Unsupervised Machine Learning Approach for Identifying Biomechanical Influences on Protein-Ligand Binding Affinity

Arjun Singh
Student, Watchung Hills Regional High School
Warren, New Jersey, United States

***Abstract*—Drug discovery is incredibly time-consuming and expensive, averaging over 10 years and $985 million per drug. Calculating the binding affinity between a target protein and a ligand through Virtual Screening is critical for discovering viable drugs. Although supervised machine learning (ML) can predict binding affinity accurately, models experience severe overfitting due to an inability to identify informative properties of protein-ligand complexes. This study used unsupervised ML to reveal underlying protein-ligand characteristics that strongly influence binding affinity. Protein-ligand 3D models were collected from the PDBBind database and vectorized into 2422 features per complex. Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), K-Means Clustering, and heatmaps were used to identify groups of complexes and the features responsible for the separation. ML benchmarking was used to determine the features' effect on ML performance. The PCA heatmap revealed groups of complexes with binding affinity of pKd<6 and pKd>8 and identified the number of CCCH and CCCCCH fragments in the ligand as the most responsible features. A high correlation of 0.8337, their ability to explain 18% of the binding affinity's variance, and an error increase of 0.09 in Decision Trees when trained without the two features suggests that the fragments exist within a larger ligand substructure that significantly influences binding affinity. This discovery is a baseline for informative ligand representations to be generated so that ML models overfit less and can more reliably identify novel drug candidates. Future work will focus on validating the ligand substructure's presence and discovering more informative intra-ligand relationships.**

Keywords*—*Drug discovery; unsupervised machine learning; feature engineering; protein-ligand binding affinity; virtual screening

I. Introduction

Drug discovery is the basis of the modern pharmaceutical market and encompasses most of the industry's research and development funding [1]. On average, it takes 12-15 years and $985 million to deliver a drug to market, demonstrating the exhaustive time and effort required to complete the drug discovery process [2, 3]. Drug-Target Interaction (DTI) analysis is one of the most critical parts of drug discovery, and it involves calculating the binding affinity between a target protein and a ligand molecule so that appropriate ligand candidates for drugs can be chosen. These ligand candidates go on to be included in in vitro experimentation in order to identify lead compounds for the final drug. The affinity of a ligand to bind with a protein depends on the atomic interactions between the ligand and the binding region (referred to as the "binding pocket") on the protein, as shown in Fig. 1 [4]. Calculating the binding affinity between a protein and ligand can be completed through Virtual Screening (VS), shown in Fig. 2, where compounds are screened and binding affinity calculated using molecular simulation software [5].

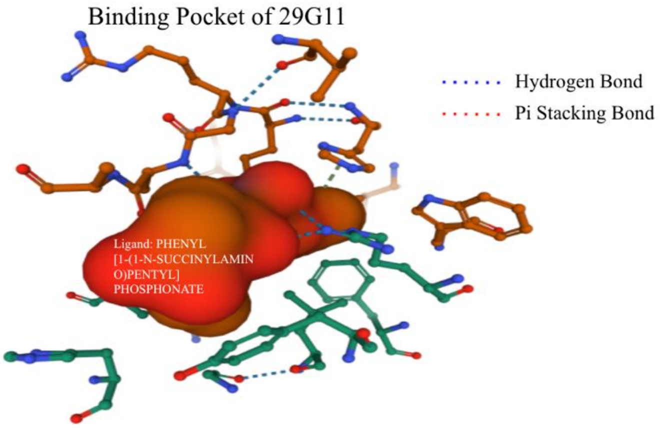

Fig. 1. Molecular view of Complex between 29G11 Protein and PHENYL [1-(1-N-SUCCINYLAMINO) PENTYL] PHOSPHONATE, Generated using Mol*.

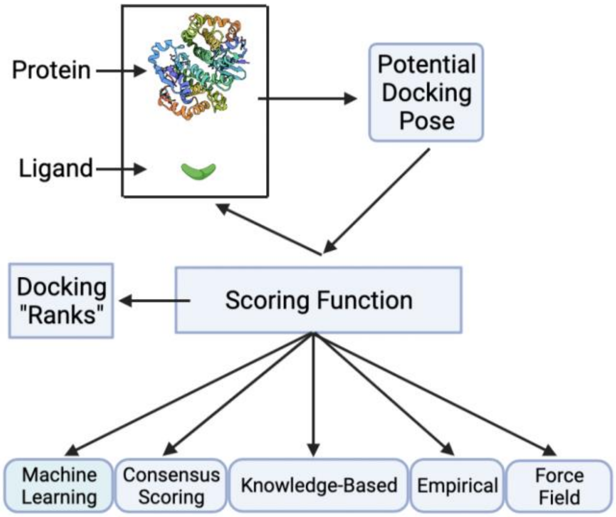

Fig. 2. Virtual Screening Workflow.

The "Scoring Function", which is the function used to calculate binding affinity, is critical for VS. Machine Learning (ML) algorithms have demonstrated considerable promise as a scoring function compared to other standard function types [6]. Given a set of training data, ML algorithms are able to learn pharmaco-like features from protein-ligand models through supervised learning functions. This allows them to accurately predict the binding affinity based on learned features that have statistically high influence [7-9, 11]. However, ML algorithms "overfit", or learn patterns that do not correlate to a physical phenomenon but still decrease error by chance [7-9, 11, 12]. This reduces their ability to generalize to out-of-distribution (OOD) data, making them unreliable for analyzing novel ligand candidates [7]. It is necessary to uncover underlying relationships between the features of protein-ligand data in order to inform the development of ML models that experience less overfitting [8].

Supervised learning techniques used to predict binding affinity can also analyze features, yet the results suffer from inconsistency and unreliability due to the overfitting of their parent algorithms [10, 13, 14, 18]. In comparison, unsupervised learning techniques such as Principal Component Analysis (PCA) are effective at identifying important features from protein-ligand models without overfitting because they are not designed to only minimize prediction error [15, 17]. t-Distributed Stochastic Neighbor Embedding (t-SNE) is also useful at visualizing the features of proteins due to its ability to retain high-dimensional information [16]. However, unsupervised learning has not been applied to analyze the differences between protein-ligand complexes in regard to binding affinity. This research can be filled help develop ML models that overfit considerably less.

The paper is structured as follows: Section II discusses the methodology, Section III presents and discusses the results, and Section IV concludes the study and proposes future work.

A. *Objectives*

There is a pressing need to reliably identify specific biomechanical features that influence binding affinity and quantify their effect on ML performance. Current literature either suffer from drawbacks in reliability and consistency caused by supervised learning or do not specifically analyze the variance in binding affinity caused by protein/ligand features. The objectives of this study are three-fold: 1) Discover the presence of underlying biomechanical interactions that influence binding affinity, 2) Identify specific pharmaco-like features responsible for high variance in binding affinities, and 3) Determine the effect of these features on the performance of ML models in predicting binding affinity.

Gathering a greater understanding of which features influence binding affinity is necessary for designing ML models that do not overfit to training data and interpret noisy features as important patterns. Models will thereby be more generalizable to OOD data, and more successful at identifying lead compounds for inclusion in innovative drugs.

II. Methodology

A. *Dataset Preprocessing*

In this study, protein-ligand models were collected from the PDBBind database [19, 41]. The 2015 "Refined" set and the 2015 "Core" set were downloaded. In order to extract relevant quantitative features of each model, a workflow described in [40] was utilized, as shown in Fig. 3.

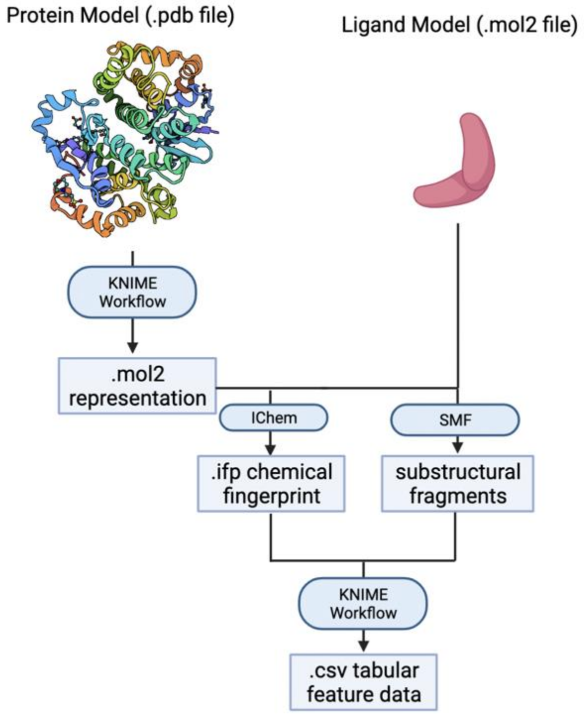
Protein Model (.pdb file)
Ligand Model (.mol2 file)
KNIME Workflow
.mol2 representation
IChem
SMF
.ifp chemical fingerprint
substructural fragments
KNIME Workflow
.csv tabular feature data

Fig. 3. Computational Workflow used to Translate 3D Molecular Models into 1D Tabular Data.

For each complex, 2422 quantitative features were collected. The frequency of 2282 unique substructural molecular fragments was collected. The remaining 140 features were frequencies of amino-acid interactions, with seven types of interactions per amino acid: 1) Hydrophobic, 2) Face-to-face aromatic, 3) Edge-to-edge aromatic, 4) H-bond accepted by ligand, 5) H-bond donated by ligand, 6) Ionic bond (ligand partially negative), and 7) Ionic bond (ligand partially positive). Files with a resolution of <2.5 Å were retained to ensure the accuracy of all feature counts, resulting in 3481 complexes from the "Refined" set and 180 from the "Core" set.

B. *Feature Analysis*

To reveal underlying feature correlations in the dataset, a combination of PCA, t-SNE, K-Means Clustering, and heatmap projections shown in Fig. 4 were performed using Python and the Scikit-Learn, Pandas, and NumPy packages.

C. *PCA/K-Means*

PCA (n=2) was performed to transform the 2422-feature data into two dimensions for visualization and to capture the features with the highest variance. K-Means Clustering (k=10) was performed on this transformation to determine if there were categories of complexes. The similarity of the clusters was calculated using the Davies-Bouldin Score (DBS). The presence of sparse categories and a low DBS would indicate an underlying biomechanical phenomenon between features. Another PCA (n=3) with K-Means Clustering (k=10) was performed to verify the outcome of the 2D PCA.

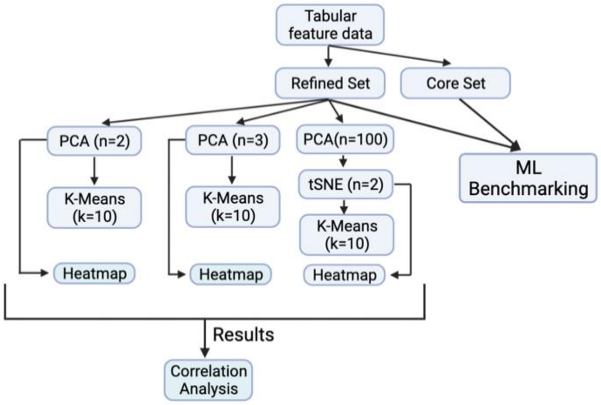

Fig. 4. Feature Analysis Workflow.

D. PCA/t-SNE/K-Means

Due to the ability of t-SNE to interpret non-linearity, a PCA (n=100) and then t-SNE (n=2) was performed to retain high-dimensional characteristics of the data. K-Means Clustering (k=10) was then performed to determine if the high-dimensional characteristics could describe separable categories of complexes. DBS was again used to score the similarity of the clusters.

E. t-SNE Heatmap

In order to determine if a biomechanical relationship could be demonstrated without clustering, a heatmap was generated of the t-SNE results where the "heat" was determined by the binding affinity. The quality of grouping was calculated using an adjusted R^2 correlation value. It is significant to note that there are 2422 features per complex; therefore what may seem to be low R^2 correlation values may actually be statistically significant due to the large number of features.

F. PCA Heatmap

In order to verify or refute the results of the t-SNE heatmap, a heatmap was generated with the PCA components in the same manner as the t-SNE heatmap. Similarly, the quality of grouping was evaluated using an adjusted R^2 correlation value.

G. Correlation Analysis

Although each clustering plot and heatmap could determine the presence of a biomechanical relationship, only the PCA plots could indicate which specific features are statistically responsible for it because each Principal Component is organized along the variance of each feature. Whichever 2D PCA plot (clustered plot or heatmap) indicated separable groups had the variance of each feature in its Principal Components returned to find the two most informative features. A covariance matrix was generated to identify the direction of the relationship between the features. The Spearman Correlation Coefficient was calculated to determine the strength of the covariance between the two features and the strength of each feature's covariance to the binding affinity. A heatmap of the features' correlation to binding affinity was generated to confirm the Spearman Correlation calculations. The results of this analysis suggested what specific biomechanical relationship may exist between the features.

H. Machine Learning Benchmarking

To determine the effect of the features on ML performance, five state-of-the-art ML models were trained/tested on two datasets: one with and one without the features. The five models were as follows: 1) Random Forests, 2) Support Vector Machine, 3) K-Nearest Neighbors, 4) Decision Tree, and 5) LightGBM Regressor. The "Refined" set was used for training and validation, and the "Core" for testing. The "Refined" set was split such that a random 80% of complexes went into the training subset and the other 20% into the validation subset. The Root Mean Squared Error (RMSE) and Pearson Correlation Coefficient (PCC) of each model's testing predictions were calculated to evaluate the model.

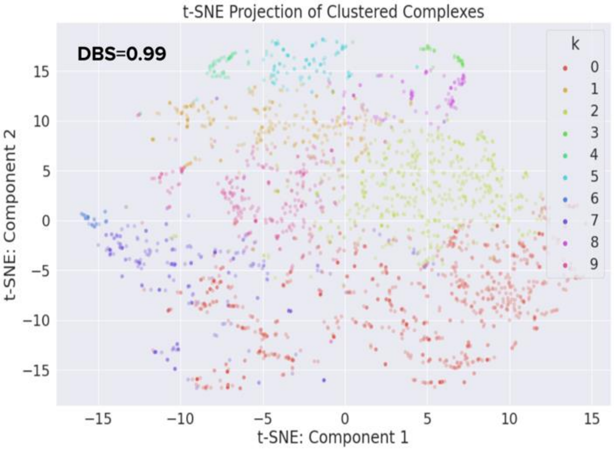

Fig. 5. Projection of t-SNE (n=2) Transformed Data after being Reduced using PCA (n=100) and Clustered using K-Means (k=10).

III. Result and Discussion

A. PCA/K-Means

A PCA (n=2) was performed and the transformed data was clustered using K-Means (k=10). Another PCA (n=3) was used to verify the 2D PCA. The 2D PCA exhibited a high DBS (>0.5) of 0.83 and dense clusters shown in Fig. 6A. The 3D PCA exhibited a similar outcome as the 2D PCA, with a higher DBS of 0.93, as shown in Fig. 6B. The clusters indicate that separable categories of complexes do not exist, suggesting that the PCA and clustering was unable to capture a biomechanical relationship between features.

B. PCA/t-SNE/K-Means

A PCA (n=100) followed a t-SNE (n=2) transformation was performed. The transformed data was clustered using K-Means (k=10). The t-SNE plot in Fig. 5 shows dense clusters and a high DBS of 0.99, suggesting that the t-SNE/clustering was also unable to identify a biomechanical relationship.

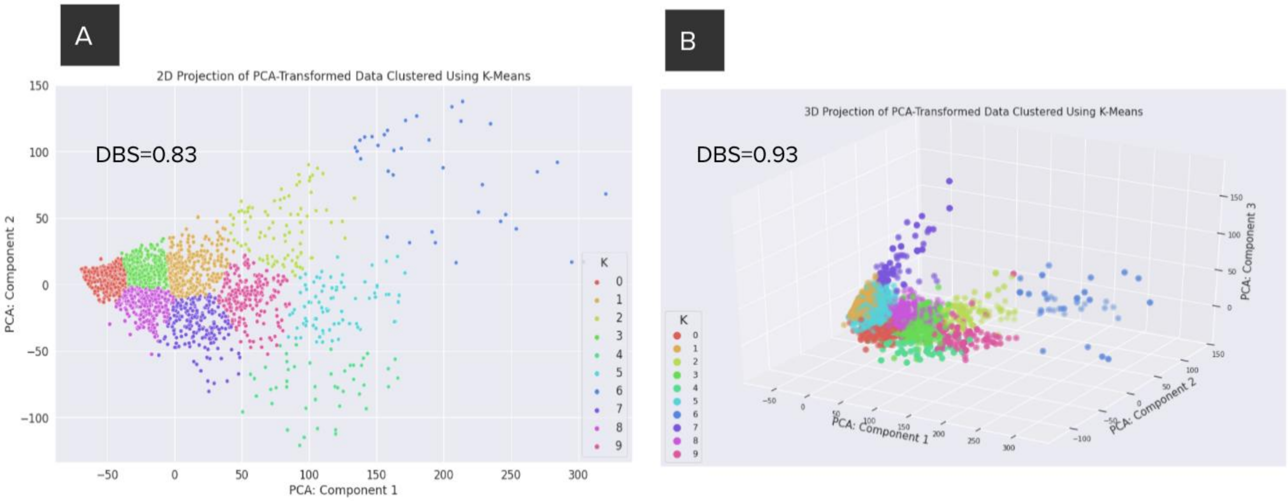

Fig. 6. Projection of PCA (n=2, A) and PCA (n=3, B) Transformed Data after being Clustered using K-Means Clustering (k=10).

C. t-SNE Heatmap

The t-SNE (n=2) transformed data was projected to a heatmap, where the "heat" was determined by the binding affinity. The plot exhibited no significant groups and an R^2 value of 0.0007, as shown in Fig. 7. The low R^2 and lack of groups reinforce the indication that the t-SNE components were unable to identify distinguished groups of complexes and therefore unable to identify a significant relationship between features.

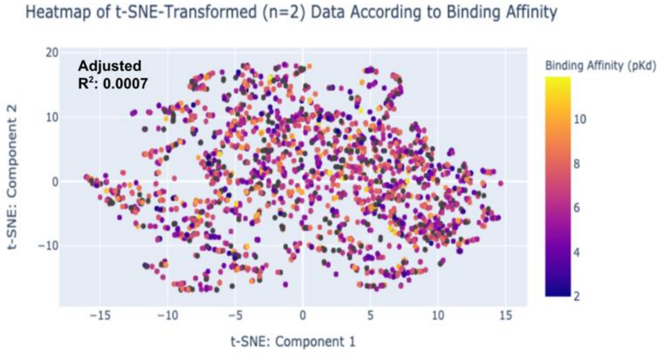

Fig. 7. Heatmap of t-SNE (n=2) Transformed Data with "Heat" Determined by binding Affinity.

D. PCA Heatmap: 2D

The PCA (n=2) results were projected to a heatmap in the same manner as the t-SNE heatmap. The PCA heatmap showed a notable difference between complexes with binding affinity of pKd<6 (blue-purple group) and those with pKd>8 (orange-yellow group) at a higher adjusted R^2 value of 0.17, as shown in Fig. 8. The R^2 supports that there does exist a biomechanical relationship between features which is significantly responsible for binding affinity. A select number of features from the Principal Components are likely to have significant chemical importance in determining binding affinity [20-25].

E. PCA Heatmap: 3D

Another PCA (n=3) was performed and projected to a 3D heatmap to verify the results of the 2D PCA. If a similar grouping was evident in the 3D PCA as the 2D, the grouping would be more statistically likely to be significant rather than by chance. The 3D heatmap did show a similar phenomenon as the 2D heatmap, with a noticeable grouping of complexes with pKd<6 (blue-purple group) and pKd>8 (orange-yellow group) at a similar R^2 correlation value of 0.18, as shown in Fig. 9.

The grouping supports the indication that the Principal Components were able to identify a biomechanical relationship that significantly affects binding affinity. High-variance features from the Principal Components are likely to be responsible for this relationship [20-25].

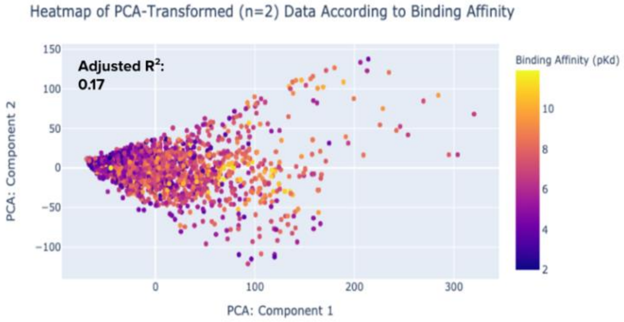

Fig. 8. Heatmap of PCA (n=2) Transformed Data with "Heat" Determined by binding Affinity.

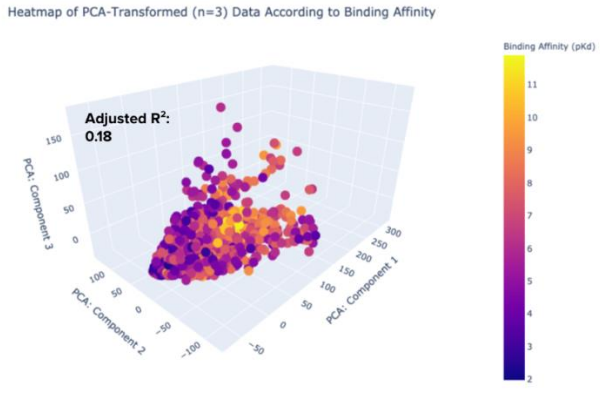

Fig. 9. Heatmap of PCA (n=2) Transformed Data with "Heat" Determined by binding Affinity.

F. Correlation Analysis

In order to determine which specific features were most likely involved in the biomechanical relationship, the feature with the highest variance in each Principal Component was returned. It was found that the CCCH and CCCCCH substructural ligand fragments features had the highest variance in the first Principal Component and the second Principal Component, respectively. In order to verify the presence of a relationship between CCCH and CCCCCH fragments, a covariance matrix was calculated between the two fragment counts. A direct (positive) relationship is evident with a covariance value of 358.34, as shown in Fig. 10. The covariance suggests that the specific relationship between the fragments is that they are both part of a larger molecular substructure within the ligand that is critical in determining binding affinity [26-28].

In order to verify the implication of the covariance matrix, the Spearman Correlation Coefficient was calculated between each combination of fragments and the binding affinity. The CCCH and CCCCCH fragments showed a high correlation of 0.8337. Each fragment and the binding affinity had a moderate correlation of 0.4286 and 0.3457, respectively, as shown in Table I. The high correlation between the fragments supports that they have a biomechanical relationship and that both fragments are part of a larger molecular substructure [26-28]. The moderate correlation between each fragment and binding affinity suggests that both fragments are involved in chemically determining binding affinity [29, 30].

The correlation calculations did not measure correlation between both fragments together and the binding affinity. Therefore, a heatmap of the fragment counts with the binding affinity was generated to verify that the fragment relationship influences binding affinity.

The same grouping that was evident in the PCA heatmaps occurred, with one group of complexes with pKd<6 and another with pKd>8 at a significant R^2 correlation of 0.18 as shown in Fig. 11. The grouping suggests that the CCCH-CCCCCH relationship is significantly responsible for determining the binding affinity with a protein. The CCCH-CCCCCH relationship is likely a critical influence on the optimal docking pose between the ligand and protein [31].

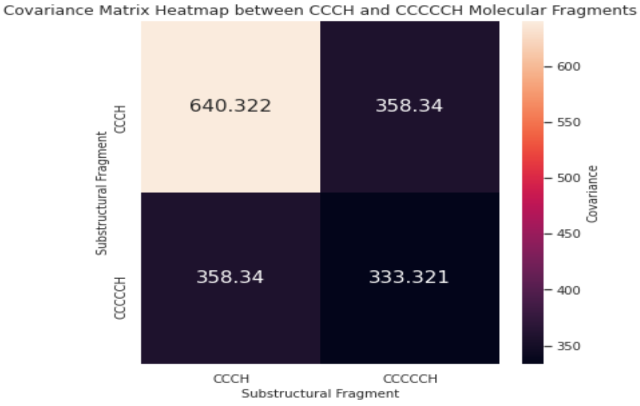

Fig. 10. Heatmap of Covariance Matrix between CCCH and CCCCCH Substructural Molecular Fragments.

TABLE I. SPEARMAN CORRELATION COEFFICIENTS BETWEEN HIGH-VARIANCE FEATURES AND BINDING AFFINITY

Rank Variable #1	**Rank Variable #2**	**Spearman Correlation Coefficient**	**P-Value**
CCCH Count	CCCCCH Count	**0.8337**	**0.0**
CCCH Count	Binding Affinity	**0.4286**	8.25e-125
CCCCCH Count	Binding Affinity	**0.3457**	5.82e-79

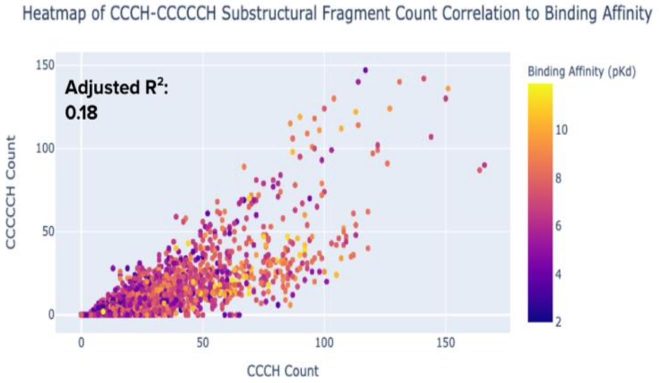

Fig. 11. Heatmap of Correlation between CCCH-CCCCCH Fragment Count and binding Affinity.

G. Machine Learning Benchmarking

In order to determine the effect of the CCCH-CCCCCH relationship on the performance of ML models in predicting binding affinity, five models were trained/tested on datasets with and without the fragment counts. The absence of the counts had an insignificant effect on most models except for the Decision Tree, which experienced an increase in RMSE of 0.09 and a decrease in PCC of 0.05, as shown in Table II. The insignificant effect on most models suggests that there are other factors with notable influence on binding affinity. The decreased performance of the Decision Tree suggests that the CCCH/CCCCCH count is an important decision rule for tree-based learning algorithms [32].

TABLE II. EFFECT OF CCCH AND CCCCCH ON MACHINE LEARNING PERFORMANCE

Model	**With CCCH and CCCCH fragment counts**		**Without CCCH and CCCCCH fragment counts**	
	RMSE	**PCC**	**RMSE**	**PCC**
Random Forests	1.49	0.77	1.50	0.77
Support Vector Machine	1.70	0.68	1.69	0.69
K-Nearest Neighbors	1.71	0.64	1.69	0.66
Decision Tree	**1.95**	**0.57**	**2.04**	**0.52**
LightGBM Regression	1.46	0.77	1.44	0.77

IV. Conclusion and Future Work

The biomechanical relationship discovered in this study serves as a baseline for further ligand interactions to be found. Including the relationship elucidated through this work, more interactions can be gathered to develop a corpus of ligand fragment relationships that influence binding affinity. This will produce a more accurate representation of ligand chemistry in regard to protein binding, improving the performance of predictive ML models [33, 34, 36]. Understanding the effect of ligand relationships on ML, as was done in this study, will also help researchers improve model performance [35].

Most importantly, uncovering specific ligand relationships will result in ML models that overfit less, making them more generalizable to new datasets and thus reliable for analyzing novel drug candidates [37-39].

The effect of generalizable ML models on effective VS is profound. It has already been demonstrated that for certain proteins such as Interleukin-1 receptor associated kinase-1 (IRAK1), ML models can increase novel ligand hit rates by over 1000% compared to standard scoring functions [40]. Developing ML models that are more generalizable can result in similar increases across wide ranges of proteins because models will be able to screen novel ligands without significant decreases in reliability. Using the relationship uncovered in this study as well as others to develop generalizable ML models is therefore critical for identifying promising drug candidates for innovative medicines.

It is significant to note that the relationship discovered in this study is useful in other scientific contexts, such as synthetic drug design. Using known information on fragments such as the two discussed in this study (CCCH and CCCCCH), synthetic ligands can be chemically designed to bind optimally to a target protein [42, 43]. Computational tools (including, but not limited to, ML models) can also be developed to design novel synthetic drugs using known relationships between ligand fragments [44-46]. Gathering a clear, data-driven understanding of ligand fragment activity is a significant method by which synthetic drug design for new medications can be improved [48].

There are several limitations in this work that present promising directions for future research. Only several unsupervised learning techniques were used in this study, yet multiple other unsupervised/self-supervised techniques such as Uniform Manifold Approximation and Projection (UMAP) and Autoencoder Networks can be used to verify the results of this study [50]. Further, multicollinearity between features was not analyzed in this study, but can significantly affect feature selection methods. Therefore, multicollinearity analysis will validate the presence of the larger substructure (containing CCCH and CCCCCH fragments) suggested in this study's results [47]. Should it exist, in-vitro experimentation can be performed to determine how the substructure affects ML performance in predicting binding affinity, revealing important information on the usefulness of such substructures in VS [49]. In addition, the protein-ligand models used in this study came from a single dataset, which introduces dataset bias and may affect the results of feature analysis. Therefore, incorporating data from other reliable datasets will verify/refute the results of this study and decrease potential bias. Future work based on this study will aid in significantly progressing protein-ligand binding affinity research.

References

[1] D. Taylor, “The Pharmaceutical Industry and the Future of Drug Development,” PiE, pp. 1-33, 2015.

[2] A. Pandey, “Drug Discovery and Development Process,” Learning Center, 2020. Available at: https://www.nebiolab.com.

[3] M. Terry, “The Median Cost of Bringing a Drug to Market is $985 Million, According to New Study,” Biospace, 2020. Available at: https://www.biospace.com/.

[4] S. Anusya, M. Kesherwani, K. Priya, A. Vimala, G. Shanmugam, D. Velmurugan, and M. Gromiha, “Drug-Target Interactions: Prediction Methods and Applications,” Current Protein and Peptide Science, vol. 19, no. 6, pp. 537-561, 2018.

[5] E. Lionta, G. Spyrou, D. Vassilatis, and Z. Cournia, “Structure-Based Virtual Screening for Drug Discovery: Principles, Applications, and Recent Advances,” Current Topics in Medicinal Chemistry, vol. 14, no. 16, pp. 1923-1938, 2014.

[6] K.A. Carpenter and X. Huang, “Machine Learning-based Virtual Screening and Its Applications to Alzheimer’s Drug Discovery: A Review,” Current Pharmaceutical Design, vol. 24, no. 28, pp. 3347-3358, 2018.

[7] D. Jones, H. Kim, X. Zhang, A. Zemla, G. Stevenson, W. F. D. Bennett, D. Kirshner, S. E. Wong, F. C. Lightstone, and J. E. Allen, “Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference,” Journal of Chemical Information and Modeling, vol. 61, no. 4, pp. 1583-1592, 2021.

[8] H. Ozturk, A. Ozgur, and E. Ozkirimli, “DeepDTA: deep-target binding affinity prediction,” Bioinformatics, vol. 34, no. 17, pp. i821-i829, 2019.

[9] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlicki, “Development and evaluation of a deep learning model for protein-ligand binding affinity prediction,” Bioinformatics, vol. 34, no. 21, pp. 3666-3674, 2018.

[10] K. Wang, R. Zhou, Y. Li, M. Li, “DeepDTAF: a deep learning method to predict protein-ligand binding affinity,” Briefings in Bioinformatics, 2021.

[11] M. A. Rezaei, Y. Li, D. Wu, X. Li and C. Li, “Deep Learning in Drug Design: Protein-Ligand Binding Affinity Prediction,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, December 2020.

[12] L. Rice, E. Wong, and J. Z. Kolter, “Overfitting in adversarially robust deep learning,” in Proceedings of the 2020 International Conference on Machine Learning (ICLM), 2020.

[13] Y. Kwon, W. Shin, J. Ko, and J. Lee, “AK-Score; Accurate Protein-Ligand Binding Affinity Prediction Using an Ensemble of 3D-Convolutional Neural Networks,” International Journal of Molecular Sciences, vol. 21, no. 22, 2020.

[14] J. Hochuli, A. Helbling, T. Skaist, M. Ragoza, and D. R. Koes, “Visualizing Convolutional Neural Network Protein-Ligand Scoring,” Journal of Molecular Graphics and Modeling, vol. 84, pp. 96-108, 2018.

[15] V. Subramanian, H. Xhaard, P. Prusis, and G. Wolfhart, “Predictive protochemometric models for kinases derived from 3D protein field-based descriptors,” Medicinal Chemistry Communications, vol. 7, no. 5, 2016.

[16] D. S. Karlov, S. Sosnin, M. V. Fedorov, and P. Popov, “graphDelta: MPPN Scoring Function for the Affinity Prediction of Protein-Ligand Complexes,” American Chemical Society Omega, vol. 5, no. 10, pp. 5150-5159, 2020.

[17] S. Khan, U. Farooq, and M. Kurnikova, “Protein stability and dynamics influenced by ligands in extremophilic complexes – a molecular dynamics investigation,” Molecular BioSystems, vol. 13, pp. 1874-1887, 2017.

[18] W. Torng and R. B. Altman, “Graph Convolutional Neural Networks for Predicting Drug-Target Interactions,” Journal of Chemical Inforamtion and Modeling, vol. 59, no. 10, pp. 4131-4149, 2019.

[19] R. Wang, X. Fang, Y. Lu, C. Yang, and S. Wang, “The PDBbind database: methodologies and updates,” Journal of Medicinal Chemistry, vol. 48, no. 12, pp. 4111-4119, 2005.

[20] G. Tang and R. Altman, “Knowledge-based Fragment Binding Prediction,” PLOS Computational Biology, 2014.

[21] E. Grant, D. Fallon, M. Hann, K. Fantom, C. Quinn, F. Zappacosta, R. Annan, C. Chung, P. Bamborough, D. Dixon, P. Stacey, D. House, V. K. Patel, N. C. O. Tomkinson, and J. T. Bush, “A Photoaffinity-Based Fragment-Screening Platform for Efficient Identification of Protein Ligands,” Angewandte Chemie International Edition, vol. 59, 2020.

[22] D. A. Erlanson, B. J. Davis, and W. Jahnke, “Fragment-Based Drug Discovery: Advancing Fragments in the Absence of Crystal Structures,” Cell Chemical Biology, vol. 26, no. 1, pp. 9-15, 2018.

[23] M. Peters, “THE APPLICATION OF SEMIEMPIRICAL METHODS IN DRUG DESIGN,” Ph.D. dissertation, DC, UF, Florida, 2007, Available at: http://etd.fcla.edu/UF/UFE0021354/peters_m.pdf.

[24] P. Kenny, “The nature of ligand efficiency,” Journal of Cheminformatics, vol. 11, no. 8, 2019.

[25] T. Pantsar and A. Poso, “Binding Affinity via Docking: Fact and Fiction,” Molecules, vol. 23, no. 8, pp. 1899, 2018.

[26] F. Chevillard, H. Rimmer, C. Betti, E. Pardon, S. Ballet, N. Hilten, J. Steyaert, W. E. Diederic, and P. Kolb, “Binding-Site Compatible Fragment Growing Applied to the Design of □2-Andrenergic Receptor Ligands,” Journal of Medicinal Chemistry, vol. 61, no. 3, pp. 1118-1129, 2018.

[27] P. Matricon, A. Ranganathan, E. Warnick, Z. Gao, A. Rudling, C. Lambertucci, G. Marucci, A. Ezzati, M. Jaiteh, D. D. Ben, K. A. Jacobson, and J. Carlsson, “Fragment optimization for GPCRs by molecular dynamics free energy calculations: Probing druggable subpockets of the A2A adenoside receptor binding site,” Scientific Reports, vol. 7, no. 6398, July 2017.

[28] J. Robston-Tull, “Biophysical screening in fragment-based drug design: a brief overview,” Bioscience Horizons, vol. 11, 2019.

[29] P. Kirsch, A. M. Hartman, A. K. H. Hirsch, and M. Empting, “Concepts and Core Principles of Fragment-Based Drug Design,” Molecules, vol. 24, no. 23, pp. 4309, 2019.

[30] F. Giordanetto, C. Jin, L. Willmore, M. Feher, and D. E. Shaw, “Fragment Hits: What do They Look Like and How do They Bind?” Journal of Medicinal Chemistry, vol. 62, no. 7, pp. 3381-3394, 2019.

[31] C. Jacquemard, M. N. Drwal, J. Desaphy, and E. Kellenberger, “Binding mode information improves fragment docking,” Journal of Cheminformatics, vol. 11, no. 24, 2019.

[32] H. Deng and G. Runger, “Feature selection via regularized trees,” in Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2012.

[33] H. Deng and G. Runger, “Feature selection via regularized trees,” in Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2012.

[34] X. Xu, C. Yan, and X. Zou, “Improving Binding mode and Binding Affinity Predictions of Docking by Ligand-based Search of Protein Confirmation: Evaluation in D3R Grand Challenge 2015,” Journal of Computer-Aided Molecular Design, vol. 31, no. 8, pp. 689-699, 2017.

[35] S. Holderbach, L. Adam, B. Jayaram, R. C. Wade, and G. Mukherjee, “RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physiochemical Features,” Frontiers in Molecular Biosciences, vol. 7, pp. 393, 2020.

[36] D. D. Wang, H. Xie, and H. Yan, “Proteo-chemometrics interaction fingerprints of protein-ligand complexes predict binding affinity,” Bioinformatics, 2021.

[37] G. G. Ferenczy and G. M. Keseru, “Thermodynamic profiling for fragment-based lead discovery and optimization,” Expert Opinion on Drug Discovery, vol. 15, no. 1, pp. 117-129, 2019.

[38] Z. Meng and K. Xia, “Persistent spectral-based machine learning (PerSpect ML) for protein-ligand binding affinity prediction,” Science Advances, vol. 7, no. 19, 2021.

[39] H. Goel, A. Hazel, V. D. Ustach, S. Jo, W. Yu, and A. D. MacKerell, “Rapid and accurate estimation of protein-ligand relative binding affinities using site-identification by ligand competitive saturation,” Chemical Science, vol. 12, pp. 8844-8858, 2021.

[40] S. Wan, A. P. Bhati, S. J. Zasada, and P. V. Coveney, “Rapid, accurate, precise and reproducible ligand-protein binding free energy prediction,” Interface Focus, vol. 10, no. 6, 2020.

[41] S. Kumar and M. Kim, “SMPLIP-Score: predicting ligand binding affinity from simple and interpretable on-the-fly interaction fingerprint pattern descriptors,” Journal of Cheminformatics, vol. 13, no. 28, 2021.

[42] Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu, and R. Wang, “PDB-wide collection of binding data: current status of the PDBbind database,” Bioinformatics, vol. 31, no. 3, pp. 405-412, 2015.

[43] A. Kashyap, P. K. Singh, O. Silakari, “Counting on Fragment Based Drug Design Approach for Drug Discovery,” Current Topics in Medicinal Chemistry, vol. 18, no. 27, pp. 2284-2293, 2018.

[44] M. Bissaro, M. Sturlese, and S. Moro, “The rise of molecular simulations in fragment-based drug design (FBDD): an overview,” Drug Discovery Today, vol. 25, no. 9, pp. 1693, 2020.

[45] Y. Bian and X. Xie, “Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications,” American Association of Pharmaceutical Scientists Journal, vol. 20, no. 59, 2018.

[46] V. D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco, and G. Melagraki, “Advances in de Novo Drug Design: From Conventional to Machine Learning Methods,” International Journal of Molecular Sciences, vol. 22, no. 4, pp. 1676, 2021.

[47] Q. Bai, S. Tan, T. Xu, H. Liu, J. Huang, and X. Yao, “MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classifical algorithm,” Briefings in Bioinformatics, vol. 22, no. 3, 2021.

[48] L. R. S. Neto, J. T. Moreira-Filho, B. J. Neves, R. L. B. R. Maidana, A. C. R. Guimaraes, N. Furnham, C. H. Andrade, and F. P. Silva, “In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery,” Frontiers in Chemistry, vol. 8, pp. 93, 2020.

[49] M. J. Caplin and D. J. Foley, “Emergeny synthetic methods for the modular advancement of sp3-rich fragments,” Chemical Sciences, vol. 12, pp. 4646-4660, 2021.

[50] M. Aldeghi, V. Gapsys, and B. L. de Groot, “Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation,” American Chemical Society Central Science, vol. 4, no. 12, pp. 1708-1718, 2018.

[51] J. O. Spiegel and J. D. Durrant, “AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization,” Journal of Cheminformatics, vol. 12, no. 25, 2020.