Structural Analysis

Sam Bolton

June 5, 2025

1 Overview

This report details the steps taken findings of exploring the pharmacophore feature space of the LXRα/β dataset. To prepare for this exploration I went through Talk 9 (T009 Ligand-Based Pharmacophores) of the Talktorial series from TeachOpenCADD. Many of the early steps in data cleaning and structuring is taken directly from this talktorial with minor alterations to fit our data structure. This report will go over the visual analysis of some pharmacophore features of interest, as well as the machine learning analysis of the full pharmacophare feature space. Additionally, due to the computational cost of some of the required steps, and certain flexabilities allowed by its use, A Jupyter Notebook was used to carry out this analysis.

1.1 Objective

The aim of this exploration was to to determine if the spatial orientation of pharmacophore features could be used to enhance machine learning efficacy.

1.2 Rationale

The current theory for LXRα/β receptor binding suggests that small differences in hydrophobicity and hydrogen bonding patterns may confer very different binding affinities and antagonistic behaviours. Through the use of the spatial data of pharmacophore feature regions, we hope to be able to more accurately predict selective antagonist properties. While this may be possible purely through a machine learning approach, pronounced trends in hydrophobic, hydrogen bond donor, and hydrogen bond acceptor regions may be helpful in developing a more comprihensive theory of how LXRα/β selectivity is attained.

1.2.1 Support Vector Machine Classifier

The support vector machine is a powerful supervised machine learning algorithm which has been used for decades. In simple terms, it defines the boundary in a dataset which maximizes the difference between two groups. In a 2 dimensional linear case, this would entail finding the line that most effectively separates a dataset into 2 groups depending on some feature. Different separation algorithms may be used in higher dimensions (such as polynomial or radial basis function) to increase predictability in more complicated datasets. This supervised learning approach will be able to determine if there are any boundaries within our spatial or property data that can be used to classify a good candidate.

1.2.2 Random Forest

Random Forest has been used at nearly every step of this project because it is a light weight, powerful machine learning algorithm which allows us to see the learned order of importance of certain features. Due to the large amount of data per molecule, this order of importance may give insight into the most important mechanisms of interaction between the ligand and receptor.

1.2.3 K-Nearest Neighbours

K-nearest neighbours has also been used at every step of this project because it is light weight and versatile. Throughout this project both the size and bias of the dataset has posed serious issues. Methods such for rebalancing the dataset, such as synthesizing data using algorithms like `SMOTE` (Synthetic Minority Over-sampling TEchnique) could be used. `SMOTE` uses the k-nearest neighbours to cluster minority classes in datasets. It then synthesize new instances of the minority set by adding new points in the cluster. The assumption here is that k-nearest neighbours can adequately cluster classification groups, which may not be the case. If k-nearest neighbours is a poor performer, then `SMOTE` likely will not assist in machine learning accuracy. Similarly if KNN is a strong performer, then `SMOTE` may be used to balance the dataset.

1.3 Methods

To determine and extract pharmacophore features and locations the `rdkit` Python Package was used. `rdkit` is a cheminformatics open source toolkit originally developed in 2006 which has been over the past two decades into a powerful tool for molecule visualization, classification, and simulation. Pharmacophore features were extracted from molecule SMILES. `matplotlib` and `Seaborn` were used for data visualization, `Pandas` was used for data storage and `Numpy` was used for simple mathematical operations. Finally, machine learning models and evaluation metrics were sourced from `Scikit-learn` (Random Forest, Support Vector Classifier, K-Nearest Neighbours). As per initial investigation into the success of classification vs regression models, classifiers were used to evaluate high LXRβ inhibition and low LXRα inhibition (one of each model per receptor). Thresholds for positives were established at 50% or lower LXRβ activity and 75% or higher for LXRα activity. Data was split into (80:20) → (training:testing) sets.

1.3.1 Feature Extraction

To extract molecular features from SMILE sequences, `rdkit` was used to generate a MOL object. The `GetFeaturesForMol` function was then used to extract the molecular features stored in the MOL object. For simplicity of use, features which had were described as a set of coordinates (such as centroids) were split into their 3 coordinate components and evaluated separately.

2 Results

2.1 ML Models

3 lightweight, supervised machine learning classification models were trained on the thresholded LXRα and LXRβ datasets. This resulted in 6 models total. Throughout these models, Random Forest (RF) seemed to perform the best on both data sets. the Support Vector Classifier (SVC) performed the poorest in the LXRα dataset while K-Nearest Neighbours (KNN) performed the poorest in the LXRβ dataset.

Model	LXRα Acc.	LXRα Pos.	LXRα Neg.	LXRβ Acc.	LXRβ Pos.	LXRβ Neg.
SVC	66.7%	N/A	N/A	64.1%	58%	67%
RF	71.8%	73%	67%	74.4%	67%	81%
KNN	69.2%	72%	60%	61.5%	55%	64%

Table 1: Machine learning model total accuracy, true positive accuracy, and true negative accuracy for LXRα and LXRβ antagonist data respectively. Models tested were the Support Vector Classifier (SVC), Random Forest Classifier (RF), and K-Nearest Neighbours Classifier (KNN).

Table 1. shows the accuracy, true positive rate, and true negative rate of each machine learning model on the LXRα and LXRβ datasets respectively. The SVC performed the worst in the LXRα set, scoring 66.7% accuracy. Due to unknown reasons, the True Positive (TP) rate and True Negative (TN) score calculation came out to be 100% and 67% respectively, which does not make sense mechanistically or mathematically and so have been omitted as to not mislead the reader. The KNN model performed close the RF model in the LXRβ set with an accuracy of 69.2% and nearly matching the RF TP score (RF = 73%, KNN = 72%). The KNN TN score was significantly lower than the TP score at 60%. RF performed the best in the LXRα set at 71.8% accuracy, 73% TP and 67% TN. For the LXRβ set RF remained the best performer with an accuracy of 74.4%, a TP of 67% and a TN of 81%. SVC performed the second best with an accuracy of 64.1%, a TP of 58%, and a TN of 67%. KNN performed the worst on the LXRβ set, with a total accuracy of 61.5%, a TP score of 55%, and a TN score of 64%.

2.2 Random Forest Importance

Random Forest (RF) performed the best on both the LXRα and LXRβ data sets out of the three models selected. Due to the mechanism of action of RF, the feature importance of the LXRα/β dataset was able to be visualized.

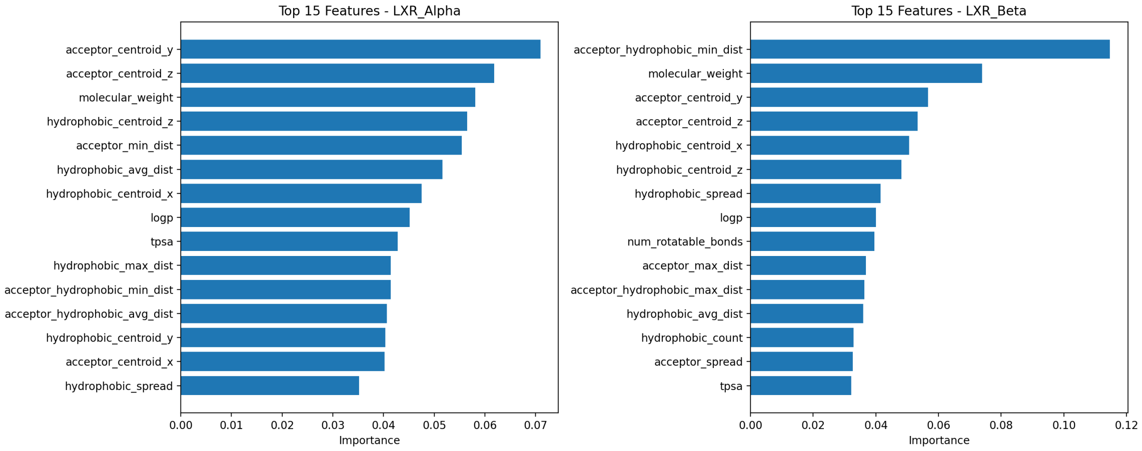


Figure 1: Top 15 features of importance of Random Forest analysis of the LXRα (left) and LXRβ (right) Pharmacophore datasets.

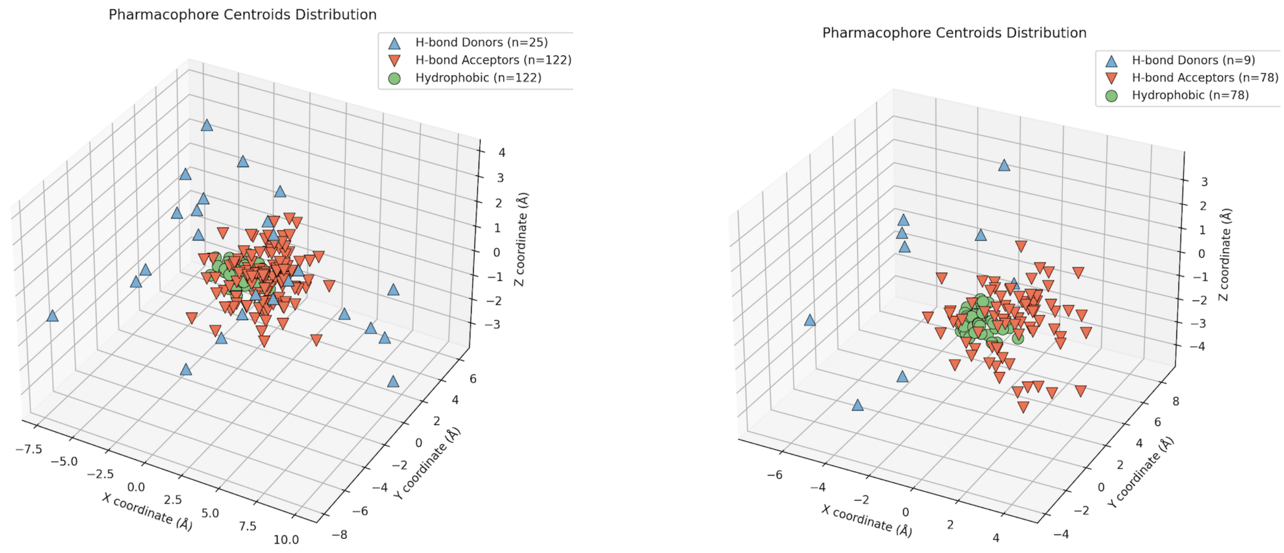
RF had the best overall accuracy on the LXRβ set, which is reflected in the importance of the number 1 feature labeled in Fig. 1. as `acceptor_hydrophobic_min_dist` at ≈ 0.11. This feature looks at the minimum distance between a hydrophobic group and the closest hydrogen bond acceptor group in the molecule. The number 1 (≈ 0.07) and 2 (≈ 0.06) features in the LXRα model was the y and z coordinate of the hydrogen bond acceptor centroid respectively. Interestingly this was the number 3 (≈ 0.06) and 4 (≈ 0.055) feature in the LXRβ. This trend of close importance of 2 of the 3 spatial coordinates is seen in the hydrophobic and hydrogen bond acceptor centroids in the LXRβ set and the hydrogen bond acceptor centroid of the LXRα set. This is encouraging, as it seems that is utilizing the 3 dimensional spatial data provided by the feature extraction. In both datasets, molecular weight still appears as one of the top predictors, which has been seen in previous iterations of the random forest model and the PCA of the original property set.

2.3 Spatial Analysis

Due to the small dataset and large amount of features, several of the extracted spatial features were visualized to determine if any visually emergent properties were present. Scatterplot's, heat maps of molecule centroids for hydrophobic, hydrogen bond donor, and hydrogen bond acceptor groups were plotted by binarized activity level.

2.3.1 Mixed Centroids

3D coordinates of select Pharmacophore features were plotted to give an idea of A) if the program was properly aligning molecules s.t. the spatial data of said features actually gave useful information to the machine learning models, and B) to see if there were any visually distinct emergent properties in the dataset. Molecules were sorted into LXRα hits and LXRβ hits according the classification threshold used in the machine learning analysis outlined above.



(a) LXRα centroid distribution

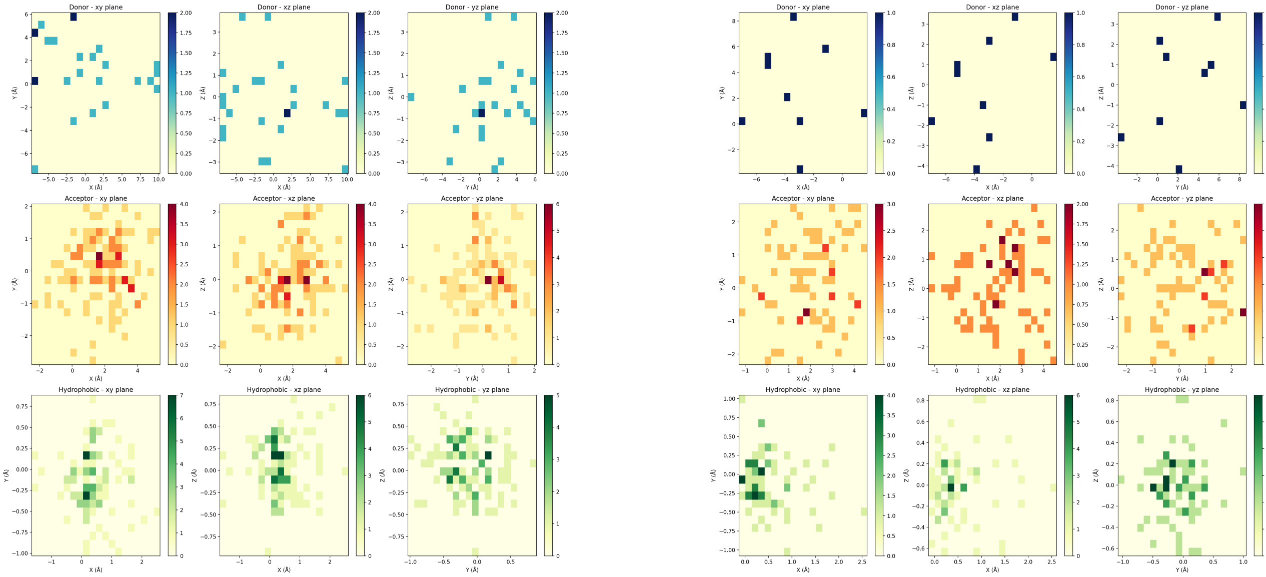
(b) LXRβ centroid distribution

Figure 2: Spatial distribution of hydrophobic (green), H-bond donor (blue), and H-bond acceptor (red) of LXRα non antagonists (>75% activity at 100nM) on the left and LXRβ antagonists (<50% activity at 100nM) on the right.

Figure 2 shows the spatial distributions of hydrophobic, H-bond donor, and H-bond acceptor groups of LXRα hits (left) and LXRβ hits (right). As expected, the similarity of molecules is reflected in the similarity between the distribution of pharmacophore features. Both groups possess a tightly clustered hydrophobic core with more diffuse H-bond acceptor groups. Between the two groups, the LXRβ group seems to have a less defined H-bond acceptor cluster (higher diffusivity) but this may just be due to the lower number of recorded molecules. While the 3D representation cannot be rotated, there appears to be some distinction in H-bond donor location. While the LXRα H-bond donors appear to be clustered fairly centrally with a slight bias to the negative X side of the plane, the LXRβ group shows a very definite bias to the negative X plane, with no centroids being present in the positive side of the plane. This discrepancy was not utilized in the Random Forest model (as seen in the importance graphs of Fig. 1) but again, this may be due to the low number of recorded H-bond donor groups in the LXRβ field.

2.3.2 Heat Maps

Due to the limitations of static images, 2D heatmaps were generated to give a better view of the spatial distribution of the hydrophobic, H-bond donor, and H-bond acceptor centroid distributions.



(a) LXRα Centroid Distribution heatmap

(b) LXRβ Centroid Distribution heatmap

Figure 3: 2D heatmaps of H-bond donor (blue), H-bond acceptor (red), and hydrophobic (green) pharmacophore centroids of LXRα (left) and LXRβ hit molecules in the XY, XZ, and YZ planes respectively. Darker colour indicates higher importance (more central centroid in the cluster). Spatial units are in angstroms (Å) from the center of the molecule

Fig. 3 shows the 2D heatmaps of the three pharmacophore features of LXRα (left) and LXRβ (right) hits, from the XY, XZ, and YZ planes respectively. Darker colored centroids generally means higher importance/more central location in the centroid cluster. Features with high dispersion and low instances (as seen in the H-bond Donor feature of LXRβ) will have a uniform importance as little correlation/importance can be drawn from the distribution. It is important to note that each graph is not necessarily centered at 0. While the Donor graphs of LXRβ seem fairly normally distributed about the origin of the plane, the cluster is biased to the negative X (mean ≈ -3Å) and positive Y axis (mean ≈ 2Å). The LXRβ H-bond acceptor cluster appears to be biased to the positive X axis (mean ≈ 2Å), with little bias in the Y and Z axis (mean ≈ 0Å). The LXRβ hydrophobic centroids show a much more skewed distribution, with a positviely skewed X axis (mean ≈ 0.5Å) but a tail up to 2.5 Å, but seemed to be normally distributed about Y and Z. LXRα showed much less bias with the H-bond donor group, with a slight bias to the positive axis axis (mean ≈ 2Å) but a significantly higher spread. The Y compontent seemed to be slightly biased to the positive direction (mean ≈ 2Å) and Z components seemed fairly normally distributed about the origin, again maintaining a high spread. The LXRα H-bond acceptor spread was much tighter compared to LXRβ and seemed to have an X component clustered around ≈ 2Å, with a slight bias to the positive Y component. The Z component was fairly normally distributed. Finally the The hydrophobic centroids seemed be fairly normally distributed about all of the axis, with mild bias towards the positive X axis.

2.3.3 Activity Distributions

Due to the generally normal distribution of data about the Z axis shown in fig. 1. and fig. 2. 2D scatter plots (X and Y) of the 3 spatial pharmacophores were plotted with feature importance gathered from the Random Forest analysis detailed below. It is important to note that only centroids from molceules which had LXRα activity greater than 75% or LXRβ activity less than 50% were visualized. Centroids of molecules that had both of these properties were visualized in both graphs.

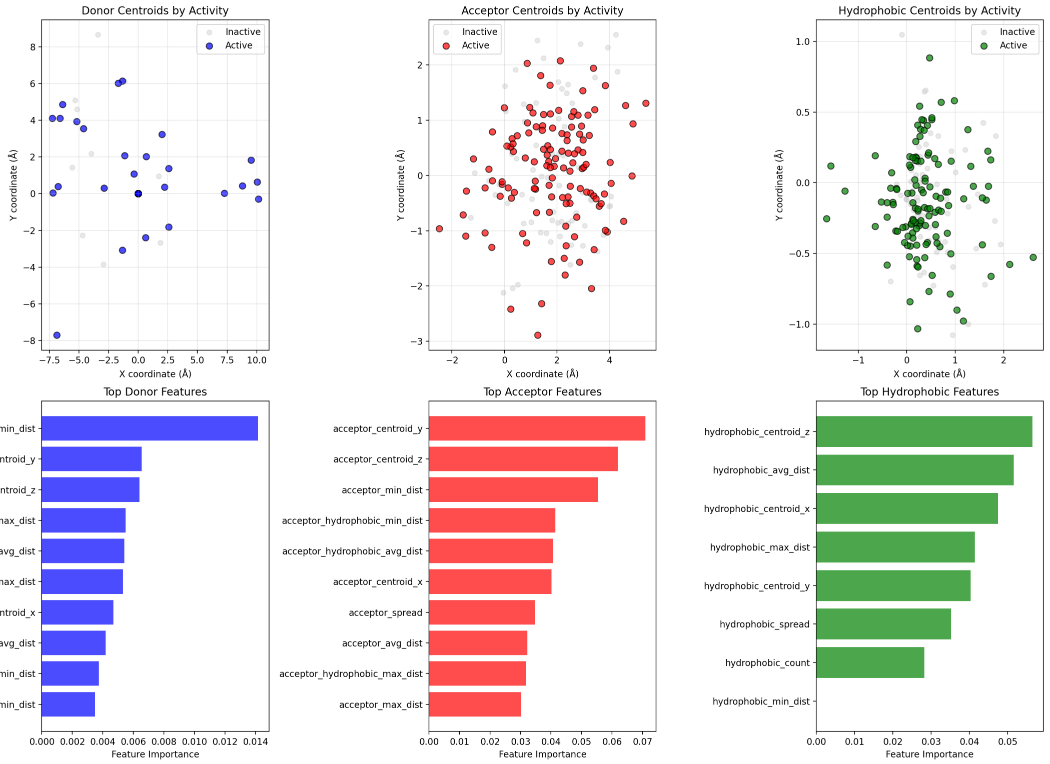


Figure 4: Spatial phamracophore centroids X and Y coordinate of molecules with high LXRα activity ($> 75\%$) in color and lower than 75% in gray. H-bond donor seen in blue, H-bond Acceptor seen in red, and Hydrophobic group seen in green. Importance was taken from `SK-learn` Random Forest importance calculated by Mean Decrease in Impurity (MDI).

Fig. 4 shows the X and Y distributions of the H-bond donor (blue), H-bond acceptor (red), and hydrophobic centroids (green) with their importance calculated by Mean Decrease in Impurity (MDI) from `SK-learn` Random Forest. Grey dots represent centroids with low LXRα activity ($< 75\%$) and colored dots represent centroids with high LXRα activity ($> 75\%$). As was seen in Fig. 3 the donor centroids appear to be fairly normally distributed about the Y axis with a high degree of spread, while there appears to be a minor bias to the positive X axis. Donor hydrophobic minimum distance, which denotes the smallest distance between a hydrophobic and H-bond donor group in a molecule. While this has a significantly higher importance than the next leading feature, it is still approximately an order of magnitude smaller than the most important features of the H-bond acceptor or hydrophobic features. The H-bond acceptor graph shows a fairly normal distribution of centroids relative to the Y axis, and a bias towards the positive X axis (mean ≈ 2Å). The feature importance was more conclusive than the donor features, but were still relatively small. Interestingly, despite the Z coordinate in the heatmap suggesting a normal distribution about 0, the Y and Z coordinates were the top 2 features which had similar feature importance (≈ 0.07 and 0.06 respectively). This importance is still very low but were the highest in the LXRα RF model. The hydrophobic centroid distribution gathered fairly tightly on the X axis, with a slight bias towards the positive X direction. The Distribution was fairly normal about the Y axis with a very slight bias to the negative Y direction. Similar to the acceptor features, despite the Z compontent showing a very normal distribution in the the heat maps of Fig. 3, the Z component of the hydrophobic centroid coordinates had the highest importance, though this was still low (between 0.05 and 0.06). Overall minor trends in centroid distribution could be found with minor biases in X direction (with varying degrees of dispersion of the cluster) for all the centroids. This however, was not reflected in the importance scores.

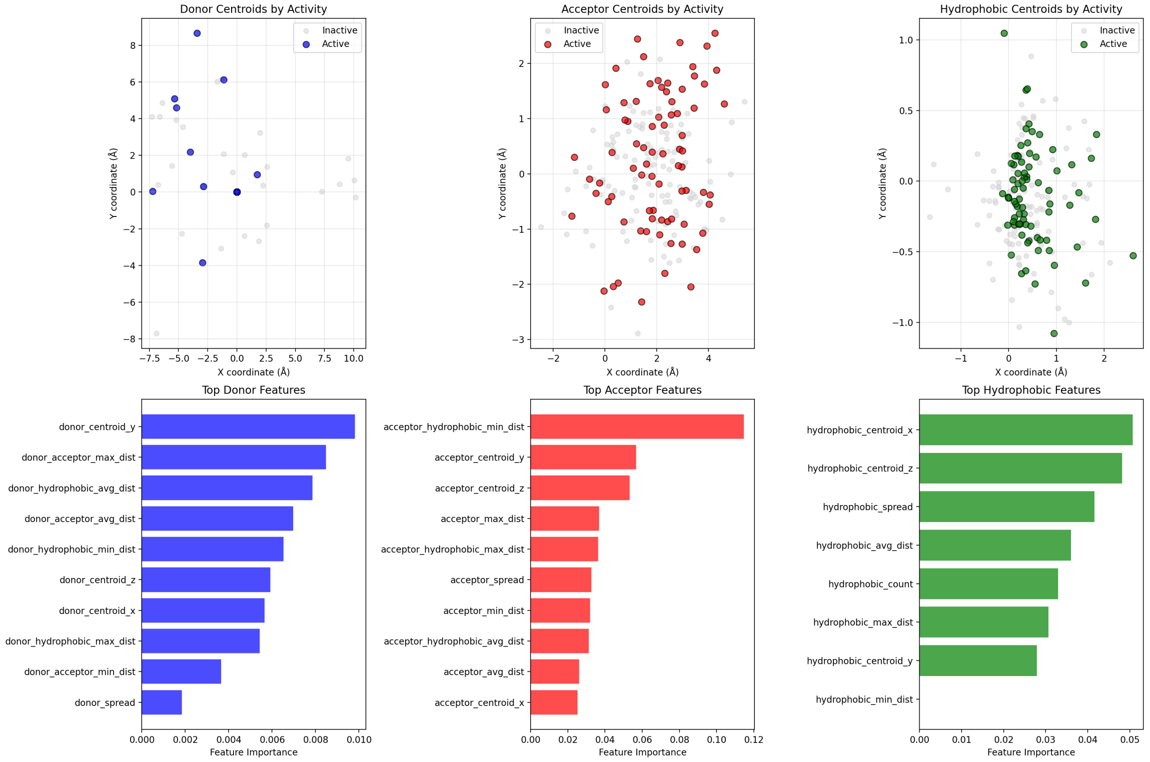


Figure 5: Spatial phamracophore centroids X and Y coordinate of molecules with high LXRβ activity $< 50\%$ in color and $> 50\%$ in gray. H-bond donor seen in blue, H-bond Acceptor seen in red, and Hydrophobic groups seen in green. Importance was taken from `SK-learn` Random Forest importance calculated by Mean Decrease in Impurity (MDI).

Fig. 5 shows the X and Y coordinate distributions of the H-bond donor (blue), H-bond acceptor (red), and hydrophobic centroids (green) with their importance calculated by Mean Decrease in Impurity (MDI) from `SK-learn` Random Forest of LXRβ antagonists. Grey dots represent centroids with high LXRβ activity ($> 50\%$) and colored dots represent centroids with low LXRβ activity ($< 50\%$). The number LXRβ donor centroids was very low (n=9), meaning little importance should be seen from their distribution which is reflected in the importance table. Visually the centroid distribution has a noticeable bias to the negative X and positive Y direction (means ≈ -3Åand ≈ 2Årespectively), this could just be an artifact of the small number of points (n=9). The highest importance feature was the Y coordinate of the centroid, but this importance was around 0.01, which is an order of magnitude smaller than the most important feature in the total feature set. The H-bond acceptor feature set showed a fairly normal clustering with respect to the Y axis, but showed some bias to the positive X axis (mean ≈ 2Å) which is mirrored in the LXRα set. The most important feature (importance ≈ 0.11) of this set was the smallest distance between a hydrophobic group and a H-bond acceptor. The second and third most important features were the Y and Z coordinates of the acceptor centroids. The hydrophobic centroids displayed a high adherence to $X = 0$ with a strong right tail but no left tail. This is interesting as it deviates quite strongly from the LXRα group. There was no discernible bias in Y axis, which had a fairly noramlly distributed spread centered at Y = 0. The most important hydrophobic features were the X and Z coordinate, and the hydrophobic spread (spread of hydrophobic groups within a molecule) respectively. This is the first instance of the RF model utilizing the X coordinate as a deciding factor, but due to the relatively low importance (≤ 0.05) it doesn't appear to be used to a great effect in decision making. Overall the spread in the hydrophobic centroids is much tighter in both the X and Y directions than the H-bond acceptors and donors, which is reflected in Fig. 2 and Fig. 3.

Generally the spatial distributions of the LXRα and LXRβ positive centroids had similiar properties, with more selectivity seen in the LXRβ set, though this may be a result of a lower number of points. The H-bond donor centroids of the LXRβ set were significantly more clustered to the negative X axis, though with only 9 instances this may just be a result of a low amount of data. The H-bond acceptors centroid distribution seemed to deviate very little between the LXRα and LXRβ sets. The hydrophobic centroids distribution displayed some deviation between the LXRα and LXRβ sets, with a noticable bias to the positive X axis for the LXRβ group, but a very defined clustering centered at X and Y = 0 for both distributions.

3 Discussion

Analysis of the spatial distribution of chemical functional groups offered some interesting insights into available avenues of pure machine learning and interesting features in LXRβ selective antagonists.

3.1 ML Models

In previous steps of this project we have used several machine learning models, but have consistently used the Random Forest model (RF) and K-Nearest Neighbours model (KNN). In this iteration we introduced the Support Vector Classifier (SVC), which has a well documented history in machine learning and classification. As has been the case with the project so far, dataset size was a significant challenge to each machine learning model, and bias between negatives and positives within the dataset lead to a noticble difference in sensitivity and specificity (True Positive and True Negative rates). Generally LXRα trained models were more accurate than LXRβ trained models and showed a marginal to significantly better ability to classify positives (LXRα activity greater than 75%) than negatives. LXRβ models were marginally less accurate on average and were more successful classifying negatives (LXRβ activity less than 50%) than positives. Random Forest was the most accurate model in both sets while the SVC and KNN models performed the poorest in the LXRα and LXRβ sets respectively.

3.1.1 Support Vector Classifier

The SVC struggled with the small dataset, but had the lowest degree of variance of accuracy between the two sets. Due to some computational issues, the True Positive and True Negative scores of the LXRα set could not be adequately calculated. The LXRβ trained model showed a minor reduction in accuracy (66.7% → 64.1%) with a 67% true negative rate and 58% true positive rate. These scores show that SVC is not well suited to this small and complicated dataset ⇒ there is no clear emergent 'separation line' in the feature space. This may just be due to the low amount of data we have or a hyperplane separation with our current feature set may not be possible. As such, in future iterations of this project the SVC should be exchanged for Gradient Boost unless some method of alternate molecule representations to artificially expand the dataset is used.

3.1.2 K-Nearest Neighbours

The KNN model performed the second best on the LXRα set with a true positive rate 72% and a true negative rate of 60%. This yeilded a total accuacy of 69.2%, owing to the high bias within the set. KNN performed the worst in the LXRβ set, with a total accuracy of 61.5% and a true postitive and true negative rate of 55% and 64% respectively. KNN was used to gauge the effectiveness of supervised clustering of the various pharmacophore properties. Through this project two main issues have hampered progress; dataset size, and dataset bias. Mechanisms for rebalancing a biased dataset exist, but rely on fundamental assumptions on the dataset. One such method that could have been used was SMOTE, which relies on the KNN algorithm to cluster majority and minority classes within a dataset and then interpolate points between minority class points to instantiate new synthetic points and rebalance a training set. Due to the low accuracy of KNN, utilization of SMOTE to rebalance our heavily biased dataset will likely prove to be ineffective.

3.1.3 Random Forest

Random Forest has been the most effective model so far, and has shown promise with the addition of spatial pharmacophore data. It performed the best in the LXRα and LXRβ sets and provided several features importance that were investigated further. Unlike the other two models, RF accuracy increased in the LXRβ set from 71.8% → 74.4%, which was also the lowest total change in accuracy for all models. The RF feature importance attribute was used to determine the most influential pharmacophores in the RF classification (Fig. 1), which showed that some of the spatial data provided by the feature extraction were most valuable when classifying molecule activity. While this doesn't explicitly confirm that the each molecule position is being properly aligned upon instantiation (rendering raw spatial data of centroids fairly useless in machine learning of this scale) it does offer evidence to suggest that the centroids are at least generally being properly aligned. Interestingly the feature with the highest importance in the LXRβ trained model was `acceptor_hydrophobic_min_dist`, which is the smallest distance in a molecule between a H-bond acceptor group and hydrophobic group. This may be a result of the mechanism of action for LXRβ inhibitors, which are hypothesized to prevent a key H-bond interaction deep in the binding pocket by steric hindrance from a large hydrophobic group. To properly align the 'intended' ligand into the LXRβ binding pocket key H-bond interaction between donors and acceptors in the pocket and ligand are needed. While the attention to the minimum acceptor-hydrophobic group distance may be indicative of this structural requirement, it may also be a bizarre artifiact of machine learning, as the accuracy is only 74.4%. In both models H-bond acceptor centroid Y and Z coordinates scored relatively high importance values, with hydrophobic centroid X and Z coordinates not far behind. These will be explored in more detail in a later section, but this lends more credence to a proper molecular alignment and specific requirements for hydrophobic and H-bond acceptor positioning for desired molecule properties. Finally, molecular weight scored in the top 3 features in the LXRα and LXRβ model. While there is no indication for how this property is being utilized (smaller or larger molecules may be more or less selective), this is a property that was revealed in the original RF analysis and held very high correlation to increased variance in the original feature set. Typically in organic molecules hydrophobic structures have higher molecular weight than polar structures, so this may also be related the hydrophobic features discussed earlier that confer LXRβ selective antagonist properties. The core of the binding pocket may also be highly hydrophobic, resulting in higher affinity for molecules with larger hydrophobic cores.

3.2 Spatial Analysis

Three mechanisms were used to visually assess the spatial distribution of the three spatial distribution pharmacophore features: H-bond donor, H-bond acceptor, and hydrophobic centroids of LXRα/β selective molecules. The three mechanisms was a general 3D overlay (Fig. 2), 2D heatmap distribution of each pharmacophore from each unique axis (Fig. 3), and a pharmacophore specific 2D scatterplot (XY plane) with feature importance (Fig. 4 and 5).

3.2.1 H-Bond Donor Centroids

In both datasets the H-bond donor groups were the least represented, with 25 instances in the LXRα group and 9 instances in the LXRβ group. Due to this very small number of data points visually discernible trends must be 'taken with a grain of salt'. In general the H-bond donor centroids had the highest degree of dispursion out of all of the centroid sets. Interestingly, in the LXRα set this level of dispersion was highest along the X axis (-7.5Åto +10Å) followed by the Y axis (-7Åto +6Å) and lowest in the Z axis (-3Åto +4Å). In the LXRβ set however, the highest dispersion was seen in the Y axis (-3Åto +8Å) followed by the X and Z axis (-6.5Åto +1Åand -4Åto +3Årespectively). While the range of the Y dispersion between the LXRα and LXRβ groups was roughly the same ($\approx$ 12Å) the LXRβ group seemed to be centered at $Y \approx +2$Åwhile the LXRα group seemed to be centered at $Y \approx 0$Å. The range in the Z axis was roughly the same, and the average of the cluster was also relatively similar, being roughly 0 in both cases. The biggest differenc came in the X axis, which, of which range and average changed dramatically. In the LXRα the average was roughly $X \approx +2$Å. The LXRβ group however had a single point greater than 0 ($\approx$ +1Å) and was roughly centered at $X \approx -4$Å. It is important to stress that the LXRβ group is very small, and is therefore subject to wild swings in distribution with each new instance. Additionally the number of donors is less acceptors and hydrophobic centroids (which have the same number of instances as each-other) in both datasets, implying that a donor group is not needed to be a selective antagonist. When comparing Fig. 4 and 5, we see that only 4 points are included in both LXRα and LXRβ sets. These findings together imply that the interesting difference in donor distribution may not be consequential when looking for selective LXRβ inhibitors. Finally, this is also supported by the very low importance score by the RF model for all donor features. Due to the small number of points, more information is needed to determine if H-bond donors contribute at all to LXRβ selectivity.

3.2.2 H-Bond Acceptor Centroids

Through both LXRα and LXRβ sets the spatial distribution of the H-bond acceptor centroid remained fairly constant. The spread and average of the X component was $\approx$ 6.5Åand $\approx$ +2Åin the LXRα set and $\approx$ 5.5Åand $\approx$ +2Åin the LXRβ set. The Y and Z components changed very little between the LXRα and LXRβ sets, with a mild increase in dispersivness relative to the Z axis in the LXRβ set and a slight left tail seen relative to the Y axis in the LXRα set. Despite this lack of real change between the datasets, the feature with the highest importance was the `acceptor_hydrophobic_min_dist` which is the smallest distance between an H-bond acceptor group and hydrophobic group present in a molecule. It is important to note that that this was only seen in the LXRβ set, and was not present in the top 3 features of LXRα set. This suggests that the model may be picking up on the proposed mechanism of action of the selective antagonist molecules, which use a hydrophobic substructure to interfere with the H-bonding of the receptor to itself deep within the binding site. The spatial importance was also confirmed by the relatively high importance of the Y and Z coordinates of the H-bond acceptor centroid. Unlike the H-bond donor group, the H-bond acceptor group had a much higher representation in the dataset. This feature set also varied significantly less between the LXRα and LXRβ groups, suggesting that the somewhat tightly bound H-bond acceptor cluster is required for adequate H-bonding. While more research will need to be done to confirm this, this could be used as a very rough 'sifting' method, to separate molecules which obviously have the wrong structure to bind to either LXRα or LXRβ from those that can.

3.2.3 Hydrophobic Centroids

The hydrophobic centroid distribution showed the most reliable change between the LXRα and LXRβ sets. Between the pharmacophore feature centroid distributions, the hydrophobic centroids displayed the highest clustering density (around the origin); Suggesting a strongly hydrophobic core is required to bind to either LXRα or LXRβ. Adherence to the Y and Z axis remained fairly constant between the LXRα and LXRβ sets, but varied considerably more along the X axis. In the LXRα set, the range of the hydrophobic cluster on each axis was 4, 2, and 2 for the X, Y, and Z axis respectively. All clusters were roughly centered at 0, aside from the X axis which was slightly biased in the positive direction. In the LXRβ set the Y and Z compontents of the cluster varied very minimally, where as the X component was dramatically more biased to the positive side of the axis, with 2 points being very slightly negative. Figure 4 and 5 showed that while all hydrophobic centroids with an X coordinate $\gtrsim 0$ have LXRβ activity levels less than 50%, some molecules exist with hydrophobic centroids in this range as well. This trend was reflected in the LXRβ feature importance, where the X coordinate was the feature with the highest importance, followed by the Z coordinate and the total hydrophobic spread in a molecule, which looks at how far each hydrophobic group is from the center of mass of the molecule. In the LXRα importance table the Z coordinate of the hydrophic centroid had the highest importance. This was followed by the hydrophobic average distance, which compares how far each hydrophobic group is away from each other, and finally the X coordinate of the hydrophobic centroid.

Overall the visual analysis of the spatial data yielded several insights that seem to be inline with current hypothesized mechanisms of action. Additionally it suggests that H-bond acceptors are not required for LXRα/β binding. It is important to note that this study fails to investigate the correlation between molecular structure and to what degree a molecule is an LXRβ selective inhibitor due to the constraints of the size of the dataset. This study also suggests that rebalancing of the dataset with methods like SMOTE will likely not be fruitful due to the failings of KNN. This study seems to confirm that we can use spatial data extracted from RDKit to train machine learning models and provide insights into the molecular structures that may confer LXRβ selective inhibition. However, the direct findings from this study, at best, may be used as a very course filter potentially separating the very incorrect molecular structures that couldn't bind either LXRα or LXRβ.

3.3 Transfer Learning in Small Datasets

Over the past several years deep learning has shown its incredible proficiency at prediction with sufficiently large labeled data. However, many problems/situations exist where large repositories of labeled data for a specific task simply do not exist. To remedy this, the transfer learning framework developed to train large deep learning models on general unlabeled datasets, and then provide a small highly specialized dataset, yielding a model with significantly higher performance than would otherwise be possible. In 2020 Li and Fourches introduced the Molecular Prediction Model Fine-Tuning (MolPMoFiT) approach, where a large molecule structure predictor is trained on over 1 million unlabeled compounds in the ChEMBL repository, and then fine tuned to a specific Quantitative structure property/activity relationship (QSPR/QSAR). To amplify training data, Li and Fourches use a method called SMILE enumeration, where each SMILE string is randomized. Each SMILE refers to a unique molecule, but multiple SMILEs can encode a single molecule, yielding SMILE enumeration as a possible method for enflating our dataset, though we would only be able to use SMILE sequences, and thus may use some form of natural language processing. For comparison, the group tested the MolPMoFiT approach to a specialized Convolutional Graph Neural Network and Random Forest model trained on several large, well studied, labeled datasets. In both regression and classification tasks the MolPMoFiT outperformed the large labeled dataset trained models. With this approach we may be able to negate the affects of our small dataset entirely.