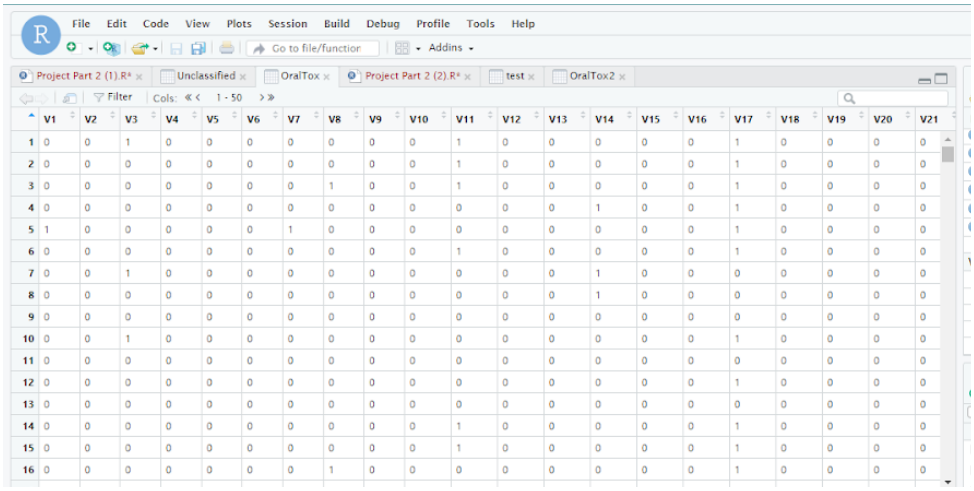
Team 5

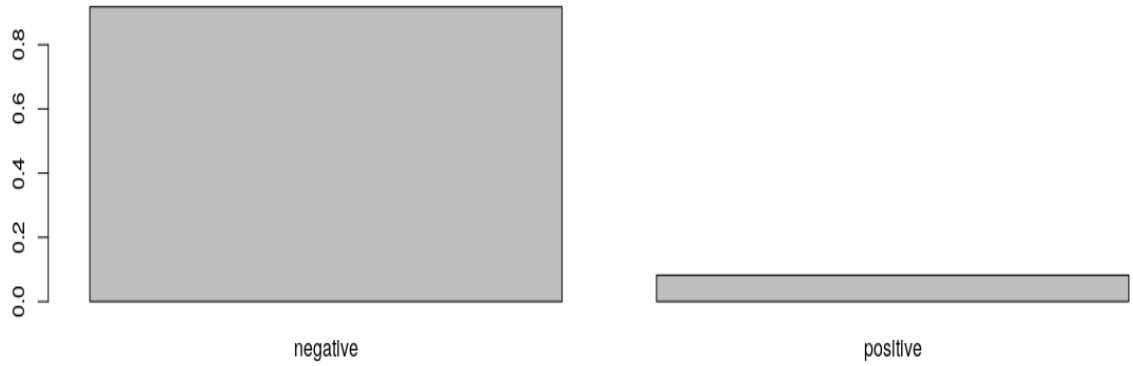
Final Report of "Let's Find Toxic Chemicals"

In recent years, the assessment of drugs and chemicals has changed from the conventional "in vivo" method, which largely relied on animal testing, to other forms that don't rely on these practices. The alternative form of testing the toxicity of these chemicals involves "in silico" methods. It is essential to develop proper models that can effectively differentiate toxic and nontoxic chemicals through the use of biomolecular fingerprints. Therefore, our goal is to identify and classify the connection between molecules' toxicity levels and their binary molecular fingerprints. This is complicated through indexing, understanding, and testing our hypotheses on our data frame.

Our imported dataset consisted of 8,992 chemicals, their binary molecular fingerprint, and a classification of each chemical, indicating whether or not it is toxic. To prepare our data for analysis, we separated each consecutive number of the binary molecular fingerprint into their own respective columns through indexing. So, the first 1024 columns of the dataset would make up the fingerprint, each of these columns composed only of 1s and 0s (binary variables). Then, the final (1025th) column indicates the classification of the chemical. If the classification is positive, that indicates that the chemical is very toxic (LD_{50} lower than mg/kg), while if the classification is negative, that indicates that the chemical is not very toxic (LD_{50} greater than or equal to 2,000 mg/kg).



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21
1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0



As seen in the model above, there is a significant difference between the proportion of non-toxic and toxic chemicals in our dataset. More specifically, there were 8,251 non-toxic chemicals and 741 toxic chemicals, representing 92% and 8% of our dataset, respectively. Due to this marked difference, we wanted to better understand the relationship between a chemical's molecular footprint and its toxicity. Thus, we came up with several hypotheses:

- Hypothesis 1: The increased presence of biomolecular components (structures) is associated with higher toxicity levels.
- Hypothesis 2: Certain binary molecular components are more important than others in determining toxicity levels. (An example of this might be if the 10th binary molecular component is highly correlated with toxicity).

To test our first hypothesis, we added an additional column that summed up the total for each row. We were quickly able to disprove our assumption that a higher sum would indicate a higher level of toxicity. For instance, in the table below, we can clearly see that some negative classifications returned higher values than the positive classifications.

V1019	V1020	V1021	V1022	V1023	V1024	V1025	Sum
0	0	0	0	0	0	negative	183
0	0	0	0	0	0	negative	74
1	0	0	0	0	0	negative	54
0	0	0	0	0	0	positive	18
0	0	0	0	0	0	negative	18
0	0	0	0	0	0	negative	18
0	0	0	0	0	0	negative	18
0	0	0	0	0	0	positive	18
0	0	0	0	0	0	negative	18
0	0	0	0	0	0	positive	18
0	0	0	0	0	0	negative	17
0	0	0	0	0	0	positive	18
0	0	0	0	0	0	negative	18
0	0	0	0	0	0	negative	18
0	1	0	0	0	0	negative	18
0	0	0	0	0	0	negative	18

We had difficulty with visualizing our data during the exploratory analysis process since our data did not have precise variables. Due to the lack of findings in our exploratory analysis, we relied more heavily on classification tests to consider our second hypotheses.

We used a general linear regression model to better understand our data and to identify any patterns that it may have. The goal was to determine whether the presence of any specific structures in the fingerprint carried more weight in determining toxicity levels than others (Hypothesis 2). Using this information, we would like to then be able to make predictions about chemicals' toxicity if given a whole new unknown dataset. To create this model, we used a train-test split to evaluate the performance of our machine learning algorithm.

Our model was able to evaluate the dataset and make predictions on whether a molecular binary fingerprint would return positive or negative with 86.96% accuracy. In the table below, the model was run, and although we do see that the model was pretty good at predicting the chemicals' toxicity, it was better at accurately predicting negatives than positives.

	Reference	
Prediction	Negative	Positive
Negative	2986	138
Positive	331	141

Since our model was relatively successful, this means that we would be able to plug in new binary molecular fingerprints and be able to predict the likelihood of them being toxic or non-toxic. The success of our model also indicates that some structures are more important than others in determining toxicity. We can make this assumption because had there not been any pattern, our accuracy would have returned at about 50%, indicative of random guessing. Therefore, our second hypothesis, although not explicitly proven, does seem to be correct. There are a few limitations to our data and our study. For instance, our data did not give us the molecular names for each fingerprint. The general linear regression model also did not return accurate coefficients for our data. This means that our model confirms that certain binary components matter more than others, but it does not tell us which binary components they are.

We were able to disprove our first hypothesis, finding that the increased presence of biomolecular components is not necessarily associated with higher toxicity levels. We were also able to provide evidence supporting our second hypothesis, therefore finding that certain biomolecular components are more important than others in determining toxicity levels. To further our initial research, we may want to apply our model to more data to further test its accuracy. Future research might also compare models by using different "families." Examples include quasibinomial or poisson.

Acknowledgment:

Alyssa Humphreys: Predictive analysis/modeling and Github repository

Sergio Chavez: exploratory analysis

Joelle Kenty: project paper

Yunseo Chang: presentation slides, research for toxicity (background information), limitations

Jennifer Garza: research of data frame and about Binary Molecular Footprints

Pablo Gomez: team organization and presentation slides

Bibliography

D. Ballabio, F. Grisoni, V. Consonni, R. Todeschini (2019), Integrated QSAR models to predict acute oral systemic toxicity, Molecular Informatics, 38, 180012; doi: 10.1002/minf.201800124

Team 5

Codes Used (also in Github):

Code For Importing and Indexing:

```
library(tidyr)

head(qsar_oral_toxicity)

names(qsar_oral_toxicity)

names(qsar_oral_toxicity) <- c("x1")

library(stringr)

OralTox2 <- as.data.frame(str_split_fixed(qsar_oral_toxicity$x1, ";", 1025))
```

Code For Adding a Sums Column to DF:

Had to get rid of last columns with negative/positive classifications so that the data was entirely numeric for sums code.

```
Unclassified <- OralTox2[-c(1025),drop=FALSE]

rowSums(sapply(No_Low_No_Intergenic_snpeff.scores[, c(20:29, 45)],

               function(x) as.numeric(as.character(x))))

OralTox2["Sum"] <-rowSums(sapply(Unclassified[, c(1:1024)],

                 function(x) as.numeric(as.character(x))))
```

Code For Clustering:

```
library(tidyverse)

clust_dat <- qsar_oral_toxicity %>% select(V1, V2)

sil_width <- vector()

 for (i in 2:10) {

   kms <- kmeans(clust_dat, centers = i)

   sil <- silhouette(kms$cluster, dist(clust_dat))

   sil_width[i] <- mean(sil[, 3])

 }
```

Code for Barplot

Team 5

prop.table(table(qsar_oral_toxicity$V1025))

Code for number of Postive and Negative Chemicals

```
qsar_oral_toxicity %>% group_by(V1025) %>% summarize(number_rows = n())
```

Code for Linear Regression

```
library(stringr)
OralTox <- as.data.frame(str_split_fixed(qsar_oral_toxicity$x1, ";", 1025))
OralTox$V1025 <- factor(OralTox$V1025, levels=c("negative","positive"))
split_row <- round(nrow(OralTox)*.60, 0)
train <- OralTox[1:split_row, ]
test <- OralTox[(split_row+1):nrow(OralTox), ]
model <- glm(V1025~., family="binomial", data=train)
summary(model)
p <- predict(model, newdata=test, type="response")
classes <- ifelse(p>=0.5, 1, 0)
p_class <- factor(classes, labels=c("negative","positive"))

confusionMatrix(data=p_class, reference=test$V1025)
```