining the designs of the active sites which demonstrated more potential into a single enzyme with more than two active sites possibly increasing the activity even further.

## References

- [1] Li, Qing, Li Yi, Peter Marek, and Brent L. Iverson. 2013. "Commercial Proteases: Present and Future." *FEBS Letters* 587(8): 1155–63.
- [2] Anissa Haddar; Rym Agrebi; Ali Bougatef; Noomen Hmidet; Alya Sellami-Kamoun; Moncef Nasri. (2009). "Two detergent stable alkaline serine-proteases from



Bacillus mojavensis A21: Purification, characterization and potential application as a laundry detergent additive.” , 100(13), 3366–3373.

- [3] Alonso, Sandra, Gerard Santiago, Isabel Cea-Rama, Laura Fernandez-Lopez, Cristina Coscolín, Jan Modregger, Anna K. Ressmann, et al. 2019. “Genetically Engineered Proteins with Two Active Sites for Enhanced Biocatalysis and Synergistic Chemo- and Biocatalysis.” Nature Catalysis, under Third Revision.

## ***Author biography***



**Miguel Luengo** received his B.E. degree in Biosystems Engineering from the Polytechnic University of Barcelona (UPC) in 2019. In 2021 he obtained an MSc. in Bioinformatics from the University Pompeu Fabra (UPF).

In 2022 he started working as a research engineer at the Electronic and Atomic Protein Modelling Group (EAPM) at the Barcelona Supercomputing Center (BSC) in the NextProt national project. Since October 2023, he has been a Ph.D. student in biotechnology in the field of enzyme engineering under the supervision of Martin Floor and Victor Guallar.

# EQUILI module in ALYA: a free-boundary Grad-Shafranov equation solver using CutFEM

Pau Manyer Fuertes<sup>\*†</sup>, Alejandro Soba Pascual<sup>\*†</sup>, Daniel Gallart<sup>\*</sup>

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {pau.manyer, alejandro.soba, daniel.gallart}@bsc.es

**Keywords**—*Grad-Shafranov, Free-boundary, CutFEM, ALYA.*

## I. EXTENDED ABSTRACT

Aiming at developing new numerical tools focused on plasma magnetic confinement in nuclear reactor's cores, a project has been initiated by the Fusion group at the Barcelona Supercomputing Center (BSC). We present the first steps towards the implementation and development of a new module EQUILI within the framework of ALYA [1], a coupled multi-physics simulation code using high performance computing techniques and specially designed to run on supercomputers such as Marenostrum hosted at BSC. The new module EQUILI will solve the Grad-Shafranov [2] equation using Finite Element Method [3] with embedded geometry, in this case implementing more specifically a CutFEM [4] algorithm.

### A. Grad-Shafranov equation and problem layout

The Grad-Shafranov equation describes the equilibrium for a two dimensional plasma in ideal magnetohydrodynamics (MHD) by means of the poloidal magnetic flux  $\psi$ . The force balance is achieved by driving a current inside the plasma that produces a Lorentz force to counter the plasma pressure gradient. Deriving from the MHD equilibrium equations and assuming toroidal axisymmetry, for instance the plasma contained in a Tokamak device, the Grad-Shafranov equation [2] can be written in cylindrical coordinates as

$$\Delta^* \psi \equiv \frac{\partial^2 \psi}{\partial R^2} - \frac{1}{R} \frac{\partial \psi}{\partial R} + \frac{\partial^2 \psi}{\partial Z^2} = \mu_0 R J_\phi(R, \psi) \quad (1)$$

where  $\Delta^*$  is the toroidal elliptic operator and  $J_\phi$  represents the plasma toroidal current. In addition to  $J_\phi$ , electrical currents are also carried in toroidal and poloidal magnetic field coils outside the plasma chamber to produce the confining magnetic field, as shown in figure 1. The red rectangle corresponds to the computational domain, in which an approximate solution [5] for the  $\psi$  field is represented. The plasma magnetic confinement coils (PF1, PF2...) and solenoids (CS1U, CS1L...) are placed outside the computational domain considering the configuration built in ITER, at locations  $(R_i^c, Z_i^c)$  [6] with current intensity  $I_i$  [7],  $i = 1, 2, \dots, n_c$  with  $n_c$  the total number of external confining magnets.

In figure 1, we observe the presence of different elements characterising the plasma domain  $\mathcal{P}(\psi)$ : the magnetic axis, which corresponds to a local extremum of  $\psi$ , and two saddle points. The *active* saddle point fixes the lowest point of the plasma domain's separatrix. The separatrix corresponds to the last closed magnetic flux surface and will be considered as the plasma domain boundary  $\partial\mathcal{P}$ . That is, outside the separatrix

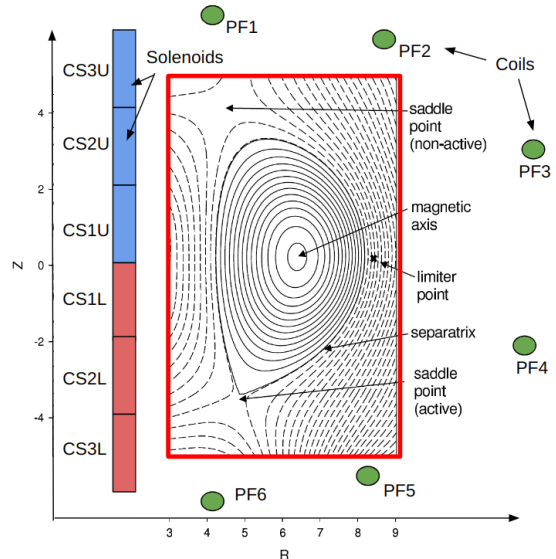


Fig. 1. Axisymmetrical plasma equilibrium problem layout according to ITER configuration (position of coils, solenoids and expected distribution of poloidal magnetic flux  $\psi$  isosurfaces). [5]

there is no plasma, and therefore we have  $J_\phi = 0$ . The Grad-Shafranov problem can hence be written as follows:

$$\Delta^* \psi = \begin{cases} \mu_0 R J_\phi(R, \psi) & (R, Z) \in \mathcal{P}(\psi) \\ \mu_0 R I_i & (R, Z) = (R_i^c, Z_i^c) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

### B. Free-boundary problem numerical treatment

While the coil current can be individually adjusted to accommodate a variety of plasma pressure and current profiles in terms of plasma positioning and shaping, the current carried by the plasma  $J_\phi$  depends directly on the shape of  $\mathcal{P}(\psi)$ , which at the same time is affected by  $J_\phi$ . Due to this coupling, the problem needs to be solved using an iterative method where the plasma shape  $\mathcal{P}(\psi)$  is not fixed and evolves towards the equilibrium configuration. This corresponds in fact to a free-boundary problem. Nonetheless, the  $\psi$  values on the computational domain's boundary (red rectangle),  $\psi_B$ , need to be fixed in order to solve the Grad-Shafranov equation in the domain. Such values are computed using the elliptic operator  $\Delta^*$ 's corresponding Green's function. The general solution strategy for solving the free-boundary problem involves thus an iterative approach based on a double loop structure: in the external loop, the method looks for the convergence on the boundary values  $\psi_B$ , computed as mentioned earlier; in the internal loop, the algorithm solves the Grad-Shafranov problem

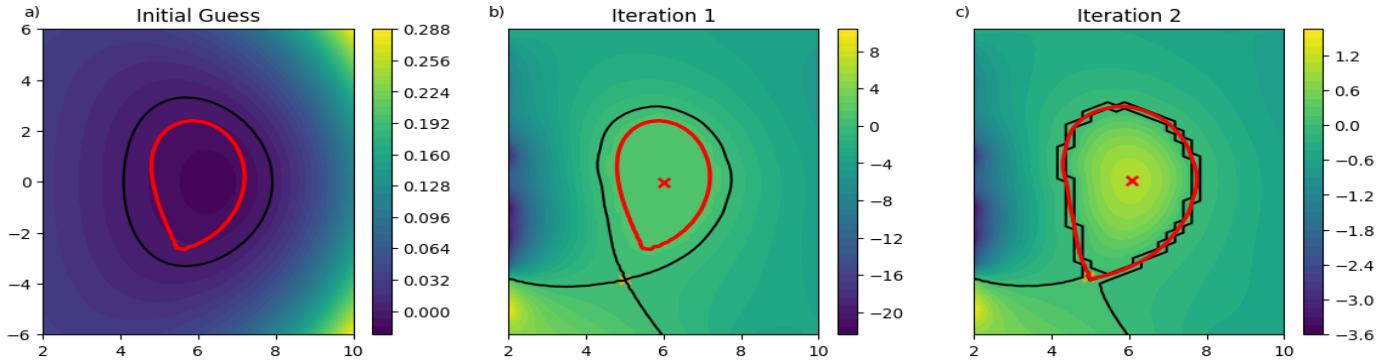


Fig. 2. Normalised  $\bar{\psi}$  obtained for the **a)** initial guess, **b)** first and **c)** second iterations. The plasma boundary  $\partial\mathcal{P}(\bar{\psi})$  used from the previous iteration is drawn with a red line, while the  $\bar{\psi}$  0-level contour has been represented using a black line. Moreover, local extremum and saddle point have been indicated using a red and orange cross respectively.

using boundary conditions  $\psi_B$  obtained during the external iteration. The fact that the plasma domain  $\mathcal{P}(\psi)$  will evolve towards the equilibrium solution calls for the implementation of a Finite Elements Method which must be able to easily track such changes in the geometry of  $\mathcal{P}(\psi)$ . That is why we have selected CutFEM [4], as boundaries are defined using level-set functions. There is no need for re-meshing or moving mesh nodes, and here lies precisely the strength in using geometry embed methods.

A usual practice consists in normalising variable  $\psi$  [5] respect to its value at the magnetic axis,  $\psi_0$ , and at the *active* saddle point,  $\psi_X$ , such that

$$\bar{\psi} = \frac{\psi - \psi_X}{\psi_0 - \psi_X} \quad (3)$$

This way, the plasma boundary  $\partial\mathcal{P}$  can be tracked by the  $\bar{\psi}$  0-level contour. Also, the source term  $J_\phi$  is typically provided as a function of  $\bar{\psi}$ .  $J_\phi$  is also normalised so that the total current circulating through the plasma domain surface remains constant [5].

Due to the easier implementation, a first stand-alone prototype for EQUILI has been build using Python, also in order to check the performance of CutFEM when solving a free-boundary problem.

### C. Numerical results

Figure 2 presents the results obtained using a first version of EQUILI developed in Python. The different subplots show the resulting  $\bar{\psi}$  obtained from solving the CutFEM system. Departing from the initial guess  $\psi^{(0)}$  [6] and an initial plasma region  $\mathcal{P}(\psi^{(0)})$  [8], we observe substantial changes in the first iteration (figure 2 **b**): first, boundary conditions  $\psi_B$  are applied, driving the emergence of both local extremum and saddle point in the  $\psi$  field; then, after normalisation we see how the 0-level contour for  $\bar{\psi}$  defines a bigger domain than the original plasma shape  $\mathcal{P}(\psi)$  (red contour), in which we assumed  $J_\phi \neq 0$  for the calculations. Hence, we can see how  $\mathcal{P}(\psi)$  is changing in order to reach the equilibrium configuration. Now, the computed  $\bar{\psi}$  0-level contour is taken as the new plasma shape  $\mathcal{P}(\bar{\psi})$  for the next iteration. For figure 2**c**) we observe in this case that  $\mathcal{P}(\bar{\psi})$  doesn't change, as the concentric  $\bar{\psi}$  isosurfaces pattern appears inside it. The stepped black contours around  $\partial\mathcal{P}(\bar{\psi})$  are due to the low mesh resolution employed for this simulation.

### D. Conclusion

In conclusion, both the double loop structure and CutFEM implementations seem to perform correctly when solving the

free-boundary Grad-Shafranov equation. Further steps will consist, on one hand in considering the Tokamak first wall geometry so that the boundary conditions  $\psi_B$  are calculated and prescribed on that interface; on the other hand, the translation of the algorithm into ALYA as a new independent module EQUILI.

## II. ACKNOWLEDGMENT

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

## REFERENCES

- [1] M. Vázquez *et al.*, “Alya: Multiphysics engineering simulation toward exascale,” *Journal of Computational Science*, 2016.
- [2] D. A. Daniel *et al.*, “Derivation and Applications of Grad-Shafranov Equation in Magnetohydrodynamics,” *Journal of Research in Applied Mathematics*, 2021.
- [3] O. Zienkiewicz *et al.*, *The Finite Element Method: its Basis and Fundamentals*. Butterworth-Heinemann, 2005.
- [4] E. Burman *et al.*, “CutFEM: Discretizing geometry and partial differential equations,” *International Journal for Numerical Methods in Engineering*, 2014.
- [5] S. Jardin, *Computational Methods in Plasma Physics*. Chapman Hall/CRC, 2010.
- [6] X.-Z. T. Shuang Liu, Qi Tang, “A parallel cut-cell algorithm for the free-boundary Grad-Shafranov problem,” 2021.
- [7] A. P. Pietro Testoni, “Introduction and validation of the APEC code for the simulation of plasma equilibrium and evolution in presence of 3D passive structures,” 2021.
- [8] C. Barril *et al.*, “MHD Equilibria of Tokamak Plasmas,” *Mathematics in Industry Reports (MIIR)*, 2021.



**Pau Manyer Fuertes** received his BSc degree in Physics Engineering from Universitat Politècnica de Catalunya, Spain in 2022. The same year, he began coursing a MSc degree in Numerical Methods in Engineering from UPC. At the same time, he was enrolled at the Fusion Research Group from the CASE department at BSC.

# Reconstructing prokaryotic metabolic contributions to the Last Eukaryotic Common Ancestor

Saioa Manzano-Morales<sup>\*†</sup>, Moisès Bernabeu<sup>\*†</sup>, Marina Marcet-Houben<sup>\*</sup>, Toni Gabaldón<sup>\*†‡§</sup>

<sup>\*</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain

<sup>†</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

<sup>‡</sup>Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

<sup>§</sup>Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona, Spain

E-mail: saioa.manzano@bsc.es, toni.gabaldon@bsc.es

**Keywords**—comparative genomics, eukaryogenesis, metabolic reconstruction, LECA.

## I. EXTENDED ABSTRACT

Eukaryotes differ from prokaryotes in deep and fundamental ways, a divide so deep it has been termed the major evolutionary transition since the origin of life itself[1]. The Last Eukaryotic Common Ancestor (LECA) stems from the endosymbiosis of an alphaproteobacteria-related bacterium[2] into an Asgard-related[3] archaeal host. However, the nature and complexity of such host remain hotly under debate, as do the relative timing and origin of the rest of hallmark eukaryotic features.

The main objective of this work is to understand the functions of the gene families brought into LECA in the major waves of gene acquisition and how they integrate into the metabolism of the organism, as well as the part they played in the complexification of cells in the path to LECA.

### A. Introduction

The origin of eukaryotes has long been a huge enigma for evolutionary biology. There is a sharp divide in the organizational complexity of the cells between eukaryotes and prokaryotes [4]. In fact, eukaryotes possess a cellular and regulative complexity unprecedented in prokaryotes: a nucleus which allows uncoupling of transcription and translation, a sophisticated endomembrane system, a complex, dynamic cytoskeleton and a unique sexual cycle[5], not too different from many extant unicellular eukaryotes, and which allowed evolution into multicellularity several times independently[6].

Many of such complex features can be traced back to the Last Eukaryotic Common Ancestor, an organism estimated to have lived in the Proterozoic (ca. 1.9–1.6 Ga)[7]. However, the path from a prokaryotic-like organism (the First Eukaryotic Common Ancestor, or FECA) to LECA (a process termed eukaryogenesis) (Figure 1) involved, at minimum, the endosymbiosis of an Alphaproteobacteria-like bacterium (which would later become the mitochondria) [2] and an Asgard host[3]. However, this scenario leaves many of LECA's features unaccounted for [8]. Studies on Asgard archaea can help elucidate which features were already present in the Asgard-like FECA [9], yet there exists a bridge between FECA and LECA that cannot be bridged with this alone.

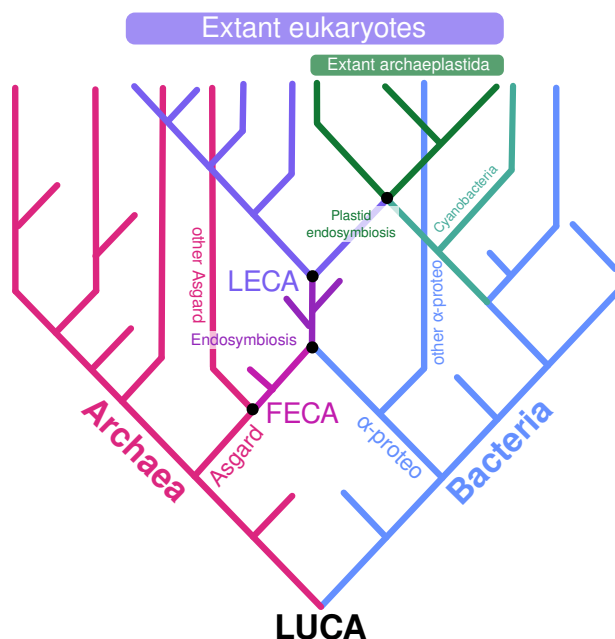


Fig. 1. Phylogenetic view of eukaryogenesis. Adapted from [10]

There is increasing evidence that eukaryogenesis likely involved events of gene acquisition other than the alphaproteobacterial endosymbiosis[11], which could even precede it in time[12].

The burst of available genomes and transcriptomes from diverse microbial eukaryotes, including newly-discovered supergroups such as *Provora* [13] and *Hemimastigophora* [14], along with an improvement of phylogenetic reconstruction pipelines and computational resources makes this a prime time for an assessment of the number and nature of the waves of gene acquisition on the path to LECA, their putative donors, and their integration into the metabolism of the proto-eukaryote.

### B. Inference of LECA Gene Families

We queried available genomes from diverse eukaryotes to form three alternative, taxonomically balanced proteome databases (termed TOLDB-A, TOLDB-B and TOLDB-

C). We then clustered the proteins into orthogroups with OrthoFinder[15], using BLAST as an aligner, and selected the orthogroups that could be assumed to have been present in LECA by taxonomic criteria, under different levels of stringency. We then searched for putative prokaryotic homologs of these gene families with a Hidden Markov Model (HMM) search against an in-house database of proteins comprising all realms (BROAD-DB). In subsequent steps, we inferred phylogenetic trees and pruned species, until left with a collection of Maximum Likelihood trees from which the prokaryotic sister to the LECA gene families could be assessed. The prokaryotic sister was then subsequently identified, and gene families were functionally annotated via KEGG[16] Orthology (KO).

### C. Metabolic reconstruction

We assessed the pathway distribution of KO objects by parsing the KGML files of the different KEGG pathways via an in-house script based on the packages KEGGREST and Pathview. In brief, per each metabolic pathway each KO was linked to their corresponding KEGG REACTION, and the distribution heatmaps were assessed as the percentage of reactions by the given module out of the total number of reactions in the pathway and module, respectively. For non-metabolic pathways, summarizing at the reaction level was not possible, and therefore the distribution was assessed by assigning the KOs to the pathways directly. We then visualized the results with the R package pheatmap.

### D. Results

Results show the implication of several prokaryotic lineages in the main waves of gene acquisition towards LECA, including known sources such as Alphaproteobacteria and Heimdallarchaeia, coupled with several others whose importance in eukaryogenesis has been underexplored. The gene families supplied by these donors belong to different KEGG Orthologs, with minimal overlap, despite being widespread across pathways. We also see enrichment of KEGG pathways both in the donor contribution to the LECA proteome and the eukaryotic innovations.

### E. Conclusion

In this study, we assess the KEGG pathway composition of the main modules of gene acquisition in LECA, observing clear differentiation of the gene donors in the function of the genes they contributed.

## II. ACKNOWLEDGMENT

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

## REFERENCES

- [1] R. Y. Stanier, *The Microbial World*. New Jersey, U.S.: Prentice-Hall, 1986.
- [2] A. J. Roger *et al.*, "The origin and diversification of mitochondria," *Current Biology*, vol. 27, no. 21, p. R1177–R1192, Nov. 2017. [Online]. Available: <http://dx.doi.org/10.1016/j.cub.2017.09.015>
- [3] K. Zaremba-Niedzwiedzka *et al.*, "Asgard archaea illuminate the origin of eukaryotic cellular complexity," *Nature*, vol. 541, no. 7637, p. 353–358, Jan. 2017. [Online]. Available: <http://dx.doi.org/10.1038/nature21031>
- [4] E. V. Koonin, "The origin and early evolution of eukaryotes in the light of phylogenomics," *Genome Biology*, vol. 11, no. 5, p. 209, 2010. [Online]. Available: <http://dx.doi.org/10.1186/gb-2010-11-5-209>
- [5] P. J. L. Bell, "Eukaryogenesis: The rise of an emergent superorganism," *Frontiers in Microbiology*, vol. 13, May 2022. [Online]. Available: <http://dx.doi.org/10.3389/fmicb.2022.858064>
- [6] S. M. Adl *et al.*, "The revised classification of eukaryotes," *Journal of Eukaryotic Microbiology*, vol. 59, no. 5, p. 429–514, Sep. 2012. [Online]. Available: <http://dx.doi.org/10.1111/j.1550-7408.2012.00644.x>
- [7] T. A. Mahendrarajah *et al.*, "Atp synthase evolution on a cross-braced dated tree of life," Apr. 2023. [Online]. Available: <http://dx.doi.org/10.1101/2023.04.11.536006>
- [8] J. B. Dacks *et al.*, "The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together," *Journal of Cell Science*, Jan. 2016. [Online]. Available: <http://dx.doi.org/10.1242/jcs.178566>
- [9] T. Rodrigues-Oliveira *et al.*, "Actin cytoskeleton and complex cell architecture in an asgard archaeon," *Nature*, vol. 613, no. 7943, p. 332–339, Dec. 2022. [Online]. Available: <http://dx.doi.org/10.1038/s41586-022-05550-y>
- [10] P. C. Donoghue *et al.*, "Defining eukaryotes to dissect eukaryogenesis," *Current Biology*, vol. 33, no. 17, p. R919–R929, Sep. 2023. [Online]. Available: <http://dx.doi.org/10.1016/j.cub.2023.07.048>
- [11] T. Gabaldón, "Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis," *IUBMB Life*, vol. 70, no. 12, p. 1188–1196, Oct. 2018. [Online]. Available: <http://dx.doi.org/10.1002/iub.1950>
- [12] A. A. Pittis and T. Gabaldón, "Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry," *Nature*, vol. 531, no. 7592, p. 101–104, Feb. 2016. [Online]. Available: <http://dx.doi.org/10.1038/nature16941>
- [13] D. V. Tikhonenkov *et al.*, "Microbial predators form a new supergroup of eukaryotes," *Nature*, vol. 612, no. 7941, p. 714–719, Dec. 2022. [Online]. Available: <http://dx.doi.org/10.1038/s41586-022-05511-5>
- [14] G. Lax *et al.*, "Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes," *Nature*, vol. 564, no. 7736, p. 410–414, Nov. 2018. [Online]. Available: <http://dx.doi.org/10.1038/s41586-018-0708-8>
- [15] D. M. Emms and S. Kelly, "Orthofinder: phylogenetic orthology inference for comparative genomics," *Genome Biology*, vol. 20, no. 1, Nov. 2019. [Online]. Available: <http://dx.doi.org/10.1186/s13059-019-1832-y>
- [16] M. Kanehisa, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, p. 27–30, Jan. 2000. [Online]. Available: <http://dx.doi.org/10.1093/nar/28.1.27>



**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 Polytechnical 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.

# Detecting non-vertical inheritance across eukaryotes

Giacomo Mutti<sup>\*†</sup>, Toni Gabaldón<sup>\*†‡§</sup>

<sup>\*</sup>Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain

<sup>†</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>‡</sup>Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

<sup>§</sup>Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Barcelona, Spain

E-mail: giacomo.mutti@irbbarcelona.org, toni.gabaldon@bsc.es

**Keywords**—Reticulation, Phylogenomics, Evolution, HGT.

## I. EXTENDED ABSTRACT

Recently, an increasing number of studies have revealed a high degree of discordance between genes phylogenies and the corresponding species histories. This phenomenon can arise from three main factors: i) analytical errors, ii) differential duplications and losses and, finally, iii) non-vertical evolution processes such as hybridization or horizontal gene transfer (HGT). Hybridization is generally an important factor on relatively short evolutionary timescales, when two different lineages can still produce viable offsprings. HGT, defined as the non-sexual movement of genetic information between genomes [1], can act on much larger evolutionary timescales instead [2].

Even though it is complicated to separate analytically these three factors, we are starting to appreciate how tangled evolution can also be in Eukaryotes. Especially considering that the absolute majority of eukaryotic organisms are unicellular (Fig. 1). The degree to which non-vertical processes affect eukaryotic evolution however, is still debated [3]. Thanks to the availability of new genomes and methodologies, we can start trying to systematically evaluate it. We will use two computational approaches to do this. First, reconciliation, which can quantify horizontal evolution within a clade by modelling evolutionary events to explain the differences between genes and species trees. Secondly, by scanning whole eukaryotic proteomes in a Tree of Life (ToL) scale database to detect patchily distributed genes, which is a distinctive feature of HGT genes. We intend to apply this pipeline across all eukaryotic groups with sufficient genomic representation.

Ultimately, given the taxonomic incompleteness, pervasive contamination and different analytical errors, our project endeavors to establish an updatable and reproducible framework that will serve as a foundation for future research efforts, facilitating the integration of new, or less contaminated, genomic resources and the refinement of analytical techniques as they become available.

### A. Gene trees-Species tree reconciliation

In order to detect intra-clade horizontal evolution we need to model the evolutionary events (*duplications*, *transfers* and *losses*) observed in the gene trees given the species history. This process is called reconciliation (see Fig. 2A). To obtain the gene trees, we first cluster proteins in gene families using OrthoFinder [5]. Each gene family phylogeny is then computed with IQ-Tree [6]. The species tree is inferred both with a concatenation method, using the markers from PhyloFisher [7],

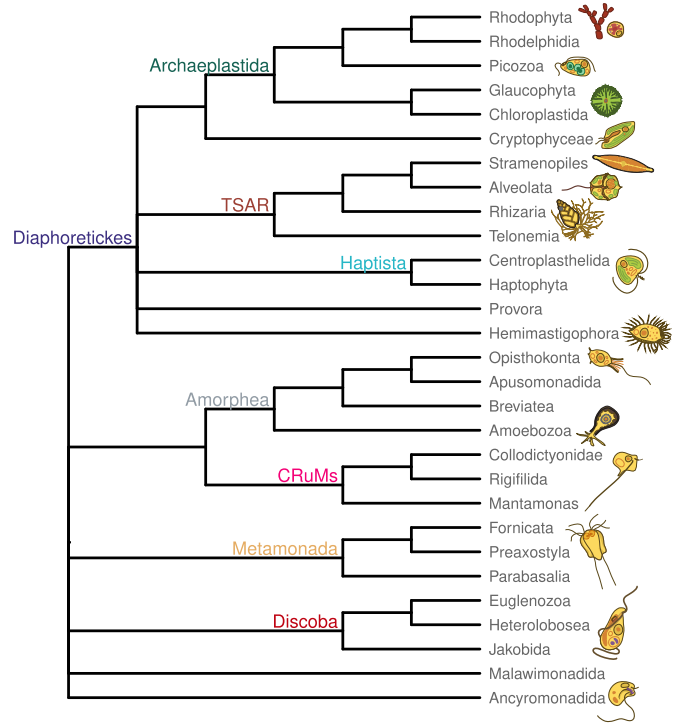


Fig. 1. Consensus of the eukaryotic tree of life. Illustrations from [4]

and with different summary methods: where the information coming from the gene trees topologies rather than from the actual sequences is used. The final species tree is the consensus of all the methods. The reconciliation is computed with AleRax [8], a recent software that can infer *duplications*, *transfers* and *losses* events with a fully probabilistic model that can also account uncertainties in the gene tree estimation.

### B. Interdomain HGTs

We get one representative sequence per orthogroup and all the orphan genes (i.e. genes without clear homologs within each dataset) and we look for homologs in a ToL scale database. If the hits to the query gene show a patchy taxonomic distribution, (for example, if most homologs are from bacteria) we compute the gene phylogeny to better understand its history and, if possible, identify the donors and acceptor involved in the putative HGT event (see Fig. 2B).

### C. Results

Currently, the computational pipeline is in active development. We created a reproducible and customisable workflow to build a representative ToL database that maximises the eukaryotic taxonomic coverage with proteomes from UniProt

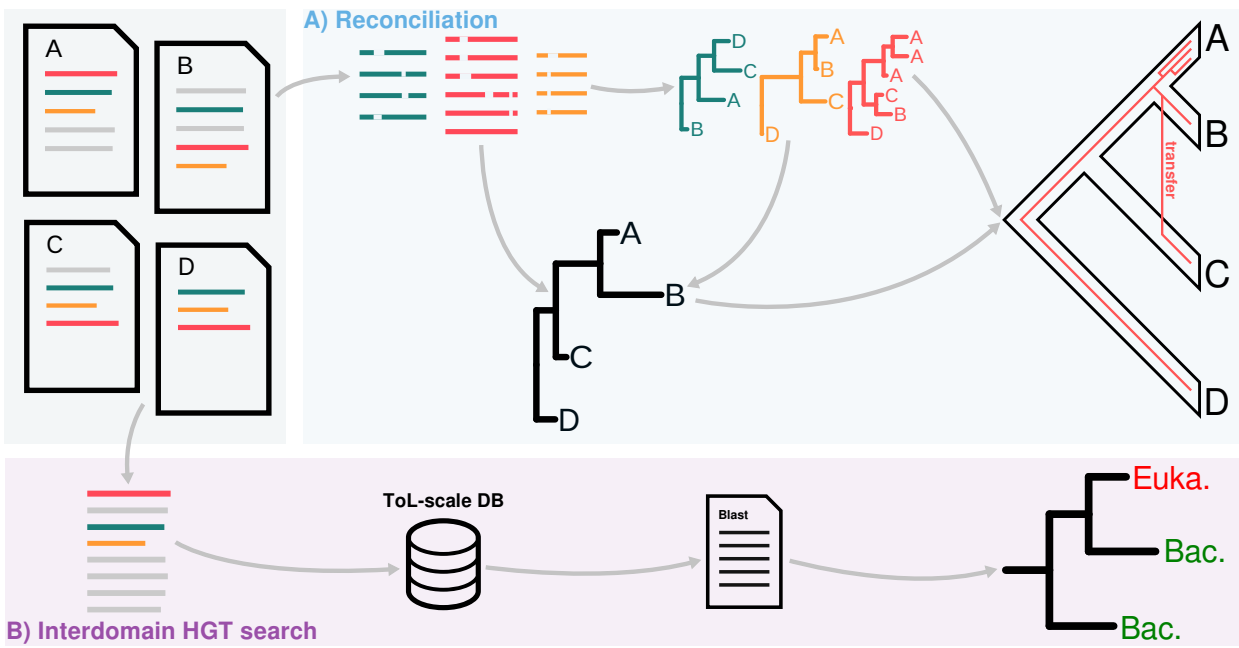


Fig. 2. The two computational approaches used to detect non vertical evolution. A) Genes are clustered in gene families and a phylogeny is computed for each. Selected alignments and gene trees are used to infer a species tree. Events in the gene trees, such as *duplications* and *transfers* are modeled with a reconciliation algorithm. B) Homologs of orphans and of a representative per gene family are searched in a Tree of Life scale DB. If the resulting hits show a patchy taxonomic distribution, a seed-based phylogeny is computed and annotated as HGT event. If possible, donors and acceptors are defined.

[9], EukProt [10] and P10K [11], prokaryotic proteomes from GTDB [12] and viral genomes from RefSeq.

#### D. Conclusions

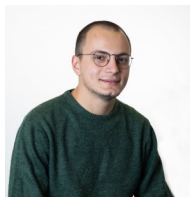
A systematic analysis of non-vertical evolution in Eukaryotes with consistent parameters is needed. This is because generally, different methods are used for each genome, making it difficult to compare results among studies [1]. Further, previous systematic efforts only explored interdomain HGT events [13].

## II. ACKNOWLEDGMENT

Giacomo's predoctoral research is supported by a Fundación "la Caixa" INPhINIT Incoming grant (code LCF/BQ/DI22/11940014).

## REFERENCES

- [1] P. J. Keeling, "Horizontal gene transfer in eukaryotes: aligning theory with data," *Nature Reviews Genetics*, pp. 1–15, 2024.
- [2] T. Gabaldón, "Patterns and impacts of nonvertical evolution in eukaryotes: a paradigm shift," *Annals of the New York Academy of Sciences*, vol. 1476, no. 1, pp. 78–92, 2020.
- [3] J. Van Etten and D. Bhattacharya, "Horizontal gene transfer in eukaryotes: not if, but how much?" *Trends in Genetics*, vol. 36, no. 12, pp. 915–925, 2020.
- [4] P. J. Keeling and Y. Eglit, "Openly available illustrations as tools to describe eukaryotic microbial diversity," *PLoS Biology*, vol. 21, no. 11, p. e3002395, 2023.
- [5] D. M. Emms and S. Kelly, "Orthofinder: phylogenetic orthology inference for comparative genomics," *Genome biology*, vol. 20, pp. 1–14, 2019.
- [6] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear, "Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era," *Molecular biology and evolution*, vol. 37, no. 5, pp. 1530–1534, 2020.
- [7] A. K. Tice, D. Žihala, T. Pánek, R. E. Jones, E. D. Salomaki, S. Nernarokov, F. Burki, M. Eliáš, L. Eme, A. J. Roger *et al.*, "Phylofisher: a phylogenomic package for resolving eukaryotic relationships," *PLoS Biology*, vol. 19, no. 8, p. e3001365, 2021.
- [8] B. Morel, T. A. Williams, A. Stamatakis, and G. J. Szöllösi, "Alerax: A tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss," *bioRxiv*, pp. 2023–10, 2023.
- [9] U. Consortium, "Uniprot: the universal protein knowledgebase in 2023," *Nucleic acids research*, vol. 51, no. D1, pp. D523–D531, 2023.
- [10] D. J. Richter, C. Berney, J. F. Strassert, Y.-P. Poh, E. K. Herman, S. A. Muñoz-Gómez, J. G. Wideman, F. Burki, and C. de Vargas, "Eukprot: a database of genome-scale predicted proteins across the diversity of eukaryotes," *Peer Community Journal*, vol. 2, 2022.
- [11] X. Gao, K. Chen, J. Xiong, D. Zou, F. Yang, Y. Ma, C. Jiang, X. Gao, G. Wang, S. Gu *et al.*, "The p10k database: a data portal for the protist 10 000 genomes project," *Nucleic Acids Research*, vol. 52, no. D1, pp. D747–D755, 2024.
- [12] D. H. Parks, M. Chuvpochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, and P. Hugenholtz, "Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy," *Nucleic acids research*, vol. 50, no. D1, pp. D785–D794, 2022.
- [13] A. Cote-L'Heureux, X. X. Maurer-Alcalá, and L. A. Katz, "Old genes in new places: a taxon-rich analysis of interdomain lateral gene transfer events," *PLoS genetics*, vol. 18, no. 6, p. e1010239, 2022.



**Giacomo Mutti** received his BSc degree in Biology from Università degli studi di Pavia in 2019. He completed his MSc degree in Bioinformatics for Computational Genomics at Università degli studi di Milano and Politecnico di Milano in 2021. After graduating he worked 9 months as research fellow in Cristian Capelli Human Evolutionary Genetics group at Università degli studi di Parma. Since November 2022, he is a PhD student in the Comparative Genomics group at Barcelona Supercomputing Center (BSC) under the supervision of Toni Gabaldón.

# Evaluating the Impact of Recurrent Mobility in Air Pollution Exposure in Catalonia

Alejandro Navarro-Martínez\*, Miguel Ponce-de-León\*, Alfonso Valencia\*†

\*Barcelona Supercomputing Center, Barcelona, Spain

†ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

E-mail: {alejandro.navarro, miguel.ponce, alfonso.valencia}@bsc.es

**Keywords**—*Air pollution, Human mobility, Mobile phone data, Dynamic population, Exposure assessment.*

## I. EXTENDED ABSTRACT

### A. Introduction

Air pollution exposure is the leading environmental health risk due to its detrimental respiratory and cardiovascular effects [1]. Assessing a population’s exposure is a challenging task because of their complex mobility patterns. Ignoring this factor could lead to systematic biases when evaluating the effect of air pollution on health outcomes [2].

### B. Methods

In this study, we used a public mobility dataset [3] representative of the population of Catalonia to estimate mobility-informed air pollution exposure (dynamic estimates) and quantify the bias committed when population dynamics are neglected (static estimates) [4][5]. The mobility dataset is extrapolated from a sample of mobile phone users and is aggregated over mobility areas at hourly resolution. We extracted the trips taking place between home and work locations (recurrent mobility) to construct the dynamic population, i.e. the distribution of residents of a home area among the rest of areas. Fine-grained air quality data of 2022 for NO<sub>2</sub>, O<sub>3</sub>, PM<sub>2.5</sub> and PM<sub>10</sub> [6] was projected over the mobility areas to compute air pollution exposure, and static and dynamic estimates were compared.

### C. Results

Between 84 and 95% of the mobility areas showed significantly different dynamic exposure estimates for any of the four pollutants (Table I). The magnitude of these differences was not large enough to entail a relevant health impact when considering the aggregated populations (the mobile population supposed 10% of the population of an area, on average). However, some of the mobile populations were exposed to unsafe air pollution levels (over the daily Air Quality Guidelines limit) for an important additional number of days than we would expect on a static setting, especially for NO<sub>2</sub> (up to 60 extra days). Spatially, the areas surrounding the Barcelona Metropolitan Area (BMA) tended to have increased dynamic exposure estimates for NO<sub>2</sub>, PM<sub>2.5</sub> and PM<sub>10</sub>—and decreased for O<sub>3</sub>—(Figure 1) caused by their flows of people going to work to the BMA (Figure 2).

TABLE I. Number of tests (or mobility areas) where the dynamic-static exposure difference was significantly different from zero (N=584 tests). In the ‘all population’ column, dynamic exposure estimates were calculated with respect to all the population of an area; whereas they were calculated with respect to ‘mobile population’ in its respective column.

Pollutant	No. rejected tests (all population)	No. rejected tests (mobile population)
NO <sub>2</sub>	554 (94.9%)	553 (94.7%)
O <sub>3</sub>	527 (90.2%)	529 (90.6%)
PM <sub>2.5</sub>	513 (87.8%)	518 (88.7%)
PM <sub>10</sub>	490 (83.9%)	493 (84.4%)

### D. Conclusions

This study evidences and quantifies the negative effect of the BMA on the exposure to air pollutants in the surrounding populations. In addition, we highlight the importance of using mobility data with high spatial resolution when assessing dynamic air pollution exposure at the population level. If the resolution is not high enough, although more accurate, the dynamic estimates will not differ much from the static ones. Due to the privacy problems of releasing finely-resolved mobility data, the most viable cases where mobility could be used to fine-tune exposure analysis are private studies at the individual level, like cohort studies.

### E. Future Work

We are underestimating recurrent mobility given that we only consider the trips taking place between home and work locations. We plan to obtain more representative mobility estimates by adding the trips between other activities and work on top of the trips between home and work (we will assume [others ↔ work] trips are proportional to [home ↔ work] trips between any two areas).

In addition, we will do a complementary analysis where we identify the areas which are contributing more to the exposure of the whole population, considering the number of people present in the area at a certain time and the pollutant concentration (we will construct an exposure indicator with units [people × concentration × time]).

We also plan to assess the health impact associated to air pollution exposure by using the dose-response functions of exposure to pollutant concentration (dose) and mortality risk (response) given by the World Health Organization [7].



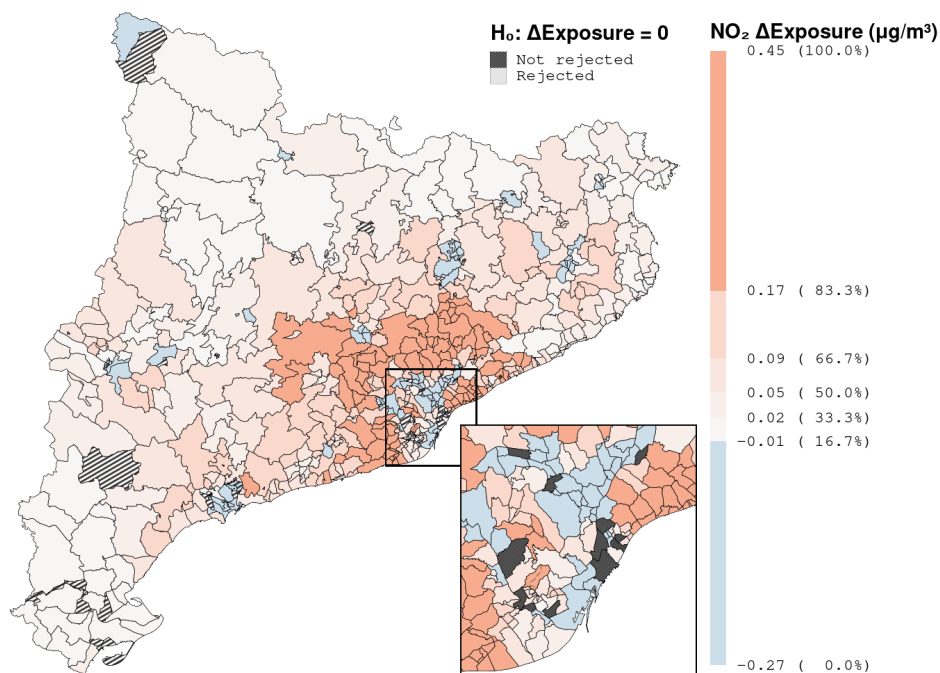


Fig. 1. Spatial pattern of the pseudomedian dynamic-static exposure difference in the 'all population' case for  $\text{NO}_2$ . Pseudomedian exposure difference (dynamic minus static) estimator of each mobility area, colored by quantiles. A striped pattern is placed over the areas where the significance test for the exposure difference was not rejected.

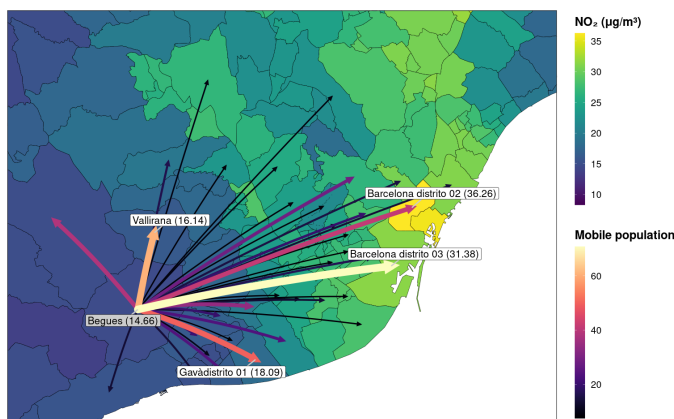


Fig. 2. Annual dynamic population components of Begues (population of 7,356), which presented higher  $\text{NO}_2$  dynamic exposure estimates than static ones.

The arrows indicate the annual mean counts of people travelling recurrently from the home area to each destination. Mobility areas are colored by annual mean  $\text{NO}_2$  concentration. Labels indicate areas that receive great flows of people together with their  $\text{NO}_2$  value.

## II. ACKNOWLEDGMENT

This work has been supported by the MePreCiSa project (record No. REGAGE22e00052915462), funded by the Ministry for Digital Transformation and of Civil Service and by the Recovery, Transformation and Resilience Plan – funded by the European Union - NextGenerationEU. The Earth System Services group of Barcelona Supercomputing Center provided the air quality data, as well as valuable advice and feedback.

## REFERENCES

[1] C. J. L. Murray *et al.*, "Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden

of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1223–1249, Oct. 2020.

- [2] Y. M. Park and M.-P. Kwan, "Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored," *Health & Place*, vol. 43, pp. 85–94, Jan. 2017.
- [3] MITMA, "Estudio de movilidad de viajeros de ámbito nacional aplicando la tecnología Big Data – versión 2," <https://www.mitma.gob.es/ministerio/proyectos-singulares/estudios-de-movilidad-con-big-data/.opendata-movilidad>, 2023.
- [4] B. Dewulf *et al.*, "Dynamic assessment of exposure to air pollution using mobile phone data," *International Journal of Health Geographics*, vol. 15, no. 1, p. 14, Apr. 2016.
- [5] M. Picornell *et al.*, "Population dynamics based on mobile phone data to improve air pollution exposure assessments," *Journal of Exposure Science & Environmental Epidemiology*, vol. 29, no. 2, pp. 278–291, Mar. 2019.
- [6] J. M. Baldasano *et al.*, "An annual assessment of air quality with the CALIOPE modeling system over Spain," *Science of The Total Environment*, vol. 409, no. 11, pp. 2163–2178, May 2011.
- [7] World Health Organization, *WHO Global Air Quality Guidelines: Particulate Matter (PM<sub>2.5</sub> and PM<sub>10</sub>), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. World Health Organization, 2021.



**Alejandro Navarro-Martínez** received his BSc degree in Biotechnology from Universitat Politècnica de València (UPV) in 2021. He completed his MSc degree in Bioinformatics from Universitat Autònoma de Barcelona (UAB) in 2022, and worked as a technician in genomic prediction at Centre for Research in Agricultural Genomics (CRAG) during the same year. Furthermore, he completed his MSc degree in Statistics and Operations Research from Universitat Politècnica de Catalunya (UPC) in January 2024, and carried out his master's thesis project about mobility and air pollution exposure at the Computational Biology group of Barcelona Supercomputing Center (BSC-CNS). He has recently started a PhD in Bioinformatics (UPC) with the same group, where he will do research in epidemiology and comorbidities.

# Big data and diversity: the specificities of analyzing discourses about refugees in Brazil

Lidia Gurgel Neves-Hora<sup>#1</sup>

<sup>#</sup>Linguistics Post Graduation Program, Universidade Federal do Espírito Santo, Brazil, visitor at Language Technologies Unit at Barcelona Supercomputing Center, Spain

<sup>1</sup>lidia.hora@edu.ufes.br

**Keywords**— Digital Discourse Analysis; Social Media Analysis; Refugees; Linguistics; Brazil

## EXTENDED ABSTRACT

The question of the refugees is a global and urgent question all around the world. In Europe, migration is a main issue in politics, as the continent counts 24.9 million forcibly displaced people or stateless people – numbers that increased during the Ukraine conflict [1], [2].

But in other parts of the world, such as Brazil, refugees have become a recent social topic, as the political conflicts in Venezuela increased, mainly since 2015. In November 2023, the R4V (Response for Venezuelans) Regional Interagency Coordination Platform counted 510.499 Venezuelans living in Brazilian territory [3], what mobilized the society and the national and local governments and other State institutions to build temporary and permanent policies for refugees [4].

In this study, we will focus on presenting the differences of analyzing discourses about refugees in Brazil, with few analysis so far, compared to other researches made in other parts of the world, mainly in Europe [5]. Those analysis bring different aspects of discourses, from specific profiles/pages or countries, that point to the migration/refugees issue as a central and polarized question. Based on this, we intend to discuss the importance of considering diversity and biases in the society, seeking for paths not only in social media, but also in other technology-mediated uses of language, including large language models [6], [7].

Based on quantitative Social Media Analysis [8], [9] and on qualitative Digital Discourse Analysis [10], we analyze, in quali-quantitative method, discourses about refugees that circulated on Facebook in the two initial years of covid-19 pandemic (2020-2021) in Portuguese, mainly in Brazil.

### A. Method of Analysis

The Perspectivist Method of Social Media Analysis involves extracting datasets, mining data and generating graphs that allow checking perspectives on a given topic at a given time, checking the traces of associative cooperation – between people, things or profiles [8, p. 91], [11, p. 3065]. For this analysis, we will use a corpus of 38 thousand posts published on Facebook, mainly from Brazilian pages. The publications were extracted through Crowdtangle, Meta Groups' platform for researchers [12]. The query used to find the corpus was "refugiados" (refugees in Portuguese).

After the data collection, we focus on the "message" of Facebook posts and use the application Ford, that was developed at Image and Cyberculture Laboratory (Labic) of the Federal University of Espírito Santo (Ufes), for data mining [13] of lexicons and cited actors.

Then, the Gephi software is used to generate graphs that identify the centrality and strength of some nodes (representing actors, lexicons or hashtags, for example), in relation to the whole corpus [14]. Each group of relationships among nodes is identified with a different color in the graph and each color is

called a perspective that can be analyzed separately. For a better comprehension, we also analyze some posts in their environment.

### B. Analyzing

In the lexical analysis, it was possible to identify many of the multiple vulnerabilities suffered by refugees during the pandemic. Discourses point to the overlapping of the situation of forced migration with those related to gender, age, racism, unemployment, poverty and poor access to health, education, work, housing, food and water. There is also a general feeling of solidarity with the global situation of this group. When it comes to refugees in Brazil, however, there is a strong presence of a stance of rejection, which can be considered xenophobic, but at the same time, shows the fragility of access to rights that also affects Brazilians in their country.

The examination of lexicons and key actors underscores the pivotal role played by institutions like international organizations and NGOs, alongside public figures and influencers—including artists, advocates, and politicians—in addressing the refugee cause, fostering solidarity and stimulating awareness-raising efforts.

At a socio-historical moment in which discourses are highly polarized, it is noteworthy that there is no main controversy regarding the topic in Brazil. This could be an opportunity to favor welcoming speeches over xenophobic ones. At the same time, it shows the importance of local studies on global themes.

### C. Conclusion and Future Enhancement

The Perspectivist Method of Social Media Analysis, associated to Digital Discourse Analysis, enabled ways of categorizing a large volume of speeches on the social network, helping to identify issues and propose solutions.

In the same way that this study focused on lexicons and actors, we intend to expand it, with the analysis of hashtags and deepening the analysis of key actors (in addition to influencers, we mention international organizations, politicians, civil society organizations, religious groups and refugee actors) and the political-ideological discourses related to the issue of refuge, as some of them are associated with other socio-political-ideological issues, such as the "Venezuelan communism" and some Brazilian affirmative policies.

As we can notice so far, although migrations and refuge are a global issue, they have some specific aspects according to the section analyzed. This is an important question when thinking about seeking for solutions to the issue and also when discussing ways of increasing data from large language models: different societies could have different issues about the same topic, and it is a challenge for large language models (LLMs) builders to reflect this in a locally-informed way that

contemplates not only the racial, gender, cultural, ethnic diversity, but also that are able to point to these varied contexts. Future enhancement could verify the possibilities of using the same method to analyze biases in LLMs.

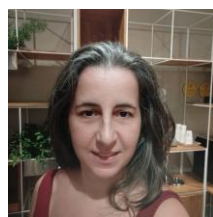
#### D. ACKNOWLEDGEMENTS

We thank Dr. Marta Villegas, leader of Language Technologies Unit at Barcelona Supercomputing Center, for the supervision during the period of 10 months as a visitor, with a scholarship offered by Brazil's government Coordination of Superior Level Staff Improvement (Capes); Dr. Fábio Malini, for the supervision of the PhD at Ufes; and bachelor in Communications Renata Coutinho, for the support for making graphs during Scientific Initiation Fellowship from Brazil's National Council of Scientific Researches (CNPq).

#### References

- [1] UNHCR, “Europe”, Global Focus, 2024. Accessed: March 15, 2024. [Online]. Available: <https://reporting.unhcr.org/global-appeal-2024/regional-overviews/europe>
- [2] L. Schulten, “Is migration the EU’s biggest challenge in 2024?”, DW, February 1, 2024. Accessed: March 15, 2024. [Online]. Available: <https://www.dw.com/en/is-migration-the-eus-biggest-challenge-in-2024/a-67859874>
- [3] R4V, “Brazil. Key Figures”. Accessed: January 16, 2024. [Online]. Available: <https://www.r4v.info/es/node/247>
- [4] D. M. Pereira, “Humanitarianism and militarisms: the role of the Armed Forces in the Brazilian State’s response to Venezuelan migrations (2018-2022)”, Thesis presented to the Postgraduate Program in Strategic Studies of Defense and Security, Federal Fluminense University, Niterói-RJ, 2023.
- [5] L. G. Neves-Hora, R. R. Coutinho, e F. Malini de Lima, “A contribuição do método perspectivista à análise do discurso digital a partir do estudo sobre refugiados na pandemia”, em *Estudos discursivos do Monjolinho: questões entorno do digital*, Campinas, SP: Editora da Abralim, no prelo.
- [6] N. Turner Lee, “Detecting racial bias in algorithms and machine learning”, *JICES*, vol. 16, no 3, p. 252–260, ago. 2018, doi: 10.1108/JICES-06-2018-0056.
- [7] R. Heilweil, “Why algorithms can be racist and sexist”, *Vox*, 18 de fevereiro de 2020. Acesso em: 25 de outubro de 2023. [Online]. Disponível em: <https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency>
- [8] F. Malini, “Um método perspectivista de análise de rede social: cartografando territórios e tempos na rede”, em *Comunicação e territorialidades: poder e cultura, redes e mídias* (e-book), D. Zanetti e R. Reis, Orgs., Vitória: Edufes, 2017, p. 83–106.
- [9] R. Recuero, *Análise de redes para mídia social*. Porto Alegre, RS: Editora Sulina, 2015.
- [10] M.-A. Paveau, *Análise do discurso digital: dicionário das formas e das práticas*. Campinas: Pontes, 2021.
- [11] J. M. R. Medeiros, “A economia de atenção vista através das centralidades em redes formadas pelas conversações do #naovaitercopa”, em *Anais...*, São Paulo: ECA-USP, 2015, p. 3064–3083.
- [12] Crowdtangle, “Crowdtangle”. 2023. Acesso em: 25 de julho de 2023. [Online]. Disponível em: <https://apps.crowdtangle.com/>
- [13] Labic-Ufes, “Ford”. 2021. Acesso em: 19 de maio de 2021. [Online]. Disponível em: <https://github.com/labic/ford-api-py>
- [14] Gephi, “Gephi: the open graph viz platform. 2021.” 2021. Acesso em: 2 de maio de 2021. [Online]. Disponível em: <https://gephi.org/>

#### Author biography



**Lidia Neves-Hora** was born in São Paulo, Brazil, in 1979. She received the bachelor's degree in Communications, from Universidade de São Paulo, in Brazil, in 2003, and the Master's degree in International Relations and Communication from Universidad Complutense de Madrid, in Spain, in 2005. Since 2021, has been a PhD candidate at Universidade Federal do Espírito Santo, in Vitoria, Brazil, and since september, 2023, has been a visitor at Barcelona Supercomputing Center, at Language and Technologies Department, under the supervision of professor Marta Villegas. Former scholarship holder at the Brazilian Center for Latin American Studies (CBEAL/Memorial da América Latina, in Brazil). Has 20 years of experience in digital communication and public communication, work for which she received the awards from Fapes/Confap 2021, A Rede 2013 and Avina 2008. Participates at the Study Group on Media Discourse (Gedim/Ufes), at the Media Observatory/Ufes and at Laboratory of Epistemological Studies and Multimodal Discursivities from Universidade Federal de São Carlos (Leedim/UFSCar). Email: [lidia.hora@edu.ufes.br](mailto:lidia.hora@edu.ufes.br) / [lidianeves@gmail.com](mailto:lidianeves@gmail.com).

# Adjusting UV-Vis Spectrum of Alizarin by Insertion of Auxochromes

Z. Noori<sup>#1</sup>, J. Poater<sup>\*#2</sup>

<sup>#</sup>*Departament de Química Inorgànica i Orgànica & IQTCUB, Universitat de Barcelona, Martí i Franquès 1–11, 08028 Barcelona (Spain)*

<sup>1</sup>zahra.noori@ub.edu, <sup>2</sup>jordi.poater@ub.edu

<sup>\*</sup>*ICREA, Passeig Lluís Companys 23, 08010 Barcelona (Spain)*

**Keywords**— Alizarin · Aromaticity · Density Functional Theory · Electronic structure · UV-vis

## ABSTRACT

First synthesized in 1868, alizarin became one of the first synthetic dyes and was widely used as a red dye in the textile industry, making it more affordable and readily available than the traditional red dyes derived from natural sources. Despite extensive both experimental and computational analyses on the electronic effects of substituents on the shape of the visible spectrum of alizarin and alizarin Red S, no previous systematic work has been undertaken with the aim to fine tune the dominant absorption region defining its color by introducing other electron-withdrawing or electron-donor groups. For such, we have performed a comprehensive study of electronic effects of substituents in position C<sub>3</sub> of alizarin by means of a time dependent DFT approach. These auxochromes attached to the chromophore are proven to alter both the wavelength and intensity of absorption. It is shown that the introduction of an electron-donor group in alizarin causes the transition bands to be significantly red-shifted whereas electron-withdrawing groups cause a minor blue-shifting.

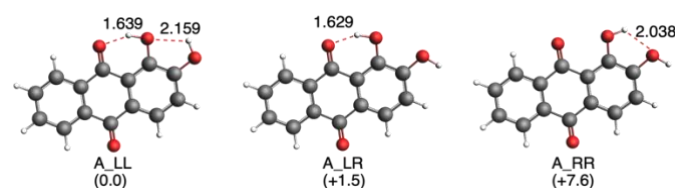
## COMPUTATIONAL METHODS

All DFT calculations were performed with the Amsterdam Density Functional (ADF) program using relativistic, dispersion-corrected density functional theory (DFT) at the ZORA-B3LYP-D3(BJ)/TZP level of theory for geometry optimizations and energy calculations, with the full electron model for all atoms (no frozen core), in gas phase. All stationary points were verified to be minima on the potential energy surface through vibrational analysis. TD-DFT calculations were carried out at the same ZORA-B3LYP-D3(BJ)/TZP level also in methanol, that was simulated by using the conductor-like screening model (COSMO). Use of a continuum solvation model for computing UV-Vis spectra has been proven to better perform than a discrete one and at a reasonable computational cost. Importantly, we consider methanol as a solvent to avoid the extreme sensitivity of the UV-Vis spectrum to pH when either alizarin or alizarin Red S are solved in water.

## RESULTS AND DISCUSSION

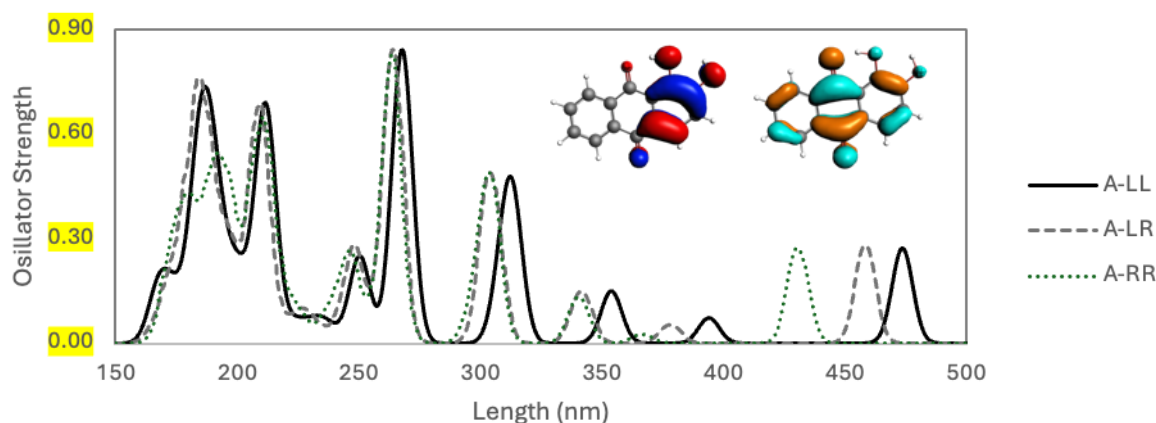
Exploration of the potential energy surface of alizarin shows that it can adopt four planar conformations based on the organization of the hydroxyl groups, three of them are found to be equilibrium geometries (LL, LR and RR), whereas one corresponds to a transition state (RL) because of the steric repulsion. [1]

Regarding alizarin, A\_LL is the most stable isomer because of the formation of two hydrogen-bonds by the hydroxyl groups (Figure 1).



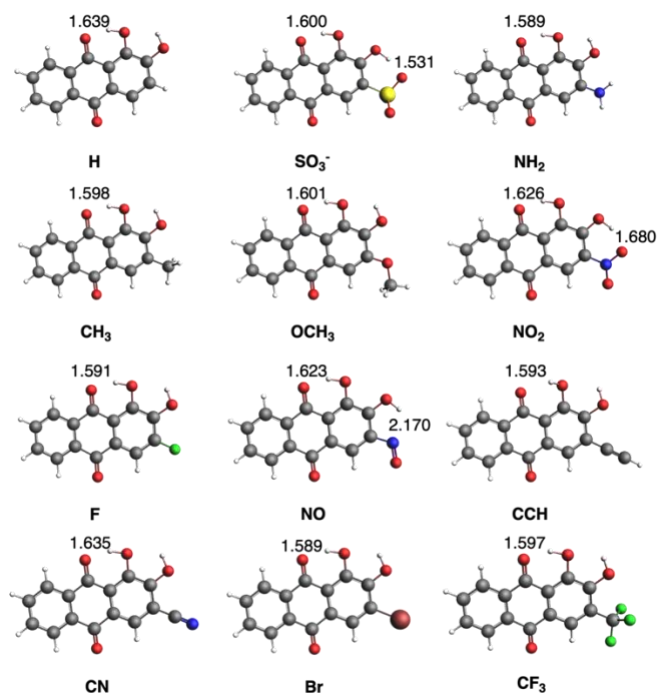
**Figure 1.** Optimized geometries of isomers of alizarin. Hydrogen-bond lengths (in Å) and relative Gibbs free energies (in parentheses, in kcal mol<sup>-1</sup>) are enclosed.

the next step is the analysis of the computed TD-DFT UV-Vis spectra for the alizarin isomers (Figure 2). For such, we mainly focus on the bands with larger wavelengths, all of them involving  $\pi$ - $\pi^*$  electronic transitions. Importantly, when going from A\_LL to A\_LR to A\_RR the bands are blue-shifted. This result can be justified by the presence of the O<sub>9</sub>...H-O<sub>1</sub> hydrogen bond and further stabilized by the O<sub>1</sub>...H-O<sub>2</sub> one. The former interaction is also present in A\_LR isomer, but not the latter, leading to relatively close stabilization values, in contrast with the A\_RR isomer without hydrogen bonds between the hydroxyl groups. At this point, it must be pointed out that in order to compare the calculated spectra with the experimental UV-Vis absorption spectrum of alizarin at ambient temperature, we have to take into account the weight of the spectra from each isomer using a Boltzmann factor for their average. Also, isomer A\_LL can easily undergo proton transfer from O<sub>1</sub> to O<sub>9</sub>, thus also contributing to the experimental spectra.



**Figure 2.** UV-Vis spectra of the three isomers (LL, LR, and RR) of alizarin. HOMO (red/blue) and LUMO (brown/turquoise) orbitals of A\_LL isomer are also show

Once the geometries, isomerization energies and UV-Vis spectra of alizarin are set up, we proceed with the introduction of substituents on C<sub>3</sub> carbon atom with a series of electron-donor (EDG) and -withdrawing (EWG) groups (Figure 3). [2]



**Figure 3.** Lowest-energy optimized geometries in methanol of all substituted alizarin systems with the main hydrogen bond lengths (in Å).

## CONCLUSIONS

The introduction of an electron-donor auxochrome in alizarin causes the transition bands to be significantly red-shifted (ca. +70 nm in methanol). At difference, the introduction of electron-withdrawing auxochromes cause a minor blue-shifting (ca. -20 nm in methanol). Analysis of valence bond structures gives a rational explanation of the above behavior based on the stability of the structures depending on the introduction of either of an EDG or EWG which hardly affect the aromaticity of the substituted alizarin rings.

- 1) Noori, Z., Moreira, I. D. P., Bofill, J. M., & Poater, J. (2024). Adjusting UV-Vis Spectrum of Alizarin by Insertion of Auxochromes. *ChemistryOpen*,
- 2) Noori, Z., Malekzadeh, A., & Poater, J. (2024). Brownmillerite Calcium Ferrite, a Promising Perovskite-Related Material in the Degradation of a Tight Dye under Ambient Conditions. *ChemistryOpen*, 13(3), e202300169.

## Author biography



**Zahra Noori** was born in Shush city, Iran, in 1979. She received the BSc degree in Pure Chemistry from Iran, in 2002, the master degree in Inorganic Chemistry from Iran, in 2009, she was teacher in Chemistry in Iran from 2003 till 2022 and the Ph.D. degree in Inorganic Chemistry from Iran, in 2024. Zahra had two PhD supervisors. One of them was Experimental and from

Iran and another one Computational and from Barcelona university of Spain. In 2021, she got a scholarship from Iran and came to the University of Barcelona (February 2022) to work on computational chemistry under the supervision of **Prof. Dr. Jordi Poater** at the IQTCUB institute in Barcelona. The second part of her thesis is based on computational chemistry and Zahra started her work with quantum chemical ADF software.

After her scholarship till now, she got a contract with Prof. Jordi Poater at Barcelona University with the aim to complete the above projects.

# Inter-individual and Inter-tissue variation of DNA methylation

Winona Oliveros Diez<sup>1#</sup>, José Miguel Ramirez<sup>2#</sup>, Marta Melé<sup>#3</sup>

<sup>#</sup>Life Sciences Department, Barcelona Supercomputing Center (BSC)

<sup>1</sup>winona.oliveros@bsc.es, <sup>3</sup>marta.mele@bsc.es

<sup>2</sup>jose.ramirez1@bsc.es

**Keywords**— DNA methylation, aging, Genetic ancestry, GTEx

## Extended ABSTRACT

### A. Introduction

DNA methylation is the major and most studied epigenetic mechanism involving direct chemical modification to the DNA. In the mammalian genome, DNA methylation involves the transfer or removal of a methyl group (-CH<sub>3</sub>) onto the C5 position of , almost exclusively, the cytosine at CpG dinucleotides [1-2].

The examination of the DNA methylation distribution throughout the genome is needed to understand its functionality. Mammalian genomes exhibit a global CpG depletion, with 60–80% of the approximately 28 million CpGs present in the human genome being typically methylated. Aside from the randomly distributed CpG sites across the genome, a majority of genes contain brief (approximately 1 Kb) CpG-rich regions identified as CpG islands (CGI) which are generally resistant to DNA methylation [3].

The deposition and maintenance of DNA methylation are essential for normal mammalian development. Hence, aberrations in DNA methylation are associated with different diseases [4].

Many studies have identified individual CpGs whose methylation status is significantly correlated with aging [5]. In fact, the accumulation of DNA methylation changes with aging measured as epigenetic clocks have been shown to accurately predict chronological age and mortality. Additionally, some recent studies have begun to identify DNA methylation changes associated with genetic ancestry [6] and sex [6]. Despite the valuable insights provided by recent studies, the study of DNA methylation variation has predominantly been confined to specific individual traits, a limited number of tissues, and small-scale DNA methylation arrays. Consequently, the collective impact of demographic traits, such as age, ancestry, and sex across diverse tissues remains largely unexplored. In this study, we leverage Genotype-Tissue Expression (GTEx) data to systematically investigate the relationships among various demographic traits and DNA methylation variation across 9 diverse human tissues.

### B. Materials and methods

#### Sample collection

All human donors were deceased, with informed consent. For details on donor characteristics, see the GTEx v8 main paper, and for details on the DNA methylation sequencing pipeline, see [7].

#### Differential methylation analysis

We downloaded normalized beta counts of the 754,054 CpGs from the Infinium Methylation EPIC array generated in [7] and stored in GEO (GSE213478). We used limma to run linear models on M values and corrected for the following set of covariates:

$M \text{ values} \sim \text{HardyScale} + \text{IschemicTime} + \text{PEER1} + \text{PEER2} + \text{PEER3} + \text{PEER4} + \text{PEER5} + \text{Ancestry} + \text{Sex} + \text{Age} + \text{BMI}$

We included 5 PEERs in the models. These factors correct for technical effects and cell-type compositions [7].

#### Annotation of DMPs

**General classification.** The DMPs were classified depending on their location at promoters, enhancers, gene bodies or intergenic based on the annotations provided in the EPIC v1.0 array manifest b5.

**Tissue-specific classification.** To annotate the chromatin states around each array probe we used the 18 chromatin states inferred with ChromHMM generated by the ROADMAP Epigenomics consortium and analyzed by EpiMap.

**TFBS.** To study the enrichment of transcription factor binding sites around DMPs, we downloaded the processed CHIP-seq data on transcription factors for the human v19 Lung, Kidney, Blood, Muscle, Breast, Digestive Tract, Prostate and a general catalog available in ChipAtlas.

#### Enrichments of annotations

We performed Fisher's exact test for each transcription factor (TF) separately for hypomethylated and hypermethylated DMPs and adjusted for multiple testing using Benjamini-Hochberg correction with  $FDR < 0.05$ . We used as background the TF-binding sites (TFBS) found in all tested CpGs. For the analysis of sex-specific TFBS we performed Fisher's exact test for each TF separately for female-DMPs and male-DMPs using as background the TFBS found in all sex-DMPs.

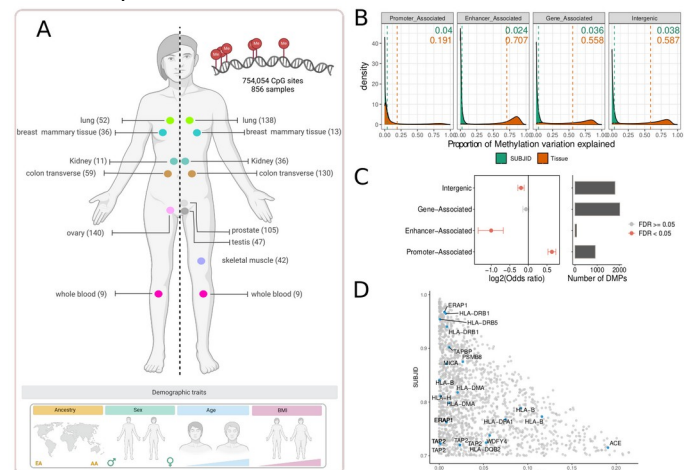
#### Functional enrichment

To perform functional enrichment with the EPIC array we need to take into account that some genes contain more probes than others. For this goal, we used the function gometh from the missMethyl package to get GO:BP terms using as background the 754,054 positions studied from the array.

### C. Results

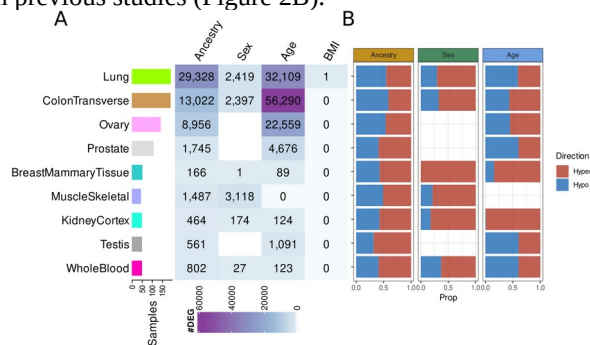
To investigate the role of individual-related traits and tissues in DNA methylation we used the data from 9 GTEx tissues from 424 individuals across 754,054 CpG sites (Figure 1A, Supplementary Figure 1A). We used linear mixed models (see methods) to quantify the contribution of individuals and tissues to DNA methylation variation. As expected, variation in DNA methylation is far greater among tissues (~62% of the total variance in DNA methylation) than among individuals (~1.4% of the total variance). The close resemblance in DNA methylation among individuals is comparable to the estimated variability in genomic sequence among individuals. The genomic location of DNA methylation changes is of interest to get insights into their putative impact on gene expression. Methylation at enhancer-associated CpGs shows higher tissue variability than other CpGs, consistent with the association of

tissue-specific methylation changes with gene enhancers (Figure 1B, Supplementary Figure 1B). Although methylation at promoter-associated CpGs is highly stable, inter-individual variable CpGs were enriched in promoter regions (Figure 1D). Highly individual-variable CpGs (4810 CpGs with > 50% of DNA methylation variation explained by individual) are enriched in genes belonging to antigen processing and presentation pathways (Figure 1E), one of the most polymorphic loci in vertebrates. On the other hand, highly tissue-variable CpGs are enriched in developmental and metabolic processes.



**Figure 1.** A. Overview of the DNA methylation data available. B. Density plot showing the proportion of DNA methylation variation explained by individual (green) and tissue (orange). The distributions are clustered by the genomic location of the CpGs. C. Left. Dotplot showing the enrichment of each genomic location on individually variable CpGs. Right. Barplot denoting the number of individually variable CpGs associated with each genomic location. D. Scatter plot representing the most individually variable CpGs. Blue dots represent CpGs associated with genes belonging to antigen processing and presentation pathways.

The impact of demographic traits on genome-wide DNA methylation variation across tissues is still not widely understood. We used linear models to simultaneously quantify DNA methylation changes with four demographic traits: genetic ancestry, sex, age, and BMI across 9 different human tissues. We identified differentially methylated positions (DMPs) while controlling for known sources of technical variation and other confounders such as cell-type composition (see methods). Age had the largest number of DMPs, followed by ancestry and sex, with some differences across tissues. Additionally, our results suggest that BMI has no impact on DNA methylation in the tissues studied, as previously hinted by other studies (Figure 2A). Lung, colon transverse, and skeletal muscle had the largest number of DMPs for ancestry, age, and sex, respectively (Figure 2A). Age- and ancestry-DMPs showed no consistent bias in the direction of methylation changes across tissues. Contrarily, sex-DMPs point toward generalized hypermethylation of the female genome, consistent with previous studies (Figure 2B).



**Figure 2.** A. Heatmap showing the number of DMPs per tissue (row) and demographic trait (column). B. Barplot showing the proportion of Hyper (red) and Hypo (blue) methylated positions per tissue and demographic trait.

## References

- [1] Smith, Z., Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 14, 204–220 (2013). <https://doi.org/10.1038/nrg3354>
- [2] Jones PA, Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation. *Cell*. 1980 May;20(1):85-93. doi: 10.1016/0092-8674(80)90237-8. PMID: 6156004.
- [3] Jones, Peter A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7), 484–492. doi:10.1038/nrg3230
- [4] Robertson, K. DNA methylation and human disease. *Nat Rev Genet* 6, 597–610 (2005). <https://doi.org.sire.ub.edu/10.1038/nrg1655>
- [5] Lu, A.T., Fei, Z., Haghani, A. et al. Universal DNA methylation age across mammalian tissues. *Nat Aging* 3, 1144–1166 (2023). <https://doi.org/10.1038/s43587-023-00462-6>
- [6] Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, Ritz BR, Chen B, Lu AT, Rickabaugh TM, Jamieson BD, Sun D, Li S, Chen W, Quintana-Murci L, Fagny M, Kobor MS, Tsao PS, Reiner AP, Edlefsen KL, Absher D, Assimes TL. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol*. 2016 Aug 11;17(1):171. doi: 10.1186/s13059-016-1030-0. PMID: 27511193; PMCID: PMC4980791.
- [7] Oliva M, Demanelis K, Lu Y, Chernoff M, Jasmine F, Ahsan H, Kibriya MG, Chen LS, Pierce BL. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet*. 2023 Jan;55(1):112-122. doi: 10.1038/s41588-022-01248-z.

## Author biography



**Winona Oliveros** was born in Andorra La Vella, Andorra, in 1995. She received the B.E. degree in Microbiology from the Autonomous University of Barcelona, Spain, in 2017, and the MSc. degree in Bioinformatics for Health Sciences from the Pompeu Fabra University (UPF) Barcelona, Spain, in 2019. Since July 2019, she has been with the

Transcriptomics and Functional Genomics Lab, BSC, where she was a Research assistant and became a PhD student in October 2020. Her current research interests include Cancer, Transcriptomics, and Epigenomics.

# A Framework and Methodology for Performance Prediction of HPC Workloads

Júlia Orteu\*, Marc Clascà\*, Marta Garcia-Gasulla\*, Jesús Labarta\*<sup>†</sup>, Elise Jennings<sup>‡</sup>

\*Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

<sup>‡</sup>ParTec AG, München, Germany

E-mail: {julia.orteu, marc.clasca, marta.garcia, jesus.labarta}@bsc.es  
elise.jennings@par-tec.com

**Keywords**—HPC Workloads, Performance Prediction, Runtime Hardware Counters, Instructions per Cycle (IPC), Performance Tools, Parallel Applications, Regression trees, ML & AI

## I. EXTENDED ABSTRACT

### A. Introduction

The presented poster outlines an approach for predicting the performance of High-Performance Computing workloads (HPC). By utilizing data gathered from runtime hardware counters across a range of HPC applications and benchmarks, we develop an artificial intelligence model based on ensemble tree algorithms. This model is capable of forecasting the performance of other HPC applications. This work differs from current research by focusing on the granularity of training and prediction. Specifically, our model is developed utilizing individual computation bursts as input samples for training. Through this approach, we prove that a prediction of the instructions per cycle (IPC) metric of unseen applications is possible based on architectural performance counters that can be obtained easily with already used and convenient performance tools.

### B. Research and Development

The research line focuses on exploring the possibility of predicting the performance, and ultimately the execution time, of known workloads in HPC machines prior to their execution.

Predicting the performance of HPC applications or workloads is a complex task with a widely discussed set of approaches and methodologies. The latest research work on this topic focuses on data-driven methodologies using machine learning, but previous studies present different types of prediction methods: analytical methods, that can be derived from a machine representation, or can be the result of a statistical work; and non-analytical methods, which comprise the artificial intelligence techniques and the simulation methods [1].

Our approach distinguishes itself from existing research by its focus on the granularity of training and prediction. Using performance analysis tools Extrae and Paraver [2] [3], developed in Barcelona Supercomputing Center (BSC), we extract data at the computational burst level and target the IPC metric of every burst as a performance measure.

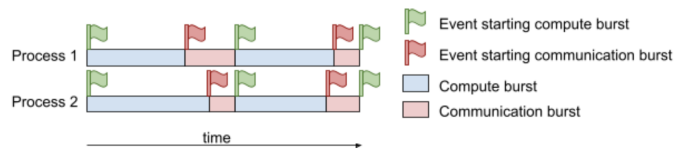


Fig. 1. Extrae and Paraver generalization of burst

A burst is defined as the time interval of active operation between two successive events in a process. In Figure 1, we provide a visual example of a timeline from an execution representation of an application running two processes. The activities within these processes have been categorized into two types: communication bursts, indicated in red, and computation bursts, represented in blue. These categorizations are derived from MPI event markers, with green flags initiating computation and red flags marking the start of communication.

In our study, we concentrate on computation bursts, which we refer to as *useful bursts*. These bursts are periods where the application is actively engaged in data processing and executing instructions. These individual useful bursts serve as the input samples to the model, as one entry of our dataset, highlighting that this approach provides a very precise performance prediction of a very specific application's part.

### C. Features and workload characterization

To facilitate our study, we consider a set of Performance Application Programming Interface (PAPI) counters [4] as the foundational data. These counters include the total number of instructions completed ( $N$ ), the total cycles seen by thread ( $Cyc$ ), memory load instructions ( $N_{LD}$ ), memory store instructions ( $N_{SR}$ ), branch instructions ( $N_{BR}$ ), and total cache miss events at both L1 and L3 levels ( $miss_{L1}$ ,  $miss_{L3}$ ). We normalize the data in each burst by using the ratio of instructions and cache misses instead of absolute counters values. This allows us to compare any individual burst from any part of a trace and from any application. From the counters, we characterize the variables using the following computations:

- Instruction mixes, which are ratios of specific instruction types to the total number of instructions, such as  $r_{LD} = \frac{N_{LD}}{N}$  for loads,  $r_{SR} = \frac{N_{SR}}{N}$  for stores, and  $r_{BR} = \frac{N_{BR}}{N}$  for branches.



- Cache miss rates, calculated as  $r_{L1} = \frac{miss_{L1}}{N_{LD} + N_{SR}}$  for the L1 cache and  $r_{L3} = \frac{miss_{L3}}{N_{LD} + N_{SR}}$  for the L3 cache, which are indicators of memory access efficiency.
- The average node concurrency during core execution, through the integral of the Parallelism function over the time from  $T_{begin}$  to  $T_{end}$ :  $\int_{T_{begin}}^{T_{end}} Par(t) dt$ .
- IPC, which is the target performance metric, given by the equation  $IPC = \frac{N}{Cyc}$ .

#### D. Data sources

The process involves the selection of a specific set of benchmarks and kernels to extract data and adapt it for the purpose of training the models. Additionally, a separate set of applications has been chosen for evaluating the performance of the trained models (testing).

The selection of kernels and benchmarks for the training set is a pivotal decision that facilitate the representation of the burst space, enabling the generalization to new applications. For each application selected for the training set, we've varied the problem sizes to capture a comprehensive dataset. The nature of the variation depends on the application's characteristics—it could be the size of an array, the granularity of a mesh, or the complexity of inputs. Additionally, we've scaled the computational workload by altering the number of processes within a single node for each variant of problem size. This methodical approach allows us to construct a training dataset that covers a wide range of scenarios.

#### E. Data extraction Framework

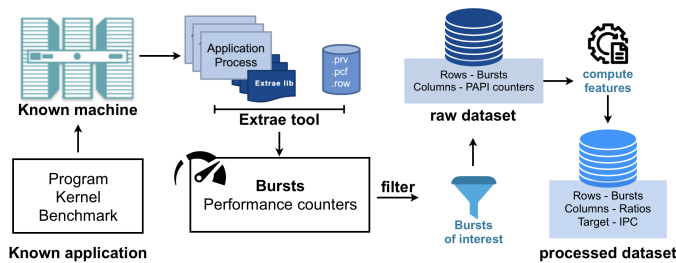


Fig. 2. Flowchart of the data extraction part of the training framework

The data extraction process for each application, both for training and testing datasets, is outlined in Figure 2.

The depicted data extraction process begins with the execution of a known application on a known machine. In this case, we have used the *MareNostrum 4* supercomputer's architecture for reference. We obtain the useful bursts from a program execution using the BSC tools, which are then processed into a features format.

#### F. AI Model

The study investigates a range of diverse machine learning algorithms with the aim of training an effective predictive model. In this exploration, we have employed a 10-fold cross-validation method on our training dataset to evaluate the performance and compatibility of these algorithms

with our type of data. The empirical results highlight a clear advantage of using Boosting Ensembles that rely on decision trees, leading to a focus on XGBoost method [5]. This also matches with previous research conclusions [6].

To ensure a fair assessment of the models, we've developed a method to inject these predictions into Paraver traces and we've devised specific error metrics into the trace tailored to evaluate the performance outcomes of each model. This allows for a more precise analysis of how well the models predict application performance.

#### G. Summary

The poster shows the main findings and discusses our exploration of this topic. We present a method of data collection and preprocessing based on low-effort program instrumentation and automatic tools, as it is well known in the literature that being able to collect data automatically and building the model effortlessly is critical to end up with valuable and convenient training and prediction workflow.

We also studied how changing the feature set, the training data size or the machine learning algorithm affects the accuracy. We discuss a method to characterize computational bursts based on instruction mix features and instantaneous machine concurrency that is later able to classify, using trees, the IPC of unseen bursts. Therefore, we prove that it is possible to foresee the performance of a whole unseen application trace based on this characterization method.

## II. ACKNOWLEDGMENT

This work has been published in proceedings of the 11th International BSC Severo Ochoa Doctoral Symposium, 2024.

## REFERENCES

- [1] J. Flores-Contreras *et al.*, "Performance prediction of parallel applications: a systematic literature review," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 4014–4055, 2021.
- [2] V. Pillet *et al.*, "Paraver: A tool to visualize and analyze parallel code," in *Proceedings of WoTUG-18: transputer and occam developments*, vol. 44, 1995, pp. 17–31.
- [3] H. Servat *et al.*, "Framework for a productive performance optimization," *Parallel Computing*, vol. 39, no. 8, pp. 336–353, 2013.
- [4] S. Browne *et al.*, "Papi: A portable interface to hardware performance counters," 1999.
- [5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.
- [6] J. Sun *et al.*, "Automated performance modeling of hpc applications using machine learning," *IEEE Transactions on Computers*, vol. 69, no. 5, pp. 749–763, 2020.



**Júlia Orteu** is a student in the first cohort of the Bachelor's degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC). Since 2023, has been a Junior Research Engineer at the Barcelona Supercomputing Center (BSC) in the Best Practices for Performance and Programmability (BePPP) group, where she currently works as a performance prediction analyst for HPC Application Workloads.

# Evaluating computational performance metrics in Climate modelling: Insights from CMIP6

Sergi Palomas, Mario C. Acosta, Gladys Utrera  
Barcelona Supercomputing Center, Barcelona, Spain  
Universitat Politècnica de Catalunya, Barcelona, Spain  
E-mail: {sergi.palomas, mario.acosta}@bsc.es, gladys.utrera@upc.edu

**Keywords**—*Performance, HPC, Climate modelling, multi-model, CMIP6*

## I. EXTENDED ABSTRACT

The Coupled Model Intercomparison Project (CMIP) is one of the biggest global efforts aimed at understanding the Earth’s climate by undergoing a multi-model analysis. In its sixth phase, CMIP6 [1], a total of 190 different climate experiments were used to simulate approximately 40000 years which produced 40 petabytes of data in the process. This work presents the primary findings for the collection of a common set of performance metrics specially designed for climate modelling, the Computational Performance Model Intercomparison Project (CPMIP). These metrics were systematically collected from production runs conducted within CMIP6, predominantly from institutions affiliated with the IS-ENES3 [2] consortium.

Through this comprehensive data collection effort, we contribute to the establishment of a robust database for future community reference, thereby setting a benchmark for evaluation and facilitating essential multi-model, multi-platform comparisons crucial for advancing climate modeling performance. Given the diverse array of applications, configurations, and hardware employed, further endeavors are imperative for standardization, for instance, by using climate and weather benchmarking codes such as the HPCW [3].

### A. The metrics

Balaji et al. [4] proposed a set of 12 performance metrics that define the Computation Performance for Model Intercomparison Project (CPMIP) which were specifically designed for climate science by taking into account the structure of ESMs and how they are executed in real experiments. This set of metrics include the climate experiment and platform properties, the computational speed and cost (core-hours and energy), measures for the coupling and I/O overhead, and for the memory consumption. Some of the most relevant ones collected are described in Table I.

### B. The data collection

The collection effort has been predominantly led by institutions affiliated with the IS-ENES3 [2], a consortium founded by a Horizon 2020 project composed of the most important weather and climate centres in Europe. This compilation is the first of its kind and constitutes a representative part of the whole CMIP6. Our data encompasses 33 distinct experiments that were used during CMIP6 to simulate almost 500 000 years across 14 different HPC machines and involving 14 independent modelling institutions.

### C. Results

1) *Parallelization and execution cost*: The *parallelisation* (i.e. the number of parallel resources allocated) is, of course, closely related to the speed of a model, thereby directly impacting the computational *cost* (*CHSY*) of model execution. As seen in Figure 1, the parallelisation and the *CHSY* are closely correlated in low-resolution models, indicating limited scalability within the current generation of HPC platforms. Otherwise, one would see that the *CHSY* of ESMs with similar resolution do not increase when using more processors given that the models run faster (i.e. higher *SYPD*). Moreover, the degree of parallelisation tends to rise as we move to higher-resolution experiments, as well as the *CHSY*. This indicates that many institutions prioritize maintaining a high to medium resolution while achieving a similar *SYPD* to that of lower-resolution configurations, albeit at the expense of augmenting the *CHSY*.

2) *Carbon footprint*: By considering the useful Simulated Years, the HPC machine efficiency, and the KWH to CO2 conversion rates provided by each energy supplier, we calculated the Carbon Footprint (in tons of CO2) using Equation 1.

$$\text{Carbon Footprint} = \text{Total Energy Cost} \times CF \times PUE \quad (1)$$

The total Carbon Footprint is 1692 tCO2, even when considering experiments conducted by only 8 out of the 49 institutions that are participating in CMIP6. Drawing from Acosta et al. [5], the Earth science group at BSC, comprising approximately 80 individuals, accounted for CO2 equivalents from commuting (29 tCO2eq per year), computing infrastructure (397 tCO2eq per year), building and infrastructure (117 tCO2eq per year), and travel (255 tCO2eq per year), with a total estimated budget of around 800 tCO2eq per year. Consequently, the carbon emissions resulting from the execution of this relatively small subset of experiments more than double our annual budget in a single year.

### D. Conclusion

Improving the performance of climate modeling is key for the future of climate sciences. In this pursuit, the collection of the CPMIP metrics for 33 different experiments conducted during CMIP6 serves as a pivotal step toward gaining comprehensive insights into performance within multi-model, multi-platform projects.

The improvement and development of benchmarks specially designed for climate science will significantly enhance multi-platform performance comparisons, like for instance the

TABLE I. SOME OF THE CPMIP METRICS COLLECTED AND THEIR DEFINITIONS

Metric	Used to evaluate
Resolution (Resol)	number of grid points NXXNYxNZ per component
Simulation Years Per Day (SYPD)	number of simulated years per day (24h) of execution time
Core-hours per Simulated Year (CHSY)	execution cost, measured in core-hours per simulated year
Parallelisation (Paral)	total number of cores allocated for the run
Joules Per Simulated Year (JPSY)	energy needed per year of simulation
Coupling Cost (Cpl C)	computing cost of the coupling algorithm and load imbalance

TABLE II. OTHER CMIP6 MEASUREMENTS. THE "USEFUL" METRIC, WHENEVER USED, ACCOUNTS ONLY FOR EXPERIMENTS THAT LED TO SCIENTIFIC VALUE. THE POWER USAGE EFFECTIVENESS (PUE) DEPENDS ON THE HPC MACHINE USED

Institution	Useful Simulated Years*	Total Simulated Years	Useful Data Produced (PB)	Total Data Produced (PB)	Useful core hours (millions)	Total core hours (millions)	Total Person/Months	Total Energy Cost (TeraJoules)	PUE	Conversion Factor (MWh - kg CO2eq)	Carbon Footprint (tons CO2)
CMCC	965		0.097		1.99		7	1.61	1.84	408	329
CNRM-CERFACS	47,000		1.350	2.48	160.00	365.00	450	6.18	1.43	40	97
DKRZ	1,276	1,321	0.600		5.52			0.41	1.19	184	24
EC-Earth	28,105	38,854	0.800	1.41	31.13	46.36	115	1.24	1.35	357	165
IPSL	75,000	165,000	1.800	7.60	150.00	320.00	200	8.72	1.43	50	172
MPI-M	24,175	35,000	1.930		16.31			0.62	1.19	184	37
NCC-NorESM2	23,096		0.600		27.23	80.00	150	1.69			
NERC	640		0.460		55.50			2.17	1.10	0	0
UKMO	37,237		10.400		683.00			26.70	1.35	87	868

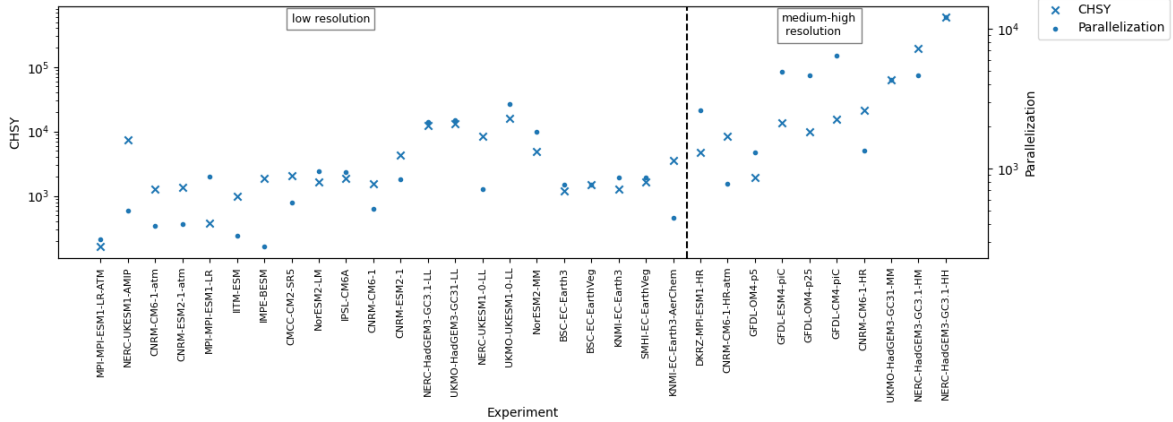


Fig. 1. Comparison between CHSY and Parallelisation for both low and medium-high resolution experiments. Experiment configurations are arranged from left to right in ascending number of gridpoints. Note that vertical axis is in logarithmic scale.

HPCW. Continuous collection of these metrics in forthcoming multi-model endeavors, such as CMIP7, promises for the establishment of a shared database accessible to both the scientific community and technology vendors.

## II. ACKNOWLEDGMENT

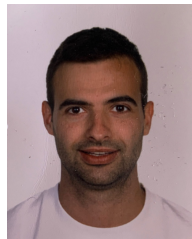
The research leading to these results has received funding from the EU H2020 ISENE3, under grant agreement n° 824084 and co-funding from the Spanish National Research Council through OEMES (PID2020-116324RA-I00)

## REFERENCES

- [1] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, "Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016. [Online]. Available: <https://gmd.copernicus.org/articles/9/1937/2016/>
- [2] S. Joussaume, "IS-ENES: Infrastructure for the European Network for Earth System Modelling," in *EGU General Assembly Conference Abstracts*, ser. EGU General Assembly Conference Abstracts, May 2010, p. 6039.
- [3] B. van Werkhoven, G. van den Oord, A. Sclocco, S. Heldens, V. Azizi, E. Raffin, D. Guibert, L. Lucido, G.-E. Moulard, G. Giuliani, B. van Stratum, and C. van Heerwaarden, "To make Europe's Earth system models fit for exascale - Deliverable D3.5," Feb. 2023, ESiWACE2

stands for Centre of Excellence in Simulation of Weather and Climate in Europe Phase 2. ESiWACE2 is funded by the European Union's Horizon 2020 research and innovation programme (H2020-INFRAEDI-2018-1 call) under grant agreement 823988. [Online]. Available: <https://doi.org/10.5281/zenodo.7671032>

- [4] V. Balaji, E. Maisonnave, N. Zadeh, B. N. Lawrence, J. Biercamp, U. Fladrich, G. Aloisio, R. Benson, A. Caubel, J. Durachta, M.-A. Foujols, G. Lister, S. Mocavero, S. Underwood, and G. Wright, "CPMIP: measurements of real computational performance of Earth system models in CMIP6," *Geoscientific Model Development*, vol. 10, no. 1, pp. 19–34, 2017. [Online]. Available: <http://www.geosci-model-dev.net/10/19/2017/>
- [5] M. Acosta and P.-A. Bretonnière, "Towards minimising carbon footprint of climate modelling: Modelling centre perspective," *C report*, 2018.



**Sergi Palomas** received his BSc degree in Computer Engineering from Universitat Autònoma de Barcelona (UAB), Spain, in 2018. He completed his MSc degree in Innovation and Research in Informatics, specializing in High-Performance Computing (HPC), Spain, in 2022. Since 2019, he has been a member of the computational Earth science group at the Barcelona Supercomputing Center (BSC). In 2023, he enrolled started his PhD studies at the department of computer architecture UPC, Spain.

# HBM performance on FPGAs

Elias Perdomo<sup>\*†</sup>, Teresa Cervero<sup>\*</sup>, Xavier Martorell<sup>\*†</sup> Behzad Salami<sup>\*</sup>,

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {elias.perdomo, teresa.cervero, xavier.martorell, behzad.salami}@bsc.es

**Keywords**—HBM, FPGA, performance, pseudo-channel, micro-switches

## I. EXTENDED ABSTRACT

Main memory access has become an increasing performance bottleneck for traditional and High-Performance Computing (HPC) applications. High Bandwidth Memory (HBM) emerged as an alternative to conventional DRAMs, offering higher bandwidth, lower power, and higher integration capabilities to meet the demands of contemporary applications. The transition of most advanced FPGAs from DDR to HBM confirms this paradigm shift. However, users face substantial challenges due to the scarcity of technical documentation on maximizing HBM features when using FPGAs.

We addressed the knowledge gap for HBM characteristics within FPGAs, aiming to standardize its utilization in the complex HPC domain. Our Memory Sandbox enables analysis within and across HBM pseudo-channels. We show that HBM achieves 99.99% of its nominal peak bandwidth with long sequential memory accesses. However, we observe a performance drop to 0.17% with reduced burst size and random data access patterns.

Our study spotlights the necessity for meticulous management of concurrent accesses and strategic data placement in HBM, offering critical considerations for optimizing HBM performance in FPGA-based systems.

### A. HBM in the current computer architecture environment

Custom hardware – from workstations to PCs– has experienced tremendous improvements in the past decades. However, while the speed of commercial microprocessors has increased by approximately 70% every year, the speed of commodity DRAM has improved by only around 50% in the past decade. As a result, computer systems are experiencing difficulties in achieving high processing efficiency[1].

The traditional approach for boosting performance in systems, particularly those at the edge where huge amounts of data need to be processed locally or regionally; consists of adding more computational capability into a chip and bringing more memory on-chip. But that approach no longer scales since we are now getting to the boundaries of the trifecta of von Neumann architectures, Moore’s Law and Dennard scaling. Consequently, engineers have begun focusing on solving the bottleneck between processors and memories by turning out new architectural designs at a rate no one would have anticipated before.

An alternative, based on recent technological advances, is moving processing elements closer to, or even into the

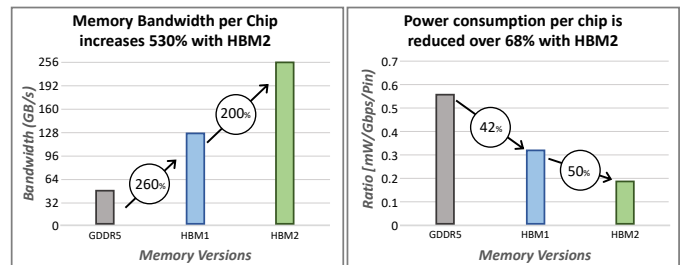


Fig. 1. Bandwidth and power consumption comparison.

memory. This solution looks to avoid the penalty for replicating processing elements, which provides an acceptable tradeoff. When utilizing wide short buses (HBM being the most common example [2], [3]), designers avoid the penalty of going outside the die for access to memory and recover some of the performance tradeoffs. HBM systems can overcome all DRAM challenges as an enabler of architectures for high-performance and/or low-power computing, while its low speed/pin consumption also improves power efficiency [19], [20](Fig. 1).

This trend is followed by Xilinx, one of the two market giants in the area of FPGAs and the leader in adaptive computing. Xilinx is firmly committed to a transition to HBM memory as a solution to memory bottlenecks, as demonstrated in recent years. In October 2018, Xilinx launched the Alveo U200 with no HBM memory [4]. Only 1 month after, in November 2018 the new Alveo U280 already included 8GB of HBM2 and halved DDR capacity [5]. Their last Alveo Data center card release, the Alveo U55C, doubled HBM capacity and rescinded the DDR memory banks’ use [6].

### B. HBM performance analysis

To shed some light on the intrinsic details of HBM, we developed the Memory Sandbox tool providing higher configurability, more control over measurements, and further insights (i.e. clock cycles of each memory transaction) than the current HBM monitor offered by Xilinx. Our configurable environment is structured in two main pieces: a front-end piece as a user interface for setting up the experiments to be executed, and a back-end piece composed of a set of hardware IPs to run the experiments in the FPGA, according to the data introduced in the front end. Thus, the most relevant IP we have developed is a highly Configurable Pattern Generator, which mimics processor threads data requests with sequential and pseudo-random memory access patterns.

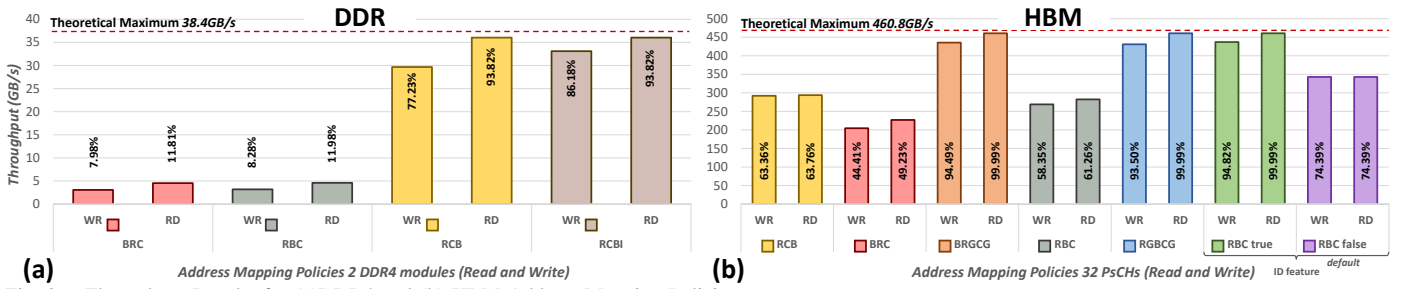


Fig. 2. Throughput Results for (a)DDR4 and (b) HBM Address Mapping Policies.

An initial analysis of typical memory access patterns allows us to implement benchmarks to reveal the subjacent characteristics of HBM and DDR in FPGAs. For this purpose, we emulate the Repetitive Sequential Traversal (RST) a typical sequential access pattern widely used in FPGA programming and sparse accesses with pseudo-random accesses. The first scenarios intend to stress HBM and DDR to measure the actual throughput peak (bandwidth) when using our Memory Sandbox. For this purpose, we perform sequential accesses (RST) in vertically attached pseudo-channels or banks. We enabled outstanding transactions and burst sizes were set to the maximum (16 and 256 beats, respectively). Address mapping policies microbenchmarks results are shown in Fig. 2.

Most modern computer applications require large amounts of memory access. In HBM, as each pseudo-channel has a size of 256MB, multiple pseudochannels will likely be accessed by most applications. From the previous experiments, we know that the performance of a single pseudo-channel is the result of any address mapping policy in Fig. 3 divided by the total amount of pseudochannels (32). Fig. 3 shows the results of accessing different HBM pseudo-channels emulating a single-threaded processing element connected to AXI Port 0. These experiments are performed with a sequential access pattern (RST), a burst size of 16 and RBC true as address mapping policy, which offers the best performance for this type of access pattern according to our experiments. Two main conclusions can be drawn from these experiments:

- Pseudo-channels on the same micro-switch show the same performance regardless of the AXI port accessing them.
- Throughput experiences an average degradation of 50% if the processing element performs memory accesses outside the pseudo-channel to which it is directly connected. This performance loss is the same for the adjacent micro-switch or the furthest one. There is no linear degradation. The performance is either the same for the 4 pseudo-channels within the same micro-switch or 50% in the other 28 pseudo-channels.

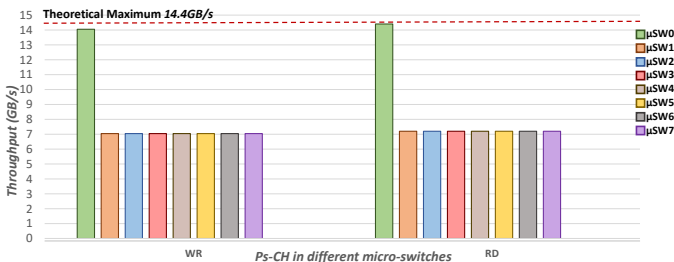


Fig. 3. Throughput Results for HBM accessing different micro-switches.

### C. Conclusion

HBM appears as a solution being integrated into FPGAs to face the memory wall issue and large companies are already committed to its use. As expected, the throughput performance was more than 12 times better when using all 32 pseudo-channels in the HBM in parallel than when using the 2 memory banks in the DDR. The different address mapping policies, the burst size, accesses within a micro-switch or external ones, and the randomization of the address can have a huge impact on the HBM throughput.

### REFERENCES

- [1] Todd Carl Mowry, "Tolerating latency through software-controlled data prefetching," PhD Thesis, Stanford University, Mar. 1994.
- [2] Copyright © 2022 Samsung. All rights reserved. "Next-level performance." Samsung HBM, Tech. Rep., 2022. [Online]. Available: <https://www.samsung.com/semiconductor/dram/hbm/>
- [3] M. Ujaldón, "HPC Accelerators with 3D Memory," *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, pp. 320–328, Aug. 2016, doi: 10.1109/CSE-EUC-DCABES.2016.203.
- [4] Xilinx Inc., "ALVEO™ Product Selection Guide Datasheet," Xilinx Inc., Tech. Rep. XMP451 (v1.7), 2021. [Online]. Available: <https://www.xilinx.com/support/documentation/selection-guides/alveo-product-selection-guide.pdf>
- [5] —, "Xilinx Extends Data Center Leadership with New Alveo U280 HBM2 Accelerator Card," Xilinx Inc., Tech. Rep., Nov. 2018. [Online]. Available: <https://www.xilinx.com/news/press/2018/xilinx-extends-data-center-leadership-with-new-alveo-u280-hbm2-accelerator-card-dell-emc-first-to-qualify-alveo-u200.html>
- [6] Xilinx Inc., "Xilinx Launches Alveo U55C, Its Most Powerful Accelerator Card Ever, Purpose-Built for HPC and Big Data Workloads," Xilinx Inc., Tech. Rep., Nov. 2021. [Online]. Available: <https://www.xilinx.com/news/press/2021/xilinx-launches-alveo-u55c-its-most-powerful-accelerator-card-ever-purpose-built-for-hpc-and-big-data-workloads.html>



**Elias Perdomo** received a B.Eng. degree in Automation Engineering in 2012, and a M.Sc. degree in Digital Systems in 2018 both from the Technological University of Havana, Cuba. In addition, he received a M.Sc. Advanced Microelectronic Systems Engineering from the University of Bristol, UK in 2019. Since 2020, he has been working with FPGA Team of the Barcelona Supercomputing Center (BSC) as well as a PhD student at the Computer Architecture Department of the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. His current research interests include Embedded Systems, RTL and RISC-V SoC design, FPGAs, Heterogeneous Computing, automatic design generation and memory management for HPC and programming models.

# Implications of the Human Oral Microbiome in Alzheimer's Disease Prognosis.

Sara Peregrina<sup>1,2</sup>, Olfat Khannous-Lleiffe<sup>1,2</sup>, Toni Gabaldón<sup>1-3\*</sup>.

<sup>1</sup>Life Science Department, Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain

<sup>2</sup>Mechanisms of Disease, Institute for Research in Biomedicine (IRB), Barcelona, Spain

<sup>3</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[sara.peregrina@irbbarcelona.org](mailto:sara.peregrina@irbbarcelona.org), [toni.gabaldon@bsc.es](mailto:toni.gabaldon@bsc.es)\*

[olfat.khannous@bsc.es](mailto:olfat.khannous@bsc.es)

**Keywords**— Alzheimer disease, oral microbiome, 16S rRNA sequencing, diagnosis, automation, microbiome perturbations, dysbiosis

## EXTENDED ABSTRACT

### 1. Introduction

Alzheimer's disease is the most common type of dementia and it has a multifactorial etiology. There is no cure for this disease, there are only a few medicines to alleviate the symptoms, which will be more optimal if the disease is detected at an early stage [1]. AD can be divided into 4 possible stages, from less to more severity: Mild Cognitive Impairment (MCI), Objective Dementia (OD), Severe Cognitive Decline (SCD) and AD (Alzheimer Disease). Its diagnosis is challenging and there is no single test that determines whether or not a patient has the disease.

However, it is worth mentioning that in patients with AD, chronic inflammation is the first sign of disease progression. Influencing factors in chronic inflammation are changes in the microbiome. Previous studies have shown associations between periodontitis, an infection of the gums due to alteration of the oral microbiome, and A $\beta$ -peptide plaques in the brain. Therefore, key risk factors implicated in the course of Alzheimer's disease include changes in the oral microbiome [2,3].

### 2. Objectives

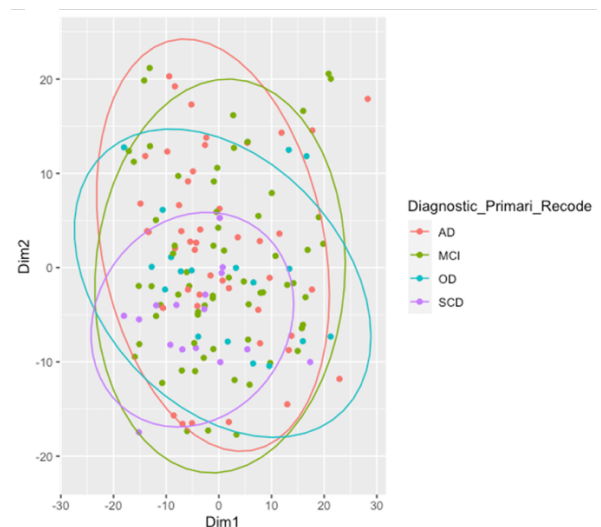
The present project has two general objectives: (2.1) the implementation of computational tools for the analysis of the oral microbiome to elucidate the impact of sample collection methods on the overall microbial composition and to establish links between the microbiome and Alzheimer's disease prognosis and (2.2) the development of automation of a database search by using a REST-API to facilitate the retrieval of previous knowledge to biological interpret the differential analysis findings.

### 3. Results and discussion

The results are based on the analysis of the sequencing of 143 oral microbiome samples from patients with different degrees of AD, of which some clinical variables was available. Then, we quantified the overall microbiome diversity by computing alpha and beta diversity metrics. As regards alpha diversity, we observed hardly any significant differences.

So then, we produced multidimensional scaling plots (MDS) using Aitchison distance between the microbial profiles of samples. No significant effect was obtained, but it can be seen

how the samples of the SCD diagnostic group tend to cluster closer together (Figure 1).



**Fig. 1** MDS plots using Aitchison distance between the microbial profiles of samples (beta diversity).

We next performed a differential analysis to detect taxa at species level with differential abundance according to the diagnostic group. All possible comparisons between the four diagnostic groups were made and we obtained for each comparison a table with the LFC, a measure describing how much a quantity changes, p-value and q-value, which is the adjusted p-value. Among all the comparisons made, species *Prevotella nanceiensis*, *Prevotella denticola* and *Anaeroglobus geminatus* were found to be differentially abundant, without being structural zeros, comparing AD and SCD groups.

The taxa reported as differential abundance are suggesting an association but can not be claimed as biomarkers as we would need a larger dataset and the evaluation of their predictivity (e.g machine learning classifiers). However, by using knowledge databases such as the dbBact database, we can biologically interpret the results by linking them with previous knowledge associating them with specific terms. Almost 60% of the species listed as differentially abundant have already been found associated with the term periodontitis in this database. *Anaeroglobus geminatus*, *Prevotella dentalis* and *[Eubacterium] nodatum* were the differentially abundant taxa with the highest F-score at the term in question. Among them is *Anaeroglobus geminatus*, which is differentially abundant without being a structural zero.

To our knowledge, this is the first study to compare oral microbiome samples from Alzheimer's patients at different stages of disease progression. As it has been seen that the SCD group appears to have a more differentiated sample type, it could be a breakthrough to be able to detect this disease before developing full-blown AD, as it would improve the prognosis and quality of life of patients, as many of the current treatments have better results. So, it is essential to detect it in early stages this disease, as has been observed, in an incipient manner, in this study. However, future studies on a larger scale are necessary to obtain clear results and biomarkers for early diagnosis of the disease, as explained above.

#### 4. Methods

In terms of materials and methods, regarding the part of analysis of oral microbiome samples, samples were sequenced using 16S amplicon variant sequencing. Raw sequencing data was preprocessed by using Dada2 pipeline [4], to obtain an amplicon sequence variant (ASV) table, which records the number of times each exact amplicon sequence variant was observed in each sample. Then, taxonomy was assigned by mapping to SILVA database (v138).

Afterwards, an analysis of microbiome data was carried out using the phyloseq R package [1.38.0 version]. We applied different necessary filters and calculated alpha and beta diversity, applying the appropriate statistical tests in each case. After that, differential abundance analysis was carried out using the ANCOM-BC method (Analysis of Microbiome Composition with Bias Correction).

Finally, with the results obtained in the differential abundance analysis, using the REST-API of dbBact [5], a script was implemented to automate searches in this database. dbBact is a collaborative central repository for bacterial knowledge that when searching for a sequence or species name, the database returns a list of terms with which the search is associated, and each term is assigned an F-score value.

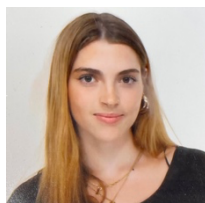
#### Acknowledgments

I would like to first thank Toni Gabaldón, for offering me the opportunity to carry out this project in his Comparative Genomics research group. Also, my thanks and appreciation to my supervisor Olfat Khannous Lleiffe. I really appreciate all her dedication and trust placed in me. I could not be more grateful for the availability shown on her part and for the constant support I have received, which has made it a pleasure to carry out this work in the Comparative Genomics group.

#### References

- [1] Blennow K, Zetterberg H. Biomarkers for Alzheimer's disease: current status and prospects for the future. *J Intern Med.* 2018;284(6):643–63.
- [2] Willis JR, Gabaldón T. The Human Oral Microbiome in Health and Disease: From Sequences to Ecosystems. *Microorganisms* [Internet]. 2020 Feb [cited 2023 Mar 22];8(2).
- [3] Seymour GJ, Ford PJ, Cullinan MP, Leishman S, Yamazaki K. Relationship between periodontal infections and systemic disease. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis.* 2007 Oct;13 Suppl 4:3–10.
- [4] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016. Jul;13(7):581–3.
- [5] dbBact REST API Documentation [Internet]. [cited 2023 Jun 19].

#### Author biography



**Sara Peregrina** was born in La Rioja, Spain, in 2000. She received the degree in Biochemistry and Molecular Biology from the Universitat Rovira i Virgili, Tarragona, Spain, in 2022, and the master's degree in bioinformatics from the Universitat Autònoma de Barcelona, Barcelona, Spain, in 2023. Since September 2023, she has been with the Comparative Genomics Group (led by Dr Toni Gabaldón), BSC-IRB, where she was a Bioinformatics Technique. In March 2024 she was awarded the FI-AGAUR fellowship to start her PhD. Her current research interests include the study of the human microbiome and development of computational tools with biomedical applications.

# Performance analysis of DLR Confined Jet High Pressure Combustor using BSC tools

Josep Pocurull Serra\*, Marta Garcia Gasulla\*

\*Barcelona Supercomputing Center, Barcelona, Spain

E-mail: {josep.pocurull, marta.garcia}@bsc.es

*Keywords—Parallelization, High-Performance Computing, Application Performance Analysis, Load Balancing, Performance Tool*

## I. EXTENDED ABSTRACT

### A. Introduction

In recent years, we have seen supercomputers evolve and reach the exascale. This has been possible thanks to different techniques, such as increasing the number of cores per node, using accelerators, or aggregating more interconnected nodes.

Now, the challenge of using all these resources efficiently relies on the software running on these platforms. CFD (Computational Fluid Dynamics) simulations are among the ones consuming most HPC resources. However, how can we know we are achieving the best performance possible and using all these resources efficiently?

The answer lies in the performance analysis of the codes. Performance analysis can provide detailed information about the application's behavior, showing bottlenecks, performance issues, or non-optimal patterns. Also, a comprehensive performance analysis provides valuable insights into how well the parallelized code scales with increased computational resources and can provide insight and suggest areas for optimization.

In a high-performance computing (HPC) environment, the factors influencing the proper performance of applications multiply due to various physical and logical elements, including network issues, system topology, parallel programming, and building configuration. Therefore, a good performance analysis is crucial for detecting issues and ensuring optimal functionality.

### B. Performance metrics and tools

Starting with the measurements of performance, the EU Centre of Excellence for Performance Optimization and Productivity (POP) project<sup>1</sup> has defined a set of performance metrics [1], offering a quantitative analysis of the factors most relevant to parallelization. These metrics reflect common causes of inefficiency in parallel programs and are calculated as percentages (on a scale from 0 to 100) in a hierarchical representation.

To get the most out of parallel computing, it's essential to understand exactly how these applications work. This is where

the BSC performance tools<sup>2</sup> come into play, providing a robust framework for performance analysis. Two of these tools are Extrae [2] and Paraver [3], which play a central role in the analysis of applications.

Extrae is a lightweight instrumentation library designed to gather detailed insights into the execution of parallel applications. By inserting code annotations into the source code, Extrae enables the collection of valuable information related to communication patterns, computation times, and synchronization events. The traces generated by Extrae include information about the code's state at a specific time, and serve as a foundation for in-depth analysis, helping developers identify areas for improvement within their parallel code.

Complementing Extrae, Paraver is a dedicated performance analysis tool designed for the visualization and analysis of the traces generated by Extrae. With its user-friendly interface, Paraver allows developers and researchers to gain a visual understanding of the behavior of parallel programs. The tool provides graphical representations, including timeline views and histograms, making it easier to identify bottlenecks and communication patterns.

Another tool for improving the efficient use of computational resources is the Dynamic Load Balancing (DLB) library [4]. This library is designed to accelerate hybrid parallel applications and optimize the use of computational resources. In DLB, TALP (Tracking Application Live Performance) [5] is a lightweight, portable, extensible, and scalable tool for online parallel performance measurement. It dynamically collects the POP efficiency metrics of a program at runtime.

Together, these tools provide developers with a clear understanding of their code behavior, making it easier to improve the performance of parallel performance.

### C. Performance analysis

The analysis was conducted on the grand challenge DLR Confined Jet High Pressure Combustor (DLR CJH) [6]. This burner is based on the Recirculation-Stabilized Jet Flame (RSJF) concept, and shows great potential for efficient and flexible combustion, particularly in its application to gas turbines. The analysis was performed on the Hawk cluster of HLRS with three different input cases, varying the number of cells in each one. The inputs have a size of 3M cells, 24M cells and 489M cells.

<sup>1</sup><https://pop-coe.eu>

<sup>2</sup><https://tools.bsc.es>



Running programs with a high number of cores using Extrac generates large trace files, which can make analysis more challenging. Therefore, the analysis begins by obtaining the POP efficiency metrics of the execution of the Grand Challenge with different numbers of nodes using the TALP tool from the DLB library. This provides some of the performance metrics without generating a trace, and it is useful to understand at what point the efficiency drops and it is therefore interesting to obtain a trace for detailed analysis. When analyzing the results, a similar trend is observed across all three inputs, indicating a correlation between the *Communication Efficiency* and the number of cells per core of each input size. Based on this observation, it is advisable to analyze the traces of the 3M case to facilitate trace manipulation. The smaller the mesh size, the smaller the trace size and the less time required for analysis.

The analysis then focuses on the 3M input traces, first obtaining an overview of the entire execution with different numbers of processes (128, 256 and 1024), to understand the behavior, the different phases, and the key aspects of the application's performance. The next step is to select a Focus of Analysis (FOA) that contains a specific segment of the application with a representative behavior, typically involving one or a few phases in iterative processes. Once the FOA is selected, in this case an iteration of the DLR CJH case, the POP efficiency metrics of the FOA are obtained and the analysis focuses on the inefficient metrics and the performance problems that the region presents. Figure I shows a table with the POP efficiency metrics for a single iteration of the DLR CJH case, where it can be seen that the case presents a problem with the *Load Balance*, the *Communication Efficiency* and the *Instruction Scalability*.

TABLE I. POP EFFICIENCY METRICS FOR ONE ITERATION OF THE DLR CJH CASE.

Number of processes	128	256	1024
Global efficiency	85.92	103.91	79.84
Parallel efficiency	85.92	78.98	56.25
Load balance	93.15	92.37	86.28
Communication efficiency	92.25	85.50	65.20
Serialization efficiency	94.62	91.00	77.99
Transfer efficiency	97.50	93.96	83.60
Computation scalability	100.00	131.57	141.93
IPC scalability	100.00	150.94	214.32
Instruction scalability	100.00	95.26	76.12
Frequency scalability	100.00	91.51	87.00

The analysis then focuses on the problematic metrics to identify their root cause. In the analysis, the Paraver tool is used to manipulate the case traces generated by Extrac.

#### D. Results

The POP performance metrics pointed to four main issues limiting the scalability: Load balancing, instruction scalability,

serialization and transfer efficiency. We identified the computational load assigned to each MPI rank as the source of the load imbalance. The performance analysis tools allowed us to identify the key functions that had a negative impact on the instruction scalability, guiding the developers in which direction to look to improve the performance of the code. We also found that the serialization problem was caused by some system noise, and that the noise disappeared when disabling the Transparent Huge Pages (THP). This change showed a 10% increase in the global efficiency. Finally, we found that the transfer efficiency was negatively affected by the bandwidth.

#### E. Conclusion

With this analysis, we show how the combination of the different tools presented allows us to perform a grand challenge's performance analysis. The POP performance metrics pointed to the main issues limiting the scalability, guiding the analysis to areas for further investigation. This serves as an example of the importance of considering all performance metrics when analyzing an inefficient application, as the specific cause of inefficiency may not be immediately apparent.

#### REFERENCES

- [1] M. Wagner *et al.*, "A structured approach to performance analysis," in *International Workshop on Parallel Tools for High Performance Computing*. Springer, 2017, pp. 1–15.
- [2] H. Servat *et al.*, "Framework for a productive performance optimization," *Parallel Computing*, vol. 39, no. 8, pp. 336–353, 2013.
- [3] V. Pillet *et al.*, "Paraver: A tool to visualize and analyze parallel code," in *Proceedings of WoTUG-18: transputer and occam developments*, vol. 44, 1995, pp. 17–31.
- [4] M. Garcia *et al.*, "Hints to improve automatic load balancing with LeWI for hybrid applications," *Journal of Parallel and Distributed Computing*, vol. 74, no. 9, pp. 2781–2794, 2014.
- [5] V. Lopez *et al.*, "TALP: A lightweight tool to unveil parallel efficiency of large-scale executions," in *Proceedings of the 2021 on Performance EngineerRing, Modelling, Analysis, and VisualizatiOn STrategy*, 2021, pp. 3–10.
- [6] S. Lesnik and H. Rusche, "DLR CJH combustor," <https://develop.openfoam.com/committees/hpc/-/tree/develop/combustion/XiFoam/DLRCJH>, 2022–2023.



**Josep Pocurull Serra** received his BSc degree in Computer Engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, in 2021. In the same year, he joined the Best Practices for Performance and Programmability (BePPP) group at the Barcelona Supercomputing Center (BSC), where he currently works as a performance analyst for parallel applications.

# Exploring the biophysical boundaries of protein families with deep learning methods

Miriam Poley-Gil \*†, R. Gonzalo Parra \*, Alfonso Valencia \*‡

\*Barcelona Supercomputing Center, Barcelona, Spain

†Universitat Politècnica de Catalunya, Barcelona, Spain

‡ICREA, Barcelona, Spain

E-mail: [mpoley@bsc.es](mailto:mpoley@bsc.es)

**Keywords - Protein Design, Deep Learning, Reverse Folding, Biophysical Characterisation**

## EXTENDED ABSTRACT

The intricate interplay between protein sequence, structure and function remains one of the central questions in Molecular Biology. But recently, Machine and Deep Learning Models have revolutionised the field allowing us to explore the protein space faster. To understand what they are capturing and generating we have combined the use of state-of-the-art protein models for inverse folding (such as ProstT5 and ProteinMPNN) and for sequence generation (such as ProtGPT2 and ZymCTRL) with biophysical analyses.

### A. Introduction

The concept of energetic frustration comes from the fact that proteins are not only optimised for folding and stability: they are also evolutionarily selected to function. This would explain that 10-15% of residue-residue native interactions in proteins are in energetic conflicts with their local structure and that these conflicts are preserved over evolutionary and physiological time scales [1]. Our novel tool called FrustraEvo measures the conservation of local energetic frustration within and between protein families [2]. Once the frustration state of each residue is calculated, we can map frustration from structures to sequences, and similarly from the structures of several evolutionarily related proteins to a multiple sequence alignment (MSA). Then we can measure how conserved each frustration state is within every position in the MSA, and check if relevant residues are energetically conserved over protein families. In this study, we design protein sequences using deep learning-based models and evaluate them using FrustraEvo to explore the biophysical limits of protein sequence and structure spaces of known natural protein families. We have studied local frustration

conservation patterns in Globins,  $\beta$ -lactamases and also RAS subfamily to shed light on the evolutionary processes leading to the diversification of proteins. Our server is fully available at <https://frustraevo.qb.fcen.uba.ar/>.

### B. Method

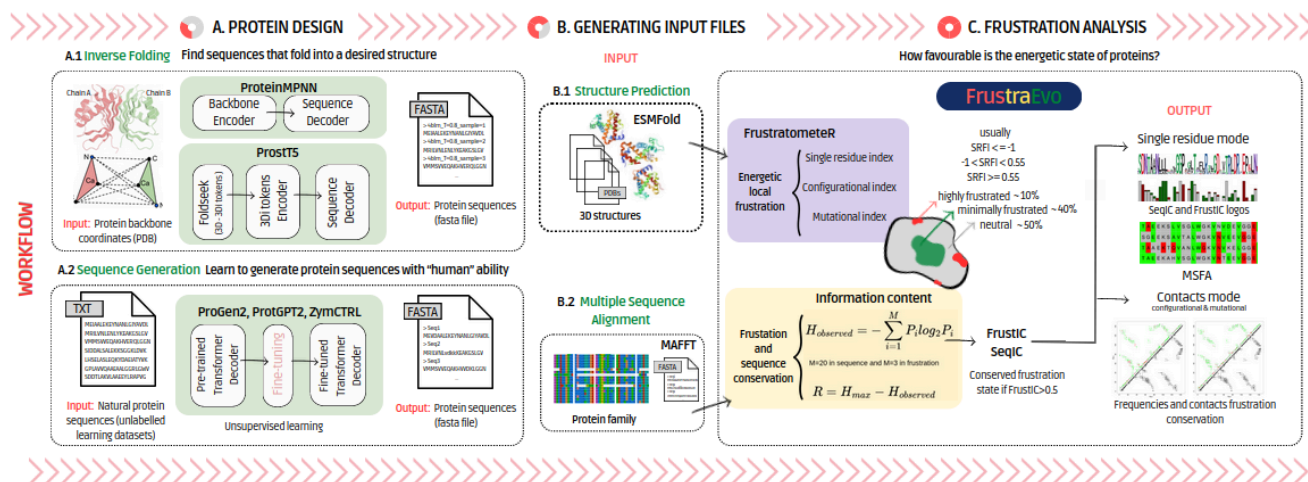
We have designed protein sequences using four different models: ProstT5 [3] and ProteinMPNN [4] for reverse folding, ProtGPT2 [5] and ZymCTRL [6] for sequence generation. To select the best designs from each model and protein family we used ESMFold to predict their structures and evaluate their pLDDT scores (predicted per-residue local distance difference test), and ProstT5 sequence embeddings, to select sequences based on how they cluster in the low dimensional protein space around their natural counterparts. The final subset of selected designs is analysed through FrustraEvo. Figure 1 shows a detailed workflow.

### C. Results and discussion

We found that most of the highly frustrated native residues are related to functional aspects. These functional residues are mostly recovered by sequence generation models, which recover almost completely the native energetic signature suggesting that there are alternative ways to design proteins instead of the one explored by evolution. In the case of catalytic sites, they are also recovered by inverse folding models. We therefore point out a selective memory concerning functionality, highly influenced by the original training of the models, where they could have learnt co-evolutionary statistics. This may involve a primary level of memory (local). However, ProteinMPNN, despite this selective memory, seems to minimise most of the highly frustrated positions but surprisingly generated diverse yet stable proteins and retained essential catalytic sites and amino acid identities. Moreover, it also recovers the main

network of frustrated contacts of the functional domains even suggesting a tertiary level of memory (contacts). While evolutionary aspects remain yet unexplored extensively, our approach promises to

effectively shed light into the intricacies of protein family boundaries and aid to explore design options for understanding protein evolution, including their implications in associated diseases.



**Figure 1. Workflow of the project.** A, B and C represent the main and common tasks applied to all protein families in the study. ProGen2 designs were discarded due to their low quality. ZymCTRL is only an enzyme model and has so far only been applied to lactamases because of its similarity to ProtGPT2.

## D. Future directions

Ongoing works intend to implement this approach across all enzymes that have documented catalytic sites as well as relevant non-globular proteins. Another valuable future approach would be to retrain these models, in order to evaluate this selective memory in combination with energetic frustration as a potential tool for biophysical characterisation of proteins.

## ACKNOWLEDGMENT

The authors are grateful for the support of the Department of Research and Universities of the Generalitat de Catalunya to the CompBio + NLP Research Group (Code: 2021 SGR 01627). They also want to thank the valuable role of Noelia Ferruz, Maria Ines Freiberger and Michael Heinzinger.

## REFERENCES

- [1] Ferreiro, D. U., Hegler, J. A., Komives, E. A., & Wolynes, P. G. (2007). Localising frustration in native proteins and protein assemblies. *Proceedings of the National Academy of Sciences*, 104(50), 19819-19824.
- [2] Freiberger, M. I., Ruiz-Serra, V., Pontes, C., Romero-Durana, M., Galaz-Davison, P., Ramírez-Sarmiento, C. A., ... & Valencia, A. (2023). Local energetic frustration conservation in protein families and superfamilies. *Nature Communications*, 14(1), 8379.
- [3] Heinzinger, M., Weissenow, K., Sanchez, J. G.,

Henkel, A., Steinegger, M., & Rost, B. (2023). ProST5: Bilingual language model for protein sequence and structure. *bioRxiv*, 2023-07.

[4] Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... & Baker, D. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615), 49-56.

[5] Ferruz, N., Schmidt, S., & Höcker, B. (2022). ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1), 4348.

[6] Munsamy, G., Lindner, S., Lorenz, P., & Ferruz, N. (2022, December). Zymctrl: a conditional language model for the controllable generation of artificial enzymes. In *NeurIPS Machine Learning in Structural Biology Workshop*.

## AUTHOR BIOGRAPHY



**Miriam Poley Gil** was born in Cadiz, Spain, in 1999. She received the BSc in Biotechnology from the Universidad de Cadiz in September 2021. She started her MSc in Bioinformatics from the Universitat Autònoma de Barcelona in September 2022, after a couple of research internships. She joined Alfonso Valencia's Computational Biology group as a master's student in March 2023 and is currently pursuing her PhD.

# Machine Learning approaches for the characterization of COPD

Iria Pose-Lagoa <sup>\*†</sup>, Alfonso Valencia <sup>\*‡</sup>

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

<sup>‡</sup>ICREA, Barcelona, Spain

E-mail: iria.poseлагоa@bsc.es

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

## I. EXTENDED ABSTRACT

Chronic Obstructive Pulmonary Disease (COPD) is a complex, heterogeneous, highly prevalent, and yet underdiagnosed disease with poor outcomes due to the difficulties of an early diagnosis. This study aims to enhance the binary patient classification of COPD using gene expression data from the Lung Tissue Research Consortium. To achieve this, we employ various feature selection criteria, including intrinsic data characteristics (data-driven), an external information source (curated COPD-related genes), and their respective biological expansions to identify the most relevant genes. Subsequently, we evaluate the performance of different classifiers: Random Forest (RF), Support Vector Machines - polynomial and radial kernel -(SVM-poly, SVM-rad), k-Nearest Neighbors (kNN), Generalized Linear Models (GLM), and XGBoost (XGB). Our results show that the data-driven and curated COPD-related expansion gene selection approaches yield the highest cross-validation and independent test data performances, respectively.

### A. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a prevalent chronic disorder characterized by airflow limitation, primarily caused by smoking, ranking as the third leading cause of death globally. It is classified into four severity subgroups that take into account the post-bronchodilator ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC), FEV1/FVC (mild  $\geq 80\%$ , moderate 50 – 80%, severe 30 – 49%, very severe  $< 30\%$ ). COPD is a heterogeneous condition comprising a wide range of nonidentical patient profiles. Its diagnosis is not straightforward, usually appearing with severe airflow obstruction profiles, leading to a need for improved strategies to identify individuals who are at greater risk of developing COPD or who have early-stage citechoi2020diagnosis.

Artificial Intelligence (AI) techniques, notably Machine Learning (ML) and Deep Learning (DL), offer promising avenues for understanding the complexity of COPD and improving diagnostic accuracy. Several studies have applied ML algorithms and penalized regression models for the analysis of COPD and the detection of possible candidate therapeutic genes, yielding molecular biomarker subgroups that have little overlap among them [1][2][3].

Here, we use gene expression data and apply several filtering selection approaches to improve the prediction of COPD and to identify informative biomarker genes and biological processes involved in the disease. Our ML techniques demonstrate their ability to accurately classify COPD patients, outperforming previous studies [2] [4] [5] with accuracies up to 84,8%, and the selected genes represent relevant biomarkers for disease prediction.

### B. Methods

Feature selection is a key step in microarray data analysis to classify new samples accurately, and some studies reveal its potential in identifying effective classification genes. Therefore, we employed various filtering approaches to reduce the number of input features and to identify the most informative genes. Firstly, we selected genes from intrinsic data characteristics (data-driven) combining the results of a Differential Expression Analysis with the minimum Relevance Maximum Redundance (mRMR) algorithm output. To complement our gene selection criteria based on analyzing expression data, we incorporated genes from other sources of information. Specifically, we extracted DisGeNET COPD information (COPD-related). Nonetheless, the overlap among these two collections was minimal, and only MMP1 was in the intersection, growing up to 22 genes when using the entire COPD-related list. Since genes work in collaboration, we investigated the potential of incorporating additional genes that are not significantly different in their expression values and have not been directly associated with COPD. Particularly, we expanded the two previous lists of genes (data-driven and curated COPD-related) with physical interactions, first neighbors, based on prior knowledge genes. As an alternative approach to using only interaction contacts, we also expanded the seed lists of genes using network-based prioritization algorithms with GUILDify. By comparing the algorithms' performance with all the feature subsets, we aimed to identify the most informative list of genes for COPD classification.

### C. Results and discussion

As the overlap among the initial list of genes (DEA, mRMR, and curated COPD-related) was small, we tried to confirm the associations of these lists of genes with COPD by conducting a literature search in PubMed DataBank. The

results reveal that curated COPD-related genes have the highest number of confirmed associations having more than 900 associations. Moreover, from the 163 data-driven genes, 45 were previously identified as being related to the occurrence and progression of COPD. To better understand the biological significance of these selected genes, we conducted a functional enrichment analysis over the DEA, curated COPD-related, and mRMR sets. We observe that curated COPD-related genes allow us to recuperate pathways not enriched in DEA and mRMR genes. Furthermore, some of these pathways have been previously linked to COPD (immune system, disease, extracellular matrix organization, and signal).

The results of ML models show that the genes prioritized by GUILDify may provide valuable information for the characterization of COPD phenotypes. Actually, these particular selections return a competitive number of genes (only 1% of the no-seed genes are chosen) able to compete with the other input gene sets, overcoming the normMCC and accuracy performance values in some cases.

In summary, our selection criteria propose the 163 data-driven genes as the ones that capture most of the relevant information for COPD prediction (highest cross-validation normMCC values). Indeed, the selected genes, such as MMP1, COMP, POU2AF1, CD19, CYP1B1, or ROR1 have previously been linked to the development and progression of the disease, supporting the reliability of our analysis.

Furthermore, the differentially expressed genes were enriched in immune system pathways, which play a key role in COPD. Inflammation is a hallmark of COPD, and cells of both the innate (activated with smoking exposure) and the adaptive immune system participate in the inflammatory response in COPD. Additionally, our cross-validation results demonstrate that some of our models outperform the metrics of previous studies using microarray gene expression data. Specifically, the kNN using data-driven genes as input achieves higher accuracy (0.85), specificity (0.82), sensitivity (0.88), and AUC (0.92) values. Moreover, various of our models outperform methods that use clinical variables as predictive factors.

#### D. Future directions

We want to explain and validate our model to understand the disease's crucial genes and biological processes. Moreover, as COPD is a very heterogeneous disease, considering unsupervised algorithms that could detect specific molecular COPD subtypes, would be a valuable future approach.

## II. ACKNOWLEDGMENT

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

## REFERENCES

- [1] P. J. Castaldi, M. Benet, H. Petersen, N. Rafaels, J. Finigan, M. Paoletti, H. M. Boezen, J. M. Vonk, R. Bowler, M. Pistolesi *et al.*, "Do copd subtypes really exist? copd heterogeneity and clustering in 10 independent cohorts," *Thorax*, vol. 72, no. 11, pp. 998–1006, 2017.
- [2] P. 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] Y. Yao, Y. Gu, M. Yang, D. Cao, and F. Wu, "The gene expression biomarkers for chronic obstructive pulmonary disease and interstitial lung disease," *Frontiers in genetics*, vol. 10, p. 1154, 2019.
- [4] K. R. Mahmudah, B. Purnama, F. Indriani, and K. Satou, "Machine learning algorithms for predicting chronic obstructive pulmonary disease from gene expression data with class imbalance," in *BIOINFORMATICS*, pp. 148–153.
- [5] M. C. Matheson, G. Bowatte, J. L. Perret, A. J. Lowe, C. V. Senaratna, G. L. Hall, N. de Klerk, L. A. Keogh, C. F. McDonald, N. T. Waidyatillake *et al.*, "Prediction models for the development of copd: a systematic review," *International journal of chronic obstructive pulmonary disease*, pp. 1927–1935, 2018.



**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. In September 2022, she joined Alfonso Valencia's Computational Biology group as a master's student and is currently pursuing her PhD.

# A Benchmark of Synthetic Transcriptomic Cancer Data Reconstruction

Guillermo Prol-Castelo<sup>\*†</sup>, Davide Cirillo<sup>\*</sup>, Alfonso Valencia<sup>\*‡</sup>

<sup>\*</sup>Barcelona Supercomputing Center (BSC), Barcelona, Spain

<sup>†</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain

<sup>‡</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

E-mail: {guillermo.prolcastelo, davide.cirillo}@bsc.es

**Keywords**—Cancer, Transcriptomics, Synthetic Patients, Deep Learning, Autoencoders.

## I. EXTENDED ABSTRACT

Cancer is the second most common cause of death worldwide, and its incidence is increasing [1]. Some methodologies have been developed to study cancer. For instance, PAM50, a collection of 50 genes important for cancer characterization, has helped categorize cancer subtypes [2]. However, the rapid growth of sequenced biological data, or omics, has made acquiring much larger amounts of genes possible. Still, the number of samples available in studies tends to be low. This combination of small sample size and high dimensionality, known as the curse of dimensionality, renders significant data analyses less efficient.

Hence, there are limitations to deep learning implementations on omics data generally and cancer data in particular [3]. In the former, the curse of dimensionality has hindered the application of deep learning, given its data-hungry nature. In the latter, our current understanding of the impact of molecular mechanisms of cancer progression challenges our interpretation of the application of deep learning algorithms to omics information [4]. In order to circumvent both of these issues, we aim to learn a low-dimensional representation of the real data, use this representation to augment the original data with improved fidelity in reconstruction and obtain meaningful insights on cancer progression along the way.

The Auto Encoder (AE) [5] is a deep learning technique that reduces data dimensionality. In this study, we define and use three types of Auto Encoders: *vanilla* Auto Encoder [5], Variational Autoencoder (VAE) [6], and Conditional Variational Autoencoder (CVAE) [7]. We discuss how we can learn from real cancer data, such as that provided by The Cancer Genome Atlas (TCGA), reconstruct the original data, and generate new data *in silico*, i.e., synthetic data.

### A. Cancer Data

We obtain a breast invasive carcinoma (BRCA) dataset from TCGA, preprocess it to remove low-variable and outlier genes, yielding a dataset with 900 female patients and 8,954 genes. Besides, we have clinical information available; specifically, we are interested in the patient stage at the time of data collection. Four stages (I, II, III, or IV) categorize the tumor’s invasive progression.

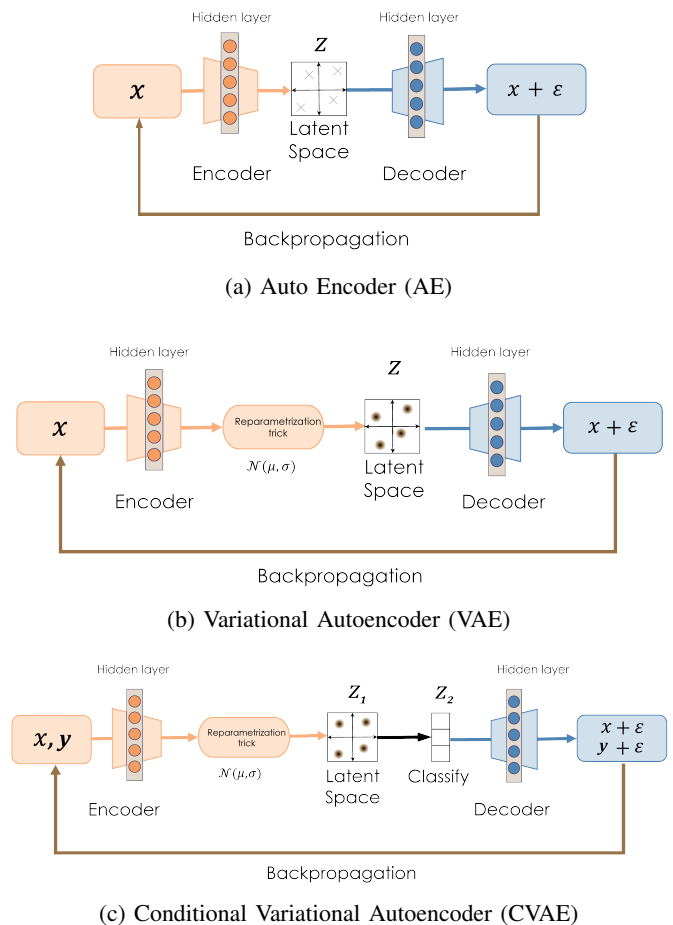
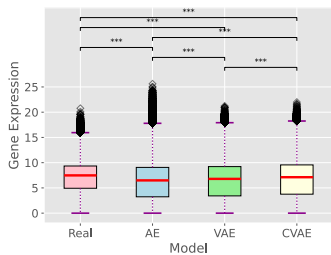
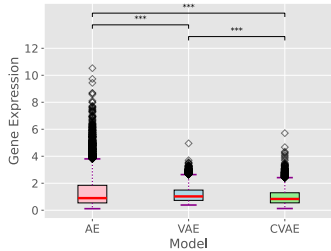


Fig. 1: Architectures of three different Autoencoders: (a) AE, (b) VAE, and (c) CVAE.  $x$  represents the input transcriptomics data,  $z$  is the latent space embedding of the input  $x$ , and  $y$  represents the class variable or the patients’ stages. The reparametrization trick in (b) and (c) yields a normalized distribution  $N(\mu, \sigma)$ , where  $\mu$  is the mean and  $\sigma$  the standard deviation of the distribution, per latent space dimension.  $\epsilon$  is the difference between the real and reconstructed data.



(a) Gene expression distributions



(b) Wasserstein distance distributions

Fig. 2: Reconstruction of BRCA data. (a) Distribution of gene expression and (b) respective Wasserstein distance distribution between real and reconstructed gene expression profiles. Red horizontal lines represent median of distributions. The Wasserstein distances medians in (b) for AE, VAE, and CVAE are 0.89, 1.02, and 0.83. The x-axis corresponds to each of the independent models used in our experiments. Triple asterisks show statistically significant differences in distributions.

### B. Auto Encoders

AEs comprise three main elements: the encoder, the latent space, and the decoder, as seen in Fig. 1. The encoder embeds a representation of the input data  $x$ —our cancer data—into a smaller subspace, the latent space, which the decoder morphs back into the original dataset.

Due to the reduced dimensionality of the latent space, AEs make classification tasks easier and faster and learn relevant characteristics from the generated subspace. Depending on the objective, a specific variation may be applied:

- Auto Encoder (AE): used when the task is reconstructing and denoising the input data. The latent space is composed of coordinates representing the original data in its respective low dimensions.
- Variational Autoencoder (VAE): relies on a Bayesian probabilistic generative model, producing a probability distribution landscape as the latent space instead of simple coordinates.
- Conditional Variational Autoencoder (CVAE): semi-supervised VAE with an additional input variable  $y$  and an additional latent layer where we classify  $y$  with the latent space variables. In our case,  $y$  are the different stages of cancer.

### C. Results

Fig. 2a compares the real and reconstructed gene expression for all patients and genes, and Fig. 2b shows the

corresponding Wasserstein distance distribution between each model’s reconstructed gene expression and the real data. The AE plot shows the most wide reconstruction profile shape. Due to its non-stochastic nature, the AE can learn extreme values which may not be representative, showing a larger range of values. The CVAE shows a shorter range of values, but its Wasserstein distance distribution is wider than the VAE, a fact that may be due to a lack of separability between the cancer stages used for classification. The VAE is an optimal intermediate that keeps a similar mean expression across all patients except those with more extreme gene expressions. This behavior does not just guarantee acceptable reconstruction but also allows for the generation of new synthetic patients while keeping a similar expression to the real cases—i.e., we are generating realistic synthetic patients.

### D. Conclusion

We show that it is possible to reconstruct transcriptomic cancer data with minimal loss while keeping enough stochasticity to allow the generation of synthetic patients that are similar, but not identical, to their real counterparts.

## 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*, vol. 100, no. 3, pp. 469–481, Jun. 2020, publisher: Elsevier.
- [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, publisher: Elsevier.
- [3] Y. Wang, Q. Chen, H. Shao, R. Zhang, and H. Shen, “Generating bulk RNA-Seq gene expression data based on generative deep learning models and utilizing it for data augmentation,” *Computers in Biology and Medicine*, vol. 169, p. 107828, Feb. 2024.
- [4] J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi, “Don’t Blame the ELBO! A Linear VAE Perspective on Posterior Collapse,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [5] H. Bourlard and Y. Kamp, “Auto-association by multilayer perceptrons and singular value decomposition,” *Biological Cybernetics*, vol. 59, no. 4-5, pp. 291–294, Sep. 1988.
- [6] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” 2013, publisher: arXiv Version Number: 11.
- [7] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, “Semi-Supervised Learning with Deep Generative Models,” Oct. 2014, arXiv:1406.5298 [cs, stat].



**Guillermo Prol-Castelo** received his BSc degree in Engineering Physics from Universitat Politècnica de Catalunya in 2019. He completed his MSc degree in Multidisciplinary Research in Experimental Sciences from Universitat Pompeu Fabra in 2021. Since 2022, he has been part of the Machine Learning for Biomedical Research Unit of the Barcelona Supercomputing Center, and a Ph.D. student at Universitat Pompeu Fabra.

# Drug Repurposing for Mammalian Heart Regeneration: Study of the Inhibition Mechanism of Neomycin and Paromomycin on Meis1-Hoxb13-DNA Trimer.

Ignasi Puch-Giner<sup>\*†</sup>, Victor Guallar<sup>\*‡</sup>

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

<sup>‡</sup>Nostrum Biodiscovery, Barcelona, Spain

E-mail: ignasi.puchginer@bsc.es, victor.guallar@bsc.es

**Keywords - Drug design, cardiomyocyte regeneration**

## I. EXTENDED ABSTRACT

The newborn mammalian heart has a remarkable inherent regenerative capacity mediated by the replication of pre-existing specific heart cells called cardiomyocytes. However, this regenerative ability diminishes shortly after birth[1], leaving the adult mammalian heart with limited capacity for self-repair following injury[2, 3].

Recent studies have seen that transcriptional factors Meis1 and Hoxb13 act cooperatively to induce postnatal cell cycle arrest[4]. These same studies suggest that pharmacological targeting of Meis1 and Hoxb13 transcriptional activity could be a viable strategy for heart regeneration.

Our collaborators performed structure-based drug repurposing screening to identify FDA-approved drugs that can inhibit Meis1 and Hoxb13 transcriptional activity based on the published crystal structure of the Meis1 and Hoxb13 DNA binding domains [5] as seen in FIG. 1. They found that Paromomycin, and Neomycin, could do it in a dose-dependent manner, inducing a significant proliferation of neonatal cardiomyocytes in vitro and in vivo.

Our work has been to understand why this happens, and what is the inhibition mechanism of these two drugs by which the transcription activity (represented in FIG. 1) is halted. We formulated three objectives: (1) identify binding energies involved in the different protagonists in the trimer; (2) characterize the binding mode of both ligands; and (3) pinpoint the process by which these ligands block the formation of the Meis1-Hoxb13-DNA trimer.

### A. Results

1) *Binding Energies:* To understand the binding mode of the ligands to the proteins, and check the energy values associated, we began by performing PELE[7] simulations for all the protein-ligand combinations.

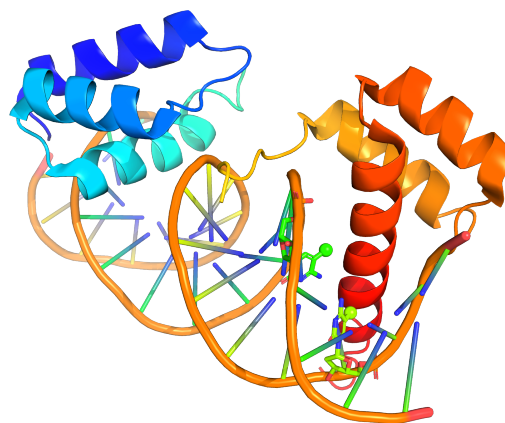


Fig. 1. Original crystal structure extracted from PDB 5EGO of the DNA methylation by the transcription factors Meis1 (blues) and Hoxb13 (reds). Methylated nucleotides have been represented with balls and sticks.

We have done a global exploration per ligand and system. The first relevant result obtained is that Neomycin reaches the lowest binding energies with both monomers when compared to Paromomycin, as seen in FIG. 2. Moreover, out of Meis1 and Hoxb13, the lowest binding energies are reached with Meis1.

2) *Binding mode characterization:* Seeing the energetic results from the previous section, we focused on the binding mode characterization, especially with Meis1 and Neomycin since they gave the most signal of interaction.

We checked the best binding energy poses of all the simulations for all the protein-ligand combinations with PELE and searched for interactions between them. Then we proceeded to perform four MD simulations.

We did 1.5  $\mu$ s simulations with the GROMACS[6] software. We performed two simulations per protein, either Meis1 or Hoxb13. One of the two simulations began with the best-scoring isomer of the PELE simulation located at the predicted binding region.



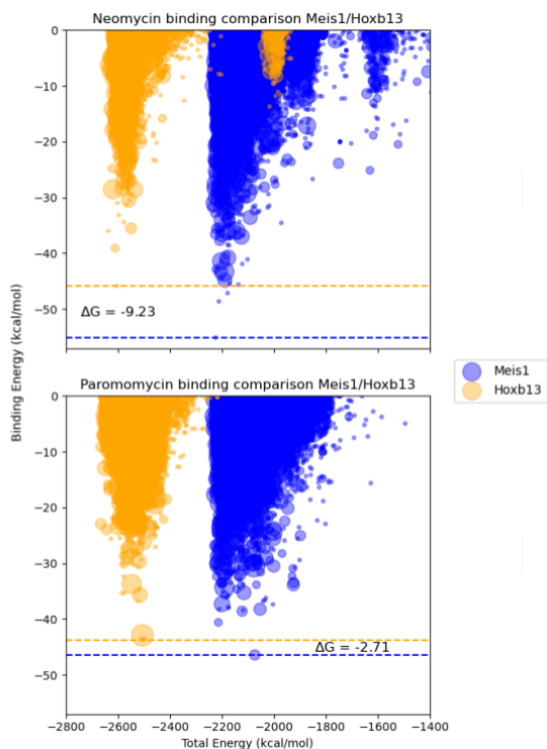


Fig. 2. Binding energy representation versus the total energy of the system for all the combinations of the ligands (Neomycin and Paromomycin) and the proteins (Meis1 and Hoxb13). The dot size represents the time spent in the conformation represented by that dot.  $\Delta G$  indicates the difference between the lowest values of binding energy between proteins.

The other simulation began with the same ligand but at a distant place from the predicted binding zone.

What we have seen is a twofold result. The first one is that the interactions are quite strong since, beginning from the best PELE pose, the ligand stays in the area for 1.2  $\mu s$  as seen in FIG. 3. The other major result is that there is a consensus between the PELE and the GROMACS prediction in three key residues highlighted in FIG. 3 with the lighter blue. These would seem to play an important role in the Neomycin binding.

3) *Further work:* We want to assess our collaborators' hypothesis: Neomycin and Paromomycin favour the formation of Meis1 dimer into a non-natural position that prevents the monomers from binding to DNA. PELE protein-protein and pyDock simulations are being used to understand the natural dimer formation. Afterwards, a study of the ligands' effect on the dimer formation will be carried out.

#### REFERENCES

- [1] Soonpaa, M. H., and LOREN J. Field. "Assessment of cardiomyocyte DNA synthesis in normal and injured adult mouse hearts." *American Journal of Physiology-Heart and Circulatory Physiology* 272.1 (1997): H220-H226.
- [2] Senyo, Samuel E., et al. "Mammalian heart renewal by pre-existing cardiomyocytes." *Nature* 493.7432 (2013): 433-436.
- [3] Bergmann, Olaf, et al. "Evidence for cardiomyocyte renewal in humans." *Science* 324.5923 (2009): 98-102.

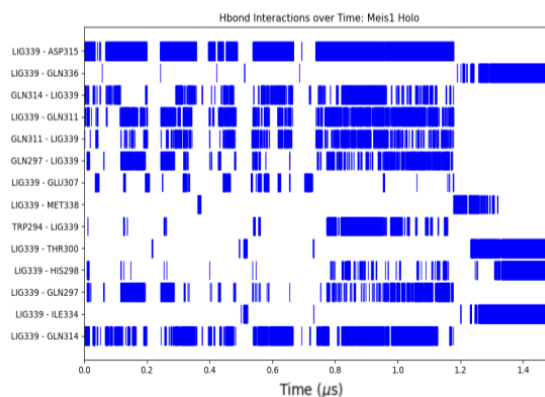
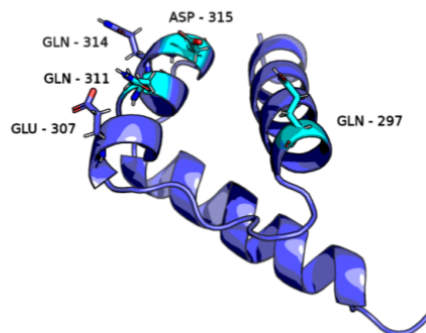


Fig. 3. In the lower figure, we see a plot of the H bond interactions between the ligand and the Meis1 that appear in at least 5% of all the simulation. In the upper figure, we see the Meis1 protein with some MD-relevant residues represented with balls and sticks, and with a lighter blue colour we can see PELE-relevant residues.

- [4] Nguyen, Ngoc Uyen Nhi, et al. "A calcineurin–Hoxb13 axis regulates growth mode of mammalian cardiomyocytes." *Nature* 582.7811 (2020): 271-276.
- [5] Yin, Yimeng, et al. "Impact of cytosine methylation on DNA binding specificities of human transcription factors." *Science* 356.6337 (2017): eaaj2239.
- [6] Van Der Spoel, David, et al. "GROMACS: fast, flexible, and free." *Journal of computational chemistry* 26.16 (2005): 1701-1718.
- [7] Borrelli, Kenneth W., et al. "PELE: protein energy landscape exploration. A novel Monte Carlo based technique." *Journal of chemical theory and computation* 1.6 (2005): 1304-1311.

#### REFERENCES



Ignasi Puch-Giner received his BSc degree in Physics at the Universitat de Barcelona (UB). He holds a MSc on Computational Modeling for Physics, Chemistry and Biochemistry from UB and Universitat Politècnica de Catalunya (UPC). Since 2021 he has been in the Electronic and Atomic Protein Modeling group led by Victor Guallar in the Life Sciences department at the Barcelona Supercomputing Center. He is currently enrolled at the UPC's Computational and Applied Physics PhD programme.

# Hybridisation in Emerging Fungal Pathogens

Álvaro Redondo-Río<sup>1,2,\*</sup>, Toni Gabaldón<sup>1,2,3,4,#</sup>

<sup>1</sup>Barcelona Supercomputing Centre (BSC-CNS), Life Sciences, Carrer de Jordi Girona 29-31, 08034 Barcelona, Spain

<sup>2</sup>Institut de Recerca Biomèdica de Barcelona (IRB-Barcelona), Carrer de Baldori Reixac 10, 08028 Barcelona, Spain

<sup>3</sup>Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

<sup>4</sup>Centro de Investigación Biomédica En Red de Enfermedades Infecciosas (CIBERINFEC), Spain

\*aredond1@bsc.es, #toni.gabaldon@bsc.es

**Keywords** — Hybridisation, Fungi, Genomics, LOH

## EXTENDED ABSTRACT

### A. Introduction

Hybridisation, the interbreeding between different species or populations, plays a significant role in species evolution by influencing genetic diversity, adaptation, and speciation. This process occurs when individuals from distinct lineages mate, resulting in offspring with a combination of traits from both parental populations. When the genomic characteristics of a given species allow it, hybridization can also occur by allopolyploidization, as often occurs with plants and fungi [1].

The genetic stress derived from these genomic rearrangements cause an accelerated evolution and adaptation to compensate and stabilise the hybrid genome. In some cases, this stabilisation leads to the exchange of advantageous genetic traits, facilitating adaptation to new environments or ecological niches. This phenomenon, known as introgression or loss of heterozygosity (LOH), can enhance the genetic diversity of populations and promote their resilience to environmental changes.

### B. Objectives and scope

Hybridization has been widely observed in fungal pathogens, from filamentous species such as *Aspergillus fumigatus* [2] to yeasts such as *Candida orthopsilosis* [3]. The prevalence of these hybrids in the clinical setting makes us pose the hypothesis that hybridization events can result in an increased fitness for infection, probably because of the combination genes encoding virulence factors or traits related to host specificity from both parentals. As a consequence, hybrid fungal pathogens may exhibit enhanced pathogenicity, broader host ranges, or increased resistance to antifungal agents. Understanding the role of hybridization in fungal pathogen evolution is essential for developing effective strategies for disease management and mitigating the impact of emerging fungal threats not only on human health, but also in agriculture and natural ecosystems.

By studying hybridization across a broad spectrum of fungal species, we expect to gain valuable insights into the mechanisms underlying pathogen evolution and adaptation. A comprehensive understanding of LOH patterns, genetic exchanges, and the ecological factors driving hybridization events can provide a foundation for predicting the emergence of novel fungal pathogens and identifying potential hotspots for disease outbreaks. Furthermore, comparative analyses of hybrid fungal lineages can reveal common genetic signatures associated with the mechanisms that allow these species to tolerate the genomic stress that is generated after a hybridisation event. This knowledge can also contribute to understanding some broader questions related to reticular evolution.

### C. Materials and methods

Analysing hybridization using genomics and inspecting loss of heterozygosity (LOH) events involves a multifaceted approach that integrates high-throughput sequencing techniques, bioinformatics tools, and population genetics analyses. Our study will be based on publicly available data of whole-genome sequencing experiments representing different fungal clades suspected of hybridization. By measuring the ploidy and heterozygosity levels of these species, we aim to identify regions of the genome where genetic variation is inherited from divergent parental lineages, indicative of hybrid ancestry. Additionally, the detection of LOH events, where heterozygous alleles are lost, can serve as a signature of genome stabilisation and can inform about the processes required for the successful generation of a new hybrid species. We will rely on bioinformatics pipelines developed in our group tailored for detecting LOH events [4]. This tool utilises read-depth and mutation frequency information to pinpoint genomic regions that have suffered LOH.

### D. Preliminary results

So far, our analyses point to hybridization as a widespread process in fungi, with a great number of independent hybridisation events spread across multiple clades. We have been able to identify hybrids that have undergone extensive LOH, showing that they have been long established as independent species and that they have evolved to stabilise their genome after the hybridization event. Polyploidization events have also been detected, showing the great flexibility of fungal genomes to tolerate the multiple incompatibilities and stresses that arise from these genomic rearrangements.

## References

- [1] Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., & Schumer, M. (2021). The genomic consequences of hybridization. *ELife*, 10. <https://doi.org/10.7554/ELIFE.69016>
- [2] Steenwyk, J. L., Lind, A. L., Ries, L. N. A., *et al.* (2020). Pathogenic Allodiploid Hybrids of *Aspergillus Fungi*. *Current Biology*: CB, 30(13), 2495-2507.e7. <https://doi.org/10.1016/J.CUB.2020.04.071>
- [3] Mixão, V., & Gabaldón, T. (2018). Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast* (Chichester, England), 35(1), 5–20. <https://doi.org/10.1002/YEA.3242>
- [4] Schiavinato, M., del Olmo, V., Muya, V. N., & Gabaldón, T. (2023). JLOH: Inferring loss of heterozygosity blocks from sequencing data. *Computational and Structural Biotechnology Journal*, 21, 5738–5750. <https://doi.org/10.1016/J.CSBJ.2023.11.003>

## *Author biography*



**Álvaro Redondo-Río** was born in Córdoba, Andalusia, Spain, in 1988. He received the B.Sc. degree in biochemistry from the University of Córdoba, in 2020, and the M.Sc. degree in bioinformatics and computational biology from the Autonomous University of Madrid (UAM), Spain, in 2022. Since February 2023, he has worked with the Computational Genomics group, led by Toni Gabaldón, first as a bioinformatic technician and now as a PhD student. He is currently studying genomic hybridisation events in fungal pathogens, although his previous work mainly revolves around microbiome analyses.

# Agile and accurate microarchitecture modeling using Python and Salabim.

Carlos Rojas Morales\*<sup>†</sup>, Adrian Cristal\*, Osman Unsal\*

\*Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {carlos.rojas, adrian.cristal, osman.unsal}@bsc.es

**Keywords**—*Discrete-event simulator, Microarchitecture, Pipeline, Out-of-Order execution, Simulation Tracers.*

## I. EXTENDED ABSTRACT

This work describes a straightforward methodology to simulate full-custom pipelined microarchitectures using a discrete-event simulation framework. We implement an out-of-order pipeline using the object-oriented approach to have parametrizable models. The simulator can generate performance counting of any traceable event and also includes a tracer to visualize the execution behavior of each micro-operation in the pipeline.

### A. Introduction

Designing and evaluating hardware always has the constraint of the necessary effort to obtain meaningful results. It is not feasible to fabricate every idea into silicon and evaluate them in real-time conditions. In this regard, computer architects rely on translating each idea into models with different levels of abstraction, usually evaluated with simulations. The closer the model is to the actual physical implementation and conditions, the more costly it is to iterate the design space exploration.

Nowadays, we have very robust tools for modeling in different abstraction levels, including trace-driven simulators, discrete-event simulators, cycle-accurate simulators, and Register Transfer Level simulators [1]. However, to start a design exploration of a new semantic and paradigm of computer architecture, starting with a complete system simulator, add a massive overhead of complexity that may not be necessary. On the other hand, starting modeling with a High-Level Language (HLL) from scratch can be difficult due to a lack of control flow mechanisms to implement a proposed design's different microarchitectural models.

Currently, programming can be very accessible due to the last developments in HLLs and frameworks. We propose using Salabim [2] simulator framework to evaluate different microarchitectural implementations. Salabim is a package for discrete event simulations in Python. It provides many objects and methods to handle resource dependencies and states. We propose to use Salabim as a High-Level abstraction simulation tool. Aside from Salabim, we have access to the Python language environment to implement diverse functionalities to integrate into our simulator. Like trace events, performance counting schemes, and system emulation functionality for the parts we are not interested in modeling at the microarchitectural level. This tool is suitable for educational purposes

to teach the basics of computer architecture in a completely interactive way. Also, we want to keep improving it to use this simulation methodology as a research tool to model new paradigms in the computer architecture field.

### B. Design flow

In order to implement a model in Salabim, we start with some basic microarchitecture specifications, Figure 1. These specifications include pipeline stages, load-store scheduling, arithmetic scheduling, etcetera. Then, from these specifications, we should describe a model of the states of each instruction across the pipeline. We use the model to implement a Salabim concurrent process that models the states of each instruction depending on the available resources and dependencies, Figure 2.

We use the Object Oriented programming paradigm to implement the required resources from the instruction. A resource is a built-in function from Salabim that is in charge of controlling the instructions flow. This way, we can obtain accurate traces of the dependencies and stall events.

We use a Fetch-engine process to spawn each instruction process. If the front-end resources are unavailable, the Fetch process stalls until the resources are free.

Finally, we use Python structures such as dictionaries and lists to model the memory caches, reorder buffers, and other functional blocks.

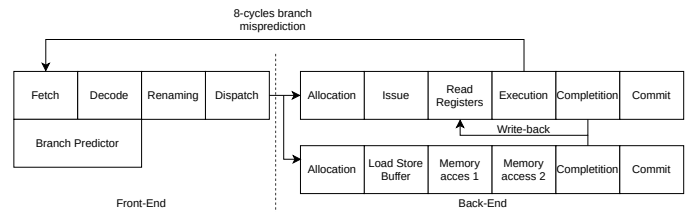


Fig. 1. Out of order superscalar pipeline specification.

### C. Evaluations

We have provided the simulator with different performance counters and parametrized implementation of the functional blocks, resources, and processes. We can do a very efficient design space exploration using different parameters for the microarchitecture model.

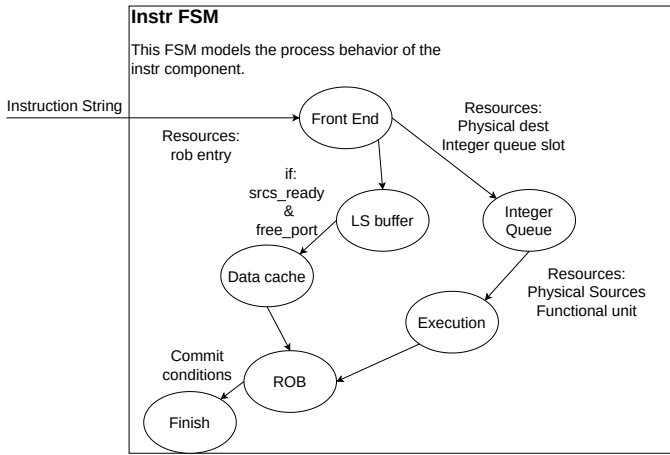


Fig. 2. Intruction states across the pipeline.

Another essential feature is the capability to trace events of the changing states in each instruction. We use these traces to verify the correctness of the program execution by comparing them with a golden reference.

Finally, we have provided our simulator with a Konata format [3] tracer generator. We use the Konata visualizer to show the execution of the instructions concurrently and how they interact in the pipeline stages. Figure 3 shows a dual issue Out-of-order pipeline execution, and Figure 3 shows an eight-issue Out-of-order trace.

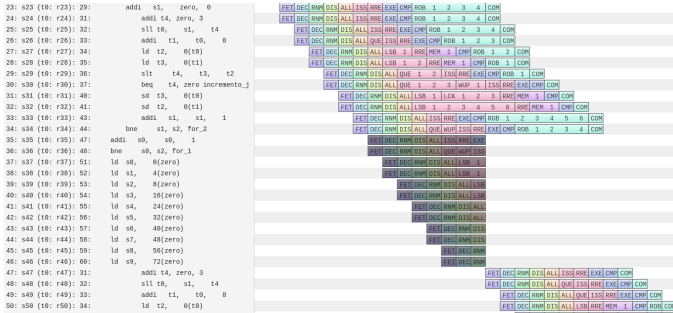


Fig. 3. Konata trace for a dual-issue width.

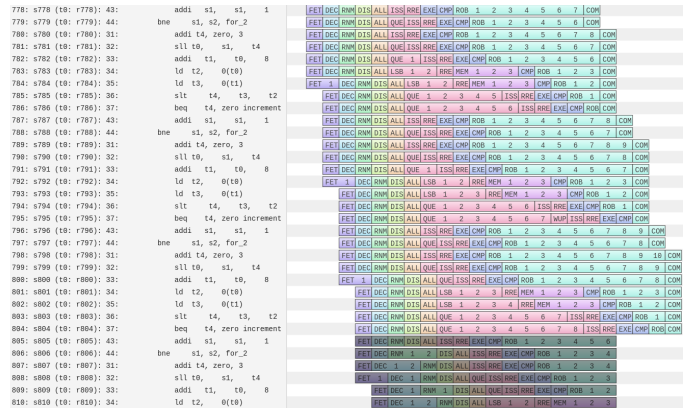


Fig. 4. Konata trace for an eight-issue width.

REFERENCES

- [1] O. Chatzopoulos *et al.*, “Towards accurate performance modeling of risc-v designs,” *arXiv preprint arXiv:2106.09991*, 2021.
- [2] R. van der Ham, “salabim: discrete event simulation and animation in python,” *Journal of Open Source Software*, vol. 3, no. 27, p. 767, 2018.
- [3] R. Shioyadan, “Visualizing the out-of-order CPU model,” GitHub Wiki, 2018. [Online]. Available: <https://raw.githubusercontent.com/wiki/shioyadan/Konata/gem5-konata.pdf>



**Carlos Rojas Morales** received his BSc degree in electronic engineering from the Instituto Politécnico Nacional (IPN), Mexico, in 2017. Then, he started a double MSc degree in Computer Engineering from IPN and Innovation and Research in Informatics from Universitat Politècnica de Catalunya (UPC), Spain, which he finished in 2020. Since 2020, he has been with the Synthesis and Physical Design of ICs group of Barcelona Supercomputing Center (BSC) and as a PhD student at the Department of Computer Architecture of UPC.

# Cooling Effect of Aerosols on Past Arctic Climate (1950-2014)

Alba Santos-Espeso\*, Pablo Ortega\*, María Gonçalves Ageitos\*†

\*Barcelona Supercomputing Center, Barcelona, Spain

†Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {alba.santos, pablo.ortega, maria.goncalves}@bsc.es

**Keywords**—Arctic, Aerosols, NTCFs, Arctic Amplification

## I. EXTENDED ABSTRACT

In a warming climate due to the increasing abundance of greenhouse gases, it is important to acknowledge and study other potentially competing climate drivers, such as Near Term Climate Forcers (NTCFs). These are atmospheric species, such as aerosols and tropospheric ozone, with atmospheric lifetimes shorter than two decades. Special attention must be paid to their effects on vulnerable systems such as the Arctic region, particularly affected by the global temperature rise due to a climate feature known as Arctic amplification [1].

The objective of this study is to isolate the effects of NTCFs using Earth System Models (ESMs). These are advanced computational tools that simulate the evolution in various components of the climate system (e.g. atmosphere, ocean, land surface, sea ice) and how they interact with each other. By comparing two historical climate simulations—one representing past emissions (*historical*) and the other constraining the emissions of NTCFs to pre-industrial levels (*hist-piNTCF*)—we can determine the effects of these species on Arctic climate.

### A. Aerosols effects on the Arctic

Our analyses indicate that among NTCFs, aerosols dominate the response of Arctic climate. These ubiquitous species directly alter the energy balance by scattering or absorbing solar radiation, and indirectly acting as cloud condensation nuclei (CCN). When aerosols contribute to the formation of clouds they change their properties and modify local precipitation. At the same time, clouds interact with radiation. Generally, clouds reflect solar radiation having a cooling effect in the system. However, in the case of the Arctic, a region characterised by high albedo and several months of scarce solar insolation, low-level clouds produce a greenhouse warming effect by trapping the radiation from the surface [2].

Another indirect impact of aerosols is the modification of Earth's heat distribution by altering atmospheric circulation and ocean heat transport [3]. All in all, we identify several competing mechanisms through which aerosols can affect Arctic climate, thus providing considerable motivation to further investigation.

### B. Results

Through the analysis of the mean surface air temperature (*tas*) of a multi-model ensemble (Fig. 1), we find that for the

## tas ensemble mean [1950-2014]

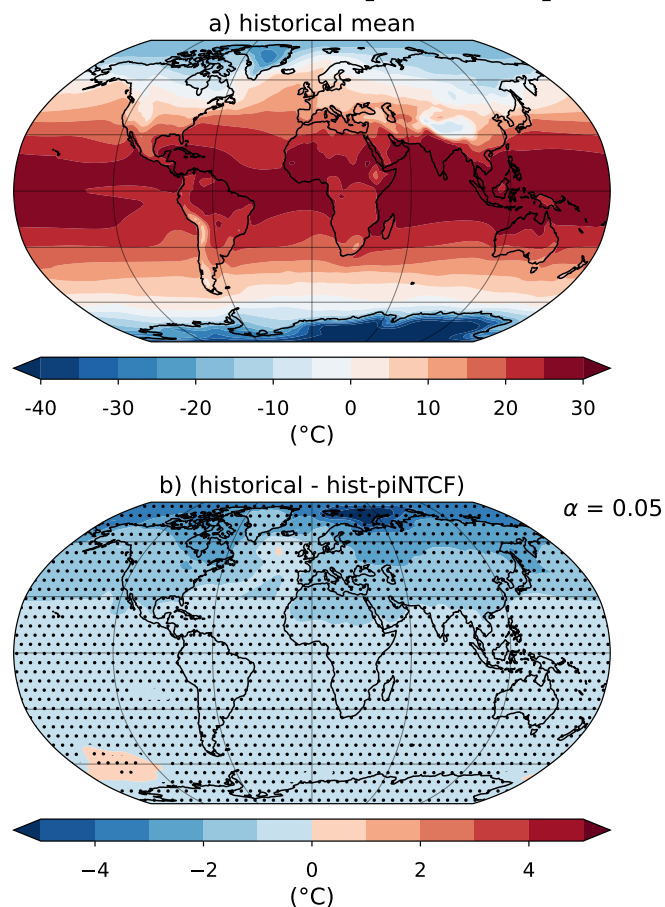


Fig. 1. Annual mean analysis of surface air temperature (*tas*) over the period 1950-2014. (a) Climatology for the *historical* experiment and (b) mean difference between *historical* and *hist-piNTCF*. The ensemble analysed is comprised of 4 models (BCC-ESM1, MRI-ESM2-0, UKESM1-0-LL and EC-Earth3-AerChem), with 3 simulations per model. Significant values in (b) are determined by a paired sample t-test with a 95% confidence and shaded with hatches.

period 1950-2014, the historical presence of NTCFs causes a global cooling, with an amplified signal in the Arctic (Fig. 1b). The most pronounced cooling is located over the Barents Sea, an area that experiences important seasonal sea ice variations. We attribute this cooling to the direct radiative effect of aerosols since they are the only NTCFs considered in our study with a negative radiative forcing.

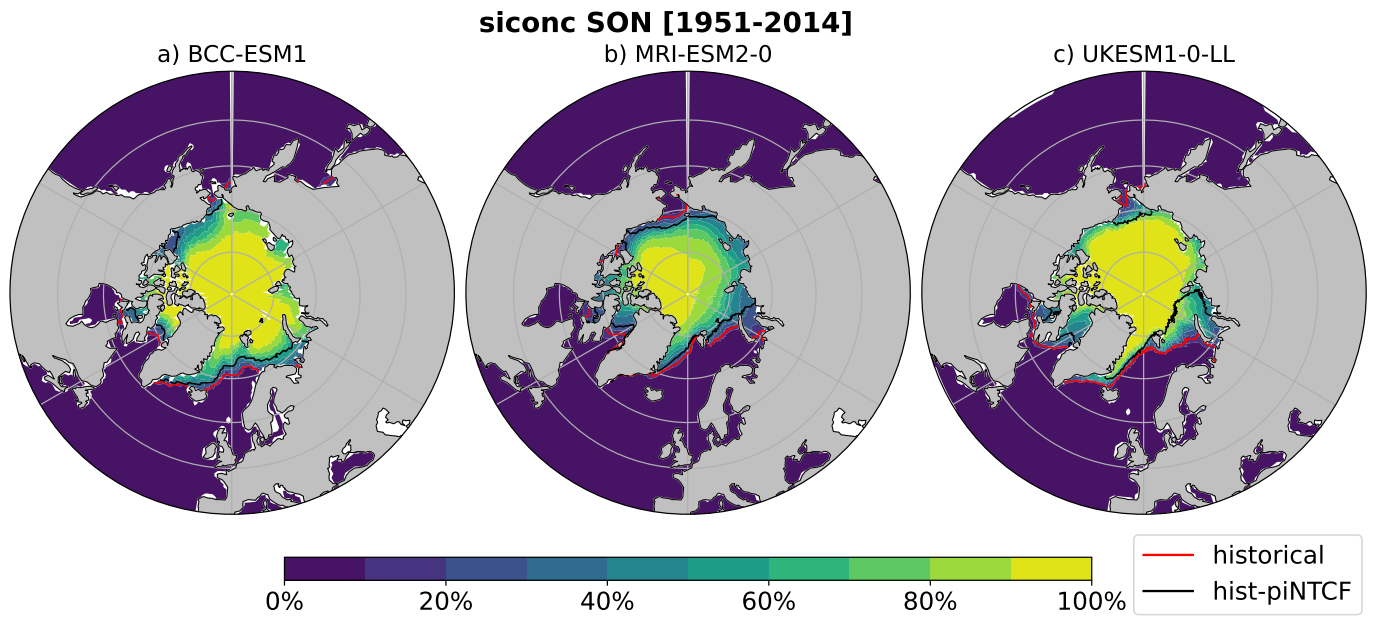


Fig. 2. Mean sea ice concentration (*siconc*) in boreal autumn (September, October and November) during the period 1951-2014. The colors represent the *historical siconc* and the contours the sea ice edge (*siconc*=15%) for the experiments *historical* (red) and *hist-piNTCF* (black). For each experiment and model we consider the mean of 3 simulations but for BCC-ESM1 *hist-piNTCF* data, only available for 1 simulation.

Lower temperatures in the Arctic region impact sea ice and vice versa. It is known that temperature changes in the Arctic are intensified through a sea ice feedback [4]. This positive feedback starts in the months of boreal summer when solar radiation is maximum in the Arctic. In the case of a temperature increase and a corresponding sea ice retreat, the Arctic surface albedo decreases (the ocean surface is darker and absorbs more radiation than the ice). During this period the Arctic ocean absorbs more energy that is later released into the atmosphere in the following seasons of autumn and winter, hence amplifying the initial Arctic warming signal. Analogously, a temperature decrease will also be amplified by the sea ice increase.

In order to assess the impact of aerosols in this feedback, we analyse the sea ice concentration (*siconc*) difference between experiments in boreal autumn (Fig. 2). Although the affected regions vary slightly between the models, they consistently show significant sea ice increase in the historical presence of NTCFs (*historical*). In essence, all models display an Arctic sea ice feedback which contributes to the intensification of the Arctic cooling signal.

### C. Conclusion

In conclusion, our study identifies aerosols as key drivers of Arctic cooling, yet the presence of competing mechanisms underscore the complexity of Arctic climate dynamics. In line with this, we observe a positive feedback: the decrease in temperature leads to increased sea ice cover, which, in turn, further cools the system due to its higher reflectivity.

Overall, we aim to improve our understanding of the complexities of Arctic climate and NTCFs, while also evaluating and improving ESMs as key tools in climate research. As we continue to deepen our knowledge, we will be better equipped to confront the coming challenges posed by climate change.

## II. ACKNOWLEDGMENT

The research leading to these results has received funding from the EU within the HE Framework Programme under grant agreement n°101056783.

## REFERENCES

- [1] M. Previdi *et al.*, “Arctic amplification of climate change: a review of underlying mechanisms,” *Environmental Research Letters*, vol. 16, no. 9, p. 093003, Sep. 2021, publisher: IOP Publishing.
- [2] J. Schmale *et al.*, “Aerosols in current and future Arctic climate,” *Nature Climate Change*, vol. 11, no. 2, pp. 95–105, Feb. 2021, publisher: Nature Publishing Group.
- [3] S. Krishnan *et al.*, “The Roles of the Atmosphere and Ocean in Driving Arctic Warming Due to European Aerosol Reductions,” *Geophysical Research Letters*, vol. 47, no. 7, p. e2019GL086681, 2020.
- [4] J. C. Acosta Navarro *et al.*, “Amplification of Arctic warming by past air pollution reductions in Europe,” *Nature Geoscience*, vol. 9, no. 4, pp. 277–281, Apr. 2016, publisher: Nature Publishing Group.



**Alba Santos-Espeso** is a PhD student at the Barcelona Supercomputing Center (BSC), within the Earth Sciences Department, and Universitat Politècnica de Catalunya (UPC). She received her BSc degree in Physics from the Universidad Autónoma de Madrid (UAM), Spain, in 2021. The following year, she completed her MSc degree in Meteorology and Geophysics from the Universidad Complutense de Madrid (UCM), Spain. In the BSC, her work focuses on how atmospheric composition affects climate variability.

# The METASAT Hardware Platform v1.1: Identifying the Challenges for its RISC-V CPU and GPU Update

Marc Solé i Bonet<sup>\*†</sup>, Aridane Álvarez Suárez<sup>‡</sup>, Leonidas Kosmidis<sup>\*†</sup>

<sup>\*</sup>Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

<sup>‡</sup> fentISS, Valencia, Spain

E-mail: marc.solebonet@bsc.es, aalvarez@fentiss.com, leonidas.kosmidis@bsc.es

**Keywords**—RISC-V, GPU, Space computing, NOEL-V, GPU, Vortex, Embedded SoC, Critical systems, FPGA

## I. EXTENDED ABSTRACT

The escalating complexity of on-board applications requires enhanced performance in space systems. Consequently, the industry is transitioning towards advanced hardware and software with increased capabilities. However, managing the complexity of critical systems within this evolving landscape poses significant challenges.

In response to these challenges, model-based approaches are being introduced in the design of space systems. The Horizon Europe project METASAT [1][2][3], funded by the European Commission, aims to develop model-based design approaches to effectively address the complexity associated with programming advanced high-performance platforms, including AI accelerators and GPUs.

Currently under development, the METASAT prototype platform will play a crucial role as a low Technology Readiness Level (TRL 3-4) demonstrator of a on-board computer of unprecedented complexity. This platform mirrors a sophisticated space system which can be used to showcase the development and validation of model-based design methods targeting multicores, GPUs and partitioned hypervisor based systems, facilitating the management of complexity in future space systems.

The METASAT hardware platform, prototyped on an Field-Programmable Gate Array (FPGA) and relying on the RISC-V Open standard Instruction Set Architecture (ISA), is designed such that it can be a candidate for qualification and use in future institutional missions, which cannot rely on high performance COTS technologies. In particular, the platform is targeted to be a mixed-criticality platform, allowing the deployment of software of different criticality on the same hardware. To this end, it takes advantage of virtualisation, using the Xtratum XNG Hypervisor from Fentiss [4], which is METASAT project partner.

### A. Multi-Core Challenges

The METASAT platform CPU multi-core is a quad-core system based on the NOEL-V processor. The NOEL-V is a RISC-V space grade processor [5], developed by Frontgrade Gaisler, which is available under a commercial license or a more restricted GPL license. For the METASAT project, the GPL release is used as it provides enough of the characteristics

required for the project needs. In particular, while the full Level-2 cache (L2) is not provided in the open sourced release, a lite design is available with limited characteristics. The first version of the METASAT platform [3] was based on Grlib's latest available release at the project start (build 4280, 2022.4 GPL release).

However, while porting the hypervisor to it, we noticed an issue with the L2-lite controller. When executing Symmetric Multiprocessing (SMP) code under linux, random exceptions appeared, likely due to errors in the cache coherence. Since the L2 cache-lite version cannot be disabled by software, a design without second-level cache was generated to verify that the issue was in the cache system.

A different error surfaced when porting the Xtratum hypervisor to the METASAT platform. As a RISC-V conforming architecture, NOEL-V supports some of the RISC-V extensions, such as the H-extension which provides full-virtualization features. This extension incorporates a set of virtual CSRs for the Guest Operating System (OS).

In a correct execution the expected behaviour would be:

- The Guest OS is initialized with the Software Interrupt enabled and no pending interrupts in Supervisor Cause (*scause*) and Supervisor Interrupt Pending (*sip*) registers.
- The hypervisor sets the software interrupt in the Hypervisor Virtual-Interrupt-Pending (*hvip*) register by setting the *VSSIP* bit.
- The Guest OS jumps to the interrupt vector and checks in *scause*, (the *SSIP* bit is set in *sip*). The exception code corresponding to Supervisor Software Interrupt is 1.
- Then the interrupt is cleared and the execution continues.

However, in the current platform the *scause* register is set to 2, which corresponds to a RESERVED exception. Then, since the cause is not identified to the Supervisor Software Interrupt, the interrupt is not handled and the execution is stuck in a loop.

After contacting Gaisler, they confirmed this behaviour to be a known bug of METASAT's NOEL-V build, suggesting to upgrade to the latest release, 2023.4 GPL - build 4288.

Despite the considerable evolution of the NOEL-V code within a year of releases, the port of the SPARROW AI



accelerator [6] it's been easy thanks to its lean and modular design. This has been very important, since updating the NOEL-V code was crucial, in order to fix critical bugs for the project success.

### B. GPU Update

The METASAT prototype was integrated with the first release of Vortex [7][8], a RISC-V soft-GPU. The design is completely open source and it targets PCI FPGA cards (i.e. Intel Arria 10 and Xilinx Alveo) to work as an accelerator in a desktop environment. For that reason the vanilla Vortex driver takes advantage of the proprietary software libraries of their FPGA vendors to communicate with the host CPU.

In the METASAT platform, however, this approach cannot be followed as the soft RISC-V CPU is also implemented within the FPGA as an embedded SoC. For the first platform implementation [3], a memory-mapped AXI-lite subordinate was configured to control the GPU. Then, Vortex, accessed the memory bank through AXI4 when in execution. In this initial design, both the GPU and the CPU were accessing the same Dynamic Random Access Memory (DRAM) unit. In order to manage the data, it was determined that the lower addresses were in CPU address space while from address 0x60000000 to 0x80000000 corresponded to GPU address space.

This approach had obvious limitations. In particular, both units were competing for memory access through an AXI-interconnect. Furthermore, in a critical-environment device a memory overflow from the CPU would violate the restrictions on memory and could access GPU-only addresses. For this reason, it was decided to move the GPU to access an different DRAM bank. Meanwhile, a new release of Vortex was published, which introduced bug-fixes and broader support. Since the design was not finalised, it was then decided to upgrade both the connection to memory and the Vortex version.

The new implementation of the GPU connection with the CPU is again done through an AXI-lite interface. However, the GPU controller has more features to mimic the default design of Vortex. This includes polling for information on the device configuration and status, writing the Device Configuration Registers (DCR), writing and reading from the GPU memory and starting the execution.

While many of this features are quality of service improvements, the memory access feature is the approach taken for transferring the data from the CPU to the GPU and vice-versa. While the initial implementation for this mechanism follows a naive approach, the CPU sends to the GPU controller the address and data to write/read, it is simple enough to have a functional design which is required for the METASAT project. In later implementations this task will be delegated to a Direct Memory Access (DMA) unit to handle the data transfers in the background.

As a result of the upgrade, the METASAT hardware platform has more than doubled its effective memory and has a simplified memory access mechanism, less prone to errors and more suitable for a critical-system.

### C. Conclusions

The METASAT hardware platform provides a representative high-performance prototype for on-board processing eval-

uation of model-based applications. During the development process some critical errors have been encountered which forced the upgrade of the system to more recent releases. At the same time, the requirement to change part of the design, motivated the update of the GPU to provide a more robust implementation. While some additional work is required, upgrading the METASAT platform has been proven to be a simple procedure which can be used to fix critical bugs. At the end of the METASAT project the full platform will be released open source[9].

### ACKNOWLEDGEMENTS

This work was supported by the European Community's Horizon Europe programme under the METASAT project (grant agreement 101082622). In addition, it was partially supported by the Spanish Ministry of Economy and Competitiveness under grants PID2019-107255GB-C21 and IJC-2020-045931-I ( Spanish State Research Agency / Agencia Española de Investigación (AEI) / <http://dx.doi.org/10.13039/501100011033> ) and by the Department of Research and Universities of the Government of Catalonia with a grant to the CAOS Research Group (Code: 2021 SGR 00637).

### REFERENCES

- [1] BSC, IKERLAN, OHB, fentISS, ALES, "METASAT: Modular Model-based Design and Testing for Applications in Satellites," <https://metasat-project.eu/>.
- [2] L. Kosmidis et al, "METASAT: Modular Model-Based Design and Testing for Applications in Satellites," in *Embedded Computer Systems: Architectures, Modeling, and Simulation - 22nd International Conference (SAMOS)*, ser. Lecture Notes in Computer Science, 2023.
- [3] L. Kosmidis, M. Solé, I. Rodríguez, J. Wolf, and M. M. Trompouki, "The METASAT Hardware Platform: A High-Performance Multicore, AI SIMD and GPU RISC-V Platform for On-board Processing," in *2023 European Data Handling & Data Processing Conference (EDHPC)*. IEEE, 2023, pp. 1–6.
- [4] fentISS, "XtratuM hypervisor," <https://www.fentiss.com/xtratum/>.
- [5] F. Gaisler, "NOEL-V Processor," <https://www.gaisler.com/index.php/products/processors/noel-v>.
- [6] M. S. Bonet and L. Kosmidis, "SPARROW: A Low-Cost Hardware/Software Co-designed SIMD Microarchitecture for AI Operations in Space Processors," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2022.
- [7] G. Tech, "Vortex GPU," <https://vortex.cc.gatech.edu/>.
- [8] B. Tine, K. P. Yalamarthy, F. Elsabbagh, and K. Hyesoon, "Vortex: Extending the RISC-V ISA for GPGPU and 3D-Graphics," in *International Symposium on Microarchitecture (MICRO)*, 2021.
- [9] M. Project, "METASAT Public repository," <https://gitlab.bsc.es/metasat-public/>.



**Marc Solé i Bonet** received his BSc degree in Computer Engineering from Universitat Politècnica de Catalunya (UPC), in 2019. The following year, he started a MSc in Innovation and Research in Informatics at the UPC. In 2021, he joined the Computer Architecture and Operating Systems (CAOS) group of Barcelona Supercomputing Center (BSC) where he has been involved in the GPU4S and METASAT projects. In 2023 he started a PhD at the department of computer architecture of UPC.

# Single-cell atlas of the aging immune system

Maria Sopena-Rios<sup>1\*</sup>, Aida Ripoll-Cladellas<sup>1\*</sup>, Marta Melé<sup>1</sup>

*1. Life Sciences Department, Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain*

maria.sopena@bsc.es, aida.ripoll@bsc.es, marta.mele@bsc.es

**Keywords**— Aging, Immune system, Single-cell transcriptomics

EXTENDED ABSTRACT

## A. Introduction

Age-associated decline in immune function, known as immunosenescence, predisposes individuals to infection, autoimmune disorders, and cancer<sup>1</sup>. Immune function decline manifests as chronic low-grade inflammation (inflammaging)<sup>2</sup> and impaired responsiveness to stimuli<sup>3</sup>. Single-cell RNA sequencing (scRNA-seq) is a powerful tool to uncover the cellular and molecular dynamics of immunosenescence among immune cell populations<sup>4</sup>.

However, studying immune cell type dynamics and cell state changes during human aging requires extremely large sample sizes. Here, we leverage a scRNA-seq dataset of 982 individuals encompassing over 1 million human peripheral blood mononuclear cells (PBMCs)<sup>5</sup> to systematically investigate the effect of aging on the human circulating immune system.

## B. Materials and Methods

### Data generation

scRNA-seq data was generated previously in<sup>5</sup>, and sequenced using 10x Genomics sequencing. The data was pre-processed and quantified using Cell Ranger<sup>6</sup>, followed by cell type annotation performed with Azimuth<sup>7</sup>.

### Differential expression analysis

Single-cell gene expression was collapsed per cell type and donor to generate pseudobulk estimates for performing differential expression analysis (DE). We employed dreamlet<sup>8</sup>, which utilizes a linear mixed model to obtain transcriptional expression changes with age while correcting for sex and batch effects.

### Cellular compositional analysis

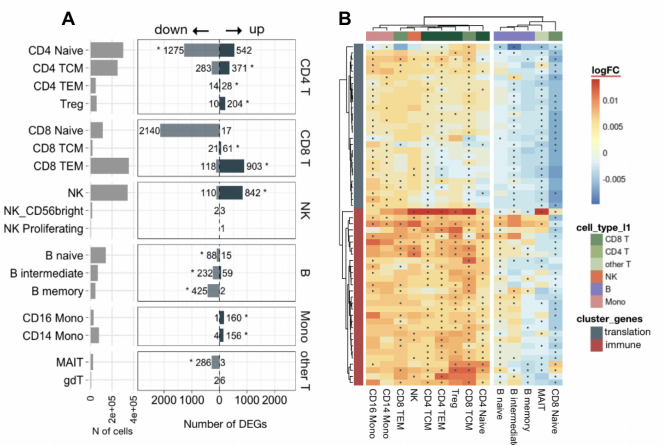
To model changes in cellular proportions, we applied cellular compositional data analysis (cellular CoDA). This approach transforms the relative cellular proportions using a centered log-ratio (CLR) method. The transformed proportions are then modeled with age using a linear mixed model while correcting for covariates including sex, donor, and batch.

## C. Results

### Dual aging trajectory across immune cell populations

Differential expression (DE) analysis with age reveals opposite expression patterns across distinct immune cell populations. Specifically, CD8<sup>+</sup> T naive and B memory cells exhibited the largest number of down-regulated genes, while Natural Killer (NK) cells and CD8<sup>+</sup> T effector memory cells (CD8<sup>+</sup> TEM) showed the largest number of up-regulated genes with age. This opposing pattern led us to identify a group of cell types, including CD8<sup>+</sup> T Naive, CD4<sup>+</sup> T Naive,

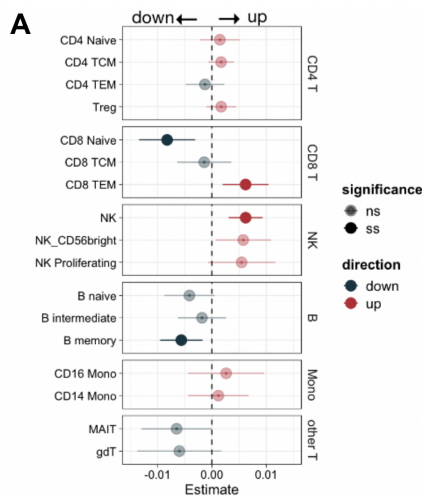
MAIT and B cells, showing a coordinated down-regulation of both inflammatory response and translation-related processes with age. Conversely, the remaining cell types, which include all CD4<sup>+</sup> T cell types, CD8<sup>+</sup> T Memory and CD8<sup>+</sup> Treg cells, monocytes and NK cells, have the same pathways up-regulated including inflammation and translation processes. When looking at genes DE in multiple cell types, we confirm that the discordant directionality pattern in the pathways is driven by the same genes that are up-regulated in one group of cell types and down in the other. These cell-type opposite responses go beyond lineage or functional classifications, highlighting a heterogeneity in aging trajectories within the immune cell repertoire.



**Fig. 1. Gene expression changes in human circulating immune cells in aging.** a. Left. total number of cells analyzed per cell type. Right. Number of age-DEGs up or down-regulated per cell type. Asterisk indicates a significant bias towards up or down-regulation. **B.** Heatmap of the relative fold-change (log<sub>2</sub>FC) of highly shared DEGs (DE > 6 cell types).

### Cell population dynamics and gene expression during aging are interconnected

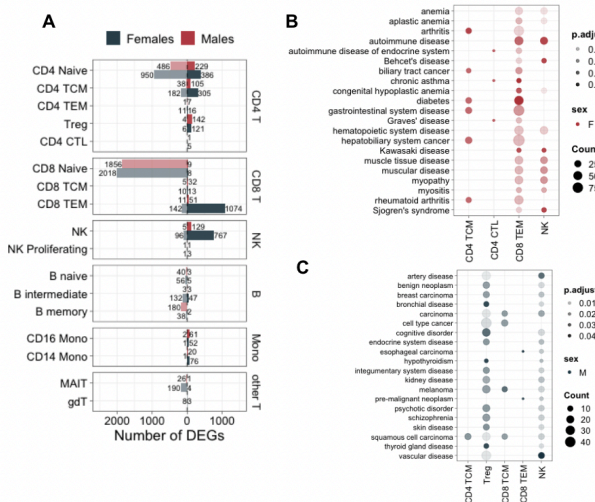
Next, we performed differential cell type composition analysis with age and observed a strong consistency between changes in cellular proportions and gene expression patterns (Fig. 2A). Specifically, cell types with significant down-regulated genes, such as CD8<sup>+</sup> T Naive and B memory, decreased in proportion. Conversely, cell types with substantial age-related up-regulation, including NK and CD8<sup>+</sup> TEM, showed a corresponding increase in their proportions. This underscores the intrinsic link between cell population dynamics and gene expression during aging.



**Fig 2. Changes in cellular proportions with age**

### Sexual-dimorphism in immune system aging

Finally, we carried out a sex-stratified differential expression analysis with age. This revealed sex-specific patterns within certain immune cell populations (Fig. 3A). Specifically, CD8+ TEM in females displayed up-regulated gene expression. Conversely, B memory cells in males showed down-regulation of numerous genes. These findings suggest that sex might be a primary driver of the previously observed age-associated gene expression changes. Finally, we find that up-regulated gene signatures in cell types with pro-inflammatory phenotypes (e.g. NK, CD8+ TEM, CD4+ CTL) are enriched for autoimmune disease exclusively in elderly females (Fig. 3B-C). These results elucidate the crucial role of sex in immune system aging, highlighting the need to include a sex perspective in immunity studies.



**Fig. 3. Sex-stratified expression changes with age.** **A.** Number of age-differentially expressed genes in males (red) and females (blue). **B.** Enrichment of female-specific up-regulated age-DEGs using Disgenet database **C.** Enrichment of male-specific up-regulated age-DEGs using Disgenet database

### D. Conclusions

Overall, our study unveils a dual aging trajectory across immune cell types coupled with joint responses in gene expression and cell type composition and provides

unprecedented insights into the cellular and molecular dynamics underlying immunosenescence.

### References

1. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
2. Franceschi, C., Garagnani, P., Parini, P., Giuliani, C. & Santoro, A. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nat. Rev. Endocrinol.* **14**, 576–590 (2018).
3. Liu, Z. *et al.* Immunosenescence: molecular mechanisms and diseases. *Signal Transduct Target Ther* **8**, 200 (2023).
4. Mogilenko, D. A., Shchukina, I. & Artyomov, M. N. Immune ageing at single-cell resolution. *Nat. Rev. Immunol.* **22**, 484–498 (2022).
5. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
6. Cell ranger - official 10x genomics support. *10x Genomics* <https://www.10xgenomics.com/support/software/cell-ranger/lat-est>.
7. Azimuth. <https://azimuth.hubmapconsortium.org/>.
8. Hoffman, G. E. *et al.* Efficient differential expression analysis of large-scale single cell transcriptomics data using dreamlet. *bioRxiv* (2023) doi:10.1101/2023.03.17.533005.

### Author biography



**Maria Sopena Rios** was born in Reus, Spain in 1998. She received the BSc degree in Human Biology from the University of Pompeu Fabra, Barcelona, in 2020, and the MSc degree in Bioinformatics for Health Sciences from the same university.

Her master's thesis was conducted at the Transcriptomics and Functional Genomics Group at the Barcelona Supercomputing Center. She studied the role of long non-coding RNAs during Ebola virus infection at single-cell resolution. After her master, she did a short research stay at the Comparative Functional Genomics group at the Institute Pasteur to study the evolution of menstruation in primates using single-cell data. She is currently a first-year PhD student at the Melé lab studying the aging immune system at the single-cell resolution.

# Design and Analysis of a Processing-in-Memory Sort Algorithm using UPMEM

Ivan Vargas-Valdivieso\*<sup>†</sup>, Osman Unsal\*, Adrian Cristal\*

\*Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail: {ivan.vargas, osman.unsal, adrian.cristal}@bsc.es

**Keywords**—*Processing in Memory (PIM), UPMEM, Sorting, High-performance computing.*

## I. EXTENDED ABSTRACT

Processing-in-memory (PIM) emerges as a promising solution to alleviate the data movement bottleneck resulting from the restricted bandwidth between host central processing units (CPUs) and main memory. Unlike conventional DRAM memories, PIM involves the integration of a small processor adjacent to the memory banks. The fundamental concept is to preprocess data within the memory prior to transmitting it through the memory hierarchy, thereby maximizing memory bandwidth utilization. Consequently, PIM aims to mitigate the bottleneck associated with memory transactions between the CPU and Memory.

Numerous research efforts have been undertaken in this domain. However, due to the recent availability of most of these technologies, much of the work conducted on Processing-in-Memory (PIM) systems relies on customized concepts validated through simulations [1]. Regrettably, simulating real-world conditions is inherently challenging and may result in inaccurate outcomes.

One of the algorithms that can take advantage of PIM technologies is sorting. Sorting algorithms are crucial for organizing data efficiently, enabling faster retrieval and analysis in various applications. For example, in database management systems, accelerating sorting algorithms can lead to faster retrieval of query results. Accelerating sorting algorithms, such as Quicksort or Radix Sort, using Processing-In-Memory (PIM) technologies like UPMEM [2], holds significant potential for enhancing performance. By integrating processing cores directly into memory banks, PIM reduces data movement between the processor and memory, minimizing latency and improving overall efficiency. This streamlined approach to data processing can lead to faster sorting times and improved system performance, making it an attractive option for applications requiring rapid data manipulation and analysis, such as database management, financial trading, and scientific computing.

In this study, we design and analyze two Processing-In-Memory sorting algorithms optimized specifically for UPMEM, which is presently the first publicly available commodity PIM-enabled technology.

## A. UPMEM architecture

The architecture of UPMEM features a groundbreaking integration of conventional DRAM technology with general-purpose processing cores, known as DRAM Processing Units (DPUs). These DPUs are embedded directly within the memory banks, facilitating seamless data processing and computation in proximity to the stored data. Leveraging this novel arrangement, UPMEM optimizes memory access and computation, thereby alleviating the traditional bottleneck between memory and processing units. Each UPMEM DIMM is equipped with either 8 or 16 PIM chips, and each PIM chip houses 8 DPUs [3].

## B. Analysis of existing sorting algorithms

This section analyzes the profiling results of quicksort and radix, with a specific focus on the memory impact. While the quicksort algorithm is categorized as a comparative sort, radix is classified as a non-comparative sort.

**Quicksort** is a sorting algorithm based on the Divide and Conquer algorithm. It selects an element as a pivot and divides the input array into subarrays in a process named partitioning. Once the pivot is selected, all the elements less than the pivot are moved to the left, and the greater elements to the right. Then, the left and right sides are both sorted recursively. Each subarray picks a pivot and is divided into subarrays, and so on until the array is sorted.

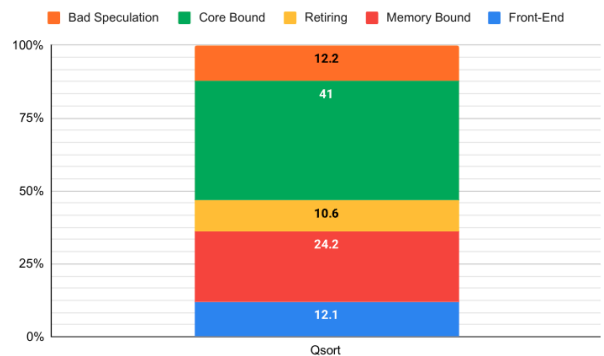


Fig. 1. Profile of the quicksort algorithm.

Figure 1 shows the results of profiling the quicksort running on an Intel Xeon E3-1240 CPU. We make the next observations: First, the execution time of Bad Speculation is high (12.2%) due the high amount of branch mispredictions. The

main reason lies on the partitioning step. As each element is compared with the pivot (carefully selected), the probability of being smaller the pivot is 50%, thereby, branches derived of this comparison are impossible to predict. Second, quick sort relies on the core execution and memory. We observed a high percentage in time execution in the core and the memory. The reason is that every time there is a branch miss prediction, the pipeline execution has to be flushed, then, the high amount of speculative execution and branch misprediction causes extra stress to the core and to the front end, this can be seen in the metrics Core Bound and Front End.

**Radix sort** is a non-comparative numerical sort with a complexity of  $O(k \cdot n)$  where the value  $k$  depends on the number of passes of the algorithm. Radix treats keys as multi-digit numbers, with each digit being an integer value. For example, a 40-bit integer can be treated as a 5-digit number.

Figure 2 shows the results of profiling the radix algorithm. Our evaluation shows that Radix reduces the Bad Speculation from 12.2% in quick sort to 2.1%. This is because radix is a non-comparative algorithm, thus, reducing the stress in the branch predictor. Radix increases the memory demand by 18.6%. As Radix iterates multiple times over the input dataset, the memory operations increased significantly compared to quicksort featuring higher memory misses.

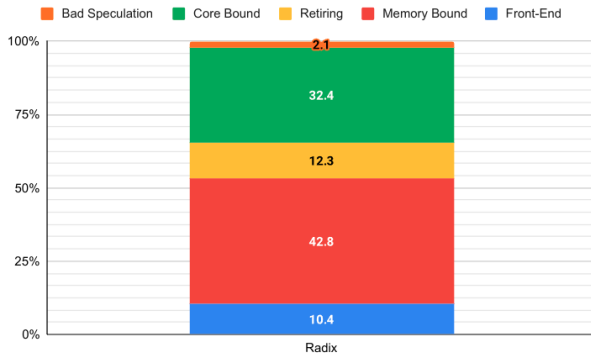


Fig. 2. Profile of the radix algorithm.

### C. Methodology

We evaluate two sorting algorithms, quicksort, and radix, using UPMEM technology, the first PIM system to be commercially available in real hardware. As input dataset we employed 3.6 million integer elements for the 1 DPU analysis and 251 million integer elements for the 256, 512, and 1024 DPU analyses. For the CPU performance comparison, we utilized the Intel Xeon E3-1240, the only compatible CPU for hosting UPMEM technology.

### D. Results and conclusion

In this study, we present the results of a performance comparison between implementations of two sorting algorithms using UPMEM. In Figure 3, we depict the performance results of quicksort and radix using 1, 2, 4, 8, and 16 tasklets (th) within one DPU. A tasklet is the name given by UPMEM to refer to a thread. The results are normalized to the performance of radix with 1 tasklet. In Figure 4, we display the performance results of quicksort and radix using 16 tasklets across 256, 512, and 1024 DPU.

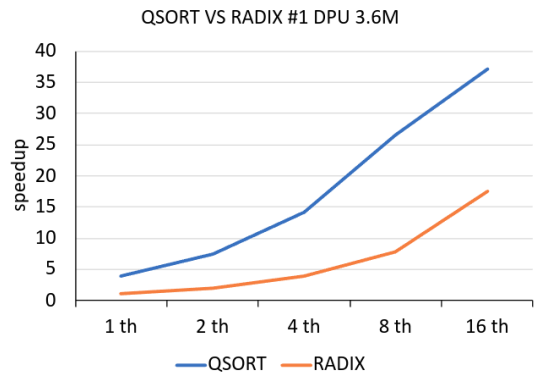


Fig. 3. Performance comparison of quicksort and radix using 1,2,4,8 and 16 tasklets(th).

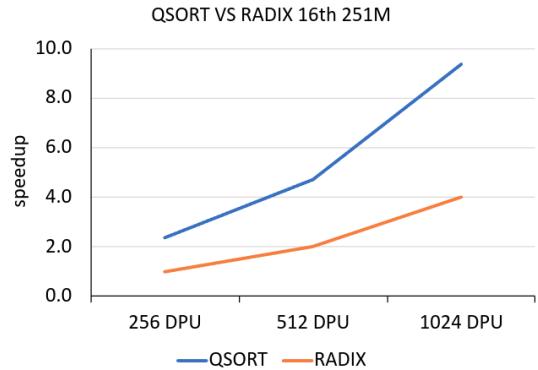


Fig. 4. Performance comparison of quicksort and radix using 16 tasklets(th) and 256, 512 and 1024 DPUs.

1024 DPUs. The results are normalized to the performance of radix using 16 tasklets and 256 DPUs. Upon analysis, we observed that despite radix encountering larger memory issues, quicksort exhibited superior performance when utilizing PIM technology. This is attributed to quicksort demonstrating superior parallel workload division, effectively leveraging the parallelism of UPMEM through the DPUs.

### REFERENCES

- [1] X. Xie, Z. Liang, P. Gu, A. Basak, L. Deng, L. Liang, X. Hu, and Y. Xie, "Spacea: Sparse matrix vector multiplication on processing-in-memory accelerator," in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2021, pp. 570–583.
- [2] UPMEM, "Upmem homepage," <https://sdk.upmem.com/2021.4.0/>, accessed: 2023-04-28.
- [3] J. G.-L. I. El Hajj, I. Fernandez, C. Giannoula, G. F. Oliveira, and O. Mutlu, "Benchmarking a new paradigm: An experimental analysis of a real processing-in-memory architecture."



**Ivan Vargas Valdivieso** is a Research Engineer at the Barcelona Supercomputing Center(BSC) and a PhD candidate in the Department of Computer Architecture at Universitat Politècnica de Catalunya(UPC) under the supervision of Osman Unsal and Adrian Cristal. He obtained his BSc degree in Mechatronic Engineering from Universidad Tecnológica de la Mixteca (UTM), Mexico, in 2010 and completed his MSc degree in Computer Science from Centro de Investigación en Computación(CIC), Mexico, in 2019.

# Methodologies for the Design and Development of Digital Twins

Fernando Vázquez-Novoa\*, Rosa M. Badia\*

\*Barcelona Supercomputing Center, Barcelona, Spain

E-mail: {fernando.vazquez, rosa.m.badia}@bsc.es

**Keywords**—*Digital Twin, DT, HPC, Workflow, Parallelism, Py-COMPSs, COMPSs, Machine learning.*

## I. EXTENDED ABSTRACT

A Digital Twin is virtual copy or representation of a physical object, and given the same input, both should produce the same output. There should be a data flow between both objects that will keep updated both objects, reflecting changes in one of them into the other.

In the paper [1], it is said that some ideas about Digital Twins have been around since the early 2000s and that the first usage of this terminology dates back to 2003 by Grieves. This topic has gained popularity in recent years, reaching a considerable presence in the literature. Despite the widespread use of this terminology in recent years, it presents a need for more standardization. In particular, there is a lack of common architectures and techniques for developing Digital Twin systems. This increases the difficulty of using this technology, as it ends in very specific implementations for each use case.

The thesis this work belongs to starts by providing a general definition that gathers all the essential requirements for the Digital Twin term. Then, a methodology for designing and developing Digital Twins will be proposed and applied to real use cases. In this work, we present a general architecture for generating the workflow that trains the models used by the Digital Twin.

### A. Standard Definition

There are publications that make a bibliographic review of the term Digital Twin, compare the different definitions used along the literature and generate their own proposal for the definition. Some of these publications are [1], [2] and [3]. In the publications, the authors consider a wide range of articles, noticing that there are a lot of systems with different requirements and specifications, all catalogued as Digital Twins.

In the literature, there are two terms in addition to the Digital Twin term that are commonly confused and treated as Digital Twin. The definition of Digital twin is: it is a system that has a digital representation of the physical object and it has a digital flow between both objects reflecting changes one of the copies to the other and keeping both objects in the same state. The two terms that are interchangeably with the term Digital Twin are: Digital Model and Digital Shadow. The difference between the three terms lies on the data flow, in the

Digital Model there is no data flow and there is only data flow from the physical to the digital object in the Digital Shadow.

### B. Life-Cycle

In the methodology proposed in the thesis, Digital Twin systems are considered to have a life-cycle. We assume that the life-cycle of the Digital Twin can be divided into four clearly separated phases: training, deployment, operation and continuous learning.

The training phase is the first step of the life-cycle, where the models that represent the behaviour of the physical object are created. Following, in the deployment phase, the models generated and the workflows for the operation are deployed on the corresponding computing infrastructure. The training may also be performed on this infrastructure, despite being considered a previous step. Then, the operation phase is the main component of the life-cycle. In the operation phase, the Digital Twin does its job; it makes predictions, changes, and acts on the physical object using the digital object and, at the same time, keeps both versions of the object in the same state. Finally, since these systems usually work with a continuous stream of data it is required a continuous learning phase or a retraining. When working with continuous streams, new patterns and/or classes tend to appear in the data, which leads to a loss of the performance of the models.

The continuous learning phase is triggered from the operation step when a loss in the performance of the model is detected. In this phase the models used in the Digital Twin are updated with the new data or they can also be retrained from scratch, generating models that work correctly on the new data.

In this thesis we plan to propose a general architecture for all the steps of this life-cycle as well as a deployment system to ease the deployment of these systems. The general architecture will reflect all the necessary objects and steps to correctly execute the different phases. The objects that make up the architecture are going to have a clear defined purpose as well as a clearly defined API for their usage. At this moment, we proposed the architecture of the training phase which is presented in the following section.

### C. Architecture Proposal for the Training phase

Figure 1 shows our proposal for the architecture of the training phase. This architecture reflects all the objects and steps that compose the workflow of the training phase. The output of this phase is a model that represents the behaviour of the object. The model can represent the general behaviour

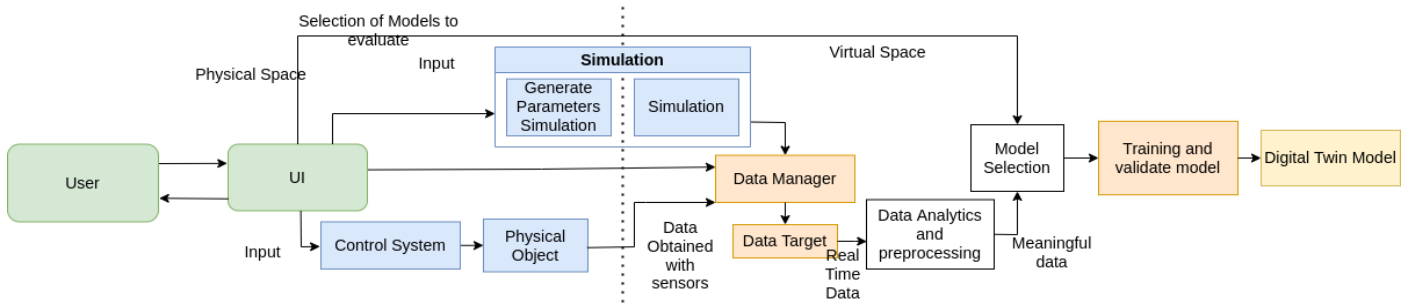


Fig. 1. Architecture proposal for training phase workflows

of the object or it can represent only one of the aspects of the object, for example its endurance.

In order to obtain an accurate representation of the physical object, it may be needed to generate several models that measure different aspects of the object. Each Digital Twin system may need more than a model, which also leads to having more than an unique training workflow.

For each of the components shown on the Figure 1, an API is going to be provided. This API will clearly define the inputs and outputs for each of the objects, generalizing the behaviour of the components across different workflows. An object called Digital Twin that will make usage of the different components and execute the whole workflow by just calling the function.

The functionality provided by the different objects will be implemented in Python. To make an efficient usage of the available resources and efficiently distributing the execution, the components will be developed on top of PyCOMPSs[4], a task-based programming model aiming to simplify the development of parallel and distributed applications. By using these components, the user will be able to generate a distributed workflow with just writing a simple Python script. The user will be completely agnostic from the application parallelism.

#### D. Experimental Environment

The evaluation of the architecture presented on this work has been evaluated in the MareNostrum 4 supercomputer (MN4). This supercomputer is made up of 3456 nodes, each node has two Intel@Xeon Platinum 8160 with 24 cores at 2.1 GHz each, leading to a total of 48 cores per node and 96 GB of main memory. Its peak performance is 11.15 Petaflops.

#### E. Results

This proposed architecture has been used in a real use case, the CAELESTIS project [5]. In the Table I are shown the Logical Lines of Code (LLOC) required to generate the training workflow in two cases. The first case is generating the code from scratch, without making usage of the proposed architecture. In the second case we measured the code of the main script that the user requires to develop using the proposed architecture and the DT object. The number of LLOC

TABLE I. CODE MEASURES.

Code case	LLOC	CC main	Max. CC	Min. CC
From scratch	491	3	8	1
Using DT object	24	1	1	1

to develop by the user is reduced at the same time than the Cyclomatic Complexity (CC).

#### F. Future work

The thesis this work belongs to is going to be focused in the developing of an architecture for the rest of the phases of the life-cycle of the Digital Twins. From the work done during the PhD thesis it is expected to make about 4 publications. These publications will be about the different phases and their usage in real use-cases.

## II. ACKNOWLEDGMENT

Author Fernando Vázquez is supported by PRE2022-104134 funded by MICIU/AEI /10.13039/501100011033 and by the FSE+. This work has been partially funded by the Spanish Government (PID2019-107255GB) y MCIN/AEI /10.13039/501100011033 (CEX2021-001148-S), and by the Departament de Recerca i Universitats de la Generalitat de Catalunya, research group MPIEDist (2021 SGR 00412).

## REFERENCES

- [1] A. Fuller, Z. Fan, C. Day, and C. Barlow, "Digital twin: Enabling technologies, challenges and open research," *IEEE access*, vol. 8, pp. 108 952–108 971, 2020.
- [2] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the digital twin: A systematic literature review," *CIRP journal of manufacturing science and technology*, vol. 29, pp. 36–52, 2020.
- [3] A. Rasheed, O. San, and T. Kvamsdal, "Digital twin: Values, challenges and enablers from a modeling perspective," *Ieee Access*, vol. 8, pp. 21 980–22 012, 2020.
- [4] E. Tejedor, Y. Becerra, G. Alomar, A. Queralt, R. M. Badia, J. Torres, T. Cortes, and J. Labarta, "Pycompss: Parallel computational workflows in python," *The International Journal of High Performance Computing Applications*, vol. 31, no. 1, pp. 66–82, 2017.
- [5] "Caelestis project," <https://www.caelestis-project.eu>, 2023, online; accessed 4 April 2024.



**Fernando Vázquez** received his BSc degree in Computer Engineering from Universidade Santiago de Compostela, Spain in 2020. He completed his MSc degree in Artificial Intelligence in the Universitat Politècnica de Catalunya (UPC) in Spain in 2022. Since January of 2022, he has been with the Workflows and Distributed Computing group of Barcelona Supercomputing Center (BSC). He started a PhD at the department of computer architecture of Universitat Politècnica de Catalunya (UPC), in February of 2024.

# Deep-learning-enhanced transcriptomic and histopathology analysis of the role of aging in female tissues

Laura Ventura\*<sup>†</sup>, Oleksandra Soldatkina\*, Marta Melé\*

\*Barcelona Supercomputing Center, Barcelona, Spain

<sup>†</sup>Universitat Pompeu Fabra, Barcelona, Spain

E-mail: {laura.ventura, oleksandra.soldatkina, marta.mele}@bsc.es

**Keywords**—Aging, menopause, female, CNN, SVM, DEA

## I. EXTENDED ABSTRACT

Life expectancy has increased from 66.8 years in 2000 to 73.4 years in 2019, according to the World Health Organization. This extended lifespan entails that people live longer in the old age stage. Aging is a degenerative and multifactorial process characterized by a progressive decline in cellular functions and morphological changes in virtually all organs [1].

Today, the estimations show that women may spend 40% of their lives post-menopause [2]. Consequently, the health and body changes derived from menopause are even more relevant nowadays, yet they haven't been thoroughly characterized. Some research on aging has shown that there are several molecular mechanisms causing senescence in the ovary, uterus, and vagina. These are advanced-glycation end-products (AGEs), DNA damage, mitochondrial and protein dysfunction, proinflammatory cytokines, oxidative stress, and telomere shortening [3]. These molecular alterations in turn cause morphological changes in the tissues mentioned above.

Given the abundant, but not interconnected, research evidence on changes in female tissue aging, we argue that a systemic approach to understanding the transformation of the female body with menopause is missing, and aim to look at the histological and transcriptomic changes on coupled samples.

Until now, tissue changes have been detected with different imaging approaches and doctors leveraged them to confirm their diagnosis. However, for some years now, Computer-Aided Diagnosis (CAD), which uses machine learning methods to analyze imaging and/or non-imaging patient data, helps clinicians in their diagnosis. Above all, this has improved cancer diagnosis [4].

However, apart from cancer diagnosis, histo(pathological) image processing constitutes a potent tool for detecting tissue features that will then be employed to feed traditional classifiers such as support vector machines (SVMs), or deep-learning approaches [5], [6], [7]. In particular, Convolutional Neural Networks (CNN) are the preferred option in deep learning for computer vision and pattern recognition tasks in images, since they extract relevant features from the input data without needing to manually design them [8]

Together with visual features, a complete scene could be drawn if linked to gene expression. [9] used RNAseq data to elucidate gene expression differences between tissues considering several demographic traits such as age, BMI, ancestry, and sex. The association between tissue features extracted by image processing or deep-learning algorithms and gene expression patterns allows a better understanding of tissue development, providing relevant insights into the molecular pathways that regulate tissue change [10].

There exist different databases containing paired samples of medical images and gene expression that allow studying both data modalities together. However, the attention has been focused on cancer, for example, creating The Cancer Genome Atlas (TCGA) [11]. For non-tumoral tissues, much less information is available, and the Genome-Tissue Expression (GTEx) constitutes the best source [12]. The GTEx provides an extensive and well-curated collection of gene expression and histological samples, obtained from deceased human donors of different ages and ancestries.

The purpose of this work is to study the changes in the ovary, uterus, vagina and female breast with age, leveraging data from GTEx. Through the integration of gene expression data and histological images, we seek to link changes in gene expression caused by age with structural tissue changes. Our approach takes advantage of both machine and deep learning algorithms to improve the knowledge of aging in female tissues and help in the understanding of age-related diseases.

## A. Experimental environment

We used RNAseq and histological data of 4 female tissues -uterus, ovary, vagina, and breast- of ages from 20 to 70. After a thorough data processing, filtering misannotated images, we classified the images into three age groups: age group 1 (20-39 y.o.), age group 2 (40-59 y.o.) and age group 3 (60-79 y.o.). Using PyHIST [13] software, we extracted tiles from the images, keeping only those containing tissue. Then, we separated an external validation set and trained and fine-tuned both SVM and CNN models with age group 1 and age group 3 as classes, to distinguish pre-menopause from post-menopause images in each of the tissues. On the one hand, SVM models, we extracted the Haralick features from the tiles and fed the SVM with them. On the other, for the CNNs we used the VGG19 model [14] pre-trained on the ImageNet dataset [15].



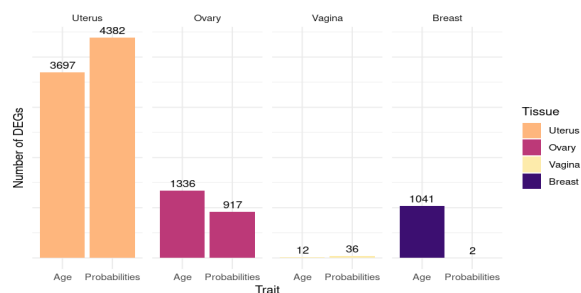


Fig. 1. Number of differentially expressed genes (DEGs) obtained with chronological and histological age in uterus, ovary, vagina, and breast

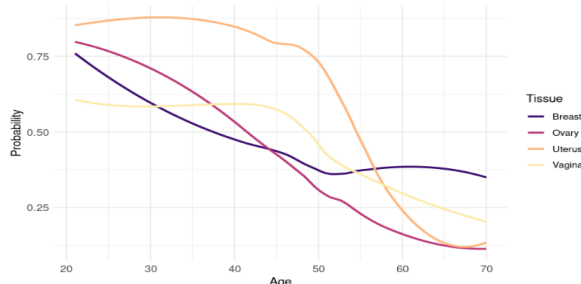


Fig. 2. Image classifier probability of belonging to Group 1 (young, premenopausal) across chronological age per tissue

For interpretability, we assessed the Haralick features that drove the classification in the case of the SVM models, whereas, for the CNN models, we leveraged the Linear Interpretability Model-Agnostic Explanations (LIME) [16] software to explain how the classification was accomplished. For each tissue, the model with the best metrics was selected and used to predict the age group 2 and, therefore, obtain a continuum of probabilities for all the donors in each tissue.

These probabilities are what we call histological age. We then performed a Differential Expression Analysis (DEA) using GTEx RNAseq gene expression data using this histological age as our trait of interest, as well as certain covariates (HardyScale, ischemic time, RNA integrity number, cohort, sequencing quality control metrics, and reads mapping with exons). Next, we accomplished a functional enrichment with the genes we found differentially expressed with histological age, to see which pathways were enriched in our gene sets.

## B. Results

The deep evaluation of the models obtained led to a selection of 4 models, one per tissue, with accuracies of 0.95, 0.95, 0.84, and 0.72, for the uterus, ovary, vagina, and breast, respectively, based on the external validation set. With these models, we predicted the middle age group (age group 2), and the obtained probabilities were used in DEA as histological age instead of chronological age.

For the uterus and vagina, the histological age allowed the discovery of more genes compared to the number of differentially expressed genes with age, but not for the ovary and breast (Fig. 1). In the case of the ovary, this is maybe due to the gradual change that this tissue suffers with time, whereas the uterus and vagina experience more drastic changes when menopause arrives, as shown in Fig. 2.

The functional enrichment of these genes unveils certain biological processes related to menopause that were not discovered when using chronological age, such as osteoporosis, and cardiovascular and muscle decline. Interestingly, it has been reported that vaginal epithelium thins with age and we found pathways related to this process in our approach that were not discovered with age alone. To investigate this further, we developed a CellProfiler [17] pipeline to identify and measure the epithelial layer on vagina images and confirmed the thinning of epithelium following the menopausal transition.

Our research helps us understand how women's bodies change as they age and go through menopause. This knowledge is essential for the development of personalized treatments that take into account age- and menopause-stage-related factors.

## REFERENCES

- [1] P. I. Deryabin and A. V. Borodkina, "Epigenetic clocks provide clues to the mystery of uterine ageing," pp. 259–271, 5 2023.
- [2] K. L. Marlatt *et al.*, "Body composition and cardiometabolic health across the menopause transition," pp. 14–27, 1 2022.
- [3] J. K. Szymański *et al.*, "Vaginal aging—what we know and what we do not know," 5 2021.
- [4] A. S. Sultan *et al.*, "The use of artificial intelligence, machine learning and deep learning in oncologic histopathology," pp. 849–856, 10 2020.
- [5] K. Bera *et al.*, "Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology," *Nature Reviews Clinical Oncology*, vol. 16, pp. 703–715, 11 2019.
- [6] S. Min *et al.*, "Deep learning in bioinformatics," pp. 851–869, 9 2017.
- [7] J. van der Laak *et al.*, "Deep learning in histopathology: the path to the clinic," pp. 775–784, 5 2021.
- [8] A. F. pour *et al.*, "Deep learning features encode interpretable morphologies within histological images," *Scientific Reports*, vol. 12, 12 2022.
- [9] R. García-Pérez *et al.*, "The landscape of expression and alternative splicing variation across human traits," *Cell Genomics*, vol. 3, 1 2023.
- [10] L. Badea and E. Stănescu, "Identifying transcriptomic correlates of histology using deep learning," *PLoS ONE*, vol. 15, 11 2020.
- [11] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," 2013. [Online]. Available: <http://www.cancergenome.nih.gov/>.
- [12] "The gtex consortium atlas of genetic regulatory effects across human tissues the gtex consortium\*," [Online]. Available: [www.gtexportal.org](http://www.gtexportal.org)
- [13] M. Muñoz-Aguirre *et al.*, "Pyhist: A histological image segmentation tool," *PLoS Computational Biology*, vol. 16, 10 2020.
- [14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: <http://www.robots.ox.ac.uk/>
- [15] *Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE Conference on : dates: 20-25 June 2009.* IEEE, 2009.
- [16] M. T. Ribeiro *et al.*, "Model-agnostic interpretability of machine learning," 6 2016. [Online]. Available: <http://arxiv.org/abs/1606.05386>
- [17] M. R. Lamprecht *et al.*, "Cellprofiler™: Free, versatile software for automated biological image analysis," *BioTechniques*, vol. 42, pp. 71–75, 1 2007.



**Laura Ventura** obtained his BSc in Biotechnology from the Pablo de Olavide University (UPO), Sevilla in 2022. This same year, she started an MSc in Bioinformatics for Health Sciences at the Pompeu Fabra University (UPF) in Barcelona. At the same time, she worked as an intern at the PharmacoInformatics group at the Barcelona Biomedical Research Park (PRBB), led by Manuel Pastor. Since 2023, she has worked as an intern at the Transcriptomics and Functional Genomics Lab at the Barcelona Supercomputing Center (BSC), led by Marta Melé, in

deep-learning enhanced transcriptomic and histopathology analysis of images.

# Generative Strategies for Multi-target Drug Design: Generating Mpro Pan-inhibitors

J Vilalta-Mor<sup>#1</sup>, I Filella-Merce<sup>#2</sup>, V Guallar<sup>#&\*3</sup>

<sup>#</sup>Life Science Department, Barcelona Supercomputing Center (BSC), Plaça d'Eusebi Güell, 1-3, 08034, Barcelona, Spain

<sup>&</sup>Nostrum Biodiscovery S.L., Av. de Josep Tarradellas, 8-10, 3-2, 08029, Barcelona, Spain

<sup>\*</sup>ICREA, Pg. Lluís Companys 23, 08010, Barcelona, Spain

<sup>1</sup>julia.vilalta@bsc.es, <sup>2</sup>isaac.filella1@bsc.es, <sup>3</sup>victor.guallar@bsc.es

**Keywords**— Polypharmacology, Generative Models, Drug Discovery, Molecular Modeling, Mpro, SARS-CoV-2

## EXTENDED ABSTRACT

Polypharmacological drugs are molecules capable of simultaneously affecting multiple targets. In the field of drug design, generative AI can be employed to train models on extensive chemical databases, enabling the generation of unseen molecules with specific properties. The presented project aims to leverage the multiobjective capability of a generative model (GM) workflow to design molecules with affinity towards multiple targets, thereby seeking to design polypharmacological drugs. Specifically, we will utilize the vast data collected during the COVID-19 pandemic and the relatively straightforward nature of viruses, to design polypharmacological inhibitors with activity against the main protease (Mpro) of SARS-CoV-2, SARS-CoV, and MERS-CoV. Using this approach, the newly designed compounds will serve as a starting point to fight against new SARS-CoV-2 variants and new virulent coronavirus species.

### A. Introduction

Polypharmacology aims to design molecules capable of simultaneously interacting with multiple targets, providing several advantages over traditional single-target molecules [1]. For instance, a molecule capable of affecting multiple targets could prove more effective in addressing complex diseases like cancer, given its cumulative efficacy across several individual targets [2]. However, these multitarget molecules present challenges stemming from their inherent promiscuity, including the necessity to prevent binding to antitargets, which could lead to off-target adverse effects.

Within Drug Discovery, Generative models (GM) are trained with extensive databases of chemical structures and their properties to learn patterns and relationships between them. Following this training, these models are able to generate new molecules with specific properties, such as enhanced efficacy toward a particular target [3]. Consequently, GM offers an innovative methodology for *de novo* drug design, facilitating the exploration of a much broader space of molecules than traditional screenings. However, GM raises several problems, including the synthesizability and druggability of the newly generated molecules [4].

Stand-alone ML implementations might fall short when designing target-specific drugs due to the applicability domain problem, which restricts the generation of valuable hits beyond the molecular training space. To overcome this limitation, solutions based on active learning have been proposed. For this purpose, Molecular Modeling (MM) techniques are an ideal partner for ML implementations. [5].

### B. Results and Workflow

In this study, we modified our in-house GM workflow, which includes a central VAE and two active learning steps focused on generating target-specific molecules [6] into a

multi-target GM workflow. Fig.1 describes this multi-target GM workflow, employed for generating novel pan-inhibitors targeting SARS-CoV-2, SARS-CoV, and MERS-CoV Mpro.

The workflow starts by introducing a general training set of molecular SMILES into a VAE to teach it how to generate feasible chemical molecules. Then, the VAE is refined with the molecular SMILES from the initial-specific training set consisting of molecules with known or predicted affinity towards the target protein. By doing so, the VAE starts learning how to generate molecules with affinity towards the target. To do so, we utilized an initial pool of known inhibitors with known experimental affinity to the three targets as a specific set. We decided to conduct two parallel workflows with distinct specific training sets: (1) a full specific set, including 477 molecules, and (2) a selective specific set, including 237 molecules, the ones with the highest affinity towards the three targets (according to an initial docking study).

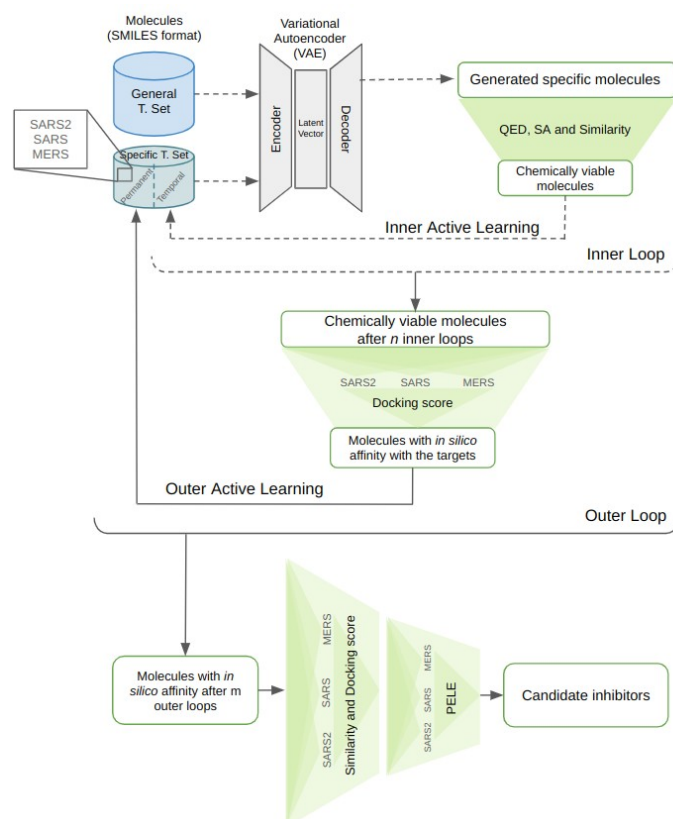


Fig. 1 Multi-target GM workflow. Starting from the top, the workflow begins with the input data provided by the two training sets (general and specific towards the targets), followed by the generation of new molecules with the VAE. Then, it proceeds through two levels of active learning (inner and outer loops) and culminates in the final selection of candidate inhibitors. SARS2, SARS, and MERS refer to the Mpro of SARS-CoV-2, SARS-CoV, and MERS-CoV as targets, respectively.

Preliminary results seem to point towards the direction that the general performance of the model using the selective specific set instead of the full specific set would yield candidate inhibitors with highest affinity to the Mpro of the three targets of study, thereby resulting in pan-inhibitor candidates with high efficiency and reduced promiscuity. This result can be due to the selectivity inherent in the composition of the initial specific set. However, further analysis should be performed in the form of replicas to validate these observations.

### References

- [1] A. Kabir and A. Muth, "Polypharmacology: The science of multi-targeting molecules," *Pharmacol. Res.*, vol. 176, p. 106055, Feb. 2022.
- [2] A. A. Antolin, P. Workman, J. Mestres, and B. Al-Lazikani, "Polypharmacology in Precision Oncology: Current Applications and Future Prospects," *Curr. Pharm. Des.*, vol. 22, no. 46, pp. 6935–6945, Dec. 2016.
- [3] Ł. Maziarka, A. Pocha, J. Kaczmarczyk, K. Rataj, T. Danel, and M. Warchoń, "Mol-CycleGAN: a generative model for molecular optimization," *J. Cheminformatics*, vol. 12, no. 1, p. 2, Jan. 2020.
- [4] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, "Generative models for molecular discovery: Recent advances and challenges," *WIREs Comput. Mol. Sci.*, vol. 12, no. 5, p. e1608, 2022.
- [5] M. Thomas, A. Bender, and C. de Graaf, "Integrating structure-based approaches in generative molecular design," *Curr. Opin. Struct. Biol.*, vol. 79, p. 102559, Apr. 2023.
- [6] I. Filella-Merce et al., "Optimizing Drug Design by Merging Generative AI With Active Learning Frameworks." arXiv, May 04, 2023

### Author biography



**Júlia Vilalta Mor** was born in La Selva del Camp, Spain, in 1999. She received a bachelor's degree in Biotechnology from the Universitat Rovira i Virgili (URV), Tarragona, in 2021, and a Master's degree in Bioinformatics for Health Sciences from the Universitat Pompeu Fabra (UPF), Barcelona, in 2023.

She has been working in the Electronic and Atomic Protein Modeling (EAPM) group in the Life Sciences department at the Barcelona Supercomputing Center (BSC-CNS) during her master's thesis studying Generative AI and Molecular Modeling techniques. Currently, she is doing a PhD conducting research on protein-protein modulating drugs.

# Active Compute Memory: Enhancing Memory and Processing in Near-Memory Architectures for Vector Classification

Victor Xirau\*, Pouya Esmaili\*, Petar Radojković\*

\*Barcelona Supercomputing Center, Barcelona, Spain

E-mail: {victor.xirau, pouya.esmaili, petar.radojkovic}@bsc.es

**Keywords**—Computer Architecture, Active Compute Memory ACM, Processing In Memory PIM, Classification in Memory

## I. EXTENDED ABSTRACT

This study evaluates the Active Compute Memory (ACM) architecture [1] for vector classification, diverging from its original use while maintaining its microarchitecture. Conducted in collaboration with La Salle Barcelona University for a final thesis, we combined analytical modeling and hardware simulation to validate our findings. We found that ACM excels in speed and achieves up to 95% accuracy in tasks with 2-3 classes. However, accuracy drops below 50% for tasks with more classes, indicating a need for further optimization for complex classifications. These results reveal a trade-off between speed and accuracy, showcasing ACM's potential as an alternative to traditional CPU-based methods for data-intensive tasks, and contributing to computer architecture advancements with practical implications for real-world applications.

### A. Enhancing ACM with Vector Classification

Introduced by BSC, the ACM architecture innovates in-memory computing by efficiently sorting key-value pairs within DRAM [1]. This system, integrating Data-RAM, Sort-RAM (with KeyTable, MetaTable, and Control Logic), significantly surpasses traditional CPU-based methods in both performance and energy efficiency.

#### a) ACM's Original Functionality and Limitations:

Initially, ACM was crafted for sorting, allowing data retrieval by sorted keys via indirect addressing through KeyTable (KT) and MetaTable (MT) manipulation. However, its basic classification algorithm—categorizing data by labels—struggles with the complexity and variety of contemporary datasets that often feature unlabeled, multidimensional data.

- *Methodological Shift:* The move towards vector classification entails a transition from processing elements with explicit labels to classifying multidimensional vectors against expected class vectors.
- *Implementation Considerations:* The vector classification process within ACM leverages the existing architecture, repurposing the MT to store expected vectors for each class and utilizing the Control Logic and Bots for dynamic vector comparison and classification. This method effectively transforms the ACM from a sorting device into a powerful classification tool.

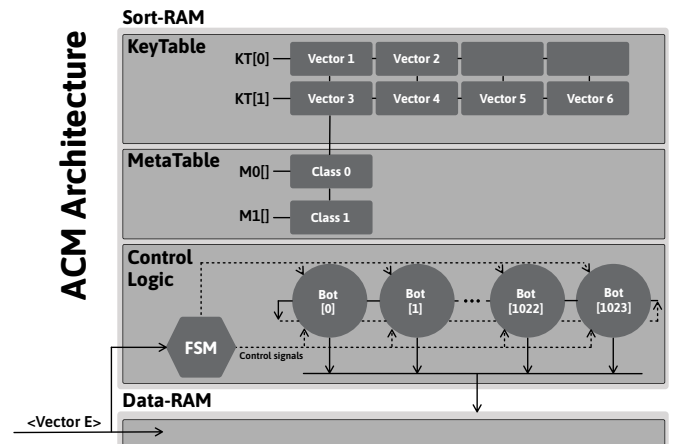


Fig. 1. Adaptation of ACM architecture for vector classification.

#### b) Exploration of Advanced Classification Algorithms:

The ACM's reliance on exact match comparisons was challenged by its limitations in handling complex, multidimensional data. To address this, a study explored alternative algorithms to improve accuracy. The Absolute Difference per Element Algorithm stood out, assessing the absolute difference between input vectors and expected class vectors against a set threshold, typically 0.75, for a more nuanced classification. This approach demonstrated superior flexibility and accuracy over other algorithms, confirmed by tests on diverse datasets like MNIST [2], Breast Cancer Wisconsin [3], Iris [4], Titanic [5], and Wine [6] and replicating the ACM's behaviour in CPU to run them.

c) *Hardware Considerations for Implementation:* Implementing the Absolute Difference per Element Algorithm within the ACM architecture required minimal additional hardware overhead. The proposed design utilizes comparators and multiplexers to compute absolute differences, with BOT units performing subtraction operations.

### B. Experimental Environment

The evaluation of ACM Vector Classification within this study adopts an Analytical Model approach, this strategy is essential for delving into the performance characteristics and efficiency of the ACM algorithms, providing a detailed examination in the absence of the actual hardware. The move towards an Analytical Model stems from the necessity to simulate

and assess algorithmic behaviors and potential optimizations realistically, circumventing the limitations inherent in purely theoretical computational complexity analyses.

To bridge the gap between theoretical analysis and tangible hardware evaluation, we enhanced the publicly available ZSim [7] and DRAMSim3 [8] simulators. These enhancements are tailored to accommodate the specific requirements of ACM evaluation, enabling practical benchmarks and performance validation of the ACM architecture. The utilization of the ZSim simulator, in particular, has been crucial for corroborating the findings of the Analytical Model, ensuring that the ACM's conceptual design does not introduce computational bottlenecks.

### C. Results

The ACM architecture's exploration in classification tasks across different datasets required custom processing for each, accommodating up to 120 elements per vector within class and dimension constraints. Through CPU emulation, variances in ACM's performance were observed, particularly on the MNIST dataset where accuracy ranged from about 10% to 19%, significantly below state-of-the-art (SOTA) methods, underscoring the algorithm's limitations and the impact of datasets with high zero-value prevalence.

Execution speeds were notably quick, mostly under 0.2 seconds, except for larger datasets like MNIST. This demonstrates ACM's fast processing but highlights a trade-off with accuracy. When compared to conventional CPU-based approaches, the ACM showed reduced execution times with about 80% accuracy in simpler scenarios (2-3 classes) but saw a decrease to below 50% in more complex classifications, indicating a pressing need for algorithmic improvement and optimization to enhance its classification performance across a broader range of applications.

### D. Conclusion

This study elucidates the potential and challenges of employing ACM for vector classification tasks. While the ACM exhibits exceptional speed, particularly in simpler classification scenarios with fewer output classes, the accuracy in more complex applications necessitates improvement. The findings from real-world dataset applications reveal a critical trade-off between execution speed and classification accuracy. Moving forward, enhancing the ACM's algorithmic framework to better

accommodate a broader spectrum of classification tasks without compromising on speed or accuracy remains a pivotal area of research.

## II. ACKNOWLEDGMENT

This thesis has been formally recognized and accepted by La Salle University, where it was awarded the distinction of Honors. While this work has not yet been published, efforts are underway to prepare a manuscript for submission.

## REFERENCES

- [1] P. Esmaili-Dokht, M. Guiot, P. Radojković, X. Martorell, E. Ayguadé, J. Labarta, J. Adlard, P. Amato, and M. Sforzin, "O (n) key-value sort with active compute memory," *IEEE Transactions on Computers*, 2024.
- [2] Ultralytics, "Mnist," n.d. [Online]. Available: <https://docs.ultralytics.com/datasets/classify/mnist>
- [3] "Breast cancer wisconsin (diagnostic) data set," n.d. [Online]. Available: <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
- [4] panData, "Unveiling the mysteries of the iris dataset: A comprehensive analysis and machine learning," March 2023. [Online]. Available: <https://levelup.gitconnected.com/unveiling-the-mysteries-of-the-iris-dataset-a-comprehensive-analysis-and-machine-learning-f5c4f9dbcd6d>
- [5] N. Donges, "Predicting the survival of titanic passengers," May 2018. [Online]. Available: <https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8>
- [6] "Uci machine learning repository," n.d. [Online]. Available: <https://archive.ics.uci.edu/dataset/109/wine>
- [7] D. Sanchez and C. Kozyrakis, "Zsim: fast and accurate microarchitectural simulation of thousand-core systems," in *40th Annual International Symposium on Computer Architecture (ISCA)*, 2013.
- [8] S. Li, Z. Yang, D. Reddy, A. Srivastava, and B. Jacob, "Dramsim3: A cycle-accurate, thermal-capable dram simulator," *IEEE Computer Architecture Letters*, vol. 19, no. 2, 2020.



**Victor Xirau** received his double BSc degree in Computer Engineering and Multimedia Engineering from La Salle Barcelona University in 2023. He was a Teacher Assistant for Operating Systems and Compiler Design from 2021 to 2023. He then went on to complete his final degree's thesis with the Memory Systems group at the Barcelona Supercomputing Center (BSC) in 2022. Since 2023, Victor is a full-time Research Engineer in the Memory Team at BSC and serves as a part-time Lecturer at La Salle Barcelona.



## 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 