Centre for AI-Fundamentals
RAEng Google DeepMind Summer Internship Programme 2025

Project proposal

Project Title	Machine Learning for Improved Estimation of Atmospheric Composition
Lead supervisor	Simon O'Meara
Project Description	Chemical transport models currently struggle to estimate abundances of particulate matter in the atmosphere within 50 % of observed abundances. But particulate matter has significant impacts on human health and climate, and chemical transport models inform our assessments of these impacts! The framework for particulate matter simulation in these regional/global models is theoretically sound, however they are limited by a data bottleneck. The model parameters are currently dependent on outputs from observations, which can only cover a small fraction of the possible physical and chemical conditions present in the atmosphere. Massive expansion of observations to cover the necessary physico-chemical space is currently impractical. However, the model parameters can alternatively be set by very detailed computer models (box models), which can cover the physico-chemical space. What is needed is a method to convert the big dataset generated by the box models into the necessary chemical transport model parameter values. Proof-of-concept work at UoM (Simon O'Meara, Gordon McFiggans and David Topping), has shown that machine learning, with techniques such as XGBoost, can indeed inform chemical transport models from box model data. Therefore, the platform is established for an internship project that advances our proof-of-concept. It is anticipated that the internship will deliver a report on the optimal setup of machine learning for the described work. The intern will work with a (already generated) big data set of box model output and will need to identify the machine learning setup that returns optimal results (considering aspects such as mean absolute error when applied to test sets). The intern will also demonstrate the impact of varying machine learning model accuracy on regional estimates of atmospheric particulate matter through the estimated human health burden (using existing estimation techniques).

	The successful intern will gain first-hand experience of machine learning in a research project, and practise quantifying how machine learning outputs affect real-world applications. The project supervisors embed EDI and project management principles in their work. Furthermore, the supervisors have abundant experience in supervising researchers of diverse characteristics (e.g. age, background).
7 Week work plan	1. Introduction from supervisory team, background reading, setting up computer environment 2. Practise reading, running and interpreting current programme for linking dataset to chemical transport model variables through XGBoost 3. Identify and implement (code) performance metrics for machine learning 4. Identify machine learning parameters and design (code) method to search over their ranges 5. Search over ranges for machine learning parameters and practise presenting the resulting performance metrics in a draft poster and draft report 6. Present results to supervisory team to gain feedback and adjust method as necessary 7. Tidying and completing final poster and report.
Summary of project objectives	1. Improved computer programme delivered for systematic searching of machine learning parameter space and returning of performance metrics 2. Evidence generated for the optimal setup of machine learning for informing chemical transport models 3. Within the available evidence, recommendation made for the optimal setup 4. Recommendation made for future work.
Learning outcomes	1. Good practise communication within a research environment 2. Computer programming 3. Planning, delivering and reporting a research project 4. Mechanics of machine learning 5. Experience of applying machine learning techniques to a real-world problem 6. Quantitative analysis of machine learning performance 7. Experience of multi-scale computer modelling of the atmosphere.