Machine learning classification of microbial community compositions to predict anthropogenic pollutants in the Baltic Sea

Bacteria are ubiquitous and live in complex microbial communities, which can react rapidly to changing environmental conditions. Their physiological variety enables microbial communities to respond in specific ways to environmental drivers, potentially resulting in distinct microbial fingerprints for a given environmental state. The thesis assessed the opportunities and limitations of machine learning to detect fingerprints indicating the presence of the herbicide glyphosate in a lab microcosm experiment and the munition compound 2,4,6-trinitrotoluene (TNT) in Baltic Sea sediments. Predictions by Random Forest and Artificial Neural Network were accurate. Furthermore, the relevant taxa were identified and means of biodegradation investigated. The interpretability of machine learning models was found of particular importance for ecological data sets. The results suggest that microbial communities can predict even minor influencing factors in complex environments, demonstrating the potential of this approach for the discovery of contamination events for environmental monitoring, where also larger data sets would become available.

Bakterien sind allgegenwärtig und leben in komplexen mikrobiellen Gemeinschaften, die schnell auf veränderte Umweltbedingungen reagieren können. Die physiologische Vielfalt ermöglicht es mikrobiellen Gemeinschaften, in spezifischer Weise auf Umweltfaktoren zu reagieren, was zu unterschiedlichen mikrobiellen Fingerabdrücken für einen bestimmten Umweltzustand führen kann. Die Dissertation untersuchte die Möglichkeiten und Grenzen des maschinellen Lernens zur Erkennung von Fingerabdrücken, die auf die Anwesenheit des Herbizids Glyphosat in einem Labormikrokosmos-Experiment und dem Sprengstoff 2,4,6-Trinitrotoluol (TNT) in Ostseesedimenten hinweisen. Die Vorhersagen von Random Forest und Artificial Neural Network waren genau. Darüber hinaus wurden die relevanten Taxa identifiziert und Wege des biologischen Abbaus untersucht. Die Interpretierbarkeit von Modellen des maschinellen Lernens wurde als besonders wichtig für ökologische Datensätze erachtet. Die Ergebnisse deuten darauf hin, dass mikrobielle Gemeinschaften selbst geringfügige Reize in komplexer Umwelt vorhersagen können, was das Potenzial dieses Ansatzes für die Entdeckung von Kontaminationsereignissen für das Umweltmonitoring demonstriert, was auch größere Probenmengen verfügbar machen würde.