Inductive bias in Machine Learning Models – consequences for causal effects and predictions under feature collinearity

Authors: Maximilian Pichler[1*] & Florian Hartig[1]

[1] Theoretical Ecology, University of Regensburg, Germany

*corresponding author: Maximilian.Pichler@biologie.uni-regensburg.de;

Abstract: The popularity of machine learning (ML), deep learning (DL), and artificial intelligence (AI) has grown rapidly in recent years. The main reason for preferring these models over statistical models such as logistic regression is their superior predictive performance due to their high flexibility (many parameters). To control their complexity and the bias-variance tradeoff, ML and DL algorithms rely on various types of inductive biases that limit their effective complexity. However, we do not know how the inductive bias and hyperparameters affect their explanatory power, especially under feature collinearity which is often necessary for reliable effect estimates and high accuracy in extrapolation tasks. Here we show that all these concerns are valid but can be mitigated by appropriate methodological choices. Similar to statistical models, a prerequisite for ML to learn correct effects is that feature selection must be based on causal principles, such as conditioning on confounders following Pearl's backdoor adjustment. We also show that this can increase the generalizability of the models, e.g. for forecasting under climate change. Finally, the choice of ML algorithm is crucial. We show that neural networks and boosted regression trees are better than random forest at reliably separating collinear effects. Moreover, as a byproduct, causally constrained ML models often exhibit lower generalization errors, which is relevant when trying to build models for extrapolation, as is done in climate change prediction.