PCA Analysis of Drug Data

Advaith Cheruvu

10/12/2021

Importance of Study

The study of pharmacology has lead to the development of new drugs, providing treatment for patients and combating deadly disease. However, the longstanding issue of pharmacology is that the development of new innovative drugs has been slowed down due to the sheer number of potential candidates and lack of a consistent pattern when discovering new drugs. Aside from smaller molecules, most new drugs present themselves in serendipitous ways. Thus, any pattern found among successful drugs would prove to be invaluable to pharmacologists. In this document, I will be analyzing 14 factors that contributed to the market success of 1270 drugs to determine which factors played the largest role. The most important factors can then be used by pharmacologists for the development of new drugs.

Pharmacology

Pharmacology is a branch of science that deals with the study of drugs and their effects on living systems. Drugs can have many effects and properties in the human body. Pharmacologists need to ensure that drugs don't cause unwanted effects in regions of the body that the drug should not be in. For example, antidepressants should cause their intended effect in the brain, and not cause damage to any other regions of the body. This study of how drugs move throughout a living system is called pharmacokinetics.

Pharmacokinetics

Pharmacokinetics, as said earlier, deals with the movement of drugs in a living system. Drugs move around the body based on the ADME processes. ADME is the acronym for absorption, distribution, metabolism, and excretion. These processes dictate how drugs move around the body, and ensure that the drugs cause their intended effects in desirable locations. Drugs that fail to reach their intended location risk serious damage, making pharmacokinetics crucial to the safety and effectiveness of a drug.

ADME: Absorption

Absorption is the process by which a drug enters the bloodstream based on the administration of the drug. Drugs can be administered orally or intravenously. Oral drugs must first travel through the gastrointestinal tract, and then may be absorbed by the lining of the stomach to reach the bloodstream. This may be useful for individuals to self-administer or if individuals have many different drugs that they must take. Intravenously administered drugs are immediately in the blood stream after administration and are typically administered by medical professionals. This form of administration is more applicable when a patient has an acute reaction or needs the immediate effect of a drug.

ADME: Distribution

Once a drug enters the blood stream, they circulate around the body and build up in various tissues and organ. This buildup is reversible and drugs are intentionally designed so that they build up in the regions where the drugs is intended to build up. Drugs enter these tissues and organs with the help of proteins that move the drug from the blood stream to the cell (uptake) and out of the cell back to the blood stream (efflux). These proteins that transfer drugs are called plasma proteins.

ADME: Metabolism

Once a drug enters the general region of cells and carries out the desired functions, metabolism processes become predominant. Metabolism, in the context of pharmacology, is the conversion from lipophilic xenobiotic compounds (drugs that dissolve in fats) to hydrophilic metabolites (waste compounds that dissolve in water). Usually, the metabolism of drugs involves enzymes which decrease the activation energy for reactions. Specifically, these reactions produce the waste products from the drugs. These waste products are removed from the cell and moved to the blood stream with the help of plasma proteins.

ADME: Excretion

Once the waste products are moved to the blood stream, they undergo excretion. Excretion is the irreversible loss of a substance (usually a waste product) from a biological system. The blood is filtered by the kidneys, producing urine that eventually is expelled from the body. The waste products can also be filtered by the liver where it is excreted in feces. Other common forms of excretion include sweat, tears, and breath.

Pharmacodynamics

Where pharmacokinetics deals with the movement of a drug, pharmacodynamics deals with the biological response of a drug. This usually occurs after a drug has been distributed. The biological response is determined by the chemistry and dosage of a drug. Though the chemistry of a drug is important for plasma protein binding, the chemistry of a drug influences the biological response more than the plasma protein binding, since there are many plasma proteins that can bind to various types of drugs. The chemistry data of various successful drugs is contained in the dataset that I will be using.

Principal Component Analysis

As mentioned earlier, the aim of this document is to analyze which of the 14 factors most contribute to the market success of 1270 various drugs. To do so, I will be doing Principal Component Analysis (PCA). PCA consolidates many different variables into either one, two, or three uncorrelated variables so that plots can be constructed. PCA removes redundant highly correlated variables and combines them into one variable so that the most information can be retained. It should be noted that PCA only works with numerical data. In the context of the drug data, I will be consolidating the 14 variables for drug data into only a few variables. There are 2 main methods to conduct PCA, including singular value decomposition (SVD) and eigenfunctions. I will be using SVD since I found it more accurate for this dataset, though either method is fairly accurate over any dataset. Singular value decomposition involves combining variables in a successive manner. Though SVD involves some complex linear algebra and matrices manipulation, it can be thought of as combining correlated variables.

Setting up Workspace and Installing Packages

The code below shows the preparation of the workspace and cleaning up the environment, as well as installing and loading packages.

```
#
# Clean up and set up
rm(list=ls())
setwd("/Users/advai/Documents/DSFS")
#
# install and load libraries
library(ggplot2)
source("myfunctions.R")
```

Loading and Looking Drug Data

The data seems to be clean and ready to work with.

```
#
# load drugs data
drugs <- read.csv(file = "C:\\Users\\advai\\Documents\\DSFS\\drugs.csv",header=TRUE)
#
# looking at the data
names(drugs)
```

```
##  [1] "X"              "logS"           "logSpH7"        "logP"
##  [5] "logD"           "X2C9pKi"        "hERGpIC50"      "BBB"
##  [9] "Pgpcategory"    "MW"             "HBD"            "HBA"
## [13] "TPSA"           "Flexibility"    "RotatableBonds"
```

```
summary(drugs)
```

```
##       X                   logS            logSpH7            logP        
##  Length:1270        Min.   :-2.750   Min.   :-2.750   Min.   :-5.0810  
##  Class :character   1st Qu.: 1.770   1st Qu.: 1.665   1st Qu.: 0.6122  
##  Mode  :character   Median : 2.755   Median : 2.611   Median : 2.2770  
##                     Mean   : 2.902   Mean   : 2.759   Mean   : 2.0912  
##                     3rd Qu.: 3.920   3rd Qu.: 3.804   3rd Qu.: 3.5545  
##                     Max.   : 9.765   Max.   :10.100   Max.   : 8.6360  
##       logD             X2C9pKi         hERGpIC50           BBB         
##  Min.   :-5.4780   Min.   :3.394   Min.   :-1.602   Min.   :-2.40000  
##  1st Qu.:-0.3665   1st Qu.:4.276   1st Qu.: 3.744   1st Qu.:-1.07800  
##  Median : 1.1280   Median :4.728   Median : 4.539   Median :-0.52290  
##  Mean   : 1.1240   Mean   :4.694   Mean   : 4.440   Mean   :-0.49389  
##  3rd Qu.: 2.5935   3rd Qu.:5.043   3rd Qu.: 5.301   3rd Qu.: 0.06151  
##  Max.   :12.8500   Max.   :6.374   Max.   : 7.977   Max.   : 1.44000  
##   Pgpcategory            MW               HBD              HBA        
##  Min.   :0.0000   Min.   :  31.01   Min.   : 0.000   Min.   :  0.000  
##  1st Qu.:0.0000   1st Qu.: 254.32   1st Qu.: 1.000   1st Qu.:  3.000  
##  Median :0.0000   Median : 328.50   Median : 2.000   Median :  5.000  
##  Mean   :0.4323   Mean   : 387.33   Mean   : 2.451   Mean   :  6.514  
##  3rd Qu.:1.0000   3rd Qu.: 428.60   3rd Qu.: 3.000   3rd Qu.:  7.000  
##  Max.   :1.0000   Max.   :4492.00   Max.   :63.000   Max.   :115.000  
##       TPSA          Flexibility     RotatableBonds  
##  Min.   :   0.00   Min.   :0.0000   Min.   :  0.000  
##  1st Qu.:  42.72   1st Qu.:0.1250   1st Qu.:  3.000  
##  Median :  72.72   Median :0.2064   Median :  5.000  
##  Mean   :  95.55   Mean   :0.2275   Mean   :  6.797  
##  3rd Qu.: 111.50   3rd Qu.:0.3000   3rd Qu.:  8.000  
##  Max.   :1903.00   Max.   :0.9091   Max.   :187.000
```

```
str(drugs)
```

```
## 'data.frame':    1270 obs. of  15 variables:
##  $ X             : chr  "ABACAVIR" "ABARELIX" "ACAMPROSATE" "ACARBOSE" ...
##  $ logS          : num  3.23 2.15 6.36 5.4 3.49 ...
##  $ logSpH7       : num  1.93 3.89 3.81 5.13 2.95 ...
##  $ logP          : num  1.39 1.39 -1.92 -2.5 1.71 ...
##  $ logD          : num  0.41 4.289 -1.844 -0.917 -0.09 ...
##  $ X2C9pKi       : num  4.71 5.03 3.87 5.22 4.4 ...
##  $ hERGpIC50     : num  5.55 1.69 3.55 2.63 4.7 ...
##  $ BBB           : num  -0.441 -1.072 -0.47 -1.586 -0.15 ...
##  $ Pgpcategory   : int  1 1 0 1 1 0 0 0 0 0 ...
##  $ MW            : num  286 1416 181 646 336 ...
##  $ HBD           : int  3 13 2 14 3 2 2 1 2 2 ...
##  $ HBA           : int  7 28 5 19 6 3 7 2 6 3 ...
##  $ TPSA          : num  101.9 425 83.5 321.2 87.7 ...
##  $ Flexibility   : num  0.167 0.453 0.5 0.192 0.458 ...
##  $ RotatableBonds: int  4 48 5 9 11 2 3 0 6 1 ...
```

Creating New Variables

By converting this data frame into a matrix, I can work with just numbers and replace the row names with the names of the drugs. This is done because PCA only works with numerical data.

```
#
# converting this data to a numeric matrix
drugs.matrix <- data.matrix(drugs[2:15])
rownames(drugs.matrix) <- drugs$X
```

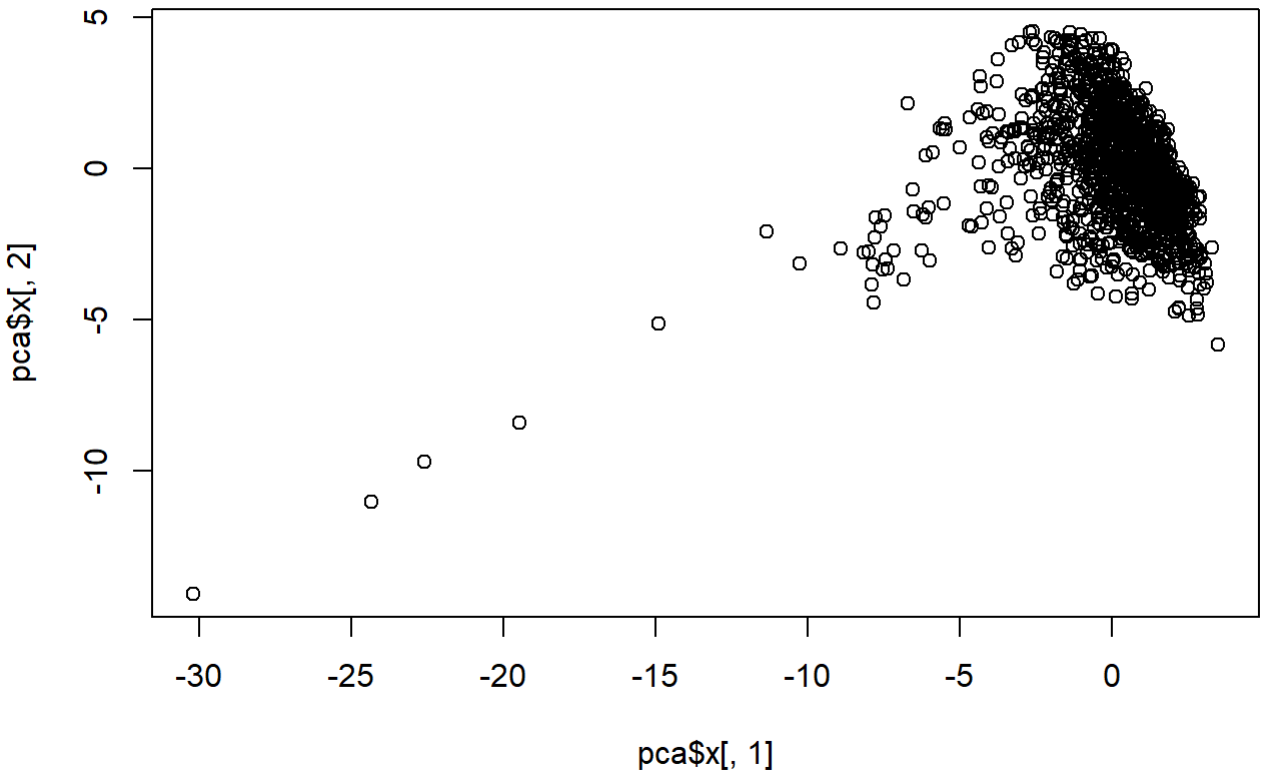
Singular Value Decomposition

I will be using a command called “prcomp” because it executes SVD. I can look at the principal components with the “summary” command. I then build a plot of PCA1 and PCA2 to examine where the variance is visually. It seems like there are a few outliers, but most of the PCA data falls within the top right corner of the plot. The variance is more for PCA1 since the spread of the data points is more horizontal than vertical. The scree plot confirms this notion since it shows that PCA1 has more variance than PCA2, though both components hold the majority of the variance for the whole data set (40.8% and 28.7% variance respectively).

```
#
# prcomp for single value decomposition
pca <- prcomp(drugs.matrix, scale=TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.3891 2.0055 1.1835 0.96972 0.76392 0.60787 0.52723
## Proportion of Variance 0.4077 0.2873 0.1000 0.06717 0.04168 0.02639 0.01986
## Cumulative Proportion  0.4077 0.6950 0.7950 0.86221 0.90390 0.93029 0.95014
##                            PC8     PC9    PC10    PC11    PC12   PC13    PC14
## Standard deviation     0.49475 0.36409 0.35536 0.29070 0.23486 0.2244 0.06595
## Proportion of Variance 0.01748 0.00947 0.00902 0.00604 0.00394 0.0036 0.00031
## Cumulative Proportion  0.96763 0.97710 0.98612 0.99215 0.99609 0.9997 1.00000
```

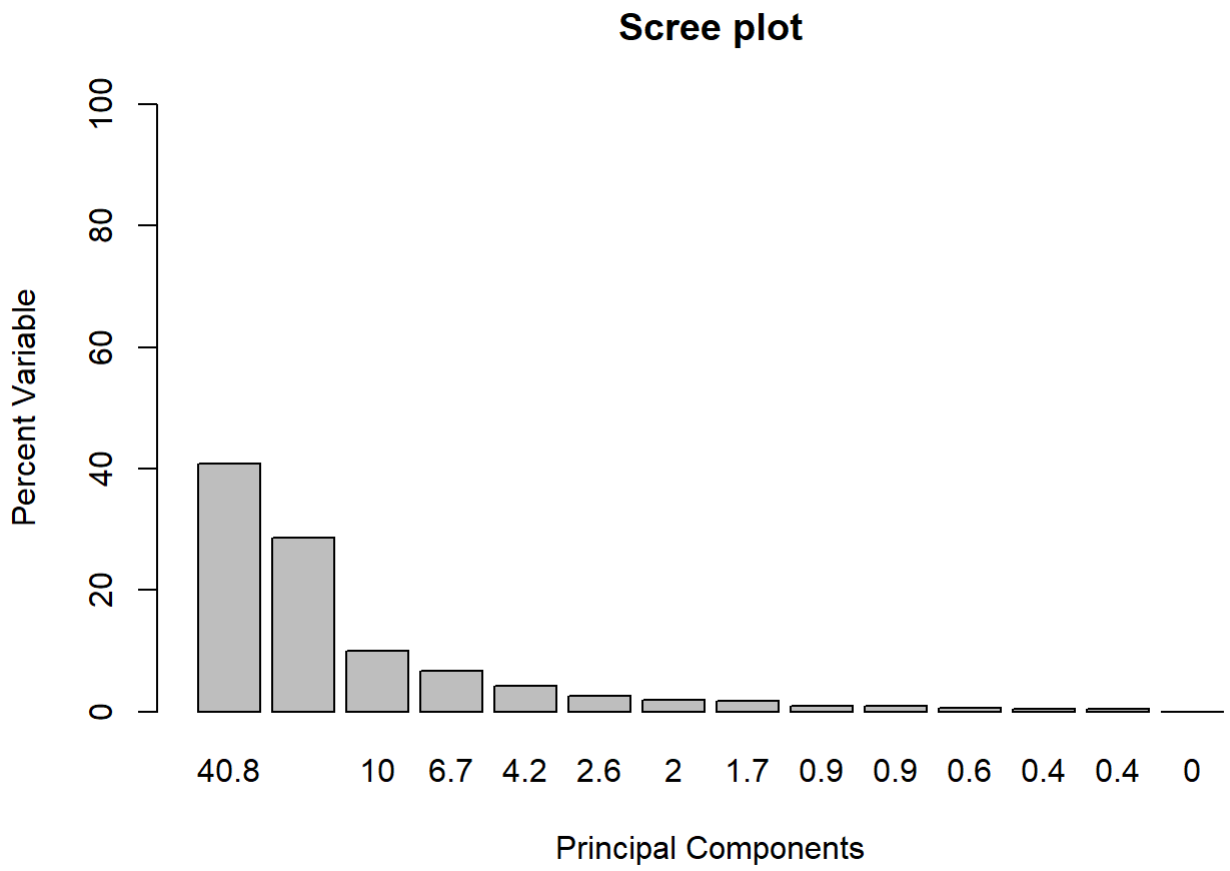
```
#
# Plot of PCA
plot(pca$x[,1],pca$x[,2])
```




```
#
# Scree Plot
pca.variance <- pca$sdev^2
pca.variance.per <-round(pca.variance/sum(pca.variance)*100,1)
pca.variance.per
```

```
##  [1] 40.8 28.7 10.0  6.7  4.2  2.6  2.0  1.7  0.9  0.9  0.6  0.4  0.4  0.0
```

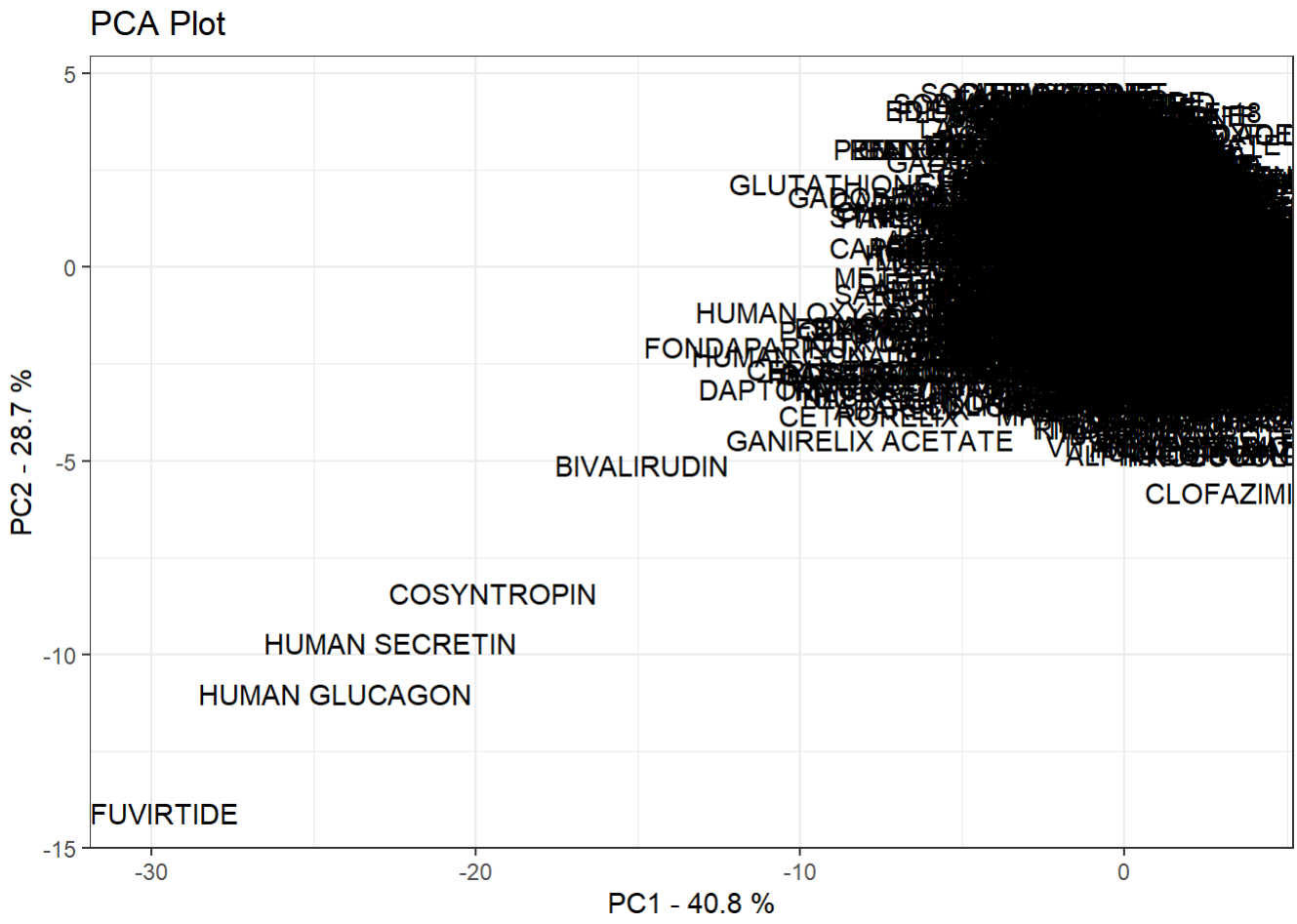
```
barplot(pca.variance.per, main="Scree plot", xlab="Principal Components", ylab="Percent Variabl
e", ylim=c(0,100),names.arg=pca.variance.per)
```



Plotting PCA Data

I combined the drug names with the PCA data by putting the PCA data into a data frame. This helps to show specifically which drugs are better determined by the principal components and which drugs are outliers. I constructed a ggplot that puts the name of a drug on a plot with PC1 as the X axis and PC2 as the Y axis. This plot tells me that most of the drugs are contained in the same region of the graph (top right), meaning that PC1 and PC2 do a sufficient job of maintaining the information from the original variables.

```
#
# Putting pca data into a data frame
pca.data <- data.frame(Sample = rownames(pca$x), X=pca$x[,1], Y=pca$x[,2])
#
# ggplot of pca.data
ggplot(data=pca.data,aes(x=X, y=Y, label=Sample)) + geom_text() + xlab(paste("PC1 - ", pca.varia
nce.per[1], " %", sep="")) + ylab(paste("PC2 - ",
pca.variance.per[2], " %", sep="")) + theme_bw() + ggtitle("PCA Plot")
```



Loading Scores

Now that it has been established that PC1 information, it is important to find out which of the original variables contribute the most to PC1. This would reveal the most crucial factor in drug development. I produce the loading scores of PC1. The loading scores of PC1 indicate that “TPSA”, “HBD”, and “HBA” are the top 3 most important variables for drug development, as they contribute the most to PC1. These three variables are topological polar surface area, hydrogen-bond donor, and hydrogen-bond acceptor, respectively.

```
#
# loading scores (which factors contribute to PC1)
loading_scores <- pca$rotation[,1]
loading_scores
```

```
##           logS        logSpH7           logP           logD       X2C9pKi
##    -0.14284891    -0.28348164     0.25755166     0.10506898     0.01090487
##      hERGpIC50            BBB    Pgpcategory             MW            HBD
##     0.28763642     0.24263729    -0.11858669    -0.30963719    -0.38317308
##            HBA           TPSA    Flexibility RotatableBonds
##    -0.38305610    -0.39204126    -0.14388054    -0.31975612
```

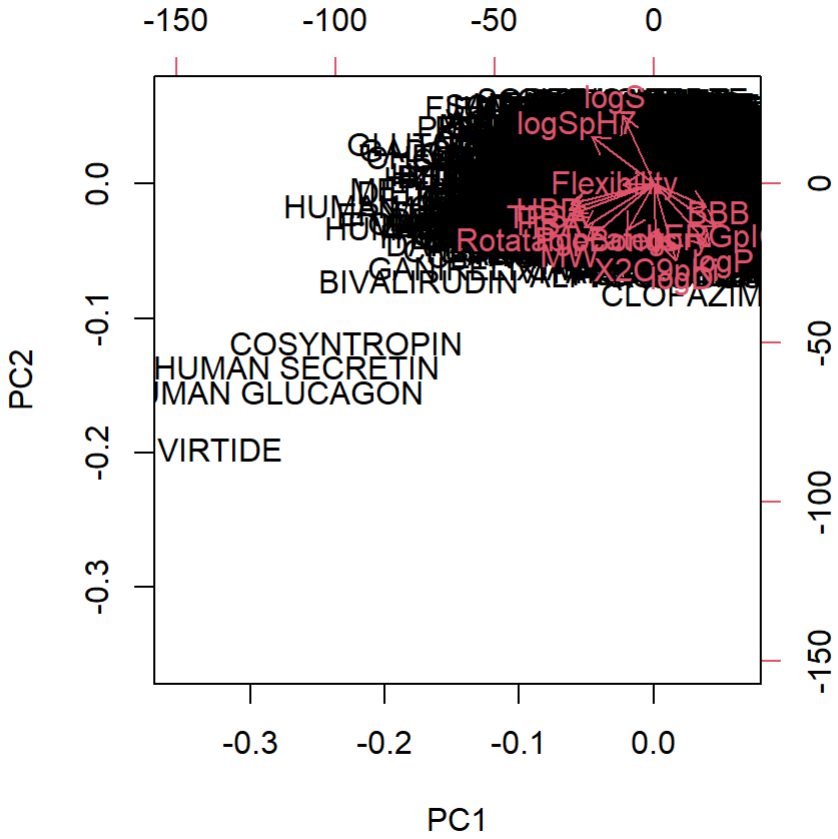
```
drug_scores <- abs(loading_scores)
drug_scores_ranked <- sort(drug_scores, decreasing = TRUE)
top3 <- names(drug_scores_ranked[1:3])
top3
```

```
## [1] "TPSA" "HBD"  "HBA"
```

Biplot

A biplot the original variables on top of the PCA plot. Based on the biplot, the original variables confirm that PC1 is more varied than PC2 since most of the arrows vary horizontally rather than vertically. Also, the arrows somewhat account for the outliers, since many of the original variables point towards the outliers.

```
#
# Biplot of PCA
biplot(pca)
```




```
# EOF (End of File)
```

Conclusion

Knowing that the topological polar surface area and hydrogen bonds of a drug are the most important factors in a successful drug, data scientists can look into what specific values for these variables are important for plasma protein binding and the actual reactions in localized parts of the body. These patterns can then be applied to the discovery of new drugs. Pharmacologists can use the generalized values for topological polar surface area and hydrogen bonds to restrict their search for new drugs, accelerating the access to new medicine and potentially saving thousands of lives.

Work Cited

"2.1 What Is Pharmacology?" Department of Health | 2.1 What Is Pharmacology?, 2004, www1.health.gov.au/internet/publications/publishing.nsf/Content/drugtreat-pubs-front6-wk-toc$_{drugtreat\text{-}pubs\text{-}front6\text{-}wk\text{-}secb}$drugtreat-pubs-front6-wk-secb-2~drugtreat-pubs-front6-wk-secb-2-1.

Doogue, Matthew P, and Thomas M Polasek. "The ABCD of Clinical Pharmacokinetics." Therapeutic Advances in Drug Safety, SAGE Publications, Feb. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC4110820/.

"The Drug Development Process." U.S. Food and Drug Administration, FDA, 2021, www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process.

Perlato, Andrea. "Principal Component Analysis." Andrea Perlato, 2020, www.andreaperlato.com/theorypost/principal-component-analysis/.

"Pharmacokinetics and Pharmacodynamics (PK/PD)." PK / PD and Clinical Pharmacology Consultants, 9 Aug. 2021, www.nuventra.com/services/pharmacokinetics-pharmacodynamics/.

Xenotech. "Absorption, Distribution, Metabolism, and Excretion (ADME)." SEKISUI XenoTech, 4 Nov. 2020, www.xenotech.com/nonclinical-studies/adme/.