1 Principle Compnent Analysis

Date: 2025–3–10
Objective: Reduce the number of features into a set of principle components. Principle components are linear combinations of the original feature sets which maximally explain the variance in the label set in descending order. for example, a feature set of 15 dimensions (distinct numerical categories of data) maybe reduced to 2 or 3 principle components, which together explain 99% of the variance in the label set. Typically 2 principle components are used.
Methods:

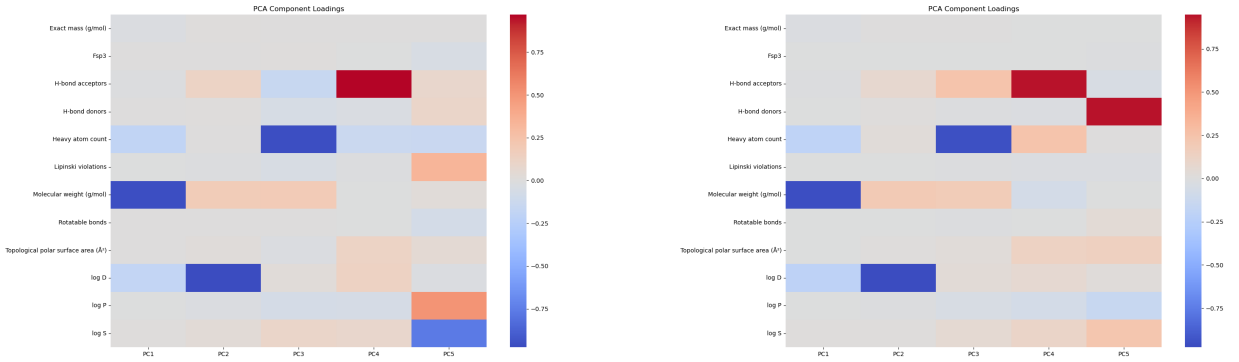
- **Algorithms/Tools:** `Scikit-learn` PCA

Rationale: As discussed earlier, PCA allows one to collapse the number of dimension of of a dataset into (typcally 2) core dimensions that explain the vast majority of variance in the given labels. These resulting dimension are eigenvectors which are linear combinations of the components of the original dataset, and so can give an idea of the major contributors, or principle components, that contribute to label variance. To this end, PCA offers a unique opportunity to assess major contributors of the LXRα/β Agonist and Antagonist profile.

Results:

1.1 Principle Component Features

PCA was used to assess the Agonist and Antagonist profiles of each molecule. While there are significantly less molecules with recorded agonist activity ($n = 131$), a somewhat more defined trend was seen with regards to Principle Component 1 (PC1). For both Agonist and Antagonist sets, the normalized molecular weight was the leading feature of PC1. Similarily the largest contributor to PC2 for both Antagonist and Agonist data was identical, though this leading contributor was Log D.

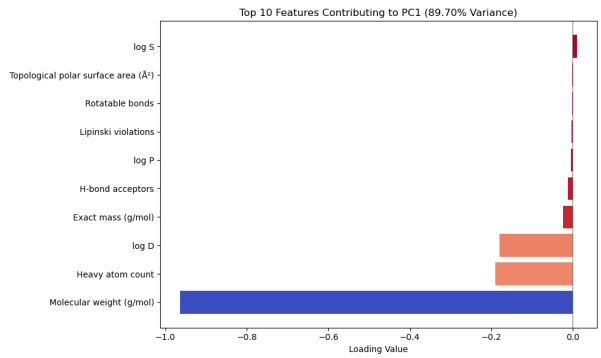


(a) Antagonist Principle Component Heat Map

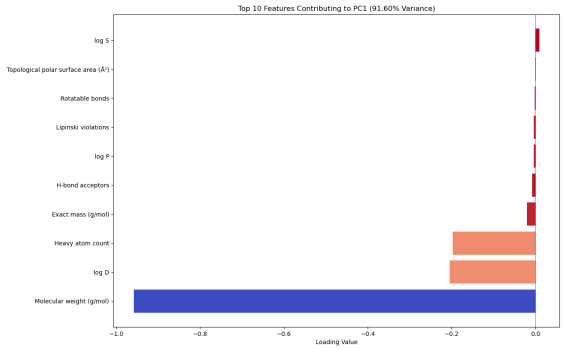
(b) Agonist Principle Component Heat Map

Figure 1: Heat Maps of the first 5 principle components of LXRα/β Antagonist (Left) and Agonist (Right) activity. Principle components were trainied on ΔLXR activity (LXRα - LXRβ). Blue indicates negative correlation with increased variance while red indicates negative correlation to increased variance.

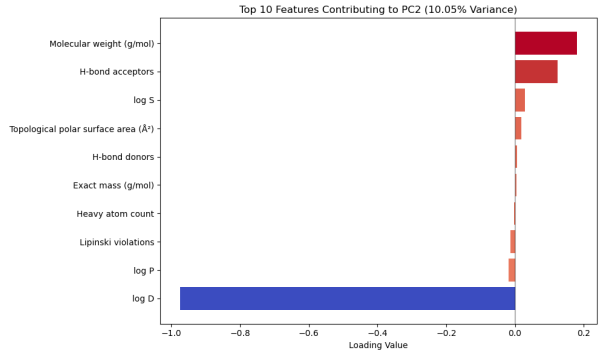
As seen, Principle Components 1 composition appears to be virtually identical between the Antagonist and Agonist data. Principle Component 2 also shares a remarkable resemblance between the Agonist and Antagonist data, but differs by reliance on the number of H-Bond acceptors (more reliance in Antagonist data) and Heavy atom count (no reliance in Antagonist data). More significant changes are seen in Principle Components 3 to 5, but due to their combined low contributions to the total variance (0.23% and 0.16% for the Antagonist and Agonist data respectively) these changes likely have little contribution to any differences in possible emergent properties.



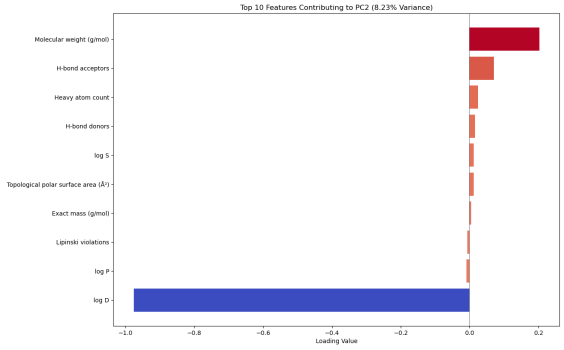
(a) Antagonist Principle Component 1



(b) Agonist Principle Component 1



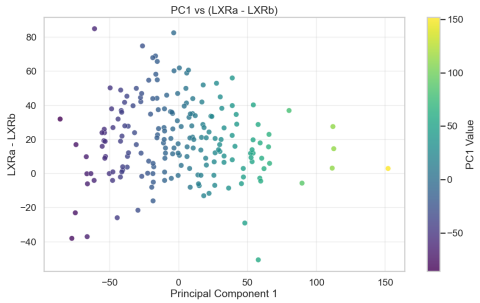
(c) Antagonist Principle Component 2



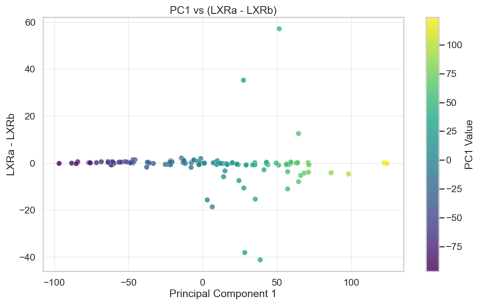
(d) Agonist Principle Component 2

Figure 2: Comparison of loading orders (importance indicated by maginitude of the bar) of the top 10 features of PC1 and PC2 in Agonist and Antagonist datasets. PC1 appears nearly identicle between the two sets, but differs in its total explanation of variance (89.7% vs 91.6% in the Antagonist a) vs Agonist b)data respectively). PC2 shows a slightly higher degree of deviation between the Antagonist and Agonist data and differs in total contributions to variance (10.05% and 8.23% for the Antagonist c) and Agonist d) data respectively).

While correlation between the first two principle axis was found, to further investigate any emergent properties that may be used to predict LXRα/β activity, the data was plotted in the reduced dimension space.



(a) Antagonist PC1 vs ΔLXR activity



(b) Agonist PC1 vs ΔLXR activity

Figure 3: Comparison of PC1 vs $LXR\alpha - LXR\beta$ Antagonist and Agonist data. Antagonist activity is generally clustered around $PC1 = 0$ though shows little correlation. Agonist activity is significantly more tightly clustered on the PC1 axis, but seems to increase in variablility with increasing PC1.

1.2 Discussion & Takeaways:

While PCA was able to determine some features of interest, the overall trends in the principle component space did not yield many noticeable trends outside of what was already clear. A clear linear correlation between LXRα and LXRβ activtiy was recorded in both Agonist and Antagonist data (see Github for more graphs).

There appeared to be some correlation between increasing PC1 and ΔLXRα/β agonist activity variability (see Fig. 3 b)), but there was no discernability in correlation between PC1 and ΔLXRα/β antagonist activity. The largest contributor to PC1 was molecular weight, which was negatively correlated to the variance. Due to the normalization step required for PCA, this indicates that slightly lower molecular weight may be indicative of more variable antagonist/agonist activity. While this is an interesting finding, due to the bias of the dataset the importance of this feature is likely due to reliance of specific substructures of the molecule. A more effective means of evaluation may be to look at the sub-structures of each molecule, which will take analysis of the provided SMILE sequences. Investigation into the various methods that are used to quantify sub-structures of molecules should be undertaken to see if a dataset of substructures can be constructed.