Feature selection for protein pockets classification into high and low Dscores

Konstantinos Panagiotakis

(Dated: January 25, 2024)

INTRODUCTION

A fundamental feature of successful drug discovery targeting protein interactions resides in the nature of its interface and druggability associated with each pocket within the surface of a protein.[1][2][3] A binding pocket is defined as a hollow space found either on the surface or inside of a protein, which possesses favorable attributes for accommodating a ligand. The specific arrangement of amino acid residues surrounding a binding pocket determines its physical and chemical properties, and in conjunction with its shape and position within the protein, determines its overall functionality. [2] In this particular scenario, druggability pertains to the probability of a compound resembling a drug to influence or hinder an interaction involving two proteins. Around 60% of drug discovery endeavors faced setbacks because of the target binding site's lack of druggability and the subsequent incapability to bind small drug-like compounds. [1]

Fpocket computes a pocket's druggability score by considering five distinct descriptors: (i) the number of Alpha spheres that match the pocket, (ii) the cavity's density, (iii) the polarity score, (iv) the mean local hydrophobic density, and (v) the portion of nonpolar alpha spheres that align with the pocket. [4]

In this paper, we characterize and reduce the number of features that are most important to classify and define the druggability scores into two classes: 1) High ($\geq 0.5\%$) and 2) Low ($< 0.5\%$). This was achieved with statistical analysis such as the calculation of a correlation matrix amongst the studied features, the viewing of the most highly correlated features through their smoothed distributions, and visualization of a scatter distribution space in which the clustering of the two classes is well defined. Druggability score classification using random forrest and K-mean clustering was then conducted on the most relevant set of features. A voxellization algorithm was then build enabling the turning of a simple 3D coordinate space of each alpha sphere into a visualization of the geometrical features of the binding pockets.

METHODS

Voxellization Algorithm

3D space voxellization can be used for utilizing data-driven approaches that rely solely on structural information to visualize ligands as spatial fields within target protein pockets. [5] Proteins exhibit a diverse range of shapes, yet nature often utilizes recurring patterns at various levels. Identifying and categorizing these similarities is crucial for comprehending evolutionary processes. [6] The data used for space voxellization was taken from fpocket. [8] The steps to complete this algorithm were as follows:

1. Filter the pockets to retain only those with a druggability score greater than 0.5.
2. Assume a single pqr file as input, containing the alpha spheres that define the pocket.
3. Calculate the bounding box (min, max, xyz) of the alpha spheres, counting their centers and radii. Name these bounding box coordinates as $x_{\min}$ and $x_{\max} \in R^3$.
4. Add a margin L to each side of the bounding box.
5. Determine the center of the bounding box and designate it as $x_0 \in R^3$.
6. Compute the coordinates of an imaginary 3D grid centered on x_0. The grid spacing must be dX (assumed to be 1 Å). Ensure the grid covers the entire bounding box + L, rounding up as necessary to maintain a square grid.
7. Assign coordinates x_i and values v_i to each cell of the grid. For each cell, set v_i to 1 if x_i is inside any alpha-sphere, otherwise set it to 0.
8. Return the grid, including its coordinates, bounding box, and values.

Further use of this voxellized representations of protein pockets enable efficient ligand binding prediction, virtual screening for drug discovery, structural analysis, and protein design and engineering by capturing the spatial distribution of pocket features for analysis, prediction, and modification.[5][7]

Statistical Analysis

To prepare for the classification of the druggability score based on biochemical and geometrical properties of the binding pocket, statistical analysis was executed to assess which features where the most highly correlated to the druggability score before engaging in machine learning classification via random forrest and K-means clustering. This was done to decrease the number of features used for the classification of the data in the two classes and for the detection of redundant descriptors. The druggability score distribution for each of the proteins was obtained and is well detailed and explained in the Results section. This allowed for a simple visualization of the general distribution of each pocket in the proteins that were sampled, and allowed us to define the High and Low bound of each Druggability Score, as referenced in [1]. A correlation matrix was then built in order to assess the most highly correlated features. This allowed us to focus on two features that had the greatest proportional correlation with the Druggability Score, as shown in the Results section. The dataset was then separated in two subsets, one that had all of the binding pockets with High ($\geq 0.5\%$) and one that had all of the Low ($< 0.5\%$) Druggability Scores. The binding pockets where then mapped to a 3D space where the coordinates where the three most highly correlated features with the Druggability Score. A smoothed distribution for high and low druggability pockets was then obtained for each of these three features and plotted in three separate histograms along with the mean and standard deviation for each one of the distributions.

Random Forest

The Random Forest Classifier is a type of machine learning model that integrates several decision trees to make predictions. [9] This method was then used for both feature importance selection and classification of the pockets into their respective classes. Every value ($< 0.5\%$) was mapped to 0 and defined as the Low Druggability Class, while every value ($\geq 0.5\%$) was mapped to 1 and defined as the High Druggability Class in the training set. The dataset was divided into a training and a test set, with an 80-20% split. The dataset was balanced to have approximately 50% of pockets in the Low Druggability Score spectrum and 50% of pockets in the High Druggability Score spectrum. This balanced approach ensured that the algorithm trained on a dataset with an equal distribution of classes, rather than a skewed distribution tailored toward the Low Druggability Score Class. The Gini criterion was used as a classifier basis, which is a measure used to assess the impurity or disorder of a node within a decision tree.

$$\text{Gini} = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{1}$$

In equation 1, C represents the total number of classes, and p_i represents the probability that an observation belongs to class i. The Gini criterion evaluates the impurity of a node by subtracting the sum of the squared probabilities of all classes from 1. The RF algorithm was trained on a balanced dataset, and then its performances were tested on the test set, and the binding pockets where classified into the 2 classes as summarized in the Results section. Feature importance was then conducted with a built in "feature importance" function from the sklearn library, and the most important features are displayed in the Results section. Feature importance in Random Forest (RF) refers to a technique used to assess the relative importance of each feature in the predictive performance of the RF model. It provides insights into which features have the most influence on the model's decision-making process.[10]

K-Means Clustering

After the statistical analysis was conducted, it was clearly noticeable that the dataset was able to be clustered into 2 different well defined clusters centered at two different points in the 2D space defined by the two most important features, Number of Alpha Spheres and Mean Local Hydrophobic Density, as we will see in the Results section. This began the exploration of using the K-means clustering for the classification of binding pockets into the two different classes. K-means clustering is an unsupervised machine learning algorithm that groups similar data points into clusters. [11] It involves initializing cluster centroids, assigning data points to the nearest centroid, updating the centroids based on the assigned data points, and repeating this process until convergence. The algorithm provides cluster assignments for each data point and the coordinates of the cluster centroids as the output. [11]

$$\mathrm{C}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_i \tag{2}$$

In this equation, C_k denotes the centroid of cluster k, N_k represents the number of data points assigned to cluster k, and x_i corresponds to the coordinates of the i-th data point within cluster k. The equation calculates the centroid by taking the mean of the coordinates of all data points assigned to that cluster. The sklearn.cluster library was used to conduct this classification. A balanced dataset was also used in this case as in the Random Forrest case. Two clusters were used as only two classes were identified for the classification of each data point. Its performances and centers of the clusters were obtained, and a confusion matrix was calculated alongside the visualization of each cluster in a 2D space defined by the Number of Alpha Spheres and the Mean Local Hydrophobic Density.

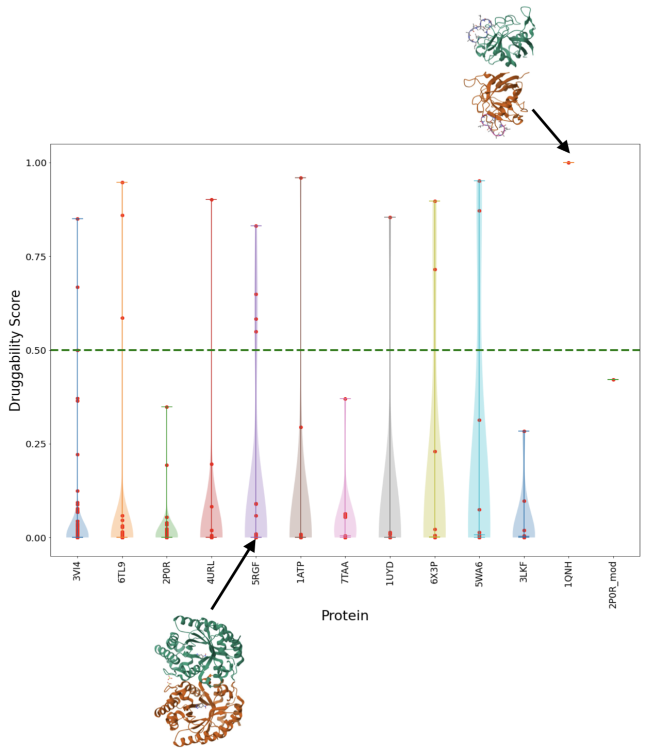

FIG. 1. Violin plot distribution of binding pockets separated in their respective protein. In green, we have the threshold line where high and low druggability scores are classified. Red dots, binding sites above the line will be labelled as High, and vice versa for the Low.

RESULTS

Dataset Class Separation

Before beginning the analysis of the data, the druggability scores where converted into two classes. Binding pocket data was obtained from the fpocket library[8], values that were ($< 0.5\%$) where mapped to 0, or to the Low Class, whether values with ($\geq 0.5\%$) where mapped to 1, or to the High Class. This allowed to perform a binary separation of the data into two separate distinct classes. 0.5 seemed like the best threshold to distinguish between a binding pocket with high and low druggability. The score analysis in [12] also helped with the threshold choice of 0.5 that states that binding pockets with high druggability scores are usually in the region of 0.8 and above. Fig.1 showcases the violin plot distribution of the druggability scores for each pocket separated in their respective protein space. As shown in the figure, the distributions of the binding pocket druggability is centered around the lower druggability score spectrum mostly due to their structural constraints and functional relevance.

Statistical Analysis

After the class separation of the dataset into two distinct Druggability Score classes, statistical analysis was conducted to firstly determine which out of the features in the dataset where the most relevant features that could be used for the classification of the Druggability Score. Looking at the correlation matrix, Fig.2, we can ultimately conclude that the Number of Alpha Spheres and the Mean Local Hydrophobic Density are the most highly correlated, while Flexibility showcases no correlation with the features in the dataset. This matrix tells us that pockets with Low Druggability Scores have a lower Mean Local Hydrophobic Density, which correctly proves the evidence demonstrated in[13], that hydrophobic grooves present on the interface of protein-protein interactions serve as significant areas for the development of small molecule inhibitors. Overall, the observed correlation suggests that larger, (as in our case the more the number of alpha spheres, the larger the volume, and the higher the druggability score generally), well-defined hydrophobic pockets are more likely to bind drug-like ligands. [1] These grooves play a crucial role in facilitating the binding and formation of stable interactions between partner proteins. [13]

After the three features of interest were selected from the correlation matrix analysis, the distribution of both druggability score classes, 0 for Low and 1 for High was plotted in a 3D space where the Number of Alpha Spheres, the Mean Local Hydrophobic Density and Flexibility where the three dimension of this space in Fig.3. As we can see for this plot, the higher the local hydrophobicity and Number of Alpha sphere, the more likely the binding pocket will be in the High druggability class, and vice versa for the Flexibility. These results are well showcased in the distributions of these three features separated between the two classes and are shown in Fig.4

Looking at Fig.4, we can easily see that the two classes are separated into two clusters in both the Mean Local Hydrophobic Density space, and the Number of Alpha Sphere space. These results further confirmed that the unsupervised learning technique of K-means clustering could be a useful tool for the classification of the binding pockets given these two features into the two respective Druggability Classes. This was also further supported by Fig.5 that clearly showcases a clear clustering of the two distributions of the classes in this 2D feature space, as the centers of the clusters distinctly show. This figure reiterates what is stated above, High Druggability Score pockets have a higher Mean Local Hydrophobic Density and Number of Alpha spheres than the Low Druggability Score pockets. An increased surface area of a protein can positively impact its druggability score as we can see from the results. Firstly, a larger surface area offers more opportunities for the presence of multiple or larger binding

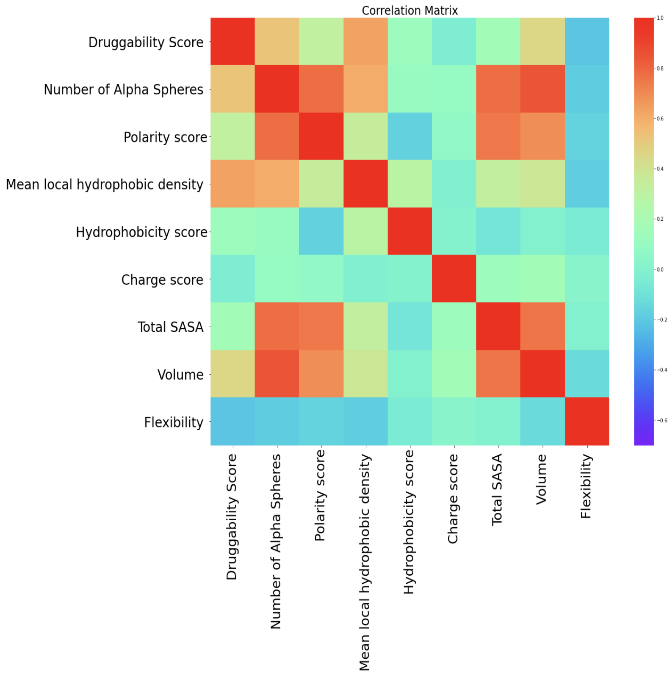

FIG. 2. *Correlation matrix of a subset of 9 features. Their relevancy was determined by analyzing both the Low and High Druggability datasets, and looking at the trends between diverse level of druggability and the values of these features. This matrix shows the correlation amongst these 9 features. We are interested in looking at the correlation with the Druggability Score. As we can see from the correlation matrix, the Number of Alpha Spheres and the Mean Local Hydrophobic Density are the most highly correlated. Volume is also highly correlated, but it's a redundant descriptor as it's almost a synonym for Number of Alpha Spheres.*

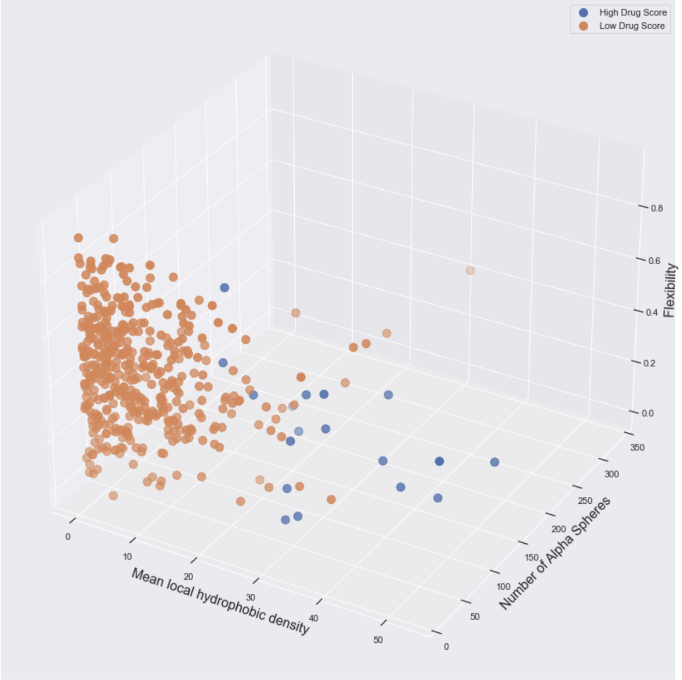

FIG. 3. *3D space where the Number of Alpha Spheres, the Mean Local Hydrophobic Density and Flexibility are the three axes defining the space. This plot was created to try to analyze the distribution of high and low Druggability Classes throughout the three most relevant features of interest to determine whether a clustering approach could be used for the classification of the pockets into one of the two classes. In Orange we see pockets partaining to the Low Druggability Score class, in blue we see pockets partaining to the High Druggability Score Class.*

sites, which enhances the chances of identifying suitable targets for drug binding. Secondly, it improves accessibility, facilitating the interaction between drugs and the protein. Thirdly, a greater surface area typically correlates with higher conformational flexibility, enabling the protein to accommodate a wider range of drug molecules with diverse shapes and sizes. Lastly, a higher surface area also raises the probability of protein-protein interactions, potentially revealing additional targets for drug development. On the other hand, a higher hydrophobicity density can also increase its Druggability Score due to several reasons. It promotes favorable binding interactions with hydrophobic drug molecules, it facilitates targeting of membrane proteins, it contributes to structural stability, and aids in drug delivery through cell membranes. [1] These factors enhance the likelihood of effective drug binding and improved therapeutic outcomes. In regards of Flexibility, looking at Fig.4 there seems to be almost an inverse correlation with the Druggability Score. This does not match with the results obtained in [14] where it's stated that flexibility of proteins, or when proteins undergo conformational changes as they bind to ligands in ligand-bound structures, has a positive correlation with their druggability scores. This increase is attributed to the presence of structural characteristics that are more conducive to drug binding. However in the results obtained in the statistical analysis the correlation between the flexibility and the Druggability Score does not present this positive correlation that has been obtained in [14].

Random Forrest

Two methods where used for the classification of binding pockets into High and Low Driggability Scores classes. Random forest was the first method used. Its performances are summarized in Table.1. The algorithm makes 93% of correct predictions, which given the size of the training set is an impressive performance. This method was also used to perform feature importance characterization and it's displayed in fig.6. Overall the random forest algorithm was a great method to both classify the binding pockets in their respective classes, and quantify the feature importance in the classification pro-

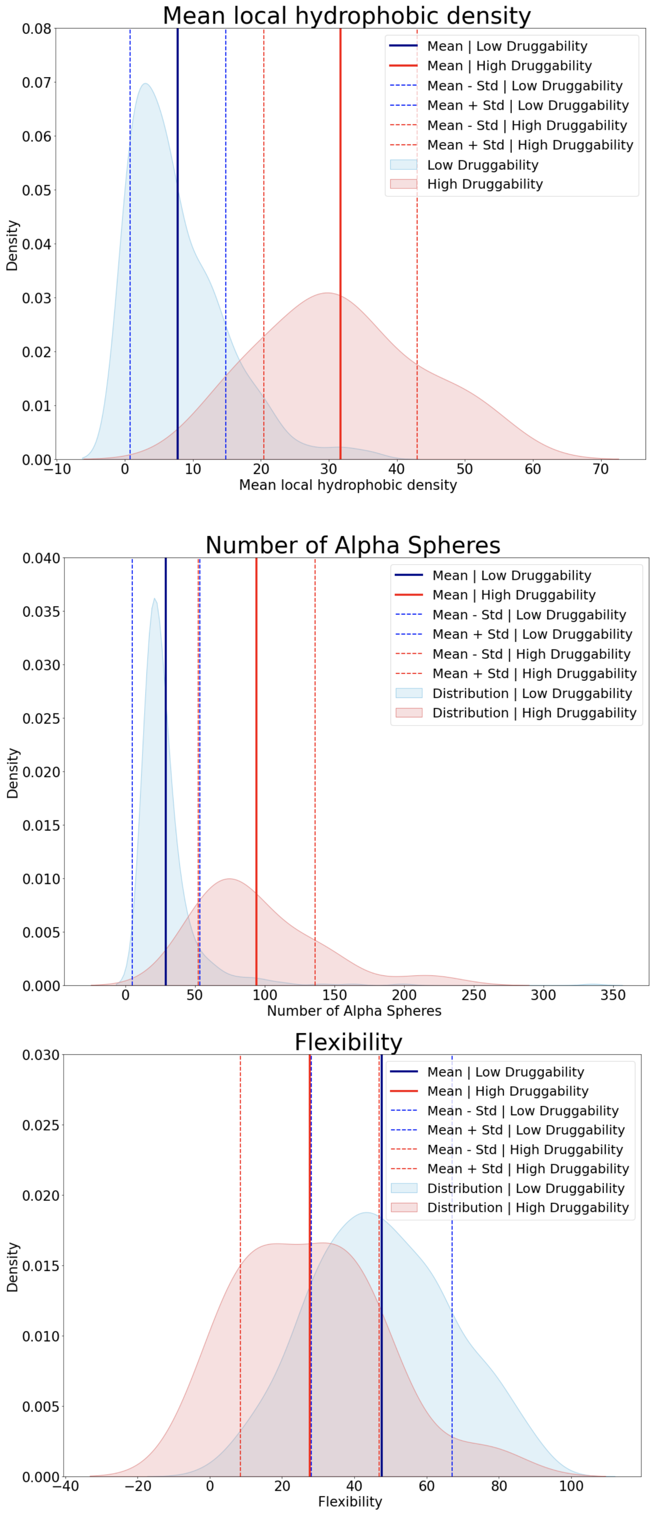

FIG. 4. Top - Smoothed Binding Pocket distribution in the Mean Local Hydrophobic Density space. Middle - Smoothed Binding Pocket distribution in the Number of Alpha Spheres Density space. Bottom - Smoothed Binding Pocket distribution in the Flexibility space. In Blue the distribution of the respective feature in the Low Druggability Score Class along with the mean and standard deviation. In red the distribution of this feature in the High Druggability Score Class along with the mean and standard deviation.

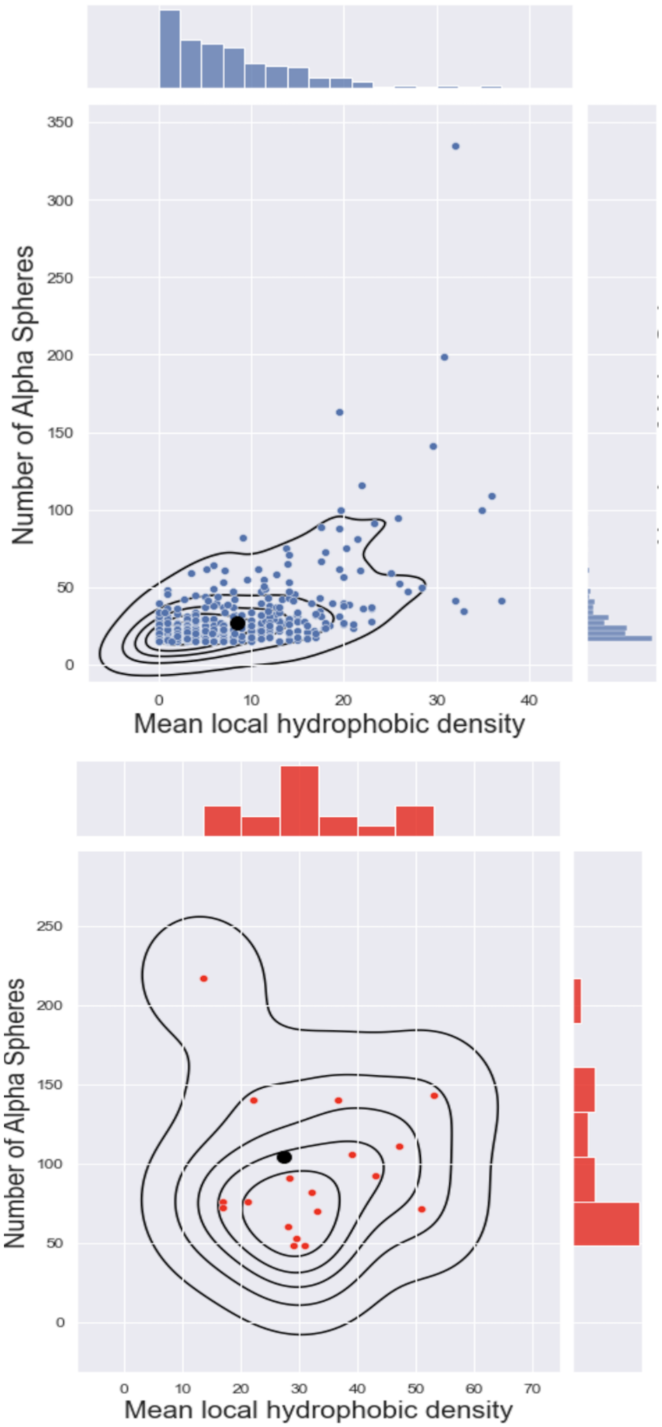

FIG. 5. Top - Low Druggability Score clustering of binding pockets. Bottom - High Druggability Score clustering of binding pockets. 'Low', blue, and 'High', red, Druggability binding pockets distribution across a 2D space where the basis of the space are the two most relevant features, Mean Local Hydrophobic Density and Number of Alpha Spheres. The centers of the two clusters were calculated with the K-mean clustering algorithm, shown in a later subsection of the results. As this figure shows, there are two distinct clusters of classes within the distribution in this 2D space. This result was a crucial factor in determining which machine learning algorithm to use for the classification of the binding pockets into the two classes.

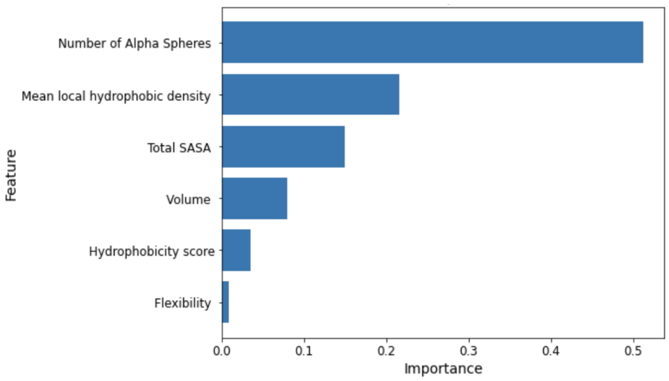

FIG. 6. Feature importance vs. Feature. This figure shows the Importance on the classification of the binding pockets into the two classes. These are 7 out of the 20+ features that were contained in the dataset.

cess. These results were used to restrict the input to the K-mean clustering even more, from 7 total input features, to 2 input features, the Number of Alpha Spheres and the Mean Local Hydrophobic Density.

Metric	**Values**
Accuracy	0.93
Precision	0.86
Recall	1
F1 Score	0.92

TABLE I. Random Forest Performances - These measures, including accuracy, precision, recall, and F1 score, are used to evaluate the performance of a Random Forest classifier, with accuracy measuring overall correctness, precision focusing on correctly predicted positives, recall assessing the classifier's ability to identify positive instances, and the F1 score providing a balanced measure of performance by combining precision and recall.

K-Means Clustering

The second classification method used was the unsupervised learning technique of K-means clustering. Its performances are summarized in the confusion matrix in fig.7. The algorithm has better performances on the classification of High Druggability Scores compared to the Low classification prediction. In fact it predicts the High Druggability Scores with an accuracy of 88% compared to the 83% of the classification of the Low class. Over the testing set it has a loss of 14.3%. It was also used for the centroid calculation of the two clusters and they are shown in fig.5. The center of the low druggability cluster is: 8.47 along the Mean Local Hydrophobic Density axis, and 27.38 along the number of alpha particles axis. On the other hand, the center of the high druggability cluster is: 31.45 along the Mean Local Hydrophobic Density axis, and 104.0 along the number of alpha particles axis. Looking at the centroids, the strong connection between hydrophobicity and drug binding sites is evident. However, the underlying assumption of a model solely based on hydrophobicity suggests that the optimal binding site should be a sealed and lipid-friendly cavity [16], and the results shocase that the higher the hydrophobicity of the site is, the higher the Druggability Score of the pocket, which solidifies this further. On the other hand, the higher Number of Alpha Spheres in a protein pocket positively influence the Druggability Score as it can indicate a larger and more complex pocket, which could potentially accommodate a wider range of drug-like molecules. A larger pocket may offer more binding opportunities and thus be considered more druggable.

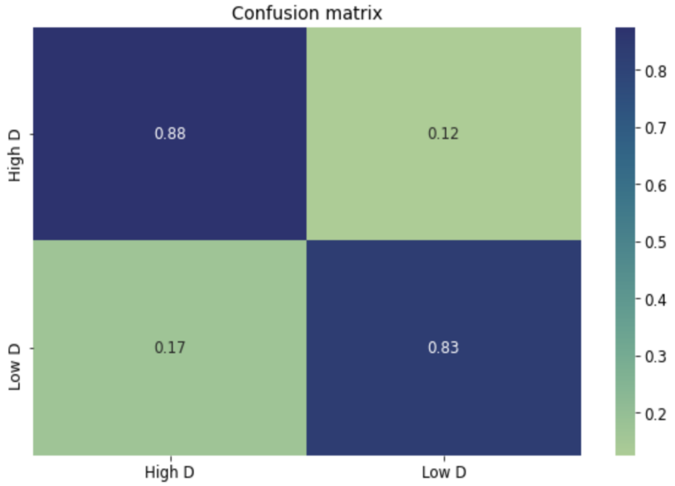

FIG. 7. Confusion matrix of K-means algorithm performance. The diagonal shows the true positives with performances of 88% and 83% respectively.

Voxellized space visualization

The final step conducted was to use the voxellization algorithm built to visualize the binding pockets. We can see from fig.8 that in blue we have a binding pocket of the High Druggability Score kind made of two sites that are substantially larger, more than 5x the amount of the Low Druggability Score represented in red. This further showcases that by increasing the surface area and the volume of the binding pocket, you directly increase the probability of finding a pocket with a high Druggability Score within its surface. A spacing of 2 units was used to make the 1QNH pocket dues its expensive computability, and a spacing of 1 unit was used for the 5RGF pocket. This results allows to visualize the pockets in a 3D coordinate grid, and it's a very useful tool for other types of pocket analysis as seen in [7].

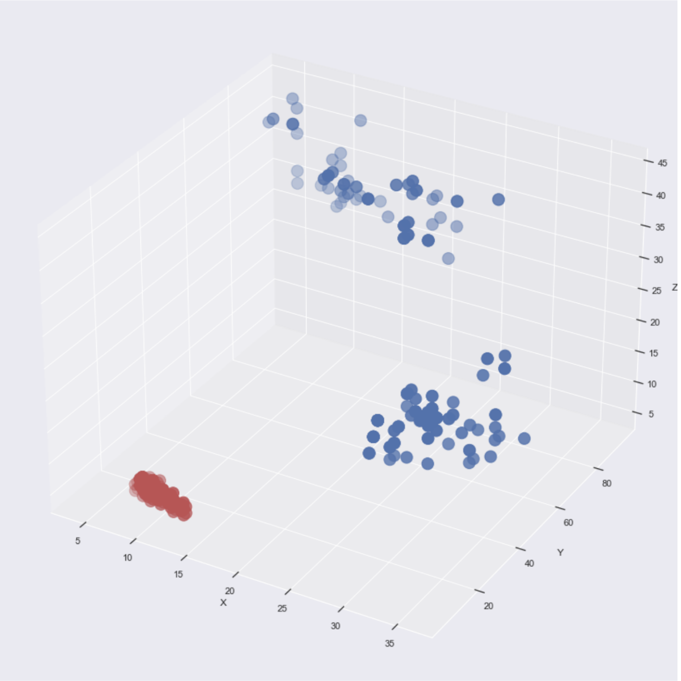

FIG. 8. Voxellized 3D space of the low drug scoring pocket from the 5RGF protein in red, and the high drug scoring pocket from the 1QNH protein in blue. The space showcases the pixels of the coordinate grid where the alpha spheres are present in the grid.

CONCLUSIONS

We introduced and showcased the classification of protein pockets into High and Low Druggability score classes with the use of K-means clustering techniques. Statistical analysis was conducted on the features of the dataset to understand the correlation amongst them for the ultimate classification of the data into the two classes. The main motivation for this classification and the reason that the prediction of pockets is a crucial area in computational biophysics is that when a molecule binds to a protein, its biochemical behaviour is affected, furthermore the identification of candidate binding sites is a key aspect to drug design as also reiterated in [16]. A 3D space voxellization of the Alpha Spheres contained in the protein pockets was designed for the ultimate purpose to obtain a visual and geometrical representation of the protein pocket in question. As mentioned in the methods section, this technique can be used in future work to enables efficient ligand binding prediction, virtual screening for drug discovery, structural analysis, and protein design and engineering by capturing the spatial distribution of pocket features for analysis, prediction, and modification.[5][7] After the design of this algorithm, the statistical analysis of the features in the dataset was conducted, and it was concluded that the Mean Hydrophobic Density and the Number of Alpha Spheres were the most highly correlated features to the Druggability Score, and Flexibility did not present correlations amongst the features in the dataset. This analysis proved to be useful for feature reduction in the classification of the pockets into High and Low Druggability Scores was conducted. It proved to be also useful to detect redundant descriptors such as Number of Alpha Spheres and Volume, which are highly correlated between one another, and only one of the two is needed for the DScore prediction task. Feature importance was then conducted with the use of the random forrest technique and the results showcased in Fig.6, show that Mean Hydrophobic Density and the Number of Alpha Spheres are the most relevant features for the Druggability Score Prediction. This showcases the relevance and the influence of geometrical and biochemical features of the protein pocket binding site in the classification of that site. The unsupervised learning technique of K-Means clustering was then used to classify the pockets into the two classes with a Loss of 12% for the High Druggability Class classification and a Loss of 17% for the Low Druggability Class classification. This technique predicts which cluster each data point in the test set belongs to based on the model's learned cluster centroids, hence enabling us to achieve a classification task. Looking at the Results section, the centroids of the cluster of the Low Druggability Score pockets had lower values of Mean Local Hydrophobic Density and Number of Alpha Sphere then the centroids in the High Druggability Class. The obtained results in this paper are further solidified by the PCA analysis in [4] of the five features used in the calculation of the Druggability Score reveals that the Number of Alpha Spheres fitting the pocket and the Mean Local Hydrophobic Density contribute the most to the Druggability Score, therefore given these two features of a pocket, one can use K-means clustering to classify the pocket into a High and Low Druggability Score class, therefore making it easier to filter out non useful binding sites for drug discovery.

[1] Alzyoud, L., Bryce, R. A., Al Sorkhy, M., Atatreh, N., Ghattas, M. A. (2022, May 13). Structure-based assessment and druggability classification of protein–protein interaction sites. Nature News. https://www.nature.com/articles/s41598-022-12105-8

[2] Antonia Stank, Daria B. Kokh, Jonathan C. Fuller, and Rebecca C. Wade, Protein binding pocket dynamics — accounts of chemical research. (2016). https://pubs.acs.org/doi/10.1021/acs.accounts.5b00516

[3] Halgren TA. Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model. 2009 Feb;49(2):377-89. doi: 10.1021/ci800324m. PMID: 19434839.

[4] Cunha, A.E.S.; Loureiro, R.J.S.; Simões, C.J.V.; Brito, R.M.M. Unveiling New Druggable Pockets in Influenza Non-Structural Protein 1: NS1–Host Interactions as An-

tiviral Targets for Flu. Int. J. Mol. Sci. 2023, 24, 2977. https://doi.org/10.3390/ijms24032977

[5] Aguti Riccardo, Gardini Erika, Bertazzo Martina, Decherchi Sergio, Cavalli Andrea, Probabilistic Pocket Druggability Prediction via One-Class Learning, Frontiers in Pharmacology 13, 2022

[6] Guzenko, D., Burley, S. K., Duarte, J. M. (n.d.). Real time structural search of the Protein Data Bank. PLOS Computational Biology. https://journals.plos.org/ploscompbiol/article?id=10.1371

[7] Qinqing Liu, Peng-Shuai Wang, Chunjiang Zhu, Blake Blumenfeld Gaines, Tan Zhu, Jinbo Bi, Minghu Song, OctSurf: Efficient hierarchical voxel-based molecular surface representation for protein-ligand affinity prediction, Journal of Molecular Graphics and Modelling, Volume 105, 2021,107865, ISSN 1093-3263, https://doi.org/10.1016/j.jmgm.2021.107865

[8] fpocket, Peter Schmidtke, Vincent Le Guilloux, Mael Shorkar, https://github.com/Discngine/fpocket/tree/master

[9] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[10] Kursa, Miron Rudnicki, Witold. (2011). The All Relevant Feature Selection using Random Forest.

[11] Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics 2020, 9, 1295. https://doi.org/10.3390/electronics9081295

[12] Shin WH, Kumazawa K, Imai K, Hirokawa T, Kihara D. Current Challenges and Opportunities in Designing Protein-Protein Interaction Targeted Drugs. Adv Appl Bioinform Chem. 2020 Nov 12;13:11-25. doi: 10.2147/AABC.S235542. PMID: 33209039; PMCID: PMC7669531.

[13] Higueruelo, A. P. et al. Atomic interactions and profile of small molecules disrupting protein-protein interfaces: The TIMBAL database. Chem. Biol. Drug Des. 74, 457–467 (2009).

[14] Alzyoud L, Bryce RA, Al Sorkhy M, Atatreh N, Ghattas MA. Structure-based assessment and druggability classification of protein-protein interaction sites. Sci Rep. 2022 May 13;12(1):7975. doi: 10.1038/s41598-022-12105-8. PMID: 35562538; PMCID: PMC9106675.

[15] Schmidtke, Peter Barril, Xavier. (2010). Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites. Journal of medicinal chemistry. 53. 5858-67. 10.1021/jm100574m.

[16] Luca Gagliardi, Andrea Raffo, Ulderico Fugacci, Silvia Biasotti, Walter Rocchia, Hao Huang, Boulbaba Ben Amor, Yi Fang, Yuanyuan Zhang, Xiao Wang, Charles Christoffer, Daisuke Kihara, Apostolos Axenopoulos, Stelios Mylonas, Petros Daras, SHREC 2022: Protein–ligand binding site recognition, Computers Graphics, Volume 107, 2022, Pages 20-31, ISSN 0097-8493, https://doi.org/10.1016/j.cag.2022.07.005.