SOUND.AI

CALL 1 - 2023

APPLICATION FORM – v. 02/10/2022

General information is provided at the end of this application form. For any question, see the SOUND.AI website or email soundai.sorbonne-universite.fr.

PROJECT PROPOSAL

(Please fill in the grayed areas)

Title of the Doctoral Research Project:

Quantifying ML uncertainties in searches for new physics at the LHC

Thesis supervisor and supervisory team:

Main supervisor	
Surname	Butter
First name	Anja
Doctoral school of affiliation	Planned affiliation with ED560 STEP'UP
Research unit	Title - Code - Tutelles - Name of the Unit director Research team within the unit - Title - Name of the team leader LPNHE - UMR 7585 - CNRS/Sorbonne Université/Université Paris Cité - Marco Zito ATLAS - Didier Lacour
Professional address	LPNHE, 4 Place Jussieu, Tour 22, 1er étage, 75005 Paris, France
Email	anja.butter@lpnhe.in2p3.fr
Phone	614957047
Doctoral students currently supervised by the thesis director	(specify the number of students and their year of first registration) Two PhD students affiliated to the Institute for Theoretical Physics in Heidelberg University in their first year (registered in 2022)

Summary of the doctoral research project

Precise estimation of uncertainties is a crucial asset in the search for new physics at the LHC. While neural network based simulation and analysis methods have enable a more efficient treatment of high-dimensional data, a rigorous treatment of network induced uncertainties remains elusive.

In the research project the PhD student will explore different methods to estimate network induced uncertainties. Starting from toy examples that highlight limitations of interpolation and extrapolation the student will analyse the properties of multiple methods including Bayesian Neural Networks, ensemble methods and mutual information. Once advantages and limitations of each method are understood, the student will apply them to complex high energy physics problems like jet unfolding and event simulation.

Title

Quantifying ML uncertainties in searches for new physics at the LHC

Main supervisor

Dr. Anja Butter
Staff researcher at LPNHE, Junior group leader at ITP in Heidelberg

Description of the project

In search of the fundamental building principles of the Universe, high energy physics combines first principle based predictions with data science methods in high precision studies. Pressing questions in high energy physics include the nature of dark matter, which shapes the formation of the universe, and the origin of the matter antimatter asymmetry that is a necessary requirement for our existence. In order to find explanations for these phenomena, the LHC searches for signs of new physics in proton-proton collisions which are able to achieve energies of 14 TeV of center of mass energy in a controlled laboratory system. The production of new particles like dark matter would lead to small differences of measurements with respect to precision predictions of the Standard Model, that describes all known particles and their interaction based exclusively on a fundamental Lagrangian. In order to find signs of physics beyond the Standard Model it is therefore crucial to establish high precision methods combining large statistics with optimized methods and reliable uncertainty treatments.

Since the nature of high energy physics is inherently probabilistic, every measurement has to be analysed using statistics methods that are able to take into account all associated uncertainties, statistical, systematic as well flat theory uncertainties. The rigorous treatment of these uncertainties enables us to put very strict quality criteria for measurements, leading to a minimal requirement of 5 sigma to declare a new discovery.

The rise of machine learning methods has been a crucial step in the development of new analysis and prediction methods at the LHC. The measured amount of data before trigger selection amounts to 1 Pb per second. Current analysis techniques strongly rely on so called trigger selections that reduce the initial amount by a factor 40000 to a manageable rate. Even at that point the high dimensionality and the large amount of statistics pose a big challenge to our data analysis methods. Machine learning has opened up new opportunities to address in particular the high dimensionality of the data which is often a limiting factor for the precision of traditional methods. Neural networks are being employed in many state of the art analyses for particle identification, signal vs background classification and calibration. Recent studies have demonstrated their aptitude for optimized analyses from likelihood free inference to optimal observables and the matrix element method. Finally neural networks are currently explored to improve the predictions from loop calculations via event generation to detector simulations with generative networks.

While machine learning has demonstrated its excellent performance in many areas, one crucial aspect of network based predictions and analyses in high energy physics remains elusive: the treatment of ML associated uncertainties. Due to the high requirements in precision and quantitative uncertainty treatments it is crucial to establish high control for the validity of methods and quantitative uncertainty estimates. Depending on the physics question, this can amount to quantifying where a method is close to optimal, in which regions predictions are limited by statistics or the ability to identify anomalies in the data. For instance in the case of quark vs gluon classification it is crucial to know whether a prediction of 50% arises from a jet that has properties that can occur in both quark and gluon jets or a jet whose features are different from both. Uncertainties offer the chance to address these questions.

The proposed project will explore how we can obtain reliable uncertainty estimations with ML methods in regions with limited statistics. More precisely, the goal of the thesis is to explore different methods to estimate uncertainties in machine learning and analyse their reliability, robustness and calibration in the context of unfolding and simulations. The following 3-year plan highlights the different steps of the project.

Year 1

- Exploration and implementation of ML methods for uncertainty estimates including
    - Bayesian Neural Networks
    - Ensemble Methods
    - Mutual Information
- Application to toy models and simplified jet calibration to explore

    - Behavior for interpolation
    - Behavior in low statistics regions (tails)
    - Behavior for outliers (out of distribution samples/ new physics)

Year 2

- Analysis of uncertainty methods in simulations with special emphasize on
    - Faithfulness for complex phase space features
    - Balance between regularization and precision
    - Uncertainties on interpolation with theory parameters

Year 3

- Analysis of uncertainty methods in unfolding with emphasize on
    - Robustness of uncertainty methods for noisy data
    - Precision in unfolding of high energy tails with rare events

Deliverables:

Each year presents a self contained section of the research project and is expected to yield a publication. The resulting publication record will be crucial for the career opportunities in particle physics. Simultaneously they will serve as references for applications in industry.

Risk assessment:

While year 2 and 3 build on the outcome of year 1 a potential risk would be the inability of the methods to estimate the uncertainties. Since several proof of concept studies have demonstrated the principal aptitude of the methods to capture uncertainties, the focus lies on the question how well they can quantify the uncertainties in different situations. Therefore any result from year 1 can be expected to be relevant and the exploration with simple examples builds a solid basis for the complex applications in year 2 and 3.

Environment:

The supervision of the student will be supported by the local administrative and social structures at LPNHE. This includes close exchange with the members of the ATLAS group on physics questions, in particular Bogdan Malaescu an expert in jet physics and Bertrand Laforge who has excellent expertise at the interface of ML with experimental physics. A close connection with SCAI will give a direct access to state-of-the-art developments in machine learning. Finally the Ecole Doctorale will provide a structured support during the entire PhD program.

Interdiscplinarity:

The research project combines multiple fields, most notably computer science, experimental and theoretical physics. Moreover statistical methods are needed to evaluate the results. While the core research questions prepare the student for a career in particle physics, the innovative methods in machine learning will enable a path in data science industry as an alternative career choice.

Internationality:
The field of high energy physics is highly international with large collaborations like ATLAS and CMS which represent each more 40 countries. However not only the creation of large collaborations depends on the support from multiple nations, international collaborations are a driving force for innovation, as they bring together different approaches and new ideas in the newly developing interdisciplinary field of ML in HEP. Moreover the international community allows us to reach a critical mass for the education and development of students and young researchers.
The supervisor has strong links with multiple international labs in particular the Institute for Theoretical Physics at Heidelberg University and the Lawrence Berkeley National Laboratory, which will open the possibility for international collaborations and exchange.
The supervisor is herself partially affiliated to the University of Heidelberg. As a result the student will have access to expertise in high-energy physics at the Institute for Theoretical Physics as well as contact to experts in computer science at the institute for scientific computing.
Finally the student will be encouraged to participate at the IRN Terascale, an international research network with bi-annual meetings in France and abroad, where the supervisor is strongly involved as convener.

PROPOSED INDUSTRIAL AND/OR INTERNATIONAL SECONDMENTS

It is foreseen that the student will have the opportunity to spend one to two months at the Lawrence Berkeley National Laboratory to work with Benjamin Nachman, a highly accomplished physicist with extensive experience in machine learning applications. Benjamin Nachman and Anja Butter have previously collaborated successfully leading to multiple joint publications on data augmentation, online training and unfolding.
Anja Butter and Benjamin Nachman are therefore currently applying for funding with the France Berkley Fund to enable an exchange program including secondments and joint workshops.

REFERENCES

Deep-learned Top Tagging with a Lorentz Layer, A. Butter et al, SciPost Phys. 5 (2018) 3, 028

How to GAN LHC Events, A. Butter et al, SciPost Phys. 7 (2019) 6, 075

Invertible Networks or Partons to Detector and Back Again, M. Bellagente et al, SciPost Phys. 9 (2020) 074

GANplifying event samples, A. Butter et al, SciPost Phys. 10 (2021) 6, 139

Publishing unbinned differential cross section results, M. Arratia et al, JINST 17 (2022) 01, P01024

General Information

Description of the call

- October - November: researchers from the academic and industrial spheres will propose their PhD research subjects to be selected by the program.
- December - January: Candidates can then respond to the call for applicants, with the ability to choose between 3 and 5 subjects.
- February - May: Eligibility check, Pre-selection process (based on paper applications), exchange with potential supervisors, determination of best 3 candidate/PhD subject pairs, interview of the candidates in front of the Selection Committee, and final selections. Start of the PhD October-November. Feedback is provided and redress/appeal procedures are open to fellows during each step of the process.

See https://soundai.sorbonne-universite.fr/dl/about > section *General Description*.

Eligibility criteria

- Proposed PhD subject clearly falls within one of the 3 main strategic domains and addresses one or more flagged sub-areas. Interdisciplinary subjects are highly encouraged.
- English description includes: abstract, objectives, concept and methodology (incl. risk/ feasibility assessment), state-of-the-art, expected impact, work plan and timeline including 3-i-secondments (interdisciplinary, intersectoral, and international), potential ethical issues and ways to mitigate them, possibilities for knowledge transfer, outreach activities, insights on typical career orientations.
- The commitment of the implementing/associated partners (i.e. host/grading institution, possibly co-funding, place for secondment, mentors,...) is effective.
- The supervisor CV/track record is provided and attest his/her habilitation to supervise research and that the limit for the PhD supervision in terms of number of students is not overpassed. To ensure a greater diversity of PhD subjects/supervisors, only one subject per open call is allowed from a potential supervisor.
- Before closing of Stage 1, the Management Team will check that each posted subject meets the Eligibility criteria.

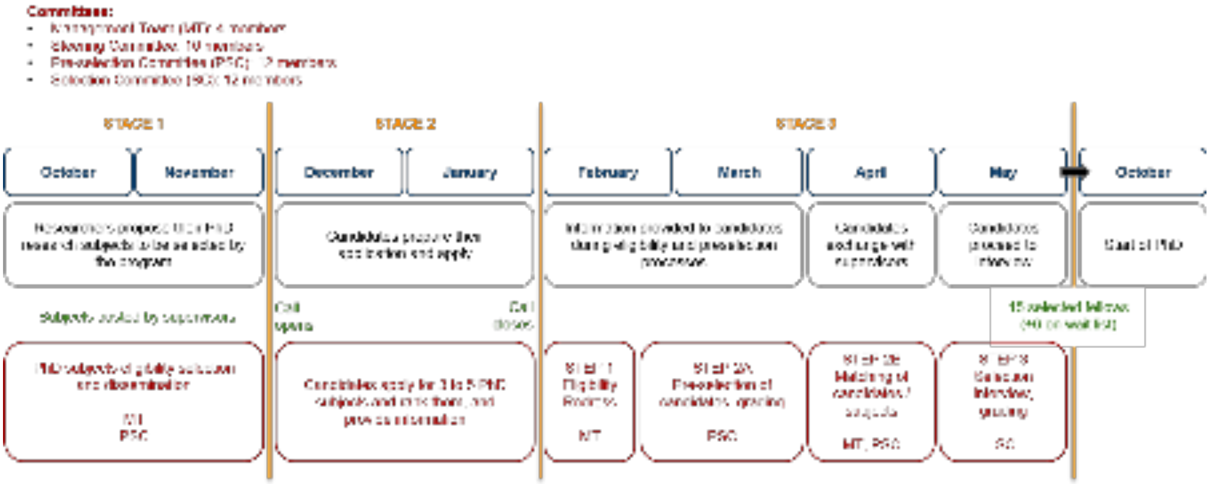
See https://soundai.sorbonne-universite.fr/dl/about > section *Scientific Network and PhD subjects*.

How to submit an application

- Registering and posting subjects are done on the SOUND.AI website

  https://supervisors.soundai.sorbonne-universite.fr
- Each supervisor and co-supervisor must register on the SOUND.AI website prior to creating a subject.
- Only the supervisor or co-supervisor who has created the subject can modify it and add/ delete co-supervisor(s)

Calendar of the call


Committees:
- Management Team (MT): 4 members
- Steering Committee: 10 members
- Pre-selection Committee (PSC): 12 members
- Selection Committee (SC): 12 members
STAGE 1
STAGE 2
STAGE 3
October
November
December
January
February
March
April
May
October
Researchers propose their PhD research subjects to be selected by the program
Candidates prepare their application and apply
Information provided to candidates during eligibility and preselection processes
Candidates exchange with supervisors
Candidates proceed to interview
Start of PhD
Subjects posted by supervisors
Call opens
Call closes
15 selected fellows (40 on wait list)
PhD subjects eligibility selection and dissemination MT PSC
Candidates apply for 3 to 5 PhD subjects and rank them, and provide information
STEP 1 Eligibility Process MT
STEP 2A Pre-selection of candidates, grading PSC
STEP 2B Matching of candidates / subjects MT, PSC
STEP 3 Selection interview, grading SC