Investigating Commerical Success in Drugs Through Principal Component Analysis

Hank Magan

10/12/2021

Abstract

Certain drugs experience widespread, commerical success, despite treating a variety of different ailments. The fact that certain drugs outperform others on the market is uncoincidental. As demonstrated in this report, certain molecular properties, such as topological polar surface area (TPSA), the presence of hydrogen bond donors (HBD), and the presence of hydrogen bond acceptors (HBA), influence the commerical success of a drug. Results are calculated using principal component analysis, or PCA. The dataset used contains data about 1,270 different commerically-successful drugs, with 14 different descriptors for each drug. It is through PCA that all of these descriptors can be observed and modeled at once; this process is known as dimension reduction. In the culmination of the report, it is concluded that the traits found to be influential in commerical success are directly correlated to ADME (the disposition of drugs within the body) as well as Lipinski's rule of five.

Introduction

Pharmacokinetics and Pharmacodynamics

Within the field of pharmacology, there are two main studies: pharmacokinectics and pharmacodynamics. From a high level standpoint, pharmacokinectics is the study of how an organism affects a drug. It is studied alongside pharmodynamics (which will be discussed in-depth shortly) in order to determine a variety of descriptors, such as drug dosing in commercial settings. When a drug (or any substance) is consumed, highly specific metabolic reactions occur, digesting and transforming the consumed drug. This is accomplished via enzymes, which, again, are highly specific. For example: lactase is responsible for digesting lactose (milk sugar), cellulase is responsible for digesting cellulose (plant cell wall), etc. Of course, the exact way that this digestion (or "breaking down") process occurs varies from enzyme to enzyme, and therefore drug to drug as enzymes are specific to their substrates. Due to this very high specificity, drug selection and dosage must be carefully considered to achieve a desired effect. With that said, there are a few general steps that can be applied to better understand the digestion process, even if it is unique to different substances. These steps are characterized by the acronym ADME, or absorption, distribution, metabolism, and excretion, and are explained in greater detail below.

Absorption When a substance enters the body, it first engages in absorption, or the process that it takes to enter the bloodstream. Absorption often occurs via mucous surfaces, such as the digestive tract. Of course, certain drugs will not absorb well in the digestive tract, and might have a greater clinical effect when administered intravenously, intramuscularly, or through inhalation (among others). Naturally, absorption is critically involved in determining a substance's bioavailability, or the fraction of the substance that actually reaches the circulatory system. Bioavailability is another very important factor to consider when determining dosage, as not all of the administered substance will reach the intended organ(s).

Distribution Once the drug enters the bloodstream, the process of distribution begins. Distribution simply describes the transfer of a drug from one location in the body to another. Distribution relies on a multitude of factors, including: vascular permeability, blood flow, and the perfusion rate of the tissue that the drug is intended to enter. The permeability of the tissue is particularly important. Some tissues, such as the blood-brain barrier (BBB), are naturally highly selective. This should be taken into consideration, as if a drug cannot exit the bloodstream and enter the desired tissues, problems may arise. Additionally, some drugs have the ability to bind with certain proteins founds in blood plasma, which can affect the drug's efficiency. Drugs that are less bound to plasma proteins can more efficiently traverse cell membranes.

Metabolism Metabolism describes all of the life-sustaining chemical reactions within the body. These reactions are responsible for a variety of life processes, one of which is the biotransformation of drugs. As soon as a compound enters the body, deconstruction begins. With drugs, a majority of small-molecule metabolism occurs in the liver and is carried out by redox enzymes. When a compound is metabolizes, it does not simply break down and disappear—rather, it is converted into new, smaller compounds named metabolites. The metabolites that are produced via metabolism may or may not be pharmacologically active. When they are active, their effects should be considered and recognized as potential side effects to the drug in question (in some cases, the metabolites may be even more active than the parent drug). When they are not, their presence can dilute the effects of the parent drug, which should also be considered.

Excretion Lastly, excretion is the process by which metabolic waste is removed from an organism. In humans, this process occurs primarily in the kidney, the liver and gut, and the lungs. Excretion is a crucial process in the disposition of a drug. Without it, the accumulation of metabolic waste (carbon dioxide, water, salts, urea, uric acid, etc) could have an adverse effect on a variety of processes in the body. The kidney is the most important site in excretion, as it is where urine is processed. Fecal excretion is the process initiated in the liver, in which waste products are packaged along with feces. The lungs are involved in excretion through the release of harmful gases, such as anesthetic gases.

While pharmacokinetics describes the body's effect on a substance, pharmacodynamics describes a substance's effect on the body. Together, the two influence dosing, benefit, and side effects of specific drugs. The drugs analyzed in this report have already been recognized as commercially successful drugs, and therefore will be examined to determine what factor has the most influence on commercial success in a drug. This is not a foreign concept; in fact, Christopher A. Lipinski developed a set of rules dubbed "Lipinski's rule of five" to determine if a drug is likely to be orally active or not. This rule describes a set of molecular properties which are important to the processes of ADME. The rule is as follows; in general, an orally active drug has no more than one violation of the following:

- No more than 5 hydrogen bond donors (the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)
- No more than 10 hydrogen bond acceptors (all nitrogen or oxygen atoms)
- A molecular mass less than 500 daltons
- An octanol-water partition coefficient (log P) that does not exceed 5

The rule mentions a few properties of molecules (such as logP, hydrogen bond acceptors [HBA; nitrogen and oxygen atoms], and hydrogen bond donors [HBD; nitrogen–hydrogen and oxygen–hydrogen bonds]) which are contained in this data set, however there are a couple more that are left out. For instance, logS describes the solubility of a substance. As one might guess, logSpH7 describes the solubility of a substance at a pH of 7. Solubility can be important in the process of absorption and distribution. Additionally, the data contains a variable called TPSA, or topological polar surface area. TPSA is the surface sum of all polar atoms within a substance; these polar atoms are primarily oxygen and nitrogen (which just so happen to be the atoms in HBA...). TPSA is also a crucial value when considering penetration of the blood-brain barrier (BBB, also measured in this dataset). Rotatable bonds are bonds which allow for free rotation around themselves; more formally, they are single, non-ring bonds, attached to non-terminal, non-hydrogen atoms. Other variables exist in this dataset (there are 14 descriptors total), however these appear to be the most relevant. The definitions of these variables will become quite important when examining results.

Principal Component Analysis (PCA)

Formally, Principal component analysis, or PCA is a technique used for making predictive models in which principal components are computed and then used to observe data variance, sometimes using only the first few principal components and ignoring the rest. Variance is a highly important concept in statistics as a whole, as well as in the context of this specific lab. Simply put, variance describes the amount of deviation in a dataset from its mean. Generally, a low variation is preferred, as this means that the data in question is highly predictable. Variation can be roughly identified visually; data with low variation will be highly clustered around a point, while data with high variation will be spread out widely.

A common use of PCA is dimension reduction, or any process which reduces a dataset composed of three or more variables into (usually) two variables so that the dataset may be properly modeled and analyzed. This is vital, as it allows one to compose graphics of data that is otherwise impossible to visualize. It accomplishes this by projecting the data onto new variables called PCs, or principal components. PC1 refers the principal component which minimizes the distance between the data and the projection. As the number following PC increases, the distance increases. As such, the first few PCs are usually the most statistically significant and represent the greatest amount of variance.

PCA is how the variable which contributes most to the success of the drugs will be identified. All 14 variables can be analyzed accurately using this method.

Methods

The analysis was conducted using R, along with the ggplot2 software library for some of the visualizations. Much of the basic data handling was done using default R commands, such as `summary` and `names`. A majority of the evaluation was completed using PCA. More specifically, single value decomposition (SVD) was composed using the `prcomp` command rather than an eigenfunction using the `princomp` command. This produced a total of 14 PCs; of these 14, the first 2 were examined graphically using a pairwise and a scree plot. The variables were then ranked based on relative influence on commercial success. Lastly, the PCA data was graphed using a biplot to further observe variance and column influence.

Results

It is always a good idea to perform a small, preliminary analysis of the data before performing the full analysis. This ensures a holistic view of the dataset.

```
## 'data.frame':    1270 obs. of  14 variables:
##  $ 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 ...
```

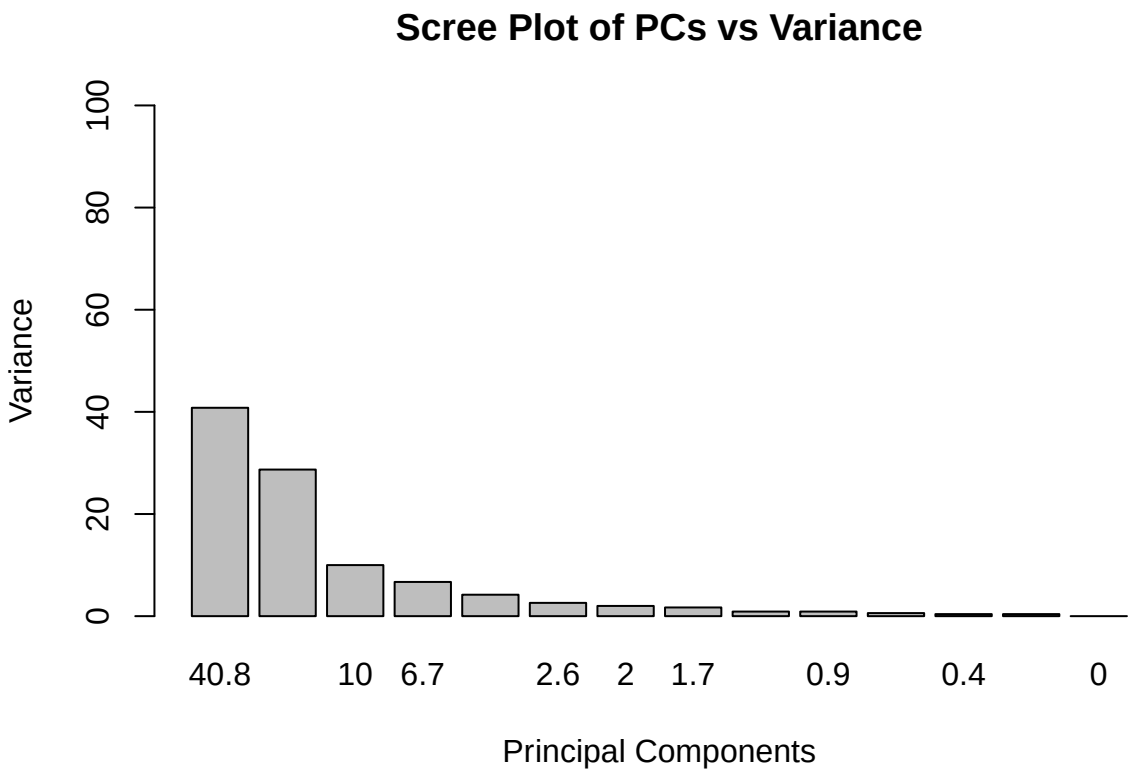
```
##       logS            logSpH7            logP              logD        
##  Min.   :-2.750   Min.   :-2.750   Min.   :-5.0810   Min.   :-5.4780  
##  1st Qu.: 1.770   1st Qu.: 1.665   1st Qu.: 0.6122   1st Qu.:-0.3665  
##  Median : 2.755   Median : 2.611   Median : 2.2770   Median : 1.1280  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

This provides a decent quick view of the data via descriptive statistics, however it does not provide data that is incredibly significant. Luckily, the data itself is already clean, so PCA can commence as discussed. The `prcomp` command was used to generate the PCA data insteadof the `princomp` command. The best way to find out the best one for the scenario is to simply try both and observe the results. In this case, the SVD-based `prcomp` (with scaling enabled) provided the best results.

```
pca <- prcomp(drugs, scale=TRUE)
summary(pca)
```

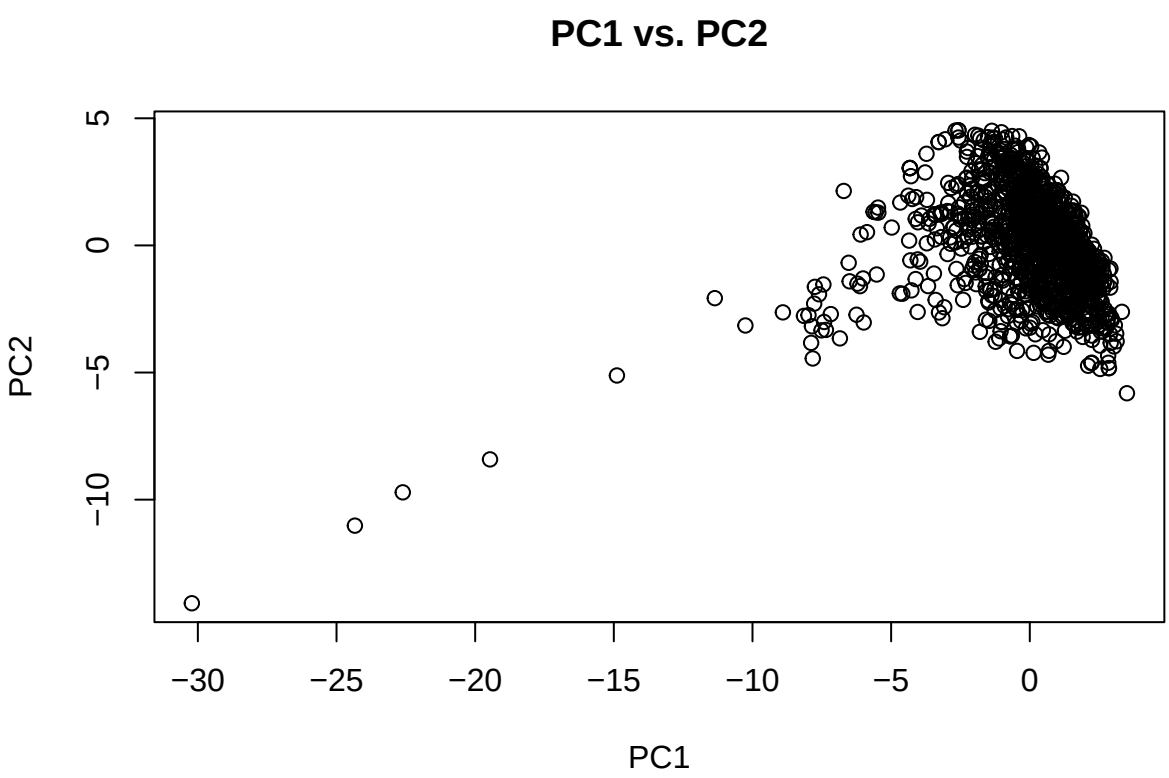
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.3891 2.0055 1.1835 0.96972 0.76392 0.60787 0.52723
## Proportion of Variance 0.4077 0.2873 0.1000 0.06717 0.04168 0.02639 0.01986
## Cumulative Proportion  0.4077 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
```

By taking the summary of this data, the PCs can be observed, as displayed above. This displays, most notably, the proportion of variance represented by each PC. Through this, it can be seen that the first two PCs represent 69.5% of the variance in the dataset. Now that the PCs have been gathered, they can be analyzed using a variety of graphs. The following is a so-called screeplot, which visualizes the variance in each of the PCs. Notice the exponential decay across the variances.

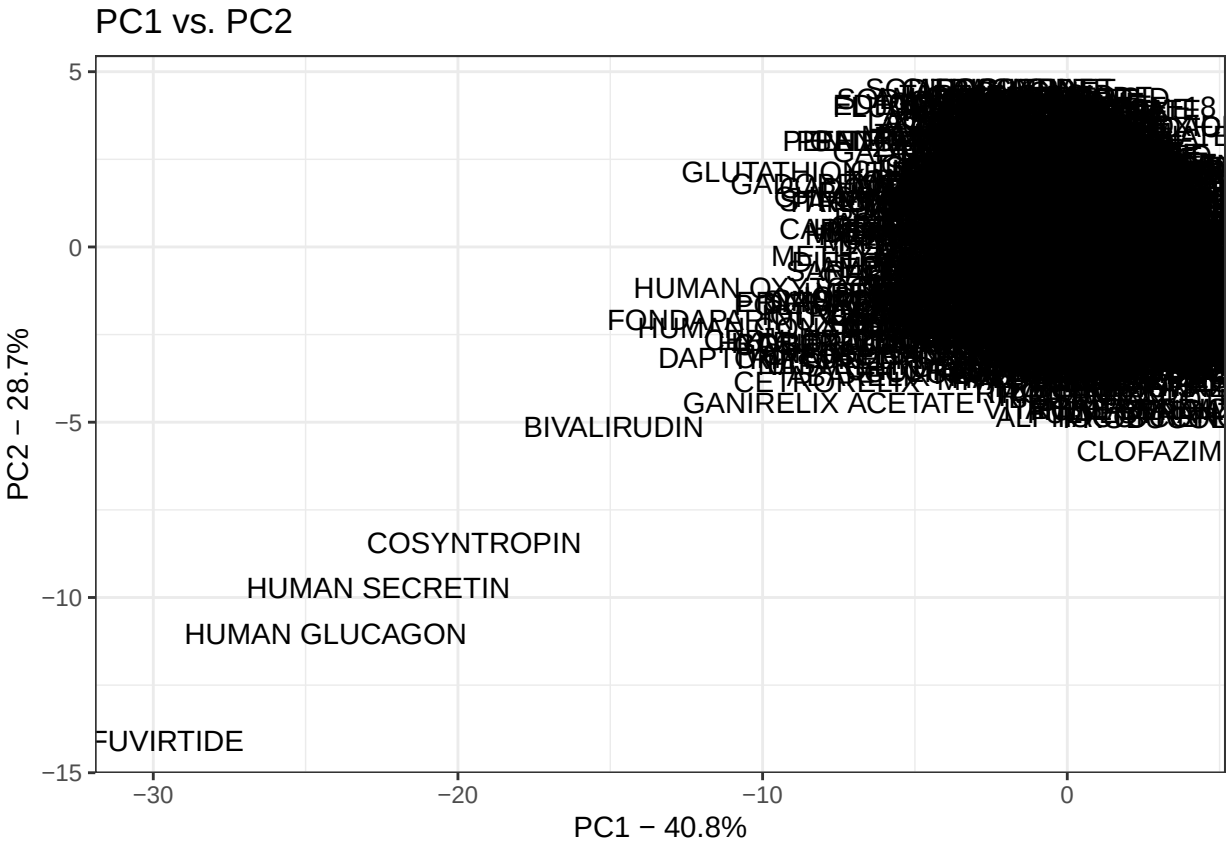


As one can see, the largest concentration of variance occurs in the first PC; then the second, and then the third, and so on. Also notice the sharp decline in variance among PCs. Additionally, the data can be modeled by creating a pairwise plot between the two most concentrated principal components, PC1 and PC2. The purpose of this is to both further understanding of the data's variance, as well as to identify any possible clusters of data.

```
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2", main="PC1 vs. PC2")
```



With this plot, it is difficult to identify clusters, as most of the data is grouped together. With that said, there are a few notable outliers running from the bottom left of the graph to the top right. As for variance, most of it seems to be around the clump in the top right, mostly running diagonally.



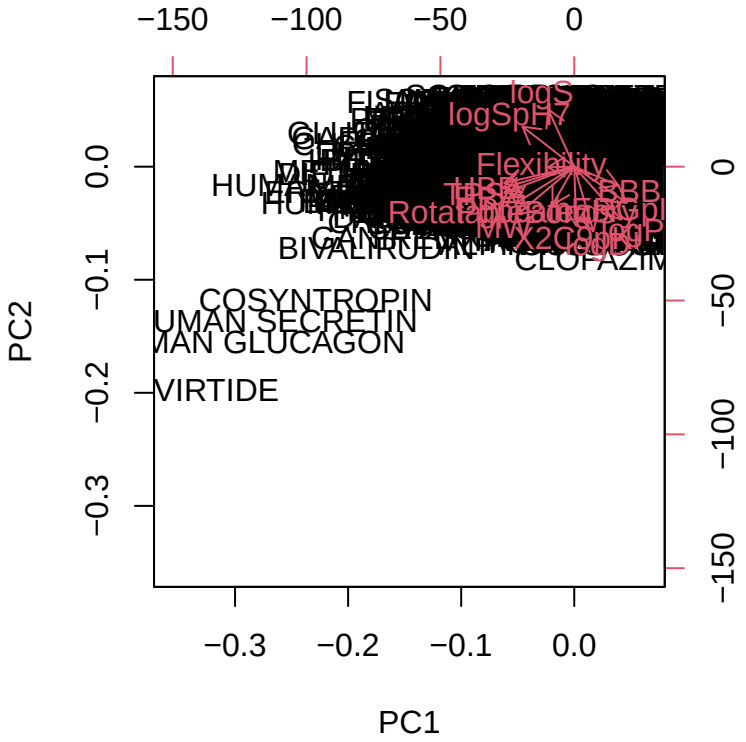
This graphic is perhaps a little more useful. In this one, specific drug names can be seen, however that clump is still near impossible to read due to the concentration of data. Regardless, it demonstrates variance. To provide the most direct answer to the question using the data, loading scores can be used to find relative influence of each variable on the PC data. By sorting this data, the rank of each variable in the dataset can be observed, and the most influential variables can be identified.

```
loading_scores <- pca$rotation[,1]
drug_scores <- abs(loading_scores)
drug_scores_ranked <- sort(drug_scores, decreasing=T)
names(drug_scores_ranked)
```

1. TPSA
2. HBD
3. HBA
4. RotatableBonds
5. MW
6. hERGpIC50
7. logSpH7
8. logP
9. BBB

10. Flexibility
11. logS
12. Pgpcategory
13. logD
14. X2C9pKi

Interestingly, the three most influential variables in the commercial success of these drugs are TPSA, HBD, and HBA. The implications of this will be discussed in further detail in the conclusion. Lastly, a biplot may be constructed to further observe the relationship between the principal components, individual drugs, and data columns, as shown below.



Conclusion

In conclusion, out of the 14 descriptors, the 3 most influential variables are TPSA, HBD, and HBA.

TPSA, or topological polar surface area, is commonly used as a metric for the optimization of a drug's ability to permeate cells (as polar molecules are hydrophilic). As the results indicate, this descriptor is quite fitting. This has direct parallels to ADME, specifically distribution. Increased polar surface area encourages permeation in cells, increasing the amount of the drug that actual enters the desired tissue. As for HBD and HBA (hydrogen bond acceptors and donors) these metrics are directly involved in Lipinski's rule of five. This further confirms the results. Moreover, TPSA is calculated using HBD and HBA atoms. All in all, these conclusions seem to support each other, which collectively support the conclusion that these metrics in particular contribute the most to commerical success in drugs.

References

1. https://en.wikipedia.org/wiki/ADME
2. https://en.wikipedia.org/wiki/PK/PD_models
3. https://en.wikipedia.org/wiki/Lipinski%27s_rule_of_five
4. https://en.wikipedia.org/wiki/Principal_component_analysis
5. https://en.wikipedia.org/wiki/Polar_surface_area