Title: *Constraining astrophysical and cosmological models with observations of the 21-cm signal from the Epoch of Reionization.*

Advisors: B. Semelin and A. Gorce

Departments: LERMA and IAS

Abstract:
This PhD project aims at developing novel methods to constrain physical processes during the Epoch of Reionization, the period in the history of the Universe when the very first galaxies were born. The data from which the constraints on the physical processes will be inferred is the redshifted 21 cm line emitted by the neutral Hydrogen in the intergalactic medium. It is currently being observed through a Key Program on the French radio-interferometer NenuFAR and will be observed, in the next decade, with the upcoming international Square Kilometer Array. The constraints will be established combining Bayesian statistics and machine learning. Field-level inference and simulation-based inference will be explored.

Context:
From 2027 on, the Square Kilometer Array radio-interferometer (SKA) will start observing the Cosmic Dawn (CD) and the Epoch of Reionization (EoR). Following the global recombination of Hydrogen in the cooling Universe at redshift z∼1080, the Universe is cold and neutral. Only with the formation of the first stars at z∼20-30 are ionizing photons emitted again. As star clusters grow into primordial galaxies, bubbles of ionized gas form around them and expand into the intergalactic medium (IGM) until they overlap. Around z∼6, the universe is completely re-ionized. During this first billion years, patches of cold and neutral Hydrogen persist in the low-density regions of the IGM. They emit and absorb photons at a wavelength of 21 cm through a hyperfine transition in the ground state of the atoms. These photons are received on Earth in a range of radio frequencies (50 to 200 MHz) that directly map to distances from the observer to the emission point (see Figure 1). The tremendous sensitivity and collecting area of the SKA will allow us to obtain 3D maps of the IGM using this faint signal, unveiling a wealth of information about structure formation and the nature of the first sources of light. In the meantime, current radio-interferometers are attempting a first detection of the power spectrum of this signal (only the SKA wil have sufficient sensitivity to yield 3D maps). So far, we have only upper limits on the power spectrum, but are getting closer and closer to a detection (see Figure 2).
Once the data are available, their interpretation will be a challenge. Indeed, the intensity of the signal depends on several properties of the Hydrogen: ionization state, density, velocity, kinetic temperature, but also on the local Lyman-alpha flux (e.g. Furlanetto et al. 2006). Thus, the interpretation will rely on two building blocks. The first is an accurate modelling of the signal, accounting for all the relevant physical processes. The second, which is at the core of this PhD proposal, is the development of inversion methods able to put constraints on model parameters, or to directly infer underlying physical quantities (density, temperature, etc.) at the field level**. In this context, the goal of this PhD project is to explore novel inference methods** making use of an existing database of numerical simulations of the signal: the LoReLi database (Meriot & Semelin 2023). While most inference work on the 21-cm signal is done using fast but approximate semi-numerical models, LoReLI was produced using full 3D radiative transfer coupled to the dynamics. Methods relying on "amortized" inference, usually using machine learning, are required to exploit such a database. The project will focus on two such methods: field-level inference and simulation-based inference.

Field-level inference:
In a bid to beat down observational noise and to speed up computations, summary statistics are commonly used to analyze 21 cm data, such as the power spectrum. The latter only exploits the information enclosed in the variance of the data called the Gaussian information. However, 21cm observations are expected to be highly non-Gaussian during the Epoch of Reionization (e.g., Watkinson et al. 2019). The missing information cannot be recovered in the models, therefore only providing a partial picture of the reionization process.
In this context, we propose to use field-based inference: rather than converting both the model and the data into summary statistics which are then compared, we compare theory and data directly in observational space, at the field level. This approach comes with a few challenges: i) A computational challenge, as the inference relies not on sampling a dozen model parameters as with a summary statistic, but thousands of them (one per pixel) and ii) A statistical challenge as there is no signal, and therefore no information, in the ionized region of the sky (with no neutral Hydrogen). To tackle these issues, we propose to investigate the potential of Hamiltonian Monte-Carlo sampling, already applied successfully to galaxy surveys (Jasche et al., 2010), combined with informed priors,

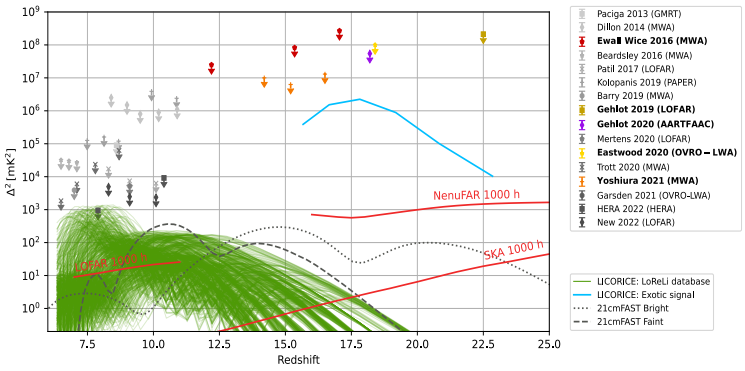

Figure 2 : Existing upper limits from radio-interferometers on the variance of the 21-cm signal's fluctuations at a fixed scale. These upper limits, at a fixed redshift, are compared with predictions from semi-numerical models in the whole redshift range and the theoretical sensitivity of different instruments if all systematics and foregrounds where perfectly removed. This figure is credited to Florent Mertens, Co-Pi of the Cosmic Dawn project on NenuFAR.

approximate marginalization (Millea & Seljak 2022), and joint analysis of complementary data sets (Zhou & Mao 2023).
With the field-based inference approach, we will be able to not only constrain models of cosmic reionization, but also reconstruct the high-redshift matter density field underlying any 21cm data, which can then be used to constrain primordial density fluctuations or to analyze lensing maps.

Simulation-Based Inference:
Whether using a Gaussian or non-Gaussian statistics for the 21cm signal we do not have a closed analytic form for their likelihood to perform Bayesian inference (e.g. Prelogovic & Mesinger 2023). One solution is to use a simple Gaussian approximation. Another is to consider the signal produced by each simulation as a realization of the likelihood of the chosen statistic given the simulation parameters, and use Machine Learning to learn this likelihood function. It is then simple (and fast) to perform the inference with usual MCMC methods. This approach opens the possibility of using any summary statistic, giving access to more information than the power spectrum alone, and, thus, obtaining tighter constraints. Defining more informative summary statistics is in itself an emerging and very open field. It can be done using novel tools such as scattering transforms (e.g. Hothi et al. 2023) or by targeting the non-Gaussian information directly in observation space (Gorce & Pritchard 2019). Depending on the pace of the progress on field-level inference, there is the possibility to explore simulation-based inference leveraging the LoReLi database as a second axis of research for the PhD project.

Research plan:
Phase 1 (~6 month): In the first few months, the PhD student will have to master the physics of the 21cm signal and the different algorithms to model the signal, and learn their way around the LoReLi database. They will also do a literature review of Bayesian inference applied to the 21cm signal, including using Machine learning techniques. They will also take in hand the HMC code developed at IAS.
Phase 2 (~18 month): The second and main phase of the PhD will be to develop a framework to apply field-level inference to the signal leveraging the LoReLi database. This will likely involve developing ML approaches to connect the density field (for example) to the signal, using LoReLi as a learning sample.
Phase 3 (~12 month): In the final part of the PhD, and depending on the results of phase 2, the student will broaden their scope to other developments in the domain of simulation-based inference. This field is growing very fast in the domain of the 21-cm signal and the advising team will target specific topics depending on the state of the art at that time. Optimal information compression and/or alternative ML inference networks are possible projects.

Positioning of the PhD project in the international context:
The proposed work will leverage the assets of the teams in LERMA and IAS – a unique database of simulations of the 21-cm signal (at LERMA), expertise in Bayesian statistics and HMC (at IAS) and Machine Learning for parameter inference (at LERMA). The PhD project is also firmly connected to the international observational context. One of the advisors is Co-Pi of the NenuFAR Cosmic Dawn project, and both have been involved for many years in the EoR-dedicated science working group of the SKA collaboration. This will ensure a firm anchoring of this PhD project in the observational context. The student will be given the opportunity to present their work at international conferences – including the SKA and NenuFAR yearly collaboration meetings.
This proposal includes a collaboration with Prof. Adrian Liu, McGill University, whose group boasts a strong expertise in Bayesian statistics and field-level inference applied to high-redshift 21cm data (Berger et al., in prep).

This collaboration will culminate in a two-month-long stay in Montreal for the student, partly funded by a Mitacs Globalink award[1], which the team will apply for in 2025.

*Doussot, Eames and **Semelin**, 2019, MNRAS, 490, 371*
*Eames, Doussot and **Semelin**, 2019, MNRAS, 489, 3655*
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rev., 433, 181
***Gorce** A. & Pritchard J. R., 2019, MNRAS, 489, 1231*
Jasche J. et al., 2010, MNRAS, 409, 355
*Mériot & **Semelin**, 2023, accepted in A&A, arXiv231002684*
Millea & Seljak, 2022, Physical Review D, 105, 10
Pregolovic & Mesinger, 2023, MNRAS, 524, 4239
***Semelin** et al., 2017, MNRAS, 472, 4508*
*Shimabukuro & **Semelin**, 2017, MNRAS, 469, 3869*
Watkinson C. A. et al., 2019, MNRAS, 482, 2653

[1] See https://www.mitacs.ca/our-programs/globalink-research-award-students-postdocs/

a mis en forme : Anglais (Canada)