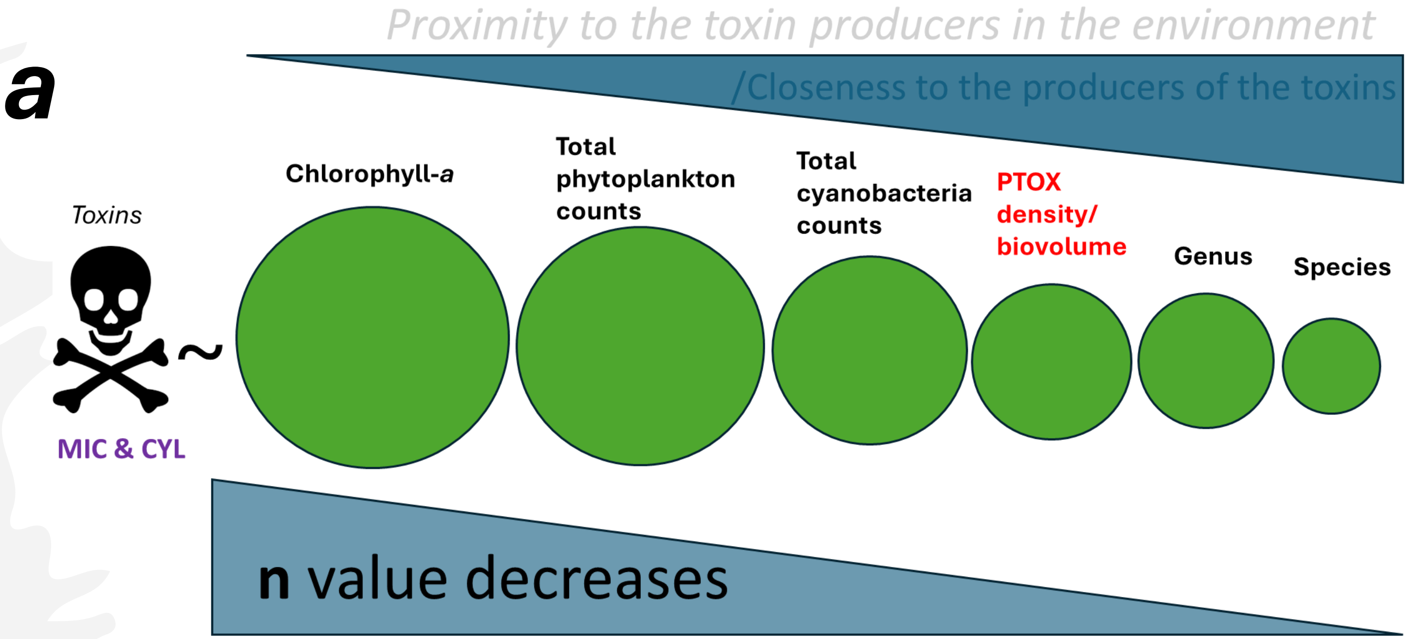
Which Fits Best? Evaluating Chlorophyll-*a* vs. Phytoplankton Composition as Predictors of Cyanotoxins in U.S. Lakes

Yusuf Olaleye

5/6/2025

In theory…



Proximity to the toxin producers in the environment

/Closeness to the producers of the toxins

Toxins

MIC & CYL

~

Chlorophyll-*a*

Total phytoplankton counts

Total cyanobacteria counts

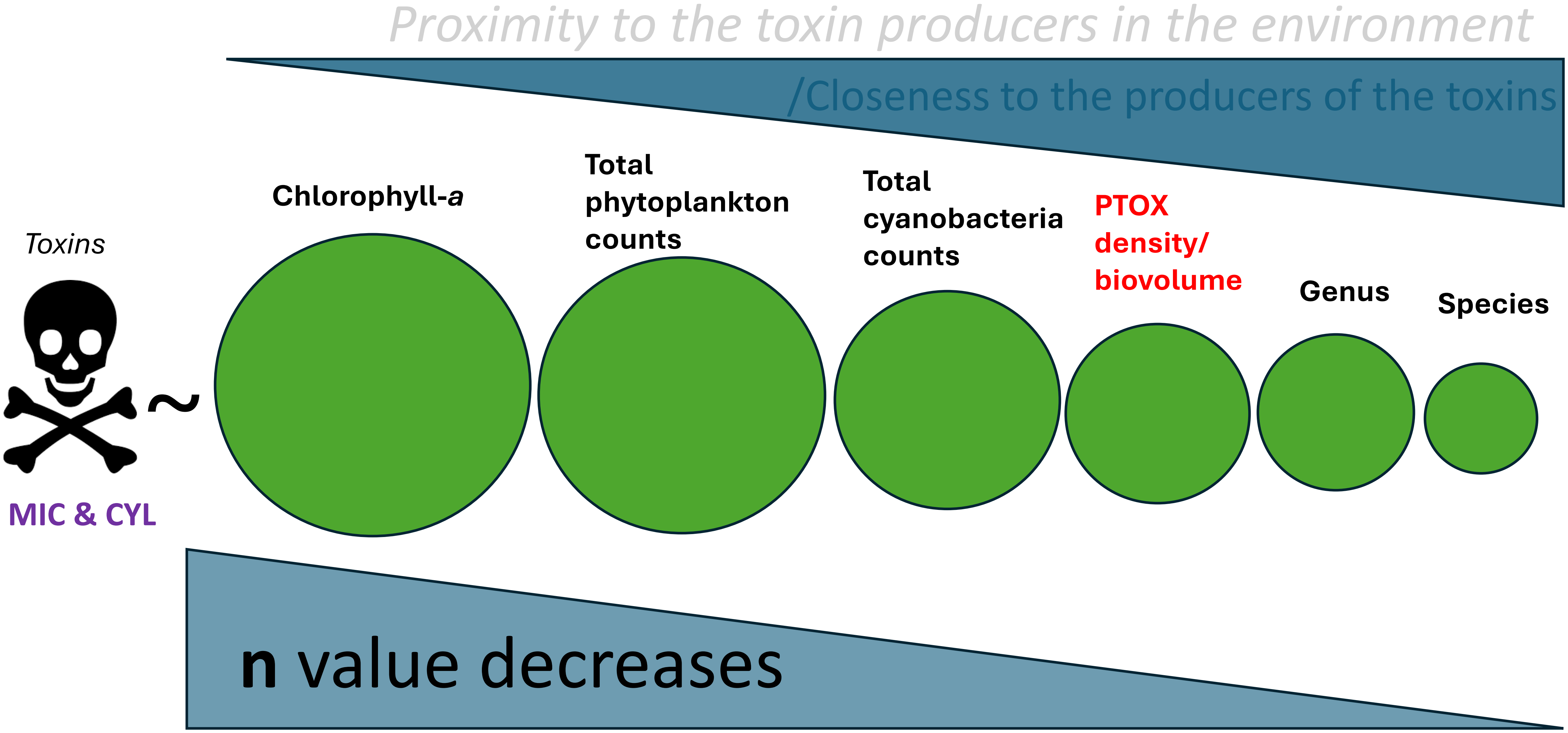
PTOX density/ biovolume

Genus

Species

n value decreases

In theory...



Proximity to the toxin producers in the environment

/Closeness to the producers of the toxins

Toxins

~

MIC & CYL

Chlorophyll-*a*

Total phytoplankton counts

Total cyanobacteria counts

PTOX density/ biovolume

Genus

Species

n value decreases

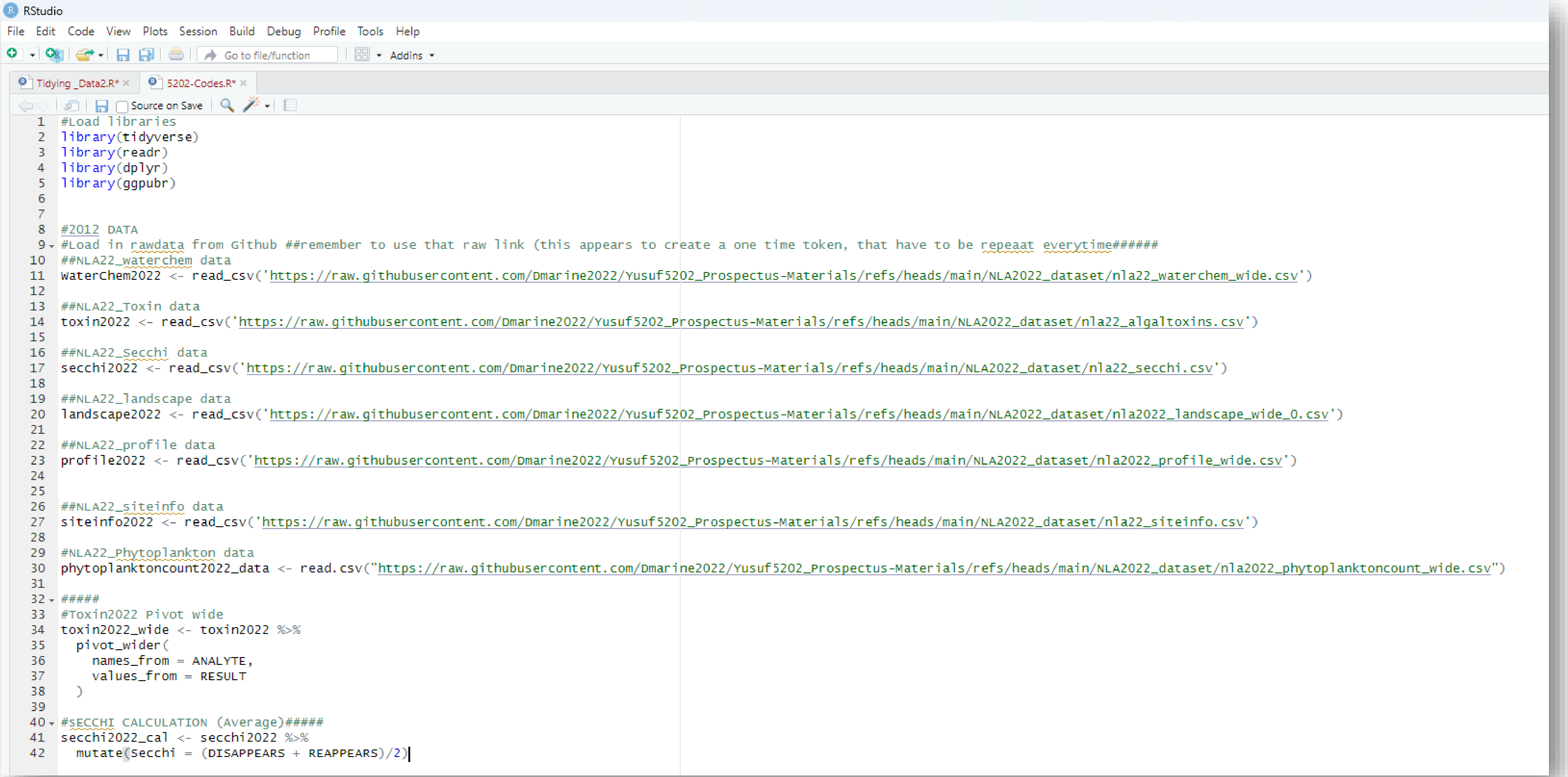
Research questions

- Which of the following biological indicators is the best predictor of cyanotoxins concentrations in U.S. lakes: chlorophyll-a (chl-a) concentration, total phytoplankton biovolume, or cyanobacterial abundance?

- Are the same biological indicators equally predictive of Microcystin and Cylindrospermopsin concentrations in U.S. lakes?

U.S. EPA's NLA 2022 Data

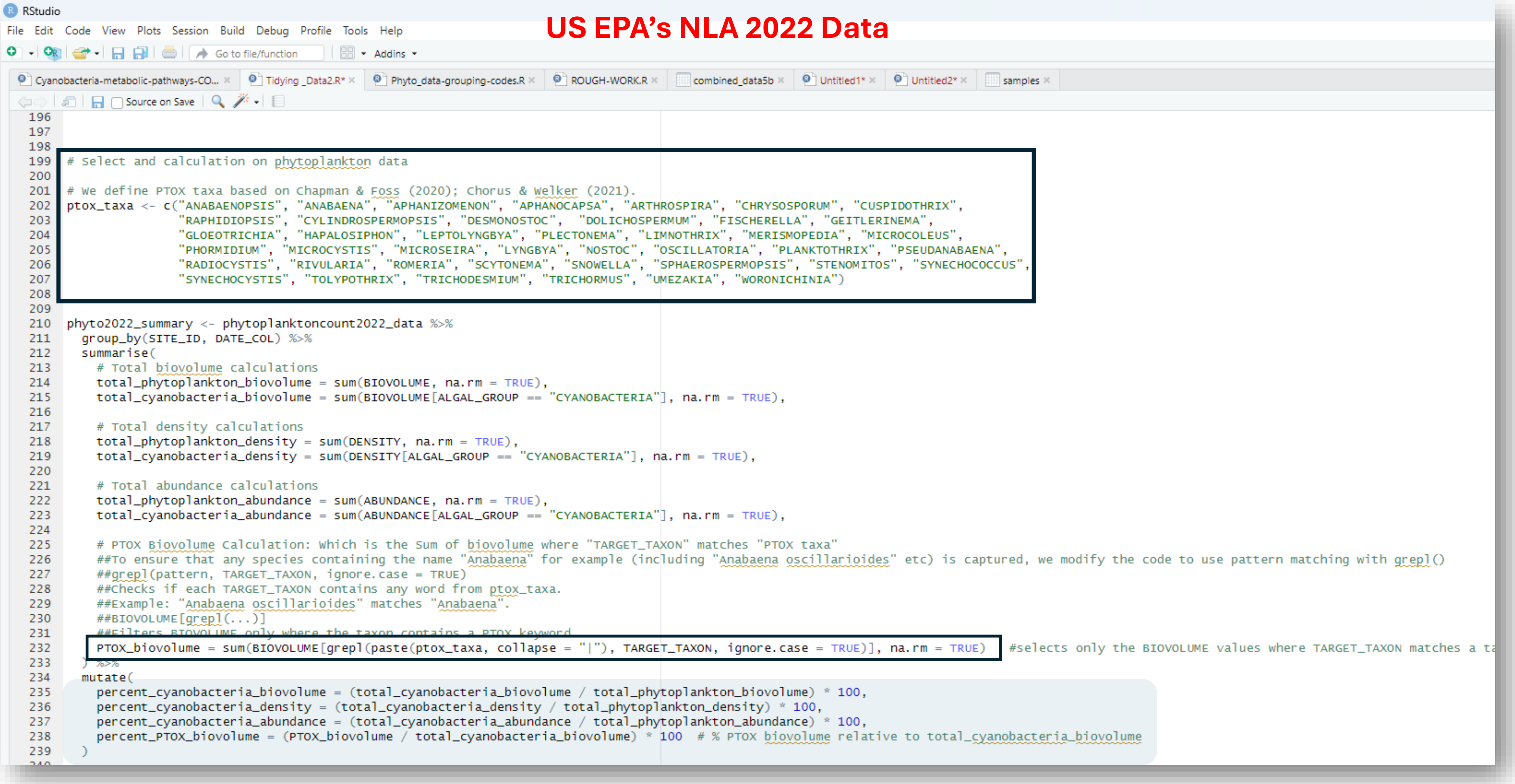
Total number of lakes surveyed = 981



```r
#Load libraries
library(tidyverse)
library(readr)
library(dplyr)
library(ggpubr)


#2012 DATA
#Load in rawdata from Github ##remember to use that raw link (this appears to create a one time token, that have to be repeaat everytime######
##NLA22_waterchem data
WaterChem2022 <- read_csv('https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla22_waterchem_wide.csv')

##NLA22_Toxin data
toxin2022 <- read_csv('https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla22_algaltoxins.csv')

##NLA22_Secchi data
secchi2022 <- read_csv('https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla22_secchi.csv')

##NLA22_landscape data
landscape2022 <- read_csv('https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla2022_landscape_wide_0.csv')

##NLA22_profile data
profile2022 <- read_csv('https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla2022_profile_wide.csv')


##NLA22_siteinfo data
siteinfo2022 <- read_csv('https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla22_siteinfo.csv')

#NLA22_Phytoplankton data
phytoplanktoncount2022_data <- read.csv("https://raw.githubusercontent.com/Dmarine2022/Yusuf5202_Prospectus-Materials/refs/heads/main/NLA2022_dataset/nla2022_phytoplanktoncount_wide.csv")

#####
#Toxin2022 Pivot wide
toxin2022_wide <- toxin2022 %>%
  pivot_wider(
    names_from = ANALYTE,
    values_from = RESULT
  )

#SECCHI CALCULATION (Average)#####
secchi2022_cal <- secchi2022 %>%
  mutate(Secchi = (DISAPPEARS + REAPPEARS)/2)
```

Grouping of Potentially toxigenic (PTOX) Cyanobacteria species

US EPA's NLA 2022 Data

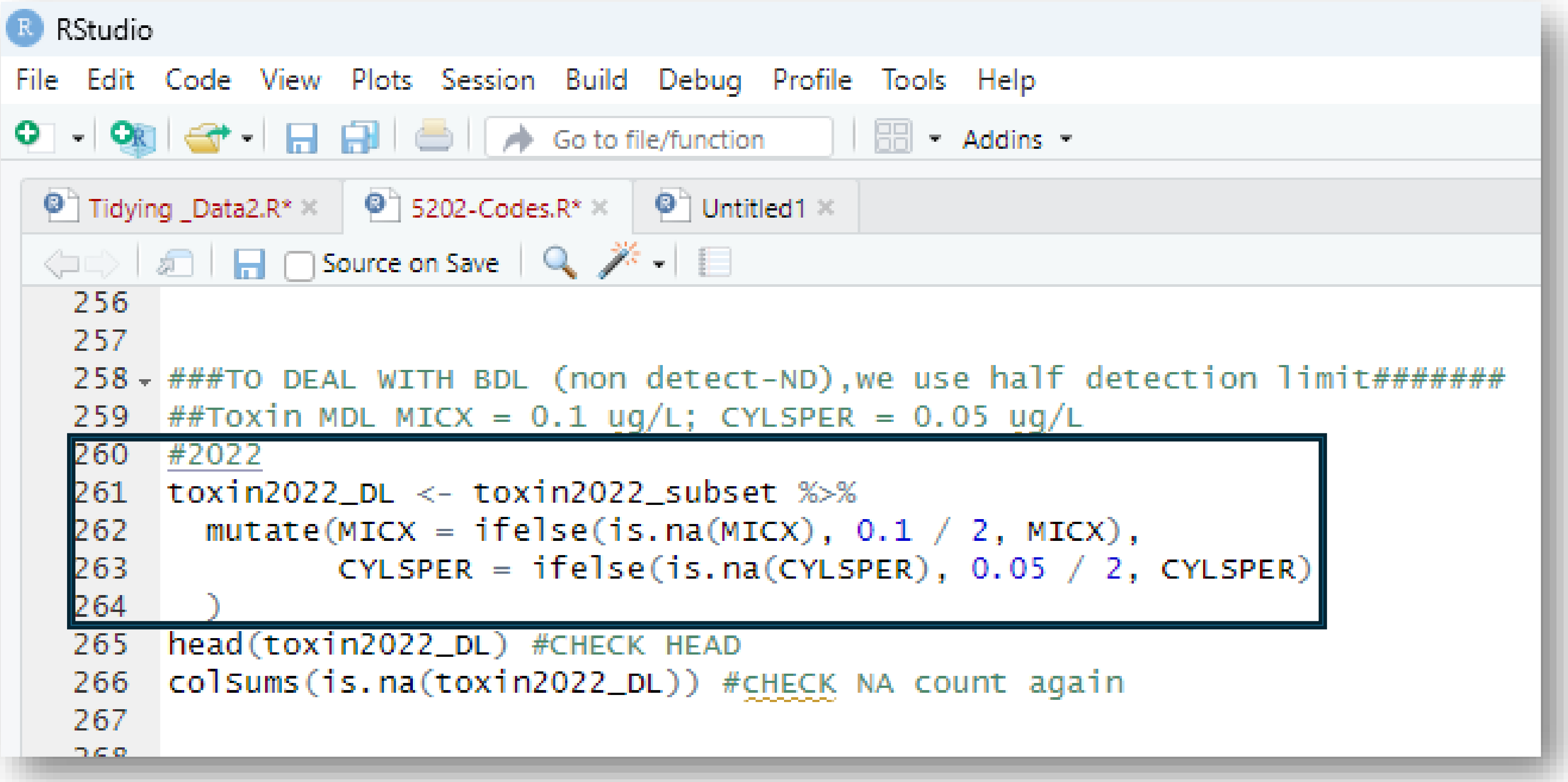


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Cyanobacteria-metabolic-pathways-CO...
Tidying _Data2.R*
Phyto_data-grouping-codes.R
ROUGH-WORK.R
combined_data5b
Untitled1*
Untitled2*
samples
Source on Save

```r


# Select and calculation on phytoplankton data

# We define PTOX taxa based on Chapman & Foss (2020); Chorus & Welker (2021).
ptox_taxa <- c("ANABAENOPSIS", "ANABAENA", "APHANIZOMENON", "APHANOCAPSA", "ARTHROSPIRA", "CHRYSOSPORUM", "CUSPIDOTHRIX",
               "RAPHIDIOPSIS", "CYLINDROSPERMOPSIS", "DESMONOSTOC",  "DOLICHOSPERMUM", "FISCHERELLA", "GEITLERINEMA",
               "GLOEOTRICHIA", "HAPALOSIPHON", "LEPTOLYNGBYA", "PLECTONEMA", "LIMNOTHRIX", "MERISMOPEDIA", "MICROCOLEUS",
               "PHORMIDIUM", "MICROCYSTIS", "MICROSEIRA", "LYNGBYA", "NOSTOC", "OSCILLATORIA", "PLANKTOTHRIX", "PSEUDANABAENA",
               "RADIOCYSTIS", "RIVULARIA", "ROMERIA", "SCYTONEMA", "SNOWELLA", "SPHAEROSPERMOPSIS", "STENOMITOS", "SYNECHOCOCCUS",
               "SYNECHOCYSTIS", "TOLYPOTHRIX", "TRICHODESMIUM", "TRICHORMUS", "UMEZAKIA", "WORONICHINIA")

phyto2022_summary <- phytoplanktoncount2022_data %>%
  group_by(SITE_ID, DATE_COL) %>%
  summarise(
    # Total biovolume calculations
    total_phytoplankton_biovolume = sum(BIOVOLUME, na.rm = TRUE),
    total_cyanobacteria_biovolume = sum(BIOVOLUME[ALGAL_GROUP == "CYANOBACTERIA"], na.rm = TRUE),

    # Total density calculations
    total_phytoplankton_density = sum(DENSITY, na.rm = TRUE),
    total_cyanobacteria_density = sum(DENSITY[ALGAL_GROUP == "CYANOBACTERIA"], na.rm = TRUE),

    # Total abundance calculations
    total_phytoplankton_abundance = sum(ABUNDANCE, na.rm = TRUE),
    total_cyanobacteria_abundance = sum(ABUNDANCE[ALGAL_GROUP == "CYANOBACTERIA"], na.rm = TRUE),

    # PTOX Biovolume Calculation: Which is the Sum of biovolume where "TARGET_TAXON" matches "PTOX taxa"
    ##To ensure that any species containing the name "Anabaena" for example (including "Anabaena oscillarioides" etc) is captured, we modify the code to use pattern matching with grepl()
    ##grepl(pattern, TARGET_TAXON, ignore.case = TRUE)
    ##Checks if each TARGET_TAXON contains any word from ptox_taxa.
    ##Example: "Anabaena oscillarioides" matches "Anabaena".
    ##BIOVOLUME[grepl(...)]
    ##Filters BIOVOLUME only where the taxon contains a PTOX keyword
    PTOX_biovolume = sum(BIOVOLUME[grepl(paste(ptox_taxa, collapse = "|"), TARGET_TAXON, ignore.case = TRUE)], na.rm = TRUE)  #selects only the BIOVOLUME values where TARGET_TAXON matches a ta
  ) %>%
  mutate(
    percent_cyanobacteria_biovolume = (total_cyanobacteria_biovolume / total_phytoplankton_biovolume) * 100,
    percent_cyanobacteria_density = (total_cyanobacteria_density / total_phytoplankton_density) * 100,
    percent_cyanobacteria_abundance = (total_cyanobacteria_abundance / total_phytoplankton_abundance) * 100,
    percent_PTOX_biovolume = (PTOX_biovolume / total_cyanobacteria_biovolume) * 100  # % PTOX biovolume relative to total_cyanobacteria_biovolume
  )
```

Toxin data

Dealing with below detection limit values....



```r


###TO DEAL WITH BDL (non detect-ND),we use half detection limit#######
##Toxin MDL MICX = 0.1 ug/L; CYLSPER = 0.05 ug/L
#2022
toxin2022_DL <- toxin2022_subset %>%
  mutate(MICX = ifelse(is.na(MICX), 0.1 / 2, MICX),
         CYLSPER = ifelse(is.na(CYLSPER), 0.05 / 2, CYLSPER)
  )
head(toxin2022_DL) #CHECK HEAD
colSums(is.na(toxin2022_DL)) #CHECK NA count again

```

Key variables

Response variables

1. -MIC concentrations
2. -CYL concentrations

~

Predictors

1. -Chl-a (Proxy for phytoplankton biomass)
2. -Total phytoplankton density/biovolume
3. -Total cyanobacteria density/biovolume

Derived predictors

4. -PTOX biovolume
   - **4a.** MIC_PTOX biovolume
   - **4b.** CYL_PTOX biovolume
5. -%Cyanobacteria biovolume
6. -%PTOX biovolume

☠ MICROCYSTINS....

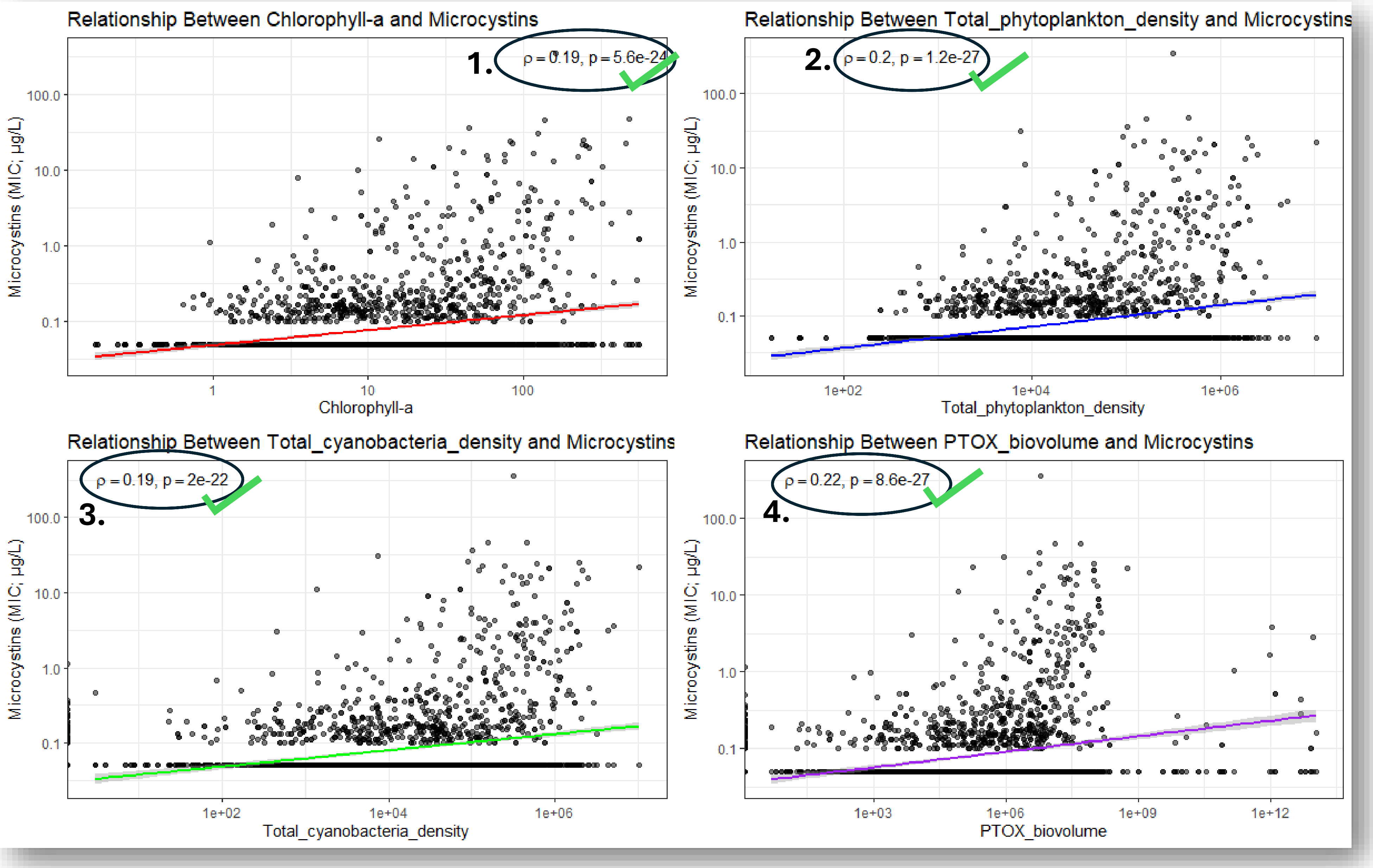


Figure. Relationships between Microcystins and common biological indicators.

CYLINDROSPERMOPSINS….

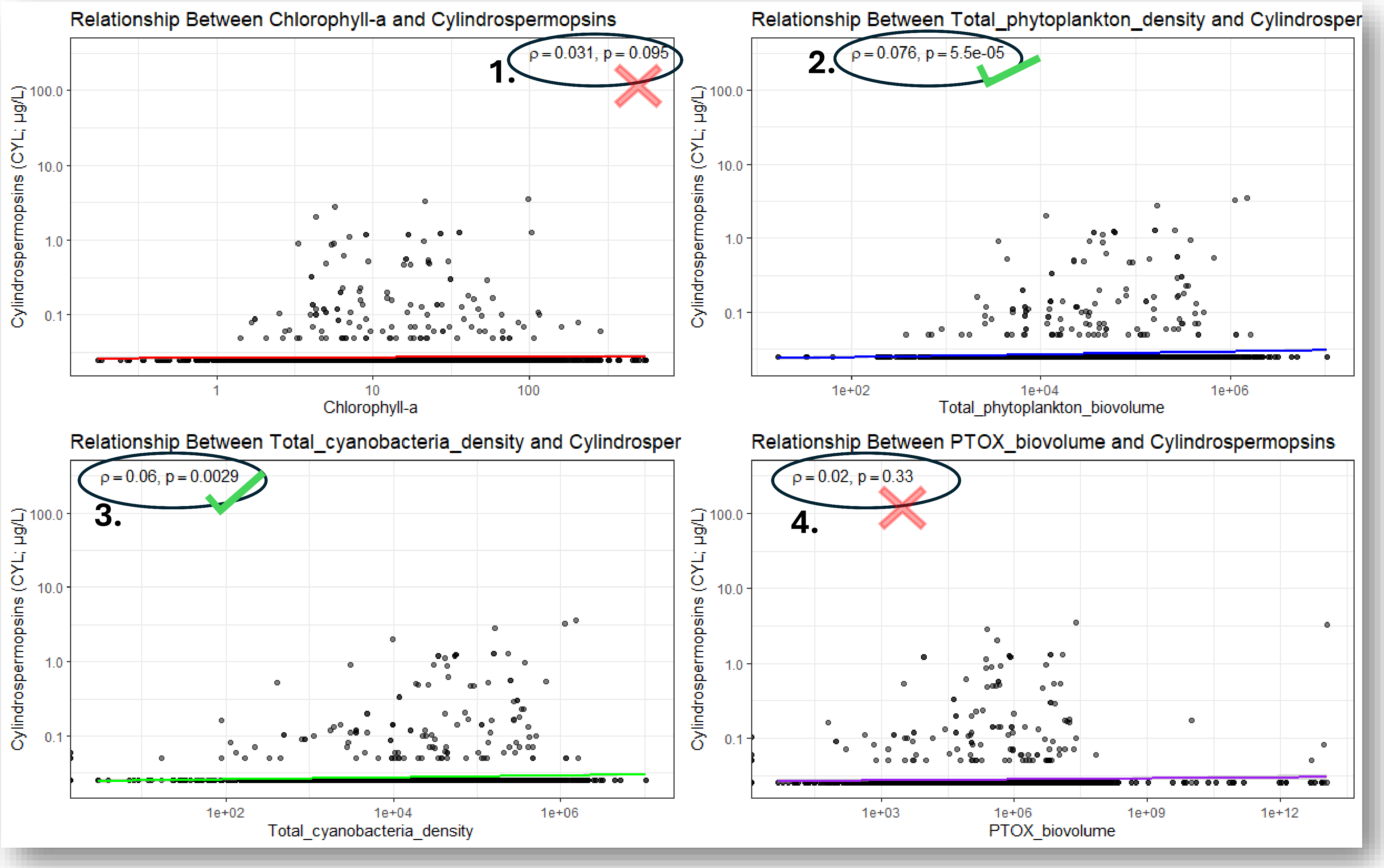
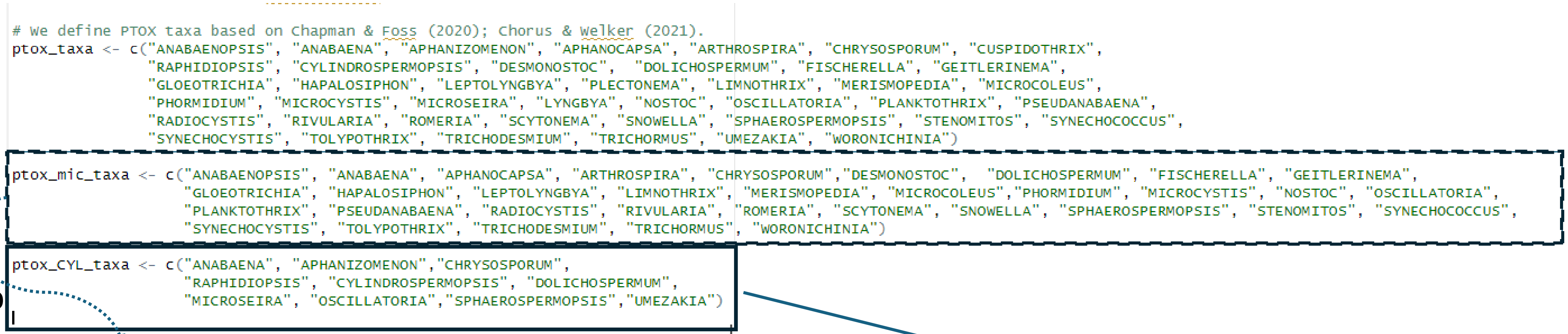


Figure. Relationships between Cylindrospermopsins and common biological indicators.



4.

```
# We define PTOX taxa based on Chapman & Foss (2020); Chorus & Welker (2021).
ptox_taxa <- c("ANABAENOPSIS", "ANABAENA", "APHANIZOMENON", "APHANOCAPSA", "ARTHROSPIRA", "CHRYSOSPORUM", "CUSPIDOTHRIX",
               "RAPHIDIOPSIS", "CYLINDROSPERMOPSIS", "DESMONOSTOC",  "DOLICHOSPERMUM", "FISCHERELLA", "GEITLERINEMA",
               "GLOEOTRICHIA", "HAPALOSIPHON", "LEPTOLYNGBYA", "PLECTONEMA", "LIMNOTHRIX", "MERISMOPEDIA", "MICROCOLEUS",
               "PHORMIDIUM", "MICROCYSTIS", "MICROSEIRA", "LYNGBYA", "NOSTOC", "OSCILLATORIA", "PLANKTOTHRIX", "PSEUDANABAENA",
               "RADIOCYSTIS", "RIVULARIA", "ROMERIA", "SCYTONEMA", "SNOWELLA", "SPHAEROSPERMOPSIS", "STENOMITOS", "SYNECHOCOCCUS",
               "SYNECHOCYSTIS", "TOLYPOTHRIX", "TRICHODESMIUM", "TRICHORMUS", "UMEZAKIA", "WORONICHINIA")
```

4.a

```
ptox_mic_taxa <- c("ANABAENOPSIS", "ANABAENA", "APHANOCAPSA", "ARTHROSPIRA", "CHRYSOSPORUM","DESMONOSTOC",  "DOLICHOSPERMUM", "FISCHERELLA", "GEITLERINEMA",
                   "GLOEOTRICHIA", "HAPALOSIPHON", "LEPTOLYNGBYA", "LIMNOTHRIX", "MERISMOPEDIA", "MICROCOLEUS","PHORMIDIUM", "MICROCYSTIS", "NOSTOC", "OSCILLATORIA",
                   "PLANKTOTHRIX", "PSEUDANABAENA", "RADIOCYSTIS", "RIVULARIA", "ROMERIA", "SCYTONEMA", "SNOWELLA", "SPHAEROSPERMOPSIS", "STENOMITOS", "SYNECHOCOCCUS",
                   "SYNECHOCYSTIS", "TOLYPOTHRIX", "TRICHODESMIUM", "TRICHORMUS", "WORONICHINIA")
```

4.b

```
ptox_CYL_taxa <- c("ANABAENA", "APHANIZOMENON","CHRYSOSPORUM",
                   "RAPHIDIOPSIS", "CYLINDROSPERMOPSIS", "DOLICHOSPERMUM",
                   "MICROSEIRA", "OSCILLATORIA","SPHAEROSPERMOPSIS","UMEZAKIA")
```

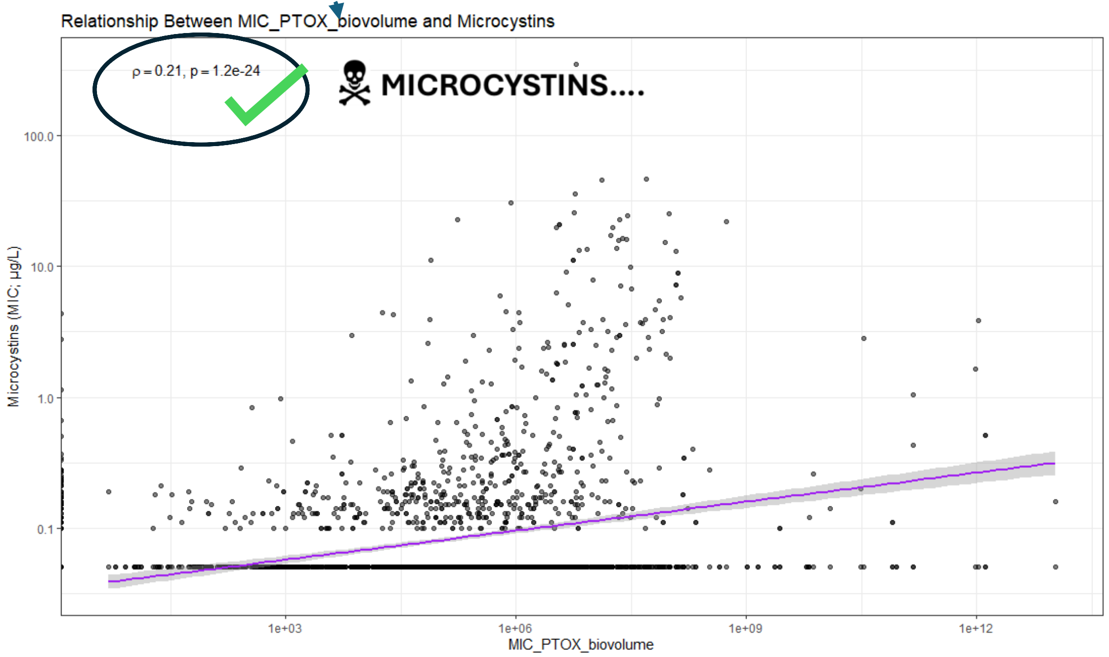


Figure. Relationships between MICs and potentially microcystins producing cyanobacteria (MIC_PTOX-biovume).

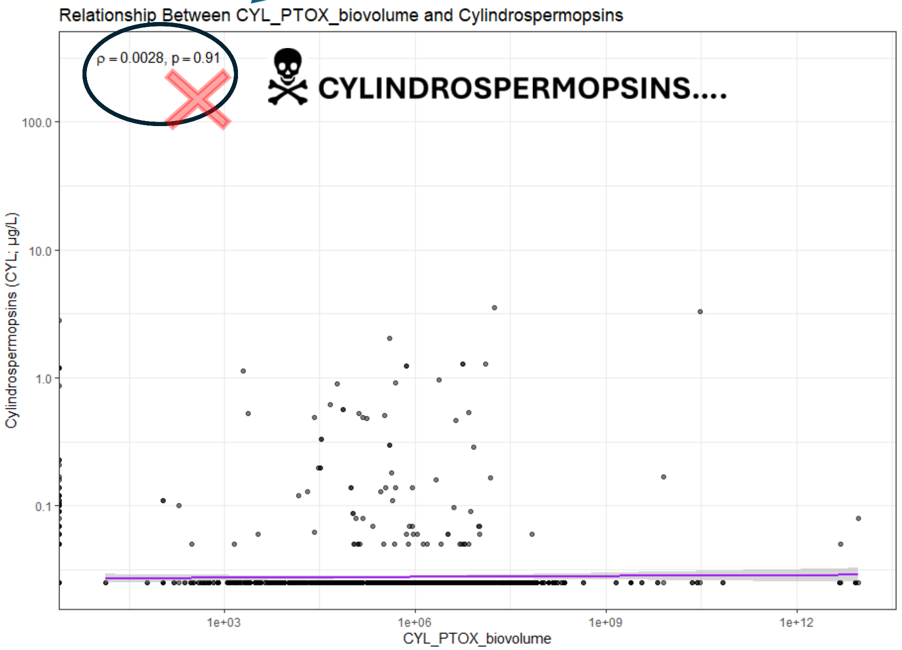
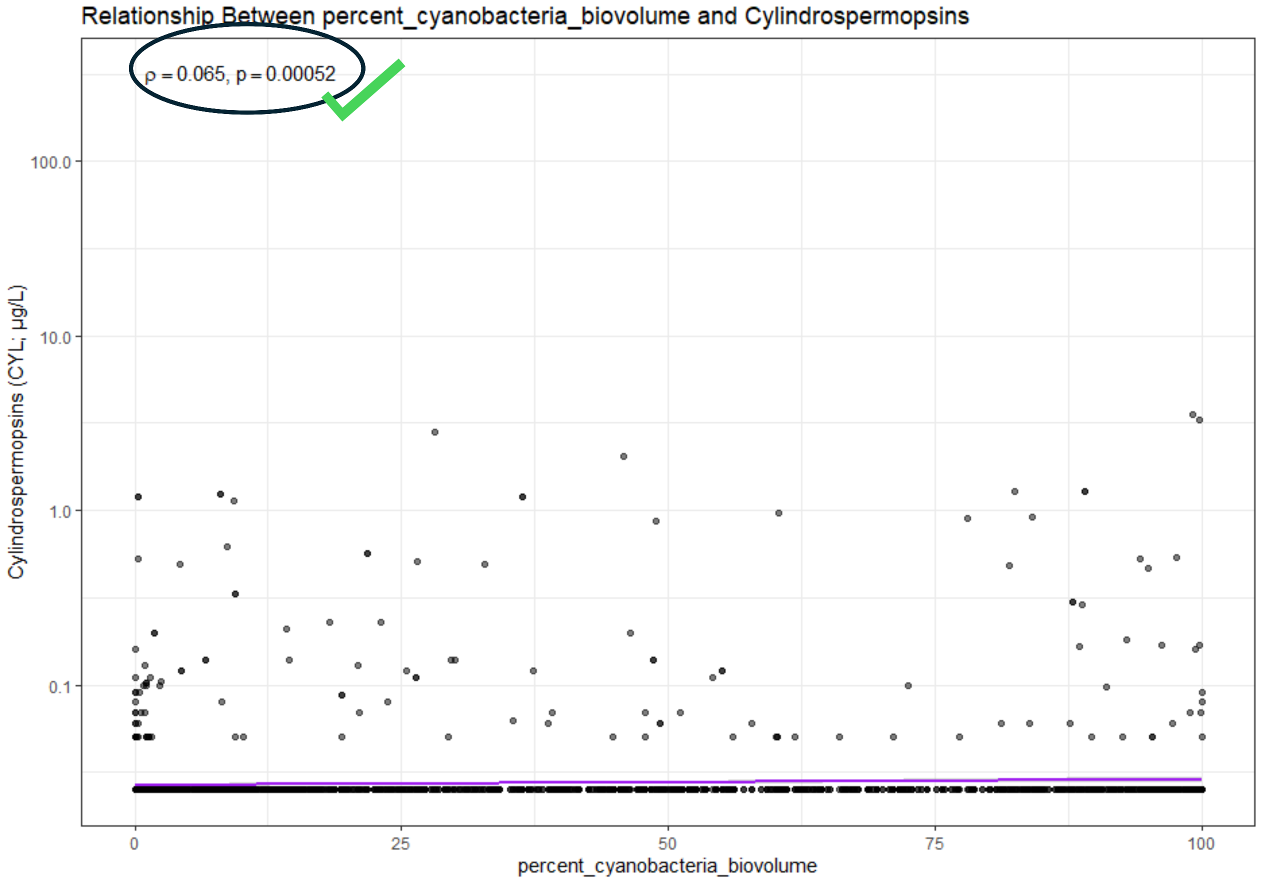


Figure. Relationships between Cylindrospermopsins and potentially cylindrospermopsins producing cyanobacteria (CYL_PTOX-biovolume)

5. percent_cyanobacteria_biovolume =

(total_cyanobacteria_biovolume / total_phytoplankton_biovolume) * 100

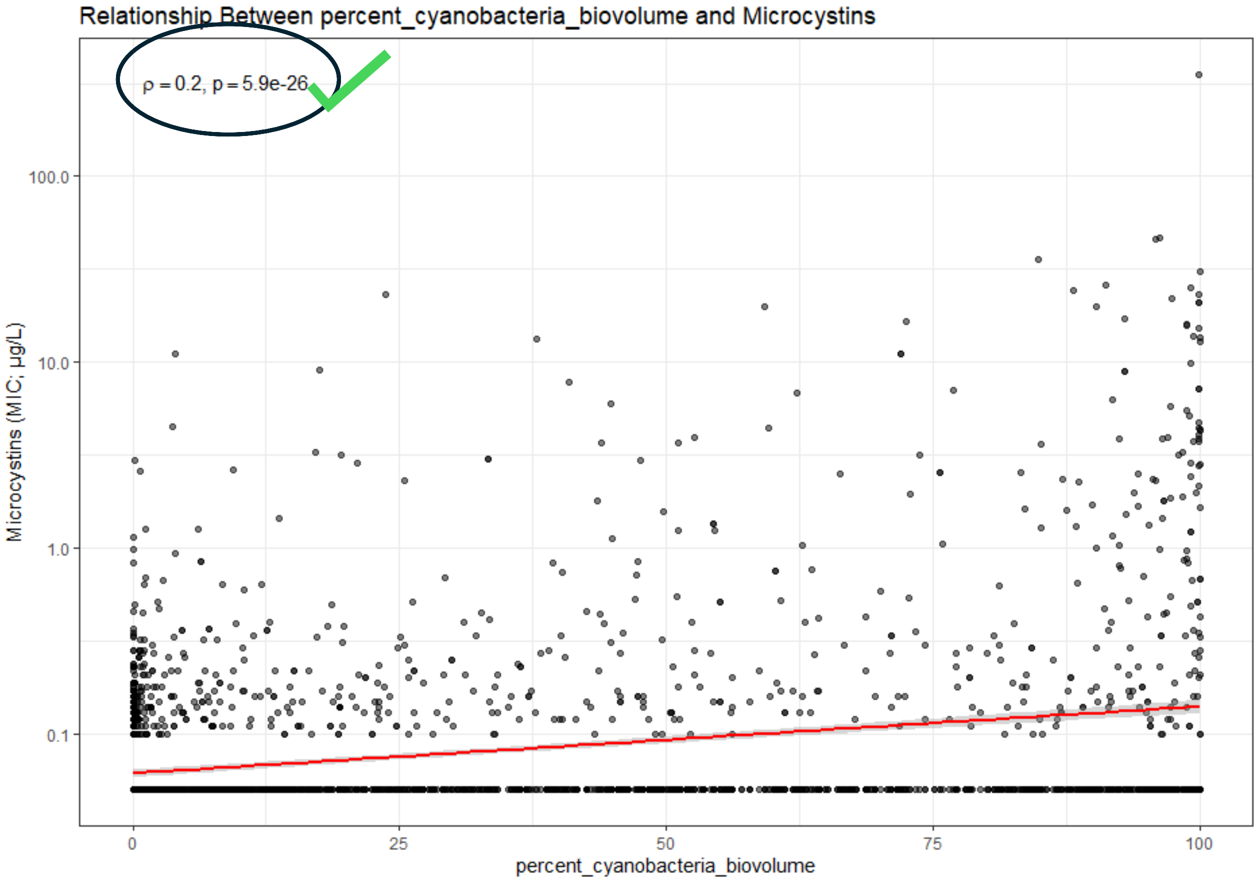
☠ CYLINDROSPERMOPSINS....



Relationship Between percent_cyanobacteria_biovolume and Cylindrospermopsins

$\rho = 0.065$, $p = 0.00052$ ✓

☠ MICROCYSTINS....



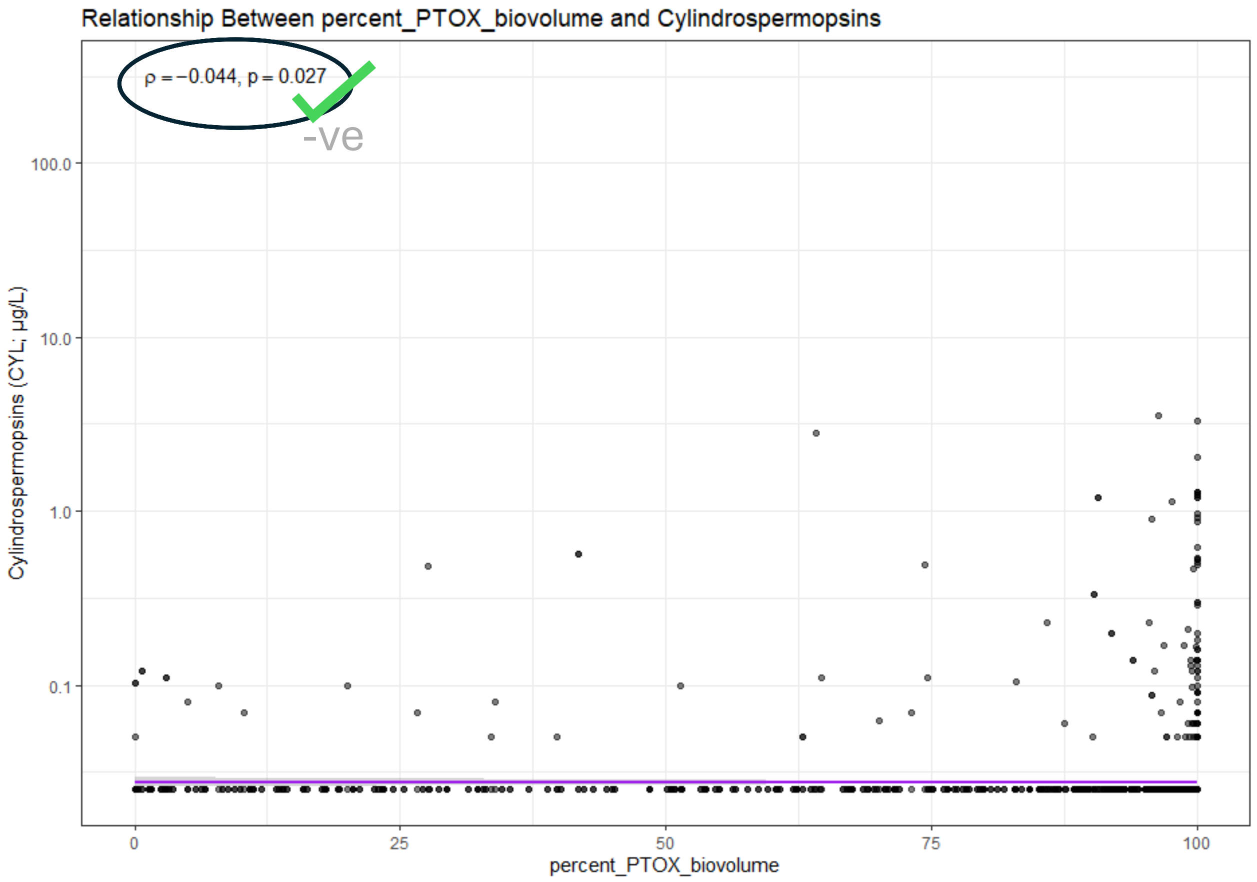
Relationship Between percent_cyanobacteria_biovolume and Microcystins

$\rho = 0.2$, $p = 5.9e\text{-}26$ ✓

6. percent_PTOX_biovolume =

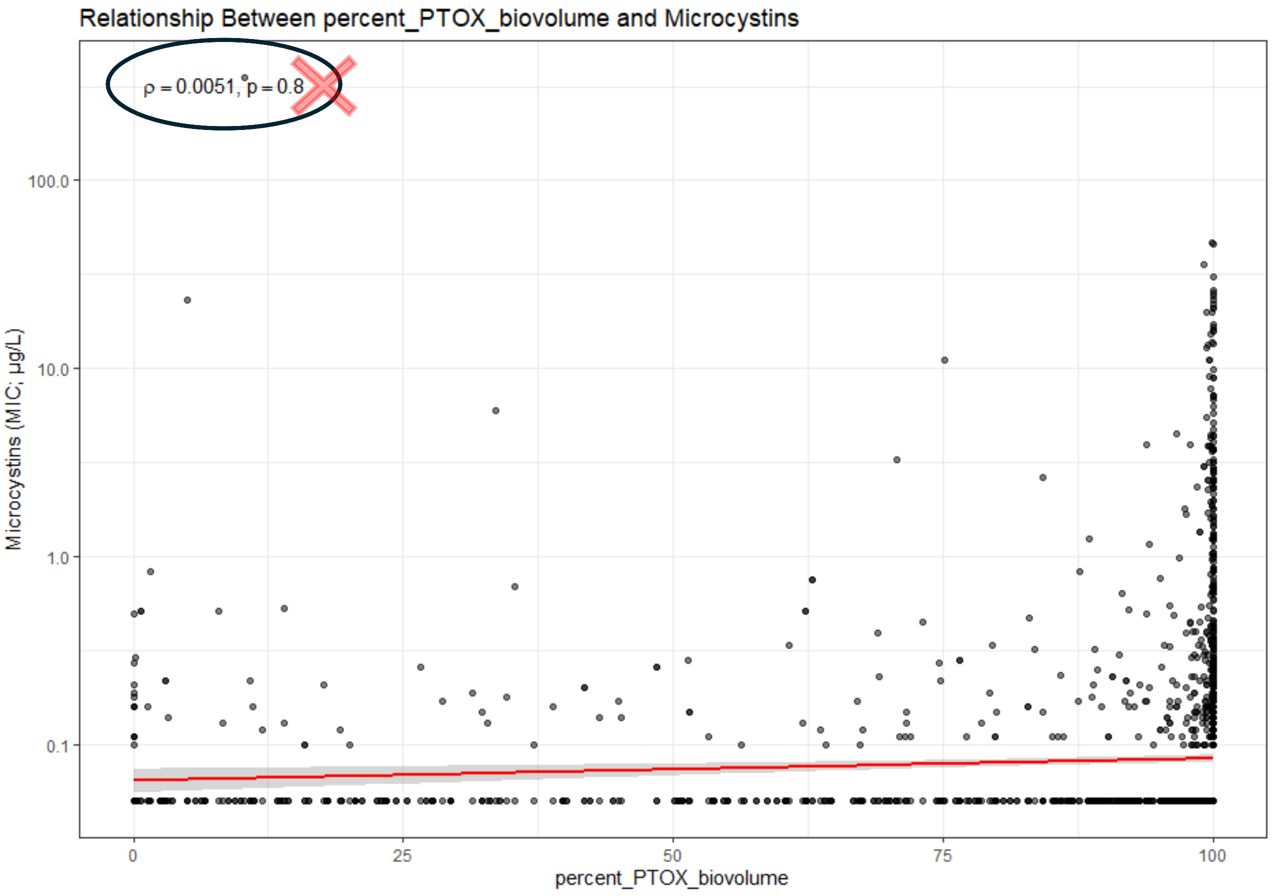
(PTOX_biovolume / total_cyanobacteria_biovolume) * 100

☠ CYLINDROSPERMOPSINS....



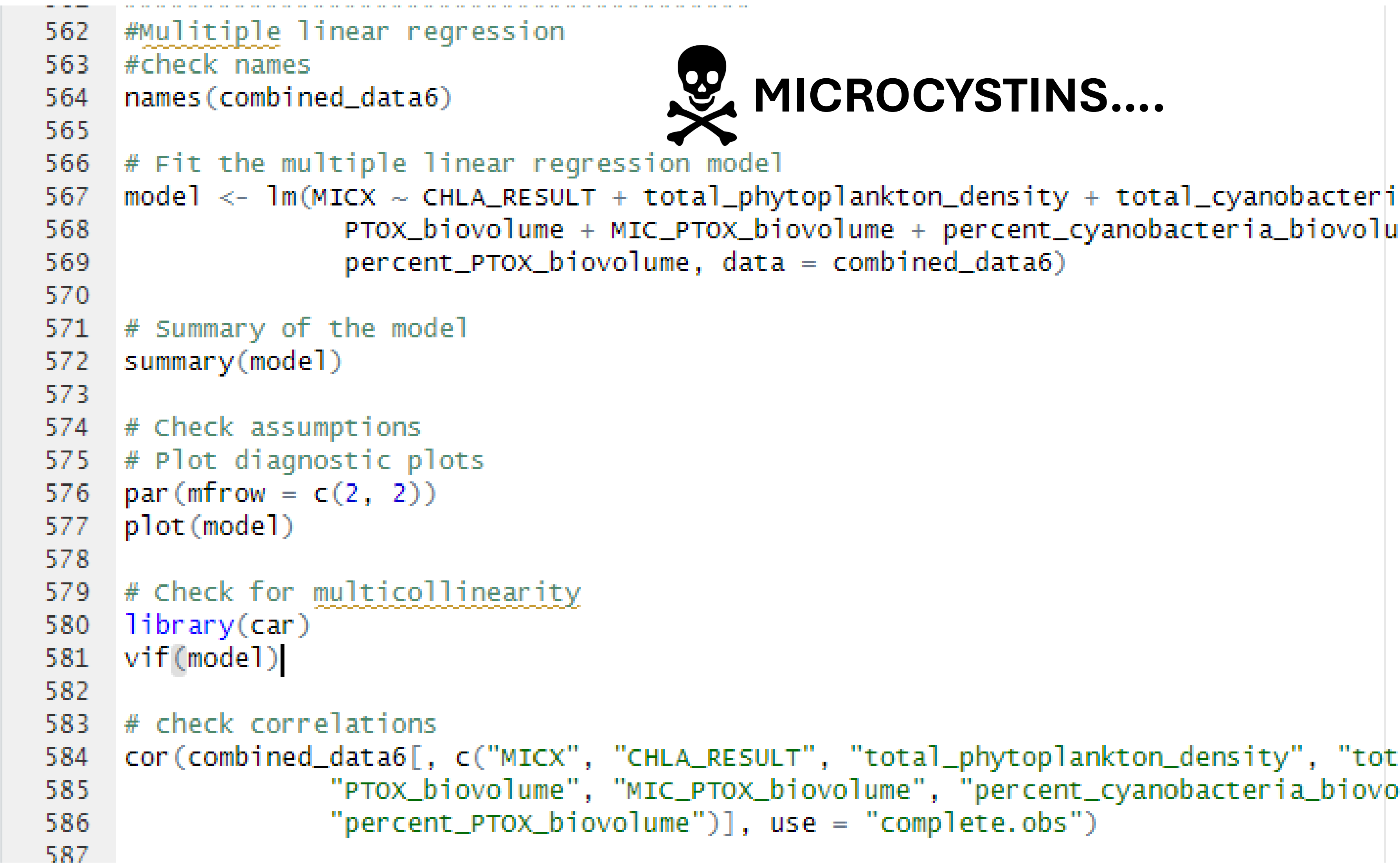
Relationship Between percent_PTOX_biovolume and Cylindrospermopsins

☠ MICROCYSTINS....



Relationship Between percent_PTOX_biovolume and Microcystins

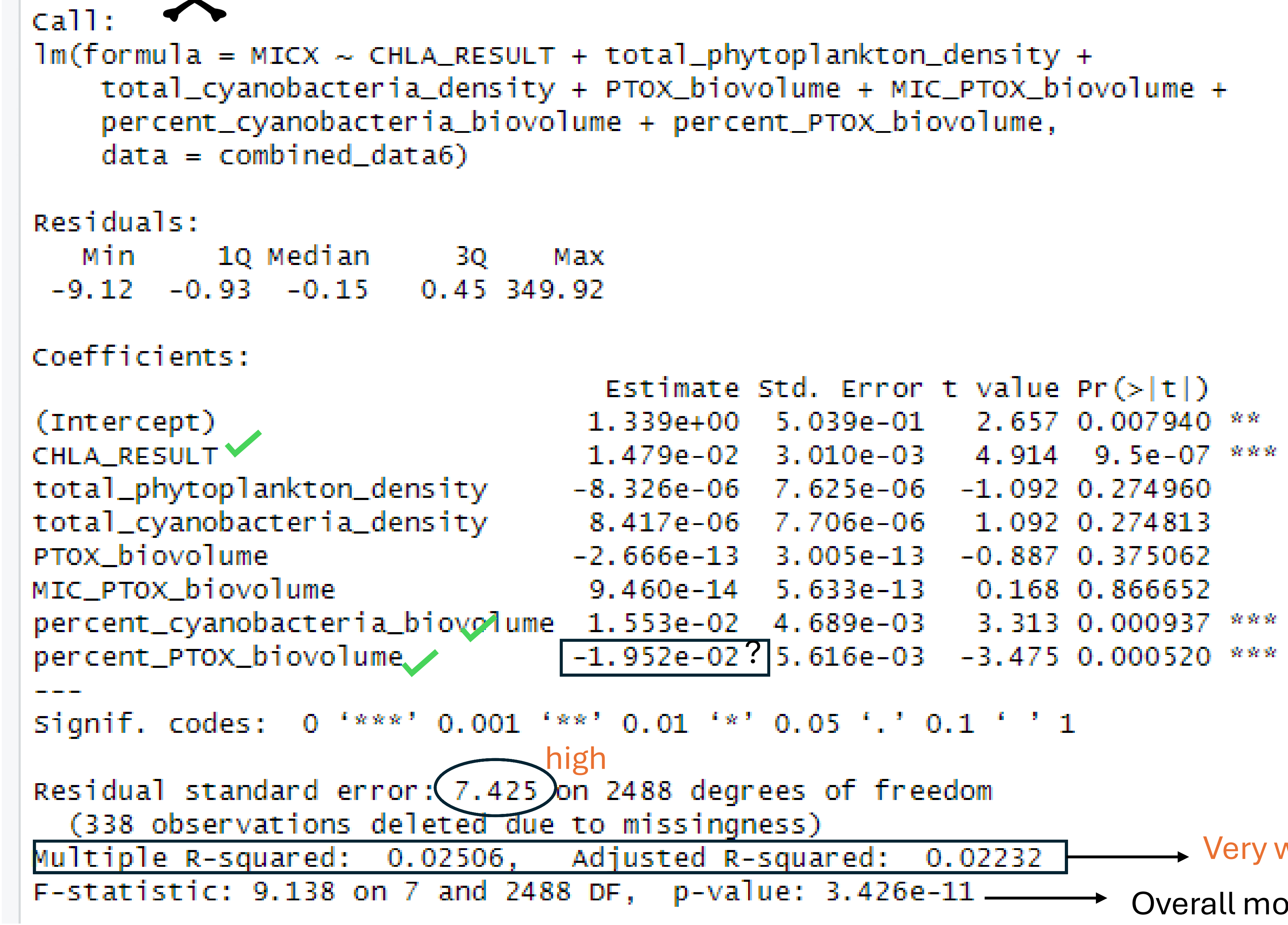
Multiple linear regression



MICROCYSTINS....

```r
#Mulitiple linear regression
#check names
names(combined_data6)

# Fit the multiple linear regression model
model <- lm(MICX ~ CHLA_RESULT + total_phytoplankton_density + total_cyanobacteria_density +
              PTOX_biovolume + MIC_PTOX_biovolume + percent_cyanobacteria_biovolume +
              percent_PTOX_biovolume, data = combined_data6)

# Summary of the model
summary(model)

# Check assumptions
# Plot diagnostic plots
par(mfrow = c(2, 2))
plot(model)

# Check for multicollinearity
library(car)
vif(model)

# check correlations
cor(combined_data6[, c("MICX", "CHLA_RESULT", "total_phytoplankton_density", "total_cyanobacteria_density",
             "PTOX_biovolume", "MIC_PTOX_biovolume", "percent_cyanobacteria_biovolume",
             "percent_PTOX_biovolume")], use = "complete.obs")
```

MICROCYSTINS....



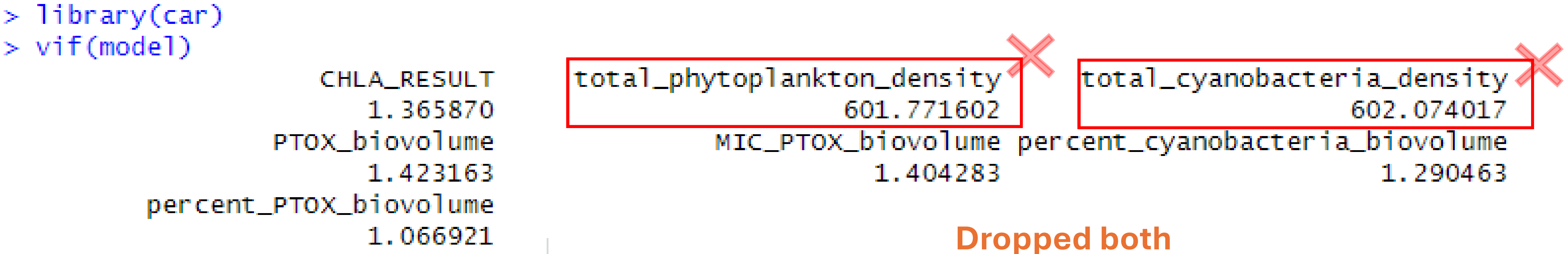
```
Call:
lm(formula = MICX ~ CHLA_RESULT + total_phytoplankton_density +
    total_cyanobacteria_density + PTOX_biovolume + MIC_PTOX_biovolume +
    percent_cyanobacteria_biovolume + percent_PTOX_biovolume,
    data = combined_data6)

Residuals:
   Min     1Q Median     3Q    Max
 -9.12  -0.93  -0.15   0.45 349.92

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.339e+00  5.039e-01   2.657 0.007940 **
CHLA_RESULT ✓                   1.479e-02  3.010e-03   4.914  9.5e-07 ***
total_phytoplankton_density    -8.326e-06  7.625e-06  -1.092 0.274960
total_cyanobacteria_density     8.417e-06  7.706e-06   1.092 0.274813
PTOX_biovolume                 -2.666e-13  3.005e-13  -0.887 0.375062
MIC_PTOX_biovolume              9.460e-14  5.633e-13   0.168 0.866652
percent_cyanobacteria_biovolume ✓ 1.553e-02  4.689e-03   3.313 0.000937 ***
percent_PTOX_biovolume ✓       -1.952e-02 ? 5.616e-03  -3.475 0.000520 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                                          high
Residual standard error: 7.425 on 2488 degrees of freedom
  (338 observations deleted due to missingness)
Multiple R-squared:  0.02506,	Adjusted R-squared:  0.02232
F-statistic: 9.138 on 7 and 2488 DF,  p-value: 3.426e-11
```

Very weak model fit.

Overall model is statistically significant

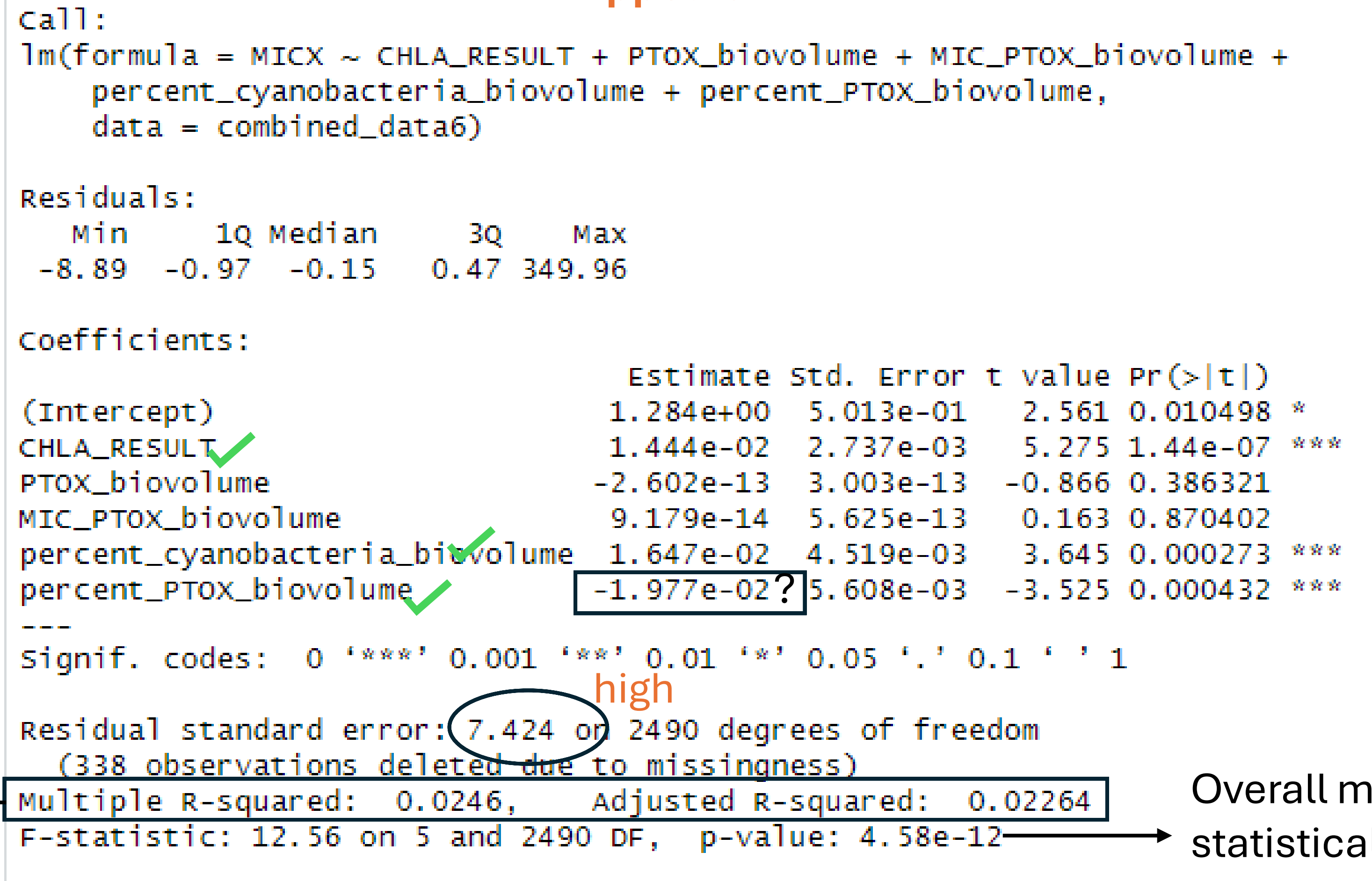
Check for multicollinearity

Variance Inflation Factor (VIF) > 10 → Red flag for multicollinearity



```
> library(car)
> vif(model)
                CHLA_RESULT   total_phytoplankton_density   total_cyanobacteria_density
                   1.365870                    601.771602                    602.074017
             PTOX_biovolume            MIC_PTOX_biovolume percent_cyanobacteria_biovolume
                   1.423163                      1.404283                       1.290463
     percent_PTOX_biovolume
                   1.066921
```

Dropped both



```
Call:
lm(formula = MICX ~ CHLA_RESULT + PTOX_biovolume + MIC_PTOX_biovolume + 
    percent_cyanobacteria_biovolume + percent_PTOX_biovolume, 
    data = combined_data6)

Residuals:
   Min     1Q Median     3Q    Max 
 -8.89  -0.97  -0.15   0.47 349.96 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      1.284e+00  5.013e-01   2.561 0.010498 *  
CHLA_RESULT                      1.444e-02  2.737e-03   5.275 1.44e-07 ***
PTOX_biovolume                  -2.602e-13  3.003e-13  -0.866 0.386321    
MIC_PTOX_biovolume               9.179e-14  5.625e-13   0.163 0.870402    
percent_cyanobacteria_biovolume  1.647e-02  4.519e-03   3.645 0.000273 ***
percent_PTOX_biovolume          -1.977e-02? 5.608e-03  -3.525 0.000432 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.424 on 2490 degrees of freedom
  (338 observations deleted due to missingness)
Multiple R-squared:  0.0246,	Adjusted R-squared:  0.02264 
F-statistic: 12.56 on 5 and 2490 DF,  p-value: 4.58e-12
```

high

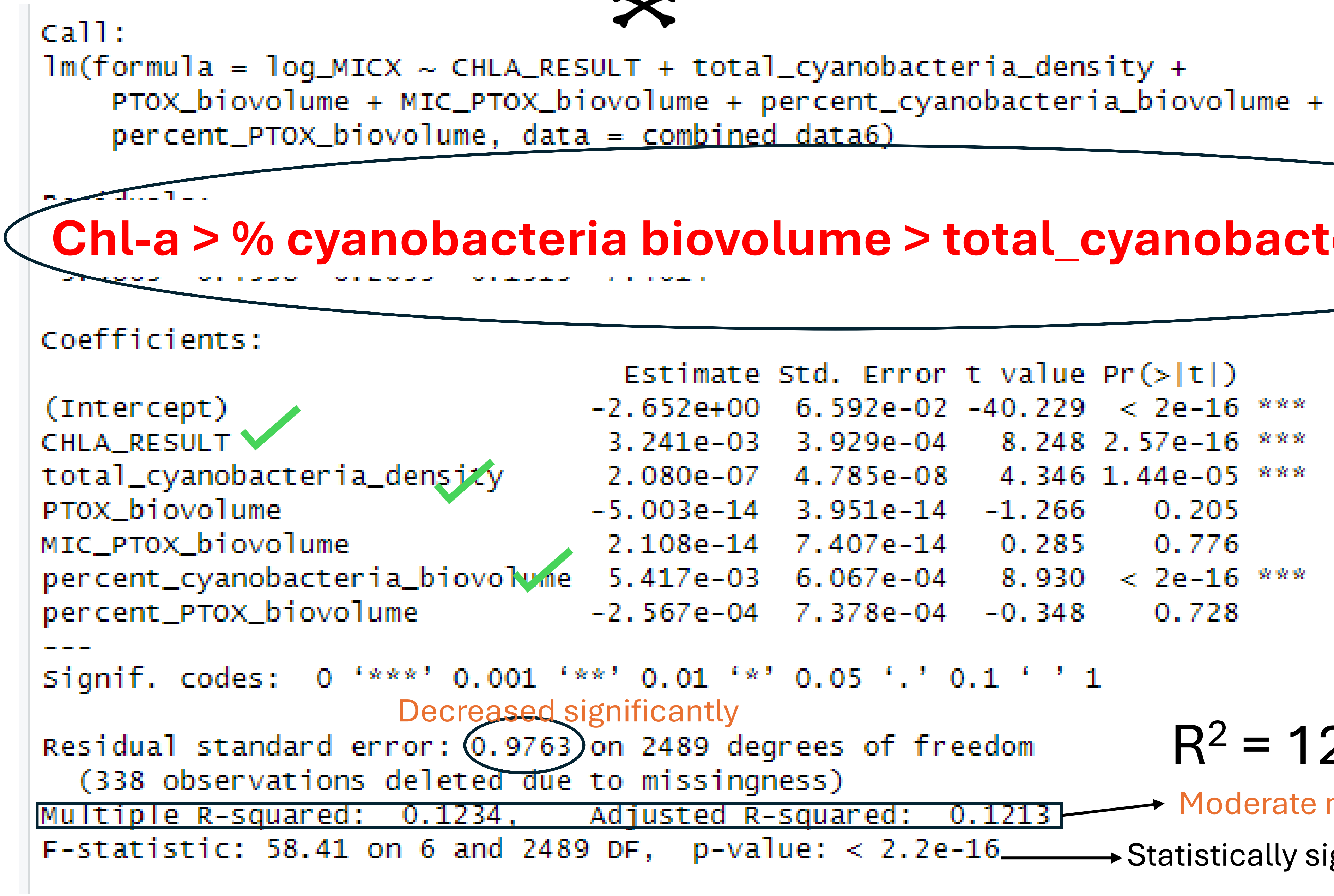
Very weak model fit

Overall model is statistically significant

☠ **MICROCYSTINS....**

Log transform Micx?

MICROCYSTINS….



```
Call:
lm(formula = log_MICX ~ CHLA_RESULT + total_cyanobacteria_density + 
    PTOX_biovolume + MIC_PTOX_biovolume + percent_cyanobacteria_biovolume + 
    percent_PTOX_biovolume, data = combined_data6)

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -2.652e+00  6.592e-02 -40.229  < 2e-16 ***
CHLA_RESULT                      3.241e-03  3.929e-04   8.248 2.57e-16 ***
total_cyanobacteria_density      2.080e-07  4.785e-08   4.346 1.44e-05 ***
PTOX_biovolume                  -5.003e-14  3.951e-14  -1.266    0.205    
MIC_PTOX_biovolume               2.108e-14  7.407e-14   0.285    0.776    
percent_cyanobacteria_biovolume  5.417e-03  6.067e-04   8.930  < 2e-16 ***
percent_PTOX_biovolume          -2.567e-04  7.378e-04  -0.348    0.728    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9763 on 2489 degrees of freedom
  (338 observations deleted due to missingness)
Multiple R-squared:  0.1234,	Adjusted R-squared:  0.1213 
F-statistic: 58.41 on 6 and 2489 DF,  p-value: < 2.2e-16
```

Chl-a > % cyanobacteria biovolume > total_cyanobacteria_density

Decreased significantly

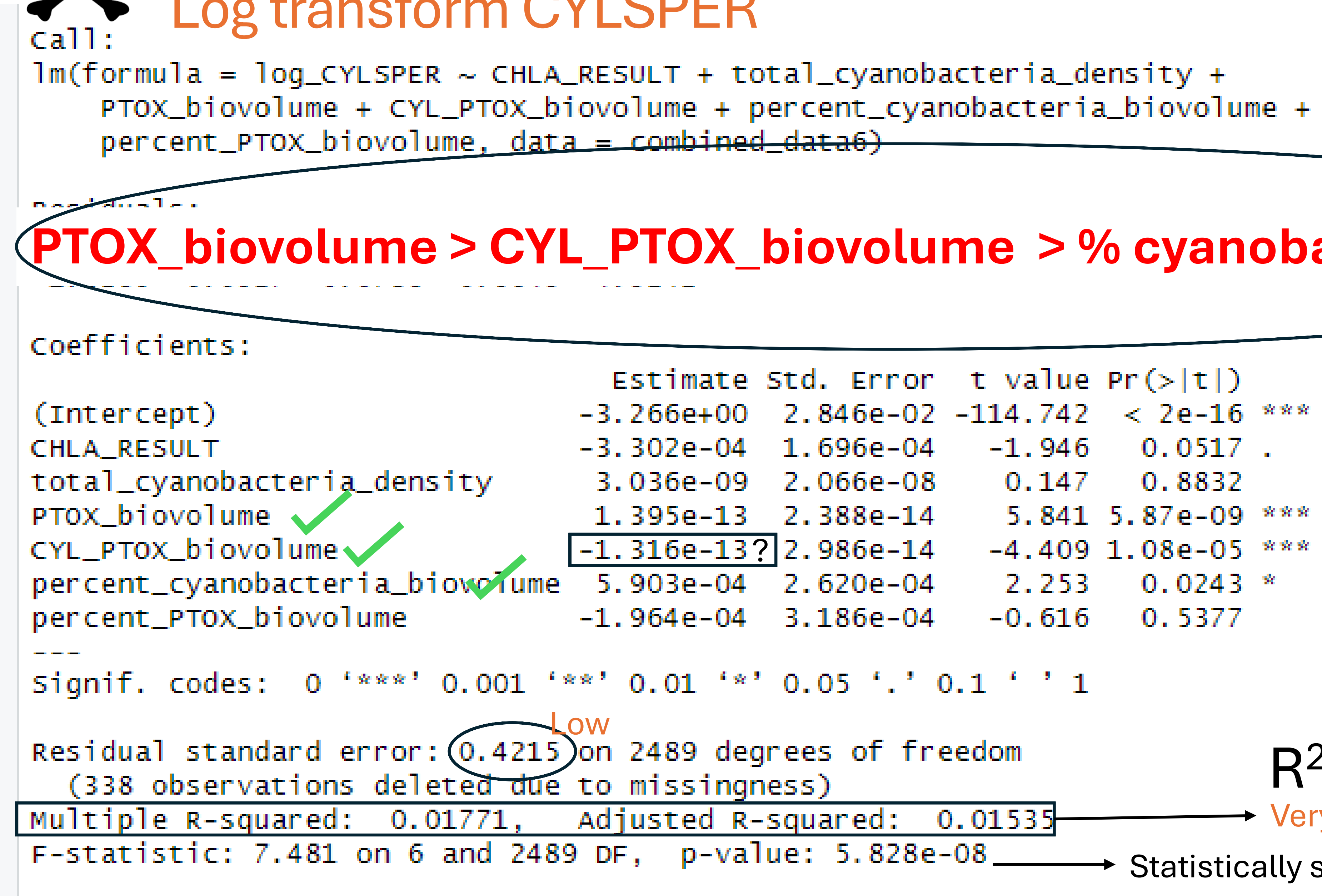
$R^2 = 12.13\%$

Moderate model fit.

Statistically significant

CYLINDROSPERMOPSINS….

Log transform CYLSPER



```
Call:
lm(formula = log_CYLSPER ~ CHLA_RESULT + total_cyanobacteria_density + 
    PTOX_biovolume + CYL_PTOX_biovolume + percent_cyanobacteria_biovolume + 
    percent_PTOX_biovolume, data = combined_data6)

Coefficients:
                                  Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                     -3.266e+00  2.846e-02 -114.742  < 2e-16 ***
CHLA_RESULT                     -3.302e-04  1.696e-04   -1.946   0.0517 .  
total_cyanobacteria_density      3.036e-09  2.066e-08    0.147   0.8832    
PTOX_biovolume                   1.395e-13  2.388e-14    5.841 5.87e-09 ***
CYL_PTOX_biovolume              -1.316e-13  2.986e-14   -4.409 1.08e-05 ***
percent_cyanobacteria_biovolume  5.903e-04  2.620e-04    2.253   0.0243 *  
percent_PTOX_biovolume          -1.964e-04  3.186e-04   -0.616   0.5377    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4215 on 2489 degrees of freedom
  (338 observations deleted due to missingness)
Multiple R-squared:  0.01771,	Adjusted R-squared:  0.01535 
F-statistic: 7.481 on 6 and 2489 DF,  p-value: 5.828e-08
```

PTOX_biovolume > CYL_PTOX_biovolume > % cyanobacteria biovolume

$R^2 = 1.535\%$

Very weak model fit.

Statistically significant

Research questions

- Which of the following biological indicators is the best predictor of cyanotoxins concentrations in U.S. lakes: chlorophyll-a (chl-a) concentration, total phytoplankton biovolume, or cyanobacterial abundance?

- **MIC → Chl-a > % cyanobacteria biovolume > total_cyanobacteria_density**
- **CYL → PTOX_biolume > CYL_ PTOX_biolume > % cyanobacteria biovolume**

Research questions

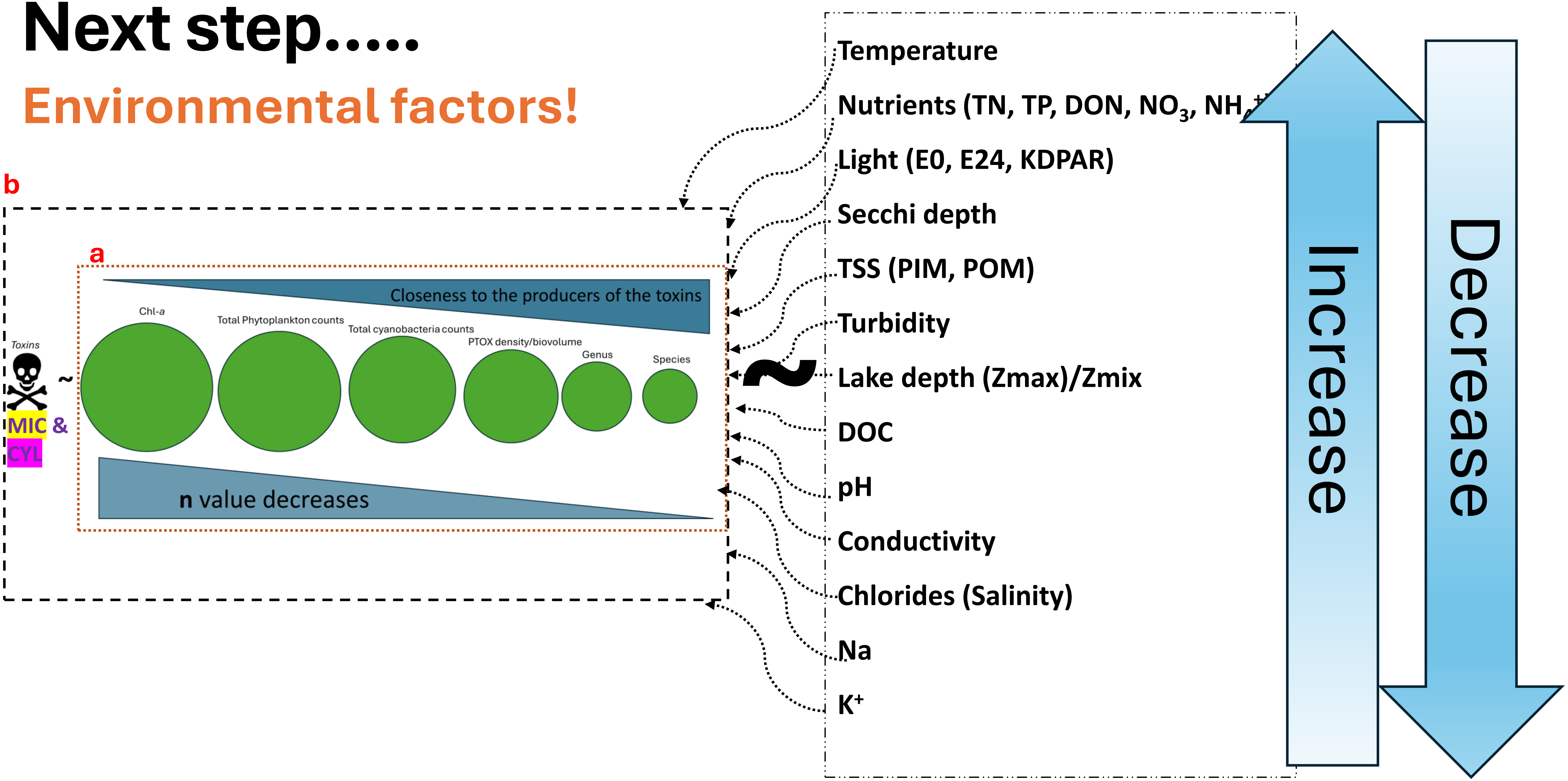
- Are the same biological indicators equally predictive of Microcystin and Cylindrospermopsin concentrations in U.S. lakes?

MIC ~ Chl-a + % cyanobacteria biovolume + total_cyanobacteria_density

CYL ~ PTOX_biovolume + CYL_PTOX_biovolume + % cyanobacteria biovolume

Next step.....

Environmental factors!



Temperature

Nutrients (TN, TP, DON, NO_3, NH_4^+)

Light (E0, E24, KDPAR)

Secchi depth

TSS (PIM, POM)

Turbidity

Lake depth (Zmax)/Zmix

DOC

pH

Conductivity

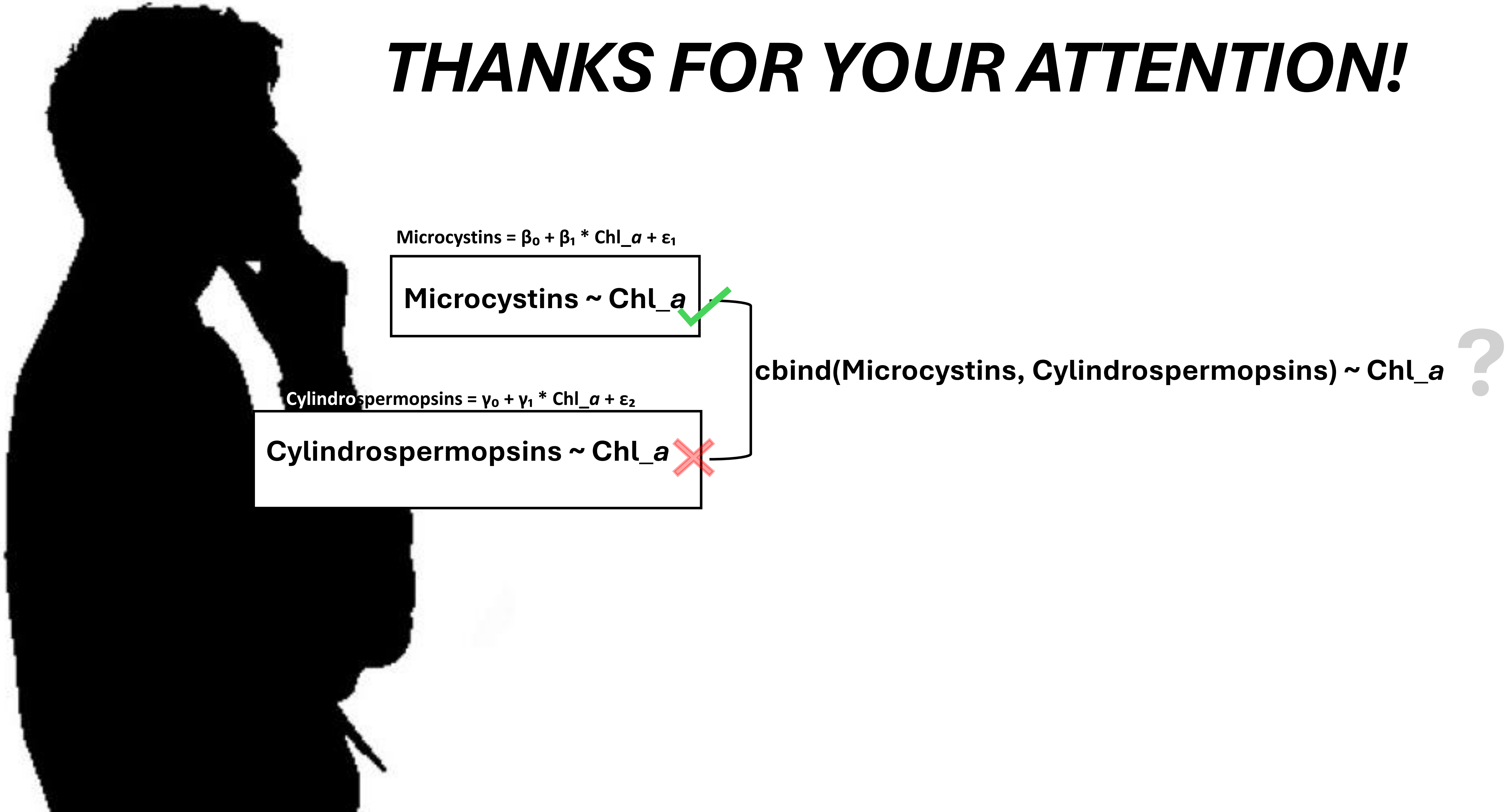
Chlorides (Salinity)

Na

K^+

- ➢ More robust models

THANKS FOR YOUR ATTENTION!



Microcystins = $\beta_0 + \beta_1$ * Chl_*a* + ε_1

Microcystins ~ Chl_*a* ✓

Cylindrospermopsins = $\gamma_0 + \gamma_1$ * Chl_*a* + ε_2

Cylindrospermopsins ~ Chl_*a* ✗

cbind(Microcystins, Cylindrospermopsins) ~ Chl_*a* ?