Self-Supervised Machine Learning Approach for Identifying Biochemical Influences on Protein-Ligand Binding Affinity

Arjun Singh
Warren, NJ, USA, 07059
arjsingh2004@gmail.com
Mentor: James Wang
iResearch Institute

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) models can predict binding affinity accurately, they suffer from lack of interpretability and inaccurate feature selection caused by multicollinear data. This study used self-supervised 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. LASSO Regression and hierarchical clustering were utilized to minimize multicollinearity between features. Correlation analyses and Autoencoder-based latent space representations were generated to identify features significantly influencing binding affinity. A Generative Adversarial Network was used to simulate ligands with certain counts of a significant feature, and thereby determine the effect of a feature on improving binding affinity with a given target protein. It was found that the CC and CCCN fragment counts in the ligand notably influence binding affinity. Re-pairing proteins with simulated ligands that had higher CC and CCCN fragment counts could increase binding affinity by 34.99-37.62% and 36.83%-36.94%, respectively. This discovery contributes to a more accurate representation of ligand chemistry that can increase the accuracy, explainability, and generalizability of ML models so that they can more reliably identify novel drug candidates. Directions for future work include integrating knowledge on ligand fragments into supervised ML models, examining the effect of CC and CCCN fragments on fragment-based drug design, and employing computational techniques to elucidate the chemical activity of these fragments.

Keywords

Machine Learning; Drug Discovery; Protein-Ligand Binding Affinity; Self-Supervised Algorithms; Feature Engineering

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 (Fig. 1) [4].

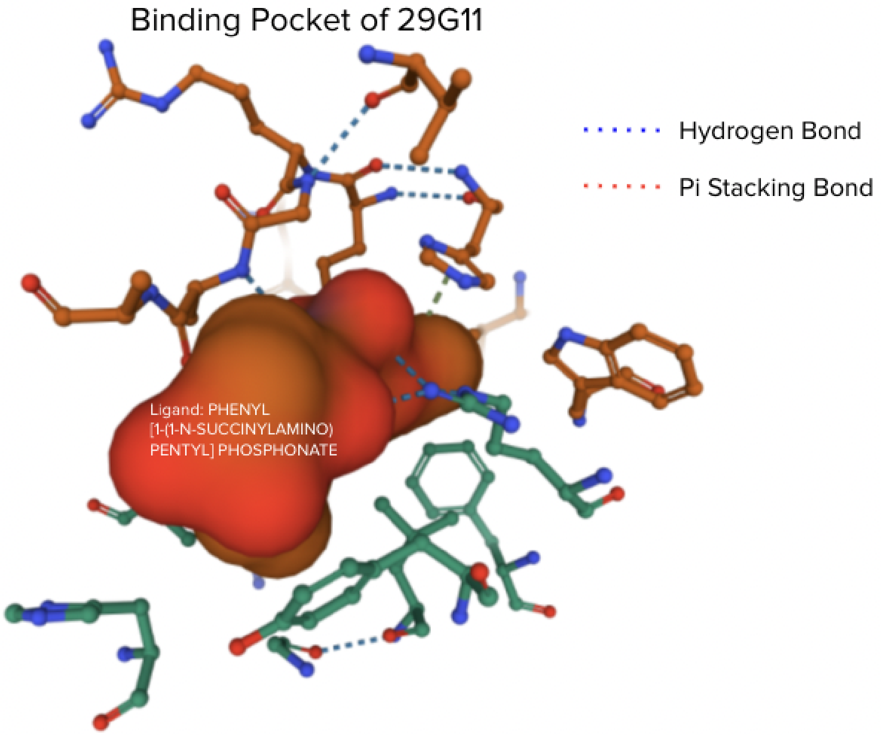


Fig 1. Molecular view of complex between 29G11 protein and PHENYL [1-(1-N-SUCCINYLAMINO)PENTYL] PHOSPHONATE, generated using Mol*. Ligand (bolded red) experiences specific interactions with the protein binding pocket (surrounding region) that are critical in determining binding affinity. Intra-ligand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

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 computer 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 chemical 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]. Beyond simply predicting the binding affinity, supervised ML algorithms can be used to determine the importance of certain pharmaco-like features in influencing the binding affinity, thereby revealing important insights that can inform the development of innovative drugs [8]. However, supervised ML algorithms suffer when multicollinearity - the presence of significant intercorrelations between two or more independent variables - exist in a dataset [12, 13]. This is because supervised feature analysis methods, the most popular being Random Forest feature

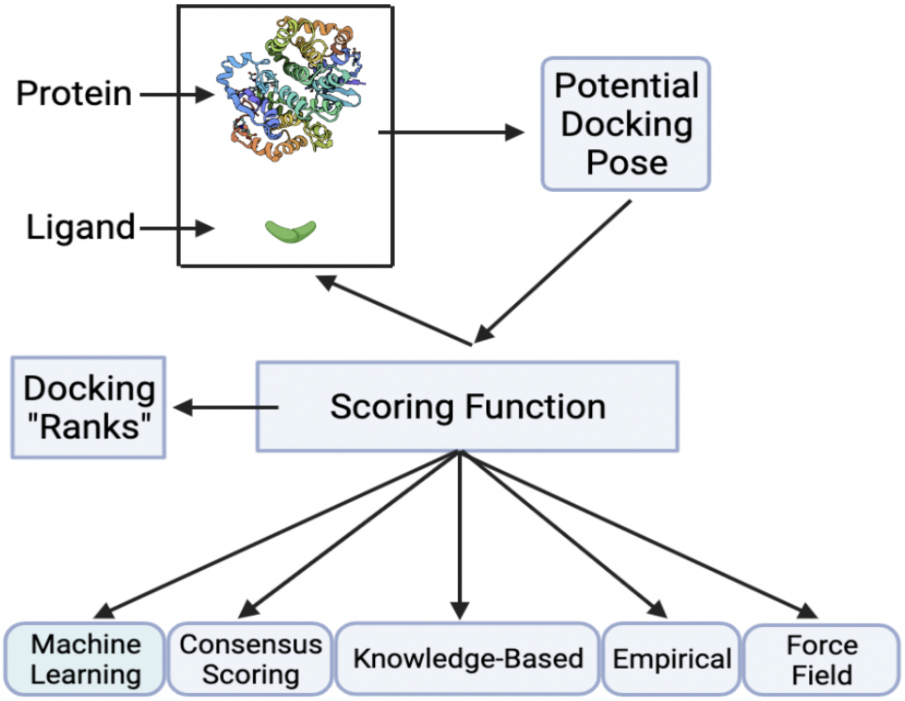


Fig 2. Virtual screening workflow. Docking poses are generated using a molecular simulation software and each pose is inputted into scoring function to calculate the binding affinity. The scoring function utilize either knowledge-based rules in physics and chemistry or learned rules to calculate the 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.

selection, rely on differential performance with/without certain predictor variables in order to determine the importance of that variable [14]. Therefore, when predictor variables are highly intercorrelated with one another, calculated "feature importance" scores do not accurately represent the independent contribution one predictor variable has to a response. In addition, supervised nonlinear models (e.g. Random Forests or Deep Neural Networks) that are successful in predicting a response output accurately suffer from lack of interpretability, meaning that informative patterns learned by the algorithm cannot be easily extracted. Therefore, supervised learning models have been given the term"black box", and are problematic for analyzing the features patterns from large-scale datasets [15, 16].

On the other hand, correlation analysis techniques such as Spearman's Rank Correlation and R^2 values have demonstrated high interpretability and computational efficiency in analyzing molecular binding properties [17, 78]. In addition, self-supervised learning techniques such as Autoencoder Networks and Generative Adversarial Networks have been shown to be useful in learning low-dimensional representations from high-dimensional datasets, thereby capturing significant patterns that can verify the quantitative importance of a feature [18, 19]. Correlation analysis and self-supervised learning techniques have not yet been applied to reveal the differences between protein-ligand complexes specifically in regards to their binding affinity. This is a research gap that can be filled to address the "black box" nature of supervised machine learning algorithms and also reveal significant biochemical insights into the most important features of protein-ligand complexes that influence their binding affinity.

Objectives:

There is a pressing need to more reliably identify and analyze biochemical features that influence binding affinity. Current literature either suffer from drawbacks in interpretability caused by supervised learning or do not account for the multicollinearity present in protein/ligand feature datasets. The objectives of this study are three-fold: 1) Account and rectify multicollinearity present between features of protein-ligand complexes, 2) Identify specific biochemical features responsible for high variance in binding affinities, and 3) Quantify the effect of these features on improving the binding affinity of drug complexes.

Gathering a greater understanding of which features influence binding affinity is necessary for developing ML algorithms that achieve higher hit rates during VS. VS will thereby be more efficient and effective, significantly improving the critical stages of early drug discovery.

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 extract relevant quantitative features of each model, a workflow described in [40] was utilized (Fig. 3).

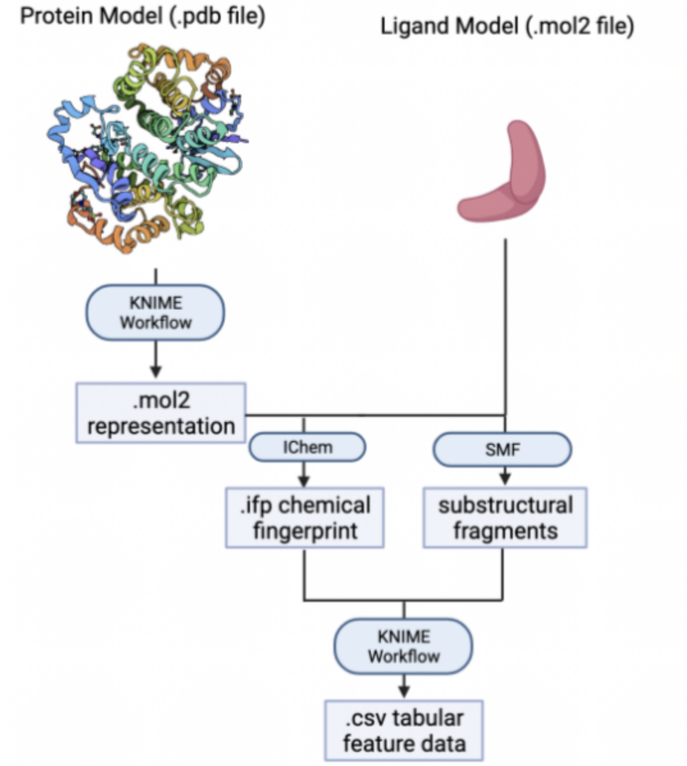

Protein Model (.pdb file)
Ligand Model (.mol2 file)
KNIME Workflow
.mol2 representation
IChem
SMF
.ifp chemical fingerprint
substructural fragments
KNIME Workflow
.csv tabular feature data

Fig 3. Computational workflow used to translate 3D molecular models into 1D tabular data. Tabular data can be used for machine learning analysis. Proposed in [46]. Dataset proposed by [88]. Each protein was converted to .mol2 using Konstanz Information Miner (KNIME). The protein .mol2 file and the ligand .mol2 file were used to generate an interaction “fingerprint” using IChem. Molecular fragment collections from the ligand were also generated using Substructural Molecular Fragments (SMF). Relevant features were collected from the interaction fingerprint and molecular fragment collections using another KNIME workflow.

For each complex, 2422 quantitative features were collected. 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 experimentally determined binding affinity (in pKd) for each complex was also collected from the PDBBind database to serve as a target variable.

Minimizing Multicollinearity:

Supervised ML models suffer from lack of interpretability when multicollinearity is present in a dataset [12, 13]. Therefore, the variance inflation factor (VIF) of each feature was calculated to quantitatively determine multicollinearity. VIF measures the factor by which the variance of a feature's estimated ordinary least squares (OLS) regression coefficient is increased due to correlation with other features [20].

It was observed that the VIF of most features was significantly above the recommended value of 10 [21]. Therefore, Standard Scaling and LASSO Regression were used to identify and remove unimportant features (those with a regression coefficient of 0) from the dataset [86]. After the regression was performed, significantly high VIF were still observed in many features. Hierarchical clustering was used to cluster the remaining features. One feature from each cluster was selected so that features providing redundant information were removed [87]. The remaining 34 features all had VIF < 10.

To ensure that the 34 features could accurately predict binding affinity, two Random Forest Regressors were trained on the "Refined" set (hereon referred to as the training set) and tested on the "Core" set (hereon referred to as the testing set) to predict binding affinity. The first regression tree was trained/tested on all 2422 features whereas the second was trained/tested on the 34 selected features. The R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (PCC) between the predictions and ground-truth values on the test set was compared for each regressor to ensure that the 34 features retained significant predictive capability. A Random Forests model was used because of its exceptional predictive performance compared to more standard models such as Linear Regression, which cannot accurately represent the information stored in non-linear datasets [26].

Correlation Analysis:

Supervised feature importance calculation methods suffer from being uninterpretable, coined with the term "black box" [15, 16]. Therefore, the correlation between features was analyzed to determine potential significant features [80, 81]. The Spearman Correlation between each feature and binding affinity was calculated. In addition, the adjusted R^2 value was calculated between each pair/triplet of features and binding affinity to ensure that features with high independent influence were not insignificant in the presence of certain other features [79, 82, 83]. Each of the three resulting lists were ordered from highest correlation to lowest. The features that were common among the top 25th percentile of correlations in each list were extracted (Fig. 4). The 25th percentile was chosen as the threshold because it was the maximum percentile in which not every one of the 34 features were common among that percentile between all the three lists. It was observed that 8 features were common among the top 25th percentile of each list. The description of each feature is discussed in the Results section.

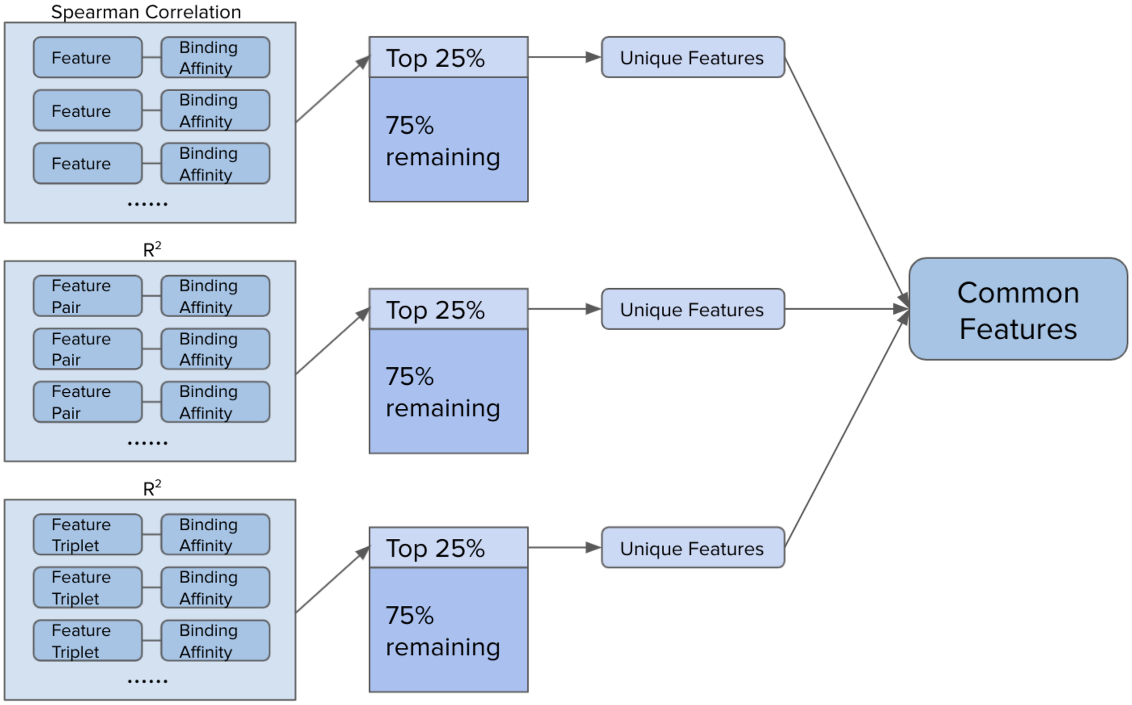


Fig. 4. Correlation analysis workflow. The Spearman Correlation between each feature and binding affinity was calculated. The R^2 value between each pair and triplet of features with binding affinity was calculated. The features that appeared in the top 25% of correlations in each group were extracted. The features that were common among the unique features of each group were selected as potentially significant features for further analysis.

Autoencoder to Filter Features:

It has been shown in current literature that autoencoders are effective self-supervised models for learning low-dimensional representations of multivariate data [18]. Therefore, an autoencoder was designed and implemented to analyze complexes with high and low counts of each of the 8 features. A three-stage autoencoder was designed and implemented to compress the 34-feature dataset into two dimensions (Fig. 5).

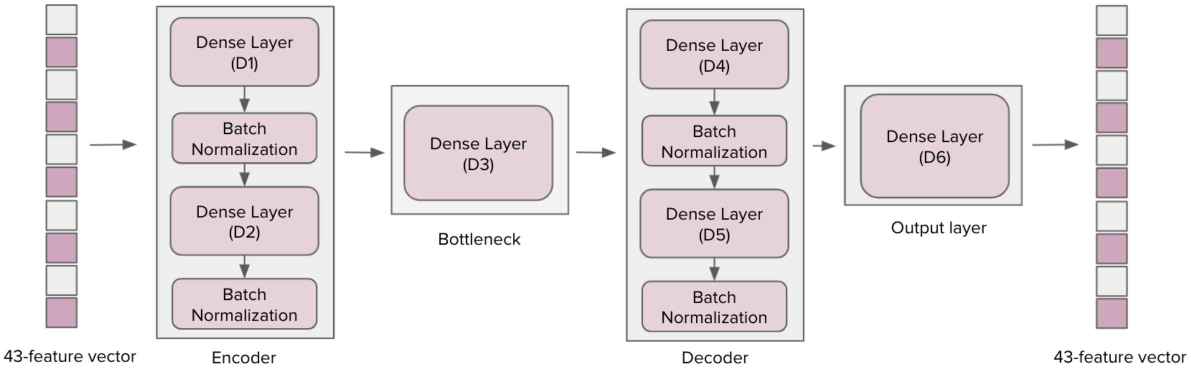


Fig 5. Autoencoder design. The autoencoder took a 43-feature vector (representing one protein-ligand complex) as its input. Two Dense layers (D1 and D2) and Batch Normalization were used to encode the input vector. D1 compresses the 43-feature vector into 21 features. D2 compresses the 21-feature vector into 10 features. A third Dense Layer (D3) was acted as a “bottleneck”, compressing the 10-feature vector into 2 features. Two more Dense Layers (D4 and D5) and Batch Normalization were used to decode the bottleneck vector. D4 expanded the 2-feature vector into 10 features. D5 expanded the 10-feature vector into 21 features. The last Dense Layer (D6) was used to expand the 21-feature vector back to its original 43-feature state. Each Dense layer utilized a Leaky ReLU activation function. The model utilized the Adam optimization function and Mean Squared Error loss function.

The autoencoder was trained on the training set for 30 epochs with a batch size of 32. The testing set was used for validation at each training epoch. In order to use the autoencoder to verify if the 8 features were influential, the “latent space” of the autoencoder was extracted. This refers to the 2-dimensional output produced by the third Dense Layer. This is because if the latent space showed a difference in binding affinity between complexes with high and low counts of a certain feature, then that feature must be influential on the complex’s binding affinity [18]. After training concluded, the Encoder and Bottleneck layers were used to calculate the latent space representation for all complexes in the training set. This resulted in a 2D representation of every complex in the training set.

For each of the 8 features, the 90th and 10th percentile of counts in the training set were calculated. Only complexes with a feature count greater than or equal to the 90th percentile or less than or equal to the 10th percentile for that feature were extracted. Each of these complexes' latent space was graphed on a 2D heatmap, with the “heat” determined by the binding affinity of that complex. This was repeated for each of the 8 features to determine if high/low counts of a certain feature held a significant influence on binding affinity in the latent space representation. Following Cohen’s Effect Size, which has shown to be useful for analyzing biomedical and molecular datasets, any feature whose heatmap exhibited an adjusted R^2 value greater than 0.25 was determined to have a "large" Effect Size on binding affinity [22, 77, 84]. It was determined that out of the 8 features, the CC and CCCN substructural molecular fragments had large Effect Sizes. No augmented atoms or amino acid residue counts exhibited large Effect Sizes.

Generative Adversarial Network to Simulate Novel Higher-Affinity Complexes:

Although the autoencoder was utilized to determine the significance of CC and CCCN fragment counts, the effect of these features on improving binding affinity was not examined. Analyzing the ability of the feature counts to improve the binding affinity of weakly-bound complexes is important to understand the application of these features to real-world VS and drug design. Generative Adversarial Networks (GANs) have been shown to be effective at generating synthetic data for molecular datasets and learning representations of protein-ligand complexes [19, 71-73]. They are also successful at generating *de novo* molecules, accurately simulating the development/discovery of new ligands for inclusion in novel drugs [74-76]. Therefore, in order to quantify the effect of significant features on increasing binding affinity, a GAN proposed in [23] was implemented and trained on the 34-feature training set. This GAN utilizes a Long Short-Term Memory (LSTM)-based generator and Multi-Layer Perceptron (MLP)-based discriminator. After training concluded, it was used to generate 3481 synthetic protein-ligand complexes, mirroring the size of the training set. Following [23], the quality of these synthetic complexes was determined by calculating a "similarity score": the Spearman Correlation between the logarithmic transformations of the means and standard deviations of all features in each dataset (real and synthetic). The log-transformed means and standard deviations for each dataset were concatenated to produce two lists that were used to calculate the Spearman Correlation.

After it was verified that the synthetic data was sufficiently representative of the training set, each of the high-Effect-Size features were used to identify ligands in the synthetic dataset that could improve the binding affinity of low-affinity (chosen as pKd<4 in this study) complexes from the testing set.

First, 26 low-affinity complexes were identified in the testing set. Every ligand in the synthetic dataset that had a certain feature count less than or equal to *n* higher than the same feature count in the low-affinity complex's ligand was selected. For each selected ligand, its features were concatenated into a list with the amino acid residue counts of the low-affinity complex's protein. This list was a 34-feature vector, representing a new protein-ligand complex made up of the low-affinity complex's protein and a synthetic ligand. A Grid-Search-optimized Random Forests model, which was trained on the training set and achieved a high PCC on the testing set of 0.75 and low RMSE of 1.58, was utilized to predict the binding affinity of this new complex.

Although the binding affinity of the original complex was already available, the Random Forests model was used to predict the binding affinity of the original complex as well, in order to account for prediction error when comparing it with the predicted values of the synthetic complexes. The percent increase from the original complex's predicted binding affinity to the mean of the synthetic complex binding affinities was calculated and recorded for each low-affinity complex. Each result was checked for statistical significance using the Cumulative Distribution Function (CDF) test on the Z-Score-normalized values of the entire distribution of binding affinities (predicted values of synthetic complexes and predicted value of original complex) [24]. Only percent increases with an associated CDF value below 0.05 were considered statistically significant. Before performing the CDF test, the distribution was first determined to be approximately normal through the Kolmogorov–Smirnov test [25]. Distributions were concluded to be approximately normal if the p-value of the Kolmogorov-Smirnov test was below 0.05. This entire workflow was repeated for all integer values of *n* from 2-10 for a given feature, with the CC fragment count and CCCN fragment count being the features of interest (Fig. 6).

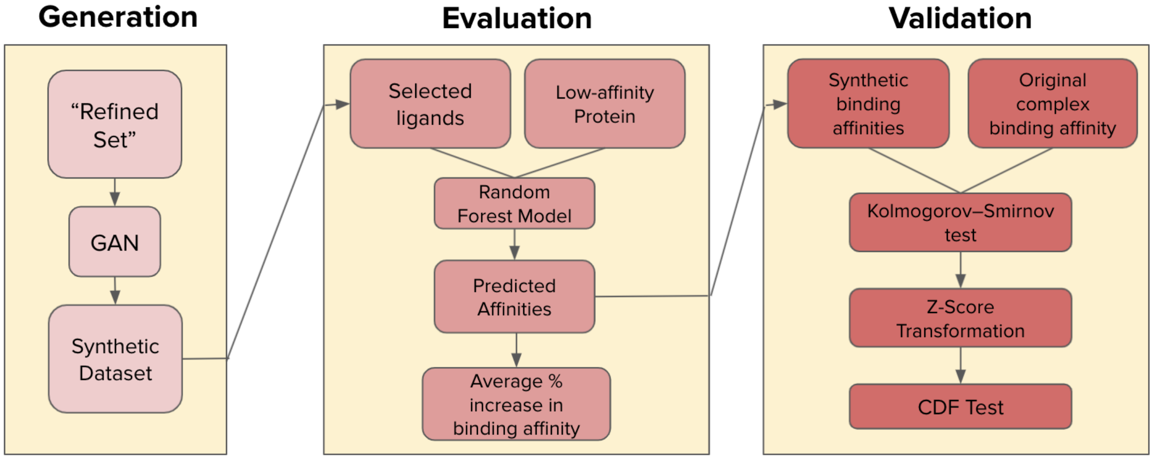


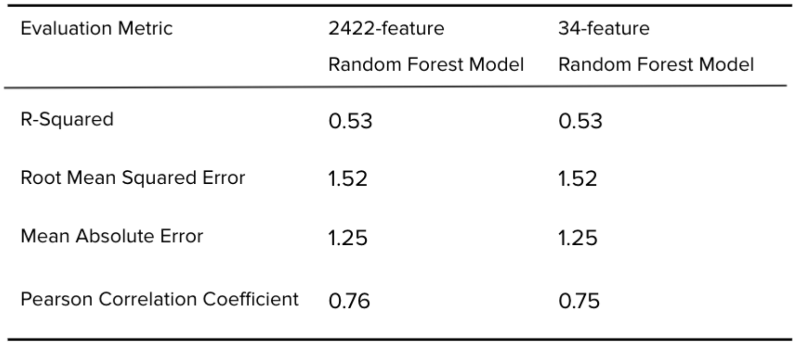
Fig. 6. Simulated GAN-based testing workflow. A Generative Adversarial Network (GAN) was trained on the "Refined Set" and used to generate a synthetic dataset of 3481 complex. The dataset was filtered for ligands with a feature count (CC or CCCN fragment count) that is less than or equal to *n* greater than a certain low-affinity complex's ligand corresponding feature count. For every selected ligand, its features were paired with the low-affinity complex's protein and the binding affinity was predicted. The % increase from the low-affinity complex's predicted binding affinity and the mean of the synthetic complex predicted binding affinities was calculated. The result was checked for statistical significance using a Kolmogorov-Smirnov test and CDF Test preceded by Z-Score Transformation. The entire workflow was repeated for the CC and CCCN fragment count feature independently.

RESULTS AND DISCUSSION

Minimizing Multicollinearity:

The two Random Forests models trained on the 2422-feature dataset and 34-feature dataset, respectively, did not show significantly different predictive performance on their corresponding testing sets (Table #1). All performance statistics were equivalent between both models with the exception of 0.01 increase in Pearson Correlation Coefficient for the 2422-feature model. This suggests that the 34 non-collinear features accurately represent the chemical properties influencing binding affinity stored in the original "Refined" dataset. Therefore, the 34 features were deemed sufficient for use in further analysis.

Table #1. *Evaluation Metrics of Random Forest model trained on 2422-feature dataset and 34-feature dataset*

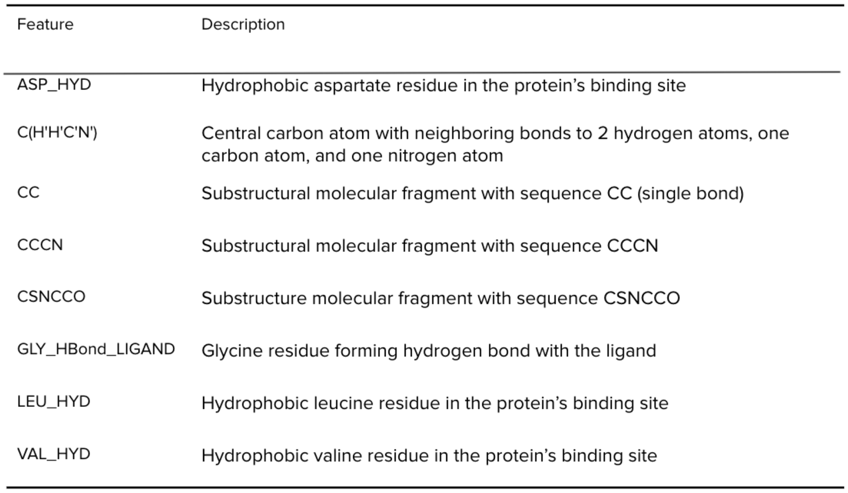


Evaluation Metric	2422-feature Random Forest Model	34-feature Random Forest Model
R-Squared	0.53	0.53
Root Mean Squared Error	1.52	1.52
Mean Absolute Error	1.25	1.25
Pearson Correlation Coefficient	0.76	0.75

Correlation Analysis:

The correlation analysis described in the Methods section resulted in eight features that were common among the top 25% of correlations in each group (Table #2). These eight features were considered to likely hold significant influence on the binding affinity of a complex.

Table #2. *Description of each feature selected for potential significance from correlation analysis*



Feature	Description
ASP_HYD	Hydrophobic aspartate residue in the protein's binding site
C(H'H'C'N')	Central carbon atom with neighboring bonds to 2 hydrogen atoms, one carbon atom, and one nitrogen atom
CC	Substructural molecular fragment with sequence CC (single bond)
CCCN	Substructural molecular fragment with sequence CCCN
CSNCCO	Substructure molecular fragment with sequence CSNCCO
GLY_HBond_LIGAND	Glycine residue forming hydrogen bond with the ligand
LEU_HYD	Hydrophobic leucine residue in the protein's binding site
VAL_HYD	Hydrophobic valine residue in the protein's binding site

Autoencoder to Filter Features:

The autoencoder described in the Methods section was evaluated by measuring the training and testing loss/accuracy over the training period. The autoencoder learned to accurately reconstruct a complex's features from a 2D representation, as it exhibited a high testing accuracy of 0.9667, a low testing loss of 0.8735, and exponential-like training loss with a final training loss of 0.9297 (Fig. 7). This suggests that the autoencoder was effective in extracting an accurate latent space representation of all complexes in the training set.

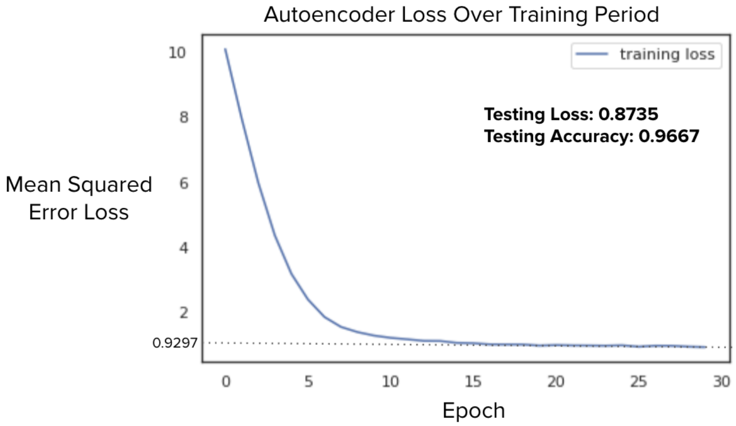


Fig. 7. Autoencoder performance over training period. The Mean Squared Error Loss on the training set was graphed over each training epoch. The final training loss achieved was 0.9297. The autoencoder learned to accurately reconstruct a complex's features, as evidenced by the low testing loss of 0.8735 and high accuracy of 0.9667. Plotted using keras utils python library.

As described in the Methods section, complexes with feature counts above the 90th percentile or below the 10th percentile were graphed on a heatmap using both latent space dimensions and the binding affinity as the "heat". The CC and CCCN fragment count features demonstrated high Effect Size, with R^2 values of 0.39 and 0.30, respectively (Fig. 8C-D). No other features demonstrated high Effect Size (Fig. 8) [22]. This suggests that the CC and CCCN fragment count features play a significant chemical role in increasing binding affinity, and that they are distinctly more influential than any other feature [27].

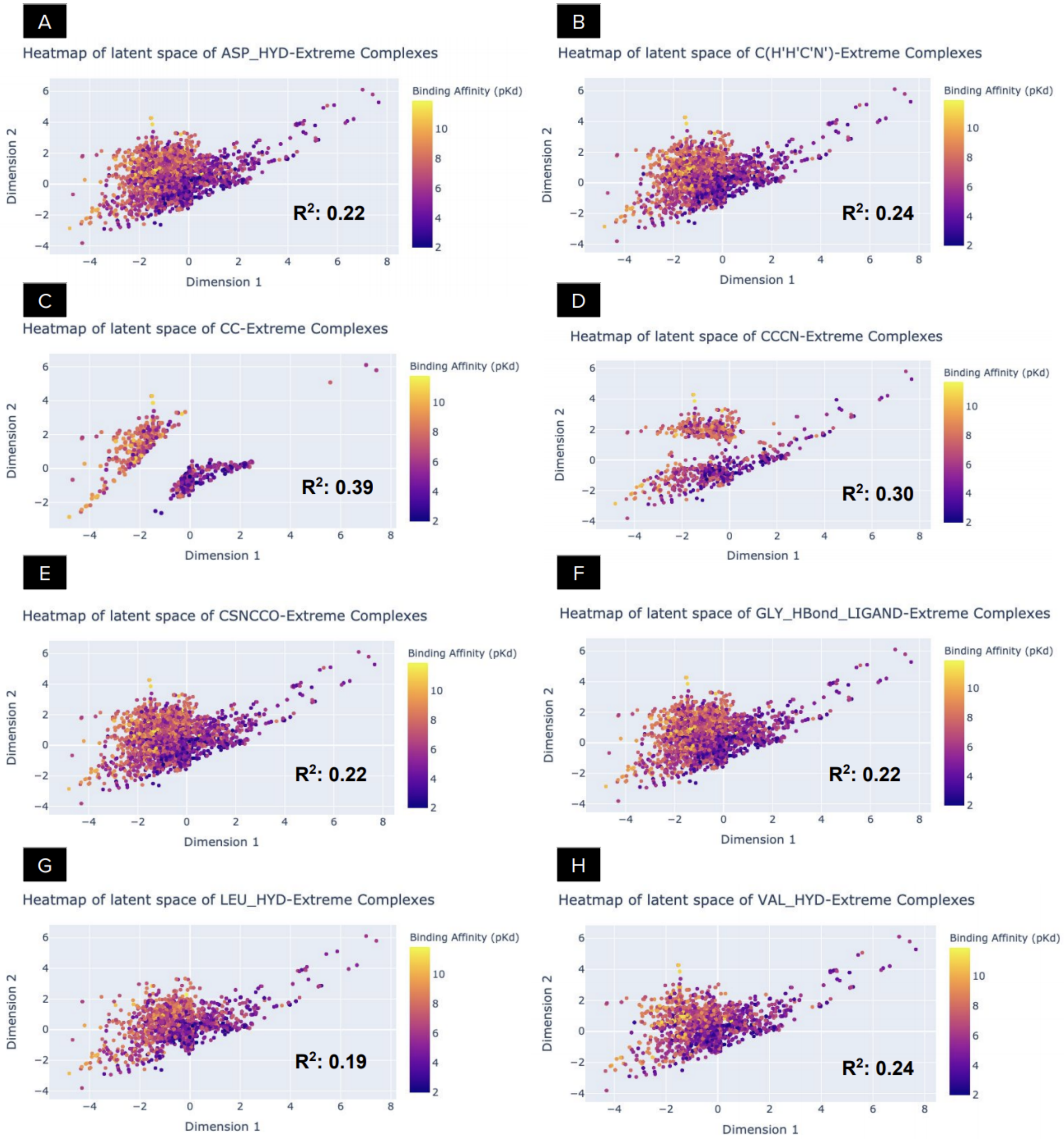


Fig. 8. Latent space heatmap of complexes with high/low feature values. For each feature, only complexes with the count of that feature greater than or equal to the 90th percentile or less than or equal to the 10th percentile were graphed (percentiles calculated using the distribution of a feature count in the training set). Dimension 1 and Dimension 2 are the two dimensions of the latent space (two features outputted by the bottleneck of the autoencoder). R^2 value was calculated to measure correlation between latent space dimensions and binding affinity. Plots C and D show the CC and CCCN fragment counts to be high-Effect-Size features (R^2 >= 0.25) [22]. No other feature demonstrated a high enough R^2 to have a high Effect Size. This suggests that the CC and CCCN fragment counts are significantly influential on binding affinity. Heatmaps plotted using the Plotly Python library.

Generative Adversarial Network to Simulate Novel Higher-Affinity Complexes:

The GAN described in the Methods section produced 3481 synthetic complexes, the quality of which was calculated by comparing the means and standard deviations of each feature between the original and synthetic dataset. It was observed that there is a strong correlation between the properties of the original and synthetic dataset, as evidenced by a high “similarity score” of 0.9912 (Fig. 9) [23]. This suggests that the synthetic data was accurately representative of the real dataset.

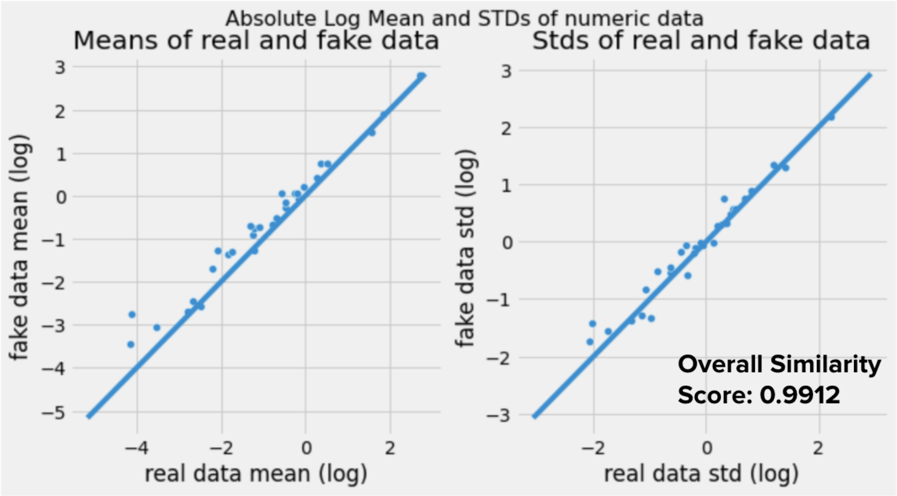


Fig. 9. Statistical comparison of synthetic (“fake”) and original (“real”) data. The logarithmic transformations of the means of each feature was graphed, as well as the logarithmic transformations of the standard deviations of each feature. The overall similarity score, which is the Spearman Correlation between the concatenated transformed means and standard deviations for the real dataset and fake dataset, was calculated. The high similarity score of 0.9912 indicates that the fake data accurately represents the real data.

26 Low-affinity (pKd < 4) complexes in the testing set were selected (Table #3). Ligands

Table #3. *Description of each low-affinity (pKd < 4) complex in the testing set*

Complex ID	PDB ID	Description	Complex ID	PDB ID	Description
1	10gs	HUMAN GLUTATHIONE S-TRANSFERASE P1-1, COMPLEX WITH TER117	121	3f3c	LeuT bound to 4-Fluoro-L-Phenylalanine and sodium
5	1f8c	Native Influenza Neuraminidase in Complex with 4-amino-2-deoxy-2,3-dehydro-N -neuraminic Acid	125	3g2n	N-acylglucosylamine with glycogen phosphorylase
14	1lol	Crystal structure of orotidine monophosphate decarboxylase complex with XMP	132	3huc	Human p38 MAP Kinase in Complex with RL40
24	1p1q	GluR2 ligand binding core (S1S2J) L650T mutant in complex with AMPA	133	3i3b	E-Coli (lacz) Beta-Galactosidase (M542A) in Complex with D-Galactopyranosyl-1-on
27	1q8u	The Catalytic Subunit of cAMP-dependent Protein Kinase in Complex with Rho-kinase Inhibitor H-1152P	137	3k5v	Abl kinase in complex with imatinib and GNF-2
33	1u33	D-gluconohydroximino-1,5-lactam novel alpha-amylase inhibitor	141	3l4u	N-terminal Human Maltase-Glucoamylase with de-O-sulfonated kotalanol

Table #3 (Continued)

50	2gss	HUMAN GLUTATHIONE S-TRANSFERASE P1-1 IN COMPLEX WITH ETHACRYNIC ACID	142	3l4w	N-terminal Human Maltase-Glucoamylase with miglitol
65	2qmj	N-terminal Subunit of Human Maltase-Glucoamylase in Complex with Acarbose	143	3lka	Catalytic domain of human MMP-12 complexed with hydroxamic acid and paramethoxy-sulfonyl amide
66	2r23	S25-2 Fab in complex with Kdo analogues	145	3mss	Abl kinase in complex with imatinib and fragment (FRAG2) in the myristate site
84	2xbv	Factor Xa in complex with a pyrrolidine-3,4-dicarboxylic acid inhibitor	147	3myg	Aurora A Kinase complexed with SCH 1473759
93	2yki	Tricyclic series of Hsp90 inhibitor	162	3su5	NS3/4A protease variant D168A in complex with vaniprevir
102	3ag9	Complex of PKA with the bisubstrate protein kinase inhibitor ARC-1012	169	3zso	Small molecule inhibitors of the LEDGF site of HIV type 1 integrase identified by fragment screening and structure based design
113	3dd0	Carbonic Anhydrase II, IX Active-Site Mimic Isozyme Complex	178	4gid	beta-secretase complexed with inhibitor

in the synthetic dataset were filtered for high CC/CCCN counts and concatenated with the proteins from the low-affinity complexes to determine the effect of each feature on increasing the binding affinity of weakly-bound complexes. Due to non-collinearity, the effect of other features on binding affinity as lurking variables was minimized. The Kolmogorov-Smirnov test and CDF test were utilized to determine the statistical significance of a given increase in binding affinity. Replacing ligands in low-affinity complexes with ligands that had higher CC fragment counts resulted in an average increase in binding affinity of 34.99-37.62% (Table #4). This suggests that

Table #4. *Mean of statistically significant increases in binding affinity of low-affinity complexes based on CC-fragment increases*

Increase index (N)	Mean increase in binding affinity	% of increases that were statistically significant	ID's of complexes that experienced statistically significant increase in binding affinity
2	36.15%	26.92%	27, 66, 102, 132, 141, 145, 169
3	36.73%	26.92%	27, 66, 102, 132, 133, 169, 178
4	37.06%	30.77%	27, 66, 102, 132, 133, 141, 169, 178
5	35.35%	38.46%	27, 33, 66, 102, 132, 133, 141, 145, 169, 178
6	36.51%	38.46%	27, 33, 66, 102, 132, 133, 141, 145, 169, 178
7	37.62%	38.46%	27, 33, 66, 102, 132, 133, 141, 145, 169, 178
8	36.96%	42.31%	5, 27, 33, 66, 102, 132, 133, 141, 145, 169, 178
9	36.29%	46.15%	5, 24, 27, 33, 66, 102, 132, 133, 141, 145, 169, 178
10	34.99%	53.85%	5, 24, 27, 33, 66, 102, 121, 132, 133, 137, 141, 145, 169, 178

a higher CC fragment count can significantly increase the binding affinity of a ligand with a target protein [28-30]. No significant correlation between increase index (value of *n* by which for a generated ligand to be selected, it must have had a CC count less than or equal to *n* greater than the CC count in the low-affinity ligand) and mean increase in binding affinity was observed. This suggests that increasing the CC fragment count can consistently increase binding affinity, but that increasing the amount by which the CC fragment count is increased does not necessarily result in a greater binding affinity. A positive relationship between CC increase index and percent of increases that were statistically significant was observed. This is supported by the fact that increasing the value of *n* also resulted in a greater number of low-affinity complexes experiencing statistically significant increases. This collectively suggests that ligands with greater CC fragment counts more reliably result in greater binding affinities.

The same evaluation process was conducted on the CCCN fragment count feature. Replacing ligands in low-affinity complexes with ligands that had higher CCCN fragment counts results in an average statistically significant increase in binding affinity of 36.83%-39.64% (Table #5). This suggests that a higher CCCN fragment count can significantly increase the binding affinity of a ligand with a target protein [28-30].

Table #5. *Mean of statistically significant increases in binding affinity of low-affinity complexes based on CCCN-fragment increases*

Increase index (N)	Mean increase in binding affinity	% of increases that were statistically significant	ID’s of complexes that experienced statistically significant increase in binding affinity
2	36.83%	38.46%	24, 27, 102, 132, 133, 137, 141, 145, 162, 169
3	37.57%	38.46%	24, 27, 33, 102, 132, 133, 141, 145, 162, 169
4	38.41%	38.46%	24, 27, 102, 132, 133, 141, 143, 145, 162, 169
5	39.60%	38.46%	24, 27, 102, 132, 133, 141, 143, 145, 162, 169
6	38.41%	46.15%	24, 27, 33, 102, 132, 133, 137, 141, 143, 145, 162, 169
7	39.46%	46.15%	24, 27, 33, 102, 132, 133, 137, 141, 143, 145, 162, 169
8	39.04%	50.00%	5, 24, 27, 33, 102, 132, 133, 137, 141, 143, 145, 162, 169
9	39.64%	50.00%	5, 24, 27, 33, 102, 132, 133, 137, 141, 143, 145, 162, 169
10	39.12%	53.85%	5, 24, 27, 33, 102, 132, 133, 137, 141, 143, 145, 147, 162, 169

No significant correlation between CCCN increase index and mean increase in binding affinity was observed. This suggests that increasing the CCCN fragment count can consistently increase binding affinity, but that increasing the amount by which the CCCN fragment count is increased does not necessarily result in a greater binding affinity. A positive relationship between CCCN increase index and percent of increases that were statistically significant was observed. This is supported by the fact that increasing the value of *n* also resulted in a greater number of low-affinity complexes experiencing statistically significant increases. This collectively suggests that ligands with greater CCCN fragment counts more reliably result in greater binding affinities.

It was also observed that certain complexes experienced statistically significant increases in binding affinity in both groups. These complexes are those with I.D. 5, 24, 27, 33, 102, 132, 133, 137, 141, 145, and 169. Several complexes did not experience any statistically significant increase in binding affinity. These complexes are those with I.D. 1, 14, 50, 65, 84, 93, 113, and 125. This suggests that CC and CCCN fragments may play a varied chemical role in different low-affinity complexes depending on the ligand's binding mode properties [31, 32].

CONCLUSIONS

In this study, three main objectives were achieved: 1) Multicollinearity was minimized in a dataset consisting of 2422 features per protein-ligand complex, 2) Specific ligand fragments were discovered to have a notable chemical influence on increasing binding affinity, and 3) It was shown that increasing the count of these fragments can significantly improve the binding affinity of protein-ligand complexes. The methods utilized in this study improved upon current literature by: 1) Minimizing the bias in results occurring from multicollinear datasets, and 2) Increasing the reliability and interpretability of feature selection by utilizing a unique pipeline of self-supervised learning and correlation analysis techniques instead of supervised learning methods. It is concluded in this study that CC and CCCN ligand fragments are chemically significant in determining the binding affinity of protein-ligand complexes and can determine ligands that notably improve binding affinity when bound to a target protein.

There are several applications of this work to real-world drug discovery. Understanding the influence of CC and CCCN fragments on binding affinity will produce a more accurate representation of ligand chemical activity, aiding researchers to build ML algorithms that more accurately predict binding affinity [33-36]. Integrating previous knowledge into supervised ML algorithms has been shown to increase predictive performance by an entire order of magnitude, demonstrating the significance of the knowledge uncovered in this study to supervised ML research [37, 39, 40]. More importantly, pre-training knowledge on ligand fragments will result in ML models that overfit less, making them more generalizable to new datasets and thus reliable for analyzing novel drug candidates [43-45]. Incorporating prior knowledge can also increase the explainability of supervised ML models, further reducing their "black box" effect [38]. Increasing the accuracy, generalizability, and explainability of supervised ML models using knowledge such as that concluded in this study will improve industry-wide VS processes by more reliably identifying ligand hits for inclusion in novel drug compounds [41, 42].

The effect of improving ML models for 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 [46]. Using the information revealed in study to develop more accurate, generalizable, and explainable ML models 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 conclusion of this study as well as others to develop more robust ML models is therefore critical for identifying promising drug candidates for innovative medicines.

It is significant to note that the discoveries of this study is useful in other scientific contexts, such as synthetic drug design. Using known information on fragments such as the two focused on in this study (CC and CCCN), 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.

In industrial fields such as process chemistry, ligand fragment activity with target proteins can be used to direct enzyme evolution in biocatalytic reactions [52-55]. Biocatalysis involves the binding of a small molecule (ligand) and an enzyme (protein) to catalyze the chemical reaction of the small molecule [52]. It has been shown that ligand features can guide the development of metathesis catalysts and nitrile hydration catalysts, demonstrating the varied applications of ligand feature information [56-58].

Therefore, it is evident that the significant ligand features elucidated in this study have important applications to Virtual Screening research, synthetic drug design, and industrial processes such as biocatalysis.

LIMITATIONS AND FUTURE WORK

It is significant to note that there are limitations and directions for future work based on this study's methodology and results.

In this study, several arbitrary thresholds were chosen for the purpose of experimentation. For example, it was chosen that complexes with feature counts above the 90th percentile or below the 10th percentile would be considered in determining that feature's influence on binding affinity. It was also chosen that the range of *n* values for the GAN-based simulative testing would be 2-10. Using statistical techniques to guide the decision of all thresholds may lead to slightly different results [75]. Further, employing additional statistical tests such as confidence intervals may increase the reliability of the results [77]. In addition, due to online availability and computational limits, only the PDB-Bind database was analyzed in this study. However, conducting the same methodology on different datasets may support or refute the results of this study [85].

There are several other directions for future work based on the methodology of this study. For instance, other self-supervised and unsupervised learning algorithms such as Principal Component Analysis, t-Distributed Stochastic Neighbor Embedding, and Variational Autoencoders can be used to reveal significant feature-based influences on binding affinity [60-63]. Analyzing the relationship between features instead of just independent features' influence can also reveal significant chemical phenomena that influence binding affinity [64, 65].

In addition to the methodology, the results of this study can lead to future work. For example, integrating the importance of CC and CCCN fragments as prior knowledge into ML models may lead to improved predictions of binding affinity [37, 39, 40, 59]. Investigating the effect of these fragments on fragment-based drug design can lead to *in-silico* techniques that directly improve synthetic drug design [66, 67]. Most importantly, utilizing methods such as 3D quantitative structure–activity relationships (3D-QSAR) will help expand on this study by revealing the specific chemical reason why CC and CCCN fragments improve binding affinity [68-70]. Therefore, there are several exciting directions for future chemical research based on the methodology and conclusions of this study.

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, and 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]: W. Yoo, R. Mayberry, S. Bae, K. Singh, Q. He, and J. W. Lillard, “A Study of Effects of MultiCollinearity in the Multivariable Analysis,” Int. J. Appl. Sci. Technol., vol. 4, no. 5, pp. 9-19, October 2014.

[13]: B. F. Darst, K. C. Malecki, and C. D. Engelman, “Using recursive feature elimination in random forest to account for correlated variables in high dimensional data,” BMC Genet., vol. 19, no. 65, September 2018.

[14]: R. Chen, C. Dewi, S. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” J., vol. 7, no. 52, July 2020.

[15]: Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, “A Survey on Neural Network Interpretability,” IEEE Trans. Emerg. Top. Comput. Intell., vol. 5, pp. 726-742, July 2021.

[16]: F. Drobnič, A. Kos, and M. Pustišek, “On the Interpretability of Machine Learning Models and Experimental Feature Selection in Case of Multicollinear Data,” Electronics, vol. 9, no. 5, May 2020.

[17]: M. M. Mukaka, “A guide to appropriate use of Correlation coefficient in medical research,” Malawi Med J., vol. 24, no. 3, pp. 69-71, September 2012.

[18]: Q. Fournier and D. Aloise, “Empirical comparison between autoencoders and traditional dimensionality reduction methods,” Proc. - 2019 IEEE 2nd Int. Conf. Artif. Intell. Knowl. Eng. AIKE 2019, pp. 211-214, June 2019.

[19]: J. M. Wolterink, K. Kamnitsas, C. Ledig, and I. Išguma, “Chapter 23 - Deep learning: Generative adversarial networks and adversarial methods,” Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 547-574, October 2019.

[20]: L. Murray, H. Nguyen, Y. Lee, M. D. Remmenga, and D. W. Smith, “VARIANCE INFLATION FACTORS IN REGRESSION MODELS WITH DUMMY VARIABLES,” Proc. - 24th Conf. App. Stat. Ag., 2012.

[21]: K. M. Marcoulides and T. Raykov, “Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods,” Educ. Psychol. Meas., vol. 79, no. 5, pp. 874-882, October 2019.

[22]: G. M. Sullivan and R. Feinn, “Using Effect Size—or Why the P Value Is Not Enough,” J. Grad. Med. Educ., vol. 4, no. 3, pp. 279-282, September 2012.

[23]: Y. Hille, “On the Generation and Evaluation of Tabular Data using GANs,” Radboud University, December 2019.

[24]: J. Lawrence, S. Rusinkiewicz, and R. Ramamoorthi, “Adaptive Numerical Cumulative Distribution Functions for Efficient Importance Sampling,” EGSR, pp. 11-20, June 2005.
[25]: F. J. Massey, “The Kolmogorov-Smirnov Test for Goodness of Fit,” J. Am. Stat. Assoc., vol. 46, no. 253, pp. 68-78, March 1951.
[26]: R. Couronné, P. Probst, and A. Boulesteix, “Random forest versus logistic regression: a large-scale benchmark experiment,” BMC Bioinform., vol. 19, no. 270, July 2018.
[27]: S. Nakagawa and I. C. Cuthill, “Effect size, confidence interval and statistical significance: a practical guide for biologists,” Biol. Rev. Camb. Philos. Soc., vol. 82, no. 4, pp. 591-605, November 2007.
[28]: 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.
[29]: 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.
[30]: J. Robson-Tull, “Biophysical screening in fragment-based drug design: a brief overview,” Biosci. Horiz., vol. 11, February 2019.
[31]: Y. Bian and X. Xie, “Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications,” AAPS J., vol. 20, no. 59, April 2018.
[32]: A. K. S. Romasanta, P. van der Sijde, I. Hellsten, R. E. Hubbard, G. M. Keseru, J. van Muijlwijk-Koezen, and I. J. P. de Esch, “When fragments link: a bibliometric perspective on the development of fragment-based drug discovery,” Drug Discov., vol. 23, no. 9, pp. 1569-1609, September 2018.
[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]: C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, “Integrating Machine Learning with Human Knowledge,” iScience, vol. 23, no. 11, November 2020.

[38]: K. Beckh, S. Muller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Reuden, “Explainable Machine Learning with Prior Knowledge: An Overview,” Arxiv (Preprint), May 2021.

[39]: M. Diligenti, S. Roychowdhury, and M. Gori, “Integrating Prior Knowledge into Deep Learning,” Proc. - 2017 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017, pp. 920-923, December 2017.

[40]: N. Muralidhar, M. R. Islam, M. Marwah, A. Karpatne, and N. Ramakrishnan, “Incorporating Prior Domain Knowledge into Deep Neural Networks,” Proc. - 2018 IEEE Int. Conf. Big Data 2018, pp. 36-45, December 2018.

[41]: J. Ricci-Lopez, S. A. Aguila, M. K. Gilson, and C. A Brizuela, “Improving Structure-Based Virtual Screening with Ensemble Docking and Machine Learning,” J. Chem. Inf. Model., October 2021.

[42]: T. B. Kimber, Y. Chen, and A. Volkamer, “Deep Learning in Virtual Screening: Recent Applications and Developments,” Int. J. Mol. Sci., vol. 22, no. 9, pp. 4435, April 2021.

[43]: 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.

[44]: 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.

[45]: 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.

[46]: 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.
[47]: 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.
[48]: 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.
[49]: Y. Bian and X. Xie, “Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications,” AAPS J., vol. 20, no. 59, April 2018.
[50]: 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.
[51]: 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.
[52]: E. L. Bell, W. Finnigan, S. P. France, A. P. Green, M. A. Hayes, L. J. Hepworth, S. L Lovelock, H. Niikura, S. Osuna, E. Romero, K. S. Ryan, N. J. Turner, and S. L. Flitsch, “Biocatalysis,” Nat. Rev. Dis. Primers, vol. 1, no. 46, June 2021.
[53]: R. Martinez and U. Schwaneberg, “A roadmap to directed enzyme evolution and screening systems for biotechnological applications,” Biol. Res., vol. 46, no. 4, pp. 395-405, 2013.
[54]: A. Kumar and S. Singh, “Directed evolution: tailoring biocatalysts for industrial applications,” Crit. Rev. Biotechnol., vol. 33, no. 4, pp. 365-378, December 2013.
[55]: D. Petrovic, V. A. Risso, S. C. L. Kamerlin, and J. M. Sanchez-Ruiz, “Conformational dynamics and enzyme evolution,” J. R. Soc., vol. 15, no. 144, July 2018.
[56]: J. O. Krause, O. Nuyken, K. Wurst, and M. R. Muchmeiser, “Synthesis and Reactivity of Homogeneous and Heterogeneous Ruthenium-Based Metathesis Catalysts Containing Electron-Withdrawing Ligands,” Chem. Eur. J., vol. 10, no. 3, pp. 777-784, February 2004.
[57]: L. Yang, M. Mayr, K. Wurst, and M. R. Buchmeiser, “Novel Metathesis Catalysts Based on Ruthenium 1,3-Dimesityl-3,4,5,6-tetrahydropyrimidin-2-ylidenes: Synthesis, Structure,

Immobilization, and Catalytic Activity," Chem. Eur. J., vol. 10, no. 22, pp. 5761-5770, October 2004.

[58]: R. Garcia-Alvarez, S. E. Garcia-Garrido, J. Diez, P. Crochet, and V. Cadierno, "Arene-Ruthenium(II) and Bis(allyl)-Ruthenium(IV) Complexes Containing 2-(Diphenylphosphanyl)pyridine Ligands: Potential Catalysts for Nitrile Hydration Reactions?" Eur. J. Inorg. Chem., vol. 2012, no. 26, pp. 4218-4230, September 2012.

[59]: S. Roychowdhury, M. Diligenti, and M. Gori, "Regularizing deep networks with prior knowledge: A constraint-based approach," Knowl. Based. Syst., vol. 222, June 2021.

[60]: 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.

[61]: 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.

[62]: 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.

[63]: S. Lin, Y. Lu, C. Y. Cho, I. Sung, J. Kim, Y. Kim, S. Park, and S. Kim, "A review on compound-protein interaction prediction methods: Data, format, representation and model," Comput. Struct. Biotechnol. J., vol. 19, pp. 1541-1556, March 2021.

[64]: J. Heaton, "An empirical analysis of feature engineering for predictive modeling," SoutheastCon., pp. 1-6, July 2016.

[65]: E. C. Blessie and E. Karthikeyan, "Sigmis: A Feature Selection Algorithm Using Correlation Based Method," J. Algorithm Comput. Technol., vol. 6, no. 3, pp. 385-393, September 2012.

[66]: Q. Li, "Application of Fragment-Based Drug Discovery to Versatile Targets," Front. Mol. Biosci., vol. 7, no. 180, August 2020.

[67]: A. Bancet, C. Raingeval, T. Lomberget, M. Le Borgne, J. F. Guichou, and I. Krimm, "Fragment Linking Strategies for Structure-Based Drug Design," J. Med. Chem., vol. 63, no. 20, pp. 11420-11435, June 2020.

[68]: N. Schaduangrat, S. Lampa, S. Simeon, M. P. Gleeson, O. Spjuth, and C. Nantasenamat, “Towards reproducible computational drug discovery,” vol. 12, no. 9, January 2020.

[69]: R. Ragno, V. Esposito, M. Di Mario, S. Masiello, M. Viscovo, and R. D. Cramer, “Teaching and Learning Computational Drug Design: Student Investigations of 3D Quantitative Structure–Activity Relationships through Web Applications,” J. Chem. Educ., vol. 97, no. 7, pp. 1922-1930, June 2020.

[70]: S. Brogi, T. C. Ramalho, K. Kuca, J. L. Medina-Franco, and M. Valko, “Editorial: In silico Methods for Drug Design and Discovery,” Front. Chem., vol. 8, pp. 612, August 2020.

[71]: R. Rong, S. Jiang, L. Xu, G. Xiao, Y. Xie, D. J. Liu, Q. Li, and X. Zhan, “MB-GAN: Microbiome Simulation via Generative Adversarial Network,” GigaScience, vol. 10, no. 2, February 2021.

[72]: L. Zhao, J. Wang, L. Pang, Y. Liu, and J. Zhang, “GANsDTA: Predicting Drug-Target Binding Affinity Using GANs,” Front. Genet., vol. 10, pp. 1243, January 2020.

[73]: O. Mendez-Lucio, B. Bailiff, D. Clevert, D. Rouquie, and J. Wichard, “De novo generation of hit-like molecules from gene expression signatures using artificial intelligence,” Nat. Commun., vol. 11, no. 10, January 2020.

[74]: A. E. Blanchard, C. Stanley, and D. Bhowmik, “Using GANs with adaptive training data to search for new molecules,” J. Cheminformatics, vol. 13, no. 14, February 2021.

[75]: Y. Bian and X. Xie, “Generative chemistry: drug discovery with deep learning generative models,” J. Mol. Model., vol. 27, no. 71, February 2021.

[76]: E. Lin, C. Lin, H. Lane, “Relevant Applications of Generative Adversarial Networks in Drug Design and Discovery: Molecular De Novo Design, Dimensionality Reduction, and De Novo Peptide and Protein Design,” Molecules, vol. 25, no. 14, pp. 3250, July 2020.

[77]: P. Monsarrat and J. Vergnes, “The intriguing evolution of effect sizes in biomedical research over time: smaller but more often statistically significant,” Gigascience, vol. 7, no. 1, January 2018.

[78]: R. Rodriguez-Perez and J. Bajorath, “Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics,” Sci. Rep., vol. 11, no. 14245, July 2021.

[79]: M. Hall, “Correlation-based Feature Selection for Machine Learning,” University of Waikato [PhD thesis], April 1999.

[80]: S. Kumar and I. Chong, “Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States,” Int. J. Environ. Res. Public. Health., vol. 15, no. 12, pp. 2907, December 2018.

[81]: A. Wosiak and D. Zakrzewska, “Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis,” Complexity, vol. 2018, October 2018.

[82]: M. Vettoretti and B. Di Camillo, “A Variable Ranking Method for Machine Learning Models with Correlated Features: In-Silico Validation and Application for Diabetes Prediction,” Appl. Sci., vol. 11, no. 16, pp. 7740, August 2021.

[83]: A. Yassine, C. Mohamed, and A. Zinedine, “Feature selection based on pairwise evaluation,” Proc. - 2017 Int. Sys. Comp. Vision. ISCV 2017, October 2017.

[84]: M. F. Pescosolido, Q. Ouyang, J. S. Liu, and E. M. Morrow, “Loss of Christianson Syndrome Na+/H+ Exchanger 6 (NHE6) Causes Abnormal Endosome Maturation and Trafficking Underlying Lysosome Dysfunction in Neurons,” J. Neurosci., vol. 41, no. 44, pp. 9235-9256, November 2021.

[85]: A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. A. Elwafa, and H. Kurdi, “Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain,” Appl. Sci., vol. 11, no. 2 pp. 796, January 2021.

[86]: N. Herawait, K. Nisa, E. Setiawan, and N. Tiryono, “Regularized Multiple Regression Methods to Deal with Severe Multicollinearity,” Int. J. Stat. App., vol. 8, no. 4, pp. 167-172, 2018.

[87]: C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carre, J. R. G. Marquez, B. Gruber, B. Lafourcade, P. J. Leitao, T. Munkemuller, C. McClean, P. E. Osborne, B. Reineking, B. Schroder, A. K. Skidmore, D. Zurell, and S. Lautenbach, “Collinearity: a review of methods to deal with it and a simulation study evaluating their performance,” Ecogeg., vol. 36, no. 1, pp. 27-46, January 2013.

[88]: 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.