Principal Component Analysis: Discovering Principal Components of Biologically Relevant Therapeutics

Charles Abel, Leo Abfalterer, Jenny Chen, Layth Marabeh
University of California, Santa Barbara

Abstract

Drug development is a crucial aspect of healthcare and has the potential to improve human health, generate economic benefits, advance scientific knowledge, and promote public health. Drug development, however, is highly nuanced and unfortunately drug toxicity results in the failure of around 30% of drug candidates that enter clinical trials [1] costing valuable money, resources, and time. Here Principal Component Analysis, Correlation Analysis, and Linear Regression were used to investigate a data set containing established drugs/substances with known toxicity values (LD50 values). This was done in order to extract the most important dimensions, corresponding to the most valuable physical characteristics of drugs and further investigate the relationships between physical parameters. It was found that molecular weight, XlogP3, and topological polar surface area are the most important dimensions of the data set used. The supporting data is presented below as well as all relevant figures and code.

Introduction

Global society relies heavily on the pivotal role of drugs. Drugs are essential for alleviating symptoms, slowing progression, or even curing a wide range of diseases and conditions, such as infections, cancer, heart disease, and diabetes. Many drugs are used to relieve pain, including familiar over-the-counter products such as ibuprofen and acetaminophen, or for more severe cases, prescribed opioids. More recently drugs play a central role in mediating mental health and are used to treat a variety of conditions, including depression, anxiety, and schizophrenia. They offer improved quality of life and can help people manage chronic conditions, such as asthma or arthritis. As exemplified by the recent COVID-19 pandemic, drugs are also integral to public health and can help prevent the spread of infectious diseases, such as vaccines for viruses and antibiotics for bacterial infections. Looking beyond the patients affected, the development and production of drugs offer immense economic benefits, the pharmaceutical industry is a major contributor to the global economy, creating jobs and generating revenue. Overall, drugs play an essential role in improving public health, reducing suffering, and promoting economic growth.

Considering these positive points, it has been estimated that toxicity is responsible for the attrition of ~ 1/3 of all drug candidates and is therefore a major contributor to the high cost of drug development, particularly when not recognized until late in the clinical trials or post-marketing. The causes of drug toxicity include mechanism-based (on-target) toxicity, immune hypersensitivity, off-target toxicity, and bioactivation/covalent modification. Although covalent binding of drugs to proteins as well as other drug-target interactions have been described and intricately characterized for decades, the significance of contributing molecular drug characteristics and their resultant effects on toxicity have been difficult to establish and can often prove costly to navigate. [2] Here, the question of what physical characteristics of a given drug contribute the most to toxicity will be explored.

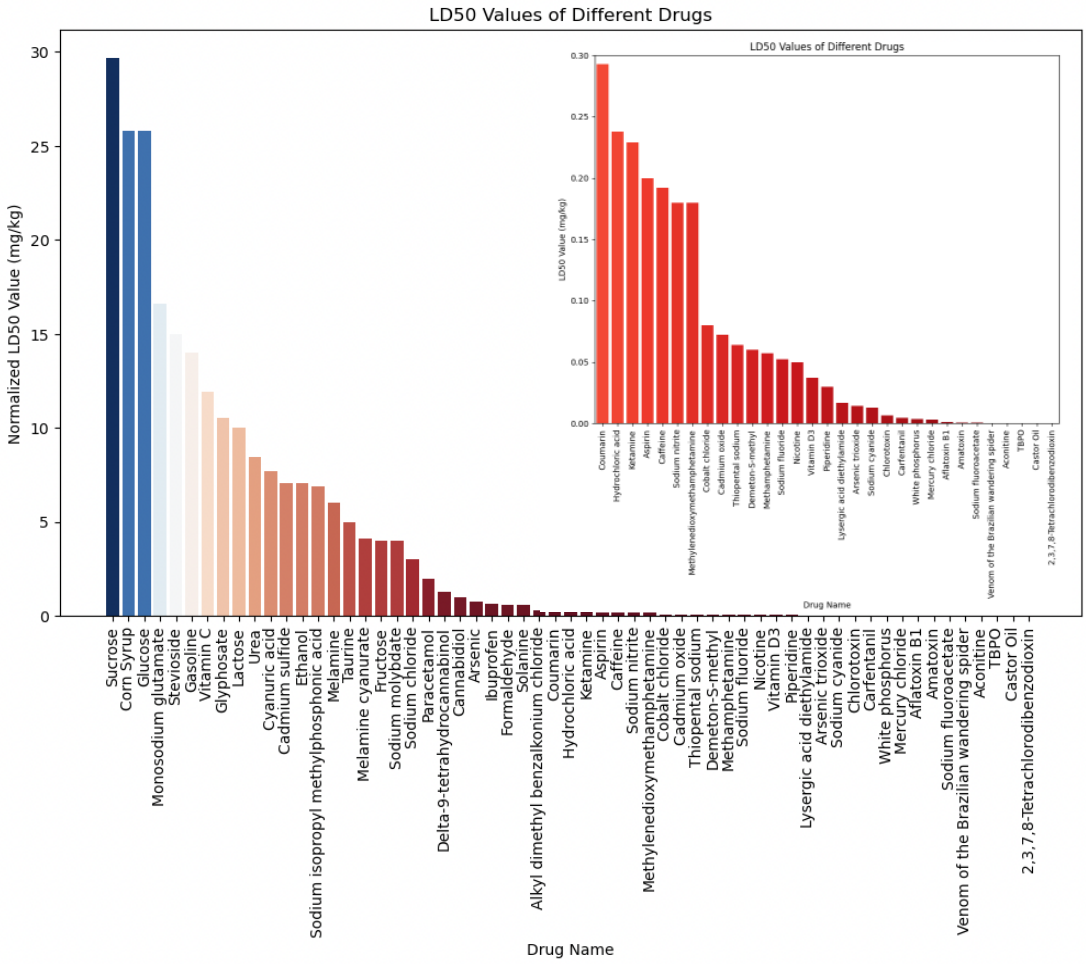

Figure 1 Shows the Normalized LD50 values (y-axis) and compound/molecule name (x-axis). The figure reflects the increase in toxicity going across the graph with the right-most compounds requiring the lowest dosage needed to kill.

Historically small-molecule drugs, such as insulin, aspirin, acetaminophen, and ibuprofen, among others have been molecules with a molecular weight of < 900 *Da*. These small molecules currently dominate the market making up 90% of pharmaceutical drugs [3], and this dominance is predominantly due to small molecules being better tolerated, causing less off-target toxicity as a result of their physically smaller form. In recent years, therapeutics have expanded production to include extensive repertoires of biologics which can be thousands of times larger in molecular size than chemically-synthesized small-molecule counterparts but often come with more expensive development costs and more complex off-target drug interactions.

In order to examine how structure dictates potential toxicity Principal Component Analysis (PCA) will be utilized. Principal Component Analysis is by far the most widespread multidimensional data analysis technique and is widely applicable across a variety of fields ranging from theoretical physics to meteorology, psychology, biology, chemistry, engineering, etc…[4] Concerning pharmacological and biomedically relevant applications, PCA becomes an extremely powerful tool due to its ability to simply and efficiently sort large amounts of often complex data. Here PCA will be applied to a data set containing 38 known drugs/substances with associated toxicity scores in order to extract the most important dimensions. The dimensions investigated are associated with physical characteristics such as size, complexity, and

charge among others. By determining the most important dimensions of a data set it becomes possible to suggest what the most important factors regarding drug toxicity are likely to be. While this data set is relatively small, larger datasets could be surveyed in order to solidify experimental findings and gain confidence in their credibility.

Given these factors, drug toxicity will be investigated below utilizing Python in order to link a drug's physical characteristics with its potential off-target effects. Considering how crucial drug development is to the treatment of billions of people over a massive range of illness, being able to lower development costs and improve patient outcomes is a priority. For these drugs to make it past clinical phases it is crucial that they elicit the smallest possible side effects, drugs that are well tolerated can be given in higher dosages making them not only safer, but often more efficient at treating disease. Here Principal Component Analysis in unison with Correlation Analysis and Linear Regression are used to compare a wide range of small-molecule drugs/compounds as well as some biologics in order to quantify and qualify the most important physical characteristics of a potential therapeutic with regards to its potential toxicity. This information is of interest because when designing therapeutics, being able to select for favorable characteristics would give researchers a higher chance of success, save time and money, and more efficiently allow for progression throughout drug development.

Methods + Results

A substance's lethal dose 50 (LD50) is defined as the dose at which 50% of a sample population passes away. [5] Molecules with selected LD50 values were experimentally determined and obtained by analyzing prior research from rodent experiments. Initially a total of 59 molecules were selected with an LD50 range from 20 *μg/kg* to 29,700 *mg/kg*. Based on the molecules chosen, a thorough database search was conducted utilizing *PubChem*, among other sources, to check for consistency across reported experimental values. Values for molecular weight, XLogP3, topological polar surface area, complexity, formal charge, heavy atom count, H-bond donor count, and H-bond acceptor account were procured from the *National Center for Biotechnology Information Storage* [6]. The information was logged in an *Excel* document with each chemical taking a row and each characteristic a column. For XLogP3, values that were unknown were left blank which resulted in only 38 of the initial 59 molecules having recorded XLogP3 values. LogP3 values are used as an indication to the absorbance of molecules by living tissues, which could otherwise easily be washed away by water. This is reflective of conditions expected in a drug administration environment so preserving this data was important. All other characteristics were easily obtainable for all molecules.

Using the *Excel* spreadsheet, principal component analysis (PCA) was performed. Three different packages in python were used including "numpy", "matplotlib.pylplot", and "pandas". As previously mentioned, molecules with missing values for XLogP3 were removed in order to perform PCA utilizing python. A list was created to store values corresponding to each characteristic of each molecule. The first value of each list was removed as it corresponded to a string of the molecule name. These lists were converted into an array to perform principal component analysis with python's "numpy" package.

The first step in performing PCA was to transpose the array. This was done in order to ensure each characteristic was represented by a row. This allowed for the mean of each row to be taken and for centering of data by subtracting the mean from it. Without mean-centering, the primary eigenvalue could correspond to distance from the mean. Next, the covariance matrix of the mean-centered data was found using "np.cov". The eigenvalues and eigenvectors of the covariance matrix were discovered using numpy's "linealg.eig" function and eigenvalues were sorted from greatest to least value. The eigenvectors were also reshuffled to ensure that eigenvalues

corresponded with their respective eigenvectors. Finally, the transpose of the eigenvectors was matrix-multiplied by the initial mean-centered matrix. This yielded another matrix with reprojected data points in which each subsequent row corresponded to principal components of lower eigenvalues. The first principal component was plotted against the second principle component by plotting the first row vs the second row of the reprojected data.

From *Figure 2*, the plot of principal component 1 vs principle component 2, a cluster of data points can be seen towards the left of the figure. A few points are scattered towards the right of the figure corresponding to large PC1 values. There is seemingly an equal distribution of PC2 values throughout the plot.

To analyze the extent of contribution to variance for different characteristics, the unsorted eigenvalues were examined. Each eigenvalue position corresponded to the position of characteristics from the table from left to right. To clearly view differences between eigenvalues corresponding to their characteristics, a bar chart was created. The characteristics were plotted against their eigenvalues using the "matplotlib.pylplot" package. The largest eigenvalue corresponding to molecular weight results in an inability to visualize other eigenvalues. In order to contrast the numerical eigenvalues of different characteristics, multiple plots were created with distinct sets of characteristics. Looking at *Figure 3b*, the range of values were still too big to see differences between eigenvalues of complexity, formal charge, heavy atom count, hydrogen bond count, and hydrogen bond acceptor account. Therefore, *Figure 3c* was created using just those characteristics. A more detailed relationship could be seen within the smaller range of eigenvalue numbers.

Looking at the bar charts in *Figure 3*, the largest eigenvalue corresponds to molecular weight. The order of eigenvalues from largest to smallest is molecular weight, XLogP3, topological polar surface area, complexity, H-bond acceptor count, H-bond donor count, heavy atom count, and formal charge. Formal charge has an eigenvalue of zero which reflects the fact that the data set used did not supply molecules with interesting formal charge values.

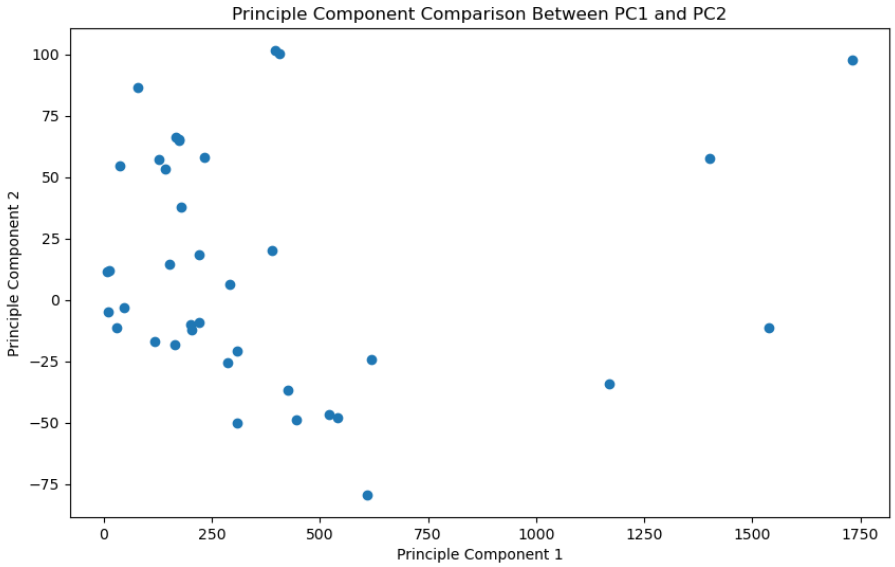

Figure 2. This scatter plot shows reprojected data points using principal component one (PC1) and principal component two (PC2) corresponding to the two highest eigenvalues. A graph was plotted with PC1 vs PC2 with a PC1 range of 0 to1750 and a PC2 range of -75 to 100 to highlight the area with data.

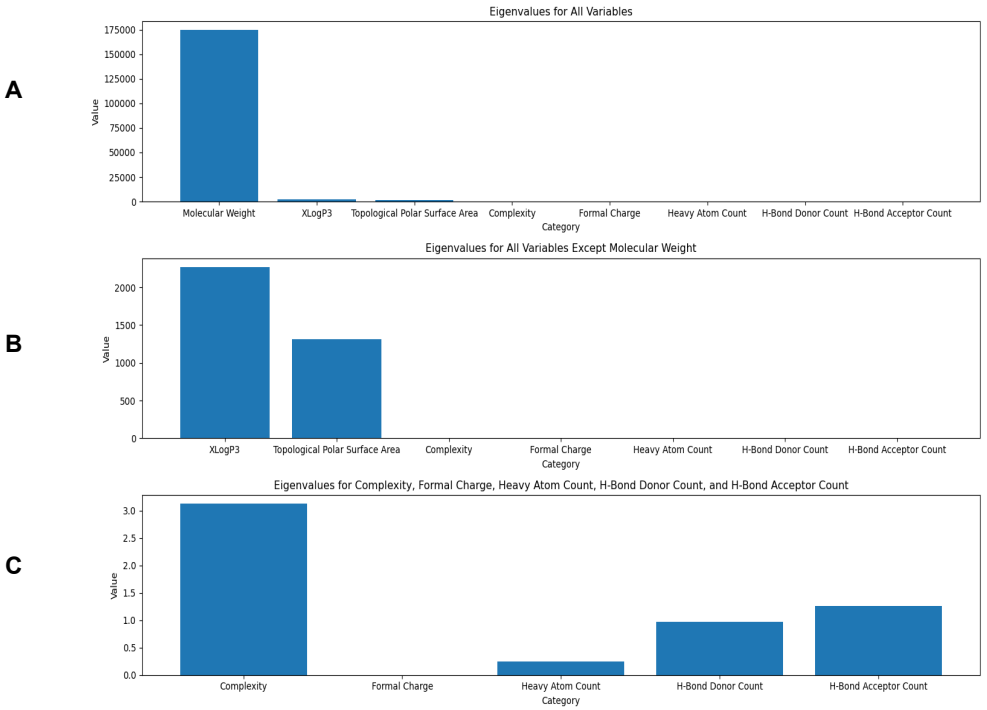

Figure 3. The top graph shows eigenvalues for all corresponding characteristics analyzed.Middle graph only shows eigenvalues for all characteristics except XLogP3. The bottom graph shows eigenvalues for complexity, formal charge, heavy atom count, H-bond donor count, and H-bond acceptor count.

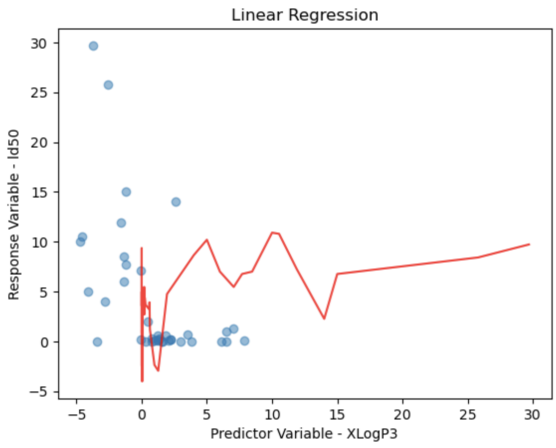

Figure 4

Figure 4. Linear regression scatter plot of predictor variable XLogP3 against LD50 as the response variable. The line of best fit is shown in red.

To analyze the linear relationship between the characteristics and LD50, Linear Regression was performed using the "StatsModel" package in python. The same eight characteristics were pulled from the pandas data frame and fit to a linear regression model. The numerical representation of the relationship between the characteristics and LD50 was obtained using the Ordinary Least Squares regression model. This model compares one or more predictor characteristics and a response characteristic. The linear regression model was plotted against LD50 for a visual representation of the linear relationship. A line of best fit was applied to each graph. The linear regression model returned eight scatter plots, one for each characteristic.

Utilizing the Spearman Correlation Analysis, correlation scores and p-values were obtained for each of the eight characteristics. As shown in *Figure 4*.

Molecular Feature	Correlation Score	p-value
Molecular Weight	-0.32	0.05
XLogP3	-0.54	0
TPSA	0.29	0.08
Complexity	-0.34	0.04
Formal Charge	nan	nan
Heavy Atom #	-0.37	0.02
H-Bond Donor #	0.38	0.02
H-Bond Acceptor #	0.14	0.41

Figure 5. Table listing the molecular feature, correlation score, and p-value stemming from Spearman Correlation Analysis.

Discussion

According to the California Department of Public Health, the three main contributing factors that determine toxicity are chemical structure, the extent to which the substance is absorbed by the body, and the body's ability to detoxify a given substance. [7] From this data set, molecular weight accounts for the most variance with an eigenvalue of around 175000. This makes sense as the drugs listed have a large distribution of molecular weights due to the presence of simple to complex compounds. Using data with a large range of toxicity values is still meaningful in order to investigate whether molecular weight may have implications for the difference among toxicity values. This is supported by previous research indicating that greater absorption corresponds to larger molecular weights [8].

The interesting contributors to variance were XLogP3 and Topological Polar Surface Area. The second highest eigenvalue of 2700 corresponds to XLogP3. Since XLogP3 corresponds to the absorbance of the drug, this analysis reaffirms what was reported by the Department of Public Health that absorbance plays a key role in toxicity. Topological Polar Surface Area also had an eigenvalue of around 1300. This is a significant value and accounts for an aspect of chemical structure that clearly contributes to toxicity in a significant way. Previous studies are in agreement with this prediction showing a correlation between polar surface area and oral drug absorbance [9].

Meanwhile, formal charge likely plays little to no role in the degree of toxicity of a molecule. Formal charge accounts for the least variance in our data with an eigenvalue near zero. This also makes sense since every molecule used had a formal charge of zero and while some drugs do have formal charge values other than zero these drugs are not as attractive for development due to their often high reactivities in the body.

When analyzing the plots and data that linear regression and ordinary least squares produced, it is clear that no characteristic has an especially well-defined linear relationship to LD50. This, however, does not mean that there is not a statistical model that does fit the relationship between any given characteristic and LD50. This is evidenced by the p-values generated using Spearman's correlation score (*Figure 4*). Molecular weight, complexity, heavy atom count, and hydrogen bond donor count all have statistically significant p-values of <= 0.05; meaning that another model could possibly be used to describe

what relationship the characteristics have with LD50.

The analysis of chemical structure's contribution to toxicity is crucial to the field of pharmaceuticals and provides valuable insight into the safety of all ingestible compounds. Ultimately, the analyzed characteristics with the largest contribution to variance in toxicity all affect the rate of absorption and the molecules' tolerability as a therapeutic. The rate of absorption is crucial to determining the toxicity of a compound. Therefore, molecular weight, XLogP3, and Topological Polar Surface Area can be examined within a compound and a correct dosage can be determined. In addition, the potential chemical safety of a compound can be strongly suggested through analysis of these characteristics.

Further studies could be conducted to ensure the impact of these characteristics on toxicity by implementing the python script to a larger data set. Currently, only 38 chemicals were used to determine the largest contributor to variance. A larger data set could provide answers that would likely be more representative and generalizable to a larger sample of drugs and would improve the confidence in suggested findings. In addition, further modification could be done by adding a larger set of characteristics to be examined or utilizing SMILES in order to compare large sets of data in a more efficient way. Using a larger set of characteristics could result in the discovery of more novel chemical characteristics that can be used to determine the toxicity of a compound.

Works Cited

1. Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., & Pangalos, M. N. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nature reviews Drug discovery, 13(6), 419-431.
2. Guengerich F. P. (2011). Mechanisms of drug toxicity and relevance to pharmaceutical development. Drug metabolism and pharmacokinetics, 26(1), 3–14.
3. Govardhanagiri, S., Bethi, S., & Nagaraju, G. P. (2019). Small Molecules and Pancreatic Cancer Trials and Troubles. In Breaking Tolerance to Pancreatic Cancer Unresponsiveness to Chemotherapy (pp. 117-131). Academic Press.
4. Giuliani, A. (2017). The application of principal component analysis to drug discovery and biomedical data. Drug discovery today, 22(7), 1069-1076.
5. *Chapter IV. guidelines for toxicity tests - fda.gov*. (1993). Retrieved March 22, 2023, from https://www.fda.gov/files/food/published/1993-Draft-%22Redbook-II%22-Chapter-IV-C-2.-Acute-Oral-Toxicity-Tests.pdf
6. U.S. National Library of Medicine. (n.d.). National Center for Biotechnology Information. Retrieved March 22, 2023, from https://www.ncbi.nlm.nih.gov/
7. Hazard Evaluation System & Information Service. (2008). *Understanding toxic substances: An introduction to chemical hazards in the workplace*.
8. Chae, S. Y., Jang, M. K., & Nah, J. W. (2005). Influence of molecular weight on oral absorption of water soluble chitosans. *Journal of controlled release : official journal of the Controlled Release Society*, *102*(2), 383–394. https://doi.org/10.1016/j.jconrel.2004.10.012
9. Palm, K., Luthman, K., Unge, A.-L., Strandlund, G., & Artursson, P. (1996). Correlation of drug absorption with molecular surface properties. *Journal of Pharmaceutical Sciences*, *85*(1), 32–39. https://doi.org/10.1021/js950285r