AGRICULTURAL RESEARCH FOUNDATION
INTERIM REPORT
FUNDING CYCLE 2019 – 2021

TITLE: Developing chemical-forensic tools for tracking nitrate sources of pollution across agricultural landscapes

RESEARCH LEADER: Gerrad Jones

COOPERATORS: Derek Godwin

EXECUTIVE SUMMARY: Nitrate is a widespread pollutant that is common in groundwater surrounding agricultural lands. Although groundwater nitrate pollution has been researched for >50 years, it remains a problem today because no good strategies exist to pinpoint the source of nitrate pollution. Nitrate has many sources including fertilizer, road runoff, deposition, leaky septic tanks, and others. Therefore, the presence of nitrate provides no indication of the potential source. As a result, implementing management strategies that reduce groundwater nitrate pollution is challenging at best. Herein, we report on a novel approach to fingerprinting sources of nitrate pollution using high-resolution mass spectrometry (HRMS) data and machine learning algorithms. In the interim report, we discussed sampling of four pollution sources as well as initial machine learning analyses. Since the interim report, we have finalized all aspects of our workflow. Within the samples we collected from four different pollution sources, we detect ~8500 chemical features. From these features, our tool is able to select the 10-100 that best predict the presence or absence of each source. To the maximum extent possible with the dataset we have, the workflow is capable of predicting source presence/absence with 100% cross validation accuracy. When we screened creek samples for each source, the workflow was able to detect the presence of wastewater treatment plant discharge in Rickreall Creek, which was confirmed based on further inspection. This work generated considerable amounts of data, and we are currently working on 4 manuscripts that will be submitted for publication in 2021.

OBJECTIVES:

1. Develop a machine learning workflow capable of identifying the chemical features that are most predictive of the presence or absence of different pollution sources.
2. Identify the chemical features associated with different pollution sources.
3. Screen creek water samples for various chemical fingerprints to determine if fingerprints can be detected in the environment.

PROCEDURES: During the second year of this project, most of our efforts focused on collecting groundwater samples with Oregon Department of Environmental Quality (ODEQ) from the Southern Willamette Groundwater Management Area (GWMA) as well as further developing our machine learning scripts. ODEQ collects quarterly samples from the GWMA. We delivered empty bottles to ODEQ personnel the day before sampling and collected samples from ODEQ on the following afternoon. Non-polar organic chemicals were extracted from all samples using

protocols established used in PI Jones' lab. These samples were analyzed at OSU's Mass Spectrometry Center. Using at high-resolution instrument (AB Sciex tTOF), we were able to quantify virtually all chemical features that hit the instruments detector. These chemical features are termed non-target features. Once we collected the data, we used support vector classification in Python to identify the non-target chemical features that were most predictive of each source. Each chemical feature was ranked based on its ability to predict the presence or absence of each source. We retained the top 10, 25, 50, and 100 most predictive chemical features, which we considered as diagnostic chemical fingerprints. With these chemical features, we were able to screen samples for a particular pollution source instead of screening for a particular chemical feature. We screened previously collected creek samples (collected in 2019) as well as the ODEQ groundwater samples (collected in 2020) for each fingerprint. Using the machine-learning model, we were able to determine the probability that a source was present or absent from each location.

SIGNIFICANT ACCOMPLISHMENTS: We made considerable advancements on our machine learning workflow. This workflow allows us to calculate the importance of each chemical feature based on its ability to predict the presence or absence of each source (Figure 1). Similar to linear regression, support vector regression utilizes coefficient weights, which are indicative of the relative importance of each variable. These weights were normalized to range from -1 to +1, with positive values increasingly predictive of source presence and negative values increasingly predictive of source absence. Upon visual inspection, of the ~8500 chemical features present across all samples, the vast majority were near 0 indicating that those features have little predictive power. A relatively small number (~500) are strongly predictive. Taking the absolute value of the normalized coefficient, the top 10-100 chemical features were isolated as chemical fingerprints (Figure 2). Based on these diagnostic chemical features, the machine-learning model could correctly predict the presence and absence of each source with 100% cross validation accuracy.

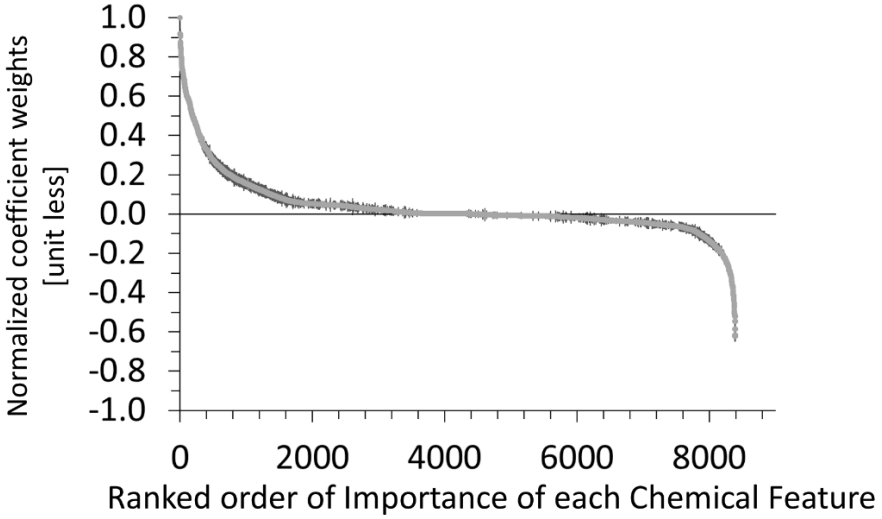

Figure 1. All chemical features were rank sorted based on their ability to predict the presence or absence of each source. This data represents the sorting of those features that are predictive of waste water treatment plants

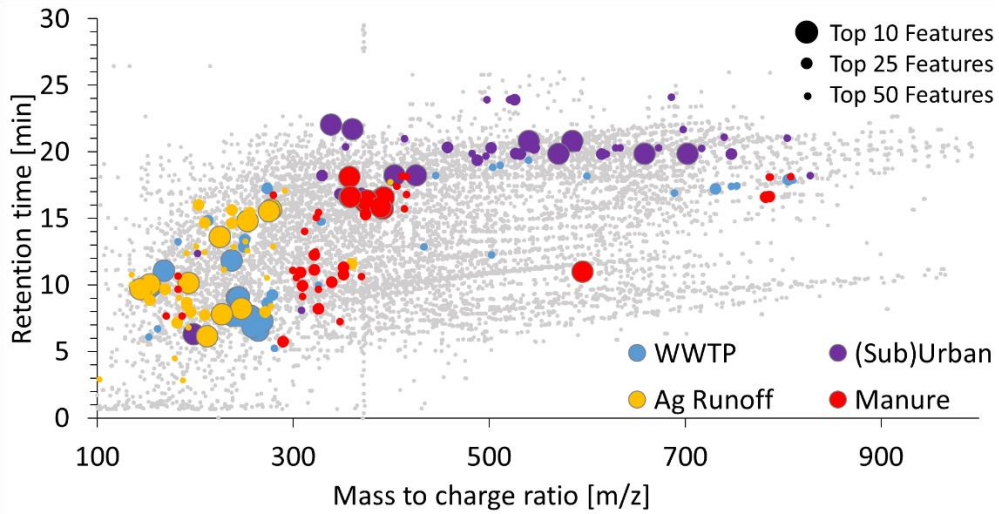

Figure 2. Diagnostic chemical fingerprints consisted of the top 10-100 features most predictive of each source, which included wastewater treatment plants (WWTP), runoff sources from parking lots and residential gutters ((sub)urban runoff), manure from dairy and beef cattle, and runoff from agricultural surfaces.

Our aim was to screen groundwater well samples for each of these fingerprints, but no fingerprint was detected in groundwater. Unfortunately, non-polar organics readily absorb to soils and are relatively immobile in the subsurface. This might explain why no sources were detected in groundwater samples. In hindsight, when screening for groundwater samples, it is more appropriate to consider polar anionic molecules, which should readily transport in the subsurface. Alternatively, we screened surface water samples for each of the fingerprints in various watersheds surrounding Corvallis. The probability that different pollution sources were present in surface waters was relatively low except for wastewater in Rickreall Creek (Figure 3).

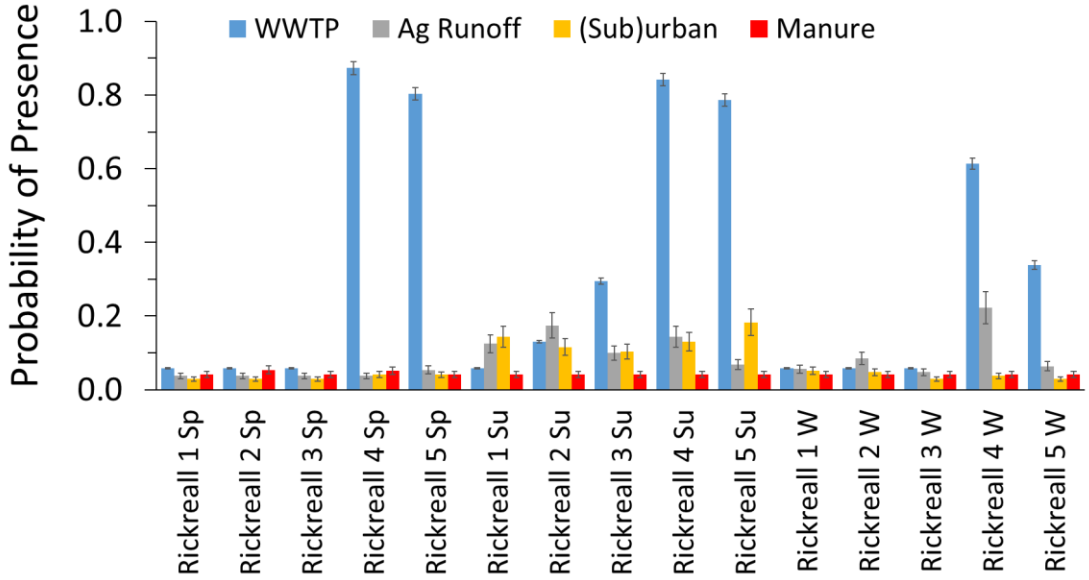

Figure 3. The probability that each pollution source was present was calculated for each source at five longitudinal sampling locations at Rickreall Creek. Samples were collected in the spring

(sp), summer (su), and winter (wi). Location 1 is the most upstream while location 5 is the most downstream.

The source with the highest presence probability in all creek samples was wastewater, most notably in Rickreall Creek. This signature was detected with the highest frequency downstream of location 3 during all three seasons (Figure 3). Upon inspection, the wastewater treatment plan for the city of Dallas is located between locations 3 and 4. This treatment facility was never sampled, yet we were still able to detect its presence suggesting that the WWTP chemical fingerprint, and perhaps the other fingerprints, are broadly applicable in other geographic areas.

BENEFITS & IMPACT: Instead of tracking individual chemicals in the environment, these fingerprints allow us to predict the probability of source contamination in a receiving body of water. This is particularly important for common pollutants, like nitrate, that have multiple pollution sources. Although we were unable to use the chemical fingerprints developed within this study due to our extraction protocol, it is relatively easy to adapt our protocols to capture chemical features that are mobile in the subsurface.

We expect our research to have a significant impact on various variety of disciplines. While the diagnostic fingerprints will be developed in western Oregon, some tested pollution sources are expected to be present worldwide (e.g., animal agriculture, treated effluent, road-water runoff, etc.). Furthermore, the optimized workflow will be immediately transferable to fields as diverse as medicine (detecting hard-to-diagnose diseases), ecosystem science (quantifying changes in ecosystem stress from climate change), microbiology (predicting the risk of inducing antibiotic resistance) and others, simply by recognizing the patterns present within non-target chemical signatures. Thousands of processes occur within ecosystems, yet we surmise that most cannot be quantified with off-the-shelf sensors. With this workflow, scientists can begin to quantify these processes, and we expect that the collective results of this workflow will provide insights into the natural world that have hitherto been unreachable.

ADDITIONAL FUNDING RECEIVED DURING PROJECT TERM: The preliminary data generated from this project was the foundation for three successful grants including US DOD/DOE/EPA Strategic Environmental Research and Development Program (SERDP; Co-PI [Jones]), NSF Chemical, Bioengineering, Environmental, and Transport Systems (CBET; Lead PI [Jones]), and USDA AFRI BNRE Water Quantity and Quality Program (Lead PI [Jones]). This data also served as the foundation for a proposal that was recently submitted to the US EPA National Priorities: Evaluation of Pollutants in Biosolids program (Co-PI [Jones]).

FUTURE FUNDING POSSIBILITIES: As mentioned above, I (Jones) believe this work has broad applicability across several disciplines. Therefore, I expect numerous outlets for this research. Currently, I am working on a proposal that I plan to submit to NSF-CBET in March and an NSF-CAREER proposal I plan to submit in July. Both of which will use preliminary data generated from this project.