# 8th BSC Doctoral Symposium

## Online | 11th - 13th May 2021

## Book of Abstracts

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

*Book of Abstracts*
8th BSC Doctoral Symposium

*Editor*
Carolina Olmo
Michelle Pinto

*Cover*
Design based on artwork created by macrovector.com

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

*published by*
Barcelona Supercomputing Center

8th Edition, May 2021

# ACKNOWLEDGEMENTS

The BSC Education & Training team gratefully acknowledges all the PhD candidates, advisors, Postdoc researchers, experts and especially the Keynote Speaker Prof. Wen-mei W. Hwu for contributing to this Book of Abstracts and participating in the 8th BSC Doctoral Symposium 2021. We also wish to thank the volunteers that supported the organisation of the event: Mónica Cabrera, Christian Guzman, Peini Liu, Héctor Navarro, Winona Oliveros, Alejandro Ortiz, Rubén Sousse, Georgios Vavouliotis.

BSC Education & Training team
education@bsc.es

# EDITORIAL COMMENT

We are proud to present the Book of Abstracts for the 8th BSC Doctoral Symposium.

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

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

In this 8th edition of the BSC Doctoral Symposium we are offering a keynote talk titled “Redesigning Computing Systems in the Age of Huge Data and Sparse Computation" by Prof. Wen-mei W. Hwu and a tutorial on Cognitive Abilities for Team Innovation ¨CATI¨.

The talks will be held in six different sessions and will tackle the topics of: Life Sciences & Genomics, Modelling & HPC, Computer Architecture, HPC applications in Earth and Life Sciences, HPC and Modelling for Earth Sciences and Machine Learning and Quantum computing and there will be two poster sessions.

This 8th edition of the BSC Doctoral Symposium has moved to the virtual scenario due to the Covid-19 ongoing limitations, however, it didn’t prevent us from giving visibility to our Phd student’s research. This Book of Abstracts is the result of their contributions.

# WELCOME ADDRESS

I am delighted to welcome all the PhD students, Postdoc researchers, advisors and experts to the 8th BSC Doctoral Symposium.

This 8$^{th}$ edition of the BSC Doctoral Symposium has substantially changed its format given the current Covid-19 circumstances. Nevertheless, the goal of the occasion continues to be providing a framework to share research results of the projects developed by PhD thesis that use High Performance Computing in some degree. The symposium provides an interactive forum for PhD students considering both the ones just beginning their research and others who have developed their research activities during several years.

I am very grateful to the BSC directors for supporting the symposium, to the group leaders and to the advisors for encouraging the participation of the students in the event.

I would also like to thank all PhD students and Postdoc researchers for their papers and presentations. I wish you all the best for your career and I really hope you enjoy this great opportunity to meet other colleagues and share your experiences even on a virtual scenario.

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

Finally, given the difficult situation consequence of Covid-19 I very much appreciate and I am very thankful for the interest and commitment of all BSC community with the Doctoral Symposium. Only their engagement has made this year’s edition possible.

Dr. Maria Ribera Sancho
Manager of BSC Education & Training

# KEYNOTE SPEAKER

## Wen-mei Hwu

Senior Distinguished Research Scientist, NVIIA Professor, and
Sanders-AMD Chair Emeritus, ECE University of Illinois at Urbana-Champaign

## Redesigning Computing Systems in the Age of Huge Data and Sparse Computation

We have been experiencing two very important developments in computing. On the one hand, a tremendous amount of resources have been invested into innovative applications such as first-principle based models, deep learning and cognitive computing. On the other hand, the industry has been taking a technological path where traditional scaling is coming to an end and application performance and power efficiency vary by more than two orders of magnitude depending on their parallelism, heterogeneity, and locality. A "perfect storm" has been formed from the fact that data movement has become the dominating factor for both power and performance of high-valued applications. It will be critical to match the compute throughput to the data access bandwidth and to locate the compute at where the data is. Much has been and continuously needs to be learned about of algorithms, languages, compilers and hardware architecture in this movement. What are the killer applications that may become the new driver for future technology development? How hard is it to program existing systems to address the date movement issues today? How will we program future systems? How will innovations in memory devices present further opportunities and challenges in designing new systems? What is the impact on long-term software engineering cost on applications (and legacy applications in particular)? In this talk, I will present our vision for and initial results from the IBM-Illinois C3SR Erudite system inside this perfect storm.

Wen-mei W. Hwu is a Senior Distinguished Research Scientist and Senior Director of Research at NVIDIA. He is also a Professor Emeritus and the Sanders-AMD Endowed Chair Emeritus of ECE at the University of Illinois at Urbana-Champaign. His research is in the architecture, algorithms, and infrastructure software for data intensive and computational intelligence applications. He directed the IBM-Illinois Center for Cognitive Computing Systems Research Center (c3sr.com) from 2016 to 2020. He was a PI of the NSF Blue Waters supercomputer project. He received the ACM SigArch Maurice Wilkes Award, the ACM Grace Murray Hopper Award, the IEEE Computer Society Charles Babbage Award, the ISCA Influential Paper Award, the MICRO Test-of-Time Award, the IEEE Computer Society B. R. Rau Award, the CGO Test-of-Time Award, numerous best paper awards, numerous teaching awards, and the Distinguished Alumni Award in CS of the University of California, Berkeley. He is a Fellow of IEEE and ACM.

# TUTORIAL

## QUERALT PRAT-I-PUBILL

Management professional with more than 25 years of experience. She has worked for recognized international companies as JPMorgan, in M&A and Equity Derivatives, for Reuters technology incubator and she has also developed several businesses in technology.

## Cognitive Abilities for Team Innovation ¨CATI¨ Exploring our cognitive abilities to build our collaboration capabilities

Goals:

1. INQUIRY
We will inquiry about our mechanisms of interpretation.
2. AWARENESS
We will become aware of the two dimensions of reality
3. SUCCESS
We will apply what we learn from the ¨get go¨

Session 1 (Wednesday May 12th, 2021)

We will use the way we deal with conflict as an inquiry entry to our mental models.
We will work on understanding the creative implications of automatically responding to our particular models.

Session 2 (Thursday May 13th, 2021)

We will practice differentiating our automatic mechanisms of interpretation, what we call the relative dimension from the absolute dimension.
Inquiring about our mechanism and being able to distance and silence them will foster a different approach to collaboration.

"Team engagement is not spontaneously produced but rationally and qualitatively created"

This is an introductory workshop for fostering collaboration in teams. There are no prerequisites to participate.

# PROGRAM

## DAY 1 (May 11th)

| Start time | Activity | Speaker/s | Chair |
|---|---|---|---|
| **14.45h Doctoral Symposium waiting room** | | | |
| 15.00h | **Welcome** | **Maria-Ribera Sancho,** Education&Training Manager | |
| 15.10h | **Opening** | **Josep Mª Martorell,** BSC Associate Director | |
| 15.30h | **Keynote talk:** Redesigning Computing Systems in the Age of Huge Data and Sparse Computation | **Dr. Wen-mei W. Hwu**<br>Senior Distinguished Research Scientist, NVIDIA Professor and Sanders-AMD Chair Emeritus, ECE University of Illinois at Urbana-Champaign | **Maria-Ribera Sancho** |
| | Abstract: We have been experiencing two very important developments in computing. On the one hand, a tremendous amount of resources have been invested into innovative applications such as first-principle based models, deep learning and cognitive computing. On the other hand, the industry has been taking a technological path where traditional scaling is coming to an end and application performance and power efficiency vary by more than two orders of magnitude depending on their parallelism, heterogeneity, and locality. A "perfect storm" has been formed from the fact that data movement has become the dominating factor for both power and performance of high-valued applications. It will be critical to match the compute throughput to the data access bandwidth and to locate the compute at where the data is. Much has been and continuously needs to be learned about of algorithms, languages, compilers and hardware architecture in this movement. What are the killer applications that may become the new driver for future technology development? How hard is it to program existing systems to address the date movement issues today? How will we program future systems? How will innovations in memory devices present further opportunities and challenges in designing new systems? What is the impact on long-term software engineering cost on applications (and legacy applications in particular)? In this talk, I will present our vision for and initial results from the IBM-Illinois C3SR Erudite system inside this perfect storm | | |
| **16.30h Event screenshot** | | | |
| **16.30h Break** | | | |
| **16.40 First Poster Session: Data-based solutions for medicine, astronomy and society** | | | |
| | 1. Adaptive Optics Control with ReinforcementLearning: First steps **Bartomeu Pou** | | |
| | 2. Lindaview: An OBDA-based tool for self-sufficiency assessment **Victor-Alejandro Ortiz** | | |
| | 3. Multiplex network uncovers Chronic Obstructive Pulmonary Disease endotypes **Núria Olvera Ocaña** | | |
| **17h Break** | | | |
| **17.10h First Talk Session: Life Sciences & Genomics** | | | |
| 17.10h | Epigenetic Characterization of Cholangiocarcinomas | **Winona Oliveros Diez** | **David Torrents** |
| 17.30h | Unveiling the Transcriptional and Cellular Landscape of Age across Human Tissues | **Aida Ripoll Cladellas** | |
| 17.50h | From Comorbidities to Gene Expression Fingerprints and Back | **Beatriz Urda** | |
| 18:10h | perSVade: personalized Structural Variation detection in your species of interest | **Miquel Àngel Schikora Tamarit** | |
| **18.30h Adjourn** | | | |

## DAY 2 (May 12th)

| Start time | Activity | Speaker/s | Chair |
|---|---|---|---|
| 9.00h | Opening of the second day | | |

**9.20h Second Talk Session: Modelling & HPC**

| | | | |
|---|---|---|---|
| 9.20h | startR: A tool for large multi-dimensional data processing | **An-Chi Ho** | **Rosa Badia** |
| 9.40h | Curved geometry modeling: interpolation of subdivision features | **Albert Jiménez Ramos** | |
| 10.00h | Optimizing Execution on Large-scale Infrastructures by Integrating Task-based workflows and MPI | **Hatem Elshazly** | |
| 10.20h | VIA: A Smart Scratchpad for Vector Units with Application to Sparse Matrix | **Julián Pavón** | |

**10.40h Break**

**10.50h Third Talk Session: Computer Architecture**

| | | | |
|---|---|---|---|
| 10.50h | Predicate-Based Filtering for Multi-GPU Utilization in Directive-Based | **Kazuaki Matsumura** | **Petar Radojkovic** |
| 11.10h | Pushing the Envelope on Free TLB Prefetching | **Georgios Vavouliotis** | |
| 11.30h | Cost-Aware Prediction of Uncorrected DRAM Errors in the Field | **Isaac Boixaderas** | |
| 11.50h | Optimizing the SpMV kernel on long-vector accelerators | **Constantino Gómez** | |

**12.10h Break**

**12.20h Fourth Talk Session: HPC applications in Earth and Life Sciences**

| | | | |
|---|---|---|---|
| 12.20h | Determining the structure of small molecules via their pseudo-electrons and atoms 3D models using FPGA | **Cesar Gonzalez** | **Xavier Martorell** |
| 12.40h | Mining the essential motions of pyruvate kinase | **Luis Jorda** | |
| 13.00h | Constraining the chemical composition of particulatematter in an atmospheric chemistry model | **Hector Navarro Barboza** | |
| 13.20h | The multilayer community structure of medulloblastoma | **Iker Núñez Carpintero** | |
| 13.40h | Tsunami inundation forecast in central Chile using stochastic earthquake scenario | **Natalia Zamora** | |

**14.00h Lunch break**

**15.00h Tutorial part 1**

| | |
|---|---|
| Title: Cognitive Abilities for Team Innovation ¨CATI¨ part 1 | **Queralt Prat-i-Pubill** |

Content&Goals

1. INQUIRY. We will inquiry about our mechanisms of interpretation.
2. AWARENESS. We will become aware of the two dimensions of reality
3. SUCCESS. We will apply what we learn from the ¨get go¨

We will use the way we deal with conflict as an inquiry entry to our mental models.
We will work on understanding the creative implications of automatically responding to our particular models.

**17.30h Adjourn**

## DAY 3 (May 13th)

| Start time | Activity | Speaker/s | Chair |
|---|---|---|---|
| 9.30h | Opening of the third day | | |

**10.00h Fifth Talk Session:  HPC and Modelling for Earth Science**

| | | | |
|---|---|---|---|
| 10.00h | High Resolution Decadal Prediction - Impacts on the predictability of the Pacific variability | **Aude Carréric** | **Pablo Ortega** |
| 10.20h | Climate Forecast Analysis Tools Framework | **Núria Pérez-Zanón** | |
| 10.40h | Bias-adjustment method for street-scale air quality models | **Jan Mateu Armengol** | |
| 11.00h | Exploiting parallelism for CPU and GPU linear solvers on chemistry for atmospheric models | **Christian Guzman Ruiz** | |
| 11.20h | Super-resolution for downscaling climate data | **Carlos Alberto Gómez Gonzalez** | |

**11.40h  Break**

**11.50 Second Poster Session: Dust modelling and genome sequencing**

1. Analysis of Hybrid Genomes in the Candida parapsilosis Clade **Valentina del Olmo**
2. Sensitivity of soluble iron deposition to soilmineralogy uncertainty **Elisa Bergas-Massó**
3. Modeling nitric acid uptake by mineral dust **Rubén Sousse Villa**

**12.10h  Break**

**12.20h Sixth Talk Session: Machine Learning and Quantum computing**

| | | | |
|---|---|---|---|
| 12.20h | An architecture for autonomic ML/AI workflow management and supervision | **Peini Liu** | **Josep Lluís Berral** |
| 12.40h | Quantum Singular Value Decomposer | **Diego García-Martín** | |
| 13.00h | Algebraic Linelet Preconditioner for the solution of the Poisson equation on boundary layer flows | **Ramiro de Olazábal** | |
| 13.20h | TunaOil: A Tuning Algorithm Strategy for Reservoir Simulation Workloads | **Felipe Portella** | |
| 13.40h | A Machine Learning based Wall Model for LES of Turbulent flows | **Sarath Radhakrishnan** | |

**14 00h Lunch break**

**15.00h Tutorial part 2**

| | |
|---|---|
| Title: Cognitive Abilities for Team Innovation ¨CATI¨ part 2 | **Queralt Prat-i-Pubill** |

Content&Goals

2. AWARENESS. We will become aware of the two dimensions of reality
3. SUCCESS. We will apply what we learn from the ¨get go¨

We will practice differentiating our automatic mechanisms of interpretation, what we call the relative dimension from the absolute dimension.

Inquiring about our mechanism and being able to distance and silence them will foster a different approach to collaboration.

**17.30  End of the Doctoral Symposium**

# TABLE OF CONTENTS

# Talk Abstracts

# Cost-Aware Prediction of Uncorrected DRAM Errors in the Field

Isaac Boixaderas*, Paul M. Carpenter*, Petar Radojković*, Eduard Ayguadé*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {isaac.boixaderas, paul.carpenter, petar.radojkovic, eduard.ayguade}@bsc.es

***Keywords*—Memory system, Reliability, Error prediction, Machine learning, Random forest, Cost–benefit analysis.**

## I. Extended Abstract

One of the main causes of hardware failure in large-scale clusters is an uncorrected error in main memory [1]–[4]. Node failures are especially problematic in high-performance computing (HPC), where a single tightly-coupled job may execute for days on thousands of nodes. If any node fails, the whole job is terminated, typically wasting all CPU hours since the last checkpoint. Memory system reliability is therefore an important limit on the ability to scale to larger systems.

This abstract summarizes our study, published in SC20 [5], which aims to increase effective use of HPC systems by reducing the compute time lost due to memory system failures. Firstly, our study presents and evaluates a method to predict DRAM uncorrected errors (UEs) that can enable the system to take active mitigation measures, e.g. checkpointing or live job migration. We concentrate on uncorrected errors (UEs), which cause the node to fail, rather than corrected errors (CEs), which do not have a direct connection to UEs.

Secondly, we discuss and clarify several aspects of methodology, relating to cost–benefit analysis and potential sources of bias, that are essential for such a prediction method to be useful in practice. We show that standard metrics for data prediction, such as precision, recall and F1-score, are not correlated with saved compute time or mitigation costs, and therefore are insufficient to decide whether and for which model parameters the prediction is useful in practice. Instead, we use a cost–benefit analysis which directly compares the system resources needed for training, failure prediction and failure mitigation against the saved compute time due to successful failure prediction and mitigation [6].

Overall, our open source method [7], reduces lost compute time by up to 57%, a net savings of 21,000 node–hours per year for a real production job distribution. We encourage the community to adopt our methodology for pre-processing, model training, parameter exploration and evaluation, so that future DRAM error prediction methods are also free from training bias and supported by a cost–benefit calculation.

### A. Environment description

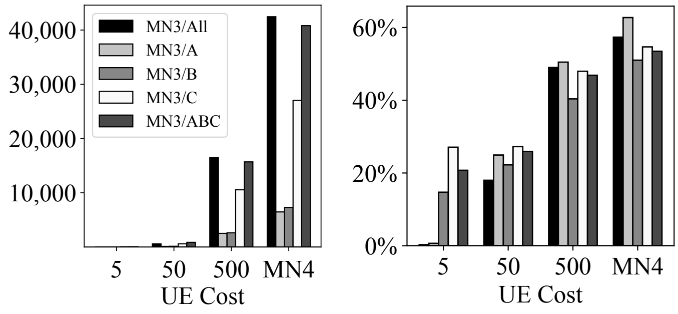
Our prediction method is trained and evaluated on memory error logs from the PRACE Tier-0 MareNostrum 3 supercomputer [8] over a production period from October 2014 to November 2016. At the time, MareNostrum 3 comprised 3056 compute nodes and more than 25,000 DDR3-1600 DIMMs from all three major memory manufacturers. These manufacturers have been anonymized, and are referred to as *Manufacturer A* (6694 DIMMs), *B* (5207 DIMMs) and *C* (13,419 DIMMs).

### B. Prediction method and methodology

We perform uncorrected DRAM error prediction using a random forest classifier. The classifier makes one prediction for each DIMM as to whether or not it will experience a UE in the upcoming "prediction window". The features were obtained directly from the CE and event logs. We use offline learning, with hyperparameter tuning and time series cross-validation.

### C. Results

Figure 1 summarizes the results of the cost–benefit analysis for the prediction model and mitigation. The $x$-axis is the UE cost: fixed at 5, 50 and 500 node–hours or calculated using the job size distribution from production MareNostrum 4 HPC job logs. The $y$-axis in Figure 1a is the saved node–hours, which is the reduction in lost compute time due to UE prediction and mitigation compared with the baseline system. The $y$-axis in Figure 1b is the saved node–hours normalized to the number of node–hours lost in the baseline system. Results are shown for MareNostrum 3 as a whole (MN3/All) and its different subsystems corresponding to DRAM manufacturer: MN3/A, MN3/B and MN3/C. Bars MN3/ABC show the overall results for the whole system treated as the sum of its three DRAM manufacturer subsystems. The results show that the effectiveness of the method is similar across all these scenarios but that the cost–benefit calculation is strongly influenced by the average UE cost. For a small UE cost of 5 node–hours, UE prediction has zero effect on the saved compute time. For medium and large UE costs, however, savings are seen in Figure 1a, of 586 node–hours (medium) and 16,541 node–hours (large). These are reductions of 18% and 49% respectively (Figure 1b). The node–hours savings computed based on the production job logs reach 57% which is equivalent to 42,000 node–hours or 21,000 node–hours per year.



(a) Number of saved node–hours (b) Percentage saved node–hours

Fig. 1: The model cost-efficiency depends on the UE cost. For large UE cost, the savings are significant, measured in thousands of node–hours over the two-year production period.

### D. Conclusions

This paper summarized our method to predict DRAM uncorrected errors and our cost–benefit methodology. We see that the effectiveness of our prediction scheme is highly dependent on system and workload characteristics, pointing the way to future work on adaptive resiliency techniques. The full paper [5] provides full details on the prediction method, the features used for prediction and the detailed cost–benefit methodology. It also analyzes the effect of parameters such as prediction window, prediction frequency, and decision threshold. It evaluates the method also with standard data prediction methods, explores the relative importance of prediction features and compares the random forest approach with five other machine learning classifiers. Overall, we hope that future researchers will build on our work to improve the throughput of production HPC systems as demonstrated by a clear cost–benefit calculation.

## Acknowledgment

This work was supported by the Spanish Ministry of Science and Technology (project PID2019-107255GB / AEI / 10.13039/501100011033), Generalitat de Catalunya (contracts 2014-SGR-1051 and 2014-SGR-1272), the Ministry of Economy and Competitiveness of Spain (Ramon y Cajal fellowships YC2018-025628-I and RYC2017-23269) and the European Union's Horizon 2020 research and innovation programme (grant agreement No 754337, EuroEXA).

## References

[1] HP, "How memory RAS technologies can enhance the uptime of HPE ProLiant servers," Hewlett Packard Enterprise, Technical white paper 4AA4-3490ENW, Feb 2016.

[2] I. Giurgiu, J. Szabo, D. Wiesmann, and J. Bird, "Predicting DRAM Reliability in the Field with Machine Learning," in *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference: Industrial Track*, 2017, pp. 15–21.

[3] B. Schroeder, E. Pinheiro, and W.-D. Weber, "DRAM Errors in the Wild: A Large-scale Field Study," in *Proceedings of the International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, 2009, pp. 193–204.

[4] A. A. Hwang, I. A. Stefanovici, and B. Schroeder, "Cosmic rays don't strike twice: understanding the nature of dram errors and the implications for system design," *ACM SIGPLAN Notices*, vol. 47, no. 4, pp. 111–122, 2012.

[5] I. Boixaderas, D. Zivanovic, S. Moré, J. Bartolome, D. Vicente, M. Casas, P. M. Carpenter, P. Radojković, and E. Ayguadé, "Cost-aware prediction of uncorrected dram errors in the field," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–15.

[6] P. Radojkovic, M. Marazakis, P. Carpenter, R. Jeyapaul, D. Gizopoulos, M. Schulz, A. Armejach, E. Ayguade, F. Bodin, R. Canal, F. Cappello, F. Chaix, G. Colin de Verdiere, S. Derradji, S. Di Carlo, C. Engelmann, I. Laguna, M. Moreto, O. Mutlu, L. Papadopoulos, O. Perks, M. Ploumidis, B. Salami, Y. Sazeides, D. Soudris, Y. Sourdis, P. Stenstrom, S. Thibault, W. Toms, and O. Unsal, "Towards Resilient EU HPC Systems: A Blueprint." European HPC resilience initiative. White paper, April 2020. [Online]. Available: https://resilienthpc.eu/blueprint2020

[7] I. Boixaderas, D. Zivanovic, S. Moré, J. Bartolome, D. Vicente, M. Casas, P. M. Carpenter, P. Radojković, and E. Ayguadé, "UEPREDICT: A method for predicting DRAM Uncorrected Errors and evaluating its model's performance," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3872777

[8] Barcelona Supercomputing Center, *MareNostrum 3 User's Guide*, Apr. 2016.

**Isaac Boixaderas** is a Research Engineer at Barcelona Supercomputing Center (BSC). He received his BSc degree in Computer Science from Universitat Politècnica de Catalunya (UPC) in 2016. Currently, he is pursuing a MSc in Data Science at Universitat Oberta de Catalunya (UOC). Prior to starting his career as a researcher, he worked as a Software Engineer at Universitat Internacional de Catalunya (UIC) and Inbenta Holdings Inc. in Barcelona.

# High-Resolution Decadal Prediction - Impacts on the predictability of the Pacific variability

Aude Carréric*, Pablo Ortega*
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {aude.carreric, pablo.ortega}@bsc.es

***Keywords—Keywords: Decadal Prediction, High-Resolution, EC-Earth coupled global climate model, Climate variability.***

## I. Extended Abstract

Decadal prediction is a relatively recent field of research, attracting growing interest beyond the scientific community due to its strong potential to provide key information for decision making in economic sectors (e.g. energy production, agriculture, insurance) in a context of pressing danger of climate change. Decadal climate prediction (DCP) skill can arise from two major sources. The first is related to the external radiative forcings (such as volcanic eruptions, solar activity or the anthropogenic greenhouse gases), whose past variations have caused important climate trends in recent decades. And the second is the internal low-frequency variability, usually associated with oceanic processes operating at decadal and multi-decadal timescales. The premise of DCP is that such internal variability processes, when adequately modelled and initialized, can improve our predictive capacity not only on the oceans but also over the surrounding land areas, such as in the North Atlantic region [1].

However, one major limitation common to current DCP systems is the little skill that they present over the continents, which appears to be connected to an incorrect representation of the teleconnection mechanisms that, mediated via the atmosphere, connect the ocean with the neighbouring continents. There are several indications that the current generation of models at standard resolution misrepresents those key teleconnections, and that higher resolution versions might improve them, decreasing common biases of global models and improving some regional seasonal prediction skills, e.g. in tropical sea surface temperature [2]. For decadal prediction, it is still unclear if similar improvements can be achieved through increased resolution, as these systems involve many more simulation years than the seasonal ones, which have made them computationally unaffordable until now.

In this study, we explore how the forecast skill of the DCP can be improved by increasing the spatial resolution of the model. A specific focus on ENSO predictive skill and its associated climate teleconnections will be given to investigate the predictability of the Pacific Ocean, given the promising results of the resolution of oceanic eddies and therefore their effect on the ocean variability [3].

### A. Climate model

The experiments will be run with the coupled global climate model EC-Earth v3.3, using its HR configuration [4]. The specific atmosphere-ocean grids are T511-ORCA025, which corresponds to a resolution of approximately 39 km in the atmosphere and 25 km in the ocean. As mentioned above, increasing the horizontal resolution of the model leads to a better representation of previously unresolved processes (e.g. ocean eddies) that are important for ocean-atmosphere interaction (Figure 1). We can expect to better reproduce both the climate mean state and its variability.

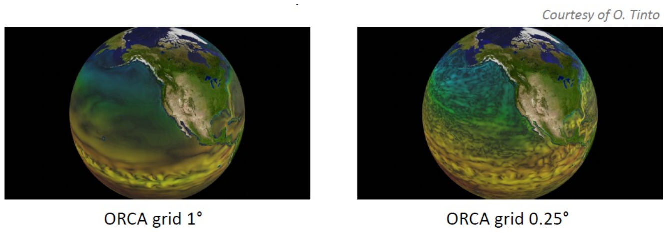


Fig. 1. Different representation of the oceanic circulation depending on the resolution of the model with (left) ORCA 1 degree, (right) ORCA 0.25 degree.

### B. Methodolody

The ability of climate models to make accurate predictions is usually tested by performing retrospective predictions or hindcasts. These are ensembles of predictions with forecast horizons from months to up to ten years that are started from different past initial states (or start dates). The set of initial states is equi-probable and aim to represent best estimates of the observational uncertainty. The ensemble of past initial conditions is generated by introducing random perturbations in the temperature fields of both the ocean and the atmospheric initial conditions This process is repeated with start dates that sample different years (ideally all years if computing resources allow for it) covering the last few decades until present. By comparing these with observations from the same period, we can then produce an assessment of the forecast quality at different forecast horizons.

Each retrospective forecast will consist of 10 ensemble members, each of them 10-year long, initialised once every 5 years, on the 1st of November over the period 1960-2010 (the last initialised ensemble will start in 2010 and will last 10 years to cover the recent decade of observations), which corresponds to 550 years of simulations.

### C. Results

One particularly interesting feature of this HR model version is the improvement in the simulation of the deep

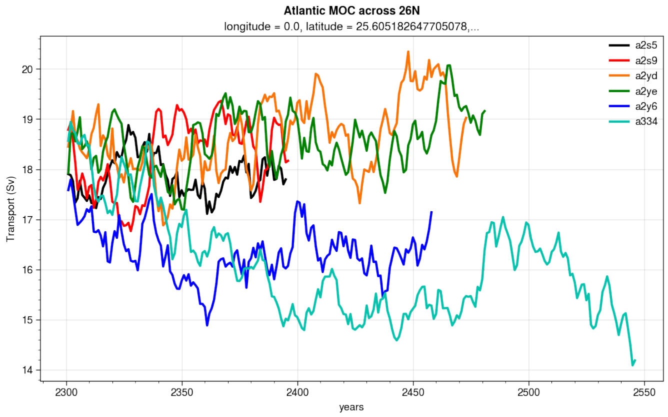


Fig. 2. Strength of the AMOC at 26N for different tuning experiments at HR compared to the SR simulation (light blue line).

convection in the Labrador Sea and the Atlantic Meridional Overturning Circulation (AMOC), compared to the standard resolution version (SR, of approximately 100 kms in both the atmosphere and the ocean). In a decadal prediction system based on EC-Earth3.3-SR, all start dates show a consistent collapse of the Labrador Sea convection [5]. This collapse is caused by the model drift towards its preferred mean state, a drift that induces a quick decrease in the predictive skill in the Subpolar North Atlantic, a source region of decadal variability and predictability [6]. Some initial tests show that this problem is not present in EC-Earth3.3-HR, for which the Labrador Sea convection remains active and stable all along the simulations, directly impacting the strength of the AMOC (Figure 2).

### D. Conclusion and perspectives

Different tests are currently being conducted, both on the observation products used to represent the initial state for the hindcast initialisation and on the tuning of the HR version of the model. Further efforts are also required for performing the HR decadal prediction system, computationally expensive, to save CPU hours via an optimal energy-to-solution configuration of the different components of the model.

The analysis of skill (evaluation against observations) and reliability (characterization of uncertainty) of the HR DCP system will then be performed to (1) assess the impact of the increase of the horizontal resolution on the prediction quality of the model and (2) investigate the predictability of the Pacific Ocean.

## II. Acknowledgment

We are grateful for the contributions of Francisco J. Doblas-Reyes*, Markus G. Donat*, Vladimir Lapin*, Etienne Tourigny*, Miguel Castrillo*, Juan C. Acosta-Navarro*, Arthur Amaral-Ramos*, Pierre-Antoine Bretonnière*, Valentina Sicardi*, Roberto Bilbao*, Javier Vegas*, Saskia Loosveldt*, An-Chi Ho* and Núria Perez Zanon*.

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020- MSCA-COFUND-75433. The EUCP project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no 776613.

## References

[1] D. Smith *et al.*, “Robust skill of decadal climate predictions,” *npj Clim Atmos Sci*, 2019.

[2] C. Prodhomme *et al.*, “Benefits of Increasing the Model Resolution for the Seasonal Forecast Quality in EC-Earth,” *J Clim*, 2016.

[3] R. Haarsma *et al.*, “HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR – description, model computational performance and basic validation,” *Geosci. Model Dev.*, 2020.

[4] ——, “Sensitivity of winter North Atlantic-European climate to resolved atmosphere and ocean dynamics,” *Sci Rep*, 2019.

[5] R. Bilbao *et al.*, “Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth,” *Earth Syst. Dynam.*, 2021.

[6] D. Smith *et al.*, “North Atlantic climate far more predictable than models imply,” *Nature*, 2020.

**Aude Carréric** was born in Hennebont, France, in 1984. She received an engineering diploma in Hydraulics and Fluid Mechanics from Toulouse INP-ENSEEIHT (France) in 2007. She subsequently completed her M.Sc degree in Climate Sciences in 2015 and received her PhD in Oceanography in 2019 from the University of Toulouse III - Paul Sabatier (France). She is currently working at the Barcelona Supercomputing Center (BSC), within the Climate Prediction (CP) Group of the Earth Sciences Department.

# Algebraic Linelet Preconditioner for the solution of the Poisson equation on boundary layer flows.

Ramiro de Olazábal[*][†], Oriol Lehmkuhl[*], Ricard Borrell[*]
[*]Barcelona Supercomputing Center, Barcelona, Spain
[†]Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: ramiro.deolazabal@bsc.es,oriol.lehmkuhl@bsc.es,ricard.borrell@bsc.es

***Keywords—Preconditioned Conjugate Gradient, Krylov Methods, Parallel Algebraic Linelet Preconditioner, Boundary Layer Flow.***

## I. Abstract

An algebraic linelet preconditioner is presented, which works in parallel regardless of the mesh's geometry and without imposing constraints on the domain partition. It is designed to deal with highly anisotropic meshes. A key aspect of this work is developing an algorithm that generates the preconditioning matrix by purely algebraic considerations. This preconditioned is coupled to Alya, the in-house HPC multi-physics code developed at Barcelona Supercomputing Center.

## II. Extended Abstract

Navier-stokes equations for incompressible flows can be solved using the fractional step projection method, by which the pressure solution is decoupled from the rest of the equations. In this context, a Poisson's equation for the pressure correction equation needs to be solved at least once per time-step on this scheme [1]. This step represents the primary source of performance bottlenecks of the code and is one of the most time-consuming and difficult to parallelize. Furthermore, when problems involving boundary layer flow need to be simulated, highly anisotropic meshes are employed to accurately describe this critical region. Such compression of the mesh degrades the conditioning of the linear system associated with the Poisson equation. This makes the solver even more expensive in terms of time and computational resources employed. Figure 2 shows an example of a mesh for the numerical simulation of an airplane wing; here, the nodes located in the prismatic boundary layer represent the $46.7\%$ of the total nodes. Thus its impact on Poisson's equation discretization is not negligible.

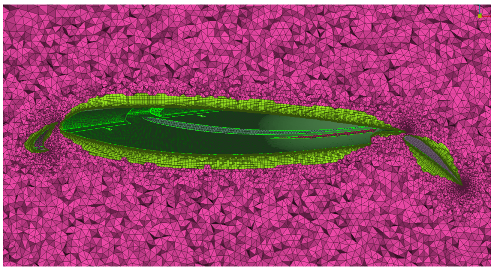


Fig. 1. Example of a mesh for the numerical simulation of an airplane wing.

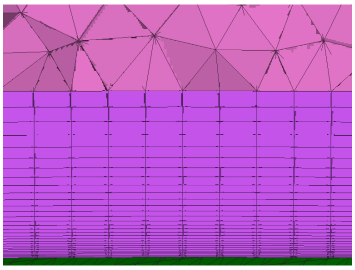


Fig. 2. Zoom of the mesh shown in figure 1 showing the prismatic boundary layer.

Motivated by these facts, we developed a parallel preconditioner able to boost the performance of the Preconditioned Conjugate Gradient (PCG) solver for meshes with high anisotropy in the boundary layer. In particular, we aimed to extend the capabilities of the linelet preconditioner [2]. This kind of preconditioner considers only the two strongest couplings for each node of the boundary layer. As a result, the system is decomposed into a set of one-dimensional tridiagonal subsystems. An algebraic approach independent of the mesh partitioning is presented, which works for Finite Elements (FE), Finite Differences (FD), and Finite Volumes (FV) methods. The authors are not aware of publications with linelet preconditioner implementations of similar capabilities.

## III. Algorithm

The linelet preconditioner can be thought of as drawing lines in the mesh joining nodes with couplings higher than a specific cut-off value, where each node can only belong either to one or no linelet. Then, the preconditioning matrix is built by assembling the system matrix's diagonal entries and the non-diagonal entries of the edges belonging to the linelets. After a proper renumeration of nodes, the preconditioning matrix consists of a set of one-dimensional tridiagonal subsystems. In particular, we propose a fully algebraic approach independent of the mesh partitioning.

The main idea of the algorithm is as follows. The preconditioner gets as input the system matrix $A$, the right hand side $b$, the ID of the global boundary nodes (i.e., the ID of the nodes in the prismatic boundary layer adjacent to the airfoil), and the maximum number of processes we want every linelet to go through (which will be called $N_P$). In the first step, every process builds its linelets following the maximum couplings between nodes. Then via a process of $2(N_P - 1)$ successive

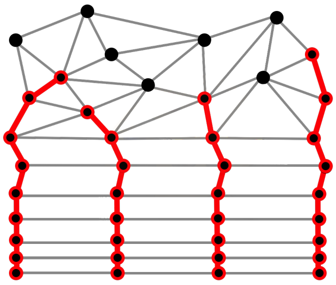


Fig. 3. Example of an anisotropic mesh, linelets are shown in red, joining nodes of high coupling.

halo updates, every subdomain joins each one of its linelets to the corresponding ones built in the rest of the domain. Next, only the diagonal element is preserved for every node that does not belong to any linelet. A communication scheme was also developed in order to deal with the situation of linelets traversing various subdomains.

The final step is to solve the system $Ms = r$ for $s$, for which the TDMA [3] is combined with a Dual Schur Decomposition Method [4]. It is important to note that this operation requires just one communication step to replicate each subdomain's total interface values.

## IV. Preliminary Results

Alya, our in-house HPC multi-physics code developed in our research group, already has an in-built linelet preconditioner to deal with highly anisotropic meshes. Nevertheless, it works by assembling the linelets within each subdomain without allowing communications between them. In this work, a Preconditioned Conjugate Gradient (PCG) solver was developed, which allows the linelets to spread over an arbitrary number of processes.

Comparing both cases mentioned above, it was studied how cutting the linelets affects the convergence of PCG. For this to be done, the global domain was partitioned in the $z$ direction. An example was run for a cubic mesh whose dimensions were $25 \times 25 \times 56$ nodes. The mesh is refined in the z-direction, being the mesh highly compressed near $z = 0$ and with refinement function

$$\Delta z_i = 2H - H\left\{1 + \frac{1}{tanh(\rho)} tanh\left[\rho(1 - \frac{i+i_0}{N_z})\right]\right\},$$

i=1,...,$N_z$.

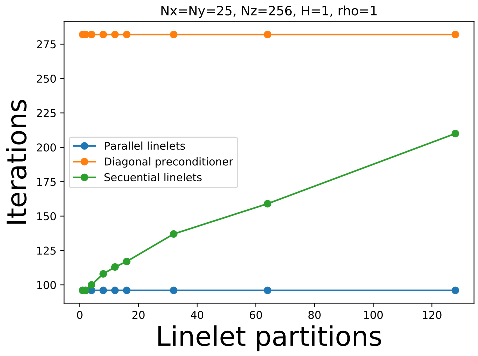


Fig. 4. Number of iterations required to achieve a residual lower than $10^{-6}$ for different number of partitions in the $z$ direction, $\rho = 1$.

Figure 4 shows the number of iterations needed for a different amount of partitions. It can be seen from Figure 4 that if linelets are spread over many processes but there is no communication between its different sections, the solver needs to perform more iterations to reach the same tolerance (green plot in Figure 4). On the other hand, if communications are allowed between different linelets sections, there is no incidence of the partition in the number of iterations required (blue plot in Figure 4). Finally, both previous cases were compared to the diagonal preconditioner(orange plot in Figure 4). Note that when no communications are allowed, the number of iterations needed to converge gets closer to the diagonal case.

## V. Conclusions

To sum up, in this work an algebraic linelet preconditioner is presented, together with all the computational and algorithmic tools required for it to work in parallel regardless of the geometry of the problem and the domain partition. It has been seen that the iterations of the PCG solver can be considerably reduced if a proper communication scheme is developed for the preconditioning stage. This preconditioner will then be coupled to Alya, the in-house HPC multi-physics code developed in our research group. In the final work, the solver will be assessed from the scalability and robustness point of view using meshes similar to those shown in Figure 2, coming from relevant for aeronautical applications.

## References

[1] J. Perot, "An analysis of the fractional step method," *Journal of Computational Physics*, vol. 108, pp. 51–58, 09 1993.

[2] O. Soto *et al.*, "A linelet preconditioner for incompressible flows," *International Journal of Numerical Methods for Heat Fluid Flow*, vol. 13, pp. 133–147, 2003.

[3] R. S. F. Quarteroni, Alfio; Sacco, *Numerical Mathematics*, 2007.

[4] M. Soria *et al.*, "A direct algorithm for the efficient solution of the poisson equations arising in incompressible flow problems," in *Parallel Computational Fluid Dynamics 2001*, P. Wilders *et al.*, Eds. Amsterdam: North-Holland, 2002, pp. 331 – 338. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780444506726500931

**Ramiro de Olazábal** received his M.Sc. in Physics from University of Buenos Aires (UBA), Argentina, in 2017. The next year, he worked as a freelancer for FRONT, which is a startup created to advise the vast majority of people that do not have advanced financial literacy to self manage their investments or the minimum amount of capital required to hire a traditional financial advisor. He worked together with an interdisciplinary team to develop an algorithm for optimizing investment portfolio when clients deposit or withdraw money in the context of model portfolio theory, from Henry Markowitz. Since 2019, he has been working with CASE group of Barcelona Supercomputing Center (BSC) as well as a PhD student at the department of Applied Mathematics of Universitat Politècnica de Catalunya (UPC), Spain, under the supervision of Ricard Borrell and Oriol Lehmkhul.

# Optimizing Execution on Large-scale Infrastructures by Integrating Task-based workflows and MPI

Hatem Elshazly*†,Francesc Lordan*†,Jorge Ejarque*†, Rosa M. Badia*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {hatem.elshazly, francesc.lordan, jorge.ejarque, rosa.m.badia}@bsc.es

***Index Terms*—Hybrid Programming Models, MPI, Task-based Parallel Programming Models, Performance, Productivity, High Performance Computing**

## I. Extended Abstract

While MPI [1] + X (where X is another parallel programming model) has been proposed and used by the community, we propose a hybrid programming model that combines task-based model + MPI. Task-based workflows offer the necessary abstraction to simplify the application development for large scale execution, and supporting tasks that launch MPI executions enables to exploit the performance capabilities of many-core systems. Hence, application programmers can get the maximum performance out of the underlying systems without compromising the programmability of the application.

We present an extension to PyCOMPSs framework [2], a task-based parallel programming model for the execution of Python applications. Throughout this paper, we name the tasks that natively execute MPI code as *Native MPI Tasks*, as opposed to tasks that call external MPI binaries. Having *Native MPI* tasks as part of the programming model means that in the same source file users can have two types of task: tasks that execute MPI code and other tasks that execute non-MPI code. PyCOMPSs organizes the tasks in Directed Acyclic Graph (DAG) and manages their scheduling and execution, hence users can focus only on the logic of the task.

### A. Native MPI in PyCOMPSs

Tasks are defined in PyCOMPSs by annotating application's method with Python decorators. Through the @task annotation, developers indicate that a function in the code becomes a task. Following the same approach, a method is declared as *Native MPI* task by means of the @mpi decorator. The number of MPI processes per *Native MPI* task can be specified using @constraints decorator as shown in the sample code snippet in Figure 1.

PyCOMPSs runtime will manage the input and output data of *Native MPI* tasks like any non-MPI task in a completely transparent manner to the user. The runtime will ensure that all the processes in the MPI environment have access to all the input data of the task. The return output of a *Native MPI* task – if any – is a list containing the output of all the MPI processes invoked for the task.

```
@constraints(computingUnits=4)
@mpi(runner='mpirun', computingNodes=1)
@task(returns=int)
def return_ranks(random_num):
    from mpi4py import MPI
    rank = MPI.COMM_WORLD.rank
    return rank*random_num
```

Fig. 1. Simple *Native MPI* task in PyCOMPSs. return_ranks task will be executed by 4 MPI processes as specified in computingUnits on 1 node. It returns a list of each MPI rank multiplied by the random_num input value.

Similar to non-MPI PyCOMPSs tasks, the execution details of *Native MPI* tasks are completely abstracted from the runtime; the MPI environment is encapsulated within the *Native MPI* task that launched it. Thus, one workflow can have multiple *Native MPI* tasks, each with different configuration parameters (i.e., number of computing nodes and MPI processes) and combine them with other tasks in the task execution graph.

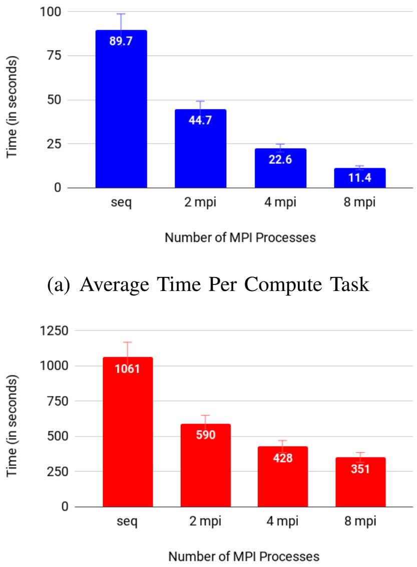
PyCOMPSs runtime launches special Python worker processes for *Native MPI* tasks at the time of the task execution to launch the MPI environment and manage the task execution. If two *Native MPI* tasks are scheduled for execution at the same time, the runtime launches an exclusive MPI worker for each of them. Hence, each of the tasks will have its own isolated execution environment.

### B. Evaluation

In this section, we evaluate performance benefits and trade-offs of using *Native MPI* tasks in PyCOMPSs. Experiments were conducted on the MareNostrum 4 supercomputer; which includes a set of high-memory computing nodes with 48 cores and 370 GB of memory each. Each experiment was run multiple times: using sequential implementation of the targeted tasks and a parallel implementation with an increasing number of MPI processes (2, 4 and 8). In all experiments, the sequential implementation of the task is used as the baseline.

For the purpose of this evaluation, we developed an application that calculates the term frequency (TF-IDF) of a web archive file. We used an input web archive file of a total size of 186 Gbytes. The application consists of a reading task which reads a record from the file and a compute task that calculates TF-IDF. The total number of tasks for this application is 1440 tasks; 720 read tasks and 720 corresponding compute tasks.

Figure 2 shows the performance results of the application. As shown in Figure 2(a) the average time per compute task decreases while increasing the number of MPI processes per compute task. Using 8 MPI processes per compute task, we obtained up to 7x speedup in the average time per compute task. In addition to that, as shown in Figure 2(b), the performance improvement per compute task is reflected as up to 3x speedup improvement in the total execution time.



(a) Average Time Per Compute Task

(b) Total Execution Time

Fig. 2. Performance Results for Web Archive Analysis Application

To further understand the performance and behaviour of *Native Python MPI* tasks in PyCOMPSs, several experiments were conducted on the Web Archive Analysis. Each experiment was launched multiple times with a sequential implementation task and then a parallel *Native MPI* task implementation with different numbers of MPI processes (2, 4, 8, 16 and 48) on different number of nodes (4, 8 and 12).

As shown in Figure 3, as the number of nodes increases, task parallelism increases so the total execution time of both applications improves. For a specific number of nodes, total execution time decreases until it reaches a point after which it starts to increase as the number of MPI processes per *Native MPI* task increases. This point is 8 MPI processes for 4, 8 nodes and 16 MPI processes for 12 nodes. This is because *Native Python MPI* tasks use the @constraint decorator of PyCOMPSs to specify the number of MPI processes per task. Increasing the number of MPI processes per task (i.e. increasing task constraints) decreases task parallelism. This effect is mitigated as the number of resources increases because there are enough resources to maintain the same level or allow for more task parallelism. This can be noted in Figure 3 where for 4 and 8 nodes the total execution time degrades at 8 MPI processes but when the number of nodes is increased to 12, this point shifts to 16 MPI processes.

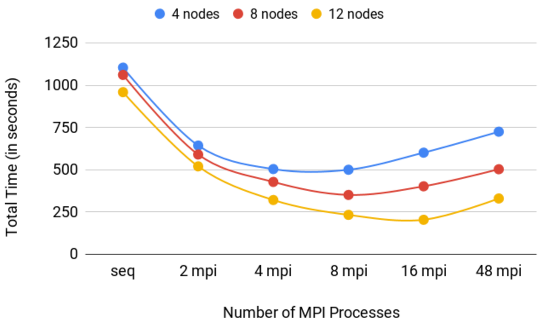


Fig. 3. Scalability Results

### C. *Conclusion*

Enabling the execution of MPI code natively in PyCOMPSs tasks offers great benefits in terms of both programmability and performance for Python applications. However, a tradeoff arises between MPI parallelism per task and task parallelism that may negatively affect the total time of the application. As future work, we plan to improve the scheduling of tasks to better utilize the underlying infrastructure.

## II. Acknowledgment

This work has been published in the 28th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), 2020 [3].

## References

[1] W. Gropp, E. Lusk, and A. Skjellum, *Using MPI: portable parallel programming with the message-passing interface*. MIT press, 1999, vol. 1.

[2] E. Tejedor, Y. Becerra, G. Alomar, A. Queralt, R. M. Badia, J. Torres, T. Cortes, and J. Labarta, "Pycompss: Parallel computational workflows in python," *International Journal of High Performance Computing Applications*, 2015.

[3] H. Elshazly, F. Lordan, J. Ejarque, and R. M. Badia, "*Performance Meets Programmability: Enabling Native Python MPI In PyCOMPSs*," in *28th Euromicro International Conference on Parallel, Distributed and Network-based Processing*, 2020.

**Hatem Elshazly** received his BSc degree in Computer Science from Cairo University, Egypt in 2012. The following year, he joined as a Masters student and research software engineer at Nile University, Egypt. He completed his MSc degree in Optimizing Data Intensive Applications in 2016. Since 2018, he has been with the workflows and distributed computing group of Barcelona Supercomputing Center (BSC) as well as a PhD student at the department of computer architecture of Universitat Politècnica de Catalunya (UPC), Spain.

# Quantum Singular Value Decomposer

Carlos Bravo-Prieto*†, Diego García-Martín*†‡1, José Ignacio Latorre†§
*Barcelona Supercomputing Center, Barcelona, Spain
†Departament de Física Quàntica i Astrofísica , Universitat de Barcelona, Barcelona, Spain
‡Instituto de Física Teórica, UAM-CSIC, Madrid, Spain
§Center for Quantum Technologies, National University of Singapore, Singapore
E-mail: 1diego.garcia@bsc.es

***Keywords*—Quantum computing, Quantum algorithms, Quantum entanglement.**

## I. Extended Abstract

### A. Context

Quantum computing is emerging as a new computational paradigm that promises substantial speed-ups (exponential in some cases [1]) for a variety of problems, with a potentially-large impact on science, knowledge and society. Quantum computers will harness the strange-looking laws of Quantum Mechanics to accelerate many computations by orders of magnitude [2]. However, they will require error-correction protocols before many important algorithms can be run on these machines. These will demand a large number of high-quality qubits, which imply a much longer time of arrival for this technology, if it is to be "fully-fledged" and incorporate these schemes.

For this reason, researchers have begun to heavily investigate about possible uses of the first generations of quantum computers that will not incorporate error correction but that nonetheless may surpass the capabilities of even the largest classical supercomputers in the world. These computers have been termed Noisy Intermediate-Scale Quantum (NISQ) computers [3]. A particularly appealing choice of quantum programs to be run on these machines are the so-called variational quantum circuits, or Quantum Neural Networks (QNN). These are hybrid algorithms that will coordinate both a classical and a quantum computer to preform a computation, and need to be specifically designed for each application.

### B. Quantum Singular Value Decomposer

In such a context, we present a variational quantum circuit that produces the Singular Value Decomposition or Schmidt decomposition of a bipartite pure state [4]. This is important for understanding the entanglement structure of quantum systems, e.g. in Condensed-Matter Physics or in Quantum Chemistry. Moreover, the algorithm is also useful to perform Principal Component Analysis (PCA) and, more generally, to diagonalize a given matrix.

The proposed circuit, that we name Quantum Singular Value Decomposer or QSVD, is made of two unitaries respectively acting on each part of the system. The key idea of the algorithm is to train this circuit so that the final state displays exact output coincidence from both subsystems for every measurement in the computational basis, see Fig. 1. Such circuit preserves entanglement between the parties and acts as a diagonalizer that delivers the eigenvalues of the Schmidt decomposition. Our algorithm only requires measurements in one single setting, in striking contrast to the $3^n$ settings required by state tomography. Furthermore, the adjoints of the unitaries making the circuit are used to create the eigenvectors of the decomposition up to a global phase.

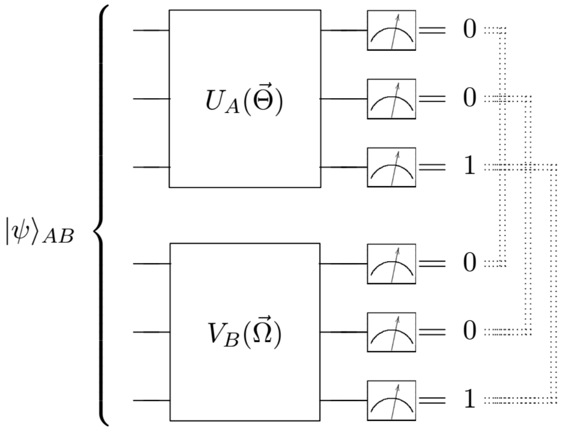


Fig. 1. Variational quantum circuit displaying exact output coincidence, used for the Quantum Singular Value Decomposer.

### C. Further applications

Some further applications of QSVD are readily obtained. The proposed QSVD circuit allows to construct a SWAP

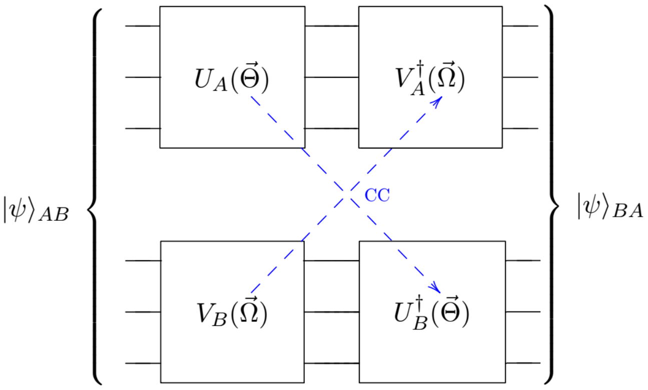


Fig. 2. An extension of the QSVD (CC) produces a long-distance SWAP between parties using only classical communication.

between the two parties of the system without the need of any quantum gate communicating them, see Fig. 2. We also show that a circuit made with QSVD and CNOTs acts as an encoder of information, compressing the original state onto one of its parties, see Fig. 3. This idea can be reversed and used to create random states with a precise entanglement structure.

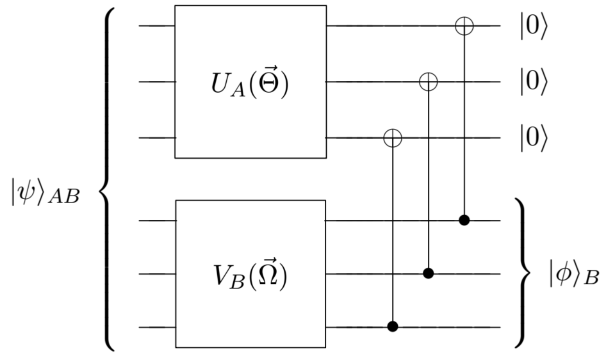


Fig. 3. Another simple extension of the QSVD converts it into a compressor of quantum information.

### D. Conclusions

We have proposed a novel hybrid quantum-classical algorithm to compute the Schmidt decomposition of a pure bipartite quantum state. This algorithm can be used to study the entanglement structure of quantum systems, e.g. in Condensed-Matter Physics or in Quantum Chemistry, to perform Principal Component Analysis or to diagonalize a given matrix, and entails an exponential reduction in the number of measurements settings as compared to state tomography. Moreover, it will likely be implementable on Noisy Intermediate-Scale Quantum computers in the coming years.

## II. Acknowledgment

This work has been published in Physical Review A 101 (6), 062310 [4].

## References

[1] P. W. Shor, “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer,” *SIAM Journal on Computing*, vol. 26, no. 5, p. 1484–1509, 1997.

[2] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information: 10th Anniversary Edition*, 10th ed. USA: Cambridge University Press, 2011.

[3] J. Preskill, “Quantum computing in the nisq era and beyond,” *Quantum*, vol. 2, p. 79, 2018.

[4] C. Bravo-Prieto *et al.*, “Quantum singular value decomposer,” *Phys. Rev. A*, vol. 101, p. 062310, 2020.

**Diego García-Martín** received a BSc degree in Biology from Universidad Autónoma de Madrid (UAM), Spain in 2013 and a BSc degree in Physics from Universidad de La Laguna (ULL), Spain in 2016. He then obtained a MSc degree in Theoretical Physics from Universidad Autónoma de Madrid (UAM), Spain in 2017. The following year he started his Ph.D. under the supervision of Germán Sierra and José Ignacio Latorre, and in 2019 he joined Barcelona Supercomputing Center (BSC) as a Junior Research Engineer to continue with his Ph.D.

# Optimizing the SpMV kernel on long-vector accelerators

Constantino Gomez[*†], Filippo Mantovani[*†], Erich Focht[‡], Marc Casas[*†]
[*]Barcelona Supercomputing Center, Barcelona, Spain
[†]Universitat Politècnica de Catalunya, Barcelona, Spain
[‡]NEC HPC Europe, Stuttgart , Germany
E-mail: {constantino.gomez, marc.casas, filippo.mantovani,}@bsc.es; erich.focht@emea.nec.com

***Keywords*—*SpMV, NEC Vector Engine, Long-Vector Architectures, Performance Optimization***

## I. Extended Abstract

Sparse Matrix-Vector multiplication (SpMV) is an essential kernel for parallel numerical applications. SpMV displays sparse and irregular data accesses, which complicate its vectorization. Such difficulties make SpMV to frequently experiment non-optimal results when run on long vector ISAs exploiting SIMD parallelism. In this context, the development of new optimizations becomes fundamental to enable high performance SpMV executions on emerging long vector architectures. In our work, we improve the state-of-the-art SELL-$C$-$\sigma$ sparse matrix format by proposing several new optimizations for SpMV. We target aggressive long vector architectures like the NEC Vector Engine. By combining several optimizations, we obtain an average 12% improvement over SELL-$C$-$\sigma$ considering a heterogeneous set of 24 matrices. Our optimizations boost performance in long vector architectures since they expose a high degree of SIMD parallelism.

### A. Background

Many different approaches have been proposed to efficiently store sparse matrices and efficiently run SpMV. One of the most common approaches, Compressed Sparse-Row (CSR), efficiently stores sparse matrices and enables simple stride-1 memory access patterns on $A$ and $y$. However, accesses on $x$ are highly irregular. Other approaches aim to mitigate the drawbacks of CSR by enlarging its storage requirements to increase the locality on $x$. SELL-$C$-$\sigma$ [1] and ELLPACK Sparse Block [2] make use of row sorting and column blocking to improve both storage requirements and locality on $x$. Our work demonstrates that, although some of these approaches are very good abstractions to represent and manipulate sparse matrices, there are many unexploited opportunities to improve their performance on long vector architectures. For that, we implement, evaluate, and discuss the performance impact of several SpMV optimizations on the VE.

### B. Optimizations

We revisit and adapt some optimizations previously proposed in the literature extending them with new approaches targeting long vector architectures.

In detail, we explore: *i)* the adequate sorting strategy based on the trade-off between performance and preprocessing overhead as the σ parameter increases; *ii)* the use of task-based parallelism and the impact of the task granularity in the scaling performance of SELL-$C$-$\sigma$; and *iii)* the impact of column blocking in matrices to improve locality on vector $x$.

In addition, our proposals to accelerate SpMV on long vector architectures are: *i)* the use of cache allocation to improve the reuse of $x$ and deprioritization of store dependencies; *ii)* divergence flow control adapting the length of vector operations to avoid loading and computing *zero-padded* elements; *iii)* enabling loop unrolling in SELL-$C$-$\sigma$ using partial loop fusion; *iv)* efficient computation of gather and scatter addresses with special instructions.

TABLE I: Optimizations applied on each of the implementations evaluated in our work.

| Optimization | SELLCS | SELLCS<br>DFC | SELLCS<br>U8-DFC | SELLCS<br>U8-NC | SELLCS<br>U8-NC-DFC |
|---|---|---|---|---|---|
| Sorting strategies | ● | ● | ● | ● | ● |
| Task-Based Parallelism | ● | ● | ● | ● | ● |
| Column Blocking | | | | | |
| Cache Allocation & Store relaxation pol. | | | | ● | ● |
| Divergent Flow Control | | ● | ● | | ● |
| Loop unrolling | | | ● | ● | ● |
| Efficient gather/scatter address computation | ● | ● | ● | ● | ● |

### C. Results

The test-bench for our experiments is the NEC Vector Engine 10B. Figure 1 shows GFLOP/s performance results for a wide set of matrices. We evaluate six different implementations of the SpMV kernel: *NLC, SELLCS, SELLCS-DFC, SELLCS-U8-DFC, SELLCS-U8-NC* and *SELLCS-U8-NC-DFC*. The *NLC* category represents results obtained with the math library developed by NEC which is particularly tailored for the VE.

The improvements added by the three main optimizations visualized vary across the different matrices, as its effectiveness depends on specific matrix layout characteristics like size, sparsity or shape.

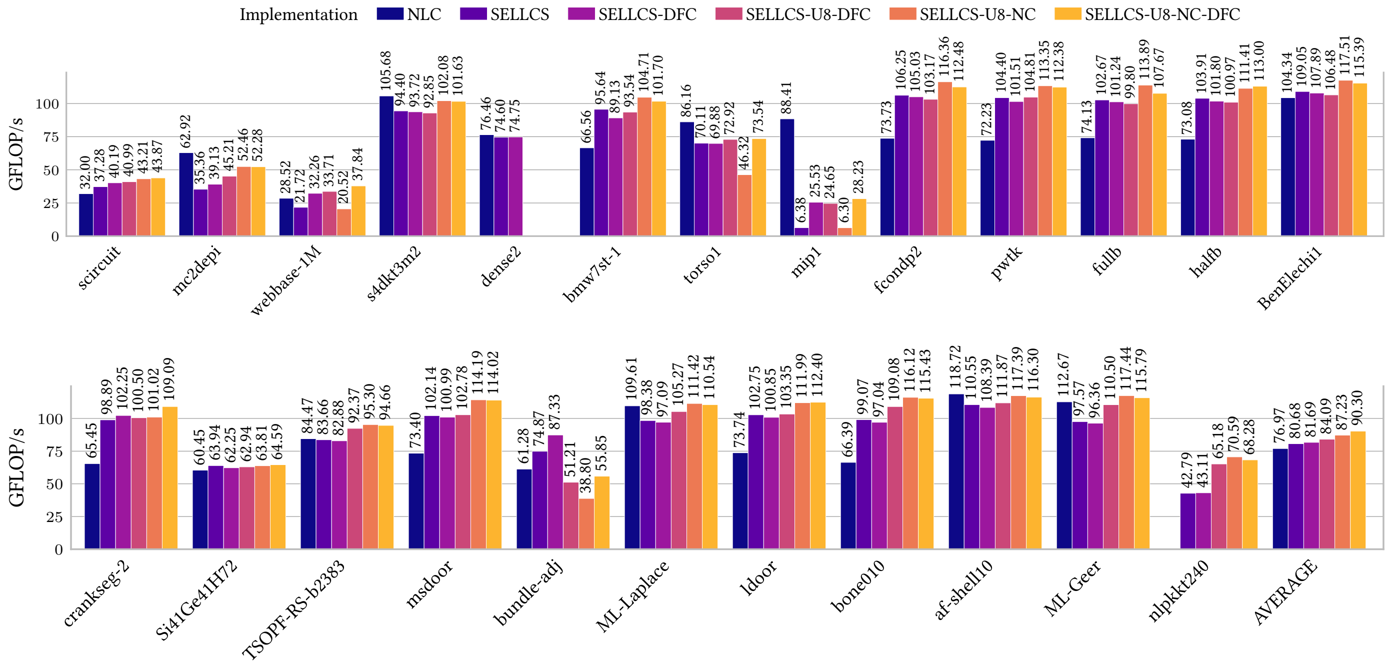


Fig. 1: Performance comparison of NLC vs our SELL-$C$-$\sigma$ implementations for regular matrices.

In short, we can draw the following insight from Figure 1. The cache allocation and store relaxation policies obtain improvements ranging between 5% to 12%, in two thirds of the matrices, when using *SELLCS-U8-NC-DFC* compared to *SELLCS-U8-DFC*. Moreover, unrolling by 8 slices yields, in general, gives benefits ranging from 1% to 15%, with a favorable trend for bigger matrices. In the particular case of nlpkkt240 it brings a 51% performance increase. Finally, to understand the impact of the *DFC* optimization, we compare the performance of *SELLCS* with *SELLCS-DFC*. These two implementations only differ in the use of the *DFC* optimization. Only the second one includes it. On average, the overall performance gains of adapting each vector length instruction to the optimal size are almost negligible. However, it has a large impact in some scenarios. For example, when considering webbase-1M, which represents a website connectivity matrix and has a very low non-zero element density, *SELLCS-DFC* is 50% faster than *SELLCS*.

We obtain in average 90.3 GFLOPs across all matrices by enabling all optimizations, which constitutes a significant improvement of $\sim$12% and $\sim$17% compared to the baseline SELL-$C$-$\sigma$ and NEC math library implementations, respectively. The significant performance increase that we obtain over the NEC proprietary software, which is specially tailored to SX-Aurora VE, demonstrates the relevance of our optimizations in long vector architectures.

### *D. Conclusion*

In this work, we developed an implementation of SpMV for the SX-Aurora long vector architecture shows very competitive performance results which mostly overtake the highly optimized proprietary vendor implementation found in the NEC Library Collection. Additionally, we explore a set of optimizations targeting long-vector accelerators, and provide insight on how to exploit them in similar platforms.

## II. Acknowledgment

This work has been published as a conference paper is published in the proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming[3].

## References

[1] M. Kreutzer *et al.*, “A Unified Sparse Matrix Data Format for Efficient General Sparse Matrix-Vector Multiplication on Modern Processors with Wide SIMD Units,” *SIAM Journal on Scientific Computing*, vol. 36, no. 5, pp. C401–C423, Jan. 2014.

[2] X. Liu *et al.*, “Efficient sparse matrix-vector multiplication on x86-based many-core processors,” in *Proceedings of the 27th international ACM conference on International conference on supercomputing*. Eugene, Oregon, USA: Association for Computing Machinery, Jun. 2013, pp. 273–282.

[3] C. Gómez *et al.*, “Efficiently running spmv on long vector architectures,” in *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 292–303. [Online]. Available: https://doi.org/10.1145/3437801.3441592

**Constantino Gómez** is a last year Ph.D student at the Barcelona Supercomputing Center. He received the BSc and MSc degrees in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 2014 and 2016. He has been involved as a researcher in the European Processor Initiative since the beginning. His research interests include long vector architectures and co-design for future massively parallel systems.

# Super-resolution for downscaling climate data

Carlos Gómez-Gonzalez*, Kim Serradell*
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {carlos.gomez, kim.serradell}@bsc.es

***Keywords*—Deep learning, Downscaling, Climate data.**

## I. Extended Abstract

A common task in Earth Sciences is to infer climate information at local and regional scales from global climate models. An alternative to running expensive numerical models at high resolution is to use statistical downscaling techniques. Statistical downscaling aims at learning empirical links between the large-scale and local-scale climate, i.e., a mapping from a low-resolution gridded variable to a higher-resolution grid that incorporates observational data.

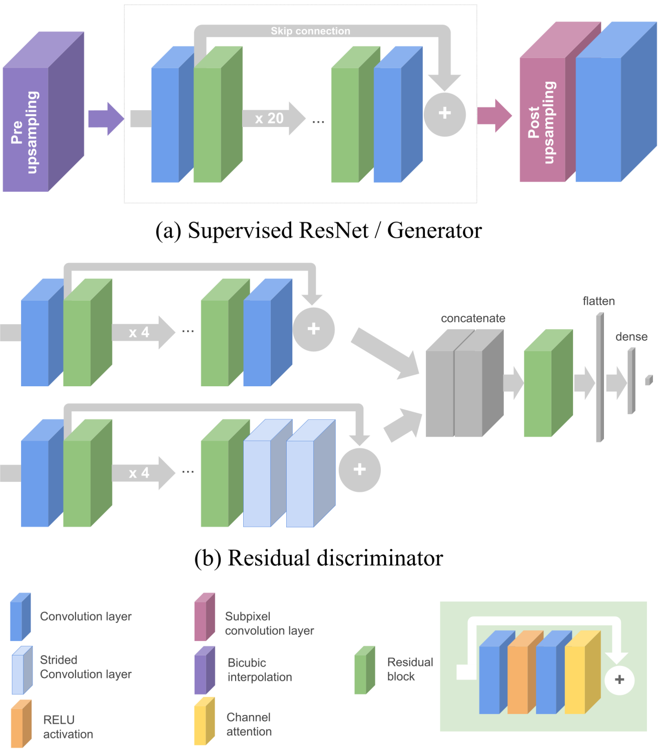
Seasonal climate predictions can forecast the climate variability up to several months ahead and support a wide range of societal activities. The coarse spatial resolution of seasonal forecasts needs to be downscaled or refined to the local scale for specific applications.

In this study, we present super-resolution (SR) techniques for the task of downscaling climate variables with a focus on temperature over Catalonia. Our models are trained using high and medium resolution ($\sim$5 and $\sim$25 km) gridded climate datasets with the ultimate goal of increasing the resolution of coarse resolution ($\sim$100 km) seasonal forecasting systems. Taking the gridded data from $\sim$100 to $\sim$5 km implies a 20x upscaling factor. It is worth pointing out that handling such large upsampling factor is not typical in computer vision, where most applications focus in 4x factors while 16x is considered as extreme SR.

### A. *Super-resolution for statistical downscaling*

Statistical downscaling of gridded climate variables is a task closely related to that of SR in computer vision, considering that both aim at learning a mapping between low- and high-resolution images. Unsurprisingly, several deep learning-based approaches have been explored by the climate community in recent years [1], [2].

For this study, we work with data from two reanalysis datasets: ERA5 [3], produced by the European Centre for Medium-range Weather Forecasts (ECMWF), and UERRA MESCAN-SURFEX [4]. ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables at a resolution of 0.25° ($\sim$25 km). UERRA MESCAN-SURFEX provides temperature, precipitation and wind at a resolution of 0.05° ($\sim$5 km). We focus on temperature at two meters above the ground from both ERA5 and UERRA, selecting the period between 1979 and 2019 at a daily temporal resolution, resulting in about 14k temporal samples. The inference is performed on the ECMWF's seasonal forecast system, SEAS5 [5] to downscale its coarse temperature grids from $\sim$100 to $\sim$5km resolution in a transfer learning fashion.



(a) Supervised ResNet / Generator

(b) Residual discriminator

Fig. 1. Panel (a) shows the architecture of our SR CGAN generator, while panel (b) shows the architecture of our SR CGAN discriminator. The architecture of the supervised ResNets is the same of the CGAN generators.

### B. *Methods*

Four different deep learning-based methods were implemented for downscaling temperature gridded fields: ResNet-INT, ResNet-SPC, and their conditional adversarial counterparts, CGAN ResNet-INT and CGAN ResNet-SPC. The ResNet-SPC is based on the EDSR [6] SR model, with residual blocks using skip connections and without batch-normalization. On the panel (a) of Fig. 1, we show the architecture of our ResNets. These networks share the main section, inside the dotted-line box, composed of convolutional layers and a stack of twenty residual blocks. The ResNet-INT, short for residual neural network with pre-upsampling via bicubic interpolation, is a model that learns an end-to-end mapping from interpolated LR images to HR images. HR UERRA/ERA5 images are downsampled by a given factor to create LR counterparts. These are then upsampled to match the size of the HR image before entering the network. The ResNet-SPC works in a post-upsampling framework using subpixel convolution layers [6]. This model learns a mapping from LR to HR images, of different sizes, in low-dimensional space and

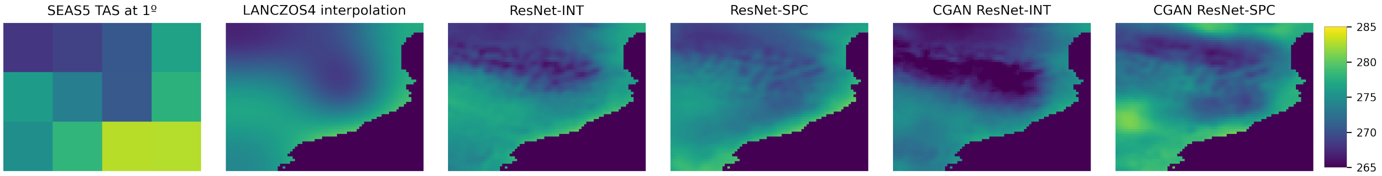


Fig. 2. Comparison of the SR models proposed in this study with respect to a LANCZOS4 interpolation for a single SEAS5 temperature grid.

requiring less computations.

Our CGAN ResNet-INT and CGAN ResNet-SPC are Conditional Generative Adversarial Networks (CGAN) [7] that use either the ResNet-INT or ResNet-SPC as generators. GANs are generative models that rely on a generator that learns to generate new (HR) images (from a LR counterpart), and a discriminator that learns to distinguish synthetic (HR) images from reference (HR) images. CGANs are supervised GANs that are trained with paired samples. We concatenate to all our input samples a topographical map and a land-ocean binary mask, as proposed in [1]. The addition of these fields, as image channels, improves the reconstruction of high-frequency details while downscaling the temperature fields.

To achieve a 20x upscaling factor, our models are composed of a stack of two networks, each one trained separately: LR to MR (4x, using ERA5) and MR to HR (5x, using UERRA). The inference is performed progressively. We tested training single models to jump from LR to HR resolution directly but the results were poor in general. All the networks were trained with sixty-four filters per layer and convolution kernels of size 3x3. The supervised ResNets and the CGAN ResNets were trained for 180 and 60 epochs respectively using the Adam optimizer. The supervised ResNet optimize a mean absolute error (MAE) loss function. A holdout of eight years was used for testing the performance of the trained models. During training, a validation dataset was used to monitor the behavior of the loss function and avoid overfitting.

### C. Results

Figure 2 shows a side-by-side comparison of the four different SR algorithms developed for downscaling SEAS5 temperature from its native 1° to the 0.05° resolution. This temperature grid corresponds to a single date and a single SEAS5 ensemble member. Table I summarizes the performance of each model in terms of the spatial RMSE and Pearson correlation. These metrics are computed per each pair of images: the model prediction and its reference from the holdout UERRA dataset. Based on these metrics and on visual inspection, we argue that the CGAN ResNet-SPC stands out at recovering high-frequency details while downscaling SEAS5 grids not seen during training. Additionally, we have compared our results with those of a traditional technique for statistical downscaling, the KNN-based analogs method. This method delivers higher RMSE and lower correlation, but is on par with the deep learning-based models in terms of the ranked probability skill score, a metric used for validating seasonal forecasts.

TABLE I. VALIDATION METRICS FOR EACH SR METHOD

| CNN model | MSE | Pearson correlation |
|---|---|---|
| ResNet-INT | 0.6379 | 0.9865 |
| ResNet-SPC | 0.5153 | 0.9912 |
| CGAN R-INT | 0.6472 | 0.9860 |
| CGAN R-SPC | **0.4960** | **0.9917** |

### D. Summary

In this study, we developed SR models for the task of downscaling temperature fields and showed their superior performance with respect to a LANCZOS4 interpolation baseline. We thoroughly tested different architecture choices, such as the type of upsampling or the training strategy (adversarial vs non-adversarial). In the future, we will perform more rigurous ablation studies for tuning these networks and explore tailored loss functions (beyond MAE and reconstructive losses) for improving the skill of the seasonal forecasts.

## II. Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 2020 R&D programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433 and the EU H2020 Framework Programme under grant agreement n° GA 823988.

## References

[1] T. Vandal *et al.*, "Deepsd: Generating high resolution climate change projections through single image super-resolution." New York, NY, USA: Association for Computing Machinery, 2017.

[2] J. Leinonen *et al.*, "Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13, 2020.

[3] H. Hersbach *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049.

[4] E. Bazile *et al.*, "Mescan-surfex surface analysis. deliverabled2.8 of the uerra project," 2017. [Online]. Available: http://www.uerra.eu/publications/deliverable-reports.html

[5] S. J. Johnson *et al.*, "Seas5: the new ecmwf seasonal forecast system," *Geoscientific Model Development*, vol. 12, no. 3, pp. 1087–1117, 2019.

[6] B. Lim *et al.*, "Enhanced deep residual networks for single image super-resolution," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul 2017.

[7] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

**Carlos Gómez-Gonzalez** is a STARS (MSCA-COFUND) postdoctoral fellow at the Earth Sciences department of the Barcelona Supercomputing Center (BSC). He holds a Ph.D. in Science from the University of Liège, Belgium. Before joining the BSC, he spent two years as a "research chair in Data Science for Earth and Space sciences" at the Université Grenoble Alpes, France. He joined the Computational Earth Sciences group at the BSC to establish a research line on Artificial Intelligence for Earth Sciences, which involves the development of machine and deep learning techniques for topics, such as statistical downscaling, bias correction techniques, or the study of extreme events.

# Determining the structure of small molecules via their pseudo-electrons and atoms 3D models using FPGA

César González[#1], Simone Balocco[2], Ramon Pons[*3]

[#]*Barcelona Supercomputing Center (BSC), c/Jordi Girona,31 Barcelona (Spain)*
[1]`cesar.gonzalez@bsc.es`, [2]`simone.balocco@ub.edu`

[*]*Institut de Química Avançada de Catalunya CSIC, Jordi Girona 18 Barcelona (Spain)*
[3]`ramon.pons@iqac.csic.es`

***Keywords*— Molecular structures, HPC, FPGA, particle particle-distance, Debye, X-ray spectra**

## EXTENDED ABSTRACT

The particle-pair or particle-particle distance problem (pp-distance) appears in several scientific fields. The pp-distance calculation is a computationally demanding task involved, for instance, in the calculation of X-ray spectra, as shown in Fig-1.

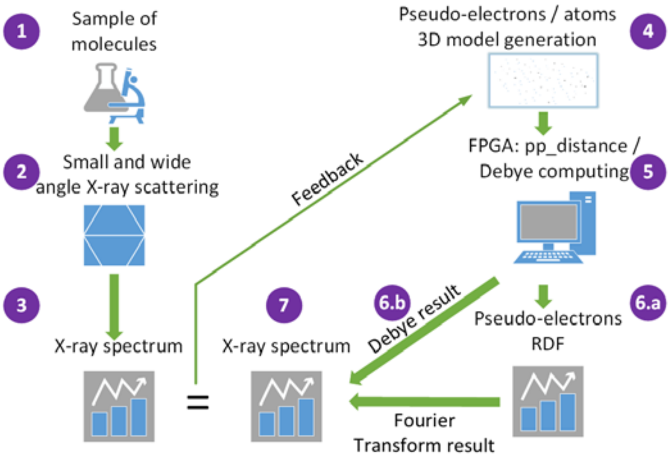


Fig. 1. Full experiment layout.

Protein structure is one of the bases of the biomedicine and nanotechnology. Different methods are used to determine the structures. One of them is X-ray which is usually limited to highly crystalline structures. We propose to extend the method to low crystallinity samples.

We solve the pp-distance problem by calculating the theoretical X-Ray spectra for non-crystalline structures and comparing it against the real one. Our C program reduces the computational time using High Performance Computing over FPGA.

We use pseudo-electrons 3D models to accelerate the convergence of the computed spectrums to the referenced one, (see Fig-1, subplot 6.a). Then, we use atoms 3D models to determine exactly the structure of the sample of molecules using Debye approximation and equations [4], (see Fig-1, subplot 6.b).

### A. *The pseudo-electrons option, fast track to find an approximation to the structure.*

The pp-distance algorithm has been customized for different FPGA with the following results:

- over an Intel D5005 Programmable Acceleration Card, computing a model of 2 million particles in 81.57 seconds, that is, 24.5 billion atoms pairs per second (*bapps*).

  In this case we use OpenCL kernels and OmpSs [2] programming model.

- over a Xilinx Alveo U200 board that computes the same 2 million particles in 34 seconds, that is 31.2 *bapps*.

  In this case we use C language for kernels with OmpSs too and Picos++ runtime manager.

Fig-2 illustrates the accuracy of the 3D model when a random distribution of 2 million pseudo-electrons is considered.

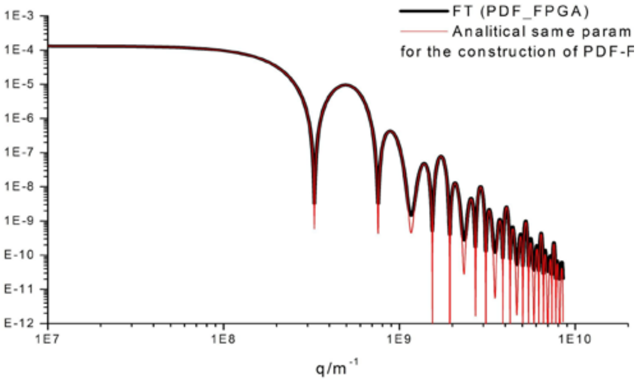


Fig. 2 Scattered intensity (in arbitrary units) as a function of scattering vector modulus.

### B. *The atoms option, the proper track to precise the exact structure.*

Debye [4] computation is used to directly determine the 3D model spectra. The Debye computation includes the previous calculation of the atomic form factors (*aff*) of the atoms at the main program as

$$f(q) = \left( \sum_{i=1}^{4} a_i \, e^{\left(-b_i\left(\frac{q}{4\pi}\right)^2\right)} \right) + c$$

Where $a_i$, $b_i$ and $c$ are coefficients characteristic of each atom type and $q$ is the scattering vector module. See more at http://lampx.tugraz.at/~hadley/ss1/crystaldiffraction/atomicformfactors/formfactors.php and performing the Debye formula calculation with these *aff* at the FPGA kernel

$$I(q) = \sum_{i=1}^{N}\sum_{j=1}^{N} f_i f_j \frac{\sin(q \cdot r_{ij})}{q \cdot r_{ij}}$$

Where *I* is the scattered intensity, $r_{ij}$ is the distance between *i* and *j* atoms with $f_i$ and $f_j$ atomic form factors.

The direct result is the data that could be represented as a plot like that shown in Fig. 3.

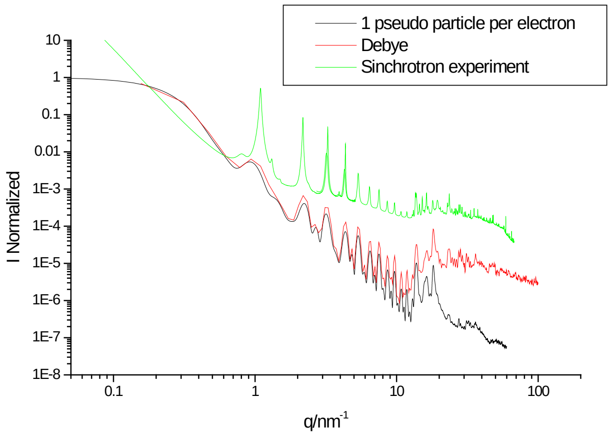


Fig. 3 SAXS-WAXS Scattered intensity of silver behenate crystallites as a function of q compared with atoms versus pseudo-electrons models as described.

In this example we show the spectrum of a model crystallite of Silver Behenate composed by 25x25x3 unit cells with a total size of 10x12x17 nm calculated using Debye procedure (red line) and using the pseudo-electrons option (black line, 1 pseudo-electron per electron). In the same figure, the green line shows the ALBA-Synchrotron experimentally obtained spectra. Note that the peak width is related to crystallite size. Applying Debye-Scherrer formula [5] the size of crystallite of the experiment corresponds to 2 µm while the calculation corresponds to 19 nm.

Further, we have also evaluated the Debye procedure with and without considering the hydrogen atoms of the 3D model as an accelerating procedure. The number of atoms in the first case is 255,000 while in the second case is 93,750 atoms, that represents a calculation acceleration of 87% with the same results.

With Debye option each generation of 64 points in the spectrum graph takes 400 seconds approximately, for a model of 93,750 atoms., the spectra shown in figure 3 took 2,800 s. while for a model of 255,000 took 21,427 s.

The calculation to get the spectrum, in our D5005 FPGA, using 1 pseudo-electron per electron took 18.38 s including the 1.01 s from the Sine Fourier Transform (FT) shown below.

$$\mathrm{P}(q) = 4\pi \int_0^\infty p(r)\frac{\sin(qr)}{qr}\,dr$$

Where *p(r)* is the radial distribution function, *q* the scattering vector modulus and *r* is the radius.

We expect half time, approx. 9 s, at our Alveo U200 board.

### C. *Conclusions and future work*

The high-performance speed of FPGA implementation allows this method to be used to search the structures of small or medium size molecular ensembles (hundreds of thousands of atoms), where the calculated spectrum is compared to an experimental one to generate new 3D models, to improve the fit. The generation of the new models, based on feedback and the relation between spectrums, will need to use new algorithms that have to be implemented via heuristics or AI.

### *ACKNOWLEDGMENTS*

We would like to thank Dr. Xavier Martorell[ac] and Dr. Harald Servat[b] for their contributions to this publication, without which it would not have been possible. Also, thanks to Dr. Xavier Teruel[c] and Dr. Nicola Cadenellic for their personal advice and support. SAXS-WAXS experiments were performed at NCD beamline at ALBA Synchrotron with the collaboration of ALBA staff (special thanks to Juan Carlos Martínez).

[a] Faculty of Computer Science of Barcelona, Polytechnic University of Catalonia.

[b] Intel Corporation Iberia, S.A. Madrid, Spain.

[c] Barcelona Supercomputing Center (BSC), Spain.

### *References*

[1] C. González, J. Bosch, J.M. Haro, M. Paolini, S. Balocco, C. Álvarez, R. Pons, “ACCELERATING PP-DISTANCE ALGORITHMS ON FPGA USING DIFFERENT STRATEGIES AND RUNTIME MANAGERS” Pre-print.

[2] Alejandro Duran, Eduard Ayguadé, Rosa M. Badia, Jesús Labarta, Luis Martinell, Xavier Martorell, and Judit Planas. 2011. OmpSs:A Proposal for Programming Heterogeneous Multi-Core Architectures. Parallel Processing Letters 21, 02 (2011), 173–193. https://doi.org/10.1142/ S0129626411000151.

[3] Fahimeh Yazdanpanah, Carlos Álvarez, Daniel Jiménez-González, Rosa M. Badia, Mateo Valero. 2015. Picos: A hardware runtime architecture support for OmpSs. Future Gener. Comput. Syst. 53: 130-139 (2015). https://doi.org/10.1016/j.future.2014.12.010

[4] Debye, P. Zerstreuung von Röntgenstrahlen. Ann. Phys. 351, 809–823 (1915). https://doi.org/10.1002/andp.19153510606

[5] Patterson, A. (1939). "The Scherrer Formula for X-Ray Particle Size Determination". Phys. Rev. 56 (10): 978–982. doi:10.1103/PhysRev.56.978.

## *Author biography*

**César González** BD in Computer Science from Universitat Oberta de Catalunya (2016). His final project, obtained with honors, was done at the IBMB-CSIC. MD in Computational Engineering and Mathematics from Universitat Rovira i Virgili (2018). Currently he is a PhD student at the Universitat de Barcelona https://orcid.org/0000-0002-6977-4413.

# Exploiting parallelism for CPU and GPU linear solvers on chemistry for atmospheric models

Christian Guzman-Ruiz, Mario Acosta, Oriol Jorba
Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {christian.guzman, mario.acosta, oriol.jorba}@bsc.es

***Keywords*—Parallelism and concurrency, Linear solvers, Chemistry software, Performance**

## I. Extended Abstract

Atmospheric models are a representation of dynamical, physical, chemical, dynamical, and radiative processes in the atmosphere [1]. Most of the algorithms used in these models have been developed decades ago. With the increasing demand on supercomputing performance, these models are gradually being updated to new performance techniques and hardware options.

These improvements typically focus on high time consuming components. One of these components corresponds to the resolution of chemical processes, which can take up to 90% of the total time execution [2]. Recent studies reported a relevant speedup by translating a chemical module to GPUs. For example, the Kinetic Preprocessor library (KPP) from the EMAC Earth system model achieves up to 20x speedup compared to the single-thread CPU version [3]. Other studies developed algorithms specifically for GPU computing, like the Runge-Kutta-Chebyshev for moderately stiff kinetics [4]. This algorithm achieves up to 59x speedup, which is notably greater in comparison to the KPP GPU. However, the KPP is able to solve higher levels of stiffness and its solving method has been proven multiple times on the CPU. Also, the KPP GPU strategy uses the same solving algorithm on both versions, which requires less programming effort.

However, both studies needs to compute a large amount of chemical systems in order to offer a performance increase over CPU algorithms [5]. These chemical systems are provided by the domain decomposition of atmospheric models, which generates multiple systems to solve typically known as grid-cells. These grid-cells are typically parallelized through MPI decomposition, assigning multiple grid-cells load to each MPI thread. As a result, a node has a large quantity of systems to solve.

This large amount of grid-cells per node can be solved in the GPU. However, it can be possible that the domain decomposition generates less cells than the required to outperform CPU algorithms. On the other hand, it can generate enough load to make a simple GPU function more efficient than its CPU alternative. Therefore, developers can integrate this small part of code instead of translating and maintaining the entire chemical module to GPUs.

Nevertheless, integrating this small GPU code can be difficult. The inner GPU function needs to compute multiple grid-cells in order to be efficient. However, the rest of the module is still on a CPU version, solving these systems one by one. The CPU code also needs to solve the grid-cells simultaneously in order to take advantage of this GPU function.

In this work we show a method to solve this issue, which we name as Multi-cells from now on. The strategy consists of solving the multiple chemical systems as a single one. In consequence, the solver needs to allow different system sizes depending on the numbers of cells. For this purpose, we use the chemical module Chemistry Across Multiple Phases (CAMP) as our test-bed. Previously, we claimed that the Multi-cells method achieves up to 12x speedup on CAMP in comparison to the original CPU version [6]. In addition, the computation of chemical reactions in GPU reached up to 18x speedup. However, in this previous study we left for future work the development of GPU versions for the rest of the solving process.

In this contribution, we focus on improving the linear solving method for a GPU computation with Multi-cells. The solving algorithm requires to solve a linear system multiple times during solving, which can take a considerable part of the total time execution. We use the Biconjugate Gradient for the GPU version, achieving up to 13x speedup in front of the original CPU implementation. In addition, we tested a different implementation from Multi-cells, which we name as Block-cells. In Multi-cells, all the cells are solved as a unique system. In consequence, it creates a big system of equations which need to be distributed among multiple GPU blocks. Therefore, the linear system needs to communicate between these blocks, which can take more than 50% of the total execution time [7].

In the Block-cells optimization, we solve each cell individually in a GPU block. This produces an increase of the speedup up to 27x compared to the CPU version. However, there are some differences in precision with respect to the Multi-cells version, since the rest of the solver is still computing all the cells as a whole. These differences can produce extra iterations on the ODE solving process, which can produce differences on the overall speedup. In the future, we will investigate fully GPU alternatives to the ODE solving following the Block-cells idea.

## References

[1] M. Z. Jacobson, “Fundamentals of Atmospheric Modeling, Second Edition,” p. 437.

[2] M. Christou, T. Christoudias, J. Morillo, D. Alvarez, and H. Merx, “Earth system modelling on system-level heterogeneous architectures: EMAC (version 2.42) on the Dynamical Exascale Entry Platform (DEEP),” *Geoscientific Model Development*, vol. 9, no. 9, pp. 3483–3491, Sep. 2016, publisher: Copernicus GmbH. [Online]. Available: https://gmd.copernicus.org/articles/9/3483/2016/

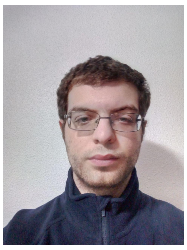
[3] M. Alvanos and T. Christoudias, “GPU-accelerated atmospheric chemical kinetics in the ECHAM/MESSy (EMAC) Earth system model (version 2.52),” *Geoscientific Model Development*, vol. 10, no. 10, pp. 3679–3693, Oct. 2017. [Online]. Available: https://www.geosci-model-dev.net/10/3679/2017/

[4] K. E. Niemeyer and C.-J. Sung, “Accelerating moderately stiff chemical kinetics in reactive-flow simulations using GPUs,” *Journal of Computational Physics*, vol. 256, pp. 854–871, Jan. 2014, arXiv: 1309.2710. [Online]. Available: http://arxiv.org/abs/1309.2710

[5] K. Niemeyer and C.-J. Sung, “GPU-Based Parallel Integration of Large Numbers of Independent ODE Systems,” Jun. 2014, pp. 159–182.

[6] C. Guzman-Ruiz, M. Acosta, and O. Jorba Casellas, “Accelerating atmospheric models using GPUs,” pp. 59–60, May 2020, accepted: 2020-10-30T12:47:44Z Publisher: Barcelona Supercomputing Center. [Online]. Available: https://upcommons.upc.edu/handle/2117/331028

[7] S. Xiao and W. Feng, “Inter-block GPU communication via fast barrier synchronization,” May 2010, pp. 1–12.



**Christian Guzman** received his Bachelor’s degree in Computer Engineering plus a bachelor’s degree in Telecommunication Electronic Engineering, and a Master in Modelling for Science and Engineering by the Autonomous university of Barcelona (UAB). He specialized in techniques for High-Performance Computing (HPC) and is actually developing his pre-doctoral studies on the Barcelona Supercomputing Center (BSC), working on the development of Chemistry Across Multiple Phases (CAMP) module alongside the Multiscale Online Nonhydrostatic Atmosphere Chemistry Model (MONARCH), contributing to the most-computational and logic part from a performance point of view and integrating multitudinal ways of GPU computation in search of speeding-up the system.

# startR: A tool for large multi-dimensional data processing

Núria Pérez-Zanón[*1], An-Chi Ho[*2], Nicolau Manubens[*3], Francesco Benincasa[*4], Pierre-Antoine Bretonnière[*5]

[*]*Barcelona Supercomputing Center, Barcelona, Spain*

[1]nuria.perez@bsc.es, [2]an.ho@bsc.es, [3]nicolau.manubens@bsc.es, [4]francesco.benincasa@bsc.es, [5]pierre-antoine.bretonniere@bsc.es

***Keywords*— big data, data processing, earth sciences, high-performance computing, high-dimensional data**

## Extended Abstract

Nowadays, the huge amount of data produced in various scientific domains has made data analysis challenging. In the climate science domain, with the constant increase of resolution in all possible dimensions of model output and the growing need for using computationally demanding analytical methodologies (e.g. bootstrapping), basic operations like extracting data from storage and performing statistical analysis on them must fulfill scientific and operational needs taking into account the growing volume and variety of data. A tool that facilitates data processing and leverages computational resources can largely save researchers' time and effort.

In this work, we introduce startR, an R package that allows to retrieve, arrange, and process large multi-dimensional datasets automatically with a concise workflow, smoothing the data-processing difficulty mentioned above.

### *A. Introduction*

startR provides a framework under which users can apply a self-defined procedure to a collection of multi-dimensional datasets from the repository as large as desired and take advantage of the computational resources in high-performance computing systems (HPC). It has been designed to require as few technical parameters as possible from the users, but users can tailor a number of configurations to adjust the execution according to their needs.

Under this framework, two approaches to conduct analysis are available: retrieving the data into RAM and performing the analysis, or creating a workflow -the data is not retrieved only declared- in which the last step is to run the user-defined analysis. The first one allows the user to explore the analysis at any stage whereas, in the second one, larger data can be involved in the analysis for the same available RAM. When the workflow is preferred, all the information needed for the data analysis can be encapsulated in a succinct and reusable script which is composed of four main functions: `Start`, `Step`, `AddStep`, and `Compute` (Fig. 1). The required information is all described and defined by these functions, including the involved data file distribution, the dimensions of the desired data cube, the workflow of operations to be applied, the job execution parameters and HPC configuration if needed.

The object type startR works with is array with named dimensions, a basic and widely used data structure. startR obtains data from the repository and returns an array object. The array can have an arbitrary number of dimensions and the order is not restricted as well since startR works with the dimension names rather than the index. This feature provides flexibility for data structure and makes the analysis less error-prone. When the execution finishes, startR also returns an array with named dimensions that can be processed further or stored in files for later use. Besides the data itself, metadata is also well-preserved and expanded with the operation information, ensuring the reproducibility of the analysis.

The startR framework implements the MapReduce paradigm for chunking and processing big data sets. The execution can run either locally or remotely on HPCs with multi-node and multi-core parallelism where possible. The EC-Flow workflow manager is used to dispatch tasks onto the HPCs, providing fault-tolerance and progress control through its graphic user interface (GUI).

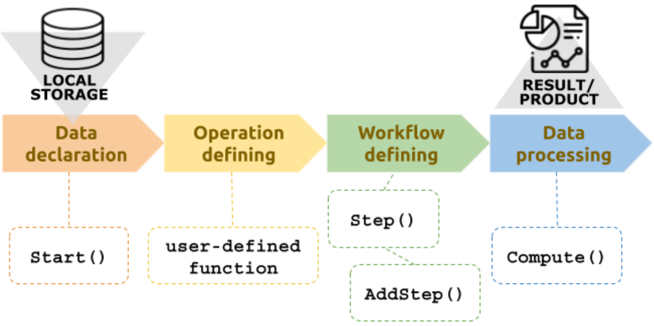


Fig. 1 The startR workflow and the corresponding functions

### *B. Data declaration*

The first step of data analysis is extracting data from storage, which is achieved by function `Start`. startR can access the repositories and the data files do not need to comply with any specific convention for their distribution pattern, file names, or the number and order of the dimensions of the variables. Though netCDF is the only data format supported in the current release, adaptors for other file formats can be plugged in startR if needed.

Start() declares the requested data as an array object. It provides two options for data retrieval: loading the entire data set in the machine or creating a pointer to the data location in the repository. With the first option, the Start call returns a multi-dimensional array that occupies the memory space of the workstation, so the involved data size is limited. On the other hand, the Start call only retrieves the dimensions and metadata of the required data and the storage location with the latter option, so the involved data size is unlimited in theory. Loading data is the conventional way for the following analysis, whereas generating a pointer can make full use of the functionality of startR.

There are several advanced parameters in Start() to handle heterogeneities across files involved in one Start() call, such as transformation, reordering, reshaping, and renaming (Table 1). In addition, ancillary data and metadata can be well-preserved through parameters considering the complex data file distribution. Start() also accepts user-defined functions for some of these purposes.

TABLE I
START() PARAMETERS AND THEIR FUNCTIONALITY

| Functionality | Parameters |
|---|---|
| **Transformation**: Interpolate data to the desired grid type. | transform<br>transform_params<br>transform_vars<br>transform_extra_cells<br>apply_indices_after_transform |
| **Reshaping and reordering**: Combine one dimension along with the files or split a dimension into multiple ones; Sort dimensions with associated values. | merge_across_dims<br>merge_across_dims_narm<br>split_multiselected_dims<br>*_depends<br>*_across<br>*_reorder |
| **Dimension and path definition**: Identify the key information required to locate files, homogenize dimension and variable names among datasets, and retrieve metadata and ancillary data. | pattern_dims<br>metadata_dims<br>path_glob_permissive<br>return_vars<br>synonims<br>largest_dims_length<br>*_var |
| **File format and data selector support**: Specify interface functions for desired file format and conversion between different selector (i.e., data of interest) types | file_opener<br>file_var_reader<br>file_dim_reader<br>file_data_reader<br>file_closer<br>selector_checker |

### C. Workflow

After the data sources are declared, the next step is to define the operation to be applied. A self-defined function in R function form and the startR functions `Step` and `AddStep` are required to build up the workflow. The workflow adopts the multiApply mechanism, which is an R package and an extension of the base apply function in R. In the apply() fashion, the function only operates on the essential dimensions but not the whole data set, and the operation will loop over the margin dimensions behind the scene. The apply family typically applies a function to a single argument, whereas multiApply efficiently takes one or a list of multi-dimensional arrays as input, suitable for the operation of the declared data arrays.

In Step(), the dimensions to be performed on and the expected returned dimensions are specified by their dimension names, which are more semantically meaningful than the indices. By this means, users can save effort on checking the dimension order and prevent the mistakes hard to detect. After the step is defined, the data and the step are ready to be bundled together as a workflow by AddStep().

### D. Data processing

Once the workflow is defined, the final step is to choose the platform to run the execution and specify the execution parameters by the function `Compute`. The platform can be either the workstation that runs the code (locally) or the HPCs that share the same file system with the workstation (remotely).

One important feature of startR, MapReduce paradigm (i.e., chunking), can be realized at this step. Users can specify which dimensions to split the data array along and the number of chunks to make for each, making each chunk fit in the RAM memory. Therefore, even if the total data size is larger than the available RAM, the execution won't overload and crash the platform. The operation defined in the workflow will be applied to each chunk, and the results will be stored in a temporary file. When all the chunks are finished, Compute() will collect and merge the results and return an object including one or multiple multidimensional data arrays as defined in Step(), and the additional metadata generated during the execution.

Since startR features in a multi-node and multi-core manner, the number of execution threads to use for data retrieval and for computation can be determined by the users. If the execution is on a remote HPC, the configuration of the machine needs to be specified in Compute(), including the number of cores per job, the amount of memory to book for each job, the type of workload used by the HPC, and the maximal wall-clock time, etc. Understanding the properties of the machine can help optimize execution efficiency. Besides, the EC-Flow server is required to be installed for job dispatch on HPCs. Through the EC-Flow graphical user interface, users can check and control the execution status of each chunk during the computation. If one chunk fails, it can be re-submitted individually without running the whole execution again, so the risk of time and resource waste is reduced.

### E. Summary and future work

startR provides a powerful framework for large multi-dimensional data retrieval and processing. With its clear workflow, a piece of data analysis work can be defined with only tens of lines, and the script is easy to reuse and adapt to other analyses. Though startR has had much flexibility for data structure and operations, it can be strengthened further. For example, only one step (i.e., one function) in the workflow is allowed now. Even though several procedures can be wrapped in one step, multiple steps can optimize the target dimension choice for different operations. Besides, as the popularity of cloud databases grows, retrieving data from cloud-based data systems is worth the development.

Several functionalities in startR, like spatial interpolation and time manipulation, are tailored for earth sciences research, with a special focus on atmospheric sciences such as climate, weather, and air quality. It is compatible with other R tools developed in BSC-CNS, forming a strong toolset for climate research. However, it is potentially competent in other research fields. With the plug-in of other interface functions, startR can be exploited in different scientific domains where large multi-dimensional data is involved.

## Author biography

**An-Chi Ho** is a researcher engineer in the Data and Diagnostic team in the Computational Earth Sciences group at BSC. She is responsible for the development and maintenance of the R-related tools in the department. She received her MA degree in Climate and Society Program from Columbia University in 2018 and MSc degree in Atmospheric Sciences from National Taiwan University in 2017.

# Curved geometry modeling: interpolation of subdivision features

Albert Jiménez-Ramos*, Abel Gargallo-Peiró*, Xevi Roca*
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
E-mail: {albert.jimenez, abel.gargallo, xevi.roca}@bsc.es

***Abstract*—We present a nodal interpolation method to approximate a subdivision model. The main application is to model and represent curved geometry without gaps and preserving the required simulation intent. Accordingly, the technique is devised to maintain the necessary sharp features and smooth the indicated ones. This sharp-to-smooth modeling capability handles unstructured configurations of the simulation points, curves, and surfaces. These surfaces are described by the initial triangulation composed of linear triangles with the same surface identifier, and determine the sharp point and curve features. Automatically, the method suggests a subset of sharp features to smooth which the user modifies to obtain a limit model preserving the initial points. This model reconstructs the curvature by subdividing the initial triangular mesh, with no need of an underlying curved geometry model. Finally, given a polynomial degree and a nodal distribution, the method generates a piece-wise polynomial representation interpolating the limit model. Numerical evidence suggests that this approximation, naturally aligned to the subdivision features, converges to the model geometrically with the polynomial degree for fair nodal distributions. We also apply the method to prescribe the curved boundary of a high-order volume mesh. We conclude that our sharp-to-smooth modeling capability leads to curved geometry representations with enhanced preservation of the simulation intent.**

***Keywords*—mesh curving, surrogate geometry, geometry modeling, subdivision, blending**

## I. Extended Abstract

### A. Introduction

The capability to model and represent curved geometry preserving the simulation intent is critical for flow simulation with unstructured high-order methods. These high-order simulations require curved meshes that approximate a curved boundary representation. Ideally, this boundary representation should be composed of smooth and sharp features agreeing with the simulation intent.

Flow simulation practitioners favor continuous normal vectors on smooth features where the intent is to obtain attached flow. In contrast, they only need model continuity on sharp features where the flow detaches. To illustrate both types of features, we can consider an aircraft model. We can find there smooth features such as the nose tip, leading edges, and wing surfaces; and sharp features such as trailing edges and points.

To model the previous features, it is standard to use CAD boundary representations based on trimmed NURBS. Unfortunately, these trim-based models might violate the simulation intent due to unintended gaps or discontinuities of the normal vector on irregular points and curves adjacent to trimmed surfaces. Nevertheless, if the element size is fine but coarser than the model tolerance, we obtain a fair second-order approximation of the CAD boundary representation without gaps. However, since the triangles are planar, this approximation does not feature the normal vector continuity through any triangular edges, and thus, it is not adequate for flow simulation. Still, we can convert the triangular mesh to a gap-free curved geometry model [1] that features normal vector continuity by using a subdivision scheme [2].

This subdivision-based conversion to a curved model is useful even when there is no underlying CAD model. The conversion only needs a model composed of triangulations, which boundaries determine the model points and polylines. Hence, this conversion is of practical interest in any application where the triangular mesh comes from legacy data or real samples, such as in onshore wind farm energy forecasting, transport of pollutants in urban areas, and bio-engineering.

In these applications, the subdivision conversion provides a curved limit model. We can query this limit model by successive refinements [1]. However, this approach requires more refinement levels when the closer is the query point to an irregular point. To avoid this unbalanced query, in our previous work [3], we proposed a method to interpolate with a continuous piece-wise quadratic or quartic mesh the limit model while exploiting the structure of iterative subdivision. Any posterior query to the interpolation model only uses the corresponding triangular-wise polynomial components, thus skipping the successive refinement step.

Although skipping posterior successive refinement, our previous approach only extends to interpolation degrees equal to powers of two on equispaced nodal distributions. Therefore, it does not allow using arbitrary polynomial degrees and nodal distributions. Recall that beyond degree four equispaced nodal distributions feature large Lebesgue constants that might hamper the corresponding interpolation accuracy.

The main contribution of this work is to propose a method to interpolate the subdivision model with any degree and nodal distribution. The method evaluates one time the limit model parameterization [4] on each interpolation point to obtain the resulting nodal curved mesh model which is ready to prescribe the boundary of a curved high-order volume mesh.

We also propose an approximation of the distance between the interpolation and the limit model to check the geometric accuracy in terms of the polynomial degree for different nodal distributions. To compute this distance, we only need to perform forward evaluations of the nodal parameterization.

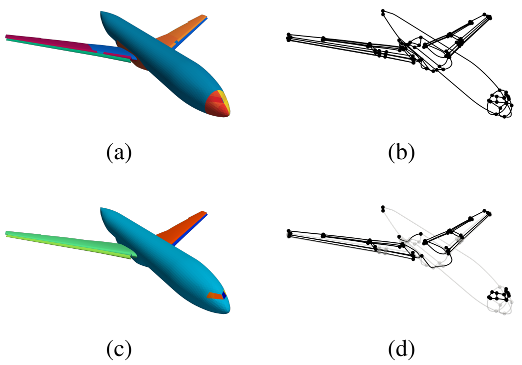


Fig. 1: Initial and final linear mesh model of an aircraft in high-lift configuration. Initial model: (a) surface features colored with their surface identifier, and (b) curve and point features. Final model: (c) virtual surface features colored with their surface identifier, (d) curve and point features recast (gray) and preserved (black).

We finally propose an assisted sharp-to-smooth modeling capability aimed to reduce the amount of human labor required to describe the simulation intent of the model. The resulting method automatically suggests a subset of sharp features to smooth which the user modifies to obtain a limit model preserving the initial points.

### B. Example

We illustrate the features of our method by modeling and curving an aircraft model in a high-lift configuration.

0) **Sharp-to-smooth modeling**. In Fig. 1(a) and 1(b), we show the initial model of the aircraft composed of 118 surface, 282 curve, and 182 point features. We devise a technique to automatically suggest the smooth features to recast which the practitioner revises to preserve the simulation intent. The recast linear model $\Omega^1$, see Fig. 1(c) and 1(d), is composed of 67 surface, 169 curve, and 122 point features.
1) **Approximate a surrogate boundary**. The hierarchical subdivision of the geometry features determines a set of $\mathcal{C}^2$-continuous limit curves and $\mathcal{C}^1$-continuous limit surfaces that serves as surrogate geometry to generate a curved high-order triangular surface mesh. This curved surface mesh preserves the sharp features and smooth regions of the linear model $\Omega^1$, and interpolates it at the nodes of the high-order mesh.
2) **Accommodate the curvature of the boundary**. We accommodate the curvature of the curved surface mesh to the boundary volume elements using an explicit hierarchical blending.
3) **Local smoothing and untangling**. If necessary, we optimize the low-quality elements locally following the approach detailed in [5].

In Fig. 2, we show the generated surface mesh of polynomial degree four using the recast model where the surfaces have been colored with their surface identifier. Since the curves describing the leading edge have been automatically recast, the surface is smooth along that feature. On the contrary, we can appreciate the desired discontinuity on the normal vectors of the surface along the curves describing the cabin windows.

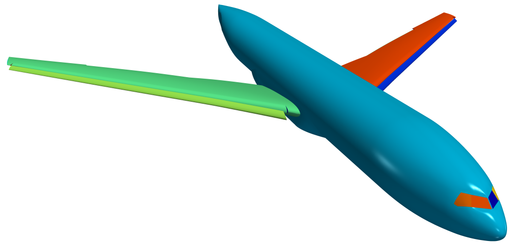


Fig. 2: Curved triangular mesh of polynomial degree four of an aircraft in high-lift configuration colored with the surface identifiers.

### C. Concluding Remarks

In conclusion, we have presented a methodology to model and represent curved geometry of practical interest for flow simulation with unstructured high-order methods. In perspective, high-order methods might benefit from using curved meshes that approximate our curved boundary representation, which we devised to describe the flow simulation intent.

## II. Acknowledgment

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 715546. This work has also received funding from the Generalitat de Catalunya under grant number 2017 SGR 1731. The work of the third author has been partially supported by the Spanish Ministerio de Economía y Competitividad under the personal grant agreement RYC-2015-01633. Special thanks to Eloi Ruiz-Gironés.

## References

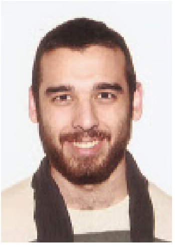
[1] P.-O. Persson, M. J. Aftosmis, and R. Haimes, "On the use of loop subdivision surfaces for surrogate geometry," in *Proceedings of the 15th International Meshing Roundtable*, P. P. Pébay, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 375–392.

[2] C. Loop, "Smooth subdivision surfaces based on triangles," Master's thesis, Department of Mathematics, The University of Utah, Masters Thesis, January 1987.

[3] A. Jiménez-Ramos, A. Gargallo-Peiró, and X. Roca, "Subdivided Linear and Curved Meshes Preserving Features of a Linear Mesh Model," in *Proceedings of the 28th International Meshing Roundtable (IMR)*. Zenodo, Feb. 2020.

[4] J. Stam, "Evaluation of loop subdivision surfaces," in *Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH*, 01 1998.

[5] A. Gargallo-Peiró, X. Roca, J. Peraire, and J. Sarrate, "Optimization of a regularized distortion measure to generate curved high-order unstructured tetrahedral meshes," *International Journal for Numerical Methods in Engineering*, vol. 103, no. 5, pp. 342–363, 2015.



**Albert Jiménez-Ramos** received his BSc degree in Mathematics from Universitat Politècnica de Catalunya (UPC), Barcelona in 2017. He completed his MSc degree in Advanced Mathematics and Mathematical Engineering from UPC in 2018. Since 2018, he has been a member of the Geometry and Meshing for simulations group of Barcelona Supercomputing Center (BSC) as a PhD student of the Applied Mathematics doctorate program of UPC, Spain.

# Mining the Essential Motions of Pyruvate Kinase

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

[#]*Barcelona Supercomputing Center (BSC)*

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

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

***Keywords* — Pathogenicity prediction, pathogenic variant, protein dynamics, molecular dynamics, pyruvate kinase.**

## I. Extended Abstract

Our current study revolves around the dynamic characterization of the human erythrocyte pyruvate kinase (PKR). The deficiency of this protein is a common cause of nonspherocytic hemolytic anemia, a rare, autosomal recessive disease. We are performing a comprehensive set of molecular dynamics simulations of both the wildtype (WT) and mutant variants of PKR in different conditions, in order to explore the dynamic behavior of the enzyme, describe the function and allosteric mechanism in terms of its dynamics fingerprint and identify altered dynamic behavior of the known pathogenic variants of the enzyme.

### *A. Introduction*

Genotyping is a powerful resource to confirm the diagnosis of Mendelian disorders (i.e. diseases caused by alterations on a single gene). The practice is increasingly becoming available and prevalent, thanks to the advances made on the techniques and their cost reduction [1]. A capital challenge of the field is the correct interpretation of the degree of pathogenicity of the found variants. The effect of a given amino acid replacement in a protein sequence can diverge from complete disruption (damaging) to innocuousness (benign). In this context, dozens of in silico predictive tools have been developing in the recent years, with the aim of providing a pathogenicity predictive score to a given missense mutation in a given gene sequence [2]. While the approaches taken in the different available tools is diverse, prediction based on sequence conservation has been the predominant strategy adopted by this kind of methods. Some predictive tools also use other complementary features such as physicochemical and biochemical properties of the replaced amino acids and functional annotations of replacement sites (if available) [3].

On the other hand, we can study the effect of a missense mutation directly at the level of the structure of the protein. By exploring the mutation site in its structural context, it is possible to estimate the impact of the amino acid substitution in terms of the expected structural alterations. Structural and dynamic features have been largely neglected in the field of pathogenicity prediction, mainly due to the difficulties and computational cost of the obtainment of a large enough dataset of metrics. However, several efforts and initiatives have been coming out in the recent years [4]. Our project revolves around the consideration of the role of dynamics features in the functional behavior of protein variants that can be used to improve pathogenicity prediction algorithms.

### *B. Use case*

To this end, we are studying the particular case of the human erythrocyte pyruvate kinase (PKR). Pyruvate kinase is one of the most widely studied enzyme families throughout the history of biochemistry, due to its major role in the regulation of glycolysis. It catalyzes the irreversible conversion of phosphoenolpyruvate (PEP) to pyruvate, generating an ATP molecule in the process. The enzyme needs cofactors ($K^+$ and $Mg^{2+}$) for proper catalytic activity, is allosterically regulated by fructose-1,6-biphosphate (FBP), an activator of its catalytic efficiency, and has a highly conserved architecture throughout evolution (Fig. 1) [5].

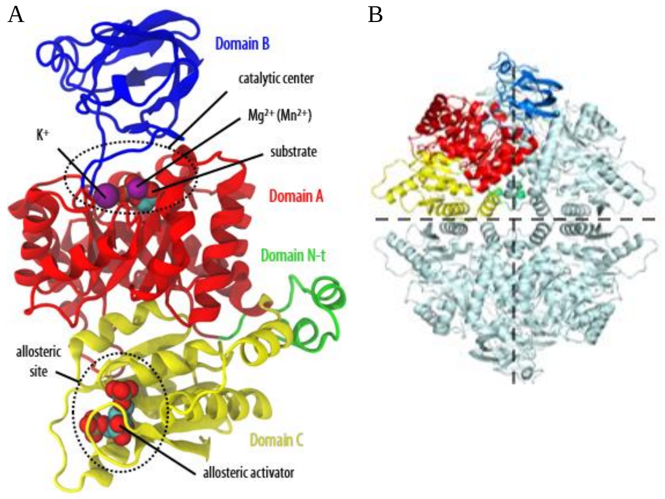


Fig. 1 Tertiary and quaternary structure of PKR. The different domains of the protein: N-terminal, A, B and C are colored green, red, blue and yellow respectively. A) View of a monomer of PKR and bound cofactors, substrate and allosteric activator. B) View of the tetramer of PKR. The two symmetry axes are shown with a dashed line. Only one subunit is colored as indicated above (the rest are colored pale cyan).

Hundreds of missense mutations in various sites of PKR have been clinically associated to a disease called hereditary nonspherocytic hemolytic anemia (5 cases per 10,000 individuals in Europe) [6]. The main aim is to go deeper into the coupling between structure and functions of this protein, as well as to provide information about its dynamic properties for both the WT form and some of its mutant variants of known pathological consequences.

### *C. Preliminary results*

We are currently conducting a thorough stage of analysis of the dynamic behavior of PKR. We work with data derived from MD simulations of PKR in a wide variety of biologically meaningful conditions. We cover both the apoenzyme form (no bound cofactors/ligands) and several variations of the holoenzyme: i) with one cofactor ($K^+$), ii) with both cofactors ($K^+$ and $Mg^{2+}$), iii) with both cofactors and the substrate PEP, iv) with both cofactors and the substrate ADP-$Mg^{2+}$, v) with both cofactors and both substrates, vi) with both cofactors and the allosteric ligand FBP, and vii) with both cofactors, both substrates and FBP. Such spectrum of conditions required an establishment of a proper MD protocol, suitable to cope both with the large dimensions of the system (around 400 000 atoms) and the treatment of coordination bonds between the protein and the metal cofactors. In this sense, one of the major highlights of the protocol has been the implementation of a QM-derived step to build parameters to model the known geometries of the binding site of the involved molecules.

In our quest for identifying and classifying the essential motions of this protein, we are applying Principal Components Analysis (PCA), a method that allows us to inspect the major concerted motions (the so-called principal components (PCs)) associated to the largest collective atomic fluctuations of each MD trajectory [7]. On top of this traditional technique, we are developing a custom method to cluster the obtained PCs from each MD trajectory. Then, we compute the centroid PC of each major cluster. The centroid PC can be interpreted as a consensus mode that: a) maintains the most relevant collective fluctuation of variables, and b) filters out the minor collective fluctuations (i.e. the fraction of correlation that is present only in particular members of the cluster). The result is a rendering of the "pure" motions of PKR occurring at the timescale of our MD simulations (400 ns). To the best of our knowledge, this approach has not been explored yet in a scenario like ours. Although still in assessment, our results hint towards an effective method of collecting and cataloguing the essential motions of the analyzed regions.

Some main collective motions show up in a consistent way in the different apo and holo conditions, potentially revealing that a fraction of the protein's main dynamics is conserved regardless of the presence or absence of the substrates and the allosteric activator. A better understanding and classification of these motions is potentially the key to revealing how the dynamic behavior of the protein is tuned in response to binding events, leading to functional enzymatic consequences.

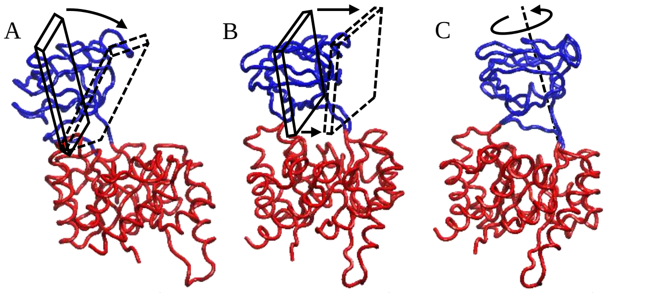


Fig. 2 Example of the essential motions of domains A and B of PKR. Domains A and B are colored red and blue, respectively. A) Opening and closing of domain B towards the catalytic site (this motion is described in the classical inactive to active transition of the enzyme). B) Lateral shaking of domain B. C) Rotation of domain B with respect to domain A.

### *D. Future development*

In the next steps of our analysis stage, we will focus on refining the study of the obtained essential motions of PKR. Some of the interesting features to collect are the average frequencies and amplitudes of motion of the motions during our MD timescale, and whether they occur simultaneously or not within the tetramer. In this sense, we will assess the known allostery and cooperativity effects of this enzyme in terms of its dynamic behavior.

Furthermore, we have begun the production of a new batch of MD simulations consisting of 33 pathogenic and 18 benign variants of PKR, simulated both in apo and some representative holo conditions. By comparing the dynamic patterns of the WT protein and its mutated forms, we expect to find and describe disruptive dynamic events on the tested pathogenic variants.

Lastly, we are interested in integrating our approaches with the increasingly popular Correlation Network Analyses (CNA), a family of methods that employ graph theory tools to analyze MD trajectories using correlation data [8].

## II. Acknowledgment

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

## References

[1] P. Bianchi *et al.*, "Addressing the diagnostic gaps in pyruvate kinase deficiency: Consensus recommendations on the diagnosis of pyruvate kinase deficiency", Am J Hematol., 94(1), 149–61, 2019.

[2] K. Shameer *et al.*, "Interpreting functional effects of coding variants: Challenges in proteome-scale prediction, annotation and assessment", Brief. Bioinform., 17, 841–862, 2016.

[3] V. López-Ferrando *et al.*, "PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update", Nucleic Acids Res., 45, W222–W228, 2017.

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

[5] G. Valentini *et al.*, "Structure and function of human erythrocyte pyruvate kinase: Molecular basis of nonspherocytic hemolytic anemia", J. Biol. Chem. 277, 23807–23814, 2002.

[6] L. Montllor *et al*., "Red cell pyruvate kinase deficiency in Spain: A study of 15 cases", Med. Clínica (English Ed. 148, 23–27, 2017.

[7] C. C. David, & D. J. Jacobs, "Principal component analysis: A method for determining the essential dynamics of proteins", Methods Mol. Biol. 1084, 193–226, 2014.

[8] X.-Q. Yao *et al.*, "Dynamic Coupling and Allosteric Networks in the Alpha Subunit of Heterotrimeric G Proteins", Biophys. J. 110, 2016.

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

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

# An Architecture for Automatic ML/AI Workflow management and supervision

Peini Liu*†, Jordi Guitart*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {peini.liu, jordi.guitart}@bsc.es

***Keywords*—Architecture, Automatic, Machine Learning Workflows.**

## I. Extended Abstract

Scientific computation problems have been faced with the need to analyze increasing amounts of data as part of their application workflows, and the science-based model is being combined with big data and machine learning models to solve complex problems and phenomena [1][2]. The machine learning workflow is composed of some reproducible steps that can be executed as a pipeline to build a model efficiently by saving iteration time, helping in debugging and detecting [3]. Currently, businesses and researchers are investigating and improving the methodology of developing and deploying machine learning workflows in both training and inference phases, which helps the data science team focus on their requirements and the data engineer team deploy and operate machine learning workflows efficiently and automatically [4].

This work presents an architecture for automatic machine learning workflows, which provides capabilities of monitoring and automatic management on the end-to-end life-cycle of machine learning workflows, including tracking and observing at the training stage, and releasing, monitoring, deployment, auto-detecting and infrastructure management at the inference stage. To validate feasibility, we have conducted a case study based on our architecture and deployed it in the cloud, and showed its automation.

### A. Architecture for Automatic Machine Learning Workflows

While working on machine learning workflows, different teams have their own focuses. The Data science team faces challenges regarding model training, especially on building and improving the model. However, the data engineer team usually works on integrating and deploying the model into production, and operating the model to provide reliable, robust, and efficient service.

The architecture for the end-to-end automatic machine learning workflows is shown in Figure 1. There are many phases and steps required to make the machine learning model in production to provide values. The top describes the steps for the data science team before a model into production. Normally, the data science team will first discover the use case and data, and then develop a machine learning workflow that contains data preparation, validation and preprocessing, as well as model training, validation and testing. Workflow manager (e.g., Scanflow) can track the metadata such as metrics and scores and the artifacts during the training phase, analyze them and automatically tune the hyper-parameters, early stopping and do neural architecture search for improving the model [5].

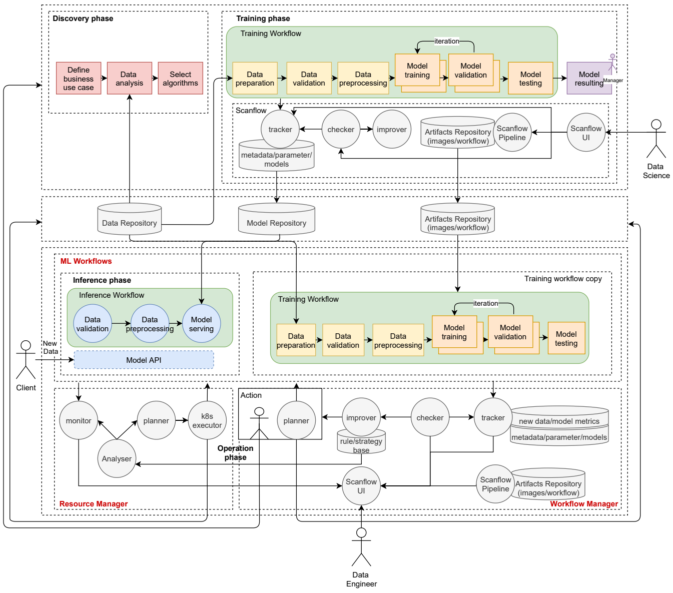


Fig. 1. Architecture for Automatic Machine Learning Workflows

The bottom describes the model in production, including the model inference workflow deployment and the operation phase that automatically manages the machine learning workflow from both the application layer (e.g., workflow manager Scanflow) and the infrastructure layer (e.g., resource manager Kubernetes). For deploying and managing the machine learning workflow at scale, the data engineer team should also build a workflow managed by the workflow manager but wrap and deploy the model as a service. From the application layer controlled view, the workflow manager could log the model metrics(such as scores) and artifacts(such as new data) to detect outliers, adversarial or drift and provide model explanations[6] and finally trigger the machine learning workflow to be re-trained or the model to be updated. From the infrastructure layer controlled view, allowing the model as a service helps it to be released, updated and rollouted independently, and can monitor the latency and failure rate of its predicted invocations at inference time [7][8]. With these observations, the resource manager can automatically scale the service to achieve the reliability and efficiency of the model.

### B. *A Case Study*

To illustrate the feasibility of our proposed architecture for automatic machine learning workflows, we used a leaf classification problem as a running example [9]. The automation scenarios for both training and inference phases are presented in Table I.

TABLE I. Automation Scenarios

| Stages | Type of events | Actions |
|---|---|---|
| training | model with hyper-parameters | different hyper-parameters tuning |
| | multiple models | model selecting |
| inference | predicting service failure rate over 90% | scale the service with more replicas |
| | the rate of the number of requests per second to the predicting service in 5 minutes over 2 | scale the service with more replicas |
| | outliers or model drift when new data comes | retain and update the model |

During this machine learning project, firstly we create a training workflow that contains data gathering, data preprocessing and model training as shown in Figure 2. The model training step uses the random-forest algorithm with different hyper-parameters tuning running in parallel. Then we end up with the best model selection based on accuracy.

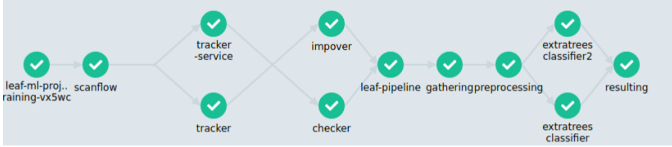


Fig. 2. Leaf training workflow

After the model training, the inference workflow is ready to be deployed in production. Figure 3 shows the inference workflow, data preprocessing and model serving are services and exposed through the cluster for clients to do invocations for predictions.

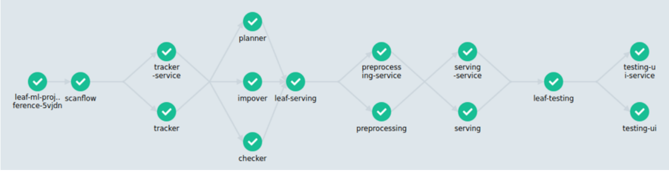


Fig. 3. Leaf inference workflow

The workflow manager(e.g., Scanflow) is started for each workflow for logging the metrics and artifacts of the workflow and debugging and managing the robustness of the model. As for the infrastructure management, we use Kubernetes as the basic container orchestration platform and Istio as a service mesh to gather all the traffic through each service and trace each invocation. Finally we use Keda event to drive autoscaling. Corresponding results are shown in Figure 4.

### C. *Conclusion*

This work presents an architecture for automatic machine learning workflows, which provides capabilities of monitoring and automatic managing on end-to-end life-cycle of machine learning workflows. The technique behind each component will continuously evolve, but we believe the architecture could be a blueprint for improving the efficiency of the machine learning model into production. The next step we will improve the management policy of the planner in each stage.

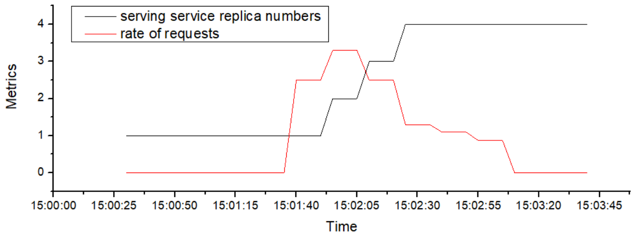


Fig. 4. Auto-scaling: The rate of the number of requests per second to the serving service over 2, system scale the service with more replicas.

## II. Acknowledgment

We thank Lenovo for providing the technical infrastructure to run the experiments in this paper. This work was partially supported by Lenovo as part of Lenovo-BSC collaboration agreement, by the Spanish Government under contract PID2019-107255GB-C22, and by the Generalitat de Catalunya under contract 2017-SGR-1414 and under grant 2020 FI-B 00257.

## References

[1] NITRD Group, "The Convergence of High Performance Computing, Big Data, and Machine Learning," September 2019.

[2] M. Asch *et al.*, *Big Data and Extreme-scale Computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry*, 2018, vol. 32, no. 4.

[3] kubeflow, "The machine learning toolkit for kubernetes," 2021. [Online]. Available: https://www.kubeflow.org/

[4] Seldon, "Machine learning deployment for enterprise," 2021. [Online]. Available: https://www.seldon.io/

[5] "Katib: a kubernetes-native project for automated machine learning (automl)," 2021. [Online]. Available: https://github.com/kubeflow/katib

[6] J. Klaise *et al.*, "Alibi: Algorithms for monitoring and explaining machine learning models," 2019. [Online]. Available: https://github.com/SeldonIO/alibi

[7] Kubernetes, "Production-grade container orchestration - automated container deployment, scaling, and management," 2021. [Online]. Available: https://kubernetes.io/

[8] Istio, "Connect, secure, control, and observe services," 2021. [Online]. Available: https://istio.io/

[9] Kaggle, "Leaf classification," 2021. [Online]. Available: https://www.kaggle.com/c/leaf-classification

**Peini Liu** received her M.S. degree in College of Computer at National University of Defense Technology (NUDT), in 2018. She is currently a Ph.D. student in Computer Architecture Department of the Universitat Politècnica de Catalunya (UPC) and collaborating with Emerging Technologies for Artificial Intelligence group of Barcelona Supercomputing Center (BSC). Her research interests include virtualization/containerization technologies, cloud native, resource management and the convergence of HPC, Big Data and AI.

# Bias-adjustment method for street-scale air quality models

J.M Armengol[#1], A. Criado[#2], J. Benavides[#3], O. Jorba[#4], M. Guevara[#5], A. Soret[#6]

[#]*Barcelona Supercomputing Center (BSC)*

[1]jan.mateu@bsc.edu, [2]acriado@bsc.edu, [3]j.perel@bsc.es, [4]oriol.jorba@bsc.es, [5]marc.guevara@bsc.es, [6]albert.soret@bsc.es

***Keywords*— air quality, bias correction, street-scale modeling**

## EXTENDED ABSTRACT

Air quality (AQ) is a growing concern, especially in urban areas where high-density populated regions are exposed to frequent exceedances of regulated pollutants. To take action in reducing citizen exposure to pollution, a reliable assessment of the pollutants' ambient concentrations across the city is required.

Street-scale AQ models are designed to capture the typical spatial variability that pollutants exhibit in the urban morphology. Such urban models are generally nested to regional AQ models and use the information of traffic emissions, together with meteorological conditions, and a geometric description of the building's layout, to provide an estimation of the dispersion of target pollutants at the street scale. However, results of urban AQ models are subjected to uncertainties, mainly due to the multiscale behavior of the phenomenon and to the challenges of characterizing the wind flow within street-canyons, which encompasses multiple emission sources and the downscaling of meteorological variables.

To minimize these uncertainties, we present a data-fusion method that combines the model results, obtained using the CALIOPE-Urban [1] model, with publicly available observations from the official monitoring network in Catalonia (XVPCA). This method is derived to preserve the spatial variability of the urban model. As a test case, we then present annual bias-corrected results of the $NO_2$ levels across the city of Barcelona for the year 2019. Results correspond to the legislated annual mean and the 19th daily maximum value of the year.

### A. Observational data

The observational data used for the data-fusion method are obtained from the 7 urban monitoring stations of the XVPCA that measure $NO_2$ within the Barcelona municipality.

### B. Model data

The spatial distribution of $NO_2$ is obtained using the CALIOPE-Urban [1] model. CALIOPE-Urban computes the dispersion of pollutants emitted by the traffic sector [2] at the street-scale using a Gaussian dispersion model. Background concentrations, used as an input in the urban model, are provided by the regional AQ model CALIOPE. In this work, the urban model computes hourly $NO_2$ concentrations at a resolution of 50 m x 50 m, while the regional model uses nested domains of resolutions: 1km x 1km for the region of Catalonia, 4 km x 4km for the Iberian Peninsula, and 12 km x 12 km for the European Union region.

### C. Data fusion methodology

Data fusion methodologies aim to correct systematic errors of the model at the entire computational grid using the available observations. In contrast with the commonly data-fusion methods used in the regional-AQ models (such as CALIOPE) that are based on the distance to the observational sites; in this work, we derive a method that preserves the spatial variability estimated by the urban model.

For this purpose, linear regressions for the daily mean concentration and the daily variability are fitted using all available data pairs of model-observation. Using these regressions, model results are corrected in terms of daily mean and daily variability, relying solely on the model output. The post-processing models are refitted every day to account for the dependency of the systematic errors on the meteorological conditions.

### D. Results

Leave-One-Out-Cross-Validation (LOOCV) has been performed to assess the validity of the present data fusion method. This validation consists in performing the bias-adjustment considering all observational data except one, which is used to cross-validate the results. In Table I, raw results from the urban model are compared with the bias adjustment and the cross validation (LOOCV) results. In average, locations for which there are no available observations are expected to have a correlation coefficient of $r^2$=0.62 and a fraction of predictions within a factor of two FAC2=0.80.

TABLE I
CROSS VALIDATION RESULTS OF THE BIAS CORRECTION FOR 2019

| Station name | Raw model | | Bias adjustment | | LOOCV | |
|---|---|---|---|---|---|---|
| | FAC2 | $r^2$ | FAC2 | $r^2$ | FAC2 | $r^2$ |
| Sants | 0.67 | 0.42 | 0.80 | 0.61 | 0.79 | 0.60 |
| Eixample | 0.82 | 0.53 | 0.91 | 0.69 | 0.87 | 0.66 |
| Ciutadella | 0.69 | 0.53 | 0.78 | 0.64 | 0.76 | 0.63 |
| Palau Reial | 0.71 | 0.47 | 0.76 | 0.62 | 0.75 | 0.61 |
| Gràcia | 0.81 | 0.54 | 0.87 | 0.69 | 0.85 | 0.65 |
| Poblenou | 0.72 | 0.58 | 0.84 | 0.70 | 0.82 | 0.68 |
| Valld'Hebron | 0.65 | 0.46 | 0.74 | 0.57 | 0.72 | 0.54 |

The corrected annual average map of $NO_2$ concentrations is shown in Fig.1. In the regions close to the main roads of the city, where the traffic is more intense, as well as the area close to the port (at the south region), the legislated threshold of 40 $\mu g/m^3$ (established in 2008/50/CE) is exceeded.

Model results of the $NO_2$ annual average at the observational sites are compared with measurements in Table II. The corrected model results are in good agreement with the observed annual mean, being that the larger relative difference is 10.6% (positive values of the relative difference stand for model over-predictions and vice-versa).

TABLE II
MODEL RESULTS AND OBSERVATIONS OF ANNUAL MEAN $NO_2$ FOR 2019

| **Station name** | **Model mean** ($\mu g/m^3$) | **Obs. mean** ($\mu g/m^3$) | **Relative diff. (%)** |
|---|---|---|---|
| Sants | 30.49 | 30.81 | -1.06 |
| Eixample | 46.65 | 48.67 | -4.33 |
| Ciutadella | 35.49 | 31.73 | 10.60 |
| Palau Reial | 29.66 | 27.46 | 7.40 |
| Gràcia-St.Gervasi | 40.37 | 42.78 | -5.97 |
| Poblenou | 33.17 | 36.02 | -8.58 |
| Vall d'Hebron | 27.62 | 28.45 | 3.02 |

The daily hourly maximum limit of $NO_2$, defined in the 2008/50/CE as the 19th daily maximum of the year, corresponds to a threshold of 140 $\mu g/m^3$. Figure 2 shows the 19th daily maximum $NO_2$ concentration across the city of Barcelona, again some exceedance of the legal limit can be observed in heavily trafficked streets. Looking at the NO2 19th daily maximum concentrations in Table III, model exceedances are observed at the traffic monitoring sites of Eixample and Gràcia.

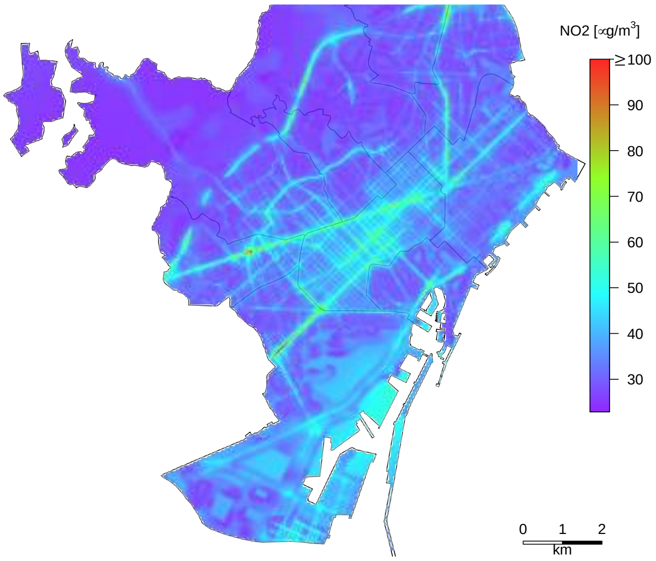


Fig. 1 Annual mean NO2 concentration levels at the municipality of Barcelona for 2019.

TABLE III
MODEL RESULTS AND OBS. OF THE 19TH DAILY MAX. OF $NO_2$ FOR 2019

| Station name | Model 19th max. ($\mu g/m^3$) | Obs. 19th max. ($\mu g/m^3$) | Relative diff. (%) |
|---|---|---|---|
| Sants | 131.70 | 111.05 | 15.68 |
| Eixample | 149.91 | 137.58 | 8.22 |
| Ciutadella | 123.98 | 107.12 | 13.60 |
| Palau Reial | 131.31 | 115.96 | 11.69 |
| Gràcia St.Gervasi | 133.98 | 156.26 | -16.62 |
| Poblenou | 120.82 | 114.97 | 4.83 |
| Vall d'Hebron | 124.54 | 114.99 | 7,68 |

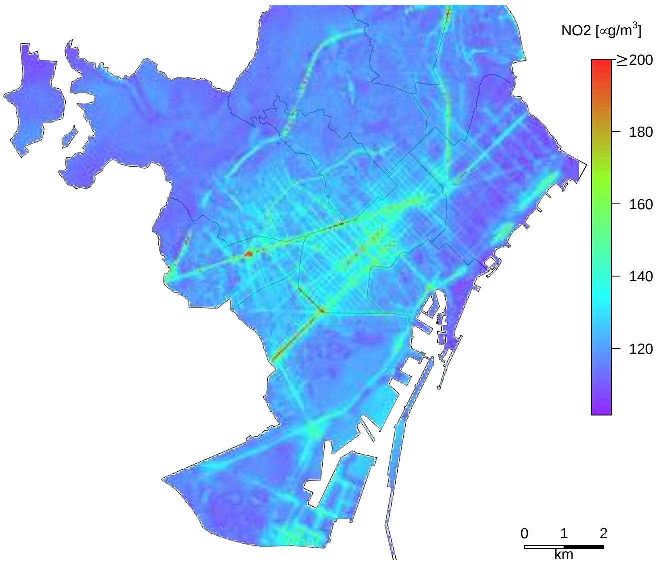


Fig. 2 19th daily maximum NO2 concentration levels at the municipality of Barcelona during 2019.

### E. *Conclusion and Future Enhancement*

The presented bias-adjustment methodology permits to perform reliable diagnosis of the annual NO2 levels. The diagnosis of Barcelona for 2019 puts in evidence that the $NO_2$ annual mean is systematically exceeded in many locations of the city, especially in the *Eixample* central district, and the region close to the port. On the other hand, the legislated hourly 19th daily maximum (140 $\mu g/m^3$) is exceeded in some heavily trafficked regions.

As future works, we plan to apply other state-of-the-art data fusion methodologies, such as Universal Kriging, to compare their performance with the present methodology. The future goal is to improve data fusion methodologies in the urban context to perform better reanalyzes which are key for health impact studies.

### F. *ACKNOWLEDGEMENTS*

The authors want to thank DTES for providing observational data through the XVPCA. This work has been carried out in the framework of a project with ATM-DTES. The first author has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433.

### *References*

[1] Benavides, J., Snyder, M., Guevara, M., Soret, A., Pérez García-Pando, C., Amato, F., ... & Jorba, O. CALIOPE-Urban v1. 0: coupling R-LINE with a mesoscale air quality modelling system for urban air quality forecasts over Barcelona city (Spain). Geoscientific Model Development, Pp. 2811-2835, 2019.

[2] Guevara, M., Tena, C., Porquet, M., Jorba, O., and Pérez García-Pando, C.: HERMESv3, a stand-alone multi-scale atmospheric emission modelling framework – Part 2: The bottom–up module, Geosci. Model Dev., Pp. 873–903, 2020.

## *Author biography*

**Jan Mateu Armengol** holds a Bachelor degree in Industrial Engineering from the Universitat Politècnica de Catalunya (Spain, 2013). He obtained a MSc. degree in numerical heat transfer and fluid flow by the Universidade Estadual de Campinas (Brazil, 2015). After receiving his joint-PhD degree in mechanical engineering from both the Universidade Estadual de Campinas and the École Centrale Paris (France, 2019), he collaborated with L'École Polytechnique (France, 2020) as a postdoctoral researcher working on Bayesian inference and uncertainty quantification on reduced chemical schemes. In 2020, he obtained a postdoctoral funding at the BSC from the STARs program (Marie-Sklodowska-Curie Action COFUND program) to join the Atmospheric composition group, as well as the Earth System Services group, to apply uncertainty quantification and data assimilation methodologies to urban air quality simulations.

# Predicate-Based Filtering for Multi-GPU Utilization in Directive-Based Programming

Kazuaki Matsumura, Simon Garcia de Gonzalo, Antonio J. Peña
Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {kazuaki.matsumura, simon.garcia, antonio.pena}@bsc.es

***Keywords*—*Multi-GPU, OpenACC, Compiler, Code Generation.***

## I. Extended Abstract

Designing and building supercomputers is a complex task in the field of high-performance computing (HPC). The hardware, middleware and algorithms need to effectively collaborate to achieve ideal results for massive and practical problems. To facilitate the easy usage of supercomputers, compiler technologies have been developed with highly automated program optimizations that use domain-specific knowledge and understandings of target architectures [1].

Directive-based programming has been employed for enabling accelerator use, while replacing vendor-specific coding with directive insertion. Keeping software portability with minimum engineering efforts upon sequential code, OpenACC and OpenMP are now widely used for accelerator programming [2], [3]. However, pursuing ideal performance is often challenging. The bare insertion of directives by the programmers exposes less program characteristics for the compilation; thus, programmers aiming at better efficiency are forced to reshape their code merely for adjusting to the environment such as compilers, software stacks and heterogeneous architecture.

While keeping the productivity, our research extends OpenACC to exploit further optimization opportunities. In a portable fashion that relies on other compilers, our approach provides an environment which enables dynamic analysis of computation and perform on-the-fly kernel specialization. Considering the high memory latency of GPUs, we add a novel code-translation technique named *predicated-based filtering* to automate multi-device utilization. We never split loop ranges nor introduce fine dependency analysis, but divide data ranges to be updated on each device. This idea allows to distribute highly-tuned code without changing code structure nor parallelism.

### A. JACC: Runtime-Extended OpenACC

We build JACC, a just-in-time compilation system for OpenACC, in which input directives are replaced with runtime routines. JACC hides every OpenACC feature behind a provided runtime library to cushion dependency to specific compilers. Once a kernel is compiled at first execution, its device code is cached to be reused for subsequent launches. Even though JACC is developed upon existing compilers, it allows calling of CUDA routines and kernels through its library. Fig. 2 shows the converted code of Fig. 1 to call runtime routines. First, combined directives (e.g. `parallel loop` of Line 2 of Fig. 1) are decomposed into three basic directives of `parallel`, `loop` and `data`. Then, for each directive, JACC inserts corresponding routines that are implemented in its library, shown in Fig. 2 (Lines 2, 5 and 12).

```
1  #pragma acc data copyout(x[0:N]) present(y)
2  #pragma acc parallel loop
3  for(int i=0; i<N; i++) x[i] = y[i] * y[i];
```

Fig. 1. Accelerator programming in OpenACC

```
1  /* Entry of #pragma acc data */
2  jacc_create(x, N * sizeof(float));
3
4  /* #pragma acc parallel loop */
5  jacc_kernel_push(
6    "#pragma acc parallel present(x, y)\n"
7    "#pragma acc loop\n"
8    "for(int i=0; i<N ; i ++) /* ... */",
9     /* args */, /* flags */);
10
11 /* Exit of #pragma acc data */
12 jacc_copyout(x, N * sizeof(float));
```

Fig. 2. Converted code by JACC (arguments omitted)

During the execution, JACC data-related routines that wrap OpenACC routines (Lines 2 and 12 of Fig. 2) assume the roles of the original directives. The routine `jacc_kernel_push` launches kernel execution while accepting source code in a string with arguments that hold runtime information (Lines 5-9 of the same figure). It should be noted that the `loop` directive is used for marking parallelism; therefore, the directive is kept in kernel strings. When the routine finds no compiled kernel for given source code or needs to update existing kernels, function code is generated to emit device code by a specified compiler and to have additional arguments for code extension. After linked dynamically, this function is called through a foreign function interface (FFI). JACC's library for each routine is extended to collect runtime information and support dynamic features.

### B. Multi-GPU Utilization with Predicates

Towards further utilization of intra-kernel parallelism, we combine multi-GPU execution with JACC. Whereas previous studies have persistently focused on loop splitting over plural GPUs [4], [5], this work divides data regions that each GPU updates to support real applications that usually entangle memory accesses among loop iterations.

Our technique, named *predicate-based filtering*, limits memory accesses depending on data regions that the GPU writes to, assuming that redundant computational code and parameters do not degrade performance due to low computational latency and high memory latency on GPUs. First, we introduce

```
1  a[i]=x; b[i]=a[i]; x=c[j]; a[k]=x; b[k]=a[k];
```

```
1  /* a[i]=x */
2  ((a_lb<=i && a_ub>=i)||
3   (b_lb<=i && b_ub>=i)) ? a[i]=x:a[i];
4  /* b[i]=a[i] */
5  ((b_lb<=i && b_ub>=i)) ? b[i]=a[i]:b[i];
6  /* x=c[j] */
7  x=((a_lb<=k && a_ub>=k)||
8     (b_lb<=k && b_ub>=k)) ? c[j]:0;
9  /* a[k]=x */
10 ((a_lb<=k && a_ub>=k)||
11  (b_lb<=k && b_ub>=k)) ? a[k]=x:a[k];
12 /* b[k]=a[k] */
13 ((b_lb<=k && b_ub>=k)) ? b[k]=a[k]:b[k];
```

Fig. 3. Example of predicate-based filtering in C code. Original (up) and filtered code (down). References to array `a` have predicates for updating array `b` and itself (Lines 2-3 and 10-11), the references to array `b` have for itself (Lines 5 and 13), and the reference to array `c` has for array `a` and `b` (Lines 7-8).

data ranges for each updated array so that array writes can be filtered based on the assigned range. For instance, in C code, array write `a[i]*=2` is rewritten to `(a_lb <= i && a_ub >= i) ? a[i]*=2 : a[i]`, where `a_ub` and `a_lb` indicate the upper and lower bound of array `a`, that are specified depending on the GPU. In Fortran, since there is no nested assignment, we use IF statement for filtering, with subsequent ELSE statement which contains an assignment of the same expression ( `a(i)=a(i)` ) that is later optimized away but facilitates compiler analysis. Additionally, we develop data-flow analysis for the innermost parallel region in each kernel to detect data dependencies between arrays. Then, we filter them to restrict accesses while solving dependencies as shown in Fig. 3. This analysis converts both array and variable references into the static single assignment (SSA) form, and iteratively finds dependencies among array accesses.

Data ranges linked to given pointers are tracked through JACC's runtime routines, and managed in a red-black tree as OpenACC compilers do [6]. Data transfers are invoked to send updated data across GPUs after each kernel execution. Device-memory allocations and host-to-GPU communications are replicated on all the GPUs and the primary GPU is used for GPU-to-host transfers. To guarantee the result of our analysis, we check kernel arguments so as to duplicate computation and disable communications on data that are referred through more than two pointers which at least one of them is read and one is written. When several pointers share the same array to update, we merge their access ranges to follow the widest. The necessary computation for array-write indexing is always duplicated. Regarding reduction or variable writes that are explicitly exported to host, we filter the computation based on the range of the outermost parallel iterator.

While being applicable to all OpenACC kernels as far as array writes are concerned, our filtering technique needs to duplicate execution on each GPU when references between split ranges are found inside the kernel. We alleviate this restriction by leveraging dimensional information.

In order to avoid lower performance due to data distribution overheads, we enable multi-GPU execution for each kernel in an adaptive way, while otherwise duplicating computation on all GPUs and performing no GPU-to-GPU communication.

### C. Results

We integrate predicated-based filtering into JACC, which translator is implemented as a XcodeML [7] converter. We measure the performance changes of our proposed techniques using the NVIDIA Tesla V100 SXM2 GPUs (16GB Memory) on NVIDIA DGX-1. Fig. 4 shows the partial results of our experiments.

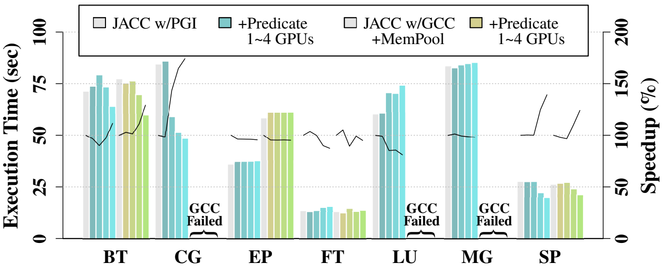


Fig. 4. Performance scaling of predicate-based filtering using NPB with PGI and GCC (top). The blue bars are results of predicate-based filtering with PGI and green bars are with GCC. The execution time is shown by the bar and the speedup by the line.

## II. Acknowledgment

We acknowledge the mentorship by a former BSC researcher Leonel Toledo.

## References

[1] M. J. Wolfe, C. Shanklin, and L. Ortega, *High Performance Compilers for Parallel Computing*. USA: Addison-Wesley Longman Publishing Co., Inc., 1995.

[2] T. O. Organization, "Openacc," 2011. [Online]. Available: https://www.openacc.org/

[3] T. O. ARB, "Openmp," 1997. [Online]. Available: https://www.openmp.org/

[4] K. Matsumura, M. Sato, T. Boku, A. Podobas, and S. Matsuoka, "Macc: An openacc transpiler for automatic multi-gpu use," in *Supercomputing Frontiers*, R. Yokota and W. Wu, Eds. Cham: Springer International Publishing, 2018, pp. 109–127.

[5] T. Komoda, S. Miwa, H. Nakamura, and N. Maruyama, "Integrating multi-gpu execution in an openacc compiler," in *2013 42nd International Conference on Parallel Processing (ICPP)*. IEEE, 2013, pp. 260–269.

[6] M. Wolfe, S. Lee, J. Kim, X. Tian, R. Xu, B. Chapman, and S. Chandrasekaran, "The openacc data model: Preliminary study on its major challenges and implementations," *Parallel Computing*, vol. 78, pp. 15–27, 2018.

[7] O. C. P. R. CCS), "Xcodeml," 2009. [Online]. Available: https://omni-compiler.org/xcodeml.html

**Kazuaki Matsumura** Kazuaki Matsumura received the BE degree from University of Tsukuba in 2017 and the MSc degree from Tokyo Institute of Technology in 2019. He is currently working at Barcelona Supercomputing Center while being affiliated with its doctoral program. His research interests include program optimization and compilers for high-performance computing.

# Constraining the chemical composition of particulate matter in an atmospheric chemistry model

Hector Navarro-Barboza*, Marco Pandolfi†, and Oriol Jorba*
*Barcelona Supercomputing Center, Barcelona, Spain
†Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Spain
E-mail: {hector.navarro, oriol.jorba}@bsc.es, marco.pandolfi@idaea.csic.es

***Keywords*—Particulate matter, chemical composition, organic aerosols.**

## I. Extended Abstract

Atmospheric aerosols have a significant impact on the climate system [1]. Depending on their chemical composition, aerosols cool the atmosphere directly by scattering solar radiation and indirectly through aerosol–cloud interactions, or warm the lower atmosphere by absorbing visible solar radiation. The most prevalent types of absorbing aerosols are black carbon (BC) and mineral dust. However, recent analyses from laboratory and field experiments have provided strong evidence for the existence of some organic aerosols (OA) with absorbing properties known as "brown carbon" [2]. The overall impact of aerosols in the atmosphere is still uncertain and can only be quantified by using numerical models. A prior step for that implies a proper representation of their abundance and chemical composition. However, atmospheric chemistry models have historically shown difficulties to simulate both properties, particularly OA [3].

In this work, we present the first step towards better constraining the burden and radiative forcing of aerosols in the atmosphere. We combine different observational datasets to evaluate the mass and chemical composition of aerosols (with special attention to OA) simulated by the Multiscale Online Nonhydrostatic AtmospheRe CHemistry (MONARCH) model developed at BSC. For that, we use a wide range of in-situ surface measurements of aerosol particulate matter (PM) and chemical composition conducted by IDAEA-CSIC at three locations in northeast Spain.

### A. Observational data

The Environmental Geochemistry and Atmospheric Research group of the Institute of Environmental Assessment and Water Research from the Spanish Research Council (IDAEA-CSIC) maintains three monitoring super-sites (urban background Barcelona station (BCN), regional background Montseny station (MSY) and remote background Montsec station (MSA); see Figure 1) with the aim to study the physical and chemical properties of atmospheric aerosol particles and their climate effects. In this work, we use data from 24h offline filter samples of PM fractions ($PM_1$, $PM_{2.5}$, and $PM_{10}$) and subsequent chemical analyses from the three sites, and online measurements of the non-refractory components of $PM_1$ with hourly resolution with an Aerosol Chemical Speciation Monitor (ACSM) available at the BCN station. Both datasets provide valuable information of the aerosol particulate matter concentration and their chemical composition, namely concentration of sulfate-nitrate-ammonium-BC-OA constituents. Further information on the primary and secondary fraction of OA is also derived from the observations.

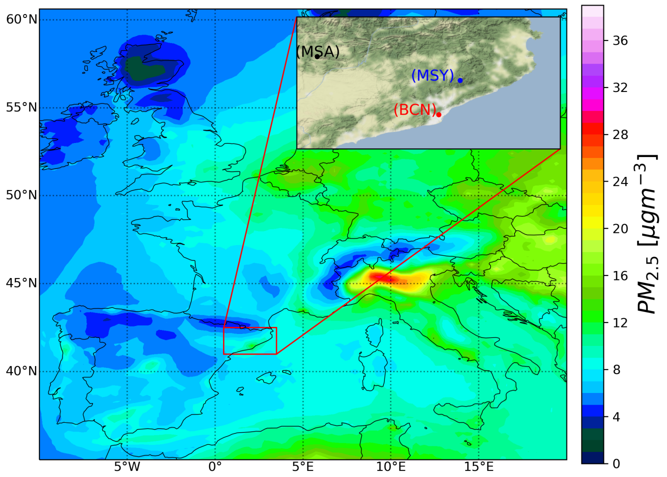


Fig. 1. Annual mean surface concentration of $PM_{2.5}$ for 2018 simulated with the MONARCH model. Top-right panel shows the IDAEA-CSIC monitoring stations.

### B. MONARCH model and simulation setup

The MONARCH model [4][5] is designed to study the chemistry processes and their interactions within the atmospheric system. The model couples the governing equations of meteorology and chemistry using an online approach and can be run on global or regional domains with telescoping nest capabilities. The system solves the tropospheric chemistry with detailed gas-phase and aerosol mechanisms.

The aerosol components considered by the model are: mineral dust, sea salt, BC, sulfate, nitrate, ammonium, primary OA, secondary organic aerosol (SOA), and the unspeciated fraction of emitted PM. A simple non-volatile SOA scheme accounts for the contribution of anthropogenic, biomass burning, and biogenic formation [6]. Emissions of chemical species are provided by the HERMESv3 emission model [7] that pre-processes anthropogenic, biomass burning, soil and ocean emissions from state-of-the-art emission inventories. Other natural emissions like desert dust, sea salt of biogenic emissions are computed online by MONARCH.

For the present study, MONARCH was configured with a regional domain covering Europe at 0.2° horizontal resolution and 24 vertical layers with the top of the atmosphere at 50 hPa. Meteorology and chemistry boundary conditions were obtained from ECMWF IFS and Copernicus CAMS global forecasting systems, respectively. The CAMS-REG-Anthrov3.1 anthropogenic emission inventory was used. An annual simulation of 2018 year was conducted and the concentrations of all aerosol components considered by the model stored with hourly resolution.

### *C. Results*

The annual mean surface concentration of $PM_{2.5}$ over Europe is shown in Figure 1. Higher concentrations are found over central and eastern Europe with hot-spots in different regions of the continent (e.g., the Po Valley). The northeast Spain is characterized by remarkable high levels.

The composition of the $PM_{2.5}$ simulated over the BCN site is presented in Figure 2 for January 2018. Good agreement with the observations is observed with important temporal variability throughout the year. Episodes of pollution are well detected, e.g. 30 January, where the fraction of SOA and nitrate to the total $PM_{2.5}$ increases significantly. Note how the organic fraction explains more than 50% of the total mass during most part of the month. From OA, the anthropogenic SOA (ASOA) represents the major contribution.

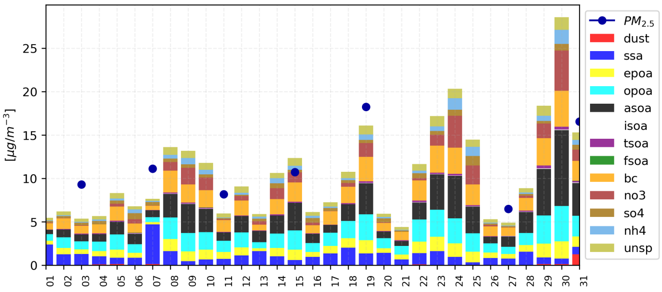


Fig. 2. Chemical composition of $PM_{2.5}$ for January 2018 in the BCN site. Colors represent daily mean concentration of each aerosol component simulated by the model: dust, sea salt, hydrophobic primary organic aerosol (EPOA), hydrophilic primary organic aerosol (OPOA), anthropogenic secondary organic aerosol (ASOA), isoprene secondary organic aerosol (ISOA), terpene secondary organic aerosol (TSOA), biomass burning secondary organic aerosol (FSOA), black carbon (BC), nitrate, sulfate, ammonium, and a fraction unspecified. Deep blue dots are the off-line filter observations.

A detailed evaluation of the OA simulated by the model is done with the ACSM measurements available in BCN site (Figure 3). An excellent agreement with the observations is obtained during May 2018, where two characteristic events of OA detected by the ACSM (red line) are well captured by the model (blue line) with a correlation coefficient of 0.9.

### *D. Conclusion and future work*

In this work, we present a detailed evaluation of particulate matter's chemical composition in three Spanish sites. The variability of the PM mass and composition is significant throughout the year and well captured by the model. The combination of different types of measurements allows identifying which components deserve further refinement in the model. Overall, the OA scheme in MONARCH performs very well, considering its simple design. Future work will extend the analysis to other European sites and target the evaluation to the optical properties.

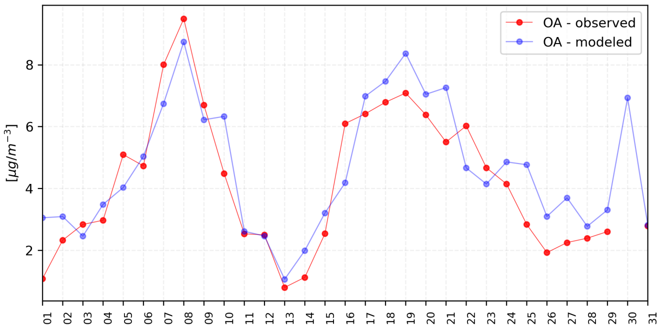


Fig. 3. Daily mean organic aerosol concentration in BCN site for May 2018. Blue: model result, Red: ACSM measurements.

## II. Acknowledgment

Hector Navarro-Barboza holds a PhD fellowship (FPI-SO PRE2018-084988) of the Severo-Ochoa project SEV-2015-0493-18-1 granted to the Barcelona Supercomputing Center. This work was supported by funding from the Ministerio de Ciencia, Innovación y Universidades as part of the BROWNING project (RTI2018-099894-BI00) and CAIAC project (PID2019-108990PB-100).

## References

[1] T. F. Stocker and co authors, "Climate change 2013: The physical science basis," *Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*, vol. 1535, 2013.

[2] A. Laskin, J. Laskin, and S. A. Nizkorodov, "Chemistry of atmospheric brown carbon," *Chemical reviews*, vol. 115, no. 10, pp. 4335–4382, 2015.

[3] K. Tsigaridis and co authors, "The aerocom evaluation and intercomparison of organic aerosol in global models," *Atmospheric Chemistry and Physics*, vol. 14, no. 19, pp. 10 845–10 895, 2014.

[4] A. Badia and co authors, "Description and evaluation of the multiscale online nonhydrostatic atmosphere chemistry model (nmmb-monarch) version 1.0: gas-phase chemistry at global scale," *Geoscientific Model Development*, vol. 10, pp. 609–638, 2017.

[5] M. Spada, "Development and evaluation of an atmospheric aerosol module implemented within the nmmb/bsc-ctm," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2015.

[6] S. J. Pai and co authors, "An evaluation of global organic aerosol schemes using airborne observations," *Atmospheric Chemistry and Physics*, vol. 20, no. 5, pp. 2637–2665, 2020.

[7] M. Guevara and co authors, "Hermesv3, a stand-alone multi-scale atmospheric emission modelling framework–part 1: global and regional module," *Geoscientific Model Development*, vol. 12, pp. 1885–1907, 2019.

**Hector Navarro-Barboza** received his BSc degree in Physics from Universidad Nacional Mayor de San Marcos (UNMSM), Peru in 2016. The same year, he started to work at Universidad del Pacifico as a research assistant. He completed his MSc degree in Physics with mention in geophysics from UNMSM, Peru in 2018. In December 2019, he enrolled the Atmospheric Composition group of Barcelona Supercomputing Center (BSC) as a PhD student.

# The Multilayer Community Structure of Medulloblastoma

I. Núñez-Carpintero*, D. Cirillo*, A. Valencia*

*Barcelona Supercomputing Center (BSC), **Nexus II Building** c/Jordi Girona, 29. 08034 Barcelona (Spain)
E-mail {iker.nunez, davide.cirillo, alfonso.valencia}@bsc.es

***Keywords* — Multilayer network, Network community analysis, Personalized medicine, rare disease multi-omics**

## EXTENDED ABSTRACT

### A. Introduction

Biomedical multilayer networks offer a wide range of possibilities for the interpretation of the molecular basis of diseases; a particularly challenging task in the case of rare diseases, where the number of cases is small in comparison with the size of the associated multi-omics datasets. In this work, we develop a dimensionality reduction methodology to identify the minimal set of genes that characterize disease subgroups based on their persistent association in the multilayer network at different levels of resolution.

We apply this approach to the study of a cohort of patients affected by medulloblastoma, a childhood brain tumor, using proteogenomic data. Our approach is able to recapitulate known medulloblastoma subtypes (accuracy > 94%) and offers a clear characterization of the associated gene functions, with the downstream implications for diagnosis and therapeutic interventions.

We verified the general applicability of our method by applying it to an independent dataset, achieving very high performances (accuracy > 98%). Overall, this approach opens the door to a new generation of multilayer-based methods able to overcome the specific dimensionality limitations of the rare disease datasets.

### B. Methods

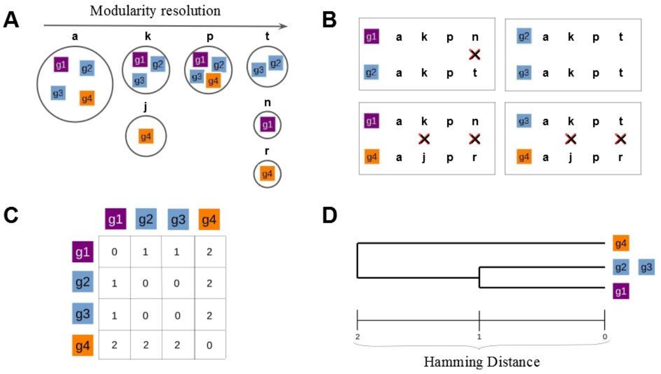
#### 1 – Multilayer Community Detection

We constructed a multilayer gene network composed of five layers: Reactome [1], Recon3D Virtual Metabolic Human [2], BioGRID molecular interactions [3], KEGG BRITE "Target-based Classification of Compounds" [4] and Monarch Disease Ontology (MonDO) [5]. In our definition of multilayer network, interlayer edges connect a gene in one layer with the same gene, if it exists, in another layer.

We performed a multilayer community trajectory analysis using the R package CmmD, that we implemented, and which depends on MolTi software [6].

By applying CmmD, it is possible to track different events throughout the process of community decomposition (**Figure 1A**) and use it as features for gene clustering (**Figure 1, B-D**) or other machine learning tasks, such classification and prediction.

The biomedical goal of the study is to identify the minimal number of genes that recapitulate the four established medulloblastoma subtypes (WNT, SHH, G3, and G4), confirmed in the original study by using bulk multi-omics data [7] (DNA methylation, RNA sequencing, proteomics and phosphoproteomics), that we use as the input data in our multilayer network based optimization procedure.



|  | g1 | g2 | g3 | g4 |
|---|---|---|---|---|
| g1 | 0 | 1 | 1 | 2 |
| g2 | 1 | 0 | 0 | 2 |
| g3 | 1 | 0 | 0 | 2 |
| g4 | 2 | 2 | 2 | 0 |

**Figure 1**. **Schematic representation of a multilayer community trajectory analysis**. For a given set of genes, we identify the multilayer communities to which they belong in a range of modularity resolution (A). We then compute the pairwise Hamming distances of the trajectories of communities visited by each gene (B). The corresponding distance matrix (C) is represented in the form of a dendrogram (D) used for clustering analysis.

#### 2 – Identification of minimal set of genes that define medulloblastoma subgroups

Identifying a minimal set of genes is crucial for both the definition of diagnostic signatures and the research on disease mechanisms. To achieve this goal, we performed a series of hierarchical clustering analyses (Ward's linkage method) where the similarity between two patients (A and B) was measured as the Jaccard index (J) of sets of altered genes selected using two parameters, θ and λ:

$$J(A_{\theta,\lambda}, B_{\theta,\lambda}) = \frac{A_{\theta,\lambda} \cap B_{\theta,\lambda}}{A_{\theta,\lambda} \cup B_{\theta,\lambda}}$$

θ defines the maximum Hamming distance (**Figure 1D**) allowed to include genes in the analysis, λ defines the maximum number of them that must co-occur in the same communities along their trajectories. For instance, with $\theta = 2$ and $\lambda = 4$, patient similarity is computed using sets of at most four genes that did not belong to the same communities at most twice along their multilayer community trajectories. For each of these clustering analyses, we identified the optimal number of patient clusters using the partitioning around medoids (PAM) algorithm [8].

Based on this approach, we formulated an optimization procedure to systematically evaluate values of θ and λ and identify the ones that maximize the accuracy of recapitulating patient stratification into the original four disease subtypes.

### C. Results

#### 1 - Multilayer community trajectories recapitulate and explain medulloblastoma subgroups

We achieved the highest accuracy (94.94%) with 5 clusters, by selecting for each patient those genes that are represented in the communities in sets of, at most, 6 ($\lambda = 6$), and that are always part of the same communities along their trajectories ($\theta = 0$). Strikingly, such high accuracy indicates that only a small portion of the genes altered in a patient is sufficient to accomplish an accurate patient segregation. This observation implies that the selected genes are tightly associated and never leave the communities they belong to along their trajectories. The identified values of $\theta$ and $\lambda$, optimized on 35 patients, correspond to an average dimensionality reduction of 87.56 % (SD = 0.44) per patient.

An important aspect of this result is that we identified 5 clusters, indicating that subtler stratas may exist (**Figure 2**), a feature we furthermore explored by expanding our methodology to a second non-overlapping multi-omics medulloblastoma patient cohort [9]. In this second dataset, patients were originally grouped into 6 biomedical subgroups (WNT, SHHa, SHHb, G3a, G3b and G4). The results of our analysis of this cohort exhibit an even higher accuracy (98.29%) with optimized parameters $\lambda = 3$ and $\theta = 0$, which corresponds to an average dimensionality reduction of 92.83% (SD=0.578).

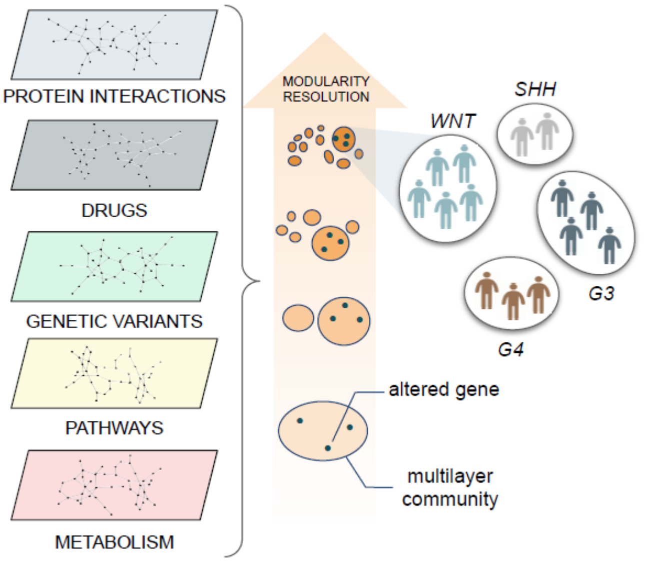


**Figure 2. Schematic representation of our dimensionality reduction optimization analysis.** Optimizing the detection of co-existing community structures of the multilayer network at different values of resolution suggested an extra stratification level within the classical medulloblastoma subgrouping (conformed by the 4 groups showed in the figure), pointing to the possible existence of a finer patient classification.

### *2- Provenance analysis of the identified gene communities*

Once we have been able to define the minimal set of genes that best define medulloblastoma subgroups, the high dimensionality reduction achieved allows for the detection of diagnostic signatures and the research on disease mechanisms underlying the associations identified through the multilayer network structure, via network enrichment analysis [10].
Overall, we found that the minimal set of genes found in all patients of WNT, SHH and G4 clusters are uniquely enriched in very specific associations in each layer, while G3-G4 and G3 clusters tend to display less specific enrichments (i.e. either several or none enriched associations). This reduction of enrichment specificity from WNT to G3 suggests an interesting parallel with the prognosis spectrum of the four classical subtypes, from best (WNT) to worst (G3) outcomes.

### *D. Discussion*

Molecular disease subtyping is a fundamental tool to achieve an effective patient stratification for clinical trials, preventive and therapeutic interventions. Rare diseases represent a challenging situation for any molecular analysis since they affect a small number of patients. Medulloblastoma is an illustrative example, with two subtypes being very well distinguishable (SHH and WNT groups) while two others are far less characterized (G3 and G4 groups).

In our vision, a meaningful molecular subtyping of rare diseases can be achieved by leveraging the wealth of biomedical information that is available in public knowledge bases and that can be integrated in the form of a multilayer network. In this work in particular, we achieved patient stratification by means of structural features (multilayer community trajectories) extracted from a general-purpose multilayer network; representing a way to both identify the minimal set of genes that characterize the subgroups and, most importantly, to retrieve and analyze the multiple associations among the identified genes, enabling the reach of a high level of interpretation of the patient subgroups and the spectrum of prognosis that characterize them, from best (WNT) to worst (G3) outcomes. This way of accomplishing two objectives with one action constitutes the main achievement of our approach [11].

#### *References*

[1] Fabregat, A. *et al.* (2017). *BMC Bioinformatics*, 18(1), p. 142. doi: 10.1186/s12859-017-1559-2.

[2] Brunk, E. *et al.* (2018). *Nature Biotechnology*, 36(3), pp. 272–281. doi: 10.1038/nbt.4072.

[3] Oughtred, R. *et al.* (2019). *Nucleic Acids Research*, 47(D1), pp. D529–D541. doi: 10.1093/nar/gky1079.

[4] Kanehisa, M. *et al.* (2019). *Nucleic Acids Research*, 47(D1), pp. D590–D595. doi: 10.1093/nar/gky962.

[5] Mungall, C. J. *et al.* (2017). *Nucleic Acids Research*, 45(D1), pp. D712–D722. doi: 10.1093/nar/gkw1128.

[6] Didier, G., Brun, C. and Baudot, A. (2015). *PeerJ*, 3, p. e1525. doi: 10.7717/peerj.1525.

[7] Forget, A. *et al.* (2018). *Cancer Cell*, 34(3), pp. 379-395.e7. doi: 10.1016/j.ccell.2018.08.002.

[8] Kaufmann, L. and Rousseeuw, P. (1987). *Data Analysis based on the L1-Norm and Related Methods*, pp. 405–416.

[9] Archer, T. C. *et al.* (2018). *Cancer Cell*, 34(3), pp. 396-410.e8. doi: 10.1016/j.ccell.2018.08.004.

[10] Signorelli, M., Vinciotti, V. and Wit, E. C. (2016). *BMC Bioinformatics*, 17(1), p. 352. doi: 10.1186/s12859-016-1203-6.

[11] Núñez-Carpintero, I. *et al.* (2021). *iScience*, p. 102365. doi: 10.1016/j.isci.2021.102365.

## *Author biography*

Iker Núñez was born in Valladolid, Spain, in 1995. He received the BSc degree in Biology from Universidad de Alcalá (UAH), Spain, in 2017, and the MSc degree in Biomedical Research from Universidad de Valladolid (UVA) and Molecular Biology and Genetics Institute (IBGM) of Valladolid, Spain, in 2018. Since October 2018, he is a Ph.D. student at Computational Biology Group within the Life Sciences Department of Barcelona Supercomputing Center (BSC), Spain. His current main research interests include Network Analysis, Deep Learning and Personalized Medicine of rare diseases and cancer.

# Epigenetic Characterization of Cholangiocarcinomas

Winona Oliveros[#1], Sandra Peiró[*2], Marta Melé[#3]

[#]*Barcelona Supercomputing Center (BSC), Spain*
[1]winona.oliveros@bsc.es, [3]marta.mele@bsc.es
[*]*Vall d'Hebron Institute of Oncology (VHIO), Spain*
[2]speiro@vhio.net

***Keywords*— Cholangiocarcinoma, RNA-Seq, ATAC-Seq, Methylation**

## Extended Abstract

Cholangiocarcinoma (CCA) is a rare type of cancer and accounts for 10-20% of primary liver diagnosed cancers, being the second most common hepatobiliary malignancy.

The only prospect of a long-term cure for these malignancies is offered by surgical resection of the tumor. However, the prognosis of all patients remains poor due to the high rate of recurrence of these tumors. In addition, the majority of patients are not diagnosed until the disease is non-resectable, thus only being suitable for palliative chemotherapy or supportive care[1].

Recently, some groups have published data on CCA describing its complex pathogenesis involving many different molecular pathways, with some of them being potential therapeutic targets.

Among the potential therapeutic targets, mutations on the IDH1/2 gene are present in around 20% of CCA patients[2]. This mutation generates what is called an "oncometabolite" (D2HG) that is responsible for many of the biological effects associated to IDH1 mutations in cancer. The main effect of D2HG is to competitively inhibit a family of alpha-KG-dependent enzymes, such as TET and JmJC, leading to a global increase of DNA and histone methylation.

Now, there is the need to establish a comprehensive understanding on how IDH1 mutations leads to the alteration of chromatin states in CCA. To this end, here we analyze the epigenome of CCA PDXs in terms of DNA accessibility, DNA methylation and transcriptome profiling in IDH1 wild-type and mutant samples. Our main goal is to shed some light on the relation between the genetic and epigenetic architecture of this type of cancer by performing an integrative analysis to characterize this unique set of CCA patient-derived models.

### A. Objectives

In this study we propose to take advantage of the unique set of CCA patient-derived xenografts we have at our disposal to characterize them molecular and epigenetically. In addition, we also plan to try different IDH1/2 inhibitors to be able to study the underlying biology of their activity and identify new potential biomarkers.

This study can be subdivided in 4 specific aims. First, the establishment and characterization of Patient Derived Xenografts (PDX) from CCA IDH1 wild-type and mutant tumors.

Second, the creation of a reference epigenome of CCA PDXs IDH1 mutant and wild-type and performing transcriptome profiling. This data will be used to perform an integrative analysis and get a better understanding of the changes at the chromatin landscape occurring at both, tumor base level and between IDH1 mutant and wild-type tumors.

Third, testing the efficacy of various treatments in IDH1 mutant and wild-type CCA PDXs.

And finally, the identification of potential biomarkers that could predict the patient's response to each treatment, and thus that would improve patient stratification.

### B. Materials and Methods

Fresh tumor samples were collected from CCA patients (>90% metastatic tumor to the liver) and implanted into nude mice. At this stage, genetic alterations in 315 genes and 28 introns are evaluated by FoundationONE. CCA PDXs from nude mice are processed within 30min after the surgical resection of the tumor, digested and the single-cell suspension is subsequently cultured in Matrigel. These samples can keep growing for 2 weeks.

DNA accessibility profiling by ATAC-seq is performed to this set of samples. Quality of data was assessed using FastQC. Reads were first aligned to mm10 using BOWTIE2 and then we re-aligned the unmapped fraction of reads to hg38. Only unique aligning reads were collected and for all samples duplicates were removed using Picard. Reads mapping to the mitochondrial genome or to the ENCODE blacklisted regions were also removed from the analysis. ATACseqQC was used to further evaluated ATAC-seq-specific quality metrics. Peak calling was performed using MACS2 using the paired-end option (-f BAMPE). The final list of peaks was obtained by intersecting called peaks for mutant samples and wild-type samples separately and then merging both lists of peaks. Finally counts per peak were obtained using featureCounts.

Transcriptome profiling by polyA+ RNA sequencing was also performed to this set of samples. Quality of data was assessed using FastQC. Reads were first aligned to mm10 using hisat2 and then we re-aligned the unmapped fraction of reads to hg38. Counts per gene were obtained using featureCounts with the latest GENCODE version as reference (GENCODE v34).

DNA Methylation by a methylation array covering over 800.000 CpGs was assessed with this set of samples. The processing of this data was done by our collaborators. Beta values, differentially methylated CpGs and DMRs were obtained.

Public transcriptomic, methylation and variation data from TCGA CHOL were downloaded from GDC webpage (https://gdc.cancer.gov/about-data/publications/pancanatlas). Data for 36 cholangiocarcinoma samples was obtained and preprocessed to be merged with our PDXs.

Differential expression analysis between mutant and wild-type IDH1 PDXs was performed using the R package DESeq2. WebGestalt was used to perform Gene Set Enrichment analysis (GSEA) and Over-representation analysis (ORA). KEGG, REACTOME and Wikipathway Cancer were used as query databases.

MOFA+ R package was used to perform an integrative analysis and uncover variation in this complex dataset containing multiple sources of heterogeneity. Minimum required sample size for this package to run is 12 samples, therefore we integrated home-made PDX data with public

TCGA samples. A total of 42 samples were used as input (6 PDXs and 35 TCGA samples).

### *C. Results*

**Differential expression analysis**. From the 58.884 genes with expression data, we found 888 genes to be differentially expressed between mutant and wild-type PDXs (Fig. 1). Gene Set Enrichment analysis revealed that pathways related to DNA repair were upregulated in IDH1-mutant PDXs, discarding our initial hypothesis of a "BRCA-ness" phenotype linked to IDH1 mutations. In addition, many signaling pathways involved in cholangiocarcinoma development and progression, such as ErbB, STAT3 and Pi3K-AKT-mTOR pathways, were also found to be upregulated in IDH1-mutant PDXs compared to IDH1 wild-type PDXs. Finally, genes linked to stem-cell capacity were downregulated in IDH1-mutant PDXs.

**Differential methylation analysis**. First, differential methylation of individual CpG sites was assessed. As expected from literature, the majority of changes were gains of methylation in mutant IDH1 PDXs. (2.840 CpG sites gained methylation in mutants from the set of 4.300 differentially methylated CpG found). Additionally, DMR calling was performed using windows of 2.5Kb long covering a median of 15 CpG sites. We found 233 differentially methylated regions, of which 95% were gains of DNA methylation in mutant samples, as expected.

**DNA accessibility analysis.** We were currently unable to perform differential peak calling because of a low number of reads per sample. Comparison of counts per peak between IDH1 mutant and wild-type PDXs showed that IDH1 mutant samples had significantly higher counts than IDH1 wild-type samples (Fig. 2). Further work needs to be carried out to know if this translates to having broader or higher peaks.

**MOFA+.** Integrative analysis of all our omics together with MOFA+ resulted in 10 learned factors explained by a combination of different omics (Fig. 3) and collectively explaining between ~40 and ~80% of variation per type of data. Factor 5 was significantly associated with IDH1 mutation and was mainly explained by methylation variation.

### *D. Conclusions*

Further work needs to be conducted to interpret MOFA results and more PDX samples are needed to increase our power to detect significant enrichments and be more confident with our main findings.

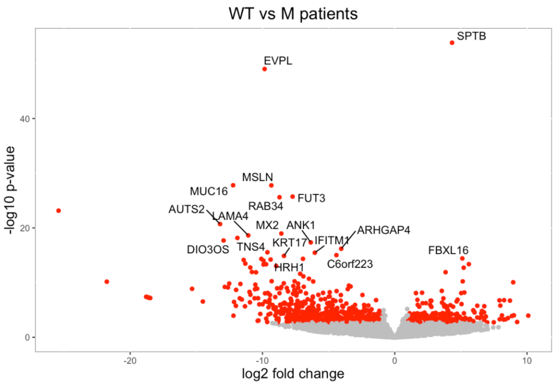


Fig. 1 Volcano plot with Differential expression analysis results. Log2 Fold changes < 0 mean the gene is overexpressed in mutant PDX. Most significant genes are specified.

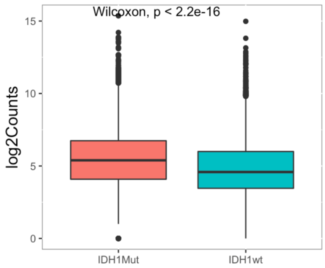


Fig. 2 Boxplot comparing counts per peak between IDH1 mutant and wild-type samples. P value from Wilcoxon test is specified.

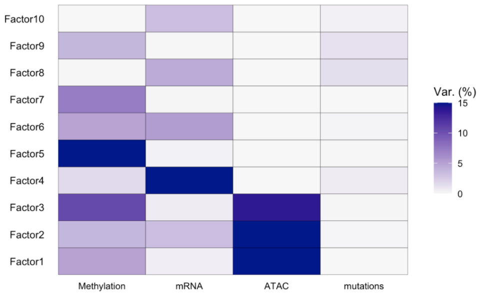


Fig. 3 Variance explained for each omic per learned factor with MOFA+.

*ACKNOWLEDGEMENTS*

We want to acknowledge our collaborators Jose Luis Sardina and Antonio Gomez for processing and analyzing DNA methylation data. We also want to acknowledge Queralt Serra for all wet-lab experiments. Finally, we want to acknowledge Tian Tian for his expertise on the wet-lab experiments and the fruitful discussions of the preliminary results.

### *References*

[1] Banales, J. M. et al. Expert consensus document: Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the European Network for the Study of Cholangiocarcinoma (ENS-CCA). Nat Rev GastroenterolHepatol 13, 261-280, doi:10.1038/nrgastro.2016.51 (2016).

[2] Rizvi, S., Khan, S. A., Hallemeier, C. L., Kelley, R. K. & Gores, G. J. Cholangiocarcinoma - evolving concepts and therapeutic strategies. Nat Rev Clin Oncol 15, 95-111, doi:10.1038/nrclinonc.2017.157 (2018).

## *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.

# VIA: A Smart Scratchpad for Vector Units with Application to Sparse Matrix Computations

Julián Pavón[1,2], Osman Unsal[1]and Adrian Cristal[1,2]
[1]*Barcelona Supercomputing Center*
*firstname.lastname@bsc.es*
[2]*Universitat Politècnica de Catalunya*
*firstname.lastname@upc.edu*
Barcelona, Spain

***Abstract*—Sparse matrix operations are critical kernels in multiple application domains such as High Performance Computing, artificial intelligence and big data. Vector processing is widely used to improve performance on mathematical kernels with dense matrices. Unfortunately, existing vector architectures do not cope well with sparse matrix computations, achieving much lower performance in comparison with their dense counterparts.**

**To overcome this limitation, we present the Vector Indexed Architecture (VIA), a novel hardware vector architecture that accelerates applications with irregular memory access patterns such as sparse matrix computations. There are two main bottlenecks when computing with sparse matrices: irregular memory accesses and index matching. VIA addresses these two bottlenecks with a smart scratchpad that is tightly coupled to the Vector Functional Units within the core. As a result, VIA achieves significant performance speedup over highly optimized state-of-the-art C++ algebra libraries. On average, VIA outperforms sparse matrix vector multiplication, sparse matrix addition and sparse matrix matrix multiplication kernels by 4.22×, 6.14× and 6.00× respectively.**

***Index Terms*—Sparse Algebra, Vector Computing, Scratchpad Memory**

## I. Introduction

Many applications can potentially benefit from vectorized execution for better performance, higher energy efficiency and greater resource utilization [4]. Ultimately, the effectiveness of a vector architecture depends on the quality of the vectorized code [5]. Sparse matrix operations are a clear example of computations difficult to vectorize [3]. Such computations are a key kernel in High Performance Computing (HPC), Artificial Intelligence (AI) and big data workloads. In particular, two such killer-applications are Sparse Matrix Vector multiplication (SpMV) and Sparse Matrix Matrix Multiplication (SpMM). There are two intertwined obstacles against efficient execution of sparse matrix computations on vector architectures: (1) existing sparse matrix representations are not easily vectorizable, and (2) current vector hardware implementations are not optimized for sparse matrix operations.

In this paper, we propose the *Vector Indexed Architecture* (VIA), a vector architecture that aims at accelerating sparse matrix computation. VIA features a smart scratchpad memory specially designed to cope with sparse-dense (SpMV) and sparse-sparse (SpMM) matrix computations.

## II. VIA: Key Design Ideas

Sparse matrix kernels present a set of challenges for current Vector Architectures: 1) High usage of inefficient memory indexed instructions and 2) index matching operations. The key idea in VIA is to attach a scratchpad memory (SPM) next to the vector functional units. The VIA SPM features two mapping techniques to tackle both challenges. For *sparse-dense* kernels, VIA performs a direct-mapped access to the SPM. The indexed instructions are executed between the Vector Functional Unit(VFU) and the SPM in VIA, thus reducing memory traffic and releasing memory bandwidth to load the low-locality sparse matrix from main memory more efficiently. For *sparse-sparse* kernels, the SPM in VIA works as a CAM memory. CAM memories are specialized hardware structures that are particularly suitable for search and index matching algorithms. The CAM-based mapping technique, allows VIA to execute index matching operations between two input vectors in a single instruction.

## III. VIA: Design Implementation

VIA is composed of two main building blocks: a *Smart Scratch-Pad Memory* (SSPM) and the *Fused Indexed Vector Unit* (FIVU) (see Figure 1). FIVU is the control interface between the Vector Functional Units (VFU) and the SSPM.

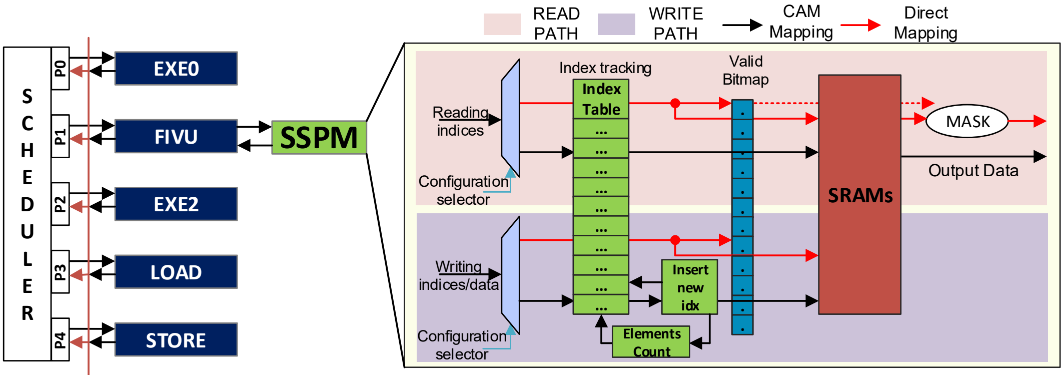
### A. The Smart Scratchpad Memory

The SSPM is a dedicated high bandwidth structure used to feed the VFU and it can be used in direct-mapped mode, or in CAM-based mode. The SSPM consists of three main building blocks: ❶ the SRAM cells to store the actual data; ❷ the valid bitmap to specify when an entry in the SRAM has been written before; and ❸ the Index tracking logic that provides the CAM-based functionality to SSPM.

**SRAM cells (SRAM):** stores the values to compute, e.g. for SpMV operations, SSPM stores the vector and for both SpMM and SpMA operations, SSPM stores the sparse row data and indices of only one of the input matrices. In our implementation, SRAM is built using four byte length blocks and each block stores a single value independently on the data length.

**Valid bitmap:** This structure is used in the direct-mapped mode as a written value indicator for the entries in the SRAM. It consists of a vector of bits, where each bit corresponds to an entry in the SRAM structure and determines whether an entry has been written.

**Index tracking logic:** This block implements SSPM-CAM functionality over the indexes. The index tracking logic consists of three key components: ❶ The index table, a CAM structure that



(a) (b)

Fig. 1. VIA building blocks: (a) interconnection between FIVU, VIA and the issue logic of an out-of-order core; (b) microarchitecture of SSPM. It consists of the index tracking mechanism (Index table, Insert new Idx and Elements Count), valid bitmap, and the storage system (SRAMs). Read and write paths are depicted separately.

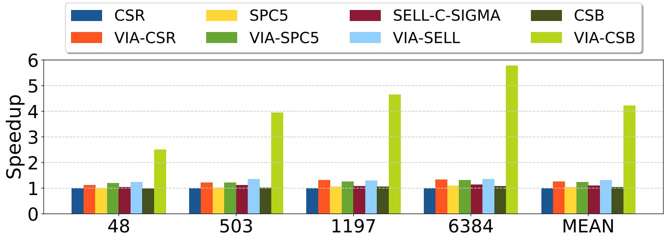


Fig. 2. Speedup for VIA SpMV kernel. Results are normalized to the CSR implementation for every category.

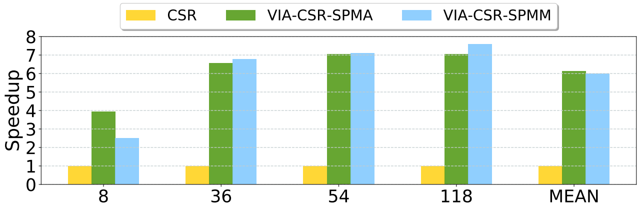


Fig. 3. Speedup for VIA SpMA and VIA SpMM. Both Kernels are normalized to their base CSR implementation.

stores the indices used to write data in the SRAM; ② The insertion logic, which inserts new indices and elements in order in the first available position in the index table and the SRAM respectively; and ③ The element count register, which holds the number of stored indices in the index table.

### B. The Fused Indexed Vector Unit

VIA introduces the *Fused Indexed Vector Unit* (FIVU) to operate over data stored in the SSPM. The FIVU works as the interface between the SSPM memory and the processor pipeline and minimally extends a generic Vector Functional Unit (VFU) with new pipeline stages to control operations to the SSPM.

## IV. Experimental Setup

We model and evaluate VIA using *Gem5* [1] to simulate an x86 full-system running an Ubuntu 16.04 OS with a 4.9.4 Linux Kernel. We simulate a single core processor using the out-of-order CPU and memory models, extended with the micro-architectural support and performance counters for VIA. We evaluate VIA efficiency using three representative sparse matrix kernels: Sparse Matrix Vector multiplication (SpMV), Sparse Matrix Addition (SpMA) and Sparse Matrix Matrix multiplication (SpMM). As input dataset, we use 1,024 sparse matrices from 56 different application domains of the *University of Florida Sparse Matrix Collection* [2].

## V. Evaluation

Figure 2 depicts performance results for VIA-SpMV kernel using different compressed representations on all the input dataset. The most noteworthy results are presented by the CSB (Compress Sparse Block) version. All the evaluated matrices were sorted by the CSB block density and evenly split among 4 categories. The x-Axis at Figure 2 shows the median non-zero values per block among each category. VIA SpMV achieves on average speedup of 4.22× with CSB; and average speedups of 1.25×, 1.24× and 1.31× over the CSR, SPC5 and Sell-C-$\sigma$ implementations repectively. The CSB format increases the locality of the input and output vectors, thus a chunk of the input vector needs to be placed in SSPM only once to compute with a single block. For the other formats, the indices to map the input or output vectors of two consecutive matrix values can be really sparse, thus the efficiency of VIA is limited to work as an accumulator for the output vector. Nevertheless, even with this limitation, VIA improves performance over the other formats by 1.26× on average. For the best usage case (executing with CSB VIA-SpMV), VIA: (1) reduces the total energy consumption (leakage + dynamic) by a factor of 3.8×. (2) increases the memory bandwidth by 2.5×.

Figure 3 shows the performance of the SpMA (VIA-CSR-SPMA column) and SpMM kernels. In a similar manner to SpMV, results were sorted and evenly split into 4 categories. As we use CSR format in both kernels construction, we used the non-zero elements per row as the criteria to sort the entire input dataset.

On average, VIA achieves 6.14× and 6.0× speedup on the input dataset for SpMA and SpMM respectively. In terms of energy, VIA reduces on average 5.6× and 5.1× of the total energy and increases the memory bandwidth by a factor of 2.1× and 3.2× for SpMA and SpMM respectively. The components of VIA allow to vectorize the index matching computation with a single vector instruction without any extra software comparisons. This capability helps to reduce the memory traffic, reduce the store-load forwarding and to increase the efficiency of the VFU over this kernel.

## VI. Author Biography

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

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

## References

[1] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib Bin Altaf, N. Vaish, M. Hill, and D. Wood, "The gem5 simulator," *SIGARCH Computer Architecture News*, vol. 39, pp. 1–7, 08 2011.

[2] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Transactions on Mathematical Software (TOMS)*, vol. 38, no. 1, p. 1, 2011.

[3] E. F. D'Azevedo, M. R. Fahey, and R. T. Mills, "Vectorized sparse matrix multiply for compressed row storage format," in *Computational Science – ICCS 2005*, V. S. Sunderam, G. D. van Albada, P. M. A. Sloot, and J. J. Dongarra, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 99–106.

[4] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Sixth Edition: A Quantitative Approach*, 6th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2017.

[5] N. Satish, C. Kim, J. Chhugani, H. Saito, R. Krishnaiyer, M. Smelyanskiy, M. Girkar, and P. Dubey, "Can Traditional Programming Bridge the Ninja Performance Gap for Parallel Computing Applications?" in *Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA)*, 2012, pp. 440–451. [Online]. Available: http://dl.acm.org.recursos.biblioteca.upc.edu/citation.cfm?id=2337159.2337210

# Climate Forecast Analysis Tools Framework

Núria Pérez-Zanón#1, An-Chi Ho#2, Francesco Benincasa#3, Pierre-Antoine Bretonnière#4, Louis-Philippe Caron*5,
Chihchung Chou#6, Carlos Delgado-Torres#7, Llorenç Lledó#8, Nicolau Manubens#9, Lluís Palma#10

*#Earth Science Department, Barcelona Supercomputing Center (BSC)*
1nuria.perez@bsc.es, 2an.ho@bsc.es,
3francesco.benincasa@bsc.es,4pierre-antoine.bretonniere@bsc.es,6chihchung.chou@bsc.es,7carlos.delgado@bsc.es,8llledo@bsc.es,9nicolau.manubens@bsc.es,10lluis.palma@bsc.es,

**Ouranos, 550 Sherbrooke St W, Montreal, Quebec H3A 1B9, Canada*
5caron.louis-philippe@ouranos.ca

***Keywords*— Climate Services, Climate forecast verification, Forecast calibration, Climate analysis tools**

## EXTENDED ABSTRACT

The climate forecast analysis tools provide functions implementing the steps required for the analysis of sub-seasonal, seasonal and decadal forecast and operational climate services, allowing researchers to manipulate climate data and apply state-of-the-art methods taking advantage of the high-performance computational resources. Researchers can share their methods while reducing development and maintenance cost. An ecosystem of R packages covering these needs is under continuous development.

### *A. Introduction*

Initialized Earth system predictions are made by starting a numerical prediction model in a state consistent with the observations corresponding to a specific day but adding perturbations on the initial conditions to generate an ensemble of simulations. The simulations run forward in time for sub-seasonal (around 30 days ahead), seasonal (from the 3 months to a year) and decadal (around ten years ahead). For each forecast horizon, the Earth system models are adapted to improve the representation of the environmental conditions that have a bigger impact on their predictability. Apart from the accuracy of initial atmospheric condition, the sources of predictability in sub-seasonal forecast comes from the stratospheric and surface representation whereas seasonal forecast is strongly affected by the sea surface temperature conditions (White et al., 2017). The source of predictability in decadal predictions is also affected by low-frequency modes on the extra-tropics sea surface temperature. However, the sub-seasonal to decadal communities share common scientific and technical challenges: initialization shock and drift; understanding the onset of model systematic errors; bias correction, calibration, and forecast quality assessment; model resolution; sources and expectations for predictability; and linking research, operational forecasting, and end-user needs (Merryfield et al., 2020).

At the Earth Sciences department of the Barcelona Supercomputing Center, the expertise in climate forecast research has traditionally been compiled in the s2dverification R package (Manubens et al., 2018) since its first release in 2009. The package provides tools implementing the steps required for the analysis of climate forecast, allowing researchers to share their methods while reducing development and maintenance cost. New packages have been developed to benefit from the available computational resources allowing researchers to conduct analysis that implies big data size and reducing the computation time.

Currently, 8 R packages (Table 1) are being maintained by the department and developed with contributions from external collaborators in the framework of European projects. In the next section, a description of the packages will be provided, whereas section 3 explains the flexibility and processing options. Finally, conclusions and future developments are summarized.

TABLE I
LIST AND DESCIRPTION OF PACKAGES IN THE CLIMATE FORECAST ANALYSIS TOOLS

| **Package name** | **Short description and link to CRAN** |
|---|---|
| easyNCDF | Read/write netCDF files into/from multidimensional R array.<br>https://CRAN.R-project.org/package=easyNCDF |
| multiApply | Apply functions to multiple multidimensional arrays or vectors allowing parallel computation<br>https://CRAN.R-project.org/package=multiApply |
| s2dverification | Functions for Forecast Verification and visualization<br>https://CRAN.R-project.org/package=s2dverification |
| s2dv | Adaptation of s2dverification to multiApply<br>https://CRAN.R-project.org/package=s2dv |
| CSTools | Methods for forecast calibration, statistical and stochastic downscaling, optimal forecast combination and tools to obtain tailored products.<br>https://CRAN.R-project.org/package=CSTools |
| CSIndicators | Sectorial Indicators for Climate Service (under-development in internal gitlab project) |
| ClimProjDiags | Climate extreme indices, evaluation of the agreement between models, weight and combination functions.<br>https://CRAN.R-project.org/package=ClimProjDiags |
| startR | Data retrieval and processing tools<br>https://CRAN.R-project.org/package=startR |

### *B. Methods in the Climate Forecast Analysis Tools*

The methods included in the climate forecast analysis tools aim to obtain a research result or climate service product from the climate forecast and reference datasets (e.g. reanalysis) by transforming the data and applying state-of-the-art methods:

- Different approaches to understand the links between climate variability and their impacts are explored by the researchers. That is the case, for instance, of the Madden-Julian Oscillation (MJO), a prominent feature of the tropical atmospheric circulation at sub-seasonal time scales, is known to modulate atmospheric variability in the Euro-Atlantic region (Lledó et al., 2020). Several atmospheric circulation patterns are already defined in the scientific literature and they are calculated by applying different methodologies (e.g.: Empirical Orthogonal Functions or area-weighted means).
- Skill metrics to assess the quality of their forecasts by comparing them against reference observation datasets can be deterministic, probabilistic and multivariate.
- Forecast post-processing refers to scientific methods to increase the quality of the forecasts, such as calibration methods to remove the systematic model error, or to increase the usability of the forecasts, such as downscaling techniques that allow achieving a finer resolution than the coarse original model resolution.

- Apart from essential climate variables, tailored indicators can help narrow the usability gap between pure science and stakeholders of key socio-economic sectors, such as renewable energy, air quality, agriculture or insurance, as well as to the general public.
- Visualization tools are a fundamental step to explore and communicate results in scientific writing and to end-users.

After reading the data from file to RAM, there is not a unique possible combination of methods, the users can skip some processes or apply multiple methods to the same data.

### *C. Tools Flexibility*

Originally, s2dverification package relied on a fixed structure of the data: an array (i.e. multi-dimensional object) in which each of the dimension corresponds to dataset identifier, member of the ensemble, start date for the initialization forecast date, lead time for the forecast time step, latitude and longitude positions defined in the data retrieval step. This structure is suitable for most of the transformations and methods, but several other needs may require extra dimensionalities, such as atmospheric level or weather type identification.

In this context, multiApply package was released as a variant of apply functions extending this paradigm. Its only function, Apply, efficiently applies functions taking one or a list of multiple unidimensional or multidimensional arrays (or combinations thereof) as input. This saves development time by preventing the R user from writing often error-prone and memory-inefficient loops dealing with multiple complex arrays. Also, a remarkable feature of Apply is the transparent use of multi-core through its parameter 'ncores'. Therefore, the latest packages (i.e. s2dv, CSTools and CSIndicators) are developed taking advantage of multiApply, and guidelines on how to develop functionalities have been written to be followed autonomous by researchers.

For the data retrieval and datasets processing steps, startR package has been developed to provide a flexible way to retrieve data from files, as well as, performing analysis when the size of involved data overcomes the available RAM memory by implementing the MapReduce paradigm, chunking the data and processing them either locally or remotely on high-performance computing systems, leveraging multi-node and multi-core parallelism where possible (Pérez-Zanón et al., 2021).

The result is that functions in the climate forecast analysis tools framework use multi-dimensional arrays with named dimensions, the users can set up parameters specifying the dimension(s) in which the functions should be applied, as well as, the number of cores to use in the computation. The learning curve is expected to show a big positive trend once the user gets an understanding of the multi-dimensional array with named dimensions as the main object in which data is stored.

### *D. Conclusion and Future Enhancement*

The success of the climate forecast analysis tools is proved by the number of peer-reviewed articles published by the department researchers, operational climate services deployed and projects successfully undertook. The tools are under continuous development exploring new methods and allowing extra transformations in a user-friendly way. Users support, internal training and discussions, as well as, dissemination are a priority at this stage.

### *E. ACKNOWLEDGEMENTS*

Part of this work was performed in the framework of the MEDSCOPE (MEDiterranean Services Chain based On climate PrEdictions) ERA4CS project (grant agreement no. 690462) funded by the European Union, the Copernicus Climate Change Service (C3S) (Metrics and Access to Global Indices for Climate Projections, MAGIC) and MEDGOLD (Turning climate-related information into added value for traditional MEDiterranean Grape, OLive and Durum wheat food systems, agreement no. 776467). We acknowledge all contributors to these tools.

### *References*

[1] L. Lledó, F.J. Doblas-Reyes, “Predicting daily mean wind speed in Europe weeks ahead from MJO status” Mon. Weather Rev. ,2020. DOI: 10.1175/mwr-d-19-0328.1.

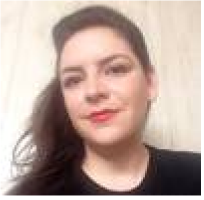
[2] N. Manubens, L-P. Caron, A. Hunter, O. Bellprat, E. Exarchou, NS. Fu Ckar, J. Garcia-Serrano, F. Massonnet, MM. En Egoz, V- Sicardi, E.C. Batt , C. Prodhomme, V. Torralba, N. Cortesi, O. Mula-Valls, K. Serradell, V. Guemas, F.J. Doblas-Reyes. “An R package for climate forecast verification” Environ. Model. Softw., 103: 29–42, 2018 DOI: 10.1016/j.envsoft.2018.01.018.

[3] W.J. Merryfield, J. Baehr, L. Batté, E.J. Becker, A.H. Butler, C.A.S. Coelho, G. Danabasoglu, P.A. Dirmeyer, F.J. Doblas-Reyes. “Current and emerging developments in subseasonal to decadal prediction”. Bulletin of the American Meteorological Society. American Meteorological Society 101(6): E869–E896. 2020 DOI: 10.1175/BAMS-D-19-0037.1.

[4] N. Pérez-Zanón, A. Ho, N. Manubens, F. Benincasa, P. Bretonnière. “startR: A tool for large multi-dimensional data processing” the 8th BSC Doctoral Symposium 2021(submitted)

[5] C.J. White, H. Carlsen, A.W. Robertson, Klein RJT, Lazo JK, Kumar A, Vitart F. “Potential applications of subseasonal-to-seasonal (S2S) predictions”. Meteorol. Appl., 315–325. 2017 DOI: 10.1002/met.1654.

## *Author biography*



**Núria Pérez-Zanón** is a postdoctoral researcher and member of the Data and Diagnostic teams in the Computational Earth Sciences group at BSC. With a background in Physics and Meteorology (degrees from the University of Barcelona; UB), she obtained her PhD from Rovira i Virgili University (URV) in 2017 on Climate variability and Change detection in the central Pyrenees using instrumental and paleoclimate proxy data. During the thesis, one year and a half of research stay in LOCEAN (Laboratoire d’Océanographie et du Climat). She is the author of 14 peer-reviewed articles in international journals. She is currently coordinating the development and maintenance of the R tools of the department.

# TunaOil: A Tuning Algorithm Strategy for Reservoir Simulation Workloads

Felipe Portella[*†‡], Josep Ll. Berral[*†]
[*]Petróleo Brasileiro S.A. (PETROBRAS), Rio de Janeiro, Brazil
[†]Barcelona Supercomputing Center, Barcelona, Spain
[‡]Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: felipe@portella.com.br, josep.berral@bsc.es

***Keywords*—Reservoir Simulation, Algorithm Configuration, Machine Learning, Parameter Tuning**

## I. Extended Abstract

The state-of-the-art techniques to tune the numerical parameters of reservoir simulators are based on running numerous simulations, specific for that purpose, to find good candidates. As the simulations for real petroleum fields require a considerable amount of time, optimizing parameters using this approach is costly in terms of time and computing resources. The main objective of this work, therefore, is to present a new methodology to optimize the numerical parameters of the reservoir simulations. It is common in the oil and gas (O&G) industry to use ensembles of models in different workflows to reduce the uncertainty associated with the forecasting of O&G production. We can leverage the runs needed to create such ensembles, to extract the information we can use to optimize the numerical parameters in future runs.

To achieve this, we mine past execution logs from many simulations with different numerical configurations and build a performance model that is based on features extracted from the data. This performance model takes general information about petroleum fields and the simulation parameters as inputs, allowing it to generalize to different unseen reservoir models. Experiments show that the presented system can correctly produce good configurations in a much-reduced time, within a history matching workflow that generates hundreds of simulations.

### A. *Reservoir Simulations and History Matching*

Essentially, reservoir simulation allows engineers to replicate the history of the production of oil, gas, and water from the reservoir over a time frame to forecast the future; this provides answers to a series of questions that are critical to different business strategies, for exploiting the oilfield.

Engineers are constantly looking for efficient tools to tune the simulation process to make it faster and achieve better decisions. The reservoir simulators available on the market allow users to tune numerical parameters, which can affect the performance and quality of the simulation significantly. However, these numerical parameters vary among simulators, making the selection of them difficult. Moreover, the parameter space is big and co-relations can exist between parameters, making this a non-trivial manual task.

The utility of the reservoir model, however, results from its ability to predict the behavior of the reservoir field in terms of the production of water, oil, gas, pressure, etc. To calibrate the model, engineers use a method called *History Matching* (HM). The explicit purpose of HM is to assign values to the parameters of the model to be optimized, such that it replicates the behavior observed during a past production period, leading to better forecasting. One tool used for automatic HM is the family of Kalman filters.

A Kalman filter (KF) is a mathematical method – robust to noise in the data – that uses all of the observed measurements, to produce estimates of unknown variables in linear systems. The ensemble Kalman filter (EnKF) is a Monte Carlo extension of the Kalman filter, capable of working on non-linear systems, which is applicable to the problem of HM in reservoir simulation. Various other extensions have been proposed for the petroleum industry. Nowadays, the standard HM method used by energy companies is the ensemble smoother with multiple data assimilation (ES-MDA) [1].

### B. *Proposal*

The objective is to develop a performance model that can achieve a faster overall ES-MDA runtime by dynamically tuning the simulations being executed. To improve the optimization process, we extract a feature vector from the output logfile of the simulation execution. This feature vector gathers important data about the underlying execution that is used to refit a performance model trained with more than 20,000 different reservoir simulations, leading to a better oracle.

### C. *Experimental Environment*

TunaOil was evaluated with three black-oil reservoir models – listed in Table I – that have multiple geological realizations. A realization is an uncertain representation of the rock-fluid properties, such as porosity, horizontal and vertical permeability, net-to-gross, and initial water saturation. Table I shows the number of realizations available, the mean elapsed time of the realizations in seconds (simulated with the default parameters or the engineer manually-tuned parameters) and the number of simulations with different configurations performed in the ES-MDA workflow for each reservoir model. The times reported are for the reservoir simulator using a 48-core (without Hyper-Threading) node.

The number of simulations represents the total number of executions required by the ES-MDA algorithm to perform the HM process, as it simulates each realization five times. Therefore, the total time to run the OLYMPUS case, considering the default numerical parameters of the reservoir simulation, was slightly more than 8 hours (119 seconds multiplied by 250 executions). Using the manual configuration selected by the reservoir engineer, the total time was over 28 hours. All these timings are cumulative, as in a computational cluster with multiple nodes, we can run some cases at the same time on different nodes, reducing the wall time (the perceived total time for the end-user). As the ES-MDA needs the results of all the realizations to apply the Kalman Gain calculation and prepare the next batch of realizations to be simulated, the maximum number of nodes to be used in parallel is defined by the number of realizations. Considering the same example of the OLYMPUS case, by using 50 nodes, we reduce the wall time to the end-user to roughly 10 minutes. The same applies to all the workloads listed in the table.

| Reservoir Model | Number of Realizations | Mean Default Time | Mean Engineer Time | Number of Simulations |
|---|---|---|---|---|
| OLYMPUS | 50 | 119 | 411 | 250 |
| UNISIM-II | 500 | 669 | 645 | 2500 |
| UNISIM-I Fine | 48 | 72.071 | 56.360 | 240 |

TABLE I. Workloads used in the evaluation (times in sec).

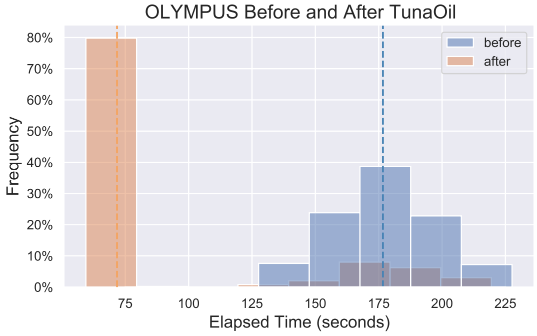


Fig. 1. Histogram of the OYMPUS simulations executed in an ES-MDA with and without TunaOil. The dashed lines represent the medium of the values.

**OLYMPUS** [2] is a synthetic reservoir model developed by TNO in 2017 for a benchmark study on field development optimization. The model has a grid of 341K cells. The **UNISIM-I** [3] is a synthetic model based on a real data sample from the Namorado Field in the Campos Basin, Brazil, while the **UNISIM-II** [4] is a synthetic model based on a carbonate offshore reservoir that represents the Brazilian pre-salt. The UNISIM-II has 190K cells, while the original UNISIM-I has 93K cells. However, the UNISIM-I directly used in our work was the fine geological model, which has more than 11M cells.

### *D. Results*

Our experiments show that TunaOil can improve the execution time of the base case by up to 40%, increasing the material balance error, on average, by less than 1% and the gas, oil, and water production error by less than 2%. Figure 1 shows the best result achieved by our methodology among the OLYMPUS workload – detailed in Section I-C – when compared to the engineer configuration. This figure shows the overall system performance impact on the ES-MDA workflow. The speed-up of the simulations was evaluated together with the impact on the quality of the outputs produced by the simulations.

### *E. Prior Unsuccessful Work*

The original proposal was to develop a general oracle that can present good numerical parameters for any unseen reservoir model. The first attempts to reach that goal were unsuccessful, leading to performance models that, in many cases, were even worse than the default parameters used by the reservoir simulator. When the suggestions led to good execution times, the results of the simulations were outside the acceptable engineer-error margin. It appears that "similar" models can perform differently due to factors that are difficult to characterize, that is, the degree of heterogeneity and non-linearities resulting from the characteristics of the problem, such as flow rates, mass transfer between phases, etc. The solution to overcome these issues was to include the oracle inside a workflow that simulates the "same" reservoir model multiple times, such as in an HM process. That way, we can use the first iterations to refit the performance model, providing the extra missing knowledge to characterize it.

### *F. Conclusion*

This work introduces the use of a performance model to dynamically tune the numeric parameters of petroleum reservoir models, reducing the overall application runtime without the need for additional simulations or a separate optimization study. Our experiments demonstrated that the oracle built was able to predict the proper effect of the changes in the solver options, in terms of simulation time and quality. The experiments have shown that our oracle makes accurate predictions in a broadly used workflow in the petroleum engineering area with black-oil models. The idea can be easily extended for other types of workflows, such as optimization processes or other types of models, such as compositional models. Ultimately, it would be feasible to couple the oracle developed in the central scheduler system of an energy company, such as Petrobras, to perform live optimization of any reservoir simulation being submitted to their HPC infrastructure, reducing time and associated costs.

## II. Acknowledgment

The author would like to thank PETROBRAS for funding this work, CMG for providing the simulator used in this research, and LNCC for providing their HPC infrastructure for the experiments. This work is under submission to the Computers and Geosciences Journal.

## References

[1] A. A. Emerick and A. C. Reynolds, "Ensemble smoother with multiple data assimilation," *Computers Geosciences*, vol. 55, pp. 3–15, jun 2013.

[2] R. Fonseca, E. Della Rossa, A. Emerick, R. Hanea, and J. Jansen, "Overview Of The Olympus Field Development Optimization Challenge," in *16th European Conference on the Mathematics of Oil Recovery (ECMOR XVI)*, Barcelona, Spain, sep 2018.

[3] G. D. Avansi and D. J. Schiozer, "UNISIM-I: Synthetic Model for Reservoir Development and Management Applications," *International Journal of Modeling and Simulation for the Petroleum Industry*, 2015.

[4] M. Correia, J. Hohendorff, A. T. F. S. Gaspar, and D. Schiozer, "UNISIM-II-D : Benchmark Case Proposal Based on a Carbonate Reservoir," in *SPE Latin American and Caribbean Petroleum Engineering Conference*, Quito, Ecuador, 2015.

**Felipe Portella** received his degree in Informatics in 2003 and an M.Sc in Computer Science in 2008 from PUC-Rio. He is an IT Consultant at the Brazilian energy company Petróleo Brasileiro S.A. (PETROBRAS), working with petroleum reservoir simulation workloads in HPC environments. He is currently a Ph.D. student at the Universitat Politècnica da Catalunya (BarcelonaTech-UPC) in partnership with the "Data-Centric Computing" group of the Barcelona Supercomputing Center (BCN-CNS).

# A Machine Learning based Wall Model for LES of Turbulent flows

Sarath Radhakrishnan*, Oriol Lehmkuhl*
*Barcelona Supercomputing Center, Barcelona, Spain
E-mail: {sarath.radhakrishnan, oriol.lehmkuhl}@bsc.es

***Keywords—Machine Learning, XGBoost, Wall Modeling, LES***

## I. Extended Abstract

The trubulent flow of fluids is still an enigma for mathematicians and engineers alike. The partial differential equation that reperesents the flow of fluids does not have an analytical solution for turbulent regimes. The closest approximation for a solution can be arrived at by using Direct Numerical Solution(DNS). But the application of DNS is quite restricted because of the broad range of scales of turbulent flow. In order to resolve the complete scales that represent the flow, the computational resourses required is beyond the current capacity for industrial flows. The second best option is to employ Large Eddy Simulation(LES). In LES, only large scales of motion, that are dependent on the boundary conditions are resolved. The smaller scales of motion which are universal are modelled. This reduces the computational demand for free flows where there are no boundaries. However, LES is very expensive when it comes to wall-bounded flows. In wall-bounded flows. Approximately 50% of the resources are used for resolving the layers close to the wall[1]. If we are able to use an appropriate model that represents the effects of the inner turbulent wall layers, that will save lots of computational resourses. Such models are know as Wall Models in LES. Wall models can be loosely classified into algebraic Equilibrim wall models and Non-Equilibrium wall models. Different types of wall models are described in [2], [1], [3], [4]. In this study Machine Learning(ML) is used to develop a wall model and compared with an algebraic equilibrium wall model(EQBWM)

## II. Background

Machine learning is a subset of Artificial Intelligence. The key idea in ML is to train the machine with a data, such that its ability to perform a preferred task improves with experience. The performance is measured using a performance metric. When the metric improves, the machine is said to have learnt the task. There are many different machine learning algorithms. Neural networks, Support Vector Machines, Clustering and Gradient Boosting Some of the well-known methods. In the current study we have chosen XGBoost[5] package. XGBoost is an 'eXtreme' Gradient Boosting[6] engine. It can be used for classification(categorical data) as well regression(continuous data) problems. It basically uses decision tree structures as additive functions to make the predictions. XGBoost has shown very good performance for a variety of problems. Since our intention is to generate a model that can be applied to a wide variety of turbulent flows, XGBoost seems a very good candidate. In addition to that XGBoost has many other features that boost the performance which makes it a great choice for training large datasets.

## III. Methodology

The data we have chosen to train the model is a turbulent channel flow with $Re_\tau = 1000$, where $Re_\tau = \frac{u_\tau}{\nu}\delta$. $u_\tau$ is a charecteristic velocity of wall bounded turbulent flows called frictional velocity and $\delta$ is the half-channel height. The number of observations is $128^3$. Only70% of the data is used training and the rest is used for validation. In addition to that, channel flow data $Re_\tau = 180$ is used for continuous testing of the model. This is done in order to avoid overfitting. The iteration where the test data stops to improve is chosen as the best iteration. The XGBoost uses many parameters to tune the model. The depth of the tree, i.e, how many branches a single tree can have was limited to three in our model. Although the algorithm lets to use regularization parameters L1 and L2, we have not used it. The number of trees were limited to 500. However, the tree fitting process is allowed to stop once the training does not make any improvement in 10 iterations. The performance of the model is measured using the metric Mean Squared Error(MSE) given by, $\frac{1}{N}\sum_{i=1}^{N}(y-\hat{y})$where N is the number of observations, $y$ is the dependent variable and $\hat{y}$ the corresponding prediction from the model. And the objective function which is minimised is $-\frac{1}{2}\sum_{i=1}^{N}(y-\hat{y})^2$. After many tests and trails, three features were selected for training. These are a Local Reynolds number, which is the Reynolds number based on local velocity and the correspoding height from the wall, distance from the wall scaled with the height of the first grid point and the velcoity scaled by the velocity at the first grid point.

## IV. Results

The model was tested on two different turbulence flows, viz., channel flow $Re_\tau = 2000$ and NASA Hump[7].

### A. *Channel Flow* $Re_\tau = 2000$

The model on a priori tests had shown very good results. This was the first a posteriori test in order to check if the model has learned what was general in a turbulent channel flow. That is, to make sure the model does not show any 'bias' to the training data set. The $Re_\tau$ for this case is twice the one which was trained. If the model is over-fitted to the training data, we will not get good results. The computational domain is $6pi\delta, 2\delta, 4\pi\delta$ in length, height and width respectively and the mesh used for the simulation is $128 \times 96 \times 96$. Fig. 1 shows

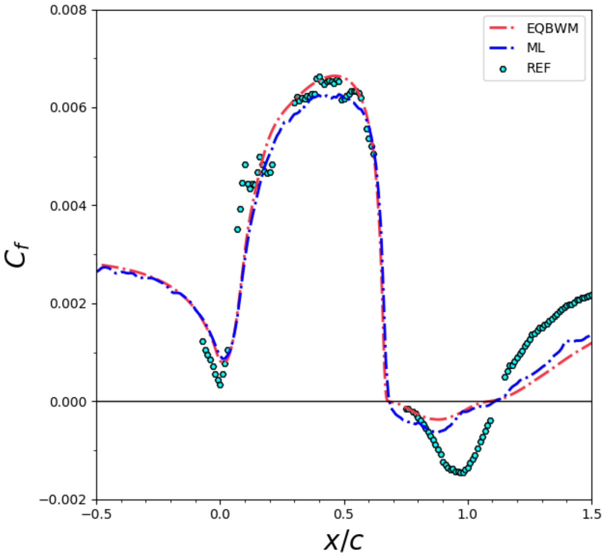


Fig. 2. **Ref** is the result from experiments. **ML** indicates the result from machine learning based model. **EQBWM** is the result from EQBWM.

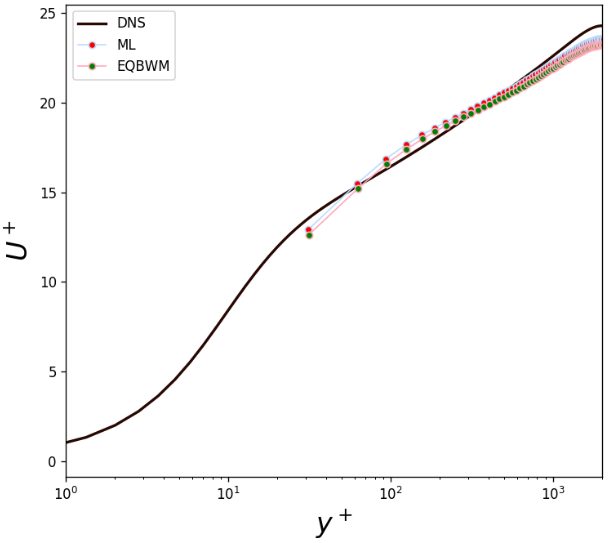


Fig. 1. $y^+$ vs $u^+$ for channel flow. **DNS** is the result from Direct Numerical Simulation. **ML** is the result from machine learning based model. **EQBWM** is the result from EQBWM.

the results for the channel flow, compared with a standard EQBWM. The model performs as good as the EQBWM. This not only proves that the model is not over-fitted, but also, it performs as good as a physics-based model. This model, in the process of the training has learned some physics of the flow. In the next test, we intend to check the extend of the physics the model has learned.

### *B. Flow over a hump*

This flow is very intersting because it has Non-Equilibrium effects(NEQBM) caused by the adverse pressure gradient acting around the aft end of the hump. Becuase of this adverse pressure gradient, there are seperation, re-attachment and recovery of the boundary layer. A model which is built on Equilibrium flows will not be able to understand the physics of such flows. This test is done in order to understand what the ML model has learnt so far and what more we have to teach the model. A mesh of approximately 8 million elements with $901 \times 111 \times 81$ divisions in stream-wise, wall-normal and span-wise directions respectively was used for the simulation. Fig. 2 shows the skin friction coefficient $C_f$. The model fairs as good as the EQBWM, but fails where the EQBWM fails. Just like the EQBWM, the model fails to capture the NEQBM featured of the flow. This means that the trained model is as good as an EQBWM and shows good variance.

## V. Conclusions

In this work, a novel machine learning based wall model is presented. The model is a posteriori tested in simple and complex flows and the performance is compared with an EQBWM. The model has no bias and high variance and it is as good as an algebraic wall model. The model fails to capture NEQBM effects as the training data lacked this feature. Further improvements may be possible if the model is given data with NEQBM effects. This is reserved for future works.

## References

[1] U. Piomelli, "Wall-layer models for large-eddy simulations," *Progress in Aerospace Sciences*, vol. 44, no. 6, pp. 437 – 446, 2008, large Eddy Simulation - Current Capabilities and Areas of Needed Research. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037604210800047X

[2] U. Piomelli and E. Balaras, "Wall-layer models for large-eddy simulations," *Annual review of fluid mechanics*, vol. 34, no. 1, pp. 349–374, 2002.

[3] J. LARSSON, S. KAWAI, J. BODART, and I. BERMEJO-MORENO, "Large eddy simulation with modeled wall-stress: recent progress and future directions," *Mechanical Engineering Reviews*, vol. 3, no. 1, pp. 15–00 418–15–00 418, 2016.

[4] S. T. Bose and G. I. Park, "Wall-modeled large-eddy simulation for complex turbulent flows," *Annual Review of Fluid Mechanics*, vol. 50, no. 1, pp. 535–561, 2018. [Online]. Available: https://doi.org/10.1146/annurev-fluid-122316-045241

[5] T. Chen and C. Guestrin, "Xgboost," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[7] G. I. Park, "Wall-modeled large-eddy simulation of a separated flow over the nasa wall-mounted hump by," 2015.

**Sarath Radhakrishnan** received his Master in Engineering degree from Ecole Central de Nantes in 2017. He started his PhD in the CASE depatment BSC in 2018 on the topic of Wall-Modeling in LES of turbulent flows.

# Unveiling the Transcriptional and Cellular Landscape of Age across Human Tissues

Aida Ripoll-Cladellas[#1], Monique G.P. van der Wijst[*2], Marta Melé[#3]

[#]*Barcelona Supercomputing Center (BSC), Spain*
[1]aida.ripoll@bsc.es, [3]marta.mele@bsc.es
[*]*University Medical Center Groningen, The Netherlands*
[2]m.g.p.van.der.wijst@umcg.nl

***Keywords*—Aging, Cell type deconvolution, Single-cell transcriptomics**

## EXTENDED ABSTRACT

As the aging population grows progressively around the globe, the need to research and develop strategies to healthy aging is ever more critical and takes on new urgency[1]. Primary hallmarks of aging include cell autonomous changes linked to epigenetic alterations, genomic instability, telomere attrition and loss of proteostasis (protein homeostasis), which are followed by antagonistic responses such as deregulated nutrient sensing, altered mitochondrial function and cellular senescence. In addition, many functions of the immune system show a progressive decline with age, referred as immunosenescence, leading to a higher risk of infection, cancer, and autoimmune diseases[2]. Although chronological age is the most powerful risk factor for most chronic diseases, the underlying molecular mechanisms that lead to generalized disease susceptibility are largely unknown[3].

In recent years, rapidly developing high-throughput omics have provided a broader insight, with the identification of a number of longevity-relevant loci based on genome-wide association studies (GWAS) and epigenome analyses. Despite this success, *APOE*, *FOXO3* and *5q33.3* are the only identified loci consistently associated with longevity[3]. Hence, the complexity of the aging phenomenon, influenced by genetic and epigenetic regulation, post-translational regulation, metabolic regulation, host–microbiome interactions, lifestyle, and many other elements, primarily explains the poor understanding of many of the molecular and cellular processes that underlie the progressive loss of healthy physiology[4].

Whether these hallmarks of aging occur across different tissues, and what are the aging changes driven by expression, splicing or cell type composition remains poorly understood. Since studies in model organisms have shown that aging is characterized by distinct alterations at the molecular, cellular and tissue level[5], a transcriptome analysis might lend greater insight than a static genetic investigation. However, since bulk samples of heterogeneous mixtures (i.e., tissues) only represent averaged expression levels, many relevant analyses are typically confounded by differences in cell type proportions[3]. Therefore, one of the major goals of this study is to disentangle the age-related gene expression changes to the cellular composition variation across tissues and individuals, as shown in Fig.1. Ultimately, this information can promote the development of personalized medicine, as well as understanding the biological mechanism of the aging process.

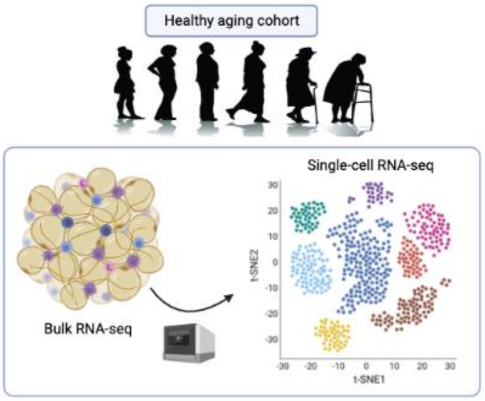


Fig. 1 Cartoon illustration showing an overview of the study workflow.

### A. *Age-related Gene Expression and Splicing Patterns Vary Among Human Tissues*

To understand how individual variation in gene expression and splicing can explain phenotypic differences (such as age, sex, ancestry or BMI) between individuals, we conducted a human transcriptome-wide analysis taking advantage of the publicly available 17,382 high-quality RNA-sequencing (RNA-seq) human samples from 838 postmortem donors across 49 tissue types of version 8 of the Genotype-Tissue Expression (GTEx) dataset[6], being the largest catalog to date of genetic regulatory variants affecting gene expression and splicing in *cis* and *trans* across tissues. Using gene-centric analysis, such as differential gene expression and splicing analysis (DEA and DSA, respectively) together with hierarchical partitioning, we discovered that age-related gene expression and splicing patterns notably vary in a tissue-wise fashion manner. Specifically, the largest gene expression changes with age were observed in arteries, while the major changes in splicing appeared in some brain regions.

### B. *Cell Type Abundance Across Tissues is Largely Associated with Age*

To assess how cell type composition could be confounding the observed gene expression variation with age across tissues, we performed a cell type enrichment analysis from GTEx gene expression data using *xCell*[7] to study the association between *xCell* enrichment scores and the different individual traits across tissues. Interestingly, we noticed a larger association between cell type abundance and age in different tissues compared to the other individual traits, which point toward relevant cellular composition changes during aging.

### C. *Single-Cell Transcriptomic Analysis Across Individuals*

To further elucidate cell-specific changes occurring across multiple cell types and organs, as well as age-related changes in the cellular composition of different organs, we will benefit from emerging single-cell RNA-sequencing (scRNA-seq) technologies. To this end, we will analyze the large number of single cell transcriptomic profiles from PBMCs (Peripherial Blood Mononuclear Cells) of many individuals provided by the single-cell eQTLGen (sc-eQTLGen) Consortium[8]. The accessibility and clinical relevance of PBMCs have made them the most studied cell types in current population-based scRNA-seq datasets. In the context of our analysis, it will help to shed light on the interplay between the age-related changes that affect different components of the immune system.

### D. *ACKNOWLEDGEMENTS*

We acknowledge the donors for their tissue donations for the GTEx research study. We thank the involvement of the large number of research groups participating in the current development of the sc-eQTLGen Consortium.

*References*

[1] Tabula Muris consortium, T. et al. A Single Cell "Transcriptomic Atlas Characterizes Aging Tissues in the Mouse." *Nature*. 2020.
[2] C.López-Otin. at al. "The hallmarks of aging". *Cell*. 2013.
[3] Peters, M. J. et al. The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* 2015.
[4] W.Zhang. et al. The ageing epigenome and its rejuvenation. *Nat. Rev. Mol. Cell. Biol*. 2020.
[5] C. J. Kenyon. et al. The genetics of ageing. *Nature*. 2010.
[6] F.Aguet et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020.
[7] D.Aran. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology*. 2017.
[8] M.G.P. Van Der Wijst. Single-cell eQTLGen Consortium: a personalized understanding of disease. *Genome Biology*. 2020.

## *Author biography*

**Aida Ripoll Cladellas** was born in Barcelona, Spain, in 1994. She received the BSc degree in Human Biology from the University of Pompeu Fabra, Barcelona, Spain, in 2017, and the MSc degree in Bioinformatics for Health Sciences from the University of Pompeu Fabra, Barcelona, Spain, in 2019.

Since September 2019, she has been with the Transcriptional and Functional Genomics Lab (TFGL) lead by Marta Melé at the Department of Life Sciences in the Barcelona Supercomputing Center (BSC), where she joined as a research engineer (RE1), and last September 2020 she started her PhD project on studying the human transcriptome changes with aging.

# perSVade: personalized Structural Variation detection in your species of interest

Miquel Àngel Schikora-Tamarit[#1], Toni Gabaldón[#]

[#]*Comparative genomics lab, Life Sciences Department, BSC*

[1]miquel.schikora@bsc.es

***Keywords:*** **genomics, structural variation, variant calling**

## EXTENDED ABSTRACT

### *Background*

Structural variants (SV) such as translocations, inversions, deletions, and other genomic rearrangements can contribute significantly to genetic and phenotypic variability across many species. The role of SV has been traditionally overlooked due to the technical limitations of SV detection and interpretation from short-read sequencing datasets. Most available algorithms yield low recall when tested on humans, but few studies have investigated the performance in non-human genomes [1]. Similarly, there are no specific indications about what parameters should be used for SV calling for most species. It is unclear whether the accuracy of each algorithm and running parameters validated on model species work equivalently for other species.

### *Results*

In order to fill this gap we have developed perSVade (personalized Structural Variation Detection), a pipeline that identifies and annotates SVs in a way that is optimized for any input sample. Starting from a set of paired-end whole-genome sequencing reads, perSVade uses simulations on the reference genome to choose the best SV calling parameters. The output includes the optimally-called SVs, a report of the accuracy and a friendly graphical interface that shows the SVs on a genome browser. In addition, perSVade allows the calling small variants and copy-number variation. In summary, this pipeline is useful to identify several types of genomic variation in short reads using a single bash command.

We validated that perSVade increases the SV calling accuracy on both simulated and real variants for five diverse eukaryotic organisms. Importantly, we find that there is no universal set of "optimal" parameters, which makes our method essential to yield accurate variant calls.

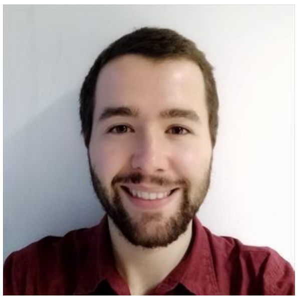
### *Conclusions*

We consider that this tool will help to understand how SVs generate phenotypes across non-human organisms.

### *References*

[1] D. L. Cameron, L. Di Stefano, and A. T. Papenfuss, "Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software.," *Nature communications*, vol. 10, no. 1, p. 3240, Jul. 2019.

## *Author biography*



**Miquel Àngel Schikora-Tamarit** was born in Lleida in 1995. In 2017 he obtains a Bachelor Degree in Human Biology by the Universitat Pompeu Fabra (UPF) of Barcelona. He works on several projects focused on understanding single-cell behavior under the supervision of Dr. Lucas Carey in the Department of Experimental and Health Sciences of the UPF between 2014 and 2018. His research, involving experimental molecular biology techniques and computational analysis, yields four first-author publications in the journals Integrative Biology, Transcription, Genome Research and Cell Reports. He pursues a Master Degree in Bioinformatics at the UPF between 2017 and 2019. He develops the thesis project in the lab of Dr. Toni Gabaldón of the Center for Genomic Regulation (CRG) understanding the evolution of Complex I from fungal genomes. He is currently conducting his PhD thesis in Biomedicine at the University of Barcelona (UB), working on the evolution of drug resistance in fungal pathogens under the supervision of Dr. Toni Gabaldón at the Barcelona Supercomputing Center.

# From Comorbidities to Gene Expression Fingerprints and Back

Beatriz Urda-García[1,2], Alfonso Valencia[1,3*]

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

[2*] *Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain,*[3*] *ICREA, Barcelona, Spain*

[1]beatriz.urda@bsc.es, [3]alfonso.valencia@bsc.es

***Keywords*— Comorbidity, gene expression, RNA-seq**

## EXTENDED ABSTRACT

Epidemiological evidence shows that some diseases tend to co-occur more than expected by chance and that patient-specific trends are observed. However, the molecular processes underlying these phenomena remain unclear.

Here we exploit the accumulating RNA-seq data on human diseases to calculate disease similarities at the transcriptomic level. We build a disease similarity network that significantly captures almost half of the medically known comorbidities, substantially outperforming previously published methods and providing biological explanations for such co-occurrences. Additionally, we group patients from a given disease with a similar expression profile into meta-patients and calculate their molecular similarities with the analyzed diseases, highlighting the need to study disease comorbidities within a personalized medicine scope. Finally, we provide a web application in which the networks and their underlying molecular mechanisms can be easily inspected.

### *A. Introduction*

Comorbidity, defined as the co-occurrence of two or more diseases in the same patient, is a complex medical problem that has become a key research area due to the associated increased Disability-Adjusted Life-Years (DALYs), complex clinical management and health care cost [1].

Accumulating evidence from epidemiological studies indicates that some diseases co-occur more than expected by chance and that patients suffering from the same disease present different risks of developing secondary conditions [2].

To tackle this problem, a better understanding of the molecular processes driving comorbidity relationships is essential. In line with this, several studies have analyzed disease similarities using molecular information (disease-associated genes in protein-protein interaction networks (PPINs) [3], microbiome, miRNA or microarrays [4]). Although these efforts were able to meaningfully capture interesting examples, they were unable to recapitulate what is known at the medical level in a considerable manner.

Here, we have reformulated the problem and we show, for the first time, that actually gene expression data – RNA-seq data – is able to reproduce medical interactions in a substantial and improved way. Additionally, we introduce the concept of meta-patients (molecularly similar patients from a given disease), that allows for the exploration of subgroup-specific patterns.

### *A. Methods*

First, we collected RNA-seq studies comprising 72 human diseases from the Gene Expression Omnibus. Then, we developed an RNA-seq pipeline destined to the parallel processing of a collection of RNA-seq studies for a given set of diseases. Afterwards, we performed Gene Set Enrichment Analyses to obtain the significantly altered gene sets and pathways for each disease.

Next, we defined a Disease Similarity Network (DSN) in which we connected diseases based on the similarities of their differential gene expression profiles. Specifically, for each disease pair, we computed the Spearman's correlation between the logFC values of the genes in the union of their significantly differentially expressed genes (sDEGs). We kept the interactions that were significant after correcting for multiple testing (FDR <= 0.05).

Since epidemiological networks only describe positive comorbidity relationships, we evaluated the overlap of the positive interactions in our DSN with the ones described by Hidalgo *et al.*[2] (based on medical records). To do so, we transformed our disease names into the International Code of Diseases, version 9 (ICD9 codes), computed the overlap of the networks and assessed its significance by shuffling the interactions while preserving the degree distribution. Next, we followed the same methodology to compare our overlap with the ones obtained with other disease-disease networks based on molecular information (microbiome, miRNAs and disease-associated genes in PPINs [3]).

Going into a deeper detail, we stratified diseases into subgroups of patients with similar expression profiles (meta-patients) by applying clustering algorithms to the normalized and batch effect corrected gene expression matrix. Both PAM (k-medoids) and Ward2 algorithms were applied independently. Next, we performed differential expression analyses and functional enrichment to the obtained meta-patients, and built a Stratified Similarity Network (SSN) by connecting meta-patients and diseases in the previously described manner.

### *B. Results and discussion*

First, we collected published studies analyzing human diseases with RNA-seq data. After quality filtering, 58% of the samples were kept, corresponding to 2.705 samples from 62 studies and comprising 45 diseases. We performed differential expression analyses to obtain sDEGs for each disease and functional enrichment analyses to better understand the transcriptomic alterations associated with them. We showed that the diseases' altered molecular processes match their known pathophysiology. We also discussed cases in which such processes can be involved in the existence of medically known comorbidities.

Next, we built a disease-disease similarity network (DSN) connecting diseases based on the similarity or dissimilarity of their gene expression profiles. The resulting network contains one single connected component and a higher percentage of positive than negative interactions (63.37% versus 36.63%). The DSN captures many known disease comorbidities, like the relationship between Chron's disease, ulcerative colitis and colorectal cancer; comorbidities between neoplasms, like lung and liver cancer; and multiple described relationships among mental and nervous system disorders, such as the one of schizophrenia with bipolar disorder, autism or Parkinson's and Huntington's diseases (HD). Interestingly, we also observe some negative correlations that reflect known inverse comorbidity patterns, defined as a lower than expected risk of disease co-occurrence. For instance, the decreased risk of developing different types of cancer (liver, lung, breast and chronic lymphocytic leukemia) in HD patients is corroborated by a negative correlation in our DSN. Moreover, since we have the gene expression fingerprint of all the diseases at different levels of granularity (genes and pathways), we can inspect the molecular mechanisms that may underlie the observed relationships. We should consider that the presence of shared molecular mechanisms does not always reflect a comorbid relationship. However, we have included detailed examples in which the dysregulation of key physiopathological pathways is shared between comorbid diseases and shows an opposite pattern for inverse comorbidities, revealing crucial aspects of such disease relationships.

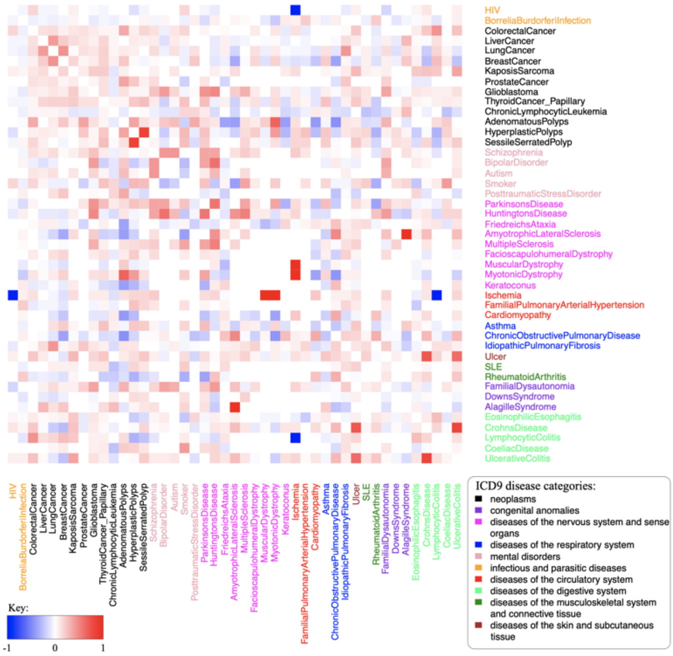


Fig. 1 Disease-disease Similarity Network (DSN). Pairwise disease correlations were computed based on the Spearman's correlation of the union of the sDEGs of each pair of diseases. A disease-disease network was built, containing the significantly positive and negative correlations (FDR <= 0.05), where the edge weights correspond to the Spearman's correlations. The heatmap shows the positive and negative disease interactions, in red and blue respectively. Diseases are coloured by ICD9 disease category.

Subsequently, we evaluated to what extent our DSN is able to capture medically known comorbidities. We found that our DSN significantly overlaps 46.53% (p-value = 0.001) of the interactions in Hidalgo et al. (based on medical records) and up to 60.48% (p-value = 0.0076) with a more stringent approach.

Next, we compared our overlap with the ones derived from previous disease-disease networks based on other molecular data. Both, the microbiome and the miRNA networks yielded non-significant overlaps with the epidemiological network. The network derived from PPINs [3] presented significant yet small overlaps with the epidemiology (8.71% for the entire network and 18.52% over the diseases in our DSN), and the one generated by Sánchez-Valle et al. using microarray presents a significant overlap of 25% [4]. This implies, for the first time, that molecular -transcriptomic- similarities can capture and meaningfully explain a sizeable percentage of medically known comorbidities.

Additionally, since patient-specific patterns are observed at the epidemiological level, we introduced the concept of meta-patients as groups of patients from a given disease with a similar expression profile. Then, we calculated the similarities between meta-patients and diseases, in an attempt to identify subgroup-specific similarities potentially reflecting comorbidity relations. Our results show that some known disease associations that are difficult to reproduce at the disease level become evident when considering disease subtypes. In fact, we observe that some diseases present meta-patients that vary greatly on their disease links. This highlights the importance of studying comorbidities within a personalized medicine scope.

A current limitation of this study is the lack of information about the patient's relevant features (e.g., sex or age). Importantly, we provide a web application in which the networks at the disease and meta-patient level, as well as the molecular mechanisms that may explain their relationships, can be easily inspected. Furthermore, the automatization of the presented analysis allows for the future integration of the fast-growing and publicly available RNA-seq studies.

## *C. ACKNOWLEDGEMENTS*

This work was supported by a Ph.D. Fellowship and funded by the Spanish Ministry of Economics and Competitiveness. We recognize Jon Sánchez Valle and Rosalba Lepore as authors of the present study.

## *References*

[1] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: Implications for understanding health and health services," *Ann. Fam. Med.*, 2009, doi: 10.1370/afm.983.

[2] C. A. Hidalgo, N. Blumm, A. L. Barabási, and N. A. Christakis, "A Dynamic Network Approach for the Study of Human Phenotypes," *PLoS Comput. Biol.*, 2009, doi: 10.1371/journal.pcbi.1000353.

[3] J. Menche *et al.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science (80-. ).*, 2015, doi: 10.1126/science.1257601.

[4] J. Sánchez-Valle *et al.*, "Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-16540-x.

## *Author biography*

**Beatriz Urda** was born in Almería, Spain, in 1993. She received the BSc in Biochemistry from the University of Granada in 2018 and the MSc in Bioinformatics for the Health Sciences from Pompeu Fabra University in 2020, Barcelona, Spain. She joined Alfonso Valencia's Computational Biology group as a master's student in 2019 and has recently started her PhD with a fellowship from the Spanish Ministry of Economics and Competitiveness.

# Pushing the Envelope on Free TLB Prefetching

Georgios Vavouliotis*†, Lluc Alvarez*†, Marc Casas*
*Barcelona Supercomputing Center, Barcelona, Spain †Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {georgios.vavouliotis, lluc.alvarez, marc.casas}@bsc.es

***Keywords*—TLB, prefetching, microarchitecture, caches.**

## I. Extended Abstract

Frequent Translation Lookaside Buffer (TLB) misses pose significant performance and energy overheads due to page walks required for fetching the translations. The address translation performance bottleneck is further exacerbated by the advent of big data and graph processing workloads due to their massive data footprints. Prefetching page table entries (PTEs) ahead of demand TLB accesses is an intuitively effective approach for alleviating the TLB performance bottleneck. However, each TLB prefetch request implies traversing the page table to fetch the corresponding PTE, triggering additional accesses to the memory hierarchy. Therefore, TLB prefetching is a promising, although costly, technique that may undermine performance when the prefetches are not accurate.

This work exploits the locality in the last level of the page table to reduce the cost and enhance the performance benefits of TLB prefetching by prefetching adjacent PTEs "for free". We design *Dynamic Free TLB Prefetching (DFTP)*, a scheme that predicts via sampling the usefulness of these "free" PTEs and prefetches only the ones most likely to save TLB misses. DFTP can be combined with any TLB prefetcher to provide further performance enhancements by exploiting page table locality for both demand and prefetch page walks.

### A. Dynamic Free TLB Prefetching (DFTP)

*1) Motivation:* Figure 1 depicts the operation of a x86-64 page walk and illustrates the locality of the PTEs in the last level of the page table. PTEs are stored contiguously in memory, and each PTE is 8B, so a single cache line can store 8 PTEs. When the requested PTE is read from memory at the end of a page walk, it is grouped with 7 neighboring PTEs and they are stored into a single 64B cache line. Hence, a cache line holds the requested PTE plus 7 more PTEs that do not require additional memory operations to be prefetched.

The naive approach is to prefetch all available free PTEs into a TLB Prefetch Buffer (PB)[1]. However, TLB prefetching is limited by the PB size, the PB area overhead, and the cost of PB lookups. Thus, naively storing all free prefetches per page walk into the PB may limit the performance benefits by evicting useful prefetches and polluting the PB with inaccurate prefetches. Hence, to exploit page table locality with a realistic PB size, a scheme that dynamically identifies and prefetches only the useful free prefetches per page walk is required.

*2) Design and Operation:* To address the findings of Section I-A1, we design *Dynamic Free TLB Prefetching (DFTP)*, a scheme that predicts via sampling the usefulness of the different free PTEs per page walk, and fetches in the PB only the most useful ones. We define *free distance* as the distance, within the cache line, between the PTE that holds the demand translation and another free PTE. Depending on the cache line position of the requested PTE, there are 14 possible free distances: from -7 to +7, excluding 0.

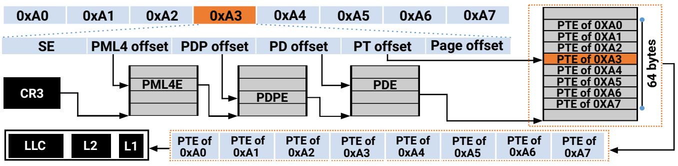


Fig. 1. Page table locality on x86-64 page table walks.

The DFTP scheme associates each free PTE with a free distance and exploits this information to predict the usefulness of the corresponding PTEs. Figure 2 presents the components and the functionality of DFTP: the *Sampling Queue (SQ)*, the *Free Distance Table (FDT)* and the *Prefetch Buffer (PB)*. The SQ is a small buffer that detects phases when free distances, which were previously useless, can provide useful prefetches. Each SQ entry stores the virtual page and its corresponding free distance for every free PTE that is decided not to be placed in the PB. The decision whether to place a free prefetch into the PB or the SQ is made by the FDT, a table with 14 counters; each counter monitors the hit ratio of one free distance. The PB is a buffer that stores the virtual page, the physical page and the corresponding free distance of the prefetches.

To explain the operation of DFTP, we consider the example presented in Figure 2 that assumes a page walk triggered by virtual page 0xF3. First, we identify the position of the requested PTE inside the cache line by extracting the 3 least significant bits of the page. Then we calculate the free distances of all PTEs residing in the same cache line and we associate each PTE with a free distance.

To determine whether a free prefetch has to be placed in the PB or the SQ, we compare the FDT saturating counter corresponding to its free distance with a threshold. If the counter exceeds the threshold, the free prefetch is fetched in the PB; otherwise, is placed in the SQ. The same procedure is followed for each free PTE in the cache line.

On PB or SQ hits, the FDT counter that corresponds to the free distance of the hit entry is increased. To prevent permanent saturation, we shift right one bit all the FDT counters when one of the counters saturates.

To summarize, DFTP adjusts the values of FDT counters depending on which free distances are frequently producing PB or SQ hits, thus DFTP is able to adapt to phase-behavior and predict the most useful free PTEs per page walk.

*3) Combining DFTP with TLB prefetching schemes:* Apart from fetching the most useful free prefetches per demand page walk, i.e., a page walk due to a demand TLB miss, DFTP is also able to operate on prefetch page walks, i.e., page walks triggered by TLB prefetch requests. Specifically, at the end of a prefetch page walk the prefetched PTE is grouped with 7 PTEs that can be prefetched for free due to page table locality.

[1] TLB prefetchers typically use a prefetch buffer to store the prefetches since prefetching directly into the TLB can negatively affect performance [1], [2].

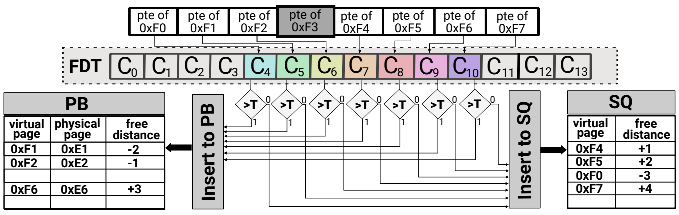


Fig. 2. Dynamic Free TLB Prefetching (DFTP) module.

At this point, DFTP is activated to decide which of the free prefetches should be placed in the PB or the SQ, essentially applying lookahead prefetching with depth 2.

*4) Methodology:* We consider a big set of industrial workloads provided by Qualcomm (QMM) for CVP1 [3], the SPEC CPU 2006 [4] and SPEC CPU 2017 [5] suites, and big data workloads included in the GAP [6] suite and the XSBench [7]. We refer to GAP and XSBench workloads as Big Data (BD) workloads. Our evaluation takes into account the workloads with a TLB MPKI of at least 1. All traces have been obtained using the SimPoint [8] methodology. For the QMM workloads we use 50M warmup instructions and 100M instructions for measuring the results. The rest of the workloads run 250M warmup instructions and 1B instructions are executed to measure the experimental results.

For evaluation we use ChampSim [9], a detailed simulator that models a 4-wide out-of-order processor. We extend ChampSim with a realistic x86 page table walker, modeling (i) the variant latency cost of page walks, (ii) the page walk references to memory hierarchy, and (iii) the cache locality in page walks.The page table walker supports up to 4 concurrent TLB misses [10], while one page walk can be initiated per cycle. Table I summarizes our experimental setup.

*TLB Prefetchers.* We consider the state-of-the-art TLB prefetchers: (i) Sequential Prefetcher (SP); SP [2] prefetches the PTE located next to the one that triggered the TLB miss, (ii) Arbitrary Stride Prefetcher (ASP); ASP [2] is a table-based prefetcher that captures miss streams with varying strides, and (iii) Distance Prefetcher (DP); DP [2] is a table-based prefetcher that correlates miss patterns with distances between pages that produce consecutive TLB misses. Our evaluation considers the most common scenario where a Prefetch Buffer (PB) is used to store the prefetched PTEs (Section I-A1).

*5) Evaluation:* To highlight the benefits of DFTP we compare it against the following scenarios: (i) free prefetches are not exploited (NoFP), *i.e.*, they are not stored in the PB; (ii) all free prefetches are naively placed in the PB (NaiveFP).

The performance impact of the above explained scenarios for the state-of-the-art TLB prefetchers is presented in Figure 3. We observe that all prefetchers achieve high performance gains for all scenarios considering free prefetching (NaiveFP, DFTP) than when free prefetching is not exploited (NoFP). We observe this behavior because (i) the free prefetches provide PB hits that reduce demand page walks, and (ii) most of the prefetch requests have already been prefetched for free, avoiding prefetch page walks. For instance, SP+DFTP outperforms SP+NoFP by 5.6% for the SPEC workloads.

Furthermore, we observe that DFTP significantly improves performance over NaiveFP for the QMM and SPEC workloads, across all prefetchers. For the BD workloads we observe that DFTP and NaiveFP provide similar performance benefits because these workloads exhibit highly irregular patterns, thus it is difficult to detect the most useful free PTEs per page walk. Finally, we expect that designing a smarter TLB prefetcher would highlight more the benefits of DFTP over NaiveFP; we leave this exploration as future work.

| **Component** | **Description** |
|---|---|
| **L1 DTLB** | 64-entry, 4-way, 1-cycle, 4-entry MSHR |
| **L2 TLB** | 1536-entry, 12-way, 8-cycle, 4-entry MSHR, 1 page walk / cycle |
| **Prefetch Buffer (PB)** | 64-entry, fully assoc, 2-cycle |
| **Sampling Queue (SQ)** | 64-entry, fully assoc, 2-cycle |
| **L1 DCache** | 32KB, 8-way, 4-cycle, 8-entry MSHR, next line prefetcher |
| **L2 Cache** | 256KB, 8-way, 8-cycle, 16-entry MSHR, ip stride prefetcher |
| **LLC** | 2MB, 16-way, 20-cycle, 32-entry MSHR |
| **DRAM** | 4GB, DDR4, 4GHz, 1600 MT/s |

TABLE I. SYSTEM SIMULATION PARAMETERS.

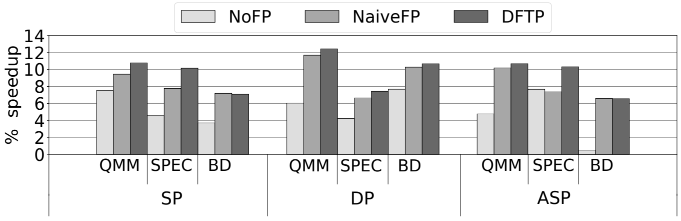


Fig. 3. Performance impact of free TLB prefetching scenarios.

## B. Conclusions

This work reveals the importance of exploiting page table locality for TLB prefetching purposes. We propose DFTP, a dynamic scheme that identifies the most useful free PTEs per page walk, and we show that DFTP can be combined with any TLB prefetcher to provide great performance enhancements.

## REFERENCES

[1] A. Bhattacharjee and M. Martonosi, "Inter-core Cooperative TLB for Chip Multiprocessors," in *Proceedings of the 15th Edition of ASPLOS on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS XV. NY, USA: ACM, 2010, pp. 359–370.

[2] G. B. Kandiraju and A. Sivasubramaniam, "Going the Distance for TLB Prefetching: An Application-driven Study," in *29th Annual International Symposium on Computer Architecture*, ser. ISCA '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 195–206.

[3] "CVP-1," https://www.microarch.org/cvp1/.

[4] "SPEC CPU 2006," https://www.spec.org/cpu2006/, [Online].

[5] "SPEC CPU 2017," https://www.spec.org/cpu2017/, [Online].

[6] S. Beamer et al., "The GAP benchmark suite," *CoRR*, vol. abs/1508.03619, 2015.

[7] "XSBench," https://github.com/ANL-CESAR/XSBench.

[8] E. Perelman et al., "Using simpoint for accurate and efficient simulation," *SIGMETRICS Perform. Eval. Rev.*, vol. 31, no. 1, pp. 318–319, Jun. 2003.

[9] "ChampSim," https://crc2.ece.tamu.edu/, [Online].

[10] Abishek Bhattacharjee, "Advanced concepts on address translation," http://www.cs.yale.edu/homes/abhishek/abhishek-appendix-l.pdf.

**Georgios Vavouliotis** received his Diploma on Electrical and Computer Engineering from National Technical University of Athens (NTUA), Athens in 2018. Since fall 2018, he has been a Ph.D. candidate at the Computer Architecture department of Universitat Politècnica de Catalunya (UPC), Spain, and he has been working on the Runtime Aware Architecture research group of Barcelona Supercomputing Center (BSC).

# Tsunami inundation forecast in central Chile using stochastic earthquake scenarios

Natalia Zamora[*][†], Alejandra Gubler[†][‡], Patricio A. Catalán[†][‡], Matías Carvajal[§][¶],
[*]Barcelona Supercomputing Center, Barcelona, Spain
[†]Research Center for Integrated Disaster Risk Management (CIGIDEN), ANID/FONDAP/15110017, Santiago, 7820436 Chile
[‡]Departamento de Obras Civiles, Universidad Técnica Federico Santa María, Valparaíso, 2390123 Chile
[§]Instituto de Geografía, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
E-mail: natalia.zamora@bsc.es

***Abstract*—As demonstrated by recent mega-tsunamis, tsunami coastal inundation could vary substantially as a result of the source characteristics and the local geomorphologic-related- effects. Here, numerical simulations are used to characterize the tsunami potential triggered by seismic sources. Particularly, available data suggesting high stressed areas along the shallow part of the interplate in central Chile are used to assess potential tsunami inundation along the highly populated coastal area of Viña del Mar and Valparaíso in Chile. The approach lies on the assessment of 1000 inundation scenarios along the region for earthquakes with $M_w$ 8.6 - $M_w$ 8.7. It is found that flow depths of 10 m can affect the region. This is crucial information for urban planning.**

***Keywords*—Tsunami potential, numerical simulations, central Chile.**

## I. Introduction

Tsunami inundation varies significantly along the affected coasts due to the characteristic of the source that triggers the tsunami or as consequence of local effects, causing enormous consequences in coastal communities. The aim of this study is to estimate the potential tsunami inundation in the coastal city of Viña del Mar and Valparaíso, the main coastal resort city and port of the country (Fig. 1). For this, earthquakes with magnitude between of $M_w$8.6 -$M_w$8.7 are considered as plausible events that could rupture in the (highly) coupled shallow part of the megathrust in central Chile[1].

## II. Seismic sources and numerical modeling

The seismic source is characterized leading to the generation of multiple scenarios considering stochastic distribution of slip. For each scenario, the initial sea-surface displacement is computed from regularly used elastic dislocation models, which are treated as initial conditions for tsunami numerical propagation and inundation using high resolution topographic and bathymetric computational grids. From these numerical simulations, the tsunami flow depths are obtained.

In this case study, only tsunamis triggered by seismic sources along the interplate are considered. First, an area of interest is identified where earthquakes of given magnitudes are expected. Stochastic rupture scenarios will be restricted over that area. The geometry of this source has been defined based on the subduction zone model of [4], extending from latitudes 31°S in the north to about 35°S in the south, which is roughly the region flanked by the main rupture zones of the 2010 $M_w$8.8 Maule and 2015 $M_w$ 8.3 Illapel earthquakes. In the dip direction, the fault region extends from the trench to about 60 km depth. These updip and downdip limits are consistent with the along-dip extent of those recent neighbouring events.

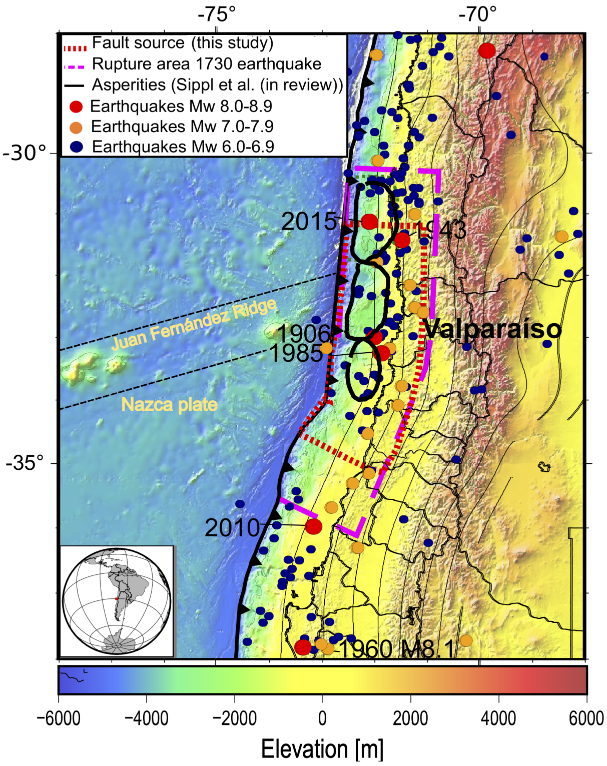


Fig. 1. Tectonic setting and the seismic source zone where stochastic slip distributions were generated. Black segments show three highly locked fault areas inferred in [1]. Two larger segments are shown: the extent of the rupture $M_w$8.6 - $M_w$8.7 (red dotted line) use in this study and the large Valparaíso maximum rupture estimated by [2] (magenta dashed line). Color-coded dots show epicenters of earthquakes greater than $M_w > 6.0$ in the region [3].

Along this source zone, a set of 1000 rupture scenarios with varying slip in both dimensions of the fault is created. All scenarios have the same target magnitudes of $M_w$8.6-$M_w$8.7, which was defined by applying earthquake scaling laws [5] to the 400-km-rupture length and 180-km-rupture width of the assumed seismic source. To generate the set of scenarios, the fault region is discretized into 192 rectangular sub-faults with

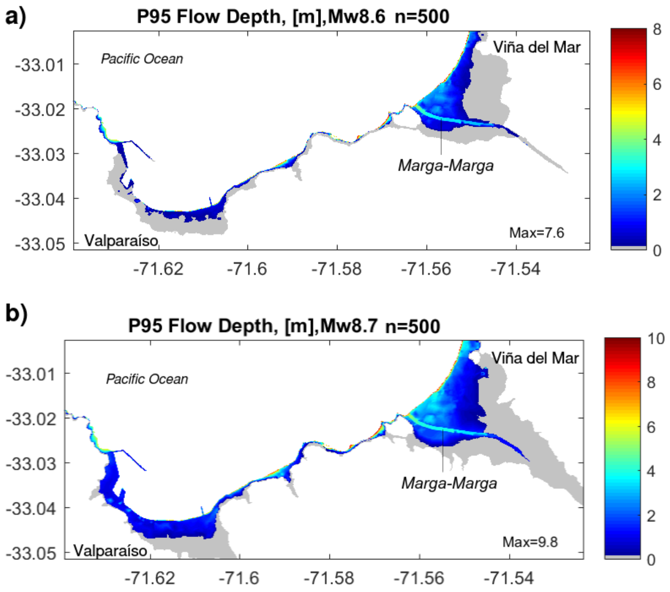


Fig. 2. Flow depths expected for Viña del Mar and Valparaíso based on percentile 95. a) Flow depth $d_m(x,y)$ for earthquakes $M_w$8.6; b) Flow depth $d_m(x,y)$ for earthquakes $M_w$8.7.

dimensions of 20 km both along strike and downdip. Next, the Karhoenen-Loeve expansion is used to generate aleatory slip distributions [6].

For each rupture seismic scenario, seafloor and land deformations were computed with the analytical solutions of a rectangular source of the Okada model [7], assuming instantaneous displacement. Once these deformations are obtained, numerical simulation are conducted for each tsunami source to obtain flow depths and arrival times using the software Tsunami-HySEA [8]. Tsunami-HySEA solves the two-dimensional shallow-water water equations (NLSWE) using a high-order path-conservative finite volume method, using high resolution bathymetry and topographic computational grids [9]. The target cities of Valparaíso and Viña del Mar are contained in the finest grid.

## III. Results

### A. Source characterization and tsunami modeling

The maximum coseismic slip considered was 20 m. This value is consistent with the offshore slip deficit accumulated since the last large earthquake in 1730 [2], considering a convergence rate of 6.5 cm/yr. From each coseismic slip distributed scenario numerical simulation were run and integrated to show flow depth along the coast of Viña del Mar.

Fig. 2 shows the results of tsunami inundation in terms of the flow depth $d_m(x,y)$. As with the slip, at each cell the $V_i(d_m,t_a)$ are used to build the cumulative density functions for each variable independently. From these, the percentile 95% are estimated, which integrates all scenario for each magnitude. In Viña del Mar (Fig. 2), inundation reaches about $d_m \approx$9 m over a very narrow band near the shoreline (red colors), which rapidly decrease to $d_m \approx$3-4 m in the surroundings of the Marga-Marga river floodplain. The largest tsunamis among the set can propagate up to 1.5 km inland, although with relatively small flow depths of $d_m \approx$0.5 m (blue colors).

## IV. Final Remarks

The present study highlights the tsunami potential along Viña del Mar and Valparaíso. A statistical spatial analysis shows the variations of flow depths triggered by seismic scenarios with $M_w$8.6-8.7. The main findings show that based on those earthquakes, flow depths up to 10 m could affect Viña del Mar; and Valparaíso will be less affected with flow depth up to 3 m. Inundation maps and time arrivals will be used to conduct tsunami vulnerability assessment in these coastal cities.

## References

[1] C. Sippl, M. Moreno, and R. Benavente, "Microseismicity appears to outline highly coupled regions on the central chile megathrust," *Preprint at EarthArXiv*, 2020.

[2] M. Carvajal, M. Cisternas, and P. A. Catalán, "Source of the 1730 Chilean earthquake from historical records: Implications for the future tsunami hazard on the coast of Metropolitan Chile," *Journal of Geophysical Research: Solid Earth*, vol. 122, no. 5, pp. 3648–3660, 2017.

[3] USGS, *USGS ComCat catalog*, United States Geological Survey, 2017, https://earthquake.usgs.gov/earthquakes/search/, last accessed 2017-01-20.

[4] G. P. Hayes, D. J. Wald, and R. L. Johnson, "Slab1. 0: A three-dimensional model of global subduction zone geometries," *J. Geophys. Res. Solid Earth*, vol. 117, no. B1, 2012.

[5] L. Blaser, F. Kruger, M. Ohrnberger, and F. Scherbaum, "Scaling relations of earthquake source parameter estimates with special focus on subduction environment," *Bull. Seismol. Soc. Am.*, vol. 100, no. 6, pp. 2914–2926, 2010.

[6] D. Melgar, R. J. LeVeque, D. S. Dreger, and R. M. Allen, "Kinematic rupture scenarios and synthetic displacement data: An example application to the cascadia subduction zone," *Journal of Geophysical Research: Solid Earth*, vol. 121, no. 9, pp. 6658–6674, 2016.

[7] Y. Okada, "Surface deformation due to shear and tensile faults in a half-space," *Bulletin of Seismological Society of America*, vol. 75, no. 4, pp. 1135–1154, 1985.

[8] J. Macías, M. J. Castro, S. Ortega, C. Escalante, and J. M. González-Vida, "Performance Benchmarking of Tsunami-HySEA Model for NTHMP's Inundation Mapping Activities," *Pure and Applied Geophysics*, vol. 174, no. 8, pp. 3147–3183, 2017.

[9] N. Zamora, P. A. Catalán, A. Gubler, and M. Carvajal, "Microzoning Tsunami Hazard by Combining Flow Depths and Arrival Times," *Frontiers Earth Science*, vol. 8, p. 591514, 2021.

**Natalia Zamora** received her PhD degree in Geosciences with focus in Geohazards from the University of Potsdam and GFZ-Potsdam, Germany in 2016. She conducted a post-doctoral research at CIGIDEN in Chile and was the main researcher of the FONDECYT project to assess the tsunami hazard along Chile. She joined the Natural and Social Hazards group in the Barcelona Supercomputing Center since September 2020 within the STARS program, funded by the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-75443.

# Poster Abstracts

# Sensitivity of soluble iron deposition to soil mineralogy uncertainty

Elisa Bergas-Massó*†, María Gonçalves-Ageitos*†, Carlos Pérez García-Pando*§

*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
§ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Spain.
E-mail: elisa.bergas@bsc.es

***Keywords—Iron cycle, Mineralogy, Climate, Dust.***

## I. Extended Abstract

Mineral dust emitted from arid and semi-arid areas has several effects on the Earth system (e.g., perturbation of the radiative budget, interaction with cloud processes, implications on ocean and land biogeochemical cycles). Mineral dust aerosols are mixtures of different minerals whose relative abundances, particle size distribution, shape, surface topography, and mixing state influence their interaction with the Earth system. However, Earth System Models (ESMs) typically assume that dust aerosols have a globally uniform composition, neglecting the known variations in the sources' mineralogical composition. This work investigates the sensitivity of a key biogeochemical cycle, the iron (Fe) cycle to uncertainties in the description of soil mineralogy in dust-producing areas.

Airborne mineral dust is the primary input of Fe to the open ocean. Fe constitutes a fundamental micro-nutrient for marine biota in its soluble form. It is, in fact, the limiting nutrient in remote regions of the open ocean known as High Nutrient Low-Chlorophyll (HNLC) regions (e.g., the Southern Ocean), where the Fe supply occurs mainly through atmospheric deposition. Ocean productivity relies on the availability of limiting nutrients. Hence, the ocean's ability to capture atmospheric $CO_2$ in HNLC regions highly depends on the atmospheric deposition of soluble Fe.

Fe abundance in soils is usually set to 3.5% [1], and its solubility is considered to be less than 0.1% [2]. However, both observations and modeling studies suggest that the solubility of Fe from dust increases downwind of the sources [3]. A primary mechanism leading to this increase in Fe solubility is acidic (proton-promoted) dissolution. Low pH conditions in aerosol water favor Fe dissolution by weakening Fe-O bonds of Fe oxides in dust [4]. Other physical and chemical mechanisms that enhance Fe solubilization involve photochemical reduction and organic ligand (e.g., Oxalate) processing [5].

Modeling the global dust mineralogical composition presents critical challenges. First, soil mineralogy atlases for dust modeling are derived by extrapolating a sparse set of mineralogical analyses of soil samples that are particularly scarce in dust source regions. Moreover, atlases are based on measurements following the wet sieving technique that tampers the undisturbed parent soil size distribution by breaking coarse particles and replacing them with smaller ones [6].

In this work, we assess the implications of soil mineralogy uncertainties on bio-available Fe delivery to the open ocean by using a state-of-the-art ESM, EC-Earthv3, where a detailed atmospheric Fe cycle and two different data sets that characterize the soil composition over dusty areas have been implemented [7] [8].

### A. Model Description and experimental setup

We performed simulations with the EC-Earthv3 ESM. EC-Earthv3 is collaboratively developed by European research centers from 10 different countries, including the Barcelona Supercomputing Center (BSC) [9]. The model configuration used here includes the Integrated Forecast System (IFS) model to represent atmospheric dynamics coupled with the Tracer Model 5 (TM5), which allows for interactive simulation of atmospheric chemistry and transport of aerosols and reactive gas species [10]. Our experiments are nudged towards the ERA-Interim reanalysis [11].

In the version of TM5 used in this work, the model further considers:

1) The primary emissions of both insoluble and soluble Fe forms, associated with mineral dust [12] [13] and combustion aerosols [14].
2) The atmospheric processing mechanism of Fe treated as a kinetic process accounting for: proton-promoted dissolution, oxalate-promoted Fe dissolution (with oxalate calculated on-line) and photo-reductive dissolution.
3) The representation of dust mineralogical composition with the introduction of two different soil mineralogy datasets (Claquin [7] and Journet [8]).
4) The fractional emission of minerals in the accumulation and coarse modes of the model follows brittle fragmentation theory [15] [16].

The Claquin et al. (1999) [7] dataset provides mineralogical information for arid dust-source regions based on 239 descriptions of soils. 8 different minerals are considered: illite, kaolinite, and smectite for the clay fraction, feldspars, hematite and gypsium for the silt fraction, and quartz and calcite in both mineral size fractions. The Journet et al. (2014) [8] dataset is based on data from 700 soil descriptions from more than 150 publications and contains information on the relative abundance of 12 minerals: quartz, feldspars, illite, smectite,

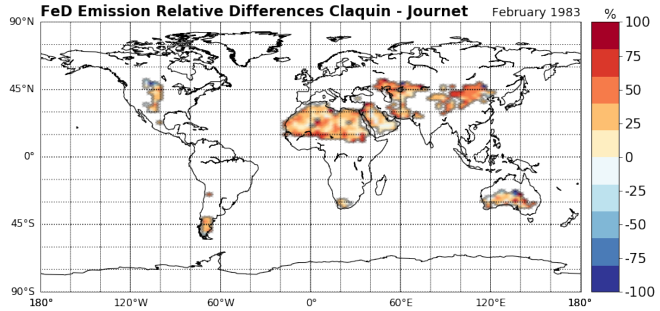


Fig. 1. Relative differences in Fe-dust emission between Claquin and Journet simulations [%] (in red Fe-dust emission is higher for Claquin and vice-versa).

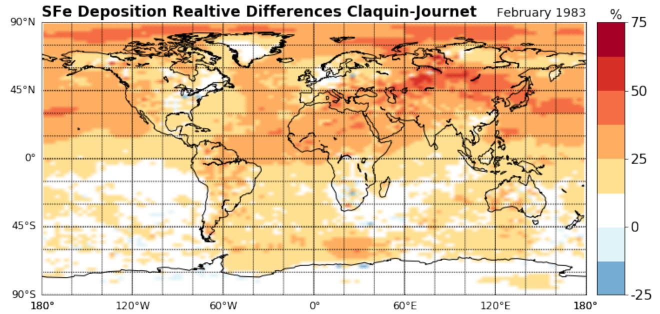


Fig. 2. Relative differences in soluble Fe deposition between Claquin and Journet simulations [%] (in red deposition is higher for Claquin).

kaolinite, chlorite, vermiculite, mica, calcite, gypsum, hematite and goethite. This set adds information in the clay fraction for some minerals considered on both data sets (e.g., feldspars and hematite). In contrast with Claquin, this data set has global coverage.

We run two equivalent 1-year-long simulations with the two soil mineralogy datasets and analyze the differences in the deposited soluble Fe.

### B. *Preliminary Results*

After analyzing the first-month results from both simulations, some clear patterns emerge. Overall, Fe-dust emissions are higher with Claquin's mineralogy, especially across the Sahel and North-East Asia (see Figure 1). This is also reflected on the soluble Fe deposition, which is higher with Claquin's mineralogy. The relative differences are more important in the Northern Hemisphere (see Figure 2). As expected, our results clearly show that soluble Fe deposition scales with the amount of emitted Fe. However we also see a slightly higher soluble to total Fe in deposition with Journet (1.13%) than with Claquin (1.04%). Further analysis regarding other mineralogical factors will be examined in this work. For instance, we aim at quantifying how atmospheric acidity is affected by mineralogy and hence Fe solubilization.

## II. Acknowledgment

We thank our collaborator Stelios Myriokefalitakis for his contribution and support. We thankfully acknowledge the RES AECT-2020-3-0020 and PRACE projects to grant this work with computer resources on Marenostrum4, and the technical support provided by the BSC and the Computational Earth Sciences team of the Earth Sciences Department. This work was funded by the Ministerio de Economía y Competitividad (MINECO) as part of the NUTRIENT project (CGL2017-88911-R). The research leading to this work has also received support from the ERC Consolidator Grant "FRAGMENT" (grant agreement No. 773051), and the AXA Chair on Sand and Dust Storms. Finally, also acknowledge financial support for this work by the National Observatory of Athens research grant (n$^{o}$ 5065).

## References

[1] R. A. Duce and N. W. Tindale, "Atmospheric transport of iron and its deposition in the ocean," *Limnology and Oceanography*, vol. 36, no. 8, pp. 1715–1726, 1991.

[2] I. Y. Fung *et al.*, "Iron supply and demand in the upper ocean," *Glob. Biogeochem. Cycles*, vol. 14, pp. 281–296, 2000.

[3] G. Zhuang *et al.*, "Link between iron and sulfur cycles suggested by fe(ii) in remote marine aerosol," in *Nature*, vol. 355, 02 1992, pp. 537–539.

[4] M. S. Johnson and N. Meskhidze, "Atmospheric dissolved iron deposition to the global oceans: effects of oxalate-promoted Fe dissolution, photochemical redox cycling, and dust mineralogy," *Geoscientific Model Development*, vol. 6, no. 4, pp. 1137–1155, 2013.

[5] S. O. Pehkonen *et al.*, "Photoreduction of iron oxyhydroxides in the presence of important atmospheric organic compounds," *Environmental Science & Technology*, vol. 27, no. 10, pp. 2056–2062, 1993.

[6] B. Chatenet *et al.*, "Assessing the microped size distributions of desert soils erodible by wind," *Sedimentology*, vol. 43, no. 5, pp. 901–911, 1996.

[7] T. Claquin *et al.*, "Modeling the mineralogy of atmospheric dust sources," in *Journal of Geophysical Research*, vol. 104256, 09 1999, pp. 243–22.

[8] E. Journet *et al.*, "A new data set of soil mineralogy for dust-cycle modeling," in *Atmospheric Chemistry and Physics*, vol. 14, 04 2014, pp. 3801–3816.

[9] W. Hazeleger *et al.*, "Ec-earth v2.2: Description and validation of a new seamless earth system prediction model," in *Climate Dynamics*, vol. 39, 12 2011, pp. 1–19.

[10] T. P. C. van Noije *et al.*, "Simulation of tropospheric chemistry and aerosols with the climate model EC-Earth," *Geoscientific Model Development*, vol. 7, no. 5, pp. 2435–2475, 2014.

[11] P. Berrisford *et al.*, "The ERA-Interim archive Version 2.0," no. 1, p. 23, 2011.

[12] S. Myriokefalitakis *et al.*, "Changes in dissolved iron deposition to the oceans driven by human activity: a 3-D global modelling study," *Biogeosciences*, vol. 12, no. 13, pp. 3973–3992, 2015.

[13] ——, "The gesamp atmospheric iron deposition model intercomparison study," *Biogeosciences Discussions*, pp. 1–50, 07 2018.

[14] A. Ito *et al.*, "Radiative forcing by light-absorbing aerosols of pyrogenetic iron oxides," in *Scientific Reports*, vol. 8, no. 1, may 2018.

[15] J. Kok, "A scaling theory for the size distribution of emitted dust aerosols suggests climate models underestimate the size of the global dust cycle," in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, 01 2011, pp. 1016–21.

[16] C. P. García-Pando *et al.*, "Predicting the mineral composition of dust aerosols: Insights from elemental composition measured at the izaña observatory," in *Geophys. Res. Lett.*, 2016.

**Elisa Bergas-Massó** received her BSc degree in Physics from Universitat de Barcelona (UB) in 2018. She completed her MSc degree in Meteorology in the same university in 2019. While doing the MSc degree, she did an internship in the Atmospheric Composition group of the Earth Sciences department of the Barcelona Supercomputing Center (BSC) working on mineral dust emission. Since September 2019, in the scope of a Ph.D., she has focused her work on the iron cycle and its implementation in climate models.

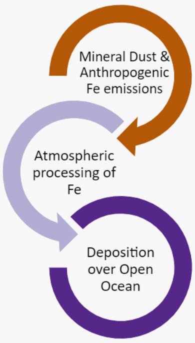
# SENSITIVITY OF SOLUBLE IRON DEPOSITION TO SOIL MINERALOGY UNCERTAINTY

Elisa Bergas-Massó(1), María Gonçalves Ageitos(1,2), Stelios Myriokefalitakis(3), Twan van Noije(4), Ron Miller(5), and Carlos Pérez García-Pando(1,6)

(1)Barcelona Supercomputing Center, (2)Universitat Politècnica de Catalunya, (3)National Observatory of Athens, Institute for Environmental Research and Sustainable Development (IERSD), (4)Royal Netherlands Meteorological Institute (KNMI), (5)NASA Goddard Institute for Space Studies, (6) Catalan Institution for Research and Advanced Studies (ICREA)

CONTACT: elisa.bergas@bsc.es

## BACKGROUND & AIM

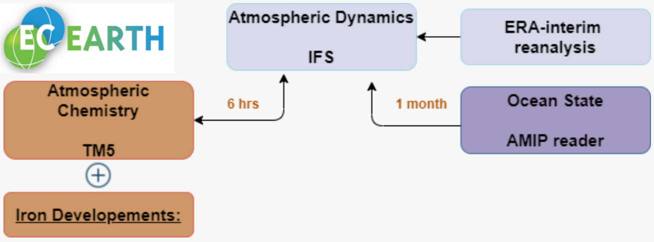

Mineral Dust & Anthropogenic Fe emissions
Atmospheric processing of Fe
Deposition over Open Ocean

Ocean productivity relies upon bioavailable iron (Fe) as nutrient, which makes the **Fe biogeochemical cycle a key modulator of the ocean's ability to uptake atmospheric CO2.**

The **main external input of Fe to the open ocean surface is atmospheric deposition**, which derive mainly from soil dust aerosol transported from arid and semi-arid regions (~95%). Fe in freshly emitted soil dust is mostly insoluble, but it is hypothesized to be partly transformed into bioavailable Fe species during atmospheric transport through a variety of dissolution mechanisms.

In this work, **we assess the implications of soil mineralogy uncertainties on bio-available Fe delivery** to the open ocean by using a state-of-the-art ESM, EC-Earthv3, where a detailed atmospheric Fe cycle and two different data sets that characterize the soil composition over dusty areas have been implemented

## MODEL DESCRIPTION & EXPERIMENTAL SETUP



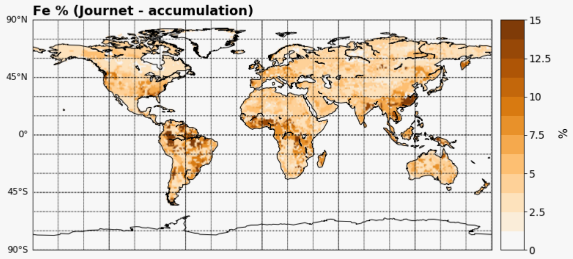
1 Fe Primary emissions associated with mineral dust and combustion aerosols.

2 Atmospheric processing mechanism of Fe :

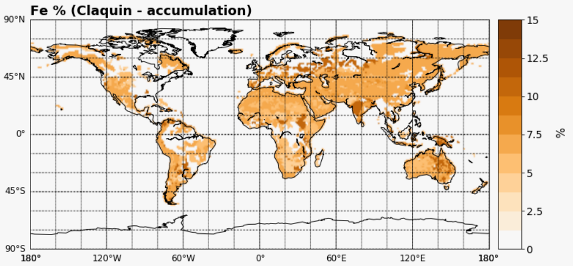
- Acidic dissolution
- Oxalate-promoted Fe dissolution
- Photo-reductive dissolution.

3 The representation of dust mineralogical composition - **two different soil mineralogy datasets:**

Journet et al. 2014



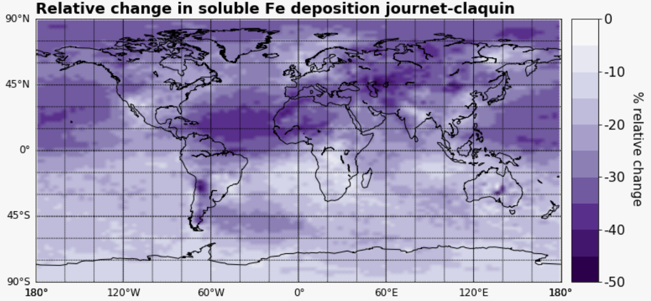
Claquin et al. 1999



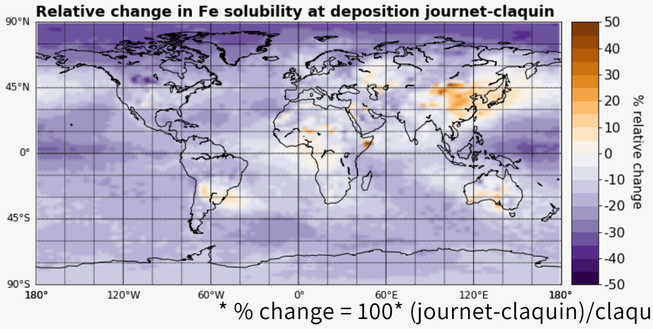
We run two equivalent 1-yr-long simulations with the two soil mineralogy datasets for the year 2011

## RESULTS

### Difference* in soluble-Fe deposition:



### Difference in Fe solubility** at deposition:



* % change = 100* (journet-claquin)/claquin

** Fe-solubity = 100* Soluble-Fe/Total-Fe

## CONCLUSIONS & FUTURE WORK

Our results show a **large sensitivity of the soluble Fe deposition to the choice of the soil mineralogy atlases**, with **differences up to 50%** downwind of major dust source regions such as North Africa. Overall, **soluble Fe deposition is larger when the Claquin dataset is applied**, particularly in the NH. However, **Journet mineralogy derives in higher solubility values for some regions**, e.g. East Asia and areas of the SH.

**The next steps in this work will include exploring how mineralogy affects the Fe-solubilization mechanisms**, e.g. by influencing atmospheric acidity

European Research Council

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

FORCeS

AXA Research Fund

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821205

# Analysis of Hybrid Genomes in the *Candida parapsilosis* Clade

V. del Olmo Toledo[#*1], T. Gabaldón[#*†2]

[#]*Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034-Barcelona, Spain*

[*]*Institute for Research in Biomedicine (IRB Barcelona), C/ Baldiri Reixac, 10, 08028 Barcelona, Spain*

[†]*ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain*

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

***Keywords*— Hybrid, Candida, whole genome sequencing, comparative genomics**

## EXTENDED ABSTRACT

The term inter-species hybridisation refers to the crossing of two divergent organisms, leading to a situation where the two parental genomes coexist in the same nuclear compartment. In higher eukaryotes, this scenario often results in incompatibilities and interference between the genetic material of the two parents, generally detrimental for the newly formed hybrid. However, hybridisation also represents a major source of genomic diversity that can drive adaptation to new niches. After a hybridisation event, the resulting hybrids have a highly heterozygous genome which can, on occasion, derive in extreme phenotypes beneficial for adaptation to new niches or confer properties by new allele combinations that are advantageous with respect to the parentals [1].

In the yeast clade of *Saccharomycotina*, hybridisation has been found to be a rather common phenomenon with numerous hybrid lineages found in industrial environments and many others isolated from clinical settings posing a serious threat to human health [2].

*Candida metapsilosis* and *Candida orthopsilosis* are two emergent fungal pathogens species that belong to the *Candida parapsilosis sensu lato* species complex and have been found to be of hybrid nature [3]. *C. metapsilosis* descends from a single hybridisation event between unknown parentals whereas for *C. orthopsilosis*, the isolates found to date originate from one of four hybridisation events from the same two parental lineages, of which only one has been identified [4,5]. The vast majority of clinical isolates from these two species are hybrids. Parental lineages are never or very rarely isolated from clinical settings suggesting that the pathogenic hybrids might have arisen from non-pathogenic parentals. In other words, that hybridisation might enhance the emergence of new hybrid lineages with an advantage to thrive in new environments, such as in the human host.

This research aims to shed light into the genomic traits that shape yeast hybrids and their evolution. In particular, we sought to understand what the environmental source of these hybrids and their parental species could be, and what are the genomic traits may have facilitated an opportunistic pathogenic behaviour. To this end, we here analyse the genomes of thirteen marine environmental strains from *C. orthopsilosis* and *C. metapsilosis*. We show that the majority of isolates (11 out of 13) are hybrids which expands the map of ecological distributions where these yeasts can be found to include aquatic environments. The fact that hybrids are overrepresented over parental strains also suggests that hybrids have an advantage over parental lineages not only in the clinical settings but in some environmental niches too. We hypothesize that the genomic features that make hybrids highly competitive in certain (perhaps extreme) environments might be also advantageous in other niches like the human body. Consistent with this statement, our phylogenetic reconstruction based on genome-wide polymorphisms shows that the new environmental hybrids fall in (or close to) previously defined clades that harbour clinical isolates.

Until now it has been a complex task to fully characterise the genome of a hybrid cell when one or both of the parentals remained unknown, and parameters like divergence between parentals or percentage of each parental haplotype in the hybrid have so far been based on estimations. In this study we find that two of the marine *C. orthopsilosis* isolates which have highly homozygous genomes represent a long-sought parental lineage so far unidentified. Thus, using a combination of short- and long-read sequencing technologies we generated a genome assembly of the new parental strain which opens a door for future research including the generation of a phased genome with resolved haplotypes that in turn, will lead to a better understanding of the hybrid genomes and a more accurate view of genetic variants.

## ACKNOWLEDGEMENTS

We acknowledge the members of the comparative genomics group for useful input and discussions, especially Dr. Marina Marcet-Houben and former member of the group Dr. Verónica Mixão. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754433.

### *References*

[1] Runemark A, Vallejo-Marin M, Meier JI. 2019. Eukaryote hybrid genomes. *PLoS Genet*. 15:e1008404

[2] Gabaldón T. 2020. Hybridization and the origin of new yeast lineages. *FEMS Yeast Res*. 20:1–8.

[3] Pryszcz LP, Németh T, Gácser A, Gabaldón T. 2014. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol Evol*. 6:1069–1078

[4] Schröder MS, Martinez de San Vicente K, Prandini THR, Hammel S, Higgins DG, Bagagli E, et al. 2016. Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species. *PLoS Genet*. 12:e1006404

[5] Pryszcz LP, Németh T, Saus E, Ksiezopolska E, Hegedűsová E, Nosek J, et al. 2015. The genomic aftermath of hybridization in the opportunistic pathogen *Candida metapsilosis*. PLoS Genet. 11:e1005626

## *Author biography*

**Valentina del Olmo Toledo** received a Bachelors degree in biology from the Universitat Autònoma de Barcelona, Barcelona, Spain, in 2012, and the M.Sc. degree in biology from the Heinrich Heine Universität, Düsseldorf, Germany, in 2014. Valentina obtained her Ph.D. in Natural Sciences at the Institute of Molecular Infection Biology of Würzburg, Germany in 2019. She has a strong background in Candida biology and the molecular techniques to study it, including large-scale NGS data. Since 2020 she is part of the comparative genomics group at the Barcelona Supercomputing Center where she focuses on the analysis of Candida hybrid genomes. Valentina is a recipient of STARS fellowship (which are part of COFUND call of the Marie Sklodowska Curie Actions).

# Analysis of hybrid genomes in the *Candida parapsilosis* clade

**Valentina del Olmo[1,2], Verónica Mixão[1,2], Ester Saus[1,2], Toni Gabaldón[1,2,3]**

*[1]Barcelona Supercomputing Center; [2]Institute for Research in Biomedicine; [3]ICREA*

contact: valentina.delolmo@bsc.es

## 1. Introduction

- Hybridisation is a common phenomenon in yeast and represents a source of novel phenotypic diversity
- *Candida orhopsilosis* and *Candida metapsilosis* are emergent fungal pathogens of hybrid nature
- The vast majority of available samples are hybrid clinical isolates
- Only one *C. orthopsilosis* parental lineage has been identified whereas both *C. metapsilosis* parentals remain unknown

## 2. Our main questions

- Does hybridisation enhance the emergence of hybrid lineages?
- Are environmental isolates more likely to be parental lineages?
- Are the genomic traits that make hybrids highly competitive in environmental niches also advantageous in other niches like the human body?

## 3. Hybrid marine isolates

Genomic analysis of 13 environmental *C. orthopsilosis* and *C. metapsilosis* strains shows a majority of hybrid (11) over non-hybrid (2) strains amongst marine isolates

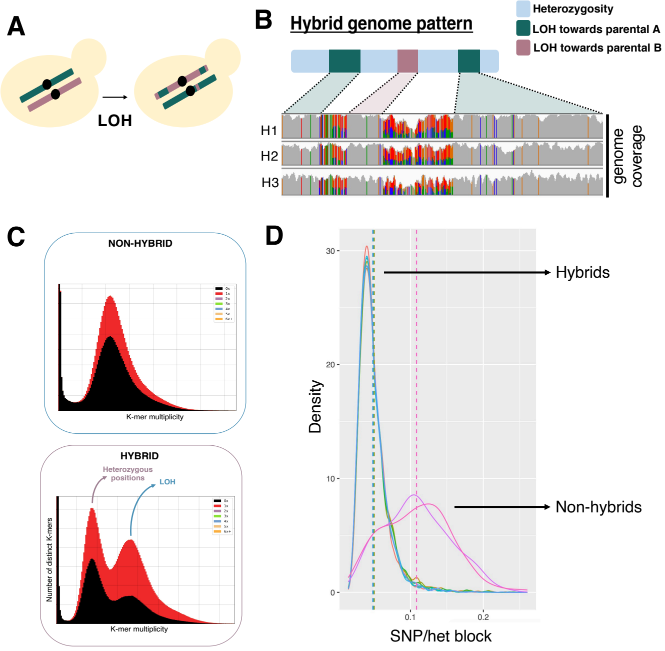


**Figure 1 (A)** After hybridisation event the resulting cells undergo loss of heterozygosity (LOH) leading to a characteristic genomic pattern in hybrids. **(B)** K-mer frequency profiles of hybrids present one peak (coverage X) corresponding to heterozygous positions and a second peak (coverage 2X) portraying LOH. Non-hybrid strains show a single peak. In the heterozygous peak of the hybrids ~50% of the k-mers are present (red) and ~50% absent (black) from the reference genome. **(C)** The density of the divergence between haplotypes in hybrids shows a single peak that translates in a single hybridisation event whereas in non-hybrid strains the divergence does not present a normal distribution.

## 4. Novel *C. orthopsilosis* parental lineage

Amongst the marine isolates we identified two highly homozygous non-hybrid strains which represent a new parental lineage of *C. orthopsilosis*

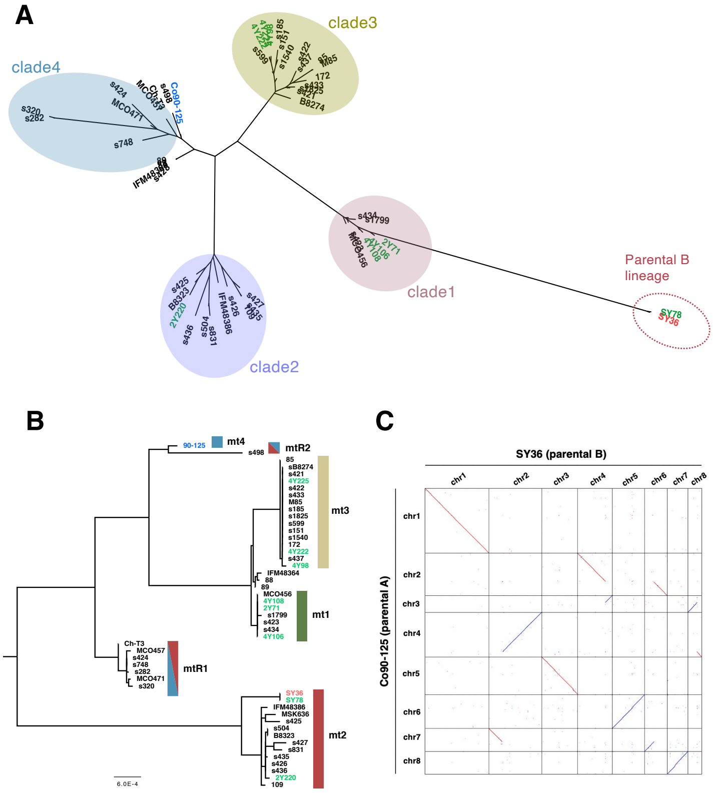


**Figure 2 (A)** Tree based on nuclear SNPs depicts the phylogenetic relationships between all known *C. orthopsilosis* isolates. The known parental strain (Co90-125) is shown in blue. Most marine isolates (green) fall into previously described clades closely related to clinical isolates (black), whereas the two strains – representing the novel parental lineage (red) – appear in a significantly distant branch. **(B)** Phylogenetic tree showing mitochondrial inheritance. The mitochondrial genome can be classified in six mitotypes, two of which (mtR1 and mtR2) are recombinant between mitotyopes 4 and 2 where parent A and B fall, respectively. **(C)** The dot plot shows the similarity between the genome assemblies of the two parental strains. Genome assembly of parent B was generated in this study.

## 5. Conclusions and future work

- Expansion of niches where hybrids can be found to now include marine environments
- The majority of hybrids over parental lineages suggests hybrids' advantage not only in clinics but also in some environments
- The finding of a long-sought *C. orthopsilosis* parental lineage opens a door for the generation of a phased genome and more accurate view of genomic variants
- Future phenotypic analyses might reveal differences between hybrids and parental lineages

## References

**1)** Gabaldón T. 2020. Hybridization and the origin of new yeast lineages. *FEMS Yeast Res.* 20:1–8.
**2)** Pryszcz LP *et al.* 2014. Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol Evol.* 6:1069–1078
**3)** Schröder MS *et al.* 2016. Multiple origins of the pathogenic yeast *Candida orthopsilosis* by separate hybridizations between two parental species. *PLoS Genet.* 12:e1006404
**4)** Pryszcz LP *et al.* 2015. The genomic aftermath of hybridization in the opportunistic pathogen *Candida metapsilosis.* PLoS Genet. 11:e1005626

**This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754433**

# Multiplex network uncovers Chronic Obstructive Pulmonary Disease endotypes

Núria Olvera[1,2,3], Rosa Faner[2,3], Alfonso Valencia[1]
[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
[3] Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Barcelona, Spain
E-mail: nuria.olvera@bsc.es, rfaner@clinic.cat, alfonso.valencia@bsc.es

***Keywords*—*COPD, network medicine, multi-omics, multiplex networks.***

## I. Introduction

Chronic Obstructive Pulmonary Disease (COPD) was the fourth leading cause of death in the world in 2019, and its burden is projected to increase in coming decades in relation to the aging of the population [1]. COPD is characterized by persistent respiratory symptoms and airflow limitation. According to the the level of airflow limitation (FEV1 % ref.), patients are classified into four categories (GOLD groups, Fig.1). Nevertheless, airflow severity is only one component of COPD, as patients with the same level of airflow limitation can present different symptoms, comorbidities and pathological processes (i.e. emphysema, cardiovascular diseases, cachexia, neutrophilic/eosinophilic inflammation) [2]. As a result, COPD is currently viewed as a heterogeneous disease with several endotypes, which are the molecular mechanisms leading to the clinical phenotype of the disease. Recognition of this disease heterogeneity is important as different endo-phenotypes may respond differently to therapies, so that more personalized therapies could be applied.

The main objective of this work is to understand the local and molecular heterogeneity of the disease integrating different types of genomic data which are known to play a role in the pathology. We jointly profiled the mRNA, miRNA and methylome in lung tissue from 135 individuals with different grades of disease severity. In order to integrate all the diversified data, a multiplex patient similarity network was built and communities were detected through unsupervised clustering. Then, these clusters of patients were characterized using the clinical and genetic data available.

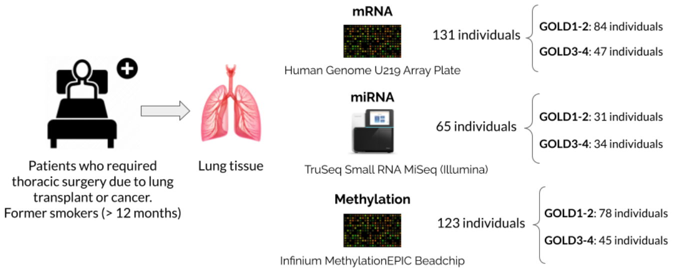


Fig. 1. Pipeline for cohort recruitment and 'omics profiling.

TABLE I. Average/Percentage of clinical features in each community

| | BMI (kg/mˆ2) | FEV1 % ref. | Packs-year | Neutrophils (%) | Eosinophils (%) | Cardiovascular risk (%) |
|---|---|---|---|---|---|---|
| Cluster3 | 27.97 | 67.77 | 55.65 | 63.89 | 2.20 | **83** |
| Cluster5 | 27.14 | 64.60 | 52.94 | 65.96 | **3.34** | 58 |
| Cluster6 | 27.75 | 64.06 | 56.29 | 65.57 | 2.20 | 68 |
| Cluster7 | **24.01** | 38.94 | 57.55 | 65.32 | 2.28 | 47 |
| Cluster8 | 26.88 | 43.12 | **78.5** | 65.76 | 1.71 | 44 |
| Cluster9 | 28.14 | 38.62 | 59 | **75.2** | 1.26 | 44 |

## II. Methods

### A. mRNA, miRNA and Methylome analysis

Lung tissue samples were obtained from COPD patients former smokers and each 'omic was profiled as shown in Figure 1. All the 'omics analysis were done in R using custom scripts. Data went through appropriate quality-control, outliers' elimination, between-sample and within-sample normalization and batch effect removal for network construction and clustering.

### B. Multiplex network

Firstly, in each 'omic data type we selected the most variable genes/probes according to the coefficient of variation. Then, for each data type we built sample-by-sample Pearson's correlation matrices to construct the unweighted undirected networks using the "backbone edges" based based on the calculation of Distance Closure [3]. Through this method, we only kept the edges that made all the nodes reachable using Dijkstra's algorithm of All Pairs Shortest Paths. Communities were detected through the optimization of the multiplex modularity (Louvain algorithm)[4]. Community assignment in each resolution was used as a feature vector in hierarchical clustering to find the definitive partition.

## III. Results

Nine communities were detected in the hierarchical clustering after bootstrapping using the community assignment in each resolution as feature vectors. Community 1, 2 and 4 were discarded since they only bounded between two and three individuals and thus, they were considered as outliers. Table 1 shows the most relevant clinical features for the remaining communities. We obtained three clusters (C3, C5,C6) with a predominance of individuals with a higher lung function (higher FEV1 % ref.) and the other three included patients with a more severe phenotype. Interestingly,we observed that the clusters were associated with well-defined clinical phenotypes, as increased % of eosinophils, or neutrophils. Eosinophilic COPD patients have a better clinical response to inhaled

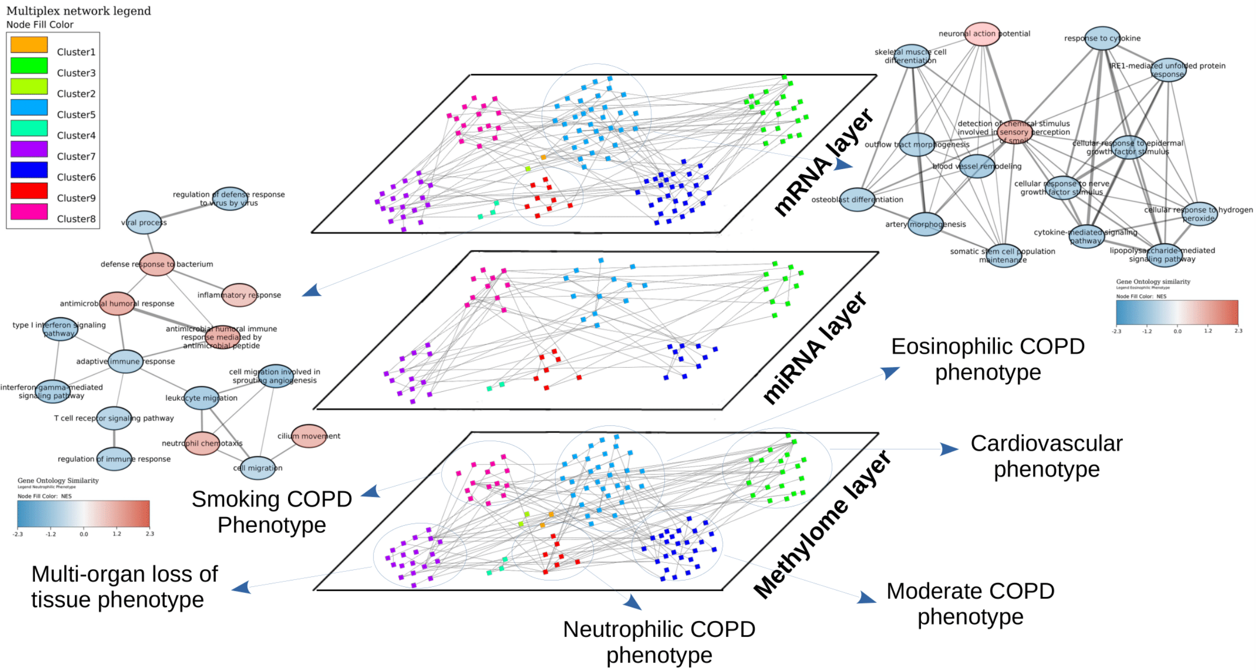

Multiplex network legend
Node Fill Color
Cluster1
Cluster3
Cluster2
Cluster5
Cluster4
Cluster7
Cluster6
Cluster9
Cluster8
mRNA layer
miRNA layer
Methylome layer
regulation of defense response to virus by virus
viral process
defense response to bacterium
inflammatory response
antimicrobial humoral response
type I interferon signaling pathway
antimicrobial humoral immune response mediated by antimicrobial peptide
adaptive immune response
interferon-gamma-mediated signaling pathway
cell migration involved in sprouting angiogenesis
leukocyte migration
T cell receptor signaling pathway
neutrophil chemotaxis
cilium movement
regulation of immune response
cell migration
Gene Ontology Similarity
Legend Neutrophilic Phenotype
Node Fill Color: NES
-2.3
-1.2
0.0
1.2
2.3
neuronal action potential
skeletal muscle cell differentiation
response to cytokine
IRE1-mediated unfolded protein response
detection of chemical stimulus involved in sensory perception of smell
outflow tract morphogenesis
cellular response to epidermal growth factor stimulus
blood vessel remodeling
cellular response to nerve growth factor stimulus
osteoblast differentiation
cellular response to hydrogen peroxide
artery morphogenesis
cytokine-mediated signaling pathway
somatic stem cell population maintenance
lipopolysaccharide-mediated signaling pathway
Gene Ontology similarity
Legend Eosinophilic Phenotype
Node Fill Color: NES
-2.3
-1.2
0.0
1.2
2.3
Eosinophilic COPD phenotype
Cardiovascular phenotype
Smoking COPD Phenotype
Multi-organ loss of tissue phenotype
Moderate COPD phenotype
Neutrophilic COPD phenotype

Fig. 2. Multiplex network representation. Nodes depict patients and are coloured according to the community they belong to. Edges represent molecular similarities (Person correlation) in each 'omic type. Edges were filtered to only keep the ones that made all the nodes reachable using an algorithm of All Pairs Shortest Paths (Distance Closure). Gene Ontology similarity networks (obtained from REVIGO software) are displayed for cluster 9 and 5, where nodes are coloured according to the normalized enrichment score in GSEA. Clinical phenotypes of the communities obtained from 3 layers' integration are outlined at the bottom.

corticosteroids and usually have a milder phenotype, whereas neutrophilic COPD includes patients with bacterial colonization of the lower airways and a worse prognostic [2]. This is in line with the clinical features found for Cluster5 (higher lung function, higher blood eosinophils) and Cluster9 (lower lung function, higher blood neutrophils). Then, we compared the gene expression of each community to the rest of them and performed gene set enrichment analysis (GSEA) in the Gene Ontology database. As shown in Figure 2, individuals of community 9 displayed a significant upregulation of neutrophil chemotaxis and antimicrobial humoral response in comparison to the rest of the communities, which is in consonance with the reported relation between neutrophilic COPD and bacterial colonization.

Community 5 had an upregulation of pathways related to olfactory receptors, which might explain the observed association between eosinophils and chronic rhinosinusitis [5]. They also showed a downregulation of lipopolysaccharides-mediated signalling pathways, which could be due to the reported inverse relation between eosinophils blood counts and bacterial airway infection [6]. The rest of the communities also matched with COPD phenotypes that have been described in the clinical setting, such as a multi-organ loss of tissue or cachexia phenotype and individuals with higher prevalence of metabolic/cardiovascular concominant diseases [7].

## IV. Conclusion and Next Steps

In this study, we report that the multilayer network based only on multi-omics patients' similarities in lung tissue uncovers for the first time communities that resemble clinical COPD endotypes. The next steps of the research will be focused on the understanding of the interplay between the molecular mechanisms in each of the layers.

## References

[1] J. B. Soriano *et al.*, "Mortality prediction in chronic obstructive pulmonary disease comparing the GOLD 2007 and 2011 staging systems: A pooled analysis of individual patient data," *The Lancet Respiratory Medicine*, vol. 3, no. 6, pp. 443–450, jun 2015.

[2] P. J. Barnes, "Endo-phenotyping of COPD patients," pp. 27–37, 2021.

[3] T. Simas and L. M. Rocha, "Distance closures on complex networks," *Network Science*, vol. 3, no. 2, pp. 227–268, jun 2015.

[4] G. Didier *et al.*, "Identifying communities from multiplex biological networks by randomized optimization of modularity." *F1000Research*, vol. 7, p. 1042, 2018.

[5] S. A. Shah *et al.*, "Pathogenesis of eosinophilic chronic rhinosinusitis," apr 2016.

[6] U. Kolsum *et al.*, "Blood and sputum eosinophils in COPD; Relationship with bacterial load," *Respiratory Research*, vol. 18, no. 1, p. 88, may 2017.

[7] A. Corlateanu *et al.*, "Chronic obstructive pulmonary disease and phenotypes: a state-of-the-art." pp. 95–100, mar 2020.

**Núria Olvera** was born in Calaf, Barcelona in 1996. She received her BSc degree in Biomedical Sciences from University of Barcelona in 2018. She completed her MSc degree in Bioinformatics for Health Sciences at Pompeu Fabra University, Barcelona in 2020. Since 2020, she is a PhD student in Àlvar Agustí's lab at Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) and Alfonso Valencia's group at Barcelona Supercomputing Center (BSC), in a collaborative framework between both research groups.

# Multiplex network in Chronic Obstructive Pulmonary Disease

Núria Olvera[1,2,3], Rosa Faner[2,3], Alfonso Valencia[1,4]

1. Barcelona Supercomputing Center (BSC), Barcelona
2. Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona,
3. Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Spain
4. Catalan Institution for Research and Advanced Studies (ICREA), Spain

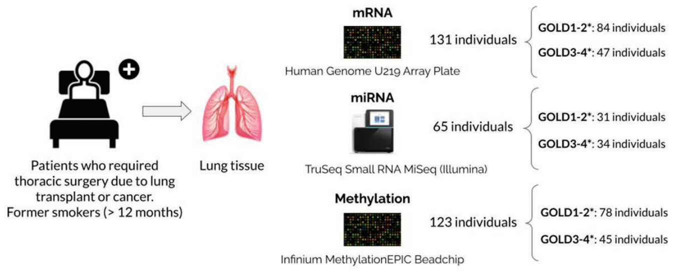
Corresponding author: nuria.olvera@bsc.es

## INTRODUCTION

**Chronic Obstructive Pulmonary Disease (COPD)** is a prevalent disease characterized by persistent respiratory symptoms and airflow limitation. Patients with the same level of airflow limitation can present different symptoms, comorbidities and pathological processes [1] (i.e. emphysema, cardiovascular diseases, cachexia, neutrophilic/eosinophilic inflammation). Thus, it is currently viewed as a **heterogeneous disease** with several **endotypes**, which are the molecular mechanisms leading to the clinical phenotype of the disease.

The main objective of this work is to **understand the local and molecular heterogeneity of the disease integrating different types of genomic data** which are known to play a role in the pathology.

We jointly profiled the **mRNA, miRNA and methylome in lung tissue** from 135 individuals with different grades of disease severity.



***GOLD1-2:** clinical group that includes individuals with mild to moderate airflow limitation.
***GOLD3-4:** clinical group that includes individuals with severe to very severe airflow limitation.

## METHODS

*mRNA, miRNA and methylome analysis*

Data went through appropriate quality-control, outliers' elimination, between-sample and within-sample normalization and batch effect removal for network construction and clustering.

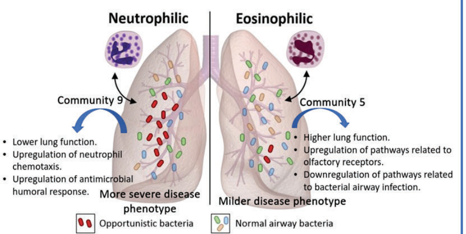
*Multiplex network construction*

For each 'omic, a single patient network was built to represent similarities between individuals.

- Feature selection based on the most variable genes/probes (coefficient of variation).
- Calculation of patient similarity with Pearson's correlation.
- The unweighted undirected network for each data type was built with the "backbone edges" (Distance closure) [2].
- Communities were detected through an adaptation of Louvain algorithm to implement the optimization of modularity in multilayer networks [3].
- Community assignment in each modularity resolution was used as a feature vector in hierarchical clustering to find the definitive partition.
- Pathways were found with Gene Set Enrichment Analysis (GSEA) in mRNA microarray.

## RESULTS

- Nine communities were detected hierarchical clustering after bootstrapping using the community assignment in each resolution as feature vectors (Figure 1 and Table 1).
- Community 1, 2 and 4 were discarded (two and three individuals) → considered as outliers.
- Three clusters (C3, C5, C6) → individuals with a higher lung function (higher FEV1 % ref.).
- Three clusters (C7, C8, C9) → patients with a more severe phenotype (lower FEV1 ref. and higher % emphysema).
- We observed that the clusters were associated with well-defined clinical phenotypes[1],as increased % of eosinophils, or neutrophils:


Neutrophilic
Eosinophilic
Community 9
Community 5
• Lower lung function.
• Upregulation of neutrophil chemotaxis.
• Upregulation of antimicrobial humoral response.
More severe disease phenotype
• Higher lung function.
• Upregulation of pathways related to olfactory receptors.
• Downregulation of pathways related to bacterial airway infection.
Milder disease phenotype
Opportunistic bacteria
Normal airway bacteria

- Other communities also matched with COPD phenotypes that have been described, such as individuals with higher prevalance of cardiovascular concominant diseases [4].

Table 1. Average/Percentage of clinical features in each community. *Kruskal-Wallis test or Fisher's exact test

| | Cluster3 | Cluster5 | Cluster6 | Cluster7 | Cluster8 | Cluster9 | P-value* |
|---|---|---|---|---|---|---|---|
| Age (yrs) | 70.09 | 65.7 | 68.82 | 62.74 | 62.5 | 60.89 | 0.00087 |
| Males, % | 91.3 | 81.82 | 96.43 | 68.42 | 68.75 | 88.89 | 0.052 |
| BMI, kg/m² | 27.97 | 27.14 | 27.75 | 24.02 | 26.88 | 28.15 | 0.068 |
| Smoking exposure (pack-year) | 55.65 | 52.94 | 56.29 | 57.55 | 78.5 | 59 | 0.95 |
| FEV1/FVC | 56.43 | 55.27 | 55.25 | 40.68 | 48.52 | 44.48 | 0.00085 |
| FEV1 % ref. | 67.77 | 64.6 | 64.06 | 38.95 | 43.12 | 38.62 | 5.8e-05 |
| % of CT- Emphysema (Y/ N) | 52.38 | 57.58 | 42.86 | 84.21 | 87.5 | 88.89 | 0.0044 |
| % of Cardiovascular risk (Y/N) | 82.61 | 59.38 | 70.37 | 47.37 | 43.75 | 44.44 | 0.066 |
| % of Dyslipidemia (Y/N) | 60.87 | 25 | 40.74 | 21.05 | 18.75 | 11.11 | 0.018 |
| Neutrophils (%) | 63.9 | 65.96 | 65.57 | 65.32 | 65.77 | 75.2 | 0.14 |
| Monocytes (%) | 6.32 | 7.02 | 6.14 | 7.22 | 6.99 | 3.96 | 0.00074 |
| Eosinophils (%) | 2.21 | 3.34 | 2.21 | 2.29 | 1.71 | 1.27 | 0.051 |

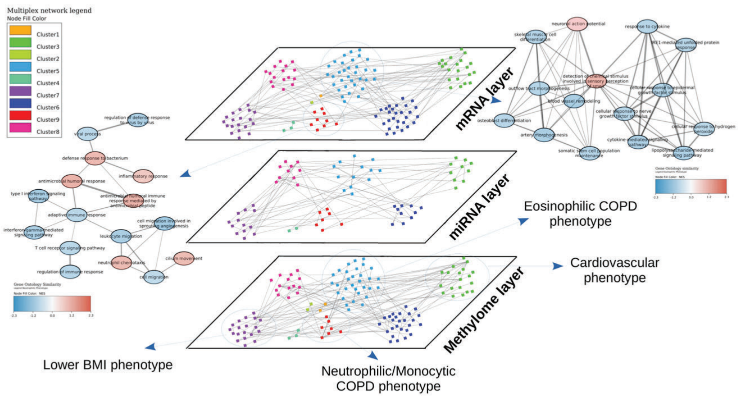

Multiplex network legend
Node Fill Color
Cluster1
Cluster3
Cluster2
Cluster5
Cluster4
Cluster7
Cluster6
Cluster9
Cluster8
regulation of defense response to virus by virus
viral process
defense response to bacterium
inflammatory response
antimicrobial humoral response
type I interferon signaling pathway
antimicrobial humoral immune response mediated by antimicrobial peptide
adaptive immune response
cell migration involved in sprouting angiogenesis
interferon-gamma-mediated signaling pathway
leukocyte migration
T cell receptor signaling pathway
neutrophil chemotaxis
cilium movement
regulation of immune response
cell migration
Gene Ontology Similarity
Node Fill Color: NES
-2.3
-1.2
0.0
1.2
2.3
mRNA layer
miRNA layer
Methylome layer
neuronal action potential
skeletal muscle cell differentiation
response to cytokine
IRE1-mediated unfolded protein response
detection of chemical stimulus involved in sensory perception of smell
outflow tract morphogenesis
cellular response to epidermal growth factor stimulus
blood vessel remodeling
cellular response to nerve growth factor stimulus
osteoblast differentiation
cellular response to hydrogen peroxide
artery morphogenesis
cytokine-mediated signaling pathway
somatic stem cell population maintenance
lipopolysaccharide-mediated signaling pathway
Gene Ontology similarity
Node Fill Color: NES
-2.3
-1.2
0.0
1.2
2.3
Eosinophilic COPD phenotype
Cardiovascular phenotype
Lower BMI phenotype
Neutrophilic/Monocytic COPD phenotype

Figure 1. Multiplex network representation. Nodes depict patients and are coloured according to the community they belong to. Gene Ontology similarity networks (obtained from REVIGO software) are displayed for cluster 9 and 5, where nodes are coloured according to the normalized enrichment score in GSEA. Clinical phenotypes of the communities obtained from 3 layers' integration are outlined at the bottom.

## CONCLUSIONS

In this study, we report that the multilayer network based only on multi-omics patients' similarities in lung tissue uncovers communities that resemble clinical COPD endotypes.

## REFERENCES

[1] P. J. Barnes, "Endo-phenotyping of COPD patients," pp. 27–37, 2021.
[2] T. Simas and L. M. Rocha, "Distance closures on complex networks," Network Science, vol. 3, no. 2, pp. 227–268, jun 2015.
[3] G. Didier et al., "Identifying communities from multiplex biological networks by randomized optimization of modularity." F1000Research, vol. 7, p. 1042, 2018.
[4] A. Corlateanu et al., "Chronic obstructive pulmonary disease and phenotypes: a state-of-the-art." pp. 95–100, mar 2020.

# Lindaview: An OBDA-based tool for self-sufficiency assessment

Victor-Alejandro Ortiz*†, Montserrat Estañol†, Maria-Cristina Marinescu*, Maria-Ribera Sancho*†, Ernest Teniente†
*Barcelona Supercomputing Center (BSC), Barcelona, Spain
†Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
victor.ortiz@bsc.es, estanyol@essi.upc.edu, maria.marinescu@bsc.es, maria.ribera@bsc.es, teniente@essi.upc.edu

***Keywords*—OBDA, Ontology, Self-sufficiency, SSM, SPARQL.**

## I. Extended Abstract

Poverty and social exclusion are a reality in every society. They are complex problems that require updated information and access to scattered data sources to make a proper assessment of a person's situation. To help social workers with these tasks, we developed the Lindaview tool at the suggestion of the Social Services Department of the Barcelona City Council.

Assessment of individuals seeking social assistance is not standardized, as it depends entirely on the social worker's perception and experience. We design a tool that provides an informed starting point for the assessment of an individual's self-sufficiency. Furthermore, we included a section of general statistics, allowing policymakers to access comprehensive, updated, and timely information, empowering them to make data-based decisions when allocating available resources.

Lindaview is an OBDA-based tool. OBDA (Ontology-Based Data Access) is a paradigm that allows accessing data from its original source without data migration or updates on the original data architecture. Moreover, with this paradigm, we can infer implicit information via ontology reasoning.

### A. Self-sufficiency matrix

S. Lauriks et al., define self-sufficiency in [1] as ”achieving an acceptable level of functioning in the essential domains of daily life”. Our tool focuses on evaluating the level of individuals' self-sufficiency. We use as a base the Self-Sufficiency Matrix (SSM), a tool that evaluates different dimensions of an individual's well-being and classifies each dimension into different levels of fulfillment. The SSM originated in the United States, based on the work of [2], where a standard is proposed to evaluate individuals' capacities to fulfill their basic needs. Based on this work, many measurement tools were generated, such as [3], [4], [5]. We base our work on the Catalan version (SSM-CAT), an adaptation of the most popular versions of the SSM, [6], the Dutch version (SSM-D).

The SSM-CAT evaluates 13 different dimensions of an individual's life, such as Finances, Lodging, Work and Education, Mental Health, Domestic relations, etcetera. It also allows classifying the level of self-sufficiency on each dimension utilizing a five-point Likert scale, where ”five” is entirely self-sufficient, and ”one” means that they have acute problems.

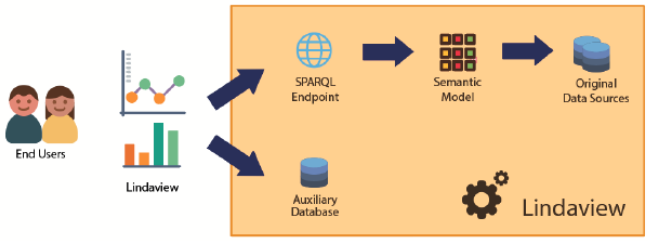


Fig. 1. Overview of the Lindaview tool architecture

### B. Data Access

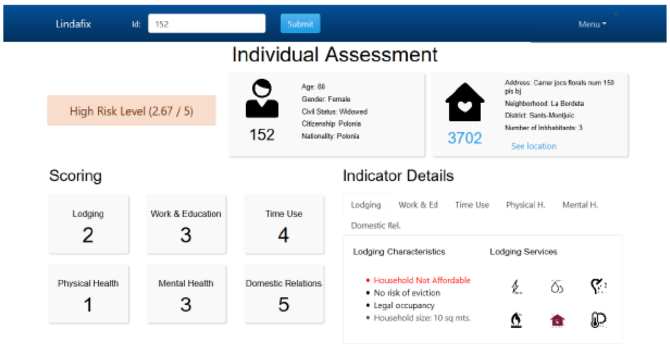
Lindaview allows access to the individuals' data from its original sources. Data may be scattered amongst separate locations or even have different formats, but it is possible to access it in a unified manner by implementing an ontology paired with OBDA. ODBA is a paradigm that consists of generating mappings between the classes in an ontology and the data instances in their original source [7]. This paradigm provides our system with independence from underlying data's technical schema by enabling a single-point of semantic data access.

An ontology is one of the Semantic Web's fundamental concepts as initially proposed by Tim Berners-Lee [8]. [9] defines an ontology as ”an explicit and formal specification of a conceptualization”. One of the major contributions of this work was developing an ontology representing key concepts involved in the self-sufficiency domain.

### C. Architecture

Figure 1 illustrates Lindaview's architecture, describing how the end-user interacts directly with the interface, a visualization developed using Python and Dash. By interacting with the interface, the end-user directly consults the original data sources through the ontology. It can also be observed that the interface queries an auxiliary database, which is generated in order to mitigate the computational power required to calculate the indicators implemented by the tool.

The tool requests data to the ontology by generating SPARQL queries, a query language defined as the standard for semantic web applications by the W3C consortium [10]. These queries are then translated to SQL using OnTop, an open-source system that exposes relational databases as virtual RDF graphs through mappings between an ontology and the original data sources [11].



Lindafix Id: 152 Submit Menu

Individual Assessment

High Risk Level (2.67 / 5)

152
Age: 88
Gender: Female
Civil Status: Widowed
Citizenship: Polonia
Nationality: Polonia

3702
Address: Carrer jocs florals num 150 pis bj
Neighborhood: La Bordeta
District: Sants-Montjuïc
Number of Inhhabitants: 3
See location

Scoring

| Lodging | Work & Education | Time Use |
|---|---|---|
| 2 | 3 | 4 |
| Physical Health | Mental Health | Domestic Relations |
| 1 | 3 | 5 |

Indicator Details

Lodging | Work & Ed | Time Use | Physical H. | Mental H. | Domestic Rel.

Lodging Characteristics

- Household Not Affordable
- No risk of eviction
- Legal occupancy
- Household size: 10 sq mts.

Lodging Services

Fig. 2. Example of the individual assessment visualization

## D. Visualizations

Our tool provides two main scopes: Individual and general statistics.

Figure 2 illustrates the individual screen, showing an overall assessment of the individual's situation, levels of self-sufficiency in all the different domains assessed by the SSM-CAT, and also details of all the concepts involved in the mentioned domains. As part of the individual scope, it is also possible to assess an individual's household - a key concept for calculating self-sufficiency. The household assessment screen provides information related to the services available in an individual's residence, and a summary of the self-sufficiency levels for all the registered inhabitants of that residence.

On the other hand, basic statistics are provided as a general overview, where a summary of all individuals in the population is shown. Moreover, the tool also offers the functionality to segment the information, displayed by nationality or area of residence (neighbourhood or district).

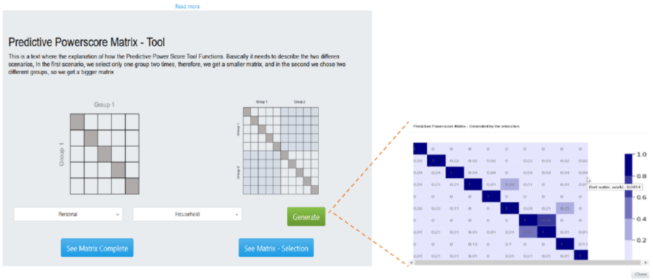
Finally, Lindaview also offers a visualization of the predictive powerscore (PPS) [12]. Figure 3 illustrates the PPS, an asymmetric correlation matrix between the concepts that are part of the different SSM-CAT's dimensions. This visualization helps both, decision-makers and social workers to find patterns of correlation between the concepts involved in self-sufficiency.

## E. Conclusions and future work

For future work, we want to explore the impact of adding a time dimension to the ontology. We believe that more complex analyses could be generated by tracking individuals' self-sufficiency over time. To scale the tool's capacity, we believe that focusing on optimizing the performance of Lindaview when dealing with big data is essential, and we propose to do this by exploring the impact of parallelizing the processing of graphics.

To the extent of our knowledge, current SSM tools work as filling forms, where the end-user must input all the data and do not provide an assessment of the individual's current situation. Providing social workers with an initial assessment can provide them with a supporting baseline to deliver more effective help to individuals in need.

Our work is easy to adapt for any country or city willing to use the SSM by simply applying some minor changes directly



Read more

Predictive Powerscore Matrix - Tool

This is a text where the explanation of how the Predictive Power Score Tool Functions. Basically it needs to describe the two differen scenarios, In the first scenario, we select only one group two times, therefore, we get a smaller matrix, and in the second we chose two different groups, so we get a bigger matrix

Fig. 3. Example of the Predictive Power Score visualization

into the ontology and mappings. This tool does not require any architectural change or data replication; therefore, we believe that this tool could positively impact any institution willing to adopt it.

## II. Acknowledgment

This work's main funding source has been Recercaixa. In addition, it has been partially supported by the Ministry of Science and Technology of Spain through REMEDiAL (TIN2017-87610-R) projects; the Generalitat de Catalunya (under 2017-SGR-1749) and the European Regional Development Fund (ERDF/FEDER)

## References

[1] S. Lauriks *et al.*, "Self-Sufficiency Matrix Manual 2017," GGD Amsterdam, Amsterdam, Tech. Rep., 2017.

[2] D. Pearce, "The Self-Sufficiency Standard for South Dakota," 2000.

[3] E. A. Schonefeld, "LifeWorks Self-Sufficiency Matrix User Manual," Arizona, p. 173, 2013.

[4] CDHS, "Self Sufficiency Matrix - Colorado (Domestic Violence Program)," Tech. Rep.

[5] M. Health and B. Healthcare, "Self-Sufficiency Matrix Guidance for Adult Community Clinical Services Providers," Tech. Rep., 2020.

[6] S. Lauriks *et al.*, "The Use of the Dutch Self-Sufficiency Matrix ( SSM-D ) to Inform Allocation Decisions to Public Mental Health Care for Homeless People," pp. 870–878, 2014.

[7] E. Kharlamov *et al.*, "Ontology Based Data Access in Statoil," *Journal of Web Semantics*, vol. 44, pp. 3–36, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.websem.2017.05.005

[8] T. Berners-Lee *et al.*, "The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 73, no. May, pp. 303–314, 2001.

[9] G. Antoniou *et al.*, *A Semantic Web Primer*, 2012, vol. 3rd Editio.

[10] S. Harris. Sparql 1.1 query language. [Online]. Available: https://www.w3.org/TR/sparql11-query/

[11] D. Calvanese *et al.*, "Ontop: Answering SPARQL queries over relational databases," *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2017.

[12] F. Wetschoreck *et al.*, "8080labs/ppscore: zenodo release," Oct. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.4091345

**Victor-Alejandro Ortiz** received his BSc. degree in Informatics from the National Autonomous University of Mexico, Mexico, and the MSc. degree in Big Data and HPC from the University of Liverpool, UK. Currently pursuing a Ph.D. in Computer Science at the Universitat Politècnica de Catalunya (UPC) and the Barcelona Supercomputing Center, Spain.

# Lindaview

## An OBDA-based tool for self-sufficiency

### Self Sufficiency

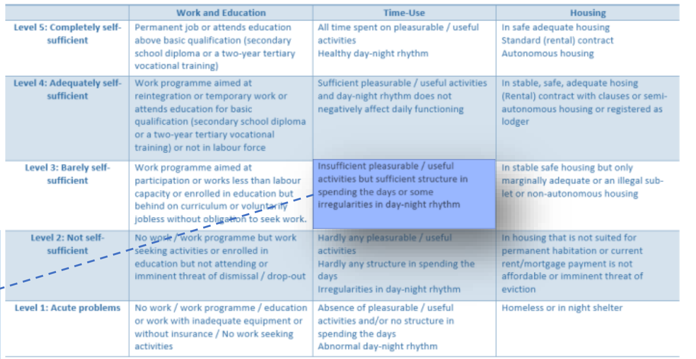
Capacity of a person to meet their basic needs on a daily basis.

### Self-Sufficiency Matrix (SSM)

Tool to standardize the assessment on an individual's self-sufficiency.

### SSM-CAT

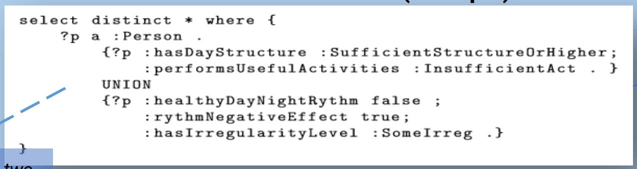
Contains 13 domains, each one categorized into 5 different levels



| | Work and Education | Time-Use | Housing |
|---|---|---|---|
| Level 5: Completely self-sufficient | Permanent job or attends education above basic qualification (secondary school diploma or a two-year tertiary vocational training) | All time spent on pleasurable / useful activities<br>Healthy day-night rhythm | In safe adequate housing<br>Standard (rental) contract<br>Autonomous housing |
| Level 4: Adequately self-sufficient | Work programme aimed at reintegration or temporary work or attends education for basic qualification (secondary school diploma or a two-year tertiary vocational training) or not in labour force | Sufficient pleasurable / useful activities and day-night rhythm does not negatively affect daily functioning | In stable, safe, adequate hosing (Rental) contract with clauses or semi-autonomous housing or registered as lodger |
| Level 3: Barely self-sufficient | Work programme aimed at participation or works less than labour capacity or enrolled in education but behind on curriculum or voluntarily jobless without obligation to seek work. | Insufficient pleasurable / useful activities but sufficient structure in spending the days or some irregularities in day-night rhythm | In stable safe housing but only marginally adequate or an illegal sub-let or non-autonomous housing |
| Level 2: Not self-sufficient | No work / work programme but work seeking activities or enrolled in education but not attending or imminent threat of dismissal / drop-out | Hardly any pleasurable / useful activities<br>Hardly any structure in spending the days<br>Irregularities in day-night rhythm | In housing that is not suited for permanent habitation or current rent/mortgage payment is not affordable or imminent threat of eviction |
| Level 1: Acute problems | No work / work programme / education or work with inadequate equipment or without insurance / No work seeking activities | Absence of pleasurable / useful activities and/or no structure in spending the days<br>Abnormal day-night rhythm | Homeless or in night shelter |

**SSM-CAT extract**

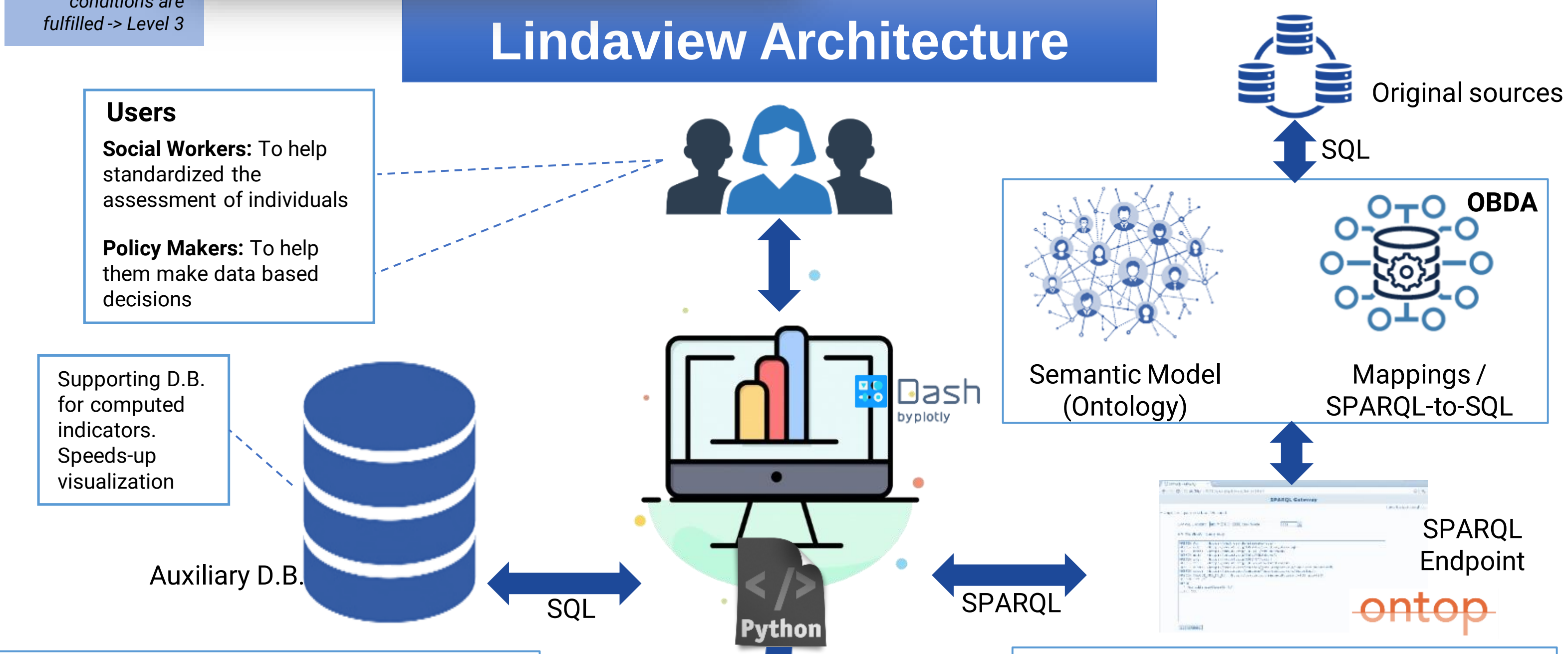
### Time-Use Indicator Evaluation (Example)



```
select distinct * where {
    ?p a :Person .
        {?p :hasDayStructure :SufficientStructureOrHigher;
            :performsUsefulActivities :InsufficientAct . }
        UNION
        {?p :healthyDayNightRythm false ;
            :rythmNegativeEffect true;
            :hasIrregularityLevel :SomeIrreg .}
}
```

*If either of the two conditions are fulfilled -> Level 3*

## Lindaview Architecture



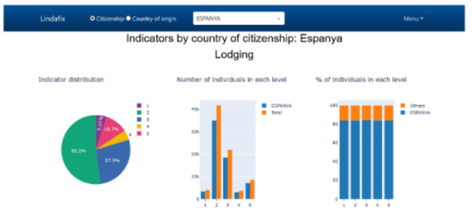
### Users

**Social Workers:** To help standardized the assessment of individuals

**Policy Makers:** To help them make data based decisions

Supporting D.B. for computed indicators. Speeds-up visualization

## General Views

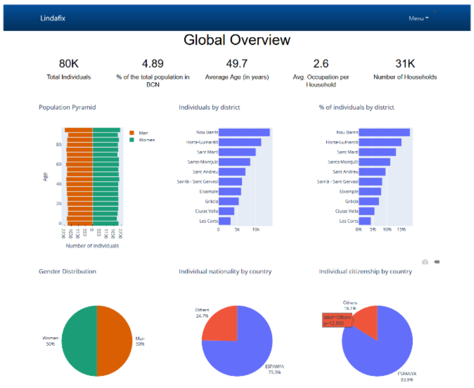


Lindafix ○ Citizenship ● Country of origin ESPANYA Menu

Indicators by country of citizenship: Espanya

Lodging

Indicator distribution
Number of individuals in each level
% of individuals in each level
52.2%
27.5%
16.7%
ESPANYA
Total
Others

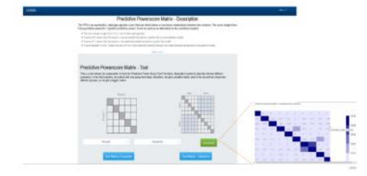


Lindafix Menu

Global Overview

| 80K | 4.89 | 49.7 | 2.6 | 31K |
|---|---|---|---|---|
| Total Individuals | % of the total population in BCN | Average Age (in years) | Avg. Occupation per Household | Number of Households |

## Predictive Power Score

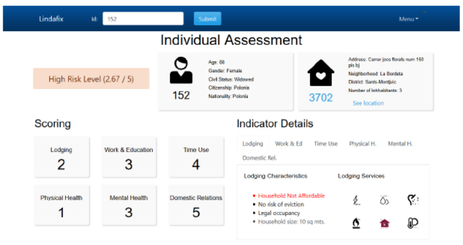


Data Analysis visual tool able to detect patterns and relations in data.

Generic views provides summarized data and statistics.

Detailed views to analyze an individual's current situation

## Individual Views



Lindafix Id: 152 Submit Menu

Individual Assessment

High Risk Level (2.67 / 5)

152

Age: 88
Gender: Female
Civil Status: Widowed
Citizenship: Polonia
Nationality: Polonia

3702

Address: Carrer jocs florals num 150 pis bj
Neighborhood: La Bordeta
District: Sants-Montjuïc
Number of Inhabitants: 3
See location

Scoring

| Lodging | Work & Education | Time Use |
|---|---|---|
| 2 | 3 | 4 |
| **Physical Health** | **Mental Health** | **Domestic Relations** |
| 1 | 3 | 5 |

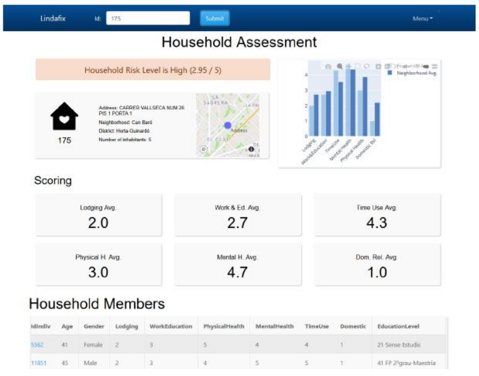
Indicator Details

Lodging Work & Ed Time Use Physical H. Mental H. Domestic Rel.

Lodging Characteristics

- Household Not Affordable
- No risk of eviction
- Legal occupancy
- Household size: 10 sq mts.

Lodging Services



Lindafix Id: 175 Submit Menu

Household Assessment

Household Risk Level is High (2.95 / 5)

175

Address: CARRER VALLSECA NUM 26 PIS 1 PORTA 1
Neighborhood: Can Baró
District: Horta-Guinardó
Number of inhabitants: 5

Scoring

| Lodging Avg. | Work & Ed. Avg. | Time Use Avg. |
|---|---|---|
| 2.0 | 2.7 | 4.3 |
| **Physical H. Avg.** | **Mental H. Avg.** | **Dom. Rel. Avg.** |
| 3.0 | 4.7 | 1.0 |

Household Members

| IdIndiv | Age | Gender | Lodging | WorkEducation | PhysicalHealth | MentalHealth | TimeUse | Domestic | EducationLevel |
|---|---|---|---|---|---|---|---|---|---|
| 5362 | 41 | Female | 2 | 3 | 5 | 4 | 4 | 1 | 21 Sense Estudis |
| 11851 | 45 | Male | 2 | 3 | 4 | 5 | 5 | 1 | 41 FP 2ºgrau-Maestria |

**Authors:**

Alejandro Ortiz, Barcelona Supercomputing Center / Universitat Politècnica de Catalunya
Montserrat Estañol, Barcelona Supercomputing Center / Universitat Politècnica de Catalunya
Maria Cristina Marinescu, Barcelona Supercomputing Center
Maria Ribera Sancho, Supercomputing Center / Universitat Politècnica de Catalunya
Ernest Teniente, Universitat Politècnica de Catalunya

**Tool open access: http://growsmarter.bsc.es:8051**

Contact to victor.ortiz@bsc.es

UPC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Adaptive Optics Control with Reinforcement Learning: First steps

Bartomeu Pou*†, Eduardo Quiñones*, Mario Martín†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {bartomeu.poumulet, eduardo.quinones}@bsc.es, mmartin@cs.upc.edu

***Keywords*—Reinforcement Learning, Adaptive Optics, Non-linear Control, Machine Learning.**

## I. Extended Abstract

### A. *Introduction*

When planar wavefronts from distant stars traverse the atmosphere, they become distorted due to the atmosphere's inhomogeneous temperature distribution. Adaptive Optics (AO) is the field in charge of correcting those distortions allowing high-quality observations of distant targets. The AO solution is composed of three main components: a deformable mirror (DM) that corrects the deformation in the wavefront, a wavefront sensor (WFS) that allows characterising the current turbulence in the wavefront and a real time controller (RTC) that issues commands to, via the deformation of the DM, correct the wavefront. Usually, the operations are performed on closed-loop with stringent real-time requirements (in the order of $10^3 - 10^4$ actions per second). At each iteration, the WFS observes the wavefront after being corrected by the DM and the RTC issues the commands to correct for the evolution of turbulence and previous uncorrected errors (Figure 1 left).

One of the primary sources of error for an AO control algorithm is the temporal error. The delay between characterising the turbulence with the WFS and setting the desired commands in the DM creates the need that any successful control approach must take into account past commands and the probable evolution of the atmosphere in this gap of time. To do that, the most common approach in AO are variants of Linear Quadratic Gaussian (LQG) with Kalman filters with one of its initial iterations presented in [1]. Usually, a linear model of the system's evolution is built with a set of parameters that are usually fitted based on observations or on theoretical assumptions, which limits the capability of the system to correct the turbulence.

In this paper, we present a novel solution based on Reinforcement Learning (RL), based on a reward signal to be optimised, that does not need any previously built model (as LQG) and is non-linear. RL has been already applied in the domain of AO, however, it has been limited to WFS-less systems (e.g. [2]) or, more recently, to control a very limited number of actuators [3]. This work's main practical objective is to be applied in the 8.2 m Subaru telescope (located in Hawaii), which includes thousands of actuators.

### B. *AO Control: Integrator with gain*

The traditional AO control algorithm is the integrator with gain. At each iteration, the WFS obtains a vector of measurements, $m$, where each element indicates a local deviation from the seen wavefront to a planar wavefront. The relationship between $m$ and commands in the DM at a timestep, $t$, can be approximated as a linear relationship:

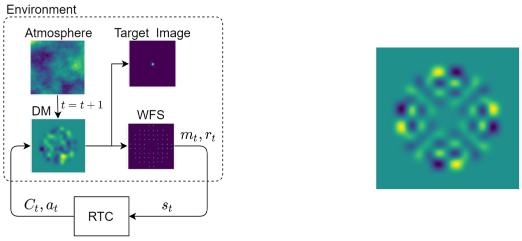


Fig. 1: Left: AO closed-loop. Right: Mode example.

$$m_t = D \cdot c_t \tag{1}$$

Where $D$ is the interaction matrix obtained with a least squares approach method on the loss $||m - Dc||^2$. By inverting the interaction matrix in equation (1), we obtain the commands to be applied to the DM to correct the current wavefront deviations on ideal conditions. To deal with non-ideal issues, such as the temporal error, integration with past commands, $C$, with a gain factor, $g$, is used:

$$C_t = C_{t-1} + gc_t \tag{2}$$

### C. *Adaptive Optics as a Reinforcement Learning problem*

RL [4] is concerned of finding a function (called policy, $\pi(\theta)$ parametrized with weights $\theta$) that maximises the expected cumulative reward ($r$) obtained by interacting with an environment. To do so, RL maps the state describing the environment ($s$) to actions ($a$) with the objective of obtaining the optimal policy, $\pi^*(\theta)$.

$$\pi^*(\theta) = \underset{\theta}{arg\,max} \underset{env}{\mathbb{E}} \left[ \sum_i r_i(\pi(\theta)) \right] \tag{3}$$

Concretely, RL requires the following elements:

*1) The states, $s$:* Defined as the union of the integrator commands, $c$, which will give information of current perturbation in the atmosphere and past integrated commands, $C$, which will give information of commands that will be applied in the next timesteps (due to delay), and the evolution of the atmosphere at every time step $t$: $s_t = (c_t, C_{t-1}, C_{t-2}, ..., C_{t-n})$.

The commands issued by the RTC are usually a vector of $n_a$ dimensions ($C \in \mathbb{R}^{n_a}$) where each element of the command vector controls one actuator in charge of deforming the mirror. This way of handling the commands is said to be zonal as each value of the command vector only modifies a particular zone of the mirror. However, one can build a modal basis, e.g. by using

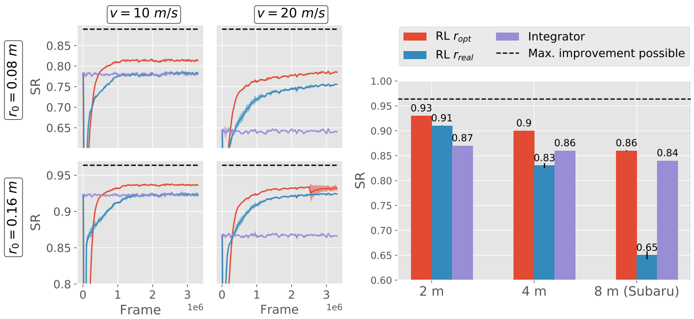


Fig. 2: Left: Training curves (77 modes, $D$=2 m). Right: Avg final performance (62 modes, $r_0$=0.16 m, v=20 m/s). Results averaged over 3 seeds.

the Zernike polynomials [5] (see Figure 1), to act globally in the whole DM with each element of the command vector. The usage of modal basis has two benefits: (1) we can just correct a subset of modes if the number of actuators to control is problematically high and (2) RL method performance depends on the feature definition [4]. Empirically, we have observed that a value of $n = 3$ for the state and using a modal basis for the commands $C_t$ leads to better performance.

*2) The actions, a:* Defined as a correction applied to the commands computed with the ”integration with gain” AO control, as follows: $C_t = C_{t-1} + gc_t + a$.

*3) The reward, r:* Defined as a function that determines how well the turbulence distortion has been corrected. To do that, we apply two different strategies: (1) A reward based on the spatial variance of the wavefront phase $\phi$, $r_{opt,t} = -var(\phi_t)$, in which a variance of 0 indicates that all the wavefront points are on phase, hence, the wavefront is planar; and (2) a reward based on the average measurements, $m_t$, squared: $r_{real,t} = -avg(m_t^2)$. The former strategy is optimal but unrealistic as it is not possible to get the variance of the wavefront at each timestep; the latter provides an approximation of the former strategy but can be obtained at each time step.

*4) The algorithm:* We choose Soft Actor Critic, which slightly modifies eq. 3 to include the entropy of the policy, $\pi(\theta)$, as a regularisation term [6].

### D. Results

This section evaluates the AO RL controller in a different range of atmospheric conditions, specifically, different values for Fried parameter, $r_0$, which a lower value denotes a higher strength of turbulence, and wind speed, $v$, which a higher value will drive up the temporal error. Moreover, it evaluates the performance of RL when increasing the telescope diameter, $D$, and thus the complexity of the problem as the number of actuators to control, and the number of measurements of the WFS to process, increases as well, when considering the optimal ($r_{opt}$) and the realistic ($r_{real}$) rewards. The performance is measured in terms of Strehl Ratio ($0 \leq SR \leq 1$), the ratio between the peak intensity of the target image and its theoretical maximum. The experiments presented use an AO control simulator named COMPASS [7], including the simulation of the atmosphere and the AO control, executed on a IBM Power9 8335-GTH CPU (40 cores) with a GPU NVIDIA V100 (16 GB).

Figure (2) *left* evaluates different atmospheric conditions. We can observe that the RL agent outperforms the traditional integrator and is both robust to variations of $v$ and $r_0$ with both reward functions, i.e., $r_{opt}$ and $r_{real}$. RL agent's quasi-constant performance in terms of wind speed may indicate that we are solving mainly temporal error.

Figure (2) *right* compares the performance of the RL agent when controlling 62 modes while increasing the telescope diameter with a fixed atmospheric configuration. While the RL agent with a limited number of modes performs better when compared with the integrator, the agent is incapable of scaling to bigger diameters with the realistic reward function, $r_{real}$. It therefore remains as future work to derive a more efficient reward function. Furthermore, while the use of a modal basis allows to significantly reduce the state's size and so avoid the problem of the curse of dimensionality [4], it remains as a future work as well, to control a higher number of modes. In addition to performance, we must take into account the inference time. Currently, for the given machine and 62 modes, the inference time is $\sim 1.2\ ms$ which is below the threshold of 2 $ms$ to not increase the delay as to affect proper operation of the telescope. However, we must take into account that for large telescopes we will probably end up controlling a higher number of modes hence increasing the inference time.

### E. Conclusion

We have presented a novel AO control based on RL that outperforms traditional controllers in a set of limited experiments. However, we must address some challenges before its application in the real world.

## II. Acknowledgment

This research has been conducted in collaboration with Dr. Damien Gratadour (Paris Observatory, PSL University and Australian National University).

## References

[1] R. N. Paschall and D. J. Anderson, “Linear quadratic gaussian control of a deformable mirror adaptive optics system with time-delayed measurements,” *Applied optics*, vol. 32, no. 31, pp. 6347–6358, 1993.

[2] K. Hu, Z. Xu, W. Yang, and B. Xu, “Build the structure of wfsless ao system through deep reinforcement learning,” *IEEE Photonics Technology Letters*, vol. 30, no. 23, pp. 2033–2036, 2018.

[3] R. Landman, S. Y. Haffert, V. M. Radhakrishnan, and C. U. Keller, “Self-optimizing adaptive optics control with reinforcement learning,” in *Adaptive Optics Systems VII*, vol. 11448. International Society for Optics and Photonics, 2020, p. 1144849.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[5] R. J. Noll, “Zernike polynomials and atmospheric turbulence,” *JOsA*, vol. 66, no. 3, pp. 207–211, 1976.

[6] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, “Soft actor-critic algorithms and applications,” *arXiv preprint arXiv:1812.05905*, 2018.

[7] F. Ferreira, D. Gratadour, A. Sevin, N. Doucet, F. Vidal, V. Deo, and E. Gendron, “Real-time end-to-end ao simulations at elt scale on multiple gpus with the compass platform,” in *Adaptive Optics Systems VI*, vol. 10703. International Society for Optics and Photonics, 2018, p. 1070347.

**Bartomeu Pou** is a PhD student at Barcelona Supercomputing Center with research focused on applying reinforcement learning to adaptive optics in large telescopes. He has a background in physics (bachelor's) and artificial intelligence (master's) and previously has worked on Accenture as a data scientist on the domains of supply chain and healthcare.

# Adaptive Optics Control with Reinforcement Learning: First steps

*B. Pou[1,2], E. Quiñones[1], D. Gratadour [3,4], M. Martín[2]*

*[1] Barcelona Supercomputing Center (BSC).*
*[2] Universitat Politècnica de Catalunya (UPC).*
*[3] Research School of Astronomy and Astrophysics, Australian National University.*
*[4] LESIA, Observatoire de Paris, Universite PSL, CNRS, Sorbonne Universite, Univ. Paris Diderot, Sorbonne Paris Cite.*

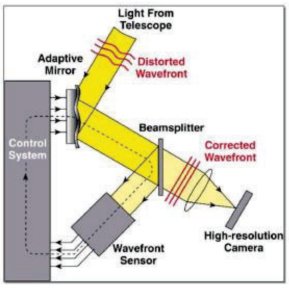
## 1. Motivation

- In ground-based telescopes, the **light from distant stars is distorted** due to small variations of index of refraction in the atmosphere
- **Adaptive Optics** (AO) systems are responsible of correcting the distortion by means of a **deformable mirror (DM).**

*Image of Ground-Based Telescopes*
*Credit: Claire E. Max, UCSC*

## 2. Adaptive Optics (AO)

- An AO system characterizes the **distortion** ($\boldsymbol{m_t}$) using a **Wavefront sensor (WFS)**
- **A Real-time Controller (RTC)** computes commands to the **DM actuators** ($\boldsymbol{c_t}$) to correct observed distortion, considering the following **linear relationship**:
  - (1) $c_t = R \cdot m_t$
  - (2) $C_t = C_{t-1} + g \cdot c_t$
- **The RTC have high-performance and real-time requirements**
  - Commands must be issued every ~2ms to ensure the correct operation



*AO Control-Loop*

*Credit: Claire E. Max, UCSC.*

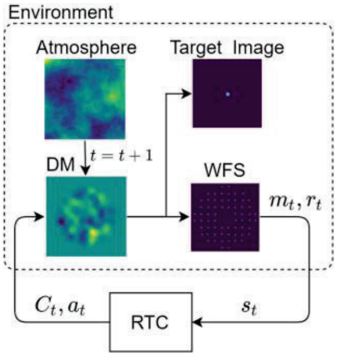
**Large telescopes includes non-linear effects not captured by current RTC that diminishes the performance of telescopes**

## 3. Reinforcement Learning (RL)

**RL allows capturing non-linear effects not addressed by linear solutions**

**RL objective**: find a function, $\boldsymbol{\pi(s)}$, that maps **states (*s*)** to **actions (*a*)** that maximizes a cumulative **reward function (*r*)** via trial and error**.**

- ***a*** corresponds to a **correction term** to the linear RTC: $C_t = C_{t-1} + g \cdot c_t + a_t$
- ***s*** **=** $(\boldsymbol{c_t, C_{t-1}, C_{t-2}, \ldots, C_{t-N}})$ corresponds to the current linear and **previous commands** and provides information about:
  - Commands that will be executed in the future
  - Statistics of evolution of the atmosphere.
- **r** is based on spatial variance of the wavefront phase, $\phi_t$ and average of measurements squared: $r_{opt} = -var(\phi_t)$ and $r_{real} = -avg(m_t^2)$, respectively



*AO loop with RL elements*

Our RL agent does not consider a single DM actuator but **global orthogonal shapes in the DM** inspired in **Zernike polynomials [1].**

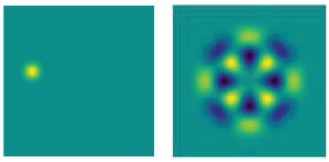
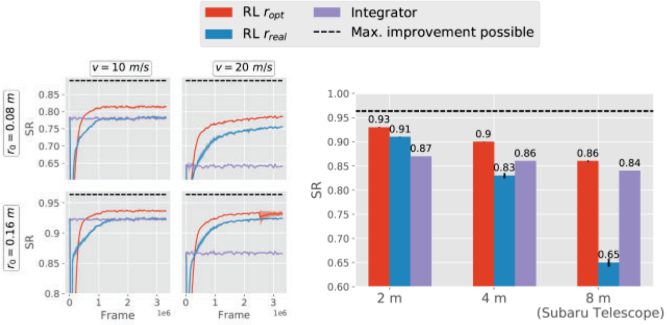


*Image of DM shape.*
*(Left) acting on a single actuator.*
*(Right) acting on a single mode.*

## 4. Results

### Experiments:

- Simulated in **COMPASS** [2] (GPU-based high-performance AO simulations).
- Characterisation of different range of atmospheric conditions.
  - **Fried parameter**, $r_0$: inverse relationship to strength of turbulence.
  - **Wind speed**, v. Related to temporal error.
- Different diameter **(D)** of telescope.
- Measuring results in **Strehl Ratio** ((*worst*) 0 ≤ SR ≤ 1 (*best*)).



*a) RL agent (77 modes) on 2m telescope with different atmospheric conditions.*

*b) RL agent controlling 62 modes with different D. Atmospheric conditions constant.*

*Results*

## 5. Conclusions

- RL agent **tackles non-linear effects such as temporal error**.
- **Dimensionality problem**.
- Realistic reward function does not work for large telescopes.

## 6. Future work

- **Multi-agent system**: each agent controls a small amount of modes.
- **Preliminary results** show an **improvement** over the integrator with a large telescope and realistic reward function.

### References:

[1] R. J. Noll, "Zernike polynomials and atmospheric turbulence,"JOsA,vol. 66, no. 3, pp. 207–211, 1976.

[2] F. Ferreira, D. Gratadour, A. Sevin, N. Doucet, F. Vidal, V. Deo, and E. Gendron, "Real-time end-to-end ao simulations at elt scale on multiple gpus with the compass platform," in Adaptive Optics Systems VI, vol. 10703.International Society for Optics and Photonics, 2018,p. 1070347.

Figure 1 and 2 extracted from https://www.ucolick.org/~max/289/ Lecture 1.

# Modeling nitric acid uptake by mineral dust

Rubén Soussé Villa*, Oriol Jorba Casellas*, Carlos Pérez García-Pando*†
*Barcelona Supercomputing Center, Barcelona, Spain
†ICREA, Catalan Institution For Research and Advanced Studies, Barcelona, Spain
E-mail: {ruben.sousse, oriol.jorba, carlos.perez}@bsc.es

***Keywords*—mineral dust aerosols, dust heterogeneous chemistry, mineralogy, nitric acid.**

## I. Extended Abstract

### A. Introduction

Mineral dust is amongst the largest contributors to the global aerosol mass load and dominates climate effects over large areas of the Earth. Dust undergoes heterogeneous chemical reactions during transport that increase its hygroscopicity, while altering its optical properties, and the associated radiative forcing. The rates of heterogeneous chemical reactions on the dust surface that form coatings of sulfate, nitrate, chloride, or organics depend strongly on the dust mineralogical composition. For example, the uptake of sulfur dioxide by calcite exceeds by at least an order of magnitude uptake by quartz, feldspar and hematite. Dust composition also affects the partitioning of semi-volatile inorganic compounds, altering their burden and radiative forcing.

### B. Objectives

In this preliminary work we first present an overview of the state-of-the-art regarding the representation of the uptake of nitric acid ($HNO_3$) by mineral dust in models. We have also implemented the uptake of $HNO_3$ that forms coarse nitrate in the *Multiscale Online Nonhydrostatic AtmospheRe CHemistry* model (MONARCH) [1] and performed a sensitivity study simulating a series of pollution events over Beijing that involve the formation of coarse nitrate using constant (0.1), null and humidity-dependent uptake coefficients. The main objective is to set a benchmark to conduct future sensitivity studies for factors affecting dust heterogeneous chemistry, such as the explicit treatment of mineralogical composition of dust or ambient relative humidity.

### C. Theoretical background

Heterogeneous dust chemistry implies mainly acidic trace gases resulting in the formation of coatings on the dust particle surface. Uptake of sulfate and nitrate are the major reactions involved [2]. For this study, the reaction evaluated is the nitric acid ($HNO_3$) uptake on coarse dust particles (diameter above $2.5\mu m$), that can be expressed by the reaction $HNO_3$ + dust $\longrightarrow NO_3{}^-$ [3], [4]. The reaction rate of this uptake can be defined as a first-order function as [3]:

$$K = \left(\frac{r}{D_g} + \frac{4}{v\gamma}\right)^{-1} \times S \tag{1}$$

Where $r$ is the aerosol bin radius, $D_g$ is the gas-phase diffusion coefficient, $v$ the mean molecular speed, $S$ the aerosol specific surface area, and $\gamma$ the uptake coefficient, defined as the ratio of the number of gas molecules depositing on the particle's surface over the total molecules colliding with the given surface.

For the uptake of $HNO_3$, the value of the uptake coefficient is typically taken as $\gamma = 0.1$ for mineral dust [4], [5]. Previous studies have shown, however, that using $\gamma = 0.1$ overestimates the particulate nitrate formation [4]. Several experimental studies highlight the strong dependence of $\gamma$ to ambient relative humidity [4], [5], [6] and with mineral dust composition (specifically on calcite percentage) [2], [5].

The uptake coefficient dependence on relative humidity ($\gamma(RH)$) has been shown to behave similarly to a Brunauer–Emmett–Teller isotherm for water adsorption on dust particles described by the function [6]:

$$\gamma = m \times \frac{cRH}{(1-RH)(1-(1-c)RH)} \tag{2}$$

Where $RH$ is relative humidity, $c$ the water adsorption scaling factor ($c = 8$ [6]) and $m$ the specific dust mineralogy scaling factor, taking different values depending on the calcite content. For example, a factor of $m = \frac{1}{30}$ for Arizona Test Dust has been proposed [6], while $m = 0.018$ for China Loess from Gobi's desert with 39% $CaCO_3$ content based on experimental measurements has been used [5], [7].

### D. Methodology

In this work we simulate 3 pollution episodes over Beijing happening in 2015 between the 28th of March and the 2nd of April, being: 1) Pure anthropogenic, 2) pure dust from Gobi desert, and 3) dust mixed with anthropogenic pollutants (numbers in figure 1, respectively). To evaluate the results, we use observations of fine and coarse nitrate surface concentration during these events from the Beijing Institute of Atmospheric Physics (IAP, 116.4ºE, 39.9ºN) [7].

We use the MONARCH model [1] over a regional domain covering Asia and the region of Beijing with an horizontal resolution of 0.2 by 0.2 degrees and 24 vertical layers up to 50hPa. The meteorological initial and boundary conditions are from the NCEP FNL analyses . Emissions are taken from the CAMSv2.1 global inventory for anthropogenic emissions and GFASv1.2 for biomass burning emissions. The Carbon Bond 2005 chemical mechanism is applied for the gas-phase chemistry, and the aerosol module describes the lifecycle of dust, sea-salt, black carbon, organic matter (both primary and secondary), sulfate and nitrate aerosols [8]. While a sectional approach is utilized for dust and sea-salt, a bulk description of the other aerosol species is adopted.

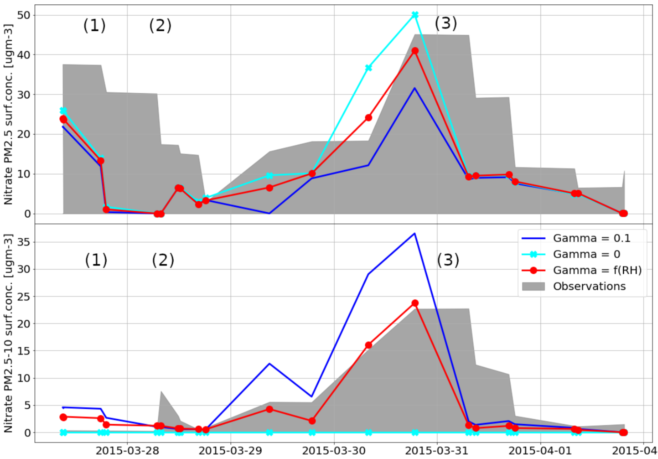


Fig. 1. (Top) Fine nitrate partition ($D < 2.5\mu m$) over Beijing for the 3 pollution events (indicated with numbers) simulated in MONARCH with uptake coefficient $\gamma = 0.1$ (blue), $\gamma = 0$ (cyan, crossed line) and $\gamma$ as function of relative humidity as described in section I-C (red, doted line), compared with the observations from IAP (gray) [7]. (Bottom) The same as top figure for coarse nitrate ($D > 2.5\mu m$).

For this study, aerosol nitrate formation in MONARCH is computed in two consecutive steps: for the fine mode (diameter under 2.5 $\mu m$), the thermodynamic equilibrium model EQSAM is used, assuming thermodynamic equilibrium. For the coarse mode (diameter greater than 2.5 $\mu m$) the specific mass transfer is computed for each of the dust and sea salt bins using the scheme described in section I-C. This methodology has been applied in the reference studies with satisfactory results [9], [7]. Three cases have been simulated: 1) assuming a constant $\gamma = 0.1$, 2) assuming there is no nitrate uptake on coarse dust ($\gamma = 0$) and 3) assuming an RH-dependent uptake coefficient as in equation 2 with $m = 0.018$, equivalent to considering the mineralogy of China Loess [7], [5].

### E. Preliminary results

The surface concentration of fine and coarse nitrate over Beijing obtained with MONARCH and its comparison against observations are shown in figure 1. The concentrations obtained with nitrate uptake coefficient equal to 0.1 overestimate coarse nitrate formation and underestimates the fine one, which indicates an exessive nitrate uptake by the coarse partition of dust during the mixed dust-anthropogenic pollutants event (number 3 in figure 1). This bias is improved when considering that gamma is a function of relative humidity, mainly for the third event. Omitting coarse nitrate formation causes an overestimation of fine nitrate during the same event. In all cases, the formation of fine nitrate during the first and second events is underestimated.

### F. Conclusion

We performed a literature review on the treatment by models of the nitric acid uptake on mineral dust. We implemented an uptake reaction on the coarse mode of dust in MONARCH for $HNO_3$ and evaluated simulations of a series of 3 dust events with strong coarse nitrate formation using: null, constant and humidity-dependent uptake coefficients. These results represent the starting point of future sensitivity studies that consider explicit mineralogy when simulating dust heterogeneous chemistry.

## II. Acknowledgements

This work is part of FRAGMENT, funded by the European Research Council (ERC) of the European Union's Horizon 2020 programme (grant agreement No. 773051), and NUTRIENT (CGL2017-88911-R) funded by the Ministerio de Economía y Competitividad (MINECO). We also thank the AXA research Fund. We acknowledge RES and PRACE for granting us supercomputing resources on Marenostrum4.

## References

[1] C. Pérez, K. Haustein, Z. Janjic, O. Jorba, N. Huneeus, J. M. Baldasano, T. Black, S. Basart, S. Nickovic, R. L. Miller, J. P. Perlwitz, M. Schulz, and M. Thomson, "Atmospheric dust modeling from meso to global scales with the online NMMB/BSC-Dust model-Part 1: Model description, annual simulations and evaluation," *Atmos. Chem. Phys*, vol. 11, pp. 13 001–13 027, 2011. [Online]. Available: www.atmos-chem-phys.net/11/13001/2011/

[2] J. N. Crowley, M. Ammann, R. A. Cox, R. G. Hynes, M. E. Jenkin, A. Mellouki, M. J. Rossi, J. Troe, and T. J. Wallington, "Atmospheric Chemistry and Physics Evaluated kinetic and photochemical data for atmospheric chemistry: Volume V-heterogeneous reactions on solid substrates," *Atmos. Chem. Phys*, vol. 10, pp. 9059–9223, 2010. [Online]. Available: www.atmos-chem-phys.net/10/9059/2010/

[3] S. E. Schwartz, "Mass-Transport Considerations Pertinent to Aqueous Phase Reactions of Gases in Liquid-Water Clouds," in *Chemistry of Multiphase Atmospheric Systems*, W. Jaeschke, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 415–471.

[4] T. D. Fairlie, D. J. Jacob, J. E. Dibb, B. Alexander, M. A. Avery, A. Van Donkelaar, and L. Zhang, "Impact of mineral dust on nitrate, sulfate, and ozone in transpacific Asian pollution plumes," *Atmospheric Chemistry and Physics*, vol. 10, no. 8, pp. 3999–4012, 2010.

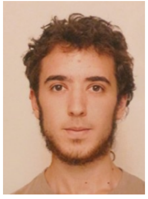
[5] C. Wei, "Modeling the effects of heterogeneous reactions on atmospheric Modeling the effects of heterogeneous reactions on atmospheric chemistry and aerosol properties chemistry and aerosol properties," 2010. [Online]. Available: https://doi.org/10.17077/etd.2xewzpnz

[6] A. Vlasenko, S. Sjogren, E. Weingartner, K. Stemmler, H. W. Gäggeler, and M. Ammann, "Atmospheric Chemistry and Physics Effect of humidity on nitric acid uptake to mineral dust aerosol particles," Tech. Rep., 2006. [Online]. Available: www.atmos-chem-phys.net/6/2147/2006/

[7] Z. Wang, X. Pan, I. Uno, J. Li, Z. Wang, X. Chen, P. Fu, T. Yang, H. Kobayashi, A. Shimizu, N. Sugimoto, and S. Yamamoto, "Significant impacts of heterogeneous reactions on the chemical composition and mixing state of dust particles: A case study during dust events over northern China," *Atmospheric Environment*, vol. 159, pp. 83–91, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.atmosenv.2017.03.044

[8] M. Spada, "DEVELOPMENT AND EVALUATION OF AN ATMOSPHERIC AEROSOL MODULE IMPLEMENTED WITHIN THE NMMB/BSC-CTM," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2015.

[9] F. Paulot, P. Ginoux, W. F. Cooke, L. J. Donner, S. Fan, M. Y. Lin, J. Mao, V. Naik, and L. W. Horowitz, "Sensitivity of nitrate aerosols to ammonia emissions and to nitrate chemistry: Implications for present and future nitrate optical depth," *Atmospheric Chemistry and Physics*, vol. 16, no. 3, pp. 1459–1477, 2016.



**Ruben Sousse** obtained his BSc in Physics at Barcelona University (UB). Afterwards, he obtained a MSc in Renewable Energies and Sustainability (UB), a Postgraduate course in Big Data and Data Science (UB) and a MSc in Environmental Physics (Bremen University). Currently he is studying for the PhD in the Atmospheric Composition group at the Barcelona Supercomputing Center.

# Modeling nitric acid uptake by mineral dust

Rubén Soussé Villa* , Oriol Jorba Casellas* , Carlos Pérez García-Pando*†

* Barcelona Supercomputing Center, Barcelona, Spain

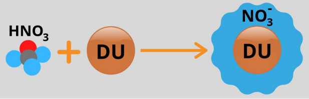
† ICREA, Catalan Institution For Research and Advanced Studies, Barcelona, Spain

E-mail: {ruben.sousse, oriol.jorba, carlos.perez}@bsc.es

## INTRODUCTION

Mineral dust is amongst the largest contributors to the global aerosol mass load and dominates climate effects over large areas of the Earth.

Dust undergoes heterogeneous chemical reactions during transport that increase its hygroscopicity, while altering its optical properties, and the associated radiative forcing. The rates of heterogeneous chemical reactions on the dust surface that form coatings of sulfate, nitrate, chloride, or organics depend strongly on the dust mineralogical composition. Dust composition also affects the partitioning of semi-volatile inorganic compounds, altering their burden and radiative forcing.



## OBJECTIVES

1. Perform an overview of the **state-of-the-art of the representation of nitric acid ($HNO_3$) uptake by mineral dust in models.**
2. **Implement the uptake of $HNO_3$** in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry model (MONARCH) [1]
3. Perform a **sensitivity study** using constant (0.1), null and humidity-dependent uptake coefficients.

This would set a benchmark to conduct future sensitivity studies for implementing explicit treatment of mineralogical composition of dust.

## THEORETICAL BACKGROUND

Heterogeneous dust chemistry implies mainly the uptake of sulfate and nitrate on dust particles, forming coatings on their surface [2]. For this study, the reaction evaluated is the nitric acid ($HNO_3$) uptake on coarse dust particles (diameter above 2.5µm), that can be expressed by the reaction [3], [4]:

$$HNO_3 + dust \rightarrow NO_3^-$$

The reaction rate of this uptake can be defined as a first-order function as [3]:

$$K = \left(\frac{r}{D_g} + \frac{4}{v\gamma}\right)^{-1} \times S$$

- *r*: aerosol bin radius
- $D_g$: gas-phase diffusion coefficient
- *v*: mean molecular speed
- *S*: aerosol specific surface area
- *γ*: the uptake coefficient

The **uptake coefficient (*γ*)**, defined as the ratio of the number of gas molecules depositing on the particle's surface over the total molecules colliding with the given surface, **is typically taken as γ = 0.1** for uptake of $HNO_3$ on mineral dust [4], [5].
However, it has been indicated that γ = 0.1 might overestimate the particulate nitrate formation [4], and some experimental studies highlight the **strong dependence of γ with relative humidity (*γ(RH)*)** [4], [5], [6], which has been shown to be similar to a Brunauer–Emmett–Teller isotherm for water adsorption on dust particles, described by the function [6]:

$$\gamma = m \times \frac{cRH}{(1-RH)(1-(1-c)RH)}$$

- *RH*: relative humidity
- *c*: water adsorption scaling factor
- *m*: specific dust mineralogy factor

Where ***c* = *8*** [6] and ***m* takes different values depending on the calcite content.** For example, a factor of *m* = *30* for Arizona Test Dust has been proposed [6], while *m* = *0.018* for China Loess from Gobi's desert with 39% $CaCO_3$ content from experimental measurements has been used [5], [7].

## METHODOLOGY

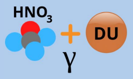
### 1 Simulations

Three pollution episodes over Beijing happening from 28th of March to 2nd of April 2015:

1) **Pure anthropogenic pollutants**
2) **Pure dust from Gobi desert composition**
3) **Dust mixed with anthropogenic pollutants composition**

Beijing Institute of Atmospheric Physics (IAP, 116.4ºE, 39.9ºN) observations of fine and coarse nitrate surface concentration are used for evaluation [7].

**Three cases for uptake coefficient values have been simulated (γ = 0.1, 0 and f(RH) with *m* = *0.018*)**:



γ=
→ **0.1** *(default value)*
→ **0** *(assuming no coarse nitrate formation)*
→ **f(RH)** *(with m = 0.018, assuming Gobi desert mineralogy [7], [5])*

### 2 Model

- **MONARCH model** [1] over a regional domain on central Asia with an horizontal resolution of 0.2 by 0.2 degrees and 24 vertical layers up to 50hPa.
- **Meteorology**: initial and boundary conditions are from the NCEP FNL analyses.
- **Emissions**: are taken from the CAMSv2.1 global inventory for anthropogenic emissions and GFASv1.2 for biomass burning emissions.
- **Chemistry**: Carbon Bond 2005 chemical mechanism is applied for the gas-phase chemistry, and the aerosol module describes the lifecycle of dust, sea-salt, black carbon, organic matter (both primary and secondary), sulfate and nitrate aerosols [8].

### 3 Nitrate mechanism

Aerosol nitrate formation in MONARCH is computed in two consecutive steps [9], [7]:

1) Fine mode (diameter < 2.5 µm): **thermodynamic equilibrium** with EQSAM model.
2) Coarse mode (diameter > 2.5 µm): **specific mass uptake** calculation for each dust size bin using the reaction rate from the *theoretical background* section.

## RESULTS

- ❖ Surface concentrations obtained with nitrate uptake coefficient equal to 0.1 overestimate coarse nitrate formation and underestimates the fine one, which indicates an excessive nitrate uptake by the coarse partition of dust during the mixed dust-anthropogenic pollutants event (number 3 in the right figure).
- ❖ This bias is improved when considering that gamma is a function of relative humidity, mainly for the third event, and it is close to observations for the coarse nitrate partition.
- ❖ Omitting coarse nitrate formation (null γ) increases the formation of fine nitrate during the event 3.
- ❖ In all cases, the formation of fine nitrate during the first and second events is underestimated compared to the observations.

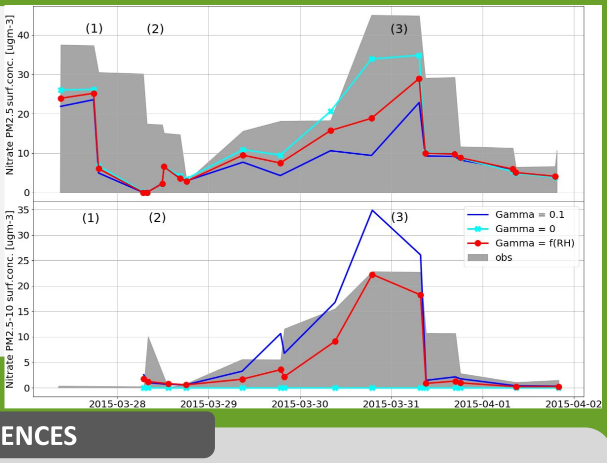


Figure: (Top) Fine nitrate partition (D < 2.5µm) over Beijing for the 3 pollution events (indicated with numbers) simulated in MONARCH with uptake coefficient γ = 0.1 (blue), γ = 0 (cyan, crossed line) and γ as function of relative humidity as described in the *theoretical background* section (red, dotted line), compared with the observations from IAP (gray) [7]. (Bottom) The same as top figure for coarse nitrate (D > 2.5µm).

## CONCLUSIONS

- A literature review on the treatment by models of the nitric acid uptake on mineral dust has been undertaken.
- We implemented an uptake reaction on the coarse mode of dust in MONARCH for $HNO_3$ and evaluated simulations of a series of 3 dust events using: null, constant and humidity-dependent uptake coefficients.
- These results represent the starting point of future sensitivity studies that consider explicit mineralogy when simulating dust heterogeneous chemistry.

## ACKNOWLEDGMENTS

This work is part of FRAGMENT, funded by the European Research Council (ERC) of the European Union's Horizon 2020 programme (grant agreement No. 773051), and NUTRIENT (CGL2017-88911-R) funded by the Ministerio de Economía y Competitividad (MINECO). We also thank the AXA research Fund. We acknowledge RES and PRACE for granting us supercomputing resources on Marenostrum4.

## REFERENCES

[1] C. Pérez, K. Haustein, Z. Janjic, O. Jorba, N. Huneeus, J. M. Baldasano, T. Black, S. Basart, S. Nickovic, R. L. Miller, J. P. Perlwitz, M. Schulz, and M. Thomson, "Atmospheric dust modeling from meso to global scales with the online NMMB/BSC-Dust model-Part 1: Model description, annual simulations and evaluation," Atmos. Chem. Phys, vol. 11, pp. 13 001–13 027, 2011. [Online]. Available: www.atmos-chem-phys.net/11/13001/2011/

[2] J. N. Crowley, M. Ammann, R. A. Cox, R. G. Hynes, M. E. Jenkin, A. Mellouki, M. J. Rossi, J. Troe, and T. J. Wallington, "Atmospheric Chemistry and Physics Evaluated kinetic and photochemical data for atmospheric chemistry: Volume V-heterogeneous reactions on solid substrates," Atmos. Chem. Phys, vol. 10, pp. 9059–9223, 2010. [Online]. Available: www.atmos-chem-phys.net/10/9059/2010/

[3] S. E. Schwartz, "Mass-Transport Considerations Pertinent to Aqueous Phase Reactions of Gases in Liquid-Water Clouds," in Chemistry of Multiphase Atmospheric Systems, W. Jaeschke, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, pp. 415–471.

[4] T. D. Fairlie, D. J. Jacob, J. E. Dibb, B. Alexander, M. A. Avery, A. Van Donkelaar, and L. Zhang, "Impact of mineral dust on nitrate, sulfate, and ozone in transpacific Asian pollution plumes," Atmospheric Chemistry and Physics, vol. 10, no. 8, pp. 3999–4012, 2010.

[5] C. Wei, "Modeling the effects of heterogeneous reactions on atmospheric Modeling the effects of heterogeneous reactions on atmospheric chemistry and aerosol properties chemistry and aerosol properties," 2010. [Online]. Available: https://doi.org/10.17077/etd.2xewzpnz

[6] A. Vlasenko, S. Sjogren, E. Weingartner, K. Stemmler, H. W. Gäggeler, and M. Ammann, "Atmospheric Chemistry and Physics Effect of humidity on nitric acid uptake to mineral dust aerosol particles," Tech. Rep., 2006. [Online]. Available: www.atmos-chem-phys.net/6/2147/2006/

[7] Z. Wang, X. Pan, I. Uno, J. Li, Z. Wang, X. Chen, P. Fu, T. Yang, H. Kobayashi, A. Shimizu, N. Sugimoto, and S. Yamamoto, "Significant impacts of heterogeneous reactions on the chemical composition and mixing state of dust particles: A case study during dust events over northern China," Atmospheric Environment, vol. 159, pp. 83–91, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.atmosenv.2017.03.044

[8] M. Spada, "DEVELOPMENT AND EVALUATION OF AN ATMOSPHERIC AEROSOL MODULE IMPLEMENTED WITHIN THE NMMB/BSC-CTM," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2015.

[9] F. Paulot, P. Ginoux, W. F. Cooke, L. J. Donner, S. Fan, M. Y. Lin, J. Mao, V. Naik, and L. W. Horowitz, "Sensitivity of nitrate aerosols to ammonia emissions and to nitrate chemistry: Implications for present and future nitrate optical depth," Atmospheric Chemistry and Physics, vol. 16, no. 3, pp. 1459–1477, 2016.

BSC

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

**Barcelona Supercomputing Center**
Jordi Girona, 31 - Torre Girona
08034 Barcelona (Spain)

education@bsc.es
www.bsc.es

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