Evaluating positional analog scanning as a method for lead optimization

Student Name: Vishnu Mothukuri
Roll Number: 2021502

BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Bioscience
on 23/04/2025

BTP Track: Research

BTP Advisor
Dr. Arjun Ray

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

I hereby declare that the work presented in the report entitled **"Evaluating positional analog scanning as a method for lead optimization"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Bioscience* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Arjun Ray**. Due acknowledgments have been given in the report for all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... **Place & Date: .....**
Vishnu Mothukuri

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date: .....**
Dr. Arjun Ray

Abstract

Understanding how molecular modifications influence biological activity is a cornerstone of drug discovery and development. This project leverages Positional Analog Scanning, a powerful technique in medicinal chemistry, to investigate how single-atom additions and functional group modifications impact the potency of lead compounds. By integrating advanced machine learning models such as Chemprop, this study uncovers complex structure-activity relationships (SAR) and statistical analyses to uncover patterns and predict biological activity changes. Morgan fingerprinting and external molecular descriptors are employed for feature engineering, while data preprocessing techniques address challenges such as skewed distributions and class imbalance. This work not only advances our understanding of SAR but also provides a scalable, data-driven framework to accelerate lead optimization, offering transformative potential for the drug development pipeline.

Keywords: Positional Analog Scanning, Lead Optimization, Medicinal Chemistry, SAR, Potency Prediction, Machine Learning, Morgan Fingerprinting, Functional Group Analysis.

Acknowledgments

I would like to thank **Dr. Arjun Ray** for his constant guidance and support throughout this project. I would like to thank **Dr. Jacob Kongsted** for giving me the opportunity to work on this project. I would also like to thank **Ms. Riddhi Sharma** for mentoring me.

Work Distribution

Vishnu Mothukuri has done all the work mentioned in this report.

Contents

Chapter 1

Introduction

Positional Analog Scanning is a widely used technique in medicinal chemistry to study the structure-activity relationship (SAR) in potential drug molecules. It involves modifying different positions of a lead compound, a compound that shows potential to be a drug, to study how these changes affect the compound's biological activity. This helps us identify the modifications or additions that can improve the pharmacokinetic properties of the drug.

Despite its advantages, implementing analog scanning poses several challenges, particularly in predicting how specific alterations to chemical structures will influence overall drug behavior. This includes effects on molecular conformation, electrostatic potential, and interaction dynamics within biological systems. The primary research problem is determining the extent to which analog scanning can enhance the pharmacological profiles of drug candidates.

Lead optimization is the most important step in the drug development process. But before drug development comes drug discovery, which includes target and lead identification.

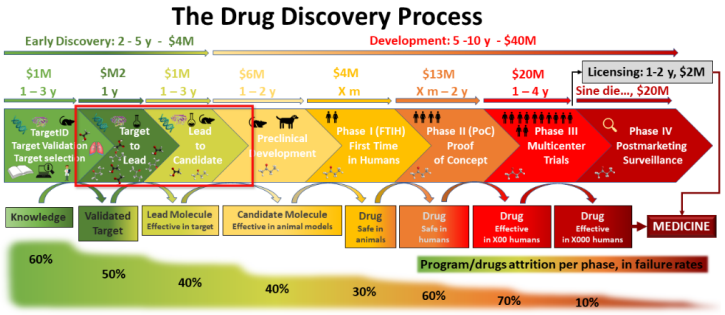

Figure 1.1: Drug Discovery Process

Ever since the human genome was mapped in 2001, a large number of potential genetic drug targets have been identified. An example of this is the GPR40 gene, which was identified in 1997. This is a class A G-protein coupled receptor that responds to free fatty acids, is expressed selectively in pancreatic β -cells, and regulates insulin secretion. In 2005 it was discovered that Arg211His mutation in human GPR40 impaired insulin secretion. This was a new potential target for the treatment of diabetes. Similarly, drug targets can also be identified by expression patterns, knock-out animals, correlation of natural mutations to diseases, and other observations connecting a target to a particular disease.

Identifying lead compounds involves molecular docking (MD) simulations, high-throughput screening(HTS), virtual screening, and structure-based de novo design. HTS is a contempo-

rary approach that has a high chance of finding new chemotypes. No knowledge of receptors or ligands is required, but it is very resource-demanding as it involves screening synthetic libraries or natural product libraries. The ligand-based approach is the classical approach, where no information about the receptor is required. The de novo approach is a futuristic approach where the lead molecule is obtained through computational homology modeling. It involves designing drugs from scratch using only the structure of the target. However, detailed knowledge of receptor structures is currently lacking.

The overall aim of lead optimization is to create a candidate molecule to be used for further clinical trials before it is launched into the market. To achieve this, the compounds at the lead optimization stage are examined and characterized. The main goals for these compounds are to:

- Improve potency / binding affinity to target
- Improve selectivity
- Change functional activity: converting agonist to antagonist and vice-versa
- Optimizing Pharmacokinetic properties:
  - Absorption
  - Digestion
  - Metabolism
  - Excretion

While positional analog scanning is a powerful approach, predicting the impact of specific molecular modifications remains challenging. Changes in potency are influenced by a variety of factors, including molecular conformation, functional group interactions, and physicochemical properties such as logP and heavy atom count (HAC). These influences are often non-linear and context-dependent, making it difficult to establish universal rules that link structural modifications to changes in activity. This complexity is compounded by the lack of tools to systematically analyze these relationships across diverse chemical scaffolds.

This project focuses on analyzing and predicting potency changes in molecular structures resulting from single-atom additions or functional group modifications. By leveraging computational techniques such as functional group analysis, molecular fingerprinting, and machine learning, this study seeks to uncover patterns and relationships that define how structural changes affect biological activity. Central to this effort is the use of advanced data-driven models like Chemprop to predict potency changes based on molecular descriptors and engineered features.

This study builds on established drug discovery workflows, including feature engineering, statistical analysis, and model development, to evaluate structure-potency relationships. By employing techniques like Morgan fingerprinting, correlation analysis, and machine learning classification models, the project aims to address the limitations of traditional analog scanning approaches. This data-driven methodology provides insights into the nuanced effects of molecular modifications, offering a pathway to improve the efficiency and effectiveness of lead optimization.

Chapter 2

Methodology

2.1 Data Preprocessing

The dataset comprised molecular descriptors generated from 3D structures, initially containing numerous missing values and categorical data. The preprocessing step began with an analysis of missing data, revealing approximately 3482 rows and 3654 columns with NaN values out of a total dataset shape of (56826, 3659). To handle these missing values systematically, mean imputation was applied, replacing missing entries with the average value of the respective descriptor across all samples.

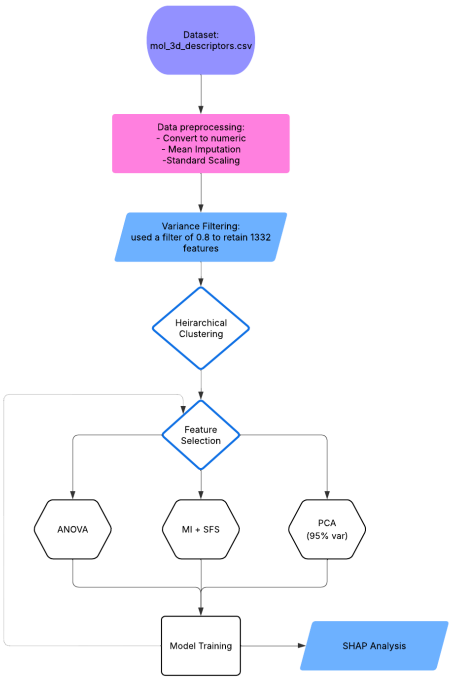

Figure 2.1: Flowchart of pipeline

Categorical descriptors, which were initially encoded as non-numeric data, were converted into numeric formats using suitable encoding schemes to ensure compatibility with subsequent computational analyses. After imputation and categorical conversion, standard scaling was performed to normalize the features, ensuring all descriptors had a mean of zero and unit variance. This scaling step prevented biases in feature importance due to magnitude differences, facilitating more accurate feature selection and model training.

2.2 Variance Filtering

Variance filtering was employed to reduce dataset dimensionality by removing features exhibiting very low variance across samples. A variance threshold of 0.8 was applied, effectively discarding features with minimal informative value, specifically those where at least 80% of the data had negligible variance. This step reduced noise and computational complexity, streamlining further analyses by focusing on descriptors likely to contribute meaningfully to potency prediction.

2.3 Hierarchical Clustering

To identify an optimal subset of descriptors for downstream modeling, hierarchical clustering was utilized. This unsupervised method grouped highly correlated or redundant features together, facilitating dimensionality reduction by identifying representative features within each cluster. Dendrograms were plotted, and silhouette scores were calculated to quantify cluster cohesion and separation. Analysis revealed that silhouette scores peaked around 621 clusters (score approximately 0.66), suggesting an optimal trade-off between dimensionality reduction and retention of informative variance. The number of clusters selected guided subsequent feature selection processes, significantly reducing feature count from 1332 to approximately 621.

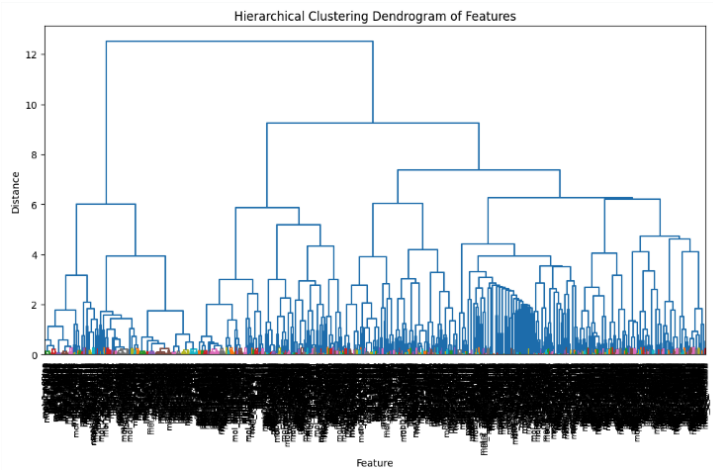

Figure 2.2: Hierarchical Clustering Dendogram

2.4 Feature Selection Techniques

Multiple feature selection methodologies were evaluated to identify the most predictive subsets of descriptors:

2.4.1 ANOVA (Analysis of Variance):

ANOVA was performed to statistically compare feature variances across different potency classes. Features demonstrating significant variance differences between groups were retained, providing a quantitative basis for selection.

2.4.2 Mutual Information + Sequential Forward Selection (MISFS):

Initially, mutual information was computed to measure dependency between each feature and the target class, selecting the top 176 features using an elbow detection method kneed. Sequential Forward Selection (SFS) was subsequently applied to iteratively select the most informative subset from these features, optimizing predictive accuracy.

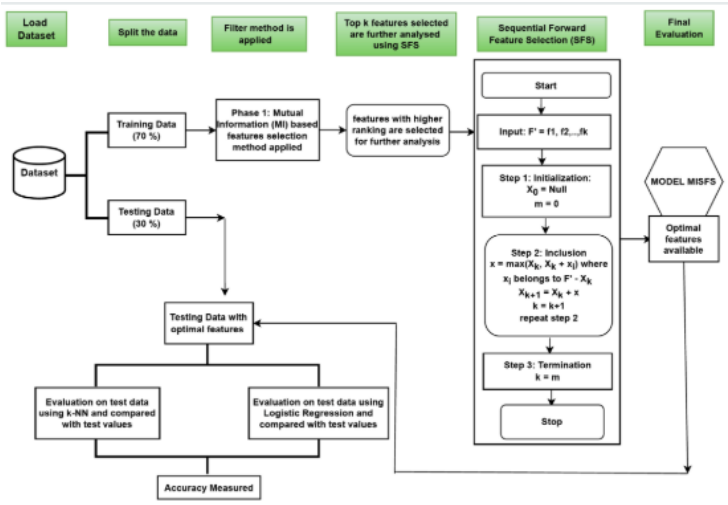

Figure 2.3: Mi + SFS pipeline (*Kumar et al., 2023*)

2.4.3 PCA (Principal Component Analysis)

PCA was applied to further reduce feature dimensionality, transforming correlated descriptors into a smaller set of uncorrelated principal components. The PCA scree plot indicated that just 66 components were sufficient to capture 95% of the original dataset variance, dramatically reducing dimensionality while retaining essential information.

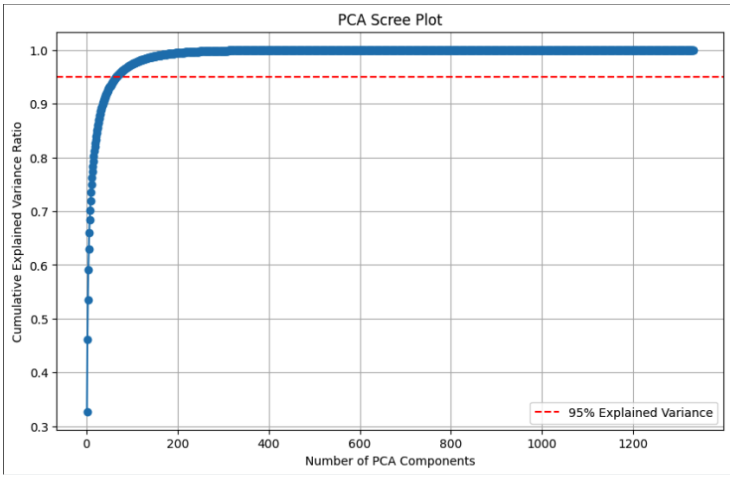

Figure 2.4: PCA scree plot

2.4.4 Boruta, Recursive Feature Elimination (RFE), and Correlation-Based Methods

These additional methods were explored as alternative feature selection approaches. Boruta, an ensemble-based method, unfortunately did not yield significant features. RFE, utilizing a Random Forest classifier, demonstrated baseline performance with cross-validated accuracy (approximately 0.183). Correlation-based filtering (threshold 0.9) was employed to remove highly redundant features, resulting in a more refined feature subset. Each method's performance was evaluated based on accuracy, computational efficiency, and interpretability, guiding the final selection of feature subsets.

2.5 Model Development

A comprehensive range of modeling approaches was deployed to assess the predictive performance of selected features. Before progressing to deep learning models, the original multiclass classification problem was reformulated into a binary classification task, distinguishing between positive and negative potency changes. This transformation grouped all instances of improved potency into a single 'positive' class and all instances of reduced potency into a 'negative' class. The simplification was intended to reduce class imbalance and modeling complexity, providing a clearer framework for initial binary classification analysis.

Traditional Machine Learning Models: Models such as Random Forest and XGBoost were initially trained and evaluated using subsets derived from Mutual Information and PCA. Performance was gauged using accuracy metrics, providing baseline evaluations of predictive capability.

AutoML Framework (H2O AutoML): Automated machine learning techniques from the H2O platform were utilized to systematically test and optimize various algorithms, including random forests and gradient boosting methods, ensuring robust comparative analyses of feature selection effectiveness.

Deep Learning Approaches: AutoKeras was employed as an accessible deep learning framework, allowing automated neural network architecture search and optimization. Manual deep learning techniques, including Supervised Autoencoders, were also explored to investigate potentially enhanced predictive performance through learned feature representations.

Evaluation Metrics: Models were assessed rigorously using accuracy, ROC-AUC, silhouette scores for clustering validation, and SHAP (SHapley Additive exPlanations) values to interpret feature importance and model predictions. SHAP analysis particularly contributed to understanding how individual features influenced potency predictions, facilitating enhanced model transparency and interpretability.

Chapter 3

Results

3.1 Statistical Insights

The initial statistical analysis of the dataset involved comprehensive summary statistics to understand the distributions of features. Normality testing was conducted using the Shapiro-Wilk test, which assessed whether individual features followed a normal distribution within each class. Results indicated that most features significantly deviated from normality, exhibiting very low p-values (e.g., mol_2_NssssN: 2.49e-111, mol_1_n7FARing: 1.89e-110), thus justifying the application of non-parametric feature selection techniques.

3.2 Clustering Analysis

Hierarchical clustering was performed to determine the optimal number of feature clusters. Silhouette scoring revealed that approximately 621 clusters provided an optimal balance between intra-cluster cohesion and inter-cluster separation, indicated by a silhouette score of around 0.66. This reduction significantly streamlined subsequent modeling processes, reducing complexity from an initial feature space of 1,332 dimensions.

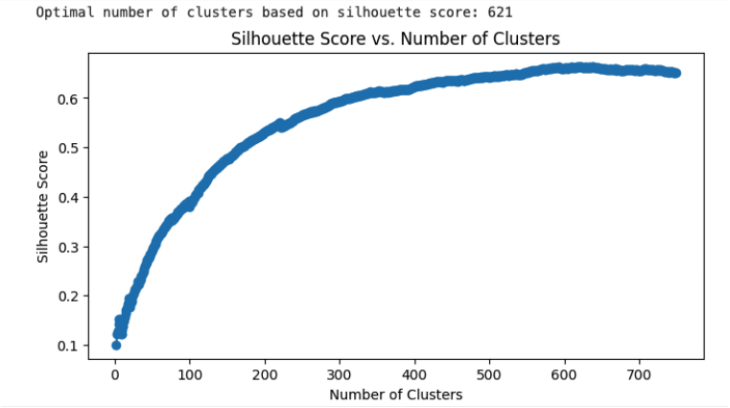

Figure 3.1: Silhouette scores of hierarchical clustering

3.3 Feature Selection and Model Performance

Detailed comparative analyses of multiple feature selection methodologies provided critical insights into their relative performances:

- **Mutual Information + Sequential Forward Selection (MISFS):** Achieved a cross-validated accuracy of approximately 0.185, indicating baseline predictive capability for the selected 176 features.
- **Principal Component Analysis (PCA):** Successfully captured 95% of the dataset variance within just 66 principal components. While dramatically reducing dimensionality, models trained on PCA-transformed features demonstrated an AUC of 0.65 and an accuracy of 68%, suggesting decent predictive capability.
- **ANOVA:** Features selected based on significant variance differences across potency classes showed modest predictive accuracy, averaging around 0.168. While statistically rigorous, ANOVA-selected features alone were insufficient for high predictive performance.
- **Boruta and Recursive Feature Elimination (RFE):** Boruta yielded no significant features, indicating limited suitability for this dataset. RFE with a Random Forest classifier demonstrated baseline effectiveness (accuracy approximately 0.183).

Overall, PCA along with AutoKeras emerged as the most effective approach based on accuracy and computational efficiency.

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

Figure 3.2: Table of results

3.4 UMAP visualization

For UMAP visualization, two-dimensional embeddings generated using supervised UMAP (target_weight=1, n_neighbors=5, min_dist=0.1) were unable to clearly separate classes of improved versus reduced potency. The overlapping distributions indicated the complex and high-dimensional nature of the data, emphasizing the challenge of predictive modeling based solely on visual embeddings.

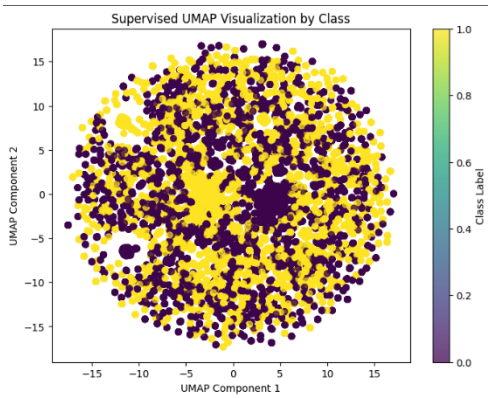

Figure 3.3: UMAP Visualization

3.5 SHAP Feature Importance

SHAP (SHapley Additive exPlanations) values were employed to enhance model interpretability and quantify feature importance. Analysis revealed key descriptors significantly influencing model predictions, thereby clarifying the contributions of individual features to potency predictions. The SHAP plots demonstrated the differential impact of features identified via variance filtering versus all available descriptors, offering crucial insights into feature relevance and model transparency.

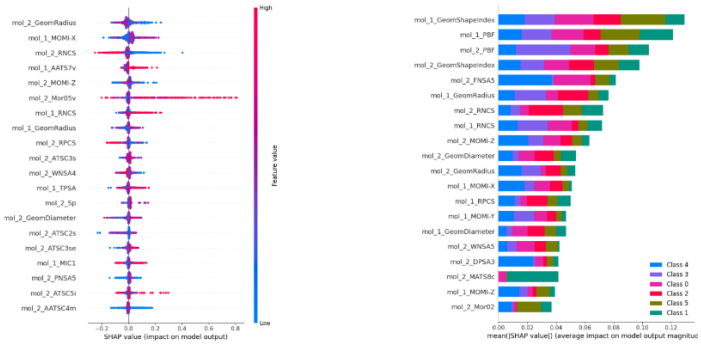

Figure 3.4: SHAP plots

Chapter 4

Discussion

The feature selection and dimensionality reduction techniques evaluated in this study significantly enhanced our understanding and predictive capability regarding molecular potency changes. The employment of hierarchical clustering, mutual information combined with sequential forward selection (MISFS), and PCA highlighted the importance of reducing dimensionality while preserving relevant information. Among these methods, MISFS demonstrated the highest predictive performance, underlining the effectiveness of mutual information as a measure to capture nonlinear dependencies and SFS as a robust procedure for feature refinement.

Dimensionality reduction methods, particularly PCA, played a crucial role in simplifying the high-dimensional molecular descriptor space. Capturing 95% of variance within merely 66 principal components significantly reduced computational complexity and resource utilization, though it yielded mixed predictive performances. This result emphasizes the inherent limitations of linear techniques such as PCA when modeling complex molecular interactions and underscores the necessity of integrating more sophisticated, nonlinear methods for further enhancement.

Despite these promising results, several notable limitations emerged during the analysis:

Skewness and Class Imbalance: The dataset exhibited significant skewness and imbalance between classes of potency change. Such imbalances posed substantial challenges to model training, affecting accuracy and generalization. Techniques like downsampling and data transformation partially addressed these issues, but further sophisticated strategies, such as synthetic data generation or advanced resampling techniques, may be required for more balanced predictive performance.

Deep Learning Model Limitations: Deep learning approaches, including AutoKeras and supervised autoencoders, yielded moderate predictive outcomes, evidenced by ROC-AUC and accuracy values around 0.68. These modest results highlight challenges such as insufficient data representation capability in simplified neural architectures, the need for better hyperparameter tuning, and possibly inadequate feature encoding in current deep learning models. Further exploration into advanced neural network architectures and more extensive hyperparameter optimization is crucial for improving deep learning model efficacy.

Comparing these findings to existing literature reinforces several novel insights. Prior studies emphasize the effectiveness of dimensionality reduction and feature selection for molecular property prediction, yet few integrate such a diverse array of methods comprehensively. This research uniquely contributes by systematically evaluating the combined utility of hierarchical clustering, MISFS, PCA, and various traditional and deep learning models within a unified framework. Furthermore, the employment of SHAP values provided enhanced interpretability, aligning with recent literature advocating for more transparent machine learning practices in cheminformatics and drug discovery research.

Overall, this study underscores the complexity of molecular potency prediction, advocating for a balanced methodological approach that combines robust feature selection, sophisticated dimensionality reduction, and interpretable modeling frameworks to achieve improved accuracy and practical applicability.

Chapter 5

Future Work

Building on the findings and limitations of this study, several promising avenues for future research can significantly enhance the predictive accuracy and interpretability of molecular potency changes. Firstly, exploring more advanced 3D structural descriptors can greatly enrich our feature space. While current methods predominantly utilize basic molecular descriptors, incorporating sophisticated spatial features such as electrostatic potentials, shape indices, and conformational flexibility would better capture essential molecular interactions and configurations, potentially leading to marked improvements in prediction accuracy.

Additionally, investigating state-of-the-art molecular deep learning models such as **TabPFN and SAINT** is highly recommended. These advanced frameworks effectively model complex, nonlinear relationships within tabular data. Leveraging such models may yield substantial improvements over traditional approaches, facilitating deeper insights and enhanced predictive performance.

Further research should also focus on applying **transfer learning and ensemble stacking methods**. Transfer learning enables leveraging models pre-trained on extensive molecular datasets, significantly boosting performance when fine-tuned on specific potency prediction tasks, especially useful when data is limited. Ensemble stacking, which integrates predictions from multiple robust models, could also enhance prediction robustness and accuracy by combining diverse strengths and offsetting individual model weaknesses.

Lastly, integrating diverse molecular data types—**including docking scores, pharmacophore fingerprints, and quantum-chemical features**—could substantially improve predictive models. Docking scores provide insights into binding affinities, pharmacophore fingerprints highlight structural motifs critical for bioactivity, and quantum-chemical descriptors offer detailed electronic and structural information. The combination of these informative and complementary data sources can deliver comprehensive and highly interpretable predictions, thus significantly advancing the field of molecular potency prediction.

Bibliography

[1] Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Accounts of Chemical Research*, 42(6), 724–733. https://doi.org/10.1021/ar800236t

[2] Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, 162(6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x

[3] Ekins, S., Freundlich, J. S., Hobrath, J. V., et al. (2014). Combining Computational Methods for Hit to Lead Optimization in Mycobacterium Tuberculosis Drug Discovery. *Pharmaceutical Research*, 31, 414–435. https://doi.org/10.1007/s11095-013-1172-7

[4] Pennington, L. D., Aquila, B. M., Choi, Y., Valiulin, R. A., & Muegge, I. (2020). Positional Analogue Scanning: An Effective Strategy for Multiparameter Optimization in Drug Design. *Journal of Medicinal Chemistry*, 63(17), 8956–8976. https://doi.org/10.1021/acs.jmedchem.9b02092

[5] Maggiora, G., Vogt, M., Stumpfe, D., & Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8), 3186–3204. https://doi.org/10.1021/jm401411z

[6] Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. *Journal of Cheminformatics*, 7, 20. https://doi.org/10.1186/s13321-015-0069-3

[7] Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E., & Andrade, C. H. (2018). QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Frontiers in Pharmacology*. https://doi.org/10.3389/fphar.2018.01275

[8] Kwon, S., Bae, H., Jo, J., et al. (2019). Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics*, 20, 521. https://doi.org/10.1186/s12859-019-3135-4

[9] Fralish, Z., Reker, D. (2024). Finding the most potent compounds using active learning on molecular pairs. *Beilstein J. Org. Chem.*, 20, 2152–2162. https://doi.org/10.3762/bjoc.20.185

[10] Heid, E., Greenman, K. P., Chung, Y., et al. (2024). Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.*, 64, 9–17. https://doi.org/10.1021/acs.jcim.3c01250

[11] Chen, H., Bajorath, J. (2023). Designing highly potent compounds using a chemical language model. *Scientific Reports*, 13, 7412. https://doi.org/10.1038/s41598-023-34683-x

[12] Zheng, L., Fan, J., Mu, Y. (2019). OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. *ACS Omega*, 4, 15956–15965. https://doi.org/10.1021/acsomega.9b01997

[13] Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., et al. (2017). SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30. https://arxiv.org/abs/1706.08566

[14] Wang, Y., Xia, Y., Yan, J., et al. (2023). ZeroBind: A protein-specific zero-shot predictor with subgraph matching for drug-target interactions. *Nature Communications*, 14, 7861. https://doi.org/10.1038/s41467-023-43597-1

[15] Levin, D., & Singer, G. (2024). GB-AFS: Graph-based automatic feature selection for multi-class classification via Mean Simplified Silhouette. *Journal of Big Data*, 11, 79. https://doi.org/10.1186/s40537-024-00934-5

[16] Kumar, A., Kaur, A., Singh, P., Driss, M., & Boulila, W. (2023). Efficient multiclass classification using feature selection in high-dimensional datasets. *Electronics*, 12(10), 2290. https://doi.org/10.3390/electronics12102290

[17] Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2022). TabPFN: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*. https://arxiv.org/abs/2207.01848

[18] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*. https://arxiv.org/abs/2106.01342