Evaluating Positional Analog Scanning as a Method for Lead Optimization

Vishnu Mothukuri
IIITD
vishnu21502@iiitd.ac.in

Dr. Arjun Ray
IIITD
arjun@iiitd.ac.in

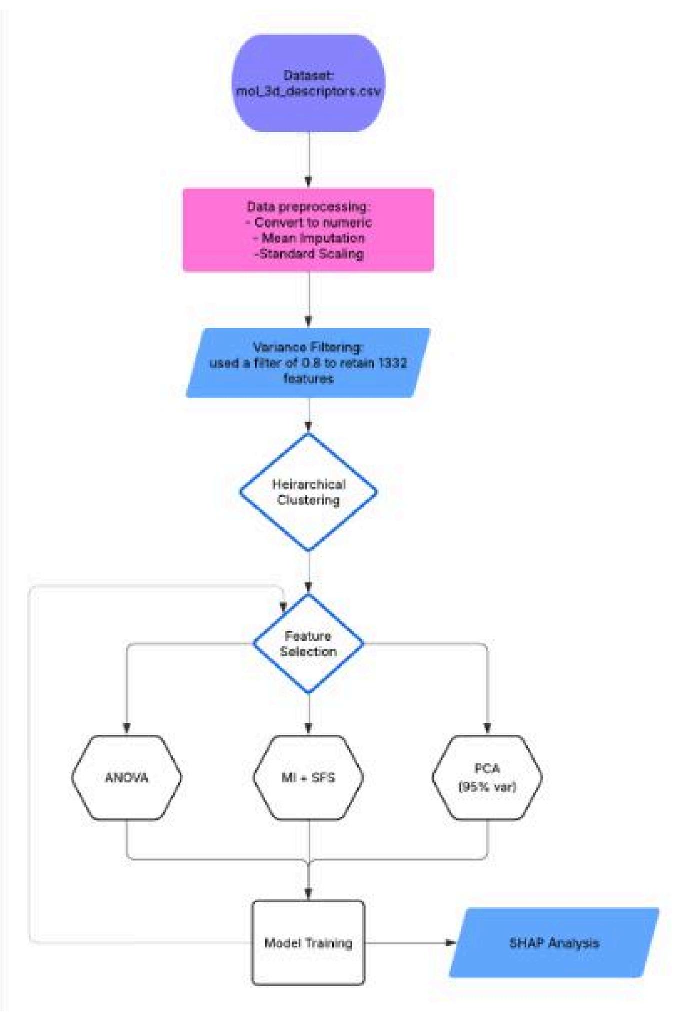
Dr. Jacob Kongsted
University of Southern Denmark
kongsted@sdu.dk

Abstract

This project uses Positional Analog Scanning and machine learning to investigate how single-atom and functional group additions affect potency change. Advanced techniques like MISFS, PCA, and hierarchical clustering were combined with Deep Learning and AutoKeras to build classification models. The binary classification and SHAP offer a robust, data-driven framework to accelerate lead optimization.

Dataset and Pipeline

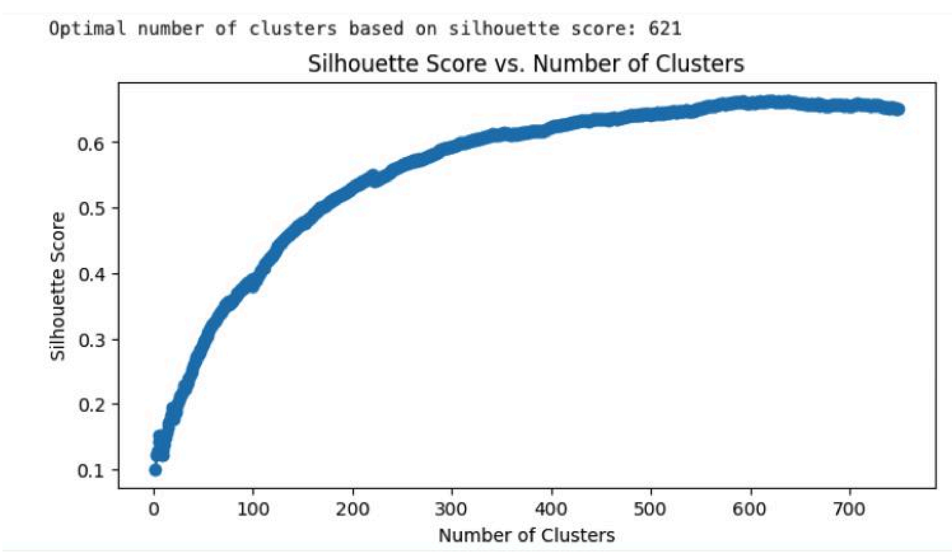
- 56,826 molecular pairs with 3,659 descriptors. Mean imputation and standard scaling were used to fill missing values and normalize the descriptors. Variance and correlation thresholds reduced noise and selected informative features.



Feature Selection and Model Building

1. Hierarchical Clustering

Hierarchical clustering was used to group correlated features based on Euclidean distance and ward criteria. Used silhouette scoring to find the optimal number of feature clusters (621, with score ≈ 0.66). This step reduced the feature space from 1,332 post-filtering features to a representative set for modeling.

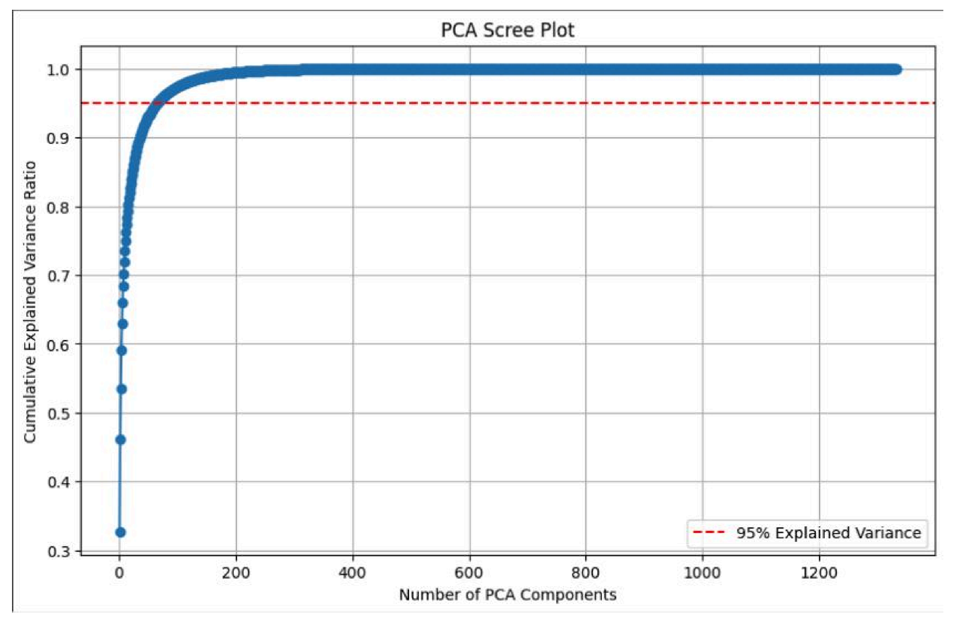


2. MISFS (Mutual Information + Sequential Forward Selection):

Identified 176 high-impact features using mutual information and then ran sequential forward selection. However, it achieved cross-validated accuracy ≈ 0.185, showing baseline performance.

3. PCA (Principal Component Analysis):

Transformed the feature space into 66 uncorrelated principal components capturing 95% of total variance. Enabled dimensionality reduction and stabilized model performance, especially for deep learning models.



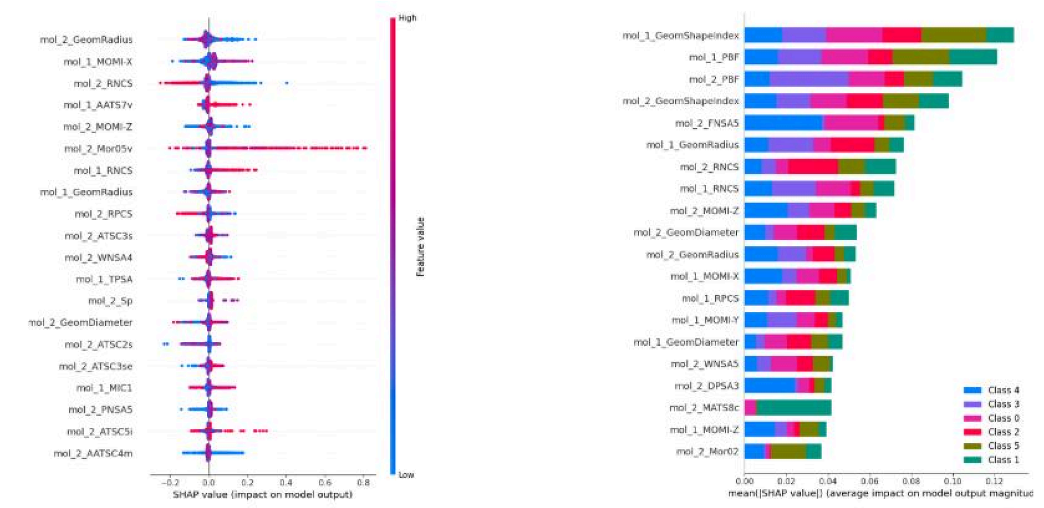
4. Model Performance

Transformed the original multiclass problem into binary classification (improved vs reduced potency) and trained Random Forest, XGBoost, AutoML, AutoKeras algorithms. AutoKeras with PCA offered a balance between dimensionality and signal strength.

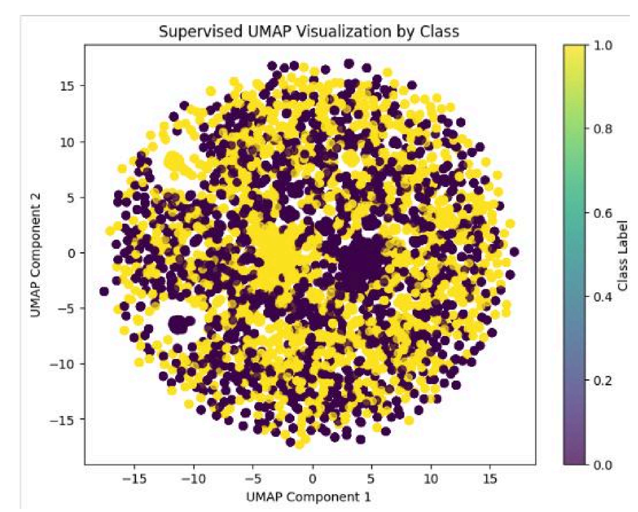
Model	Feature Selection	Number of Features	#	Accuracy (%)	ROC AUC	Key Hyperparameters / Settings
XGBoost	Variance Selection	1332		18.3	0.512	n_estimators=100, random_state=42, t
HistGradientBoosting	Variance Selection	1332		18.2	0.514	max_iter=100, random_state=42
LightGBM	Variance Selection	1332		18.4	0.512	n_estimators=100, random_state=42
CatBoost	Variance Selection	1332		17.8	-	iterations=100, random_seed=42, verb
Decision Tree	Variance Selection	1332		18.1	-	random_state=42
Random Forest, XGBoost	ANOVA	100, 200, 300, 400, 500, 600, 700		16.8	0.52	n_estimators=100, random_state=42, r
Random Forest	Tree based selection	621		16.49	0.521	n_estimators=100, random_state=42, r
Logistic Regression	L1 logistic selection	0	-		-	-
KNeighborsClassifier	MI + SFS	176		18.49	-	n_neighbors=5
XGBoost_1_AutoML_5_2025032	Correlation based selecti	501		17.37	-	max_models=20, max_runtime_secs=6
AutoKeras DNN	Clustering based selectic	621		51.23	0.5021	max_trials=10, epochs=20
XGBoost_1_AutoML_2_2025032	Variance Selection	1332 (binary classification)		64.83	0.531726	max_models=20, max_runtime_secs=6
GBM_1_AutoML_2_20250328_9	Variance Selection	1332 (binary classification)		64.83	0.522246	max_models=20, max_runtime_secs=6
AutoKeras DNN	PCA (95% variance)	66		68	0.654	max_trials=10, epochs=50
Supervised Autoencoder (DL)	Variance Selection	1332		51	0.519	validation_split=0.2, epochs=100, batc

Key Findings

Used SHAP values to interpret model predictions. Identified several high-contributing features that matched known chemical descriptors.



Supervised UMAP failed to clearly separate classes in 2D.
This highlights the high-dimensional complexity and overlap in the potency distribution.



Future Work

- Integrate 3D conformers and quantum descriptors like docking scores, pharmacophore fingerprints.
- Try TabPFN and SAINT for improved modeling for tabular data.
- Exploring transfer learning and ensemble stacking.