

## Short-Term Scientific Mission Grant - APPLICATION FORM<sup>1</sup> -

Action number: CA20111

Applicant name: Orestis Melkonian

### **Details of the STSM**

Title: Machine-learned premise selection for Agda: the prover's infrastructure

Start and end date: 19/06/2023 to 21/06/2023

### **Goals of the STSM**

The goal of this collaboration is to investigate a prototype for augmenting Agda's `auto` facility with premise suggestion from the current scope, by using machine-learning on a dataset scraped from public Github repositories. This is a joint application with 'Machine-learned premise selection for Agda: the neural part' by Konstantinos Kogkalidis, which focuses on the actual neural network architecture, the training of the model, and optimizing the tool's accuracy. To compliment that, this STSM proposal focuses on the Agda infrastructure instead, namely the backend generating the training dataset and the frontend that will eventually use the trained neural model to acquire the desired suggestions at interactive proof time.

### **Working Plan**

This is a fresh collaboration between a programming language researcher (myself), a machine learning researcher (Konstantinos Kogkalidis) and a common host (Jean-Philippe Bernardy) that has already been involved within both fields.

Some initial remote collaboration has already resulted in a preliminary Agda backend (Agda's way of adding plugins to the type-checker, e.g. the same facility used to compile to Haskell and render HTML pages) that is able to extract useful training data as JSON files from the standard library, modulo some corner cases that we are happy to ignore for now:

<https://github.com/omelkonian/agda2train/>

This process involved crucial decisions like settling on a well-defined goal and standardizing the format of the input data (e.g. how scopes, types and terms will be represented).

The primary working plan for the applicant's visit is to sort out various technical details that are still unresolved, but mandatory for a first prototype tool to be usable. Having done that, we hope to start preparing a paper publication for a suitable venue that lies in the intersection of machine learning and theorem proving. Concretely, we hope to have the following contributions by the end of the visit:

<sup>1</sup> This form is part of the application for a grant to visit a host organisation located in a different country than the country of affiliation. It is submitted to the COST Action MC via-e-COST. The Grant Awarding Coordinator coordinates the evaluation on behalf of the Action MC and informs the Grant Holder of the result of the evaluation for issuing the Grant Letter.

- A stable Agda backend for generating data, i.e. releasing a first version of agda2train.
- A web scraper that successfully retrieves adequate data from Github (using agda2train).
- On the interactive Emacs side, functionality to get the same scope/context information we get in the Agda backend, thus completing the desired workflow and allowing the neural part of the proposal to progress independently.

### **Expected outputs and contribution to the Action MoU objectives and deliverables.**

In terms of **research coordination**, this STSM is closely related to the objectives of [WG5: learning-based methods for automated theorem proving and natural language proofs](#), since it involves a learning-based method for automating proofs in an *interactive* proof assistant, which also might lead to more interaction between the two distant fields of automated and interactive theorem-proving. Furthermore, the proposal nicely fits in with the objectives of [WG4: maintenance, refactoring and search of proof libraries](#), as a stronger `auto` tactic will significantly reduce library code that is currently full of tedious proofs that could be easily automated away, as well as provide search/refactoring capabilities within the large Agda corpus we make available (e.g. searching existing code with similar goal and context).

In terms of **capacity building**, this STSM will foster stronger collaboration bonds between two European groups working on programming languages in general and interactive type-driven theorem provers in particular: the Laboratory for Foundations of Computer Science (LFCS) group in the University of Edinburgh and the Logic & Types (LT) group in the University of Gothenburg. From a personal perspective, I have already had remote collaboration with people from Gothenburg and it would be great to also get a chance to meet them in person.

Moreover, language models have recently gained a lot of traction within the space of programming languages, with tools such as Copilot and ChatGPT, and we believe there are many useful tools to be extracted from this trend for theorem-proving needs. Apart from just the premise-selection tool that constitutes the obvious main deliverable, we hope this sets off a long series of ML-based tool for the Agda proof assistant that will, in no doubt, increase the productivity of its users.

Last but not least, I am confident that this STSM will actively support my career as a young researcher by making my work visible to another group with a similar theme, disseminating my results so far, and hopefully resulting in more opportunities for my research path down the road after the end of my PhD this year.