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

Arjun Singh (arjsingh2004@gmail.com)
Warren, NJ

Research article

Keywords: Drug discovery, binding affinity, protein-ligand complex, unsupervised machine learning, feature analysis

Posted Date: September 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-919250/v1

Version of Record: A version of this preprint was published at International Journal of Advanced Computer Science and Applications on January 1st, 2021. See the published version at https://doi.org/10.14569/IJACSA.2021.0121171.

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 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.

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 $985million 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 (Fig.1) [4].

Calculating the binding affinity between a protein and ligand can be completed through Virtual Screening (VS), where compounds are screened and binding affinity calculated using software [5] (Fig.2). 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 phenomena 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 [13, 14]. 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 in low-dimensional space [16]. However, unsupervised learning has not been applied to analyze the differences between protein-ligand complexes in regards to their binding affinity. This is a research gap that can be filled to help develop ML models which learn protein-ligand feature patterns without considerably overfitting.

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.

Methods

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 acquire relevant quantitative features of each model, a dataset proposed in [40] was downloaded and utilized for this study.

For each complex, 2422 quantitative features were collected by [40]. The frequency of 2282 unique substructural molecular fragments were 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. The dataset used in this study can be found in [40].

Feature Analysis:

To reveal underlying feature correlations in the dataset, a combination of PCA, t-SNE, K-Means Clustering, and heatmap projections were performed using Python and the scikit-learn, pandas, and numpy packages (Fig.3).

PCA/K-Means:

PCA (n = 2) was performed to transform the 2422-feature datas 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 phenomena between features. Another PCA (n = 3) with K-Means Clustering (k = 10) was performed to verify the outcome of the 2D PCA.

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.

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.

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.

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.

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.

Results And Discussion

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 (Fig. 4A). The 3D PCA (Fig. 4B) exhibited a similar outcome as the 2D PCA, with a higher DBS of 0.93. 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.

PCA/t-SNE/K-Means:

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

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 (Fig. 6). The extremely 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.

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 (Fig. 7). The R^2 supports that there is 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].

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 phenomena 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 (Fig. 8). 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].

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 (Fig. 9). 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 (Table 1). The high correlation between the fragments supports that they have a biomechanical relationship and the suggestion 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 the chemical interaction that ultimately determines 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 (Fig. 10). 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 that maximizes binding affinity [31].

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 (Table2). 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].

Conclusions And Future Work

The biomechanical relationship discovered in this study serves as a baseline for further interactions within ligands 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 regards to protein binding, improving the performance of ML models that predict binding affinity [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 overift 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 are 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.

There are several promising directions for future work. Utilizing database-wide computational techniques such as 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 then reveal whether or not the substructure significantly affects binding affinity [48]. Further 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]. Lastly, in order to build a large corpus of relevant ligand activity, more unsupervised learning techniques can be used to uncover unknown ligand relationships [50]. Future work based on this study will aid in significantly progressing protein-ligand binding affinity research.

Declarations

Availability of data and materials

Protein-ligand models available at: http://www.pdbbind.org.cn/

Featurized dataset, preprocessing softwares and instructions available at: https://github.com/college-of-pharmacy-gachon-university/SMPLIP-Score.

Competing Interests

Not applicable

Funding

Not applicable

Authors' Contributions

The corresponding author of this study completed all experimentation described and wrote/reviewed/edited the manuscript independently.

Acknowledgements

The corresponding author of this study would like to thank James Wang (jameswang1279@gmail.com) for his mentorship throughout the research and the manuscript drafting. The authors of this study would also like to thank the iResearch Institute for providing the internship that facilitated this study.

Authors' Information

The corresponding author of this study, Arjun Singh, is a high school student from Warren, New Jersey, USA. Arjun has significant research experience in machine learning and biomedicine. This study was completed by Arjun as part of the iResearch Institute Summer Research Program.

References

1. D. Taylor, “The Pharmaceutical Industry and the Future of Drug Development,” PiE, pp. 1-33, September 2015.
2. A. Pandey, “Drug Discovery and Development Process,” Learning Center, June, 2020. [Online]. Available: NorthEast BioLab, https://www.nebiolab.com. [Accessed July 23, 2021].
3. M. Terry, “The Median Cost of Bringing a Drug to Market is $985 Million, According to New Study,” BioSpace, March 04, 2020. [Online]. Available: BioSpace, https://www.biospace.com/. [Accessed July 23, 2021].
4. S. Anusuya, M. Kesherwani, K. Priya, A. Vimala, G. Shanmugam, D. Velmurugan, and M. Gromiha, “Drug-Target Interactions: Prediction Methods and Applications,” Curr. Protein. Pept. Sci., vol. 19, no. 6, pp. 537-561, April 2018.
5. E. Lionta, G. Spyrou, D. Vassilatis, and Z. Cournia, “Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances,” Curr. Top. Med. Chem., vol. 14, no. 16, pp. 1923–1938, August 2014.
6. K. A. Carpenter and X. Huang, “Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review,” Curr. Pharm. Des., vol. 24, no. 28, pp. 3347-3358, August 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,” J. Chem. Inf. Model., vol. 61, no. 4, pp. 1583-1592, March 2021.
8. H. Öztürk, A. Özgür, and E. Ozkirimli, “DeepDTA: deep drug–target binding affinity prediction,” Bioinformatics, vol. 34, no. 17., pp. I821-i829, September 2018.
9. M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction,” Bioinformatics, vol. 34, no. 21, pp. 3666-3674, November 2018.
10. K. Wang, R. Zhou, Y. Li, M. Li, “DeepDTAF: a deep learning method to predict protein–ligand binding affinity,” Brief Bioinform., April 2021.
11. M. A. Rezaei, Y. Li, D. Wu, X. Li and C. Li, "Deep Learning in Drug Design: Protein-Ligand Binding Affinity Prediction," TCBB., December 2020.
12. L. Rice, E. Wong, and J. Z. Kolter, “Overfitting in adversarially robust deep learning,” The 2020 International Conference on Machine Learning (ICLM), July 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,” Int. J. Mol. Sci., vol. 21, no. 22, November 2020.
14. J. Hochuli, A. Helbling, T. Skaist, M. Ragoza, and D. R. Koes, “Visualizing Convolutional Neural Network Protein-Ligand Scoring,” J. Mol. Graph Model, vol. 84, pp. 96-108, June 2018.

15. V. Subramanian, H. Xhaard, P. Prusis, and G. Wohlfahrt, “Predictive proteochemometric models for kinases derived from 3D protein field-based descriptors,” MedChemComm., vol. 7, no. 5, April 2016.
16. D. S. Karlov, S. Sosnin, M. V. Fedorov, and P. Popov, “graphDelta: MPNN Scoring Function for the Affinity Prediction of Protein–Ligand Complexes,” ACS Omega, vol. 5, no. 10, pp. 5150-5159, March 2020.
17. S. Khan, U. Farooq, and M. Kurnikova, “Protein stability and dynamics influenced by ligands in extremophilic complexes – a molecular dynamics investigation,” Mol. Biosyst., vol. 13, pp. 1874-1887, July 2017.
18. W. Torng and R. B. Altman, “Graph Convolutional Neural Networks for Predicting Drug-Target Interactions,” J. Chem. Inf. Model., vol. 59, no. 10, pp. 4131–4149, October 2019.
19. R. Wang, X. Fang, Y. Lu, C. Yang, and S. Wang, “The PDBbind database: methodologies and updates,” J. Med. Chem., vol. 48, no. 12, pp. 4111-4119, June 2005.
20. G. Tang and R. Altman, “Knowledge-based Fragment Binding Prediction,” PLoS Comput. Biol., April 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,” Angew. Chem. Int. Ed., vol. 59, August 2020.
22. D. A. Erlanson, B. J. Davis, and W. Jahnke, “Fragment-Based Drug Discovery: Advancing Fragments in the Absence of Crystal Structures,” Cell Chem. Biol., vol. 26, no. 1, pp. 9-15, October 2018.
23. M. Peters, “THE APPLICATION OF SEMIEMPIRICAL METHODS IN DRUG DESIGN,” Ph.D. dissertation, DC, UF, Florida, 2007, Accessed on: July 30, 2021. [Digital File]. Available: http://etd.fcla.edu/UF/UFE0021354/peters_m.pdf
24. P. Kenny, “The nature of ligand efficiency,” J. Cheminformatics, vol. 11, no. 8, January 2019.
25. T. Pantsar and A. Poso, “Binding Affinity via Docking: Fact and Fiction,” Molecules, vol. 23, no. 8, pp. 1899, August 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 β2-Adrenergic Receptor Ligands,” J. Med. Chem., vol. 61, no. 3, pp. 1118-1129, January 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 A 2A adenosine receptor binding site,” Sci. Rep., vol. 7, no. 6398, July 2017.
27. J. Robson-Tull, “Biophysical screening in fragment-based drug design: a brief overview,” Biosci. Horiz., vol. 11, February 2019.
28. P. Kirsch, A. M. Hartman, A. K. H. Hirsch, and M. Empting, “Concepts and Core
29. Principles of Fragment-Based Drug Design”, Molecules, vol. 24, no. 23, pp. 4309, November 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?,” J. Med. Chem., vol. 62, no. 7, pp. 3381–3394, March 2019.
31. C. Jacquemard, M. N. Drwal, J. Desaphy, and E. Kellenberger, “Binding mode information improves fragment docking,” J. Cheminformatics, vol. 11, no. 24, March 2019.
32. H. Deng and G. Runger, "Feature selection via regularized trees," The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, June 2012.
33. X. Xu, C. Yan, and X. Zou, “Improving Binding Mode and Binding Affinity Predictions of Docking by Ligand-based Search of Protein Conformations: Evaluation in D3R Grand Challenge 2015,” J. Comput. Aided. Mol. Des., vol. 31, no. 8, pp. 689-699, August 2017.
34. S. Holderbach, L. Adam, B. Jayaram, R. C. Wade, and G. Mukherjee, “RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physicochemical Features,” Front. Mol. Biosci., vol. 7, pp. 393, December 2020.
35. D. D. Wang, H. Xie, and H. Yan, “Proteo-chemometrics interaction fingerprints of protein–ligand complexes predict binding affinity,” Bioinformatics, February 2021.
36. G. G. Ferenczy and G. M. Keseru, “Thermodynamic profiling for fragment-based lead discovery and optimization,” Expert Opin. Drug Discov., vol. 15, no. 1, pp. 117-129, November 2019.
37. Z. Meng and K. Xia, “Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction,” Sci. Adv., vol. 7, no. 19, May 2021.
38. 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,” Chem. Sci., vol. 12, pp. 8844-8858, May 2021.
39. 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, December 2020.
40. S. Kumar and M. Kim, “SMPLIP-Score: predicting ligand binding affinity from simple and interpretable on-the-fly interaction fingerprint pattern descriptors,” J. Cheminformatics, vol. 13, no. 28, March 2021.
41. 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, February 2015.
42. A. Kashyap, P. K. Singh, O. Silakari, “Counting on Fragment Based Drug Design Approach for Drug Discovery,” Curr. Top. Med. Chem., vol. 18, no. 27, pp. 2284-2293, March 2018.
43. M. Bissaro, M. Sturlese, and S. Moro, “The rise of molecular simulations in fragment-based drug design (FBDD): an overview,” Drug Discov. Today, vol. 25, no. 9, pp. 1693–1701, September 2020.
44. Y. Bian and X. Xie, “Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications,” AAPS J., vol. 20, no. 59, April 2018.
45. 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," Int. J. Mol. Sci., vol. 22, no. 4, pp. 1676, February 2021.
46. 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 classical algorithm," Brief. Bioinform., vol. 22, no. 3, May 2021.
47. L. R. S. Neto, J. T. Moreira-Filho, B. J. Neves, R. L. B. R. Maidana, A. C. R. Guimarães, N. Furnham, C. H. Andrade, and F. P. Silva, "In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery," Front. Chem., vol. 8, pp. 93, February 2020.
48. M. J. Caplin and D. J Foley, "Emergent synthetic methods for the modular advancement of sp3-rich fragments," Chem. Sci., vol. 12, pp. 4646-4660, March 2021.
49. M. Aldeghi, V. Gapsys, and B. L. de Groot, "Accurate Estimation of Ligand Binding Affinity Changes upon Protein Mutation," ACS Cent. Sci., vol. 4, no. 12, pp. 1708-1718, December 2018.
50. J. O. Spiegel and J. D. Durrant, "AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization," vol. 12, no. 25, April 2020.

Tables

Due to technical limitations, tables are only available as a download in the Supplemental Files section.

Figures

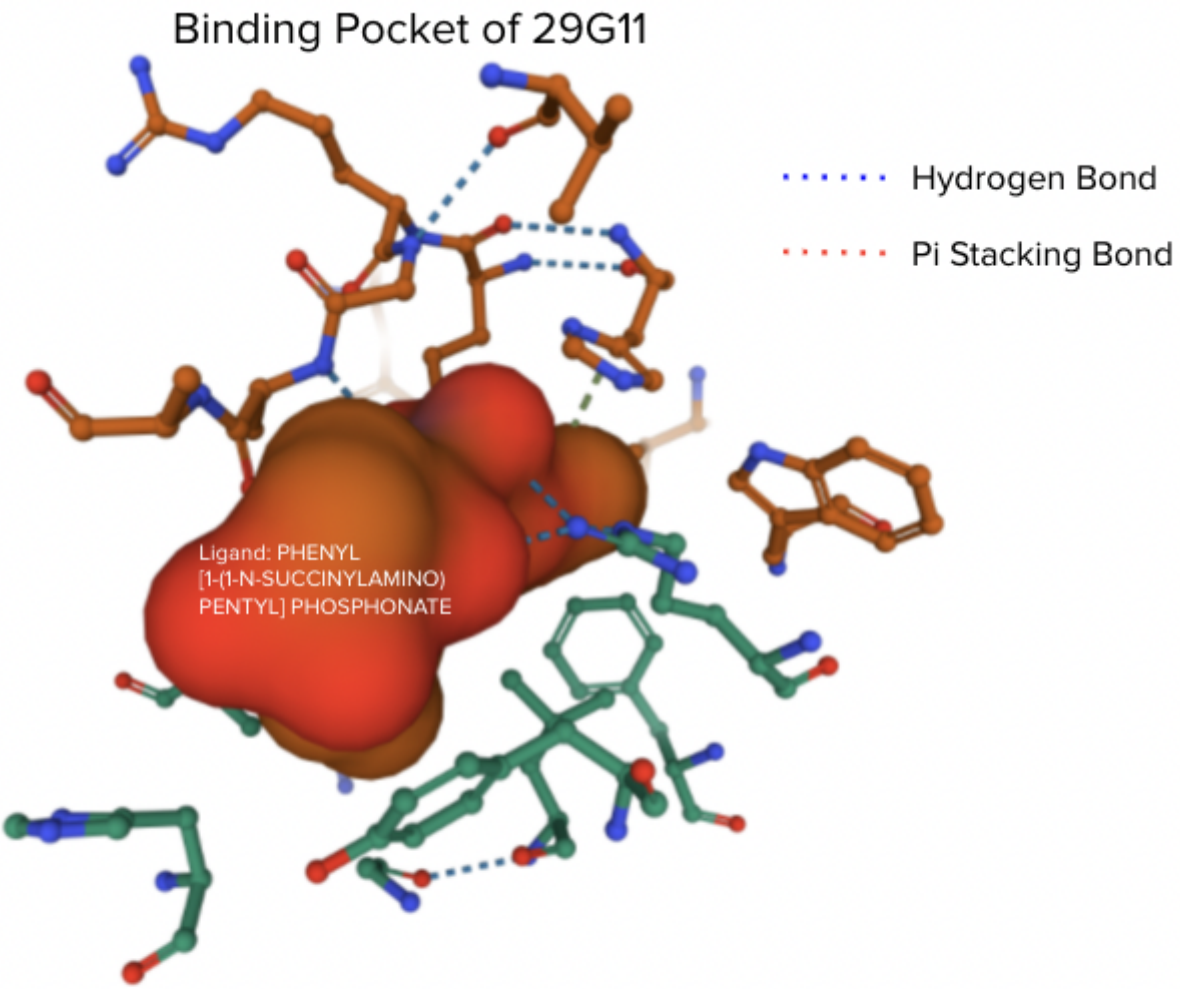

Fig 1. Molecular view of complex between 29G11 protein and PHENYL [1-(1-N-SUCCINYLAMINO)PENTYL] PHOSPHONATE, generated using Mol*. Ligand (bolded red) experiences specific interactions with the protein binding pocket (surrounding region) that are critical in determining binding affinity. Intra-ligand characteristics also determine docking pose and affinity.

Figure 1

See image above for figure legend

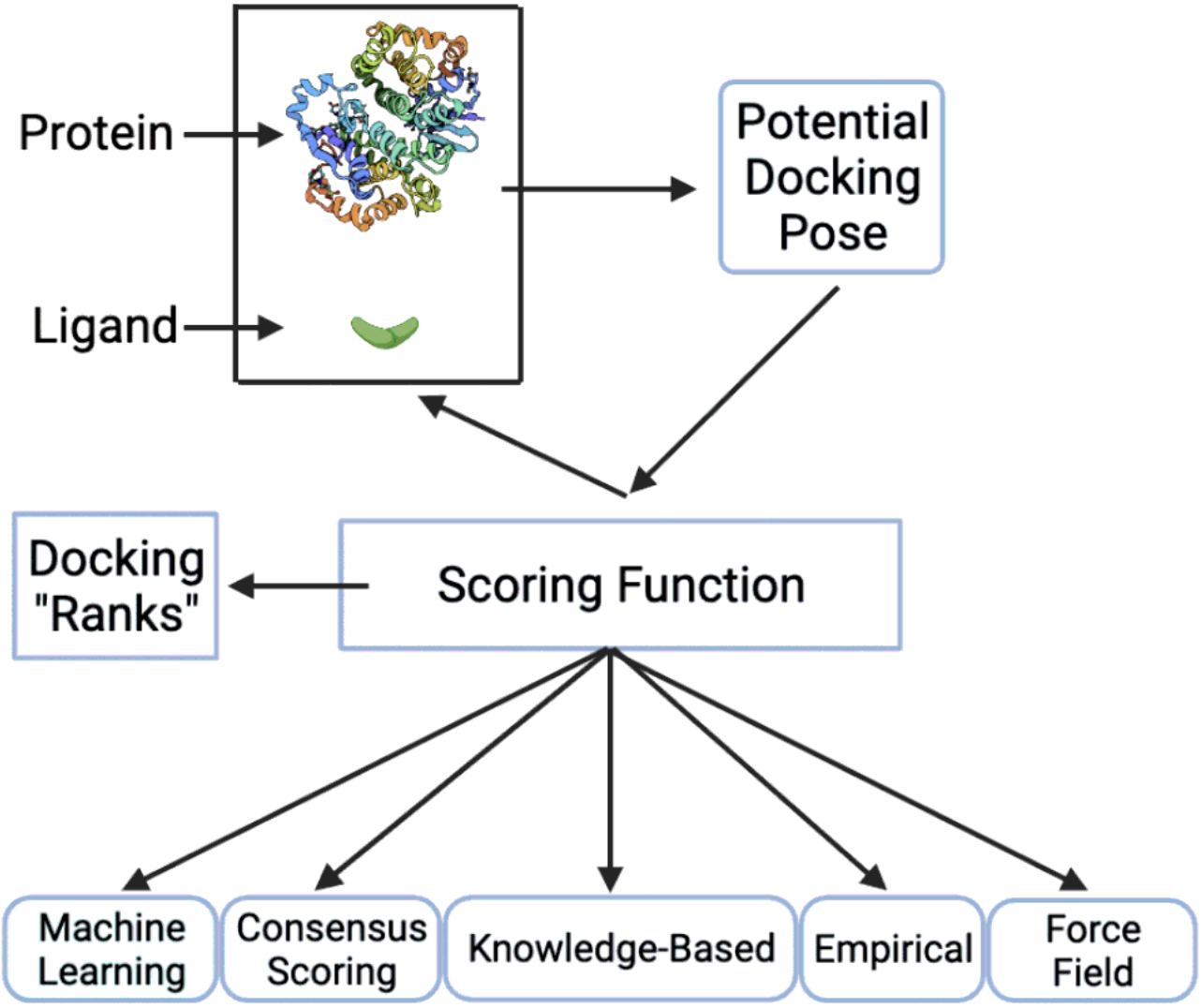

Fig 2. Virtual screening workflow. Docking poses are generated using a molecular simulation software and each pose is inputted into scoring function to calculate the binding affinity. The scoring function utilize either knowledge-based rules in physics and chemistry or learned rules to calculate the binding affinity. Machine Learning (ML), specifically, has demonstrated notable superiority to other functions in predicting binding affinity. After scoring, each pose is ranked against each other based on calculated affinity. The pose with the highest affinity is chosen as the "optimal pose" because it has the highest likelihood of acting as a stable compound in the biological system of interest.

Figure 2

See image above for figure legend

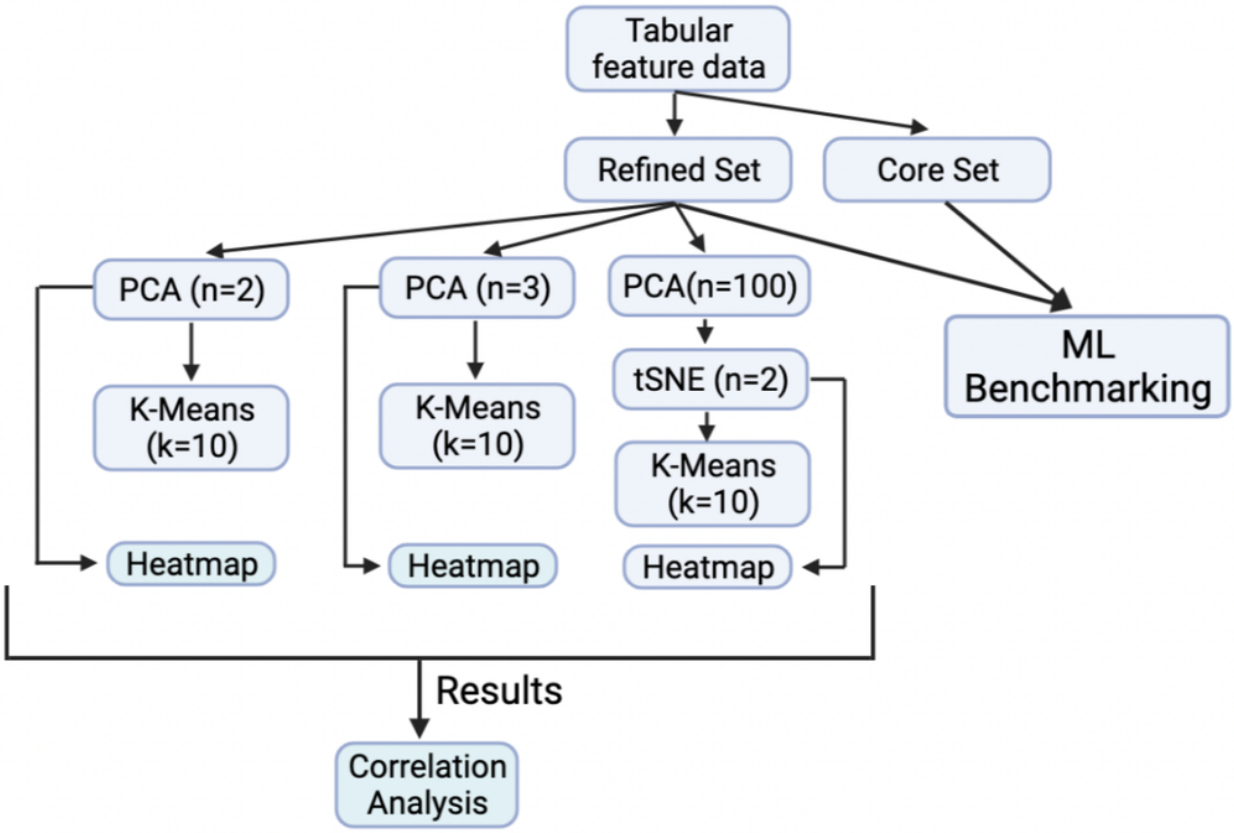

Fig 3. Sequential application of analytical functions on features of dataset. Determined 1) presence of an underlying categorical difference, 2) features responsible for the correlation, and 3) effect of the correlation on ML performance. PCA/K-Means and PCA/tSNE/K-Means supported or refuted presence of categorical difference. Heatmap projections revealed non-linear difference. Correlation analysis identified correlated features that are responsible for the difference, and ML benchmarking quantified the effect of these features on ML performance. The PCA/K-Means, PCA/tSNE/K-Means, and Heatmap plots were used to inform the correlation analysis. The "Core" set from PDBBind was held out as a testing set for ML benchmarking.

Figure 3

See image above for figure legend

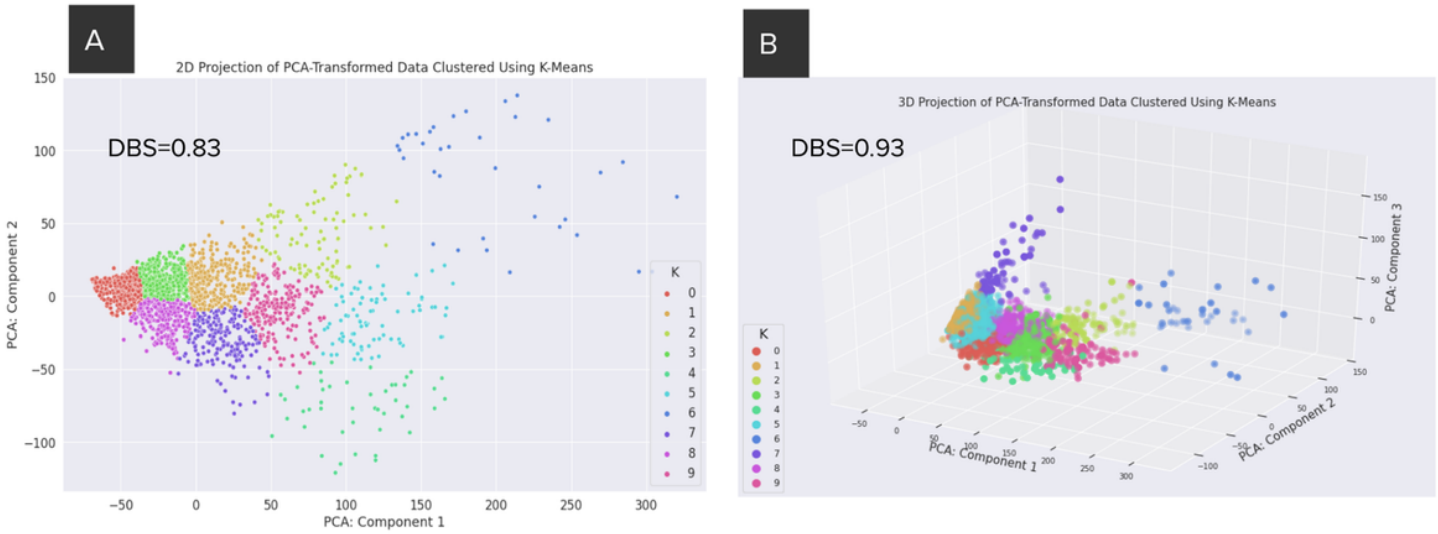

Fig 4. Projection of PCA (n=2, A) and PCA (n=3, B) transformed data after being clustered using K-Means Clustering (k=10). Plotted using Seaborn and Matplotlib Python packages. Cluster similarity calculated by Davies-Bouldin Score (DBS). Plot A shows mostly dense, close-together clusters and a high DBS (>0.5). Plot B confirms Plot A by also demonstrating mostly dense clusters and a higher DBS. Clusters indicate that there are not separable categories of complexes, and therefore no significant biomechanical phenomena that separates complexes.

Figure 4

See image above for figure legend

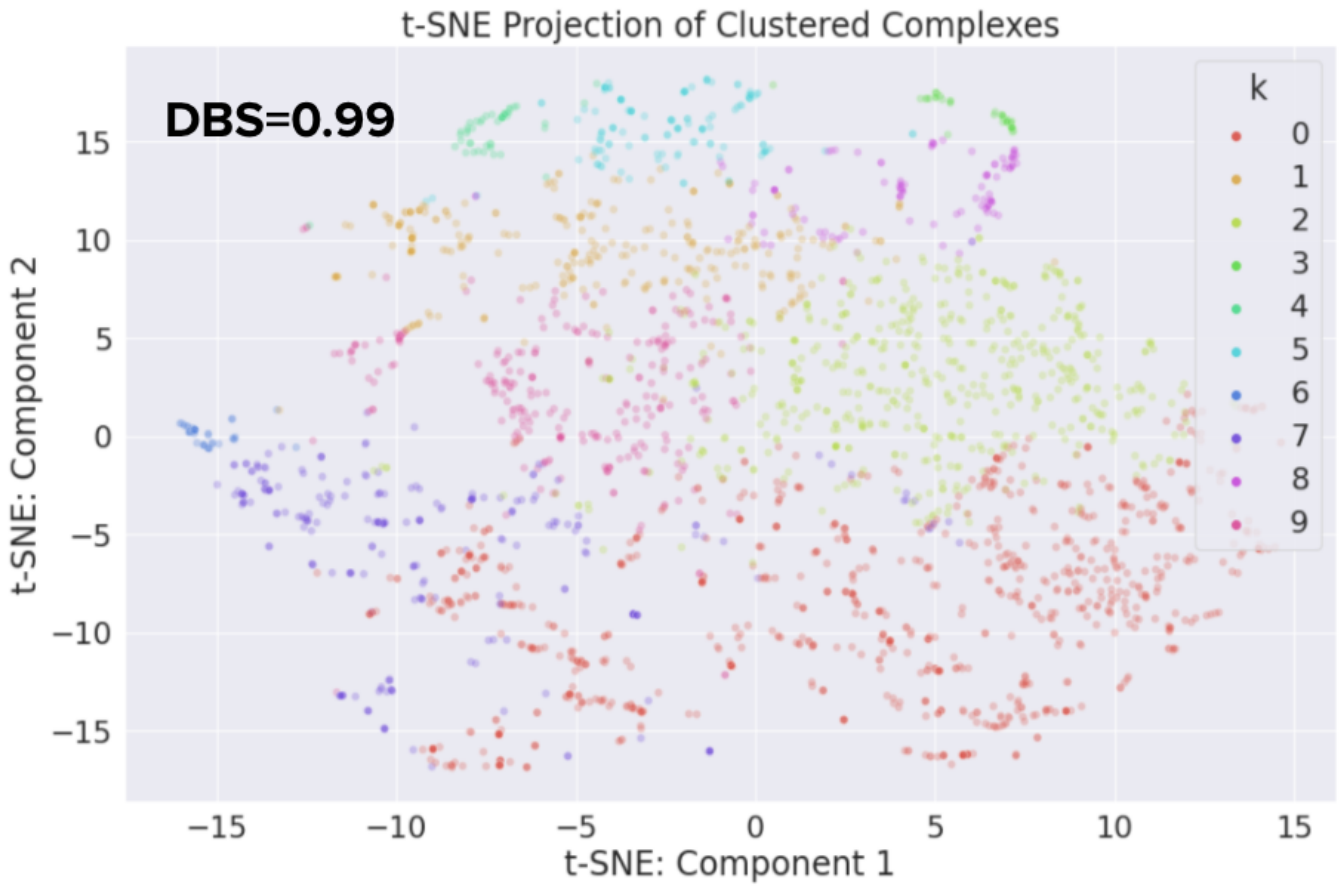

Fig 5. Projection of t-SNE (n=2) transformed data after being reduced using PCA (n=100) and clustered using K-Means (k=10). Plotted using Seaborn and Matplotlib Python packages. Cluster similarity calculated using Davies-Bouldin Score (DBS). Plot shows dense clusters, and DBS is high (>0.5). Clusters suggest that there are no separable categories of clusters, and thus a significant biomechanical interaction between features does not exist.

Figure 5

See image above for figure legend

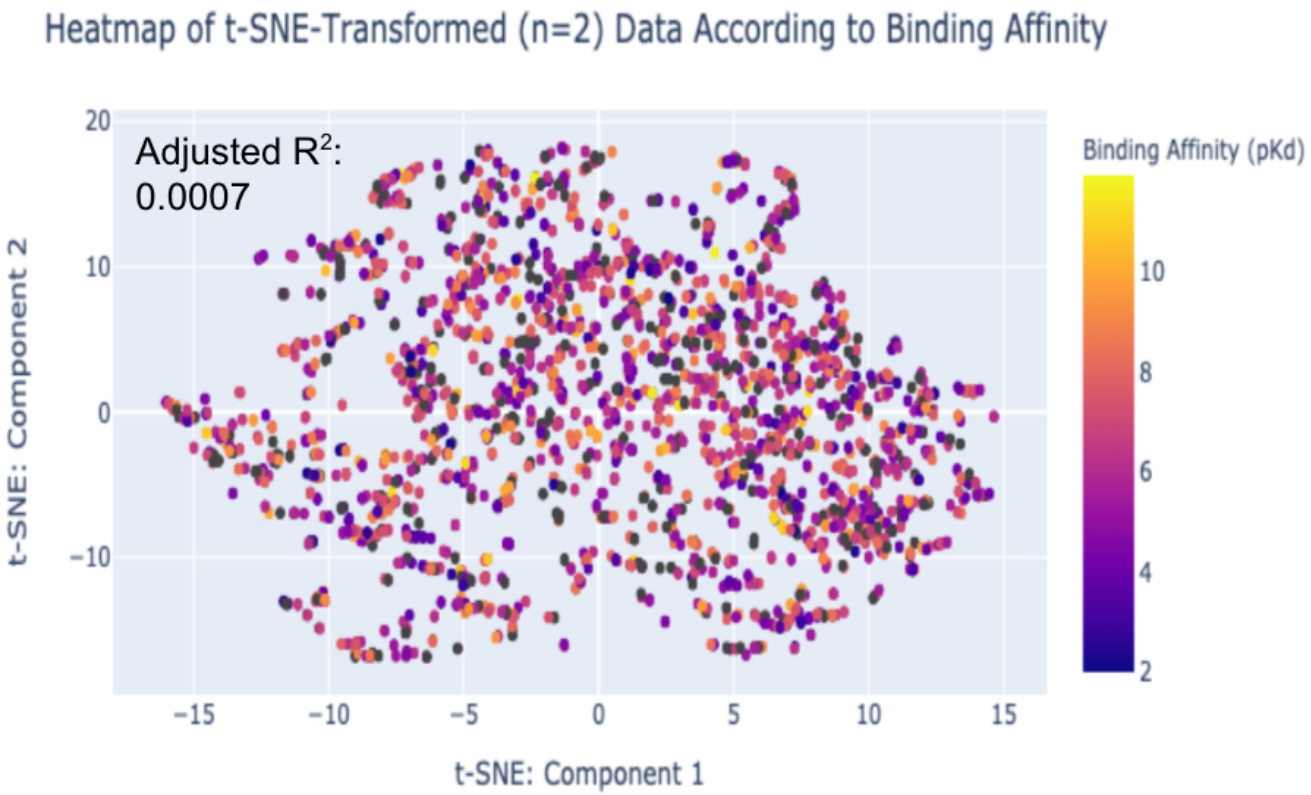

Fig 6. Heatmap of t-SNE (n=2) transformed data with "heat" determined by binding affinity. Plotted using Plotly Python package. Grouping quality calculated using adjusted R^2 correlation value. Plot shows no notable groups of clusters, extremely low R^2 of 0.0007. Indicates that separable groups of complexes do not exist, and thus no significant biomechanical interaction between features.

Figure 6

See image above for figure legend

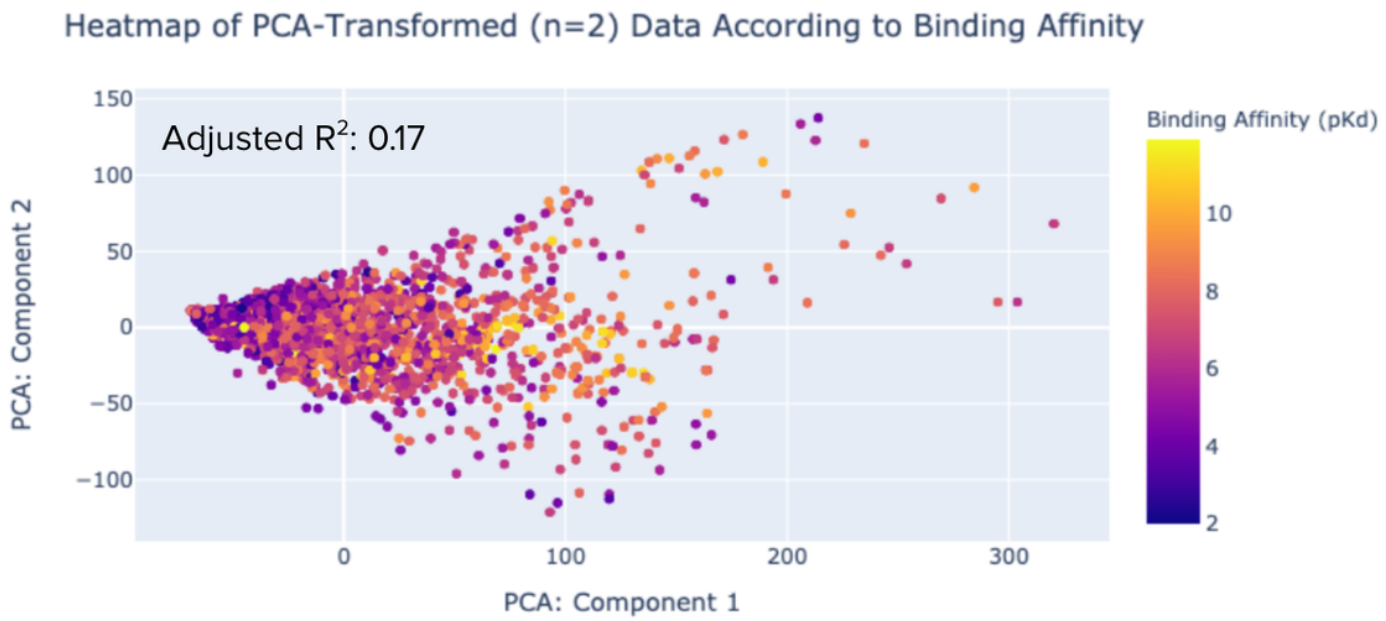

Fig 7. Heatmap of PCA (n=2) transformed data with "heat" determined by binding affinity. Plotted using Plotly Python package. Grouping quality calculated using adjusted R^2 correlation value. Plot shows notable grouping of complexes with pKd<6 and those with pKd>8, supported by higher R^2 of 0.17. Suggests that there is an underlying biomechanical interaction between features that separates complexes into both groups.

Figure 7

See image above for figure legend

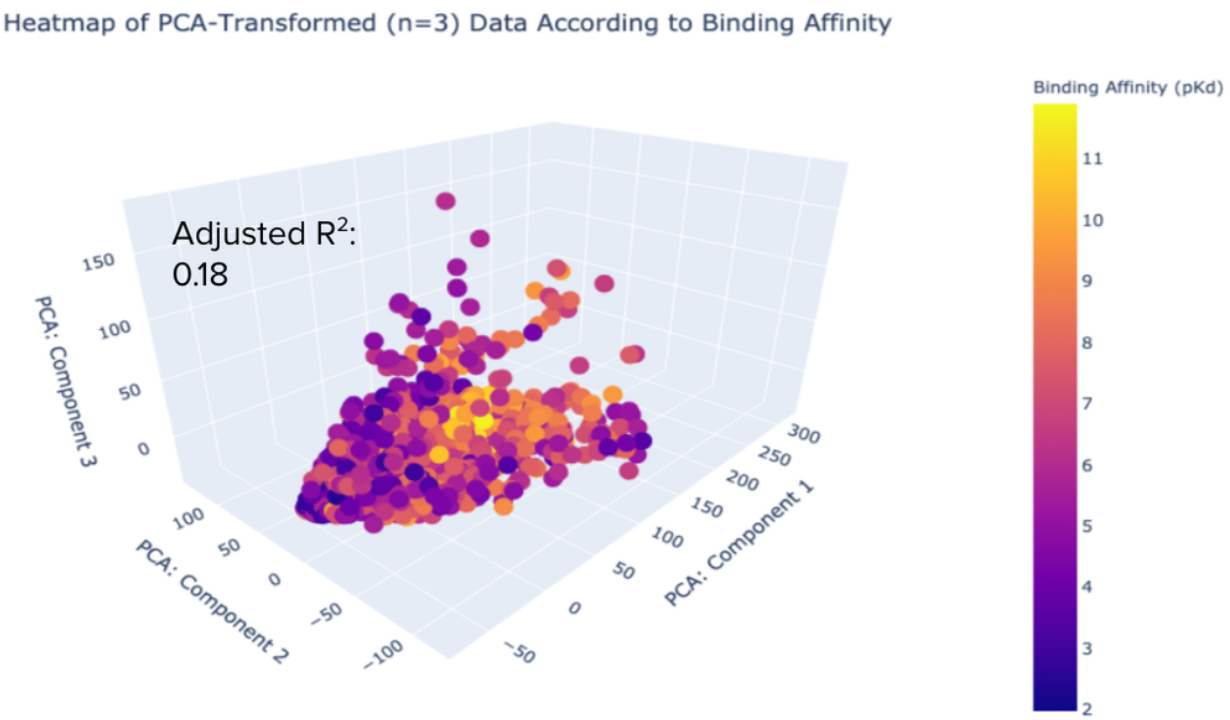

Fig 8. Heatmap of PCA (n=2) transformed data with "heat" determined by binding affinity. Plotted using Plotly Python package. Grouping quality calculated using adjusted R^2 correlation value. Plot confirms visible grouping of complexes with pKd<6 and those with pKd>8, supported by similar R^2 value with the 2D PCA at 0.18. Suggests that there is an underlying biomechanical interaction between features that separates complexes into both groups.

Figure 8

See image above for figure legend

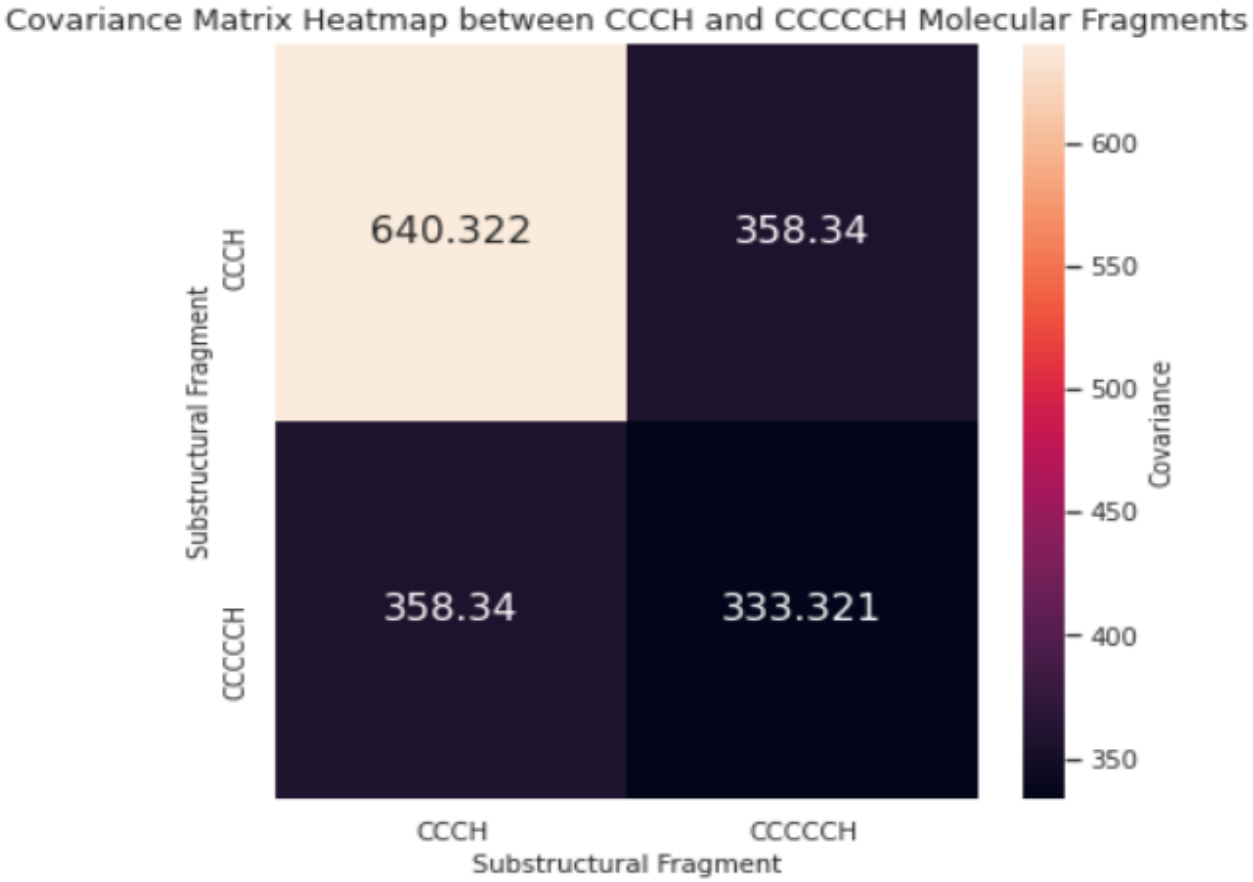

Fig 9. Heatmap of Covariance Matrix between CCCH and CCCCCH substructural molecular fragments. Plotted using Seaborn and Matplotlib Python packages. Heatmap indicates positive covariance between CCCH and CCCCCH counts. Suggests that CCCH and CCCCCH are involved in a biomechanical interaction. The fact that their relationship is direct suggests that the fragments are part of a larger molecular fragment within the ligand.

Figure 9

See image above for figure legend

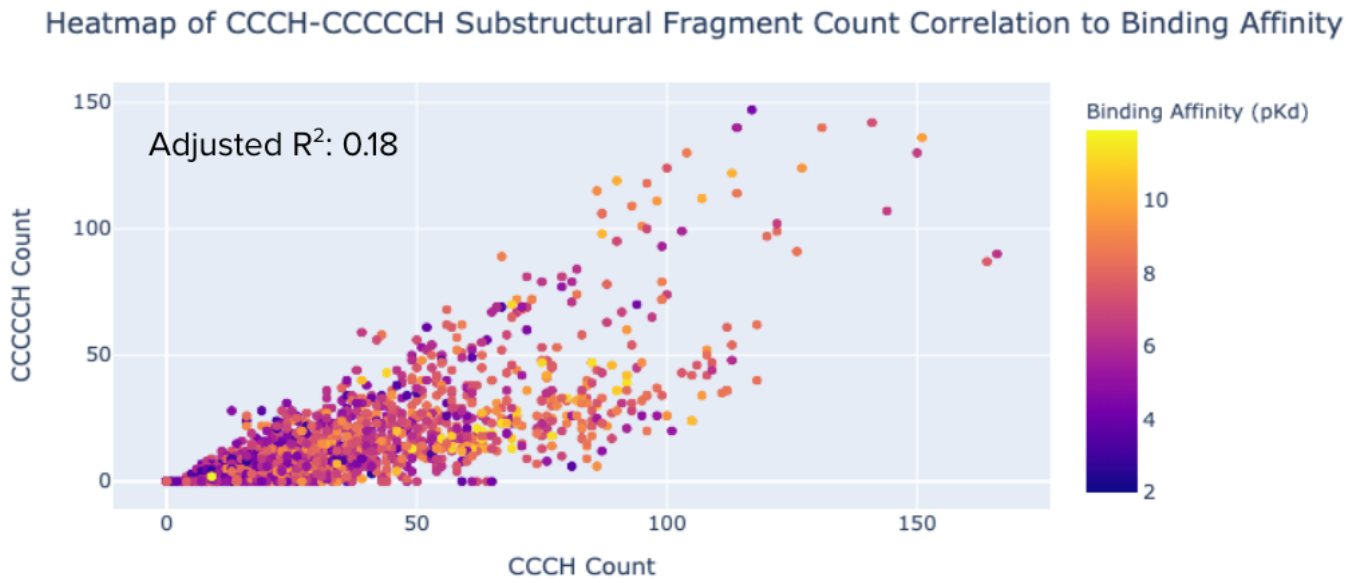

Fig 10. Heatmap of correlation between CCCH-CCCCCH fragment count and binding affinity. Plotted using Plotly Python package. Grouping quality calculated using adjusted R^2 correlation value. Plot shows notable groups of complexes with pKd<6 and pKd>8, and moderate R^2 of 0.18. Suggests that the CCCH-CCCCCH relationship explains significant levels of binding affinity variance.

Figure 10

See image above for figure legend

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.png
- Table2.png