The volatilome reveals microcystin concentration, microbial composition, and oxidative stress in a critical Oregon freshwater lake

Lindsay Collart
Duo Jiang
Kimberly Halsey (Kimberly.Halsey@oregonstate.edu)
Oregon State University

Research Article

Keywords: Cyanobacteria, volatile organic compounds, microcystin, harmful algal blooms, cyanotoxins, saturated fatty aldehydes, HAB monitoring

Posted Date: April 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2454193/v2

Version of Record: A version of this preprint was published at mSystems on August 17th, 2023. See the published version at https://doi.org/10.1128/msystems.00379-23.

Abstract

Toxins commonly produced by cyanobacterial blooms in freshwater lakes are a serious public health problem. The conditions leading to toxin production are currently unpredictable, thereby requiring expensive sampling and monitoring programs globally. We explored the potential of volatile organic compounds (VOCs) to indicate microcystin presence and concentration, and microbial community composition in Upper Klamath Lake, OR. Elastic net regularization regression selected 29 of 229 detected *m/z+1* values (corresponding to unique VOCs) in models predicting microcystin toxicity that outperformed or significantly improved upon regression models based on environmental parameters, including chlorophyll, pH, and temperature. Several *m/z+1* values selected by elastic net were putatively identified as saturated fatty aldehydes (SFAs), which are important in defending cyanobacteria against oxidative stress. Unique sets of *m/z+1* values were also identified by elastic net regression that predicted the relative abundance of the most dominant bacterial phyla, classes, and cyanobacterial genera. These results show that VOCs may be a key component of lake monitoring strategies.

Importance

Harmful algal blooms are among the most significant threats to drinking water safety. Blooms dominated by cyanobacteria can produce dangerous toxins and, despite intensive research, toxin production remains unpredictable. We measured gaseous molecules in Upper Klamath Lake, Oregon over two years and used them to predict the presence and concentration of the cyanotoxin, microcystin, and microbial community composition. Our approach shows potential for gaseous chemicals to be harnessed in monitoring critical waterways. Subsets of gaseous compounds were identified that are associated with microcystin production during oxidative stress, pointing to ecosystem-level interactions leading to microcystin contamination.

Introduction

Cyanobacterial harmful algal blooms (cyanoHABs) occur globally and are characterized by excessive growth of photosynthetic bacteria in freshwater lakes and rivers. CyanoHABs degrade water quality, negatively impacting potability, aquatic life, and agricultural and recreational activities. Some cyanoHABs produce toxins that pose direct threats to animal and human health[1–5] and also stunt crop development[6,7]. The chemical ecology of cyanotoxins is not well understood[8] but appears to alter the microbial community and disrupt multitrophic interactions[9]. Annual economic losses caused by cyanoHABs in the United States alone are conservatively valued at $2-4 billion[10,11], and the severity and consequences of cyanoHABs are predicted to be exacerbated by climate change[12–16]. These widely ranging impacts call for near to real-time monitoring of cyanobacteria and their toxins to protect the public and effectively manage cyanoHABs in source and recreational waters[17].

CyanoHAB monitoring programs are challenged because the specific toxins produced are strain-specific, and no morphological shifts or commonly measured environmental triggers are known to be reliably

associated with toxin production[17]. Quantifying the genes encoding the cyanotoxin microcystin in an ecosystem offers one approach to assess risk of cyanotoxin contamination, but may be of limited value because the presence of microcystin genes is not evidence of its expression[18]. Direct microcystin measurement in water and fish tissues by enzyme-linked immunosorbent assay (ELISA) and liquid chromatography tandem quadrupole mass spectrometry (LC-MS/MS)[19] is expensive, specific to a subset of microcystin congeners[20], and cannot identify the microcystin producer or its abundance[21]. Tools to leverage high resolution detection of cyanobacterial cells and their metabolites remain nascent in application, but are needed to address many environmental problems that are reaching a crisis status despite decades of intensive scientific effort[22].

We investigated the potential of the volatilome to provide high sensitivity detection of cyanoHABs and microcystin production[23–25]. The “volatilome” is the full range of low molecular weight (~30 - 272 a.m.u.) volatile organic compounds (VOCs) produced in an ecosystem[26]. VOCs have roles in cell signaling[27–30], predator-prey interactions[31], microbial carbon cycling[32,33], and atmospheric emissions that impact tropospheric ozone and climate[34]. Some VOCs inhibit growth and induce lysis in algal community members[35–37], thus regulating microbial interactions and community composition[37–39]. VOC production depends on the algal species present, their growth phase, and their environment[40,41,37,42,29,32].

Algae, including cyanobacteria, release a wide array of VOCs, including terpenes, fatty acids and their the 2-keto acid degradation products, alkanes and alcohols[38,43–45] as a result of primary and secondary metabolism[46,47] and indirectly through photochemical reactions with dissolved organic matter[48]. Cyanobacteria produce medium-chain aliphatic hydrocarbons that strengthen and add flexibility and fluidity to lipid bilayers[49,50], as well as provide tolerance to temperature and light stress[51,52]. Alkanes (C_n) and their corresponding alcohols and aldehydes are produced from C_{n+1} fatty aldehydes via aldehyde deformylating oxygenase (ADO) yielding aliphatic hydrocarbons of varying lengths[51]. The array of volatile hydrocarbons present in cyanobacteria appear to provide mechanisms for managing cell-level oxidative stress[51].

Upper Klamath Lake (UKL) is a large shallow lake in southern Oregon that is a hub of complex water use for agriculture, wildlife, fisheries, recreation, and Tribal subsistence and culture. Intensive farming and drought have decreased water quality in UKL over the last half century, contributing to annual cyanoHAB events. UKL cyanoHABs are typically dominated by *Aphanizomenon* and *Microcystis* and produce the hepatotoxin, microcystin, at elevated concentrations that prompt public warnings to avoid water contact in the mid- to late-summer[53]. We characterized the volatilome using proton transfer reaction time-of-flight mass spectrometry (PTR-MS) at lake and canal sites in UKL over two years and identified over 200 *m/z+1* values, corresponding to unique VOCs. Elastic net regularized regression selected small subsets of the *m/z+1* values that were effective predictors of microcystin contamination or microbial community composition in UKL. Microcystin prediction by elastic net models outperformed other models based only on environmental variables and in-water properties that are commonly used to detect cyanoHAB

development. Several *m/z+1* values recurring in our elastic net models appear to be associated with the fatty aldehyde ADO pathway, suggesting these cyanobacterial metabolites underlie lipid repair and reactive oxygen species (ROS) reduction during oxidative stress, which is thought to be associated with microcystin production. These fatty aldehydes in combination with other key VOCs may be ideal targets for cyanoHAB monitoring and indicate ecosystem interactions associated with microcystin production.

Results And Discussion

Upper Klamath Lake chemical and microbial composition

Water samples were collected from three sites on Upper Klamath Lake, one site on its northern arm, Agency Lake, and four canal sites during the months of May-December in 2018 and 2019 (Fig 1). The mean microcystin concentration among UKL samples with detectable toxin was 8.7 ppb (Table 1), surpassing the United States Environmental Protection Agency's recommended health advisory limit for drinking water of 0.3 ppb for pre-school aged children and 1.6 ppb for children and adults[54]. The minimum reporting, recreational, and drinking water limits for microcystin varies by state depending on water use and potential for exposure[55]. Of the 70 samples collected over 2018-2019, ten UKL samples and three canal samples were contaminated with microcystin at concentrations ≥0.3 ppb. Toxic samples mostly occurred in summer months (July – Sept), but occasionally in November, 2019, and occurred at all four lake sites (NAL, WBR, EPP, PEL) (Fig 2). The highest microcystin concentration was 469 ppb from NAL in September, 2019. Environmental parameters varied widely in UKL (Fig 2; Table 1) and toxic samples were sometimes associated with high temperature, chloride, pH, POC, PON, chlorophyll, ammonium, and conductivity; however, no significant correlations were observed with microcystin concentration and any single parameter measured at UKL (Fig S1).

Untargeted volatilomics detected 229 *m/z+1* values in samples collected at UKL and associated canals during 2018 and 2019. Seven *m/z+1* values were present in significantly discriminating amounts between samples with microcystin ≤0.3 ppb and samples with microcystin ≥0.3 ppb (Fig 3). Using these seven *m/z +1* values in a multiple linear regression model failed to predict microcystin contamination or concentration (R^2=0.08; p-value=0.89). Volatilomes clustered well by sampling date, and samples collected in 2018 mostly clustered separately from those collected in 2019 (Fig S2). Volatilomes of toxic samples did not demonstrate clear clustering (Fig S2).

The relative abundances of four phyla, Cyanobacteria, Bacteroidota, Pseudomonadota, and Actinobacteria represented 79-99% of the 16S rRNA sequences in all UKL samples during 2018-2019 (Fig 4). The class Cyanophyceae were only ~10% of the microbial community in May and peaked in September, 2019 when they were up to 75% of the community before decreasing in the autumn months. The four bloom-forming and potentially microcystin-producing cyanobacteria genera in ULK were *Aphanizomenon*, *Anabaena/Dolichospermum*, *Microcystis*, and *Gloeotrichia*. *Anabaena/Dolichospermum* sequences were always the dominant Cyanobacteria, contributing 75->99%

of the sequences in all samples. The relative abundance of *Microcystis* represented 5-25% of sequences in August through December and May but was absent in June and July.

Cell morphologies characteristic of *Aphanizomenon*, which is the dominant Cyanobacteria during the mid-summer in UKL[56], were commonly observed in UKL samples inspected by light microscopy (Fig S3). Nevertheless, few sequences were placed within *Aphanizomenon* and instead sequences often grouped with representatives of *Anabaena* sp. strain 90, *Dolichospermum circinate* strain ACBU02, and *Anabaena* sp. strain WA 102. 16S rRNA-based phylogenies are so far unable to resolve *Aphanizomenon* and *Anabaena/Dolichospermum*[57]. For example, addition of metagenomic data (with morphological validation) from 16 *Aphanizomenon*, *Anabaena*, and *Dolichospermum* strains collected in the Pacific northwest of the United States to collections of Cyanobacterial genomes from previous phylogenetic analyses still placed some strains, such as an *Anabaena* strain collected from Washington state, within *Aphanizomenon* clades[58].

Microcystin toxin prediction using the volatilome

Elastic net is a regression method that uses regularization and selects the input variables that are important for the prediction. We developed elastic net regularized regression models using the volatilome with outputs that were either linearly predictive of microcystin concentration (linear models) or predictive of microcystin concentration ≥ 0.3 ppb (logistic models) to facilitate different water management approaches (Table 2). Linear model M1 and logistic model M2 were developed using only the 229 *m/z+1* values. Linear model M7 and logistic model M8 were developed using the 229 *m/z+1* values and 'low-cost' environmental parameters (e.g., buoy data such as temperature, pH, conductivity, which are rapidly retrieved by current technologies) (Table 1). Across the four elastic net models, variable selection identified 24 of the 229 unique *m/z+1* as being important to predicting microcystin contamination (Table 3), and their relative concentrations are shown in figure 2. Nine *m/z+1* values were selected in two elastic net models, and four *m/z+1* values (151.119, 157.157, 199.189, and 203.185) were selected in three elastic net models (Table 3).

Four additional regression models based on the 'low cost' environmental parameters (M3, M4) or the full collection of environmental parameters ('low + high cost', M5, M6) were developed to compare against the skill of the VOC-based elastic net models. Similar to previous studies[59], 'low-cost' linear M3 was weakly predictive of microcystin concentration and retained only pH and chlorophyll (Table S1). POC, PON, and AMM strongly boosted the predictive power of linear M5. Neither logistic 'low-cost' M4 nor 'low + high cost' M6 were able to discriminate whether samples contained microcystin ≥0.3 ppb with greater than 50% probability (Fig 5, Table S2, Fig S4).

All of the VOC-based models outperformed the ability of 'low-cost' comparator models to predict microcystin in UKL (Fig 5). Addition of "low-cost" environmental parameters to the training data did not improve VOC-based model performance (Fig 5), and except for "month" in M8, were not retained in the final equations (Table S1, Table S2). The high Akaike Information Criterion (AIC) in logistic M2 and M8

are partly attributable to the number of selected variables and were strongly balanced by area under the receiver operating characteristic curve (AUC) values that were 0.78 and 0.88 compared to 0.50 (no better than chance) for M4 and only 0.22 for M6 (Fig 5; Fig S4).

VOCs were effective predictors of microcystin in UKL. Our ability to rapidly and inexpensively measure volatile metabolites in water samples (5 min PTR-MS measurement of raw water samples) provides a unique platform to explore relationships between the volatilome and ecosystem health and the potential for VOCs to be leveraged in cyanotoxin monitoring. Low volatility of toxins, including microcystin, makes their direct detection by PTR-ToF-MS unfeasible. Direct toxin measurement by ELISA or mass spectrometry is the current gold standard for monitoring but remains too expensive for the widespread and frequent application needed to provide timely public health advisories[19]. The metabolome is increasingly used to evaluate human health[60,61],[62] and ecosystem status, such as shifts in soil microbial ecology[63]. Similarly, the success of the volatilome to provide information about microcystin presence and concentration suggests that unique collections of VOCs in UKL are produced depending on cell physiology and community composition.

Predicting microbial community composition using the volatilome

Elastic net models were also developed using the relative abundances of the four most abundant phyla and classes, and toxin producing cyanobacteria genera as dependent variables and the 229 *m/z+1* values as independent variables. The 12 resulting models selected a total of 71 *m/z+1* values (Table S3). All twelve elastic net models performed well, yielding mean squared prediction errors (MSPE) of 0.75-1.02 and standard deviations (SD) of 0.08-0.54 (Fig S5). The *m/z+1* value 205.204 was an important predictor of the relative abundance of Cyanobacteria phylum, Cyanophyceae class, and all four Cyanobacteria genera (Fig 6). Eleven of the 18 *m/z+1* values predictive of the Cyanobacteria phylum relative abundance were also predictive of Cyanophyceae (class) relative abundance and 14 were predictive of the relative abundance of at least one of the Cyanobacteria genera. Similarly, seven of the eight *m/z+1* values predictive of Actinobacteriota relative abundance were predictive of Actinobacteria (class) relative abundance (Fig 6). Six *m/z+1* values identified in models predicting microcystin concentration were also identified in models predicting the relative abundances of Cyanobacteria genera (Table 3).

Elastic net regularized regression yielded a collection of VOC-based models that were highly effective at predicting the relative abundance of key cyanobacteria, including *Microcystis,* which is thought to be the primary source of microcystin in UKL. The success of these models is likely a consequence of seasonal changes in the microbial community composition and taxonomic and physiological differences leading to the collections of VOCs released[32,64–66]. We do not know if the VOCs identified here would also be detected in cultures of the different cyanotoxin-producing cyanobacteria. Fundamental differences in metabolism between strains in culture collections are likely to result in unique volatilomes and the absence of certain VOCs in cultures may be a consequence of in situ conditions rather than a clear VOC-strain association. Furthermore, the complex interactions between cyanobacteria and ecosystem processes leading to toxin production remain enigmatic and challenging to replicate in the laboratory.

Nevertheless, the subsets of VOCs identified using elastic net revealed *m/z+1* values that were maintained through coarse and finer-grained taxonomic groups. These results indicate systematic relationships between volatilomes and microbial taxa in UKL. We are unaware of studies that have leveraged the metabolome to describe microbial community composition; however, neural networks and linear regression approaches are being used to integrate metabolomic, metagenomic, and taxonomic data[67–70]. In our study, elastic net machine learning applied to volatilomes yielded models that were strongly predictive of ecosystem cyanotoxins and microbial community composition.

Selected *m/z+1* values in our models suggest that those compounds mediate interactions between cyanobacteria, microcystin, and the environment. For example, a sesquiterpene, *m/z+1* 203.185, was retained with positive coefficients by three models predicting microcystin and in models predicting relative abundances of Phylum Cyanobacteria, Class Cyanophyceae, and *Anabaena.* Sesquiterpene synthases are present in *Anabaena* species[71], and the recurrence of *m/z+1* 203.185 in our models is consistent with the abundance of *Anabaena* in UKL and release of sesquiterpenes and microcystin during cyanobacterial senescence[72].

β-ionone was assigned to *m/z+1* 193.153 based on its known PTR-ToF-MS target mass[73]. *m/z+1* 193.153 was retained with negative coefficients in M1 and M7 predicting microcystin and three models predicting relative abundance of non-cyanobacterial taxonomic groups. *m/z+1* 193.153 was positively correlated with phylum Cyanobacteria, class Cyanophyceae, and *Anabaena* (Fig 6). β-ionone and other norcarotenoids are products of carotenoid oxidation in various cyanobacteria during photo-oxidative stress, including *Anabaena, Aphanizomenon, and Microcystis,* and inhibit photosystem II in *Microcystis*[38,74,75,76,77]. Oxidative stress in UKL may have induced production of β-ionone in cyanobacteria[78,79], thereby decreasing *Microcystis* abundance and microcystin production. Nontoxic *Microcystis* strains employ peroxidases in response to oxidative stress, but toxic *Microcystis* strains may produce microcystin to combat mild, chronic oxidative stress[80]. The different pathways employed by cyanobacteria to tolerate oxidative stress point to β-ionone as a potentially important compound that mediates interactions within the cyanobacterial community, including microcystin production. β-ionone is also a taste-odor compound in potable freshwater sources[74] that can be rapidly identified using our approach.

m/z+1 137.129 is likely limonene with the molecular formula $(C_{10}H_{14})H^+$. Limonene is a monoterpene produced by planktonic and benthic cyanobacteria[81]. Other compounds with the same *m/z+1* value reported in PTR-MS databases include pinene and linalool, but neither of these terpenes are produced by wild-type cyanobacteria[82,83]. *m/z+1* 137.129 was retained with a negative coefficient in M7 and a positive coefficient in the model predicting relative abundance of *Aphanizomenon*. *m/z+1* 137.129 was also negatively correlated with *Microcystis* and *Gloeotrichia* (Fig 6). Limonene can inhibit photosynthesis[37,84] and lyse *M. aeruginosa*[85], suggesting that limonene produced by *Aphanizomenon* was associated with lower *Microcystis* abundance and perhaps consequently, lower microcystin concentrations.

The *m/z+1* values 157.157, 171.171, and 185.185 were selected in M1, M7, and M8 and differ by 14.014 mass units, suggesting these VOCs are products of sequential demethylation activity. A fourth, *m/z+1*, 199.189, is 14.004 mass units greater than 185.185 and was retained in M2, M7, and M8 with positive coefficients. The smallest *m/z+1* value in this series, 157.157, was retained with negative coefficients. Chemical formulas for these *m/z+1* values include C10-C13 saturated fatty aldehydes (SFAs), decanal, undecanal, dodecanal[86], and tridecanal (Table 3). Tridecanal is a key marker for Cyanophyceae[87], but is not yet present in PTR-MS chemical databases and has not been reported in PTR-MS-based research. Nevertheless, the longer chained SFAs (C12 and possibly C13) appear to be upregulated in concert with microcystin production.

The associations between SFAs and microcystin concentration in our elastic net models indicate that the relative abundances of SFAs shift during oxidative stress. SFAs accumulate between the lipid bilayers of cyanobacterial thylakoid and cytoplasmic membranes[50,88] where they contribute to membrane structure and help fine-tune localization of photosynthetic machinery[50] during temperature and light stress[49,50,89]. Cyanobacteria use an acyl-ACP reductase/aldehyde-deformylating oxygenase (ADO) pathway to produce fatty aldehydes of decreasing chain length (C_n, C_{n-1}, C_{n-2}…; Fig 7). Fatty aldehydes are substrates for aldehyde dehydrogenase (ALDH) yielding fatty acids that can be used to repair membrane lipids damaged by ROS (e.g., hydrogen peroxide, H_2O_2) produced during photosynthesis. ADO, like other diiron oxygenases, appears to be a powerful oxidizing enzyme with a wide substrate range[51]. The alkane products of NADH-dependent ADO activity on fatty aldehydes can also serve as electron donors to reduce ROS. This latter reaction is primed by generation of a Fe^{IV}-Fe^{IV} diiron center in ADO by H_2O_2 and alkane oxidation returns the diiron center to the Fe^{III}-Fe^{III} state[90]. However, alkane-dependent H_2O_2 reduction could also lead to $OH^{\bullet}$ accumulation causing a deleterious cycle of cell damage. During high light stress the fatty acid and alkane metabolites of the ADO pathway would be rapidly depleted requiring larger pools of longer chain length fatty acids to maintain ongoing lipid repair and H_2O_2 destruction. The genes encoding ADO and ALDH were upregulated in the model cyanobacterium, *Synechocystis* sp. PCC6803 during high light and oxidative stress[91,92]. As the cell's capacity to repair systems damaged by ROS becomes overwhelmed, longer chained SFAs may accumulate.

The currently prevailing hypothesis for the biological role of microcystin posits that it protects photosystems and peroxidases against oxidative damage[80,93–95]. Increased cyanobacterial dependence on the ADO pathway during oxidative stress is consistent with the mechanistic view that ROS can rapidly accumulate to algicidal concentrations even in the presence of microcystin production. Although the *m/z+1* values retained in our microcystin models and assigned here to SFAs need to be independently verified using standards or other mass spectrometry approaches (e.g., GC-MS), selection of this collection of related *m/z+1* values in multiple elastic net models predicting microcystin concentration suggests their ecological and biochemical interactions with microcystin production (Fig 7).

The use of VOCs to evaluate microcystin and microbial composition in UKL is time-efficient, and could be streamlined or even automated to inform agencies and water managers within a day, compared to the

weeks-long waiting periods sometimes required for results of microcystin and microscopic sample processing. The volatilome in water samples was collected directly by PTR-MS without need for pre-processing or sorption onto resins. Our models were designed to determine total microcystin concentration, and cannot at this time evaluate toxicity, which would require knowledge of the abundances of specific microcystin congeners[96,97]. Evaluation of the volatilome is a holistic and indirect measurement of the ecosystem. Many of the *m/z+1* values identified in our elastic net models provide valuable targets for future study of their roles in cell to ecosystem level processes.

Conclusion

The increasing frequency and severity of toxic cyanoHABs in waterways makes new, cost-effective monitoring strategies an urgent task. The ideal monitoring approach would yield information about cyanotoxin identity and concentration, cyanobacterial abundances, and ecosystem health. The VOCs produced in Upper Klamath Lake, OR provided information about the integrated growth environment and were leveraged using machine learning to determine the microcystin concentration and microbial community composition in UKL water samples. Specific VOCs, including SFAs, may be the smoking gun needed to quickly detect toxin production in freshwater lakes.

Cyanotoxins can now be detected in many waterways that were thought to be pristine (Kurt Carpenter, pers. comm.), suggesting that the combination of ongoing human activities and climate change are shifting many waterways towards ecological tipping points where HABs and cyanotoxin contamination are reliable annual events. Application of volatilomes and complex data analysis shows their potential for guidance of water treatment for taste-odor compounds in drinking water, monitoring of toxic and non-toxic cyanoHABs, and novel discovery of ecological interactions leading to toxin production in situ. An important next step is to determine whether the identified *m/z+1* values in our models emerge in samples from other lakes experiencing toxic cyanoHAB events. Because water manager actions are predicated upon sensitive and timely detection of cyanotoxins and their bacterial producers, future research that harnesses volatilomes in conjunction with other accessible complex data, including real-time buoy and satellite monitoring, to track and predict cyanoHAB trajectories before, during and after toxic HAB events is warranted to limit public exposures and economic hardship.

Methods

Water sample collection

Water samples were collected from three sites on Upper Klamath Lake and one site on its northern arm, Agency Lake, during the months of May-December in 2018 and 2019. Sampling sites were NAL, an agricultural-dominated terrain on the northeastern shore of Agency Lake, two wildlife and recreational areas near the peninsula at Eagle Point in UKL (EPP) and western shore of UKL at Howard Bay (WBR), and a residential area near the dam at the southern end of the lake (PEL) (Fig 1). Additional samples were collected from canals that drain from UKL for agricultural irrigation (NCA, ADY, MER, and LRI; Fig 1).

Samples were collected by pole from the surface about 2 meters from the shore or canal line. Samples for all analyses (VOCs, anions, pigments, particulate C and N, community composition) excepting microcystin concentration were collected in autoclaved 1 L polycarbonate bottles with limited to no headspace. Samples were transported in a cooler to Oregon State University, Corvallis, OR, and analyzed or processed by filtration and frozen within 24 h of collection for later analysis via Oregon State Universities Freshwater IIW Collaboratory. Samples for microcystin concentration were collected in autoclaved 10 ml glass vials and frozen upon arrival at OSU (-20°C) for later analysis.

Environmental Parameters

Temperature, pH, and conductivity were measured using an Extech pH/temperature meter (Nashua, NH) and YSI 30 Conductivity meter (Yellowsprings, OH), respectively. The anions bromide, fluoride, chloride, nitrite, nitrate, phosphate, and sulfate were measured with a Dionex ICS-1500 Ion Chromatograph Autosampler (Sunnyvale, CA). Data for bromide and fluoride are not shown because only one and two samples yielded data above the detection limits, respectively. Ammonium was measured by UV-Vis spectroscopy after three freeze thaw cycles[98]. Particulate carbon and nitrogen were determined from three volumes (3-110 ml) filtered onto pre-combusted GF/F filters to create a linear regression, and frozen until analysis by Exeter Analytical EA1 elemental analyzer[99] (Coventry, England). Non-particulate C and N was determined from sample filtrate and subtracted from the filtered samples. The median sample volume of filtrate was re-filtered onto a fourth GF/F filter, frozen, and analyzed with the sample filters. Chlorophyll concentration was measured in triplicate using 2-100 ml of sample filtered onto 25 mm GF/F filters (until green was observed on the filter) and extracted for 24-48 h at -20° in 90% acetone. Extract absorption was measured by UV-VIS spectrophotometer (Shimadzu, Kyoto, Japan), and calculated using the equation for cyanobacteria from Ritchie (2006)[100]. Microcystin concentrations were measured using Eurofins Abraxis Inc. Microcystins/Nodularins (ADDA) ELISA Kit (Product Number 520011).

Detection of VOCs

Triplicate 100 ml sub-samples were transferred to custom made 200 mL polycarbonate dynamic stripping chambers with sintered glass frits (2-2.5 µm) at the bases[32]. Chambers were kept in an incubator at the sample collection temperature. Samples were stripped of VOCs by flowing synthetic air through a hydrocarbon trap, then a flow controller (Sierra Instruments) set to 50 sccm, then through the glass frits into the samples. The carrier air with the stripped VOCs was directed into the PTR-MS (Ionicon, Austria) where the primary ion (H_3O^+) causes a proton transfer reaction, or soft ionization event, to VOCs having higher proton affinities than 691 kJ mol $^{-1}$, which is the proton affinity of water. VOCs in the mass range 18-363 a.m.u. were detected at their molecular masses plus 1 (*m/z+1*). Data were collected over 5 min. The conditions of the drift tube were 2.1 mbar, 80°C, and 500 V with an E/N value of 125 Td.

VOC data processing

PTR-MS raw peak data were processed using PTRwid[101]. The resulting output yielded tables giving each integrated *m/z+1* peak signal that incorporated a correction for overlapping peaks. PTRwid yields a unified mass list of all *m/z+1* values detected in all lake and canal samples. Known contaminants and internal standards were removed from the list prior to subsequent data processing (Table S1). The first 2.5 min of data were removed to account for contaminating air in the tubing and headspace of the stripping chambers. The remaining data were integrated over 2.5-5 min.. Differences in concentrations of *m/z+1* values and Wald test derived p-values were determined using the R package DESeq2, and Benjamini-Hochberg corrected p-values of *m/z+1* values were determined using the EnhancedVolcano package[102]. Chemical formulas were assigned using Ionicon PTR Viewer software and the Ionicon integrated database (PTR Viewer software version 3.3), PTR Viewer calculated values (version 3.4.2), or GLOVOCs database for PTR-MS[103] (update Nov. 16, 2020). Some chemical formulas were assigned based on published PTR-MS research on those compounds. The maximum mass shift (difference between the actual mass value and the detected mass value) allowed for compound assignment was 0.007 a.m.u. as determined by PTR-MS calculated RMSE.

VOC-based elastic net models predicting microcystin concentration

Two elastic net model types were developed (1) linear models that predict the continuous outcome of microcystin concentration (M1 and M7) and (2) logistical models that predict the dichotomous outcome of whether microcystin concentration ≥ 0.3 ppb (M2 and M8). The logistical models were trained with a binary output that designated a sample as toxic if the microcystin concentration was at or above the 0.3 ppb threshold. M1 and M2 utilized only *m/z+1* values and were trained using the glmnet function in R software (version 4.1.0) on 95% of the samples (total n=70). M7 and M8 utilized *m/z+1* values and “low cost” environmental variables (Table 3) and were trained on 95% of all samples for which VOC and “low cost” environmental data were available (total n=35). Cross validation with 10 (M7, M8) or 15 (M1, M2) fold was used to determine the value of the tuning parameters (Table S2), and hence the strength of regularization in M1, M2, M7, and M8. MSPE and their SD in M1 and M7 were calculated from the lambda value yielding the minimum mean cross-validated error. AUC and AIC for M2 and M8 were averaged from ten random iterations of each model. The *m/z+1* values in the final models were selected using the full data sets. The *m/z+1* values retained by the elastic net models were refit using the lm function in R to yield the coefficients in each final model (Table 2). A lower MSPE, lower AIC, and higher AUC are indicative of a preferred model.

‘Low cost’ and ‘Low + high cost’ regression models predicting microcystin concentration

Four base models were developed to predict microcystin concentration using only environmental parameters (Table 3). Outlier removal and bidirectional stepwise elimination was implemented using the MASS package in R to select the environmental parameters. Microcystin concentration in M3 was modeled by multiple linear regression using month of collection, collection site, and “low cost” environmental parameters. Multiple linear regression model, M5, was based on month of collection, collection site, and both “low cost” and “high cost” environmental parameters (Table 3). Two logistic

models were developed based on “low cost” and “low + high-cost” environmental parameters (M4 and M6, respectively) to predict microcystin concentration ≥ 0.3 ppb. The predictive performance of these linear regression base models was evaluated using the glmnet package with lambda and alpha values set to zero.

DNA extraction and sequencing

Samples (10-150 mL) were filtered onto 0.2 µm polycarbonate filters and stored at -20°C until DNA extraction using phenol:choloroform:iso-amyl 25:24:1. DNA quality and quantity was determined by Nanodrop 1000[104]. The V1-V2 region of the 16SrRNA gene was amplified using 27F (5’-AGAAGAGTTTGATCNTGGCTCAG-3’) and 338 RPL (5’-CWGCCWCCCGTAGGWGT -3’) primers with overhang adaptors according to the Illumina Inc. standard 16S sequencing library preparation protocol. Libraries were created using dual indices and Illumina sequencing adapters with a Nextera XT Index Kit (Illumina Inc.), then pooled in equimolar concentrations and sequenced using Illumina MiSeq (2X250 PE) in two batches: 54 samples were sequenced at the Center for Quantitative Life Sciences (Oregon State University, OR), and 16 samples were sequenced at Molecular Research DNA-RNA Laboratory (Shallowater, TX)[104].

16S rRNA gene amplicon analysis

The 27F (20b) and 338RPL (18 bp) primers were removed using CutAdapt, then DADA version 1.2 R package (version 3.6.1), using the SILVA database train version 138, quality filtered, dereplicated, merged, constructed an ASV table, removed chimeras, and taxonomically assigned the sample reads through the dada2 package[104]. Taxonomic assignment through dada2 was used for phylum and class classification. Taxonomic assignment of 16S rRNA sequences within the Cyanobacteria genera was done using Cydrasil and its maximum-likelihood phylogenetic tree constructed of 1327 Cyanobacteria reference sequences[105]. Sequences were aligned using reference alignments constructed using PaPaRa vs 2.0, and the alignments were placed using EPA-ng[105]. The placements were visualized via the Interactive Tree of Life (iTOL) vs. 6.5.4, and taxonomy was hand assigned (Figure S4 and S5)[106]. The dominant cyanobacteria were also confirmed by visual identification using light microscopy and morphological characteristics (Figure S3).

VOC-based models of microbial community composition

Relative abundances of the four most abundant microbial phyla, classes, and cyanobacterial genera (Anabaena, Aphanizomenon, Gloeotrichia, and Microcystis) were determined using the R package phyloseq[107]. A total of 12 VOC-based linear regression models were developed to predict microbial relative abundances using elastic net modeling. Models were trained using the glmnet function in R (v. 4.1.0) on 95% of the samples (total n=70). Cross validation with 15 fold was used to determine the value of the tuning parameter (Table S4) and strength of regularization. MSPE and their SD were calculated from the lambda value yielding the minimum mean cross-validated error. The *m/z+1* values in the final

models were selected using the full data set. The *m/z+1* values retained by the elastic net models were refit using the lm function in R to yield the coefficients in each final model (Table S4). The R package ComplexHeatMap was used to create figures 1,5 and S2, and the R package EnhancedVolcano was used to create figure 2.

Data availability

The authors declare that all data supporting the results of this study are available within the article, corresponding Supplementary Information (Supplementary Table 2, Supplementary Tables 4-5), and Supplementary Data that have been deposited and are publicly available on NCBI's SRA (link available upon request) and OSF (https://osf.io/3uctg/).

References

1. Lee, S. *et al.* Fresh produce and their soils accumulate cyanotoxins from irrigation water: Implications for public health and food security. *Food Research International* **102**, 234–245 (2017).

2. Lee, J., Lee, S. & Jiang, X. Cyanobacterial Toxins in Freshwater and Food: Important Sources of Exposure to Humans. *Annual Review of Food Science and Technology* **8**, 281–304 (2017).

3. Dreher, T. W. *et al.* Anabaena/Dolichospermum as the source of lethal microcystin levels responsible for a large cattle toxicosis event. *Toxicon: X* **1**, 100003 (2019).

4. Massey, I. Y. *et al.* Exposure routes and health effects of microcystins on animals and humans: A mini-review. *Toxicon* **151**, 156–162 (2018).

5. Badar, M. *et al.* EFFECTS OF MICROCYSTINS TOXINS CONTAMINATED DRINKING WATER ON HEPATIC PROBLEMS IN ANIMALS (COWS AND BUFFALOS) AND TOXINS REMOVAL CHEMICAL METHOD. *Buffalo Bulletin* 13 (2017).

6. Weralupitiya, C., Wanigatunge, R. P., Gunawardana, D., Vithanage, M. & Magana-Arachchi, D. Cyanotoxins uptake and accumulation in crops: Phytotoxicity and implications on human health. *Toxicon* **211**, 21–35 (2022).

7. Bouaïcha, N. & Corbel, S. Cyanobacterial Toxins Emerging Contaminants in Soils: A Review of Sources, Fate and Impacts on Ecosystems, Plants and Animal and Human Health. in *Soil Contamination - Current Consequences and Further Solutions* (eds. Larramendy, M. L. & Soloneski, S.) (InTech, 2016). doi:10.5772/64940.

8. Holland, A. & Kinnear, S. Interpreting the Possible Ecological Role(s) of Cyanotoxins: Compounds for Competitive Advantage and/or Physiological Aide? *Marine Drugs* **11**, 2239–2258 (2013).

9. Christoffersen, K., Lyck, S. & Winding, A. Microbial activity and bacterial community structure during degradation of microcystins. *Aquat. Microb. Ecol.* **27**, 125–136 (2002).

10. Anderson, D. M., Hoagland, P., Kaoru, Y. & White, A. W. *Estimated annual economic impacts from harmful algal blooms (HABs) in the United States*. (Woods Hole Oceanographic Institution, 2000). doi:10.1575/1912/96.

11. Watson, S. B. *et al.* Harmful Algal Blooms. in *Freshwater Algae of North America* 873–920 (Elsevier, 2015). doi:10.1016/B978-0-12-385876-4.00020-7.

12. Huisman, J. *et al.* Cyanobacterial blooms. *Nat Rev Microbiol* **16**, 471–483 (2018).

13. Paerl, H. W. & Huisman, J. CLIMATE: Blooms Like It Hot. *Science* **320**, 57–58 (2008).

14. O'Neil, J. M., Davis, T. W., Burford, M. A. & Gobler, C. J. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae* **14**, 313–334 (2012).

15. Jöhnk, K. D. *et al.* Summer heatwaves promote blooms of harmful cyanobacteria: HEATWAVES PROMOTE HARMFUL CYANOBACTERIA. *Global Change Biology* **14**, 495–512 (2008).

16. Anneville, O., Domaizon, I., Kerimoglu, O., Rimet, F. & Jacquet, S. Blue-Green Algae in a "Greenhouse Century"? New Insights from Field Data on Climate Change Impacts on Cyanobacteria Abundance. *Ecosystems* **18**, 441–458 (2015).

17. Schmidt, J., Wilhelm, S. & Boyer, G. The Fate of Microcystins in the Environment and Challenges for Monitoring. *Toxins* **6**, 3354–3387 (2014).

18. Pacheco, A., Guedes, I. & Azevedo, S. Is qPCR a Reliable Indicator of Cyanotoxin Risk in Freshwater? *Toxins* **8**, 172 (2016).

19. Massey, I. Y. *et al.* A Mini-Review on Detection Methods of Microcystins. *Toxins* **12**, 641 (2020).

20. Jones, M. R. *et al.* CyanoMetDB, a comprehensive public database of secondary metabolites from cyanobacteria. *Water Research* **196**, 117017 (2021).

21. Almuhtaram, H. *et al.* State of knowledge on early warning tools for cyanobacteria detection. *Ecological Indicators* **133**, 108442 (2021).

22. Alivisatos, A. P. *et al.* A unified initiative to harness Earth's microbiomes. *Science* **350**, 507–508 (2015).

23. Goutman, S. A. *et al.* Untargeted metabolomics yields insight into ALS disease mechanisms. *J Neurol Neurosurg Psychiatry* **91**, 1329–1338 (2020).

24. Zhang, H., Wang, J., Sun, Z., Zurada, J. M. & Pal, N. R. Feature Selection for Neural Networks Using Group Lasso Regularization. *IEEE Trans. Knowl. Data Eng.* **32**, 659–673 (2020).

25. Pozzer, A. C., Gómez, P. A. & Weiss, J. Volatile organic compounds in aquatic ecosystems – Detection, origin, significance and applications. *Science of The Total Environment* **838**, 156155 (2022).

26. Weisskopf, L., Schulz, S. & Garbeva, P. Microbial volatile organic compounds in intra-kingdom and inter-kingdom interactions. *Nat Rev Microbiol* **19**, 391–404 (2021).

27. Amavizca, E. *et al.* Enhanced performance of the microalga Chlorella sorokiniana remotely induced by the plant growth-promoting bacteria Azospirillum brasilense and Bacillus pumilus. *Sci Rep* **7**, 41310 (2017).

28. Fink, P. Ecological functions of volatile organic compounds in aquatic systems. *Marine and Freshwater Behaviour and Physiology* **40**, 155–168 (2007).

29. Xu, Q. *et al.* Volatile organic compounds released from Microcystis flos-aquae under nitrogen sources and their toxic effects on Chlorella vulgaris. *Ecotoxicology and environmental safety* **135**, 191–200 (2017).

30. Zuo, Z.-J., Zhu, Y.-R., Bai, Y.-L. & Wang, Y. Volatile communication between Chlamydomonas reinhardtii cells under salt stress. *Biochemical Systematics and Ecology* **40**, 19–24 (2012).

31. Schulz-Bohm, K. *et al.* The prey's scent – Volatile organic compound mediated interactions between soil bacteria and their protist predators. *ISME J* **11**, 817–820 (2017).

32. Halsey, K. H. *et al.* Biological cycling of volatile organic carbon by phytoplankton and bacterioplankton. *Limnol. Oceanogr.* **62**, 2650–2661 (2017).

33. Moore, E. R., Weaver, A. J., Davis, E. W., Giovannoni, S. J. & Halsey, K. H. Metabolism of key atmospheric volatile organic compounds by the marine heterotrophic bacterium Pelagibacter HTCC1062 (SAR11). *Environmental Microbiology* 11 (2021).

34. Aneja, V. P., Schlesinger, W. H. & Erisman, J. W. Effects of Agriculture upon the Air Quality and Climate: Research, Policy, and Regulations. *Environ. Sci. Technol.* **43**, 4234–4240 (2009).

35. Ozaki, K. *et al.* Lysis of cyanobacteria with volatile organic compounds. *Chemosphere* **71**, 1531–1538 (2008).

36. Ye, C. *et al.* Volatile organic compound emissions from Microcystis aeruginosa under different phosphorus sources and concentrations. *Phycological research* **66**, 15–22 (2018).

37. Zuo, Z. *et al.* Effects of phosphorus sources on volatile organic compound emissions from Microcystis flos-aquae and their toxic effects on Chlamydomonas reinhardtii. *Environmental geochemistry and health* **40**, 1283–1298 (2018).

38. Ikawa, M., Sasner, J. J. & Haney, J. F. Activity of cyanobacterial and algal odor compounds found in lake waters on green alga Chlorella pyrenoidosa growth. *Hydrobiologia* **443**, 19–22 (2001).

39. Yamashita, R. *et al.* Analytical Technique Optimization on the Detection of β-cyclocitral in Microcystis Species. *Molecules* **25**, 832 (2020).

40. Zuo, Z., Zhu, Y., Bai, Y. & Wang, Y. Acetic acid-induced programmed cell death and release of volatile organic compounds in Chlamydomonas reinhardtii. *Plant Physiology and Biochemistry* **51**, 175–184 (2012).

41. Liu, M. *et al.* Cyanobacteria blooms potentially enhance volatile organic compound (VOC) emissions from a eutrophic lake: Field and experimental evidence. *Environmental Research* **202**, 111664 (2021).

42. Zuo, Z. *et al.* Effects of nitrogen nutrients on the volatile organic compound emissions from Microcystis aeruginosa. *Ecotoxicology and environmental safety* **161**, 214–220 (2018).

43. Bižić, M. *et al.* Aquatic and terrestrial cyanobacteria produce methane. *Science Advances* **6**, eaax5343 (2020).

44. Santos, A. B., Fernandes, A. S., Wagner, R., Jacob-Lopes, E. & Zepka, L. Q. Biogeneration of volatile organic compounds produced by Phormidium autumnale in heterotrophic bioreactor. *Journal of applied phycology* **28**, 1561–1570 (2016).

45. Lindberg, P., Park, S. & Melis, A. Engineering a platform for photosynthetic isoprene production in cyanobacteria, using Synechocystis as the model organism. *Metabolic Engineering* **12**, 70–79 (2010).

46. O'Dowd, C. D. & de Leeuw, G. Marine aerosol production: a review of the current knowledge. *Phil. Trans. R. Soc. A.* **365**, 1753–1774 (2007).

47. Moore, E. R., Davie-Martin, C. L., Giovannoni, S. J. & Halsey, K. H. *Pelagibacter* metabolism of diatom-derived volatile organic compounds imposes an energetic tax on photosynthetic carbon fixation. *Environ Microbiol* **22**, 1720–1733 (2020).

48. Kieber, D. J., McDaniel, J. & Mopper, K. Photochemical source of biological substrates in sea water: implications for carbon cycling. *Nature* **341**, 637–639 (1989).

49. Vuorio, E. *et al.* Hydrocarbon Desaturation in Cyanobacterial Thylakoid Membranes Is Linked With Acclimation to Suboptimal Growth Temperatures. *Front. Microbiol.* **12**, 781864 (2021).

50. Lea-Smith, D. J. *et al.* Hydrocarbons Are Essential for Optimal Cell Size, Division, and Growth of Cyanobacteria. *Plant Physiol.* **172**, 1928–1940 (2016).

51. Qiao, Y., Wang, W. & Lu, X. High Light Induced Alka(e)ne Biodegradation for Lipid and Redox Homeostasis in Cyanobacteria. *Front. Microbiol.* **11**, 1659 (2020).

52. Berla, B. M., Saha, R., Maranas, C. D. & Pakrasi, H. B. Cyanobacterial Alkanes Modulate Photosynthetic Cyclic Electron Flow to Assist Growth under Cold Stress. *Sci Rep* **5**, 14894 (2015).

53. Environmental Public Health, O. H. A. *Cyanobacteria Blooms*. https://www.oregon.gov/oha/ph/healthyenvironments/recreation/harmfulalgaeblooms/pages/blue-greenalgaeadvisories.aspx.

54. United States Environmental Protection Agency. *EPA Drinking Water Health Advisories for Cyanotoxins*.

55. Recommended Human Health Recreational Ambient Water Quality Criteria or Swimming Advisories for Microcystins and Cylindrospermopsin. 249.

56. Carmichael, W. W., Drapeau, C. & Anderson, D. M. Harvesting of Aphanizomenon flos-aquae Ralfs ex Born. & Flah. var. flos-aquae (Cyanobacteria) from Klamath Lake for human dietary use. 12.

57. Gugger, M. Phylogenetic comparison of the cyanobacterial genera Anabaena and Aphanizomenon. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* **52**, 1867–1880 (2002).

58. Dreher, T. W., Davis, E. W. & Mueller, R. S. Complete genomes derived by directly sequencing freshwater bloom populations emphasize the significance of the genus level ADA clade within the Nostocales. *Harmful Algae* **103**, 102005 (2021).

59. Rousso, B. Z., Bertone, E., Stewart, R. & Hamilton, D. P. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Research* **182**, 115959 (2020).

60. Madsen, R., Lundstedt, T. & Trygg, J. Chemometrics in metabolomics—A review in human disease diagnosis. *Analytica Chimica Acta* **659**, 23–33 (2010).

61. Mallick, H. *et al.* Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat Commun* **10**, 3136 (2019).

62. Morrow, A. L. *et al.* Early microbial and metabolomic signatures predict later onset of necrotizing enterocolitis in preterm infants. *Microbiome* **1**, 13 (2013).

63. Wilson, R. M. *et al.* Soil metabolome response to whole-ecosystem warming at the Spruce and Peatland Responses under Changing Environments experiment. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2004192118 (2021).

64. Lawson, C. A., Possell, M., Seymour, J. R., Raina, J.-B. & Suggett, D. J. Coral endosymbionts (Symbiodiniaceae) emit species-specific volatilomes that shift when exposed to thermal stress. *Scientific Reports* **9**, (2019).

65. Meskhidze, N., Sabolis, A., Reed, R. & Kamykowski, D. Quantifying environmental stress-induced emissions of algal isoprene and monoterpenes using laboratory measurements. *Biogeosciences* **12**, 637–651 (2015).

66. Dani, K. G. S. & Loreto, F. Trade-Off Between Dimethyl Sulfide and Isoprene Emissions from Marine Phytoplankton. *Trends in Plant Science* **22**, 361–372 (2017).

67. Noecker, C. *et al.* Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation. *mSystems* **1**, e00013-15 (2016).

68. Morton, J. T. *et al.* Learning representations of microbe–metabolite interactions. *Nature Methods* **16**, 1306–1314 (2019).

69. Quinn, T. P. & Erb, I. Examining microbe–metabolite correlations by linear methods. *Nat Methods* **18**, 37–39 (2021).

70. Morton, J. T. *et al.* Reply to: Examining microbe–metabolite correlations by linear methods. *Nat Methods* **18**, 40–41 (2021).

71. Agger, S. A., Lopez-Gallego, F., Hoye, T. R. & Schmidt-Dannert, C. Identification of Sesquiterpene Synthases from *Nostoc punctiforme* PCC 73102 and *Nostoc* sp. Strain PCC 7120. *J Bacteriol* **190**, 6084–6096 (2008).

72. Merel, S. *et al.* State of knowledge and concerns on cyanobacterial blooms and cyanotoxins. *Environment International* **59**, 303–327 (2013).

73. Muñoz-González, C., Canon, F., Feron, G., Guichard, E. & Pozo-Bayón, M. Assessment Wine Aroma Persistence by Using an in Vivo PTR-ToF-MS Approach and Its Relationship with Salivary Parameters. *Molecules* **24**, 1277 (2019).

74. Zhang, K., Lin, T. F., Zhang, T., Li, C. & Gao, N. Characterization of typical taste and odor compounds formed by Microcystis aeruginosa. *Journal of Environmental Sciences* **25**, 1539–1548 (2013).

75. Shao, J. *et al.* First report on the allelopathic effect of Tychonema bourrellyi (Cyanobacteria) against Microcystis aeruginosa (Cyanobacteria). *J Appl Phycol* **25**, 1567–1573 (2013).

76. Liu, J. *et al.* Programmed cell death of Chlamydomonas reinhardtii induced by three cyanobacterial volatiles β-ionone, limonene and longifolene. *Science of The Total Environment* **762**,

144539 (2021).

77. Shao, J. *et al.* Elucidating the toxicity targets of β-ionone on photosynthetic system of Microcystis aeruginosa NIES-843 (Cyanobacteria). *Aquatic Toxicology* **104**, 48–55 (2011).

78. Walsh, K., Jones, G. J. & Dunstan, R. H. Effect of high irradiance and iron on volatile odour compounds in the cyanobacterium microcystis aeruginosa. *Phytochemistry* **49**, 1227–1239 (1998).

79. Suurnäkki, S. *et al.* Identification of geosmin and 2-methylisoborneol in cyanobacteria and molecular detection methods for the producers of these compounds. *Water Research* **68**, 56–66 (2015).

80. Hellweger, F. L. *et al.* Models predict planned phosphorus load reduction will make Lake Erie more toxic. *Science* **376**, 1001–1005 (2022).

81. Höckelmann, C. & Jüttner, F. Volatile organic compound (VOC) analysis and sources of limonene, cyclohexanone and straight chain aldehydes in axenic cultures of Calothrix and Plectonema. *Water Science and Technology* **49**, 47–54 (2004).

82. Yang, R. *et al.* Photosynthetic Conversion of CO2 Into Pinene Using Engineered Synechococcus sp. PCC 7002. *Front. Bioeng. Biotechnol.* **9**, 779437 (2021).

83. Pattanaik, B. & Lindberg, P. Terpenoids and Their Biosynthesis in Cyanobacteria. *Life* **5**, 269–293 (2015).

84. Zhao, J. *et al.* Inhibitory effects of eucalyptol and limonene on the photosynthetic abilities in *Chlorella vulgaris* (Chlorophyceae). *Phycologia* **55**, 696–702 (2016).

85. Hu, X. *et al.* Effects of limonene stress on the growth of and microcystin release by the freshwater cyanobacterium Microcystis aeruginosa FACHB-905. *Ecotoxicology and Environmental Safety* **105**, 121–127 (2014).

86. Ghanbari, J., Khajoei-Nejad, G., Erasmus, S. W. & van Ruth, S. M. Identification and characterisation of volatile fingerprints of saffron stigmas and petals using PTR-TOF-MS: Influence of nutritional treatments and corm provenance. *Industrial Crops and Products* **141**, 111803 (2019).

87. Jalliffier-Merlon, E., Marty, J.-C., Denant, V. & Saliot, A. Phytoplanktonic sources of volatile aldehydes in the river rhône estuary. *Estuarine, Coastal and Shelf Science* **32**, 463–482 (1991).

88. Coates, R. C. *et al.* Characterization of Cyanobacterial Hydrocarbon Composition and Distribution of Biosynthetic Pathways. *PLoS ONE* **9**, e85140 (2014).

89. Parveen, H. & Yazdani, S. S. Insights into cyanobacterial alkane biosynthesis. *Journal of Industrial Microbiology and Biotechnology* **49**, kuab075 (2022).

90. Xue, G., Fiedler, A. T., Martinho, M., Münck, E. & Que, L. Insights into the P-to-Q conversion in the catalytic cycle of methane monooxygenase from a synthetic model system. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20615–20620 (2008).

91. Mitschke, J., Vioque, A., Haas, F., Hess, W. R. & Muro-Pastor, A. M. Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20130–20135 (2011).

92. Hernández-Prieto, M. A., Semeniuk, T. A., Giner-Lamia, J. & Futschik, M. E. The Transcriptional Landscape of the Photosynthetic Model Cyanobacterium Synechocystis sp. PCC6803. *Sci Rep* **6**, 22168 (2016).

93. Dick, G. J. *et al.* The genetic and ecophysiological diversity of *Microcystis*. *Environmental Microbiology* **23**, 7278–7313 (2021).

94. Zilliges, Y. *et al.* The Cyanobacterial Hepatotoxin Microcystin Binds to Proteins and Increases the Fitness of Microcystis under Oxidative Stress Conditions. *PLoS ONE* **6**, e17615 (2011).

95. Dziallas, C. & Grossart, H.-P. Increasing Oxygen Radicals and Water Temperature Select for Toxic Microcystis sp. *PLoS ONE* **6**, e25569 (2011).

96. Díez-Quijada, L., Prieto, A. I., Guzmán-Guillén, R., Jos, A. & Cameán, A. M. Occurrence and toxicity of microcystin congeners other than MC-LR and MC-RR: A review. *Food and Chemical Toxicology* **125**, 106–132 (2019).

97. Van de Waal, D. B. *et al.* The ecological stoichiometry of toxins produced by harmful cyanobacteria: an experimental test of the carbon-nutrient balance hypothesis: Ecological stoichiometry of toxin production. *Ecology Letters* **12**, 1326–1335 (2009).

98. Sparks, D. L., Page, A. L., Helmke, P. A. & Loeppert, R. H. *Methods of Soil Anaylsis, Part 3: Chemical Methods*.

99. Halsey, K. H., Milligan, A. J. & Behrenfeld, M. J. Physiological optimization underlies growth rate-independent chlorophyll-specific gross and net primary production. *Photosynth Res* **103**, 125–137 (2010).

100. Ritchie, R. J. Consistent Sets of Spectrophotometric Chlorophyll Equations for Acetone, Methanol and Ethanol Solvents. *Photosynthesis Research* **89**, 27–41 (2006).

101. Holzinger, R. PTRwid: A new widget tool for processing PTR-TOF-MS data. *Atmospheric Measurement Techniques* **8**, 3903–3922 (2015).

102. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

103. Yáñez-Serrano, A. M. *et al.* GLOVOCS - Master compound assignment guide for proton transfer reaction mass spectrometry users. *Atmospheric Environment* **244**, 117929 (2021).

104. Bolaños, L. M. *et al.* Small phytoplankton dominate western North Atlantic biomass. *ISME J* **14**, 1663–1674 (2020).

105. Roush, D., Giraldo-Silva, A. & Garcia-Pichel, F. Cydrasil 3, a curated 16S rRNA gene reference package and web app for cyanobacterial phylogenetic placement. *Sci Data* **8**, 230 (2021).

106. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).

107. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).

Declarations

Acknowledgements

We are grateful for funding support from the College of Science SciRIS program and Agricultural Research Foundation at Oregon State University. K.H. is especially thankful for funds from an anonymous donor that established the Excellence in Microbiology Faculty Scholar award and supported this research. We thank Dr. Theo Dreher for assistance with sample collection, and Dr. Benjamin Philmus for his insight with mass spectrometry.

Contributions

L.C., D.J. and K.H. designed the study. L.C. collected the data and developed the elastic net models. K.H. contributed overall project oversight. D.J. supervised model development. L.C. and K.H. wrote the manuscript with input from D.J.

Tables

Table 1. Environmental parameters collected at Upper Klamath Lake, OR. Microcystin concentration is in bold in the top row. ‘Low cost’ parameters are shaded grey and ‘high cost’ parameters are white.

Environmental parameter	Abbreviation	Minimum	Maximum	Mean	Standard deviation
Microcystin (ppb)	TOX	0	469.51	8.72	56.03
Chlorophyll (µg/mL)	CH	0	6.8	0.24	0.94
Temperature (°C)	TEMP	-0.1	26.87	17.62	6.65
pH	PH	7.22	10.22	8.59	0.94
Conductivity (S/m)	COND	61.4	133.6	102.9	12.6
Particulate organic carbon (µg/mL)	POC	0.45	432.32	18.38	65.53
Particulate organic nitrogen (µg/mL)	PON	0.066	100.20	3.96	15.12
Chloride (ppm)	CHL	2.392	50	4.39	6.94
Sulfate (ppm)	None	1.70	50	4.46	6.97
Nitrate (ppm)	None	0.2	15.54	0.73	2.26
Phosphate (ppm)	None	0.11	50	1.37	7.35
Ammonium (ppm)	AMM	0.01	5.43	0.41	0.82

Table 2. Models developed for prediction of microcystin contamination

Model number	Model type	Input variables	Output type
M1	Linear elastic net	VOCs	Continuous
M2	Logistic elastic net	VOCs	Binary
M3	Linear regression	Low cost environmental parameters	Continuous
M4	Logistic regression	Low cost environmental parameters	Binary
M5	Linear regression	Low + high cost environmental parameters	Continuous
M6	Logistic regression	Low + high cost environmental parameters	Binary
M7	Linear elastic net	VOCs + low cost environmental parameters	Continuous
M8	Logistic elastic net	VOCs + low cost environmental parameters	Binary

Table 3. *m/z+1* values identified in models predicting microcystin contamination. A ‘C+’ indicates the *m/z+1* value was retained in the model with a positive coefficient and ‘C-‘indicates a negative coefficient; *m/z+1* values selected in two models are shaded light orange, and *m/z+1* values selected in three models are shaded dark

orange. *m/z+1* values shaded grey were also important in predicting bacterial relative abundance (Fig 6). Shift is the difference between the chemical's actual mass and detected mass. Chemical identifications were made using the Ionicon PTR viewer integrated database ('a' superscript), the PTR viewer calculated formulas ('b' superscript), GLOVOC database ('c' superscript), previously published PTR-MS research ('d' superscript), or relationships to other identified *m/z+1* values ('e').

m/z+1 value	Peak variance	Chemical	Shift [m/z]	M1	M2	M7	M8
35.042	0.009346	$(CH_4O)H^+$ [a]	0.004				x
80.045	0.02135	$(C_5H_5N)H^+$ [ac]	-0.004	C+		C-	
		$(C_3H_8S)H^+$ [a]	0.004				
		$(C_6H_5)H^+$ [a]	-0.008				
		$(C_4H_3N_2)H^+$ [b]	0.008				
83.055	0.022152	$(C_4H_6N_2)H^+$ [ac]	-0.005		C+		
		$(C_4H_4N_2)H^+$ [a]	0.004				
		$(C_5H_6O)H^+$ [ac]	0.006				
		$(C_3H_4N_3)H^+$ [b]	0.007				
83.078	0.022158	$(C_6H_8)H^+$ [a]	0.001		C+		
		$(C_6H_{10})H^+$ [ac]	-0.008				
		$(C_5H_8N)H^+$ [b]	0.005				
98.040	0.026149	$(C_5H_4FN)H^+$ [c]	0.000			C+	
		$(C_4H_4N_2O)H^+$ [a]	-0.003				
		$(C_3H_3N_3O)H^+$ [ac]	0.005				
		$(C_5H_5O_2)H^+$ [b]	0.003				
		$(C_4H_5N_2O)H^+$ [b]	-0.008				
103.070	0.027491	$(C_5H_{10}O_2)H^+$ [ac]	-0.005		C+		
		$(C_5H_8O_2)H^+$ [a]	0.006				
		$(C_3H_8N_3O)H^+$ [b]	-0.005				
		$(C_4H_8NO_2)H^+$ [b]	0.007				
111.102	0.029633	$(C_7H_{11}N)H^+$ [a]	0.002			C-	C+
		$(C_8H_{12})H^+$ [a]	-0.006				
		$(C_7H_{12}N)H^+$ [b]	-0.003				
137.129	0.036575	$(C_{10}H_{16})H^+$ [acd]	-0.003			C-	
		$(C_{10}H_{14})H^+$ [a]	0.005				
138.131	0.036842	$(C_9H_{15}N)H^+$ [ac]	0.003			C+	
		$(C_{10}H_{16})H^+$ [a]	-0.005				
148.073	0.039494	$(C_7H_{12}ClN)H^+$ [a]	0.003	C+		C-	
		$(C_9H_9NO)H^+$ [c]	-0.003				
		$(C_9H_7NO)H^+$ [a]	0.006				
		$(C_5H_{10}N_2O_3)H^+$ [a]	-0.007				
		$(C_6H_{11}O_4)H^+$ [b]	-0.001				
		$(C_7H_9N_4)H^+$ [b]	-0.002				
		$(C_2H_7N_6O_2)H^+$ [b]	0.002				
		$(C_6H_{12}N_2Cl)H^+$ [b]	-0.004				
		$(C_5H_{12}N_2OP)H^+$ [b]	-0.004				
		$(C_4H_{11}N_4S)H^+$ [b]	-0.005				
		$(C_5H_{11}N_2OS)H^+$ [b]	0.006				
		$(C_7H_{12}OCl)H^+$ [b]	0.008				
149.117	0.039772	$(C_7H_{16}O_3)H^+$ [c]	0.000	C-		C+	
		$(C_6H_{14}N_2O_2)H^+$ [a]	0.000				
		$(C_8H_{18}S)H^+$ [a]	0.001				
		$(C_{10}H_{13}N)H^+$ [a]	0.002				
			0.001				

		$(C_5H_{14}N_3O_2)H^+$ [b] $(C_{10}H_{14}N)H^+$ [b]	-0.003				
151.119	0.040306	$(C_{10}H_{14}O)H^+$ [c] $(C_9H_{14}N_2)H^+$ [c] $(C_9H_{12}N_2)H^+$ [a] $(C_8H_{12}N_3)H^+$ [b]	0.007 -0.004 0.005 0.008	C+	C-	C+	
153.095	0.040833	$(C_9H_{12}O_2)H^+$ [c] $(C_4H_{14}N_3OP)H^+$ [a] $(C_7H_{10}N_3O)H^+$ [b] $(C_6H_{10}N_5)H^+$ [b]	0.004 -0.003 0.005 -0.006			C-	
157.157	0.041917	$(C_{10}H_{20}O)H^+$ [c] $(C_9H_{18}N_2)H^+$ [a] $(C_8H_{18}N_3)H^+$ [b]	-0.002 -0.004 -0.001	C-		C-	C-
169.113	0.045106	$(C_{10}H_{16}S)H^+$ [ac] $(C_9H_{16}NO_2)H^+$ [b] $(C_5H_{10}N_7)H^+$ [b]	0.008 0.003 0.005			C-	
171.171	0.045655	$(C_{11}H_{22}O)H^+$ [c] $(C_{10}H_{20}N_2)H^+$ [a] $(C_9H_{20}N_3)H^+$ [b]	-0.003 -0.006 -0.003	C+		C-	
175.159	0.046718	$(C_{10}H_{22}S)H^+$ [c] $(C_9H_{20}NO_2)H^+$ [b] $(C_5H_{16}N_7)H^+$ [b]	0.008 0.002 0.004	C-		C-	
185.185	0.049392	$(C_{12}H_{24}O)H^+$[d] $(C_{10}H_{22}N_3)H^+$ [b] $(C_{11}H_{22}NO)H^+$ [b]	-0.004 -0.004 0.007	C-		C+	
189.151	0.05045	$(C_{14}H_{18})H^+$ [a] $(C_{13}H_{18}N)H^+$ [b] $(C_8H_{18}N_3O_2)H^+$ [b] $(C_{10}H_{22}NS)H^+$ [b] $(C_4H_{14}N_9)H^+$ [b] $(C_7H_{18}N_5O)H^+$ [b]	-0.004 -0.001 0.003 -0.004 0.006 -0.008			C-	
193.153	0.051517	$(C_{11}H_{17}N_3)H^+$ [a] $(C_{13}H_{18}O)H^+$ [a] $(C_6H_{18}N_5O_2)H^+$ [b] $(C_{11}H_{18}N_3)H^+$ [b] $(C_{12}H_{18}N_O)H^+$ [b] $(C_6H_{19}N_5P)H^+$ [b]	0.000 0.003 -0.001 -0.005 0.006 0.007	C-		C-	
199.189	0.053128	$(C_{13}H_{26}O)H^+$[e] $(C_{12}H_{24}NO)H^+$ [b]	N/A -0.005		C+	C+	C+
203.185	0.054193	$(C_7H_{20}N_7)H^+$ [b] $(C_{11}H_{24}NO_2)H^+$ [b] $(C_{11}H_{25}NP)H^+$ [b] $((C_{15}H_{22})H^+$ [c]	-0.001 -0.004 0.005 0.006	C+	C+	C+	
233.959	0.062401	Many compounds			C+		

332.868	0.088782	Many compounds		C+		C+

Figures

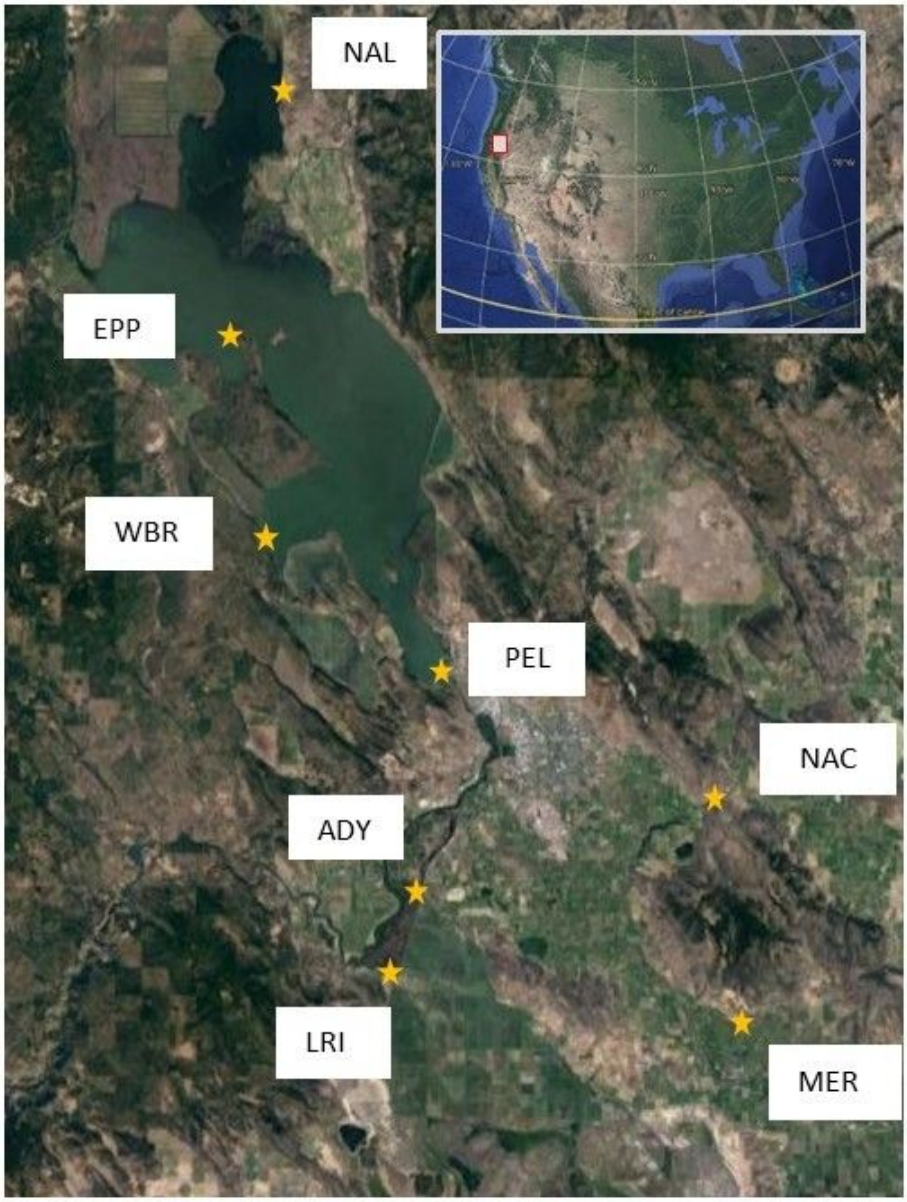

Figure 1

Sample sites on Upper Klamath and Agency Lakes, OR. Lake sites and their geospatial positions were NAL (42.559839-121.929579), WBR (42.314529-121.942224), EPP (42.430715 -121.962764), and PEL (42.2390 -121.8097). Canal sites and their geospatial positions were NCA (42.1222 -121.8289), ADY (42.0808 -121.8456), MER (42.0536 -121.6006), and LRI (42.1733 -121.6175).

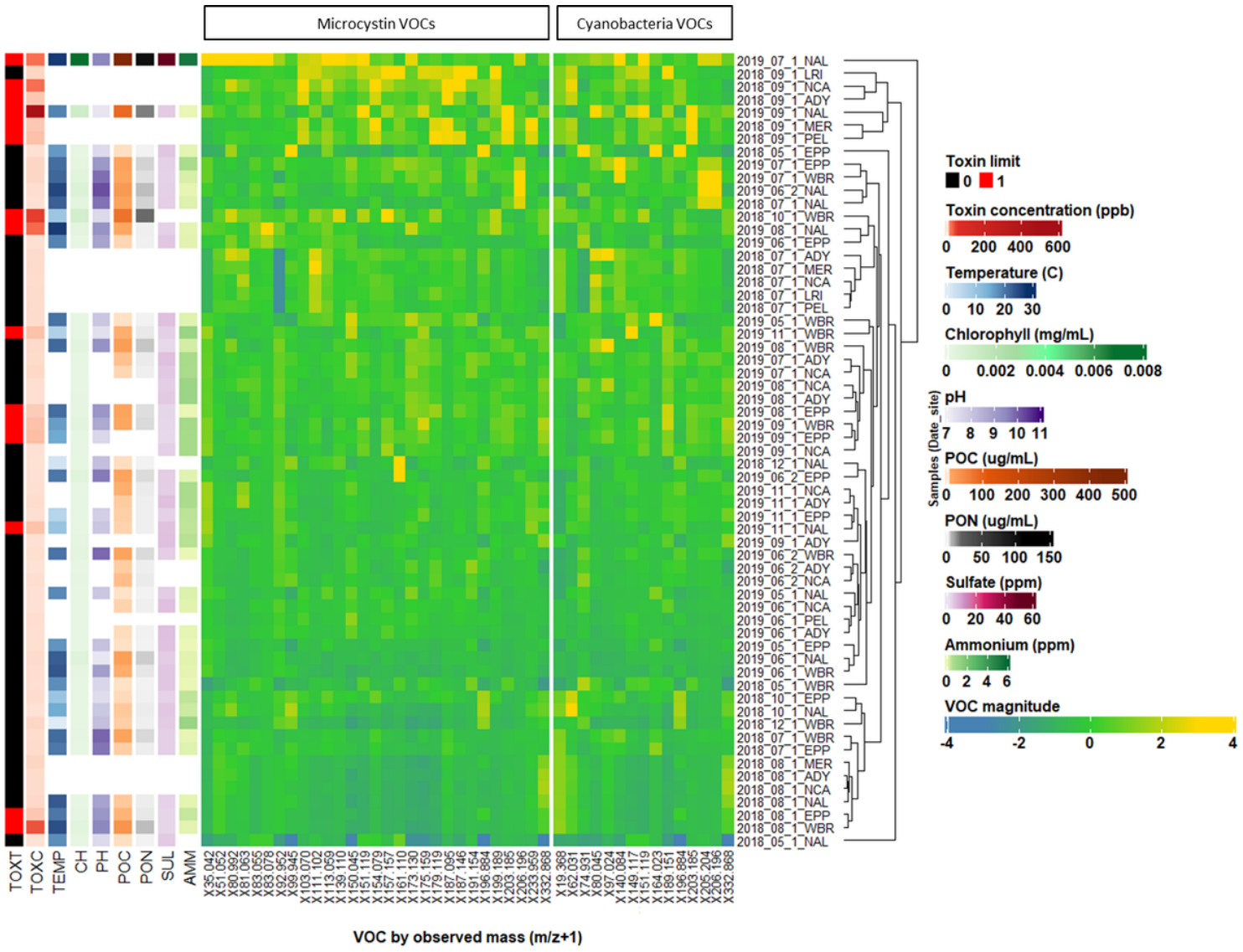

Figure 2

Unsupervised hierarchical clustering of *m/z +1* values selected by elastic net in M1, M2, M7, and M8 (left panel) and in models predicting cyanobacterial genera relative abundances (right panel) in lake and canal samples. Samples are shown in rows and labeled with site and date. The heatmap shows the Pearson correlation between each *m/z +1* value and microcystin concentration or relative abundances of cyanobacteria genera, with yellow being most positively correlated and dark blue being most negatively correlated (legend: VOC magnitude). To the left of the heatmap are environmental parameters identified by elastic net or stepwise linear regression or logistical models associated with each sample: TOXT, microcystin concentration ≥ 0.3 ppb (red) or < 0.3 ppb (black); TOXC, microcystin toxin concentration (ppb); TEMP, temperature (°C); CH, chlorophyll concentration (mg/mL); pH; POC, particulate organic

carbon (ug/mL); PON, particulate organic nitrogen (ug/ml); AMM, ammonium (ppm); and SUL, sulfate (ppm).

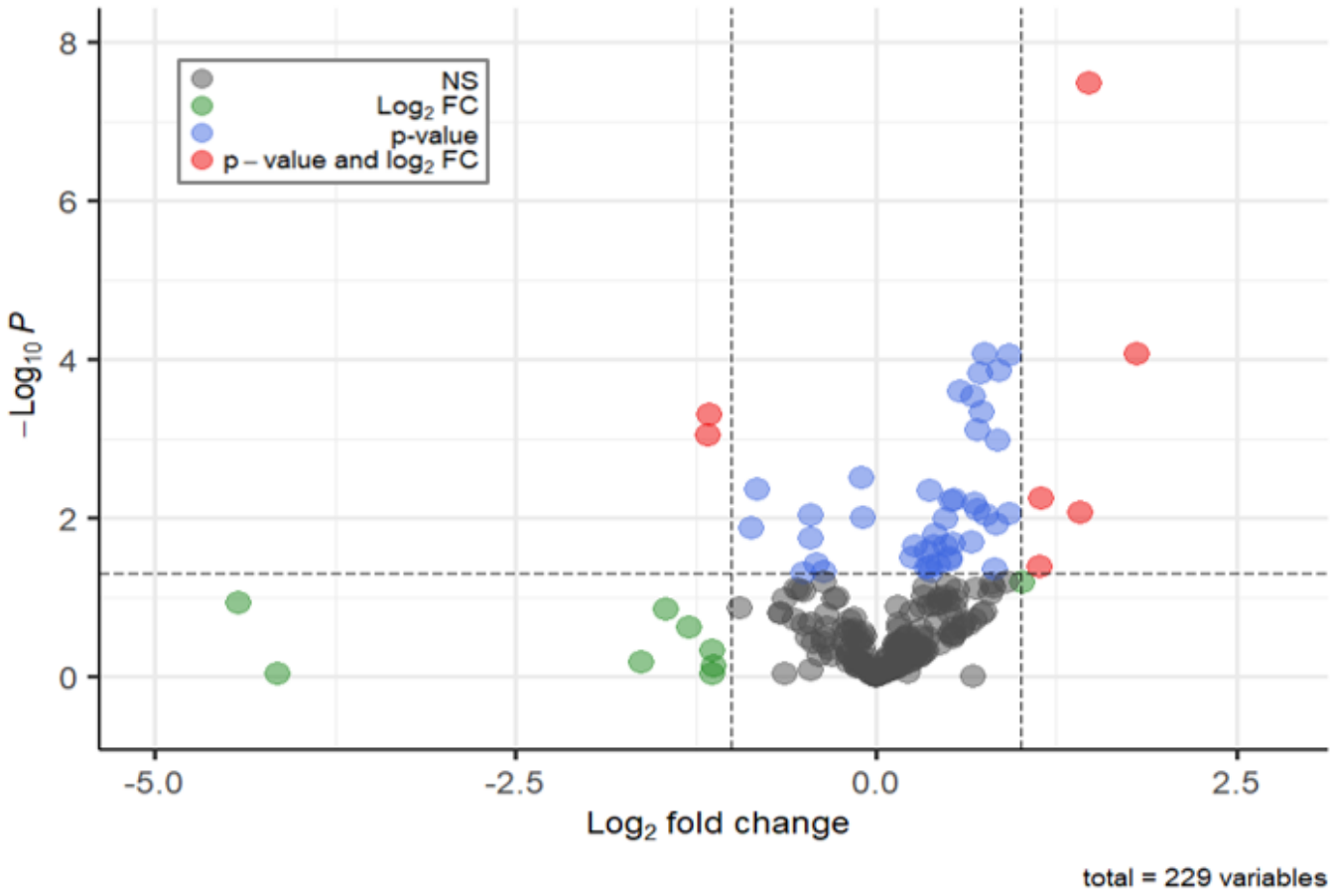

Figure 3

Volcano plot showing $\log_2$ fold-changes and p-values (significance determined via Wald test) for the 229 *m/z +1* values in toxic (≥ 0.3 ppb) vs. non-toxic samples. Multiple test correction using the Benjamin-Hochberg False Discovery Rate was applied to the p-value for each *m/z+1* value. The points are colored according to $\log_2$ fold changes and degree of significance. Points with positive fold changes greater than the dashed vertical line at +1.0 are *m/z+1* values enriched in toxic vs. non-toxic samples. Points with negative fold changes less than the dashed vertical line at -1.0 are *m/z+1* values depleted in toxic vs. non-toxic samples.

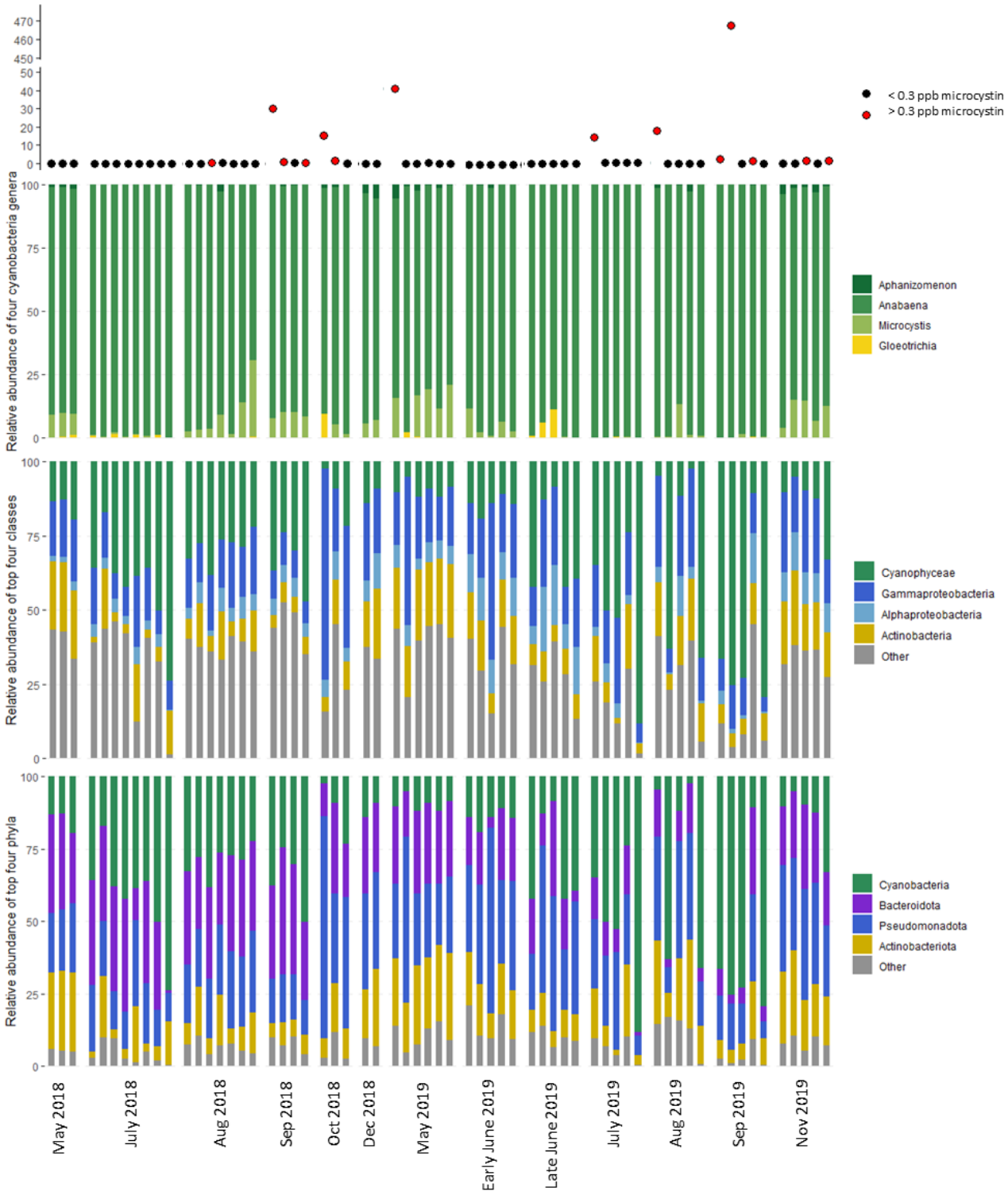

Figure 4

Coarse-level microbial community composition in UKL samples across 2018-2019. Microcystin concentration shown as symbols at top, with concentrations ≥0.3 ppb in red. Relative abundances of the four toxin producing cyanobacteria genera (top bar graph), four most abundant microbial classes (middle bar graph), and four most abundant microbial phyla (bottom bar graph). Samples are ordered by date on the X-axis.

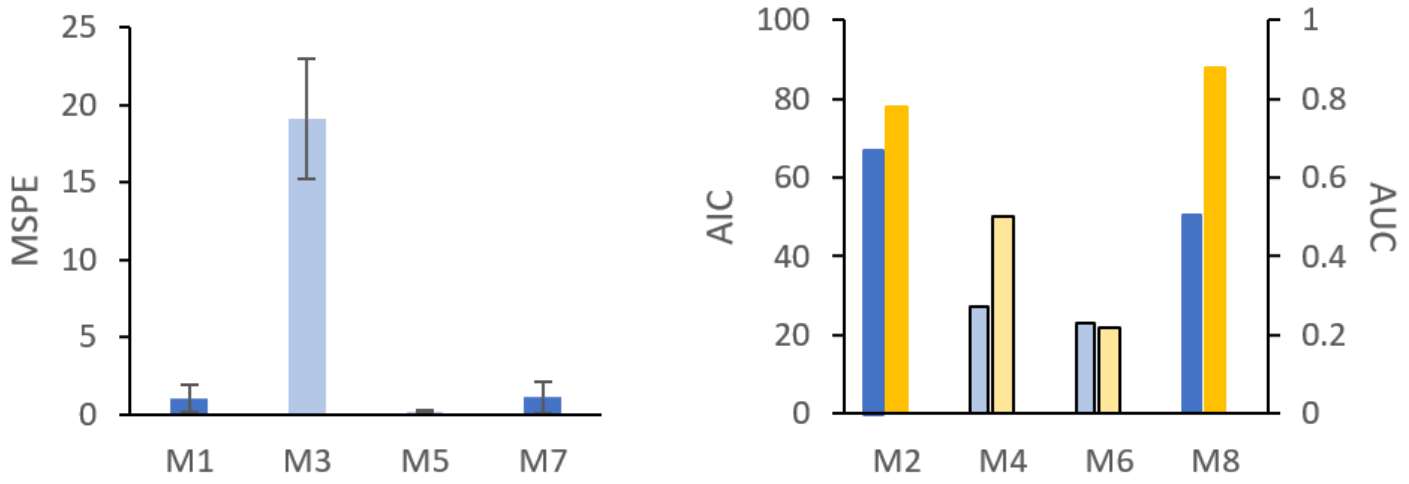

Figure 5

Statistical performances of linear models (left) and logistic models (right) predicting microcystin toxicity. VOC-based M1, M2, M7, M8 (dark blue); environmental parameter-based M3, M4, M5, M6 (light colors). AIC, blue bars; AUC, yellow bars. Error bars are SD.

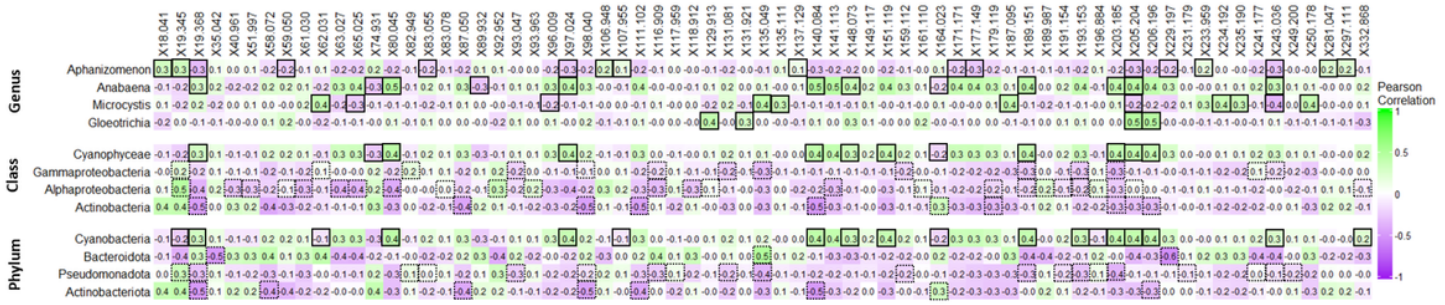

Figure 6

Pearson correlations between the relative abundances of four toxin producing cyanobacterial genera (top), four most abundant microbial classes (middle), or four most abundant microbial phyla (bottom) in UKL samples and the associated *m/z+1* values identified in elastic net models. Outlined boxes are *m/z+1* values identified in the model predicting the relative abundance of the taxonomic group in each row. Boxes with a solid outline indicate *m/z+1* values in models predicting relative abundances of cyanobacteria genera, class, or phylum. Note that some *m/z+1* values predictive of cyanobacteria genera relative abundances are also predictive of Cyanophyceae and Cyanobateria relative abundances. Boxes with a dotted outline indicate *m/z+1* values in models predicting relative abundance of other taxonomic classes or phyla. Pearson r value of 1 (green) indicates a positive correlation, a value of -1 (purple) indicates a negative correlation.

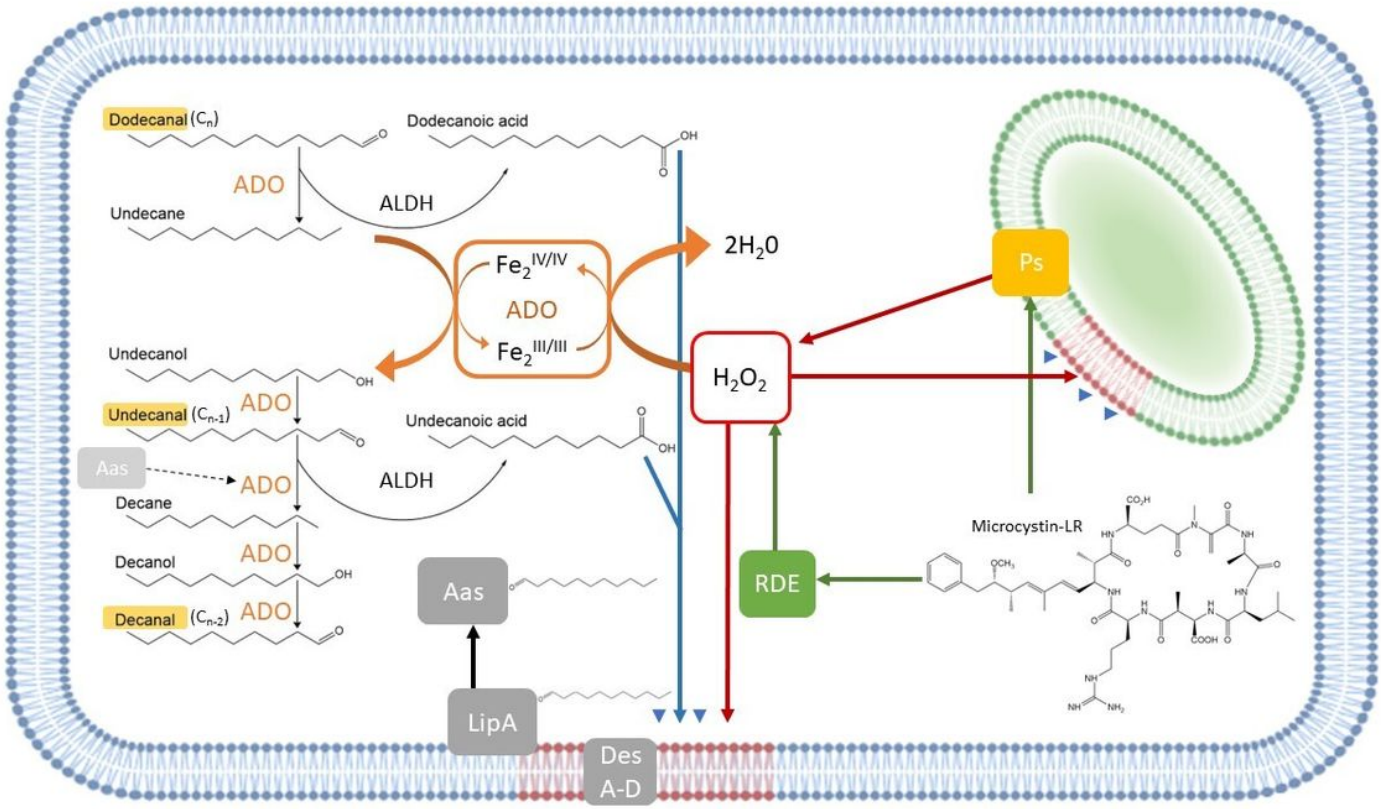

Figure 7

The role of the saturated fatty aldehyde (SFA) oxidation pathway in membrane lipid repair and depletion of reactive oxygen species. SFAs putatively identified by elastic net models predicting microcystin concentration are in yellow (m/z+1 values 157.157, 171.171, and 185.185). SFAs produced by aldehyde-deformylating oxygenase (ADO) are metabolized by aldehyde dehydrogenase (ALDH) producing SFAs. Blue arrows and arrowheads represent SFAs used to repair thylakoid (green) and plasma membrane (grey) lipids damaged by ROS (red), represented here by H_2O_2. Photosynthesis (Ps) unavoidably produces ROS. Medium chained alkanes, (C10-C12) reduce ROS via ADO activity (see text). Microcystins (MC) can protect against ROS by binding to the photosynthetic subunits and by binding to and promoting production of ROS degrading enzymes (RDE). Desaturases A-D (DesA-D) unlink glycerol from the fatty acyl moieties in the membrane. The fatty acyl is removed from the membrane by lipolytic enzyme (LipA). Acyl-ACP synthase (Aas) can reattach the fatty acyl to ADO to reenter the SFA oxidation pathway, represented by a dotted arrow.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- VOCstoxicologycommunitySupplementaryinfo04142023.docx
- VOCstoxicologycommunityTableS412102022.csv