Incorporating Protein Dynamics Through Ensemble Docking in Machine Learning Models to Predict Drug Binding

Fatemah Alghamedy[1], Jeevith Bopaiah[1], Derek Jones[1], Xiaofei Zhang[1], Heidi L. Weiss[1], Sally R. Ellingson, PhD[1]
[1]University of Kentucky, Lexington, KY, USA

Abstract

Drug discovery is an expensive, lengthy, and sometimes dangerous process. The ability to make accurate computational predictions of drug binding would greatly improve the cost-effectiveness and safety of drug discovery and development. This study incorporates ensemble docking, the use of multiple protein conformations extracted from a molecular dynamics trajectory to perform docking calculations, with additional biomedical data sources and machine learning algorithms to improve the prediction of drug binding. We found that we can greatly increase the classification accuracy of an active vs a decoy compound using these methods over docking scores alone. The best results seen here come from having an individual protein conformation that produces binding features that correlate well with the active vs. decoy classification, in which case we achieve over 99% accuracy. The ability to confidently make accurate predictions on drug binding would allow for computational polypharamacological networks with insights into side-effect prediction, drug-repurposing, and drug efficacy.

1 Introduction

Machine learning is currently being used to advance many scientific disciplines, including drug binding predictions, and shows promise in increasing accuracy enough to make reliable polypharmacological predictions. Chemical drug features have been combined with molecular docking scores in machine learning models to rescore the interactions of one candidate drug to multiple proteins [1]. Components of docking scoring functions can be used as features in a machine learning model to greatly improve the accuracy of identifying active compounds in models specific for one protein [2]. A study on protein druggability used protein features in machine learning models to predict whether or not a protein is druggable [3].

Molecular flexibility can contribute to a favorable change in free energy of binding [4, 5, 6]. Protein-ligand complexes undergo a wide range of motions ranging from small changes in binding site residues to large-scale motions of entire protein domains. Molecular docking is an efficient (but not highly accurate) computational method that predicts how and how well a drug will bind to a protein. To keep these calculations efficient in order to investigate large libraries of chemical compounds, proteins are generally kept static or mostly static in which only a few selected side chains can rotate. Commonly, a crystal structure of the protein (experimentally resolved 3-dimensional protein structure) is used in molecular docking in which a ligand was bound for crystallization but removed from the Protein Data Bank (PDB) [7] file for docking. However, proteins may exist in multiple druggable states, and potentially none of these states are those favored for crystallization and may differ from those favored by a different bound compound.

Ensemble docking is used here to describe the process of using multiple possible protein conformations in molecular docking to represent many potentially druggable states. In this case we use molecular dynamics to generate a trajectory of protein movement and select distinct conformations from the trajectory. Previous studies have shown the usefulness of this technique to improve molecular docking in general [8]. However, ensemble docking is still not perfect and there is no way to pick the best conformation used in ensemble docking without prior knowledge on sets of binding and non-binding compounds. There also still remains deficiencies in the molecular docking scoring functions which must assume a fixed functional form.

This paper explores the combination of ensemble docking, biomedical data sources, and machine learning algorithms to greatly increase the accuracy of binding predictions. In this study, we investigate Tyrosine-protein kinase Lck (LCK) which is implicated as a drug target in many cancers and also known to have toxic effects when unintentionally targeted. LCK is also the subject of a previous study in our lab that investigates theoretically more accurate binding calculations than molecular docking[9].

2 Methods

In this paper we perform ensemble docking by first collecting protein conformations from a molecular dynamics trajectory and performing molecular docking with these structures. We collect features on our drug set and also generate features on the protein conformations. We use a random forest regressor method to rate the importance of features and then investigate several machine learning models using K Nearest Neighbors (Knn) machine learning algorithm for classification of active (binding) vs decoy compounds (non-binding).

2.1 Data Collection

The various types of features collected are described below and the total number of features collected for each category is given in Table 1.

- **Binding Descriptors** The conformations of LCK used in this study are those extracted from a molecular dynamics simulation used for molecular mechanics and generalized Born and surface area continuum solvation (MM/GBSA) calculations in a previous study. Details on the ensemble docking method used can be found in the published work [9]. The conformations are extracted from the molecular dynamics trajectory using a root mean squared deviation clustering method of all the atoms near the binding site. The results of the clustering method is distinct conformations in which the first selected conformation has the most similar frames in the trajectory. Therefore, the first conformation is considered to be the conformation in which the protein is in most often, with the last conformation being a rare, but potentially important, state. The previous study used homology modeling to predict more complete structures of different states of LCK and the one used here corresponds to a homology model made using PDB structure 1QCF [10]. This structure is of HCK and chosen based on sequence identity. This model is of the inactive state. Active and decoy compounds for LCK come from the Directory of Useful Decoys (DUD-e) [11] and are prepared for docking using modified ADT scripts and a wrapper script for automation. Docking was performed using VinaMPI [12], which allows the distribution of a large number of Autodock Vina [13] docking jobs on MPI-enabled high-performance computers. The default number of models per docking job (predicted bound drug conformations with binding scores) was changed to collect up to 20 models per docking job which was for the MM/GBSA protocol. The results of the docking jobs were submitted to Autodock Vina using the "–score-only" option to collect the individual terms calculated in the scoring function. This includes terms for gauss1, gauss2, repulsion, hydrophobic, and hydrogen interactions. The values for all the models and averages of each term for all models are kept. There are also features for the final docking score and a normalized ranking.

- **Protein Descriptors** Since we are only considering one protein here, but seven different conformations, we collected features using two different webservers that calculate features using the 3-dimensional structure provided in a PDB file. These servers include Coach [14, 15] and 2struc [16]. Coach calculates features used for ligand binding site predictions and 2struc gives information on the actual secondary structure of a 3-dimensional protein structure (i.e. not predicted secondary structure from amino acid sequence).

- **Drug Descriptors** There are 28,536 different molecules (drugs) for which we collect features. Drug features are calculated using the Dragon Software [17]. Dragon can calculate over 5 thousand molecular descriptors, including the simplest atom types, functional groups and fragment counts, topological and geometrical descriptors, and three-dimensional descriptors. It also includes several property estimations like logP and drug-like alerts like Lipinski's alert. There are 1,058 descriptors excluded because their values are constant values, near-constant values, or standard deviation less than 0.0001. In this study three-dimensional descriptors are also left out because the input structures for Dragon are the predocking structures and not those predicted by molecular docking.

- **Labels** A drug is labeled as the positive class if it is a known active for LCK in DUD-e [11]. The drug is labeled as the negative class if it comes from the DUD-e decoy set for LCK. Since we do not know the conformation in which each drug binds, and it could potentially bind to all conformations, the drug is labeled the same for each conformation.

Table 1: The total number of features in each category for the base dataset, the best 100 features for the whole dataset, and the best features for each protein conformation (Conf 1-7).

Category	Base	Best	Conf 1	Conf 2	Conf 3	Conf 4	Conf 5	Conf 6	Conf 7
Drugs	2,792	99	10	54	20	7	5	10	7
Protein	103	0	0	0	0	0	0	0	0
Binding	109	1	0	1	0	8	0	0	8
The total	3004	100	10	55	20	15	5	10	15

2.2 Feature Selection

A random forest regressor method implemented in scikit-learn [18] is used to rate all of the features by importance when classifying compounds as active vs. decoy. This was done using all conformations at once and also one conformation at a time. The dataset was first balanced by randomly selecting the same number of decoy compounds as actives. We also selected some protein features based on their ability to identify the protein conformation.

2.3 Classification

The classification in this study is done using the scikit-learn [18] Knn implementation. Knn is a fairly simple machine learning algorithm in which a test object is classified by the majority vote of its K nearest neighbors, thus having the same class as most of its neighbors. There are many different measures that quantify the closeness of two objects. Several distance measures, $K = 1, 2, 3, 4,$ and 5 nearest neighbors, and $n = 10, 5,$ and 2 for n-fold cross validation were evaluated. Since the data is unbalanced (with approximately 50 decoys to every active compound), the dataset is first balanced by randomly selecting the same number of decoys as actives. We first select all active drugs in the dataset such that the $label = 1$, and get the total number of all active drugs as a $sampleSize$. For ten experiments, we randomly select a total of $sampleSize$ decoy drugs. The $sample()$ method, which is in pandas.DataFrame(), is used to randomly select the decoy set. After that, we merge the negative and positive data into one dataset and n-fold cross validation is performed on this set. To prepare the data for the machine learning, we split the data into X and y such that X holds all the features except the label and y represents the label column. For each n-fold, we provide indices (sKf) to the knn method (where neighbors = k) for testing and training using the $StratifiedKFold()$ method to ensure a balance of positive and negative classes. Finally, $cross\_val\_predict()$ method is used with knn, X, y, and sKf to predict if the drug is an active or a decoy. The metrics presented in the results is an average over the ten experiments. The process is given in Algorithm 1.

Algorithm 1

```
$positiveData, sampleSize$ = Select all drugs such as $label = 1$
**for** each experiment $e_i = 1, 2, ..10$ **do**
    $negativeData$ = select $sampleSize$ decoy drugs
    merge the $negativeData$ and $positiveData$ into one dataset
    split the data into $X$ = all features except label and $y = label$
    **for** each n=1,2,10 $n$-fold **do**
        $sKf$ = train/test indices to split data into balanced train/test sets
        set up $knn$ with $n\_neighbors = k$
        get the $y - predict$ by running $cross\_val\_predict()$ method with $knn$, $X$, $y$, and $sKf$
    **end for**
**end for**
```

In some cases we want to reduce our feature set to the smallest size that gives the best metrics. In this case, classification was performed with the top 5 features with the best predicted contributions to the classification and the next 5 most important features are incrementally added. The optimal small feature set is selected based on the AUC as defined in Table 2.

Some of our analysis did not work with the cross validation method described. In this case we still balanced the data as described, tested for the same K values, evaluated testing set sizes of 10%, 20%, and 50% of the data, and averaged results over ten experiments.

2.4 Evaluation

In this study we compare three different ways of looking at the data using the machine learning algorithm and also compare the performance to the computed docking score. Docking scores are typically used for ranking compounds from most likely to least likely to bind and there is no standard that defines an exact docking score that determines a binding prediction. In order to compare binding predictions from the docking score alone to the machine learning models, the maximum Youden's index (or J value) is calculated for each model. The best J value is calculated from the docking score receiver operator characteristic (ROC) curve and used as a cut-off to define true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values for the docking results. All the metrics presented in the Results are defined in Table 2. The three different data models are described below.

- **Model 1** In this model, we split the dataset based on the protein conformation number, giving us seven different datasets. The assumption here is that one of the conformations best represents the druggable state.
- **Model 2** In this model, we keep all entries for every drug combined with each conformation. We ensure that all entries for one drug (seven total - one for each conformation) are entirely in the testing or entirely in the training set. The assumption here is that the drug may bind to all conformations of the protein at varying rates.
- **Model 3** In this model, a consolidated dataset is made by keeping one entry per drug. The entry that is kept is the one with the best overall docking score. The assumption here is that the conformation with the best docking score is the preferred conformation for actual binding. This method has shown to improve docking results with ensemble docking [8].

Table 2: Metrics used in this study

Name	Definition	Formula
Youden's index	Performance of dichotomous test. The value 1 indicates a perfect test and -1 indicates a useless test.	$\frac{TP}{TP+FN} + \frac{TN}{TN+FP} + 1$
AUC	Area under the curve for the ROC curve. The probability P that a randomly chosen positive instance ranks higher than a randomly chosen negative one (X_0 and X_1 are ranks for negative and positive instances, respectively).	$P(X_1 > X_0)$
Accuracy	Percent of correct predictions	$\frac{TP+TN}{TP+FP+TN+FN}$
F1	Harmonic mean of precision and recall	$\frac{2TP}{2TP+FP+FN}$
Precision	Positive predictive value	$\frac{TP}{TP+FP}$
Recall	True positive rate	$\frac{TP}{TP+FN}$

3 Results

3.1 Feature Selection

The top 100 features are kept when using the entire dataset for feature selection and also the top 100 features per protein conformation. The types of features important for classifying the entire dataset are given in Table 1. The total number of features that gave the best machine learning model AUC per individual conformation and their categories are given in Table 1 as well. Only conformations 2, 4, and 7 had important binding features. Just over half of the important features for conformations 4 and 7 are binding features. There are 28 protein features that have a weight greater than zero when selected for classification of the protein conformation.

3.2 Youden's index for docking ROC curves

The maximum Youden's index (or J value) is calculated for each model. The best J value is used to define TP, FP, TN, and FN values for each model using docking scores. The best J values and the docking score cut-off for each model are given in Table 3.

Table 3: Youden's Index for each models.

Value	**Model** 1							**Model 2**	**Model 3**
	Conf 1	**Conf 2**	**Conf 3**	**Conf 4**	**Conf 5**	**Conf 6**	**Conf 7**		
Docking Score cut-off	-9.2	-8.2	-8.5	-2.8	-8.3	-8	37.4	-8.1	-8.9
Best J Value	0.1600	0.0829	0.0832	0.1052	0.1142	0.2103	0.0125	0.0874	0.1283

Table 4: Model 1. Conf = conformation; MD = molecular docking; ML = machine learning model; Acc. = Accuracy; Prec. = Precision

All Features										
Conf	**MD AUC**	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
1	0.6189	0.7964	0.5806	0.7964	0.4399	0.8098	0.6625	0.76	0.3294	0.8666
2	0.5448	0.7919	0.5389	0.7919	0.6225	0.8059	0.527	0.7555	0.7603	0.8637
3	0.5397	0.7991	0.5343	0.7991	0.4311	0.812	0.554	0.7633	0.3529	0.8674
4	0.5518	0.7874	0.5496	0.7874	0.6477	0.8017	0.5319	0.7516	0.8279	0.8591
5	0.5739	0.7907	0.5591	0.7907	0.6392	0.8042	0.541	0.7554	0.7809	0.8599
6	0.6477	0.7948	0.6082	0.7948	0.6113	0.8067	0.6066	0.763	0.6162	0.856
7	0.4209	0.7944	0.5063	0.7944	0.6688	**0.8124**	0.5032	0.7473	0.9971	0.89
Best 100 features of the whole dataset										
Conf	**MD AUC**	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
1	0.6167	0.8607	0.5785	0.8607	0.4387	0.868	0.6571	0.8253	0.3294	0.9154
2	0.5453	0.8584	0.5388	0.8584	0.6224	0.8651	0.5269	0.8262	0.7603	0.9081
3	0.5385	0.8526	0.5375	0.8526	0.4328	0.8604	0.5595	0.8176	0.3529	0.9081
4	0.555	0.8503	0.5521	0.8503	0.649	0.8579	0.5337	0.8166	0.8279	0.9037
5	0.5713	0.8549	0.5554	0.8549	0.6372	0.8617	0.5383	0.823	0.7809	0.9043
6	0.6401	0.8544	0.6044	0.8544	0.6091	0.8623	0.6022	0.8181	0.6162	0.9118
7	0.4158	0.8613	0.5063	0.8613	0.6688	**0.8689**	0.5032	0.8241	0.9971	0.919
Best 100 features for conformation										
Conf	**MD AUC**	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
1	0.6186	0.7275	0.5785	0.7275	0.4387	0.7378	0.6569	0.7112	0.3294	0.7665
2	0.5499	0.8304	0.5415	0.8304	0.6238	0.8372	0.5289	0.8052	0.7603	0.8721
3	0.5425	0.6793	0.536	0.6793	0.432	0.6921	0.5571	0.6656	0.3529	0.721
4	0.5454	0.9795	0.549	0.9795	0.6474	0.9798	0.5315	0.9642	0.8279	0.996
5	0.5679	0.7617	0.5517	0.7617	0.6353	0.7772	0.5356	0.7299	0.7809	0.8312
6	0.6422	0.8814	0.6057	0.8814	0.6099	0.8871	0.6038	0.8463	0.6162	0.9322
7	0.4102	0.9816	0.506	0.9816	0.6687	**0.9818**	0.503	0.9714	0.9971	0.9925
Best features for conformation										
Conf	**MD AUC**	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
1	0.6227	0.8959	0.5799	0.8959	0.4396	0.8999	0.661	0.8668	0.3294	0.9356
2	0.5498	0.9031	0.5422	0.9031	0.6242	0.9068	0.5294	0.8736	0.7603	0.9426
3	0.5445	0.9357	0.5379	0.9357	0.433	0.9368	0.5603	0.921	0.3529	0.9532
4	0.5542	0.9928	0.5537	0.9928	0.6498	0.9928	0.5347	0.9893	0.8279	0.9963
5	0.5704	0.914	0.556	0.914	0.6375	0.9164	0.5387	0.8914	0.7809	0.9428
6	0.6444	0.9244	0.6104	0.9244	0.6127	0.9264	0.6094	0.9027	0.6162	0.9515
7	0.4224	0.9949	0.5061	0.9949	0.6687	**0.9949**	0.5031	0.9918	0.9971	0.9979

3.3 Evaluation

Many metrics are reported here, but with a focus on maximizing the F1 score, this metric is bold in all the tables and the best models are highlighted in gray. The results for Model 1 are given in Table 4. The results presented here are for $K = 1$ nearest neighbors, $n = 10$ n-fold cross validation, and running in the 'Auto' mode (in which the algorithm

Table 5: Model 2. MD = molecular docking; ML = machine learning model; Acc. = Accuracy; Prec. = Precision

Best 100 features of the whole dataset									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5534	0.8688	0.5479	0.8712	0.5243	**0.8815**	0.5814	0.8474	0.4781	0.9202
Best 100 features of the whole dataset + 2 best protein features									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5437	0.8556	0.5388	0.8588	0.5106	**0.8687**	0.5735	0.8357	0.4618	0.9056
Best 100 features of the whole dataset + 28 informative protein features									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5396	0.8499	0.5403	0.8484	0.5002	**0.8508**	0.5261	0.8078	0.4778	0.9007

Table 6: Model 2 and 3 using best 100 features

Model	AUC	Accuracy	F1	Precision	Recall
2	0.8688	0.8712	**0.8815**	0.8474	0.9202
3	0.8665	0.8662	**0.8687**	0.8305	0.9116

Table 7: Model 3. MD = molecular docking; ML = machine learning model; Acc. = Accuracy; Prec. = Precision

All Features									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5862	0.7985	0.5552	0.7985	0.6198	**0.811**	0.5413	0.7641	0.725	0.8643
Best 100 Features of the whole dataset									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5894	0.8543	0.5561	0.8543	0.6203	**0.8614**	0.5421	0.8217	0.725	0.9051
Best 100 Features of the wholeset plus the 12 docking features									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5859	0.8227	0.554	0.8227	0.6192	**0.8321**	0.5403	0.7906	0.725	0.8784
Best 100 Features of the whole dataset plus all docking features									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5841	0.7241	0.5529	0.7241	0.6186	**0.7441**	0.5394	0.6941	0.725	0.8019
Docking features alone									
MD AUC	**ML AUC**	**MD Acc.**	**ML Acc.**	**MD F1**	**ML F1**	**MD Prec.**	**ML Prec.**	**MD Recall**	**ML Recall**
0.5881	0.6624	0.5543	0.6624	0.6193	**0.6819**	0.5405	0.6449	0.725	0.7235

attempts to choose the best metric to define the nearness of neighbors). These values usually give the best results with not a lot of variation. We evaluate Model 1 using all the features, the best 100 features from feature selection on the entire dataset, the best 100 features when doing feature selection on each individual conformation, and a subset of the last feature set that gives the best result. Interestingly, conformation 7 always performs the best by machine learning F1 even though it is the worst in many docking metrics. Since we saw a decrease in metric values when going from features selected on the entire dataset to conformations, we did a further analysis on the best subset of features and picked the subset that gave the best machine learning AUC. All conformations perform better with their smaller subset than any conformation using the best 100 features obtained using the entire dataset (the best again is conformation 7 and highlighted in yellow).

The results for Model 2 are given in Table 5. Since we had to ensure drugs were always in the testing or training set in this model we could not run Model 2 with the same cross validation method used for Model 1. The results here are for a 90:10 training/testing split and metrics are averages of running the analysis 10 times with a different random set of decoys. We wanted to test the inclusion of protein conformation features. We tested this by using a base model of the best 100 features selected using the whole dataset and then adding the 2 protein features most important when classifying the conformation and also all 28 with a non-zero importance. We see a decline in performance at each step.

We wanted to run Model 3 with cross validation but also wanted to directly compare Model 2 and Model 3. Therefore, Table 6 gives a comparison of both models using the same protocol as used for Model 2. They both have very similar metrics, with the average of Model 2 slightly higher, but some of the individual runs of each performing better than the other. Table 7 gives the results of Model 3 using cross validation. Here we compare the models using the entire feature set, the best 100 features, the best 100 features plus the base docking features (for the best predicted docking model, averages of all, the final docking score, and a ranking), the best 100 features plus all docking features (includes up to 20 docked models), and the docking features alone. The best 100 features performs the best, but all still have an improvement over docking alone. Using only docking features in the machine learning model actually has a slight decrease for recall compared to the docking calculation (highlighted in yellow), but the better precision with the machine learning model gives it a better F1.

4 Discussion

The best Model 1 is obtained using conformation 7 and the small subset of features obtained doing feature selection on just that conformation's subset of data and only taking the number of features that gives the best AUC. Conformation 7 performs the worst according to docking AUC, however the docking recall is nearly as high as the machine learning recall. This is because the best J-value is actually quite low and the docking cut-off is one that you would not expect to signify an active compound (+37.4, when the most negative scores are best). Because of this cut-off there are a large number of false positives driving down the docking precision. We believe the success of conformation 7 for machine learning is the fact that it has so many binding features that correlate with the active vs decoy classification (8 of the 15 most important features). We see the same trend in the next best Model 1 which is obtained using conformation 4 and its subset of important features, which also has 8 out of 15 as binding features. You can see from the docking cut-off value for classification that conformation 4 and 7 actually have the worst docking values. However, when using the individual components in the machine learning algorithm this information can classify active vs. decoy with great precision and recall. This shows the importance of the terms calculated for a docking scoring functions but the deficiencies in the scoring functions themselves. Since the length of the trajectory used to capture different conformations is only 200 ns, there is not a huge structural variation in the conformations. However, upon visualization it appears that conformation 7 may have a slightly smaller binding pocket which may allow more favorable interactions with the active compounds.

Model 2 looks at keeping all entries in the dataset. Model 2 performs better with the best 100 features than any single conformation by F1 metrics. We did not get any improvement in any metrics by adding protein features. This could be because there is not enough variation in these values to help with predictions. Including such features when there are larger differences may provide further insight. Information on which conformation drugs bind to and which they do not (so drugs could be marked as active to one conformation and decoy to another) could make these features more useful as well. However, creating that dataset would be difficult if not impossible and is discussed further in the limitations. Even though we were not successful using protein features in this study, we include them here for thoroughness. We are also interested in models that incorporate multiple proteins and conformations obtained from longer trajectories, therefore we are still interested in the usefulness of these features. We can see from the Model 3 results that although the binding features play a significant role in good classification when they correlate with the classification, forcing all the binding features to be in the model does not help.

We found in this study that we can successfully use machine learning to increase the prediction of drug binding and using drug features calculated from protein conformations selected from molecular dynamics trajectories increases the predictability of the models. Lessons learned in this study will be used to build models of different proteins to be used in drug discovery applications. In fact, we are already developing an “all kinase” model to predict drug repurposing and potential side-effects.

4.1 Limitations

There are several limitations of this study that warrant further investigation. We are likely overfitting the data, as feature selection was performed on the same data used in the validation. We have omitted a large number of the decoy compounds from feature selection to balance the data. However, in this initial study, we wanted to have a large number of active compounds in the selection to better understand the features that drive the classification. We plan to increase the size of our dataset to enable the omission of a large number of actives for validation in the future. This study utilizes DUD-e, in which the decoys are only assumed non-binders. However, in this study we are using the enhanced version of the dataset that has a reduced number of false decoys by having a more stringent filtering process and experimental validation when available. Another limitation is the fact that active drugs from the DUD-e dataset are labeled as active for all the possible conformations assessed in this study. However, if we only use protein-drug complexes with experimentally resolved conformations we are setting ourselves up for further limitations as they may not relate to how the drug binds to the protein when not in the experimental conditions used to resolve the structures.

4.2 Future Work

Some potential future directions of this research include, (1) testing other machine learning algorithms, (2) incorporating data on multiple possible binding sites, (3) having a strict validation set not used in feature selection, and (4) applying models to novel drug discovery applications.

5 Conclusion

We successfully incorporated ensemble docking, biomedical and biological data sources, and machine learning to improve binding predictions. The addition of protein features did not help this model but it may in cases where there is more variation (such as if multiple proteins are in the dataset). The best results seen here come from having an individual conformation that produces binding features that correlate well with the active vs decoy classification, giving models with over 99% accuracy. We also see that every way we examine the data using machine learning gives improvement over molecular docking alone.

6 Acknowledgements

This research was supported by the Cancer Research Informatics Shared Resource Facility of the University of Kentucky Markey Cancer Center (P30CA177558). This research used computational resources at the University of Kentuckys Center for Computational. This work was supported by the National Institutes of Health (NIH) National Center for Advancing Translational Science grant KL2TR000116 and 1KL2TR001996-01. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the BLUFF.

References

1. Kun-Yi Hsin, Samik Ghosh, and Hiroaki Kitano. Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *PloS one*, 8(12):e83922, 2013.

2. Sarah L Kinnings, Nina Liu, Peter J Tonge, Richard M Jackson, Lei Xie, and Philip E Bourne. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of chemical information and modeling*, 51(2):408–419, 2011.

3. Ali Akbar Jamali, Reza Ferdousi, Saeed Razzaghi, Jiuyong Li, Reza Safdari, and Esmaeil Ebrahimie. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discov. Today*, 21(5):718–724, May 2016.

4. Erika Balog, Torsten Becker, Martin Oettl, Ruep Lechner, Roy Daniel, John Finney, and Jeremy C Smith. Direct determination of vibrational density of states change on ligand binding to a protein. *Physical review letters*, 93 (2):028103, 2004.

5. Kei Moritsugu, Brigitte M Njunda, and Jeremy C Smith. Theory and normal-mode analysis of change in protein vibrational dynamics on ligand binding. *The Journal of Physical Chemistry B*, 114(3):1479–1485, 2009.

6. Erika Balog, David Perahia, Jeremy C Smith, and Franci Merzel. Vibrational softening of a protein on ligand binding. *The Journal of Physical Chemistry B*, 115(21):6811–6817, 2011.

7. Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, pages 675–684. Springer, 2006.

8. Sally R Ellingson, Yinglong Miao, Jerome Baudry, and Jeremy C Smith. Multi-conformer ensemble docking to difficult protein targets. *The Journal of Physical Chemistry B*, 119(3):1026–1034, 2014.

9. Xiaofei Zhang, Amir Kucharski, Wibe A de Jong, and Sally R. Ellingson. Towards a better understanding of on and off target effects of the lymphocyte-specific kinase lck for the development of novel and safer pharmaceuticals. *Procedia Computer Science*, 108:1222–1231, 2017.

10. Thomas Schindler, Frank Sicheri, Alexander Pico, Aviv Gazit, Alexander Levitzki, and John Kuriyan. Crystal structure of hck in complex with a src family–selective tyrosine kinase inhibitor. *Molecular cell*, 3(5):639–648, 1999.

11. Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

12. Sally R Ellingson, Jeremy C Smith, and Jerome Baudry. Vinampi: Facilitating multiple receptor high-throughput virtual docking on high-performance computers. *Journal of computational chemistry*, 34(25):2212–2221, 2013.

13. Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

14. Jianyi Yang, Ambrish Roy, and Yang Zhang. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, 29(20):2588–2595, 2013.

15. Jianyi Yang, Ambrish Roy, and Yang Zhang. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.

16. DP Klose, Bonnie A Wallace, and Robert W Janes. 2struc: the secondary structure server. *Bioinformatics*, 26 (20):2624–2625, 2010.

17. Kode srl. *Dragon (software for molecular descriptor calculation) version 7.0.6*, 2016. `https://chm.kode-solutions.net`.

18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.