

# Improving the Prediction of Kinase Binding Affinity Using Homology Models

Jeffrey Chyan  
Rice University  
Dept. of Computer Science,  
MS132  
PO Box 1892  
Houston, TX 77251  
Jeffrey.Chyan@rice.edu

Mark Moll  
Rice University  
Dept. of Computer Science,  
MS132  
PO Box 1892  
Houston, TX 77251  
mmoll@rice.edu

Lydia E. Kavradi\*  
Rice University  
Dept. of Computer Science,  
MS132  
PO Box 1892  
Houston, TX 77251  
kavraki@rice.edu

## ABSTRACT

Kinases are a class of proteins very important to drug design; they play a pivotal role in many of the cell signaling pathways in the human body. Thus, many drug design studies involve finding inhibitors for kinases in the human kinome. However, identifying inhibitors of high selectivity is a difficult task. As a result, computational prediction methods have been developed to aid in this drug design problem.

The recently published CCORPS method [3] is a semi-supervised learning method that identifies structural features in protein kinases that correlate with kinase binding affinity to inhibitors. However, CCORPS is dependent on the amount of available structural data. The amount of known structural data for proteins is extremely small compared to the amount of known protein sequences. To paint a clearer picture of how kinase structure relates to binding affinity, we propose extending the CCORPS method by integrating homology models for predicting kinase binding affinity. Our results show that using homology models significantly improves the prediction performance for some drugs while maintaining comparable performance for other drugs.

## Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: Biology and genetics; I.2.6 [Artificial Intelligence]: Learning; I.5.1 [Pattern Recognition]: Models—*Statistical*

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Protein kinases; prediction of binding affinity; semi-supervised learning; protein structure comparison; homology models

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BCB '13, September 22 - 25, 2013, Washington, DC, USA  
Copyright 2013 ACM 978-1-4503-2434-2/13/09 ...\$15.00.

## 1. INTRODUCTION

The protein kinase family is one of the largest protein families. Approximately 1.7% of the human genome consists of protein kinases [19]. Protein kinases are an important part of the human body, affecting a wide variety of cell signaling pathways. Some of the bodily functions affected by kinases include cell growth, immune responses, brain function, and blood pressure regulation [5]. As a result, developing inhibitors for protein kinases is a popular issue in drug design [4].

Our goal is to predict the binding affinity between inhibitors and kinases. Accurate prediction of kinase binding affinity for potential drugs could significantly decrease the amount of time and money needed for drug discovery. There are two challenges to predicting binding affinity of inhibitors for kinases: the highly conserved binding site of kinases and the lack of experimental structural data. We will elaborate on both below. Despite the many different processes kinases are involved in, their binding site is extremely conserved. Binding affinity predictions are thus necessarily based on very subtle variations among all kinases. Inhibitors can be classified by the two main binding modes: type I and type II [17]. Type I kinase inhibitors generally have low selectivity. The more recently developed type II inhibitors achieve higher selectivity by binding to both the ATP site as well as an adjacent, hydrophobic site. The subtle changes in the binding site make it difficult for sequence-based analysis to predict the binding affinity [3]. Thus, we turn to structural analysis of kinases to reveal more differences between them that could attribute to different binding affinities. This leads us to the next challenge: the lack of experimental structural data. Relative to protein sequential data, there is much less protein structural data. According to the PFAM:Pkinase family in release 25 of the Pfam database [24], there are 56691 sequences and only 172 (0.3%) of those sequences have one or more referenced PDB structures. As a result, structures may not fully cover the range of naturally occurring variation. In addition, with relative little structural data, spurious patterns can emerge that are unrelated to binding affinity.

We propose a method that extends the Combinatorial Clustering Of Residue Position Subsets (CCORPS) method [3] by integrating homology models. The structure-based, semi-supervised CCORPS method addresses the first challenge in our problem to some degree. We integrate homology models to address both challenges, providing more structural data

to paint an accurate picture of kinase structures to enhance prediction of binding affinities for inhibitors. The homology models are generated based on known template structures through comparative modeling [9]. In our work we use Modeller [7], which is one such method that generates homology models. More specifically, our method uses homology models from the ModBase database [23], which contains homology models calculated by the ModPipe pipeline [6] that uses Modeller [7].

Our results demonstrate significant improvement in kinase binding affinity prediction using our extended CCORPS method. We observed the prediction performance on several drugs improve with the addition of just a few homology models relative to the number of experimental structures used. We cover specific examples in our results section and highlight the different effects of homology models on prediction performance.

## 2. RELATED WORK

Interest in structural analysis of proteins has grown as more structural data has become available. Recent work shows that similar, specific structural features can be found across the kinase family, and the structural similarities can be used to analyze kinase selectivity [15, 21, 11, 14, 20]. The Pocket-FEATURE method [16] finds similar “micro-environments” between two binding sites. Using this method, 9 kinase binding site pairs that have the same inhibitor and similar micro-environments have been found. Recently, another method called CCORPS [3] has been introduced to predict functional classifications and annotations from available structural data. Existing methods have several problems that CCORPS addresses. CCORPS does not rely on a single, representative structure for individual kinases as other methods do [3]. By allowing multiple structures for individual kinases, CCORPS is able to use information from different conformations. Also, while existing methods only use information on kinase binding sites known to bind to inhibitors, CCORPS also uses data on binding sites known *not* to bind to inhibitors [3].

The use of predicted protein structures has also been the subject of recent work. All the methods discussed above are reliant on available experimental structural data for input. By contrast, the I-TASSR method is an automated method that computationally *predicts* protein structure and then uses this information to predict protein function [26]. This method defines protein function as Enzyme Commission numbers [27], Gene Ontology terms [1], or binding site. However, I-TASSR does not predict binding selectivity with inhibitors, or more specifically, kinase binding affinity. I-TASSR performs prediction by matching predicted models to known proteins. Unlike our method, it does not identify specific structural features among known structures and homology models to predict kinase binding affinity.

Homology models, generated from comparative modeling [9], are predicted structures that have great potential to impact computational methods. In [25] it was shown that docking ligands with kinase homology models performs just as well or better than crystal structures. In recent work [10] it was shown that homology models can be used to more effectively compare binding sites supporting the use of homology models in prediction methods. These works established that homology modeling methods have matured to the point that homology models can be useful in structure-based modeling and prediction of function. Building on this line of work, we

show in this paper that homology models are useful in kinase binding affinity prediction.

## 3. METHOD OVERVIEW

The method in this paper extends the previously published CCORPS method. It is essential to understand CCORPS before discussing how we integrate homology models. We will first cover an overview of CCORPS in the following section followed by a detailed description of homology model integration in the subsequent section.

### 3.1 CCORPS Summary

The CCORPS method is a semi-supervised learning method that finds correlations between protein structural features and annotation labels to predict labels for a set of proteins. The method takes as input a set of protein structures and a set of labels for some of the proteins. Corresponding substructures are extracted from each protein structure. A substructure is defined as a set of residues in the 3-D protein structure. The correspondence between substructures can be determined by a variety of methods such as multiple sequence alignment and substructure matching [22]. The output of CCORPS consists of the predicted labels for proteins with no known label. We break CCORPS down into four steps: substructure extraction, clustering, prediction, and validation.

The first step in CCORPS is to obtain substructures from a set of protein structures. We determine the substructures with an input representative structure from the input set of proteins and a set of residue positions of interest. With the residue positions, the corresponding residues from each protein structure are selected using a multiple sequence alignment to get substructures. After getting the substructures, CCORPS generates all possible subsets of size  $k$  residues from the  $r$  residue positions. Previous work in CCORPS showed that subsets of size 3 are ideal [3]. Each subset represents a specific  $k$ -tuple of residue positions. At this point, we have  $\binom{r}{k}$  subsets where each subset contains the  $k$ -tuple substructure of each substructure we have extracted from protein structures. In the case of kinases for our experiments, we have  $\binom{27}{3} = 2925$  subsets. See [3] for details on how the 27 binding site residues were selected.

After obtaining substructures and generating subsets, the next step of CCORPS is to perform clustering. For each subset an isolated clustering is performed to compare substructures on only the specific  $k$  residue positions of the subset. In each clustering, a pairwise distance matrix of the substructures is generated. Each row of the matrix is a so-called feature vector of a substructure representing how a substructure compares to other substructures structurally. Each component in the feature vector for a substructure is the distance from another substructure. CCORPS uses a distance measure that takes into account geometric and chemical differences. The side-chain centroid LRMSD between two substructures is used as the geometric difference. The chemical difference takes into account various chemical properties of the amino acids [3]. Principal Component Analysis (PCA) is used to reduce the dimensionality of the feature vectors. Once each substructure is represented by a reduced feature vector, Gaussian Mixture Model clustering is performed on the reduced feature vectors using the MCLUST package [8].

The next step in CCORPS is the prediction of labels for

the set of proteins that have unknown labels. CCORPS then selects highly predictive clusters (HPCs) from the clusterings. A cluster is considered highly predictive if all labeled substructures in the cluster have the same label. HPCs suggest that some structural feature common to all substructures in the cluster is correlated with the majority label of the HPC. For each protein substructure that we are predicting for, CCORPS goes through each subset clustering to identify the cluster that the protein substructure lies in and to tally votes for labels based on the majority label of the clusters that are HPCs. The maximum number of votes a substructure can have is equal to the number of subsets, since a substructure can only fall in one cluster per clustering. However, the clusters the substructure falls in are not necessarily HPCs, so the number of votes is typically less than the number of subsets. Once we have the votes for labels, we need a decision boundary to determine what prediction to make based on the votes. CCORPS uses a support vector machine that is trained on the training set of substructures that have known labels to determine the decision boundary.

The final step is to perform cross-fold validation to assess the ability of CCORPS to predict labels. The protein substructures are divided into test sets based on 70% sequence identity clusters. Each of these sequence identity clusters represents one fold, where the structures in the sequence identity cluster constitute the test set, and the remaining structures are the training set. The number of structures in each test set varies. For each fold, the structures in the test set are labeled as “unknown.” The test set structures are ignored when calculating purity for finding HPCs and predictions are made for the test set structures. The SVM is trained on the training set to determine the decision boundary for prediction.

### 3.2 Homology Model Integration

The CCORPS method was built to use known PDB structures. Our method extends CCORPS to use homology models and known structures in order to reveal structural features that may not be detected from known structures alone. To each of the four steps in CCORPS mentioned in our overview—substructure extraction, clustering, prediction, and validation—we make modifications to integrate homology models. We will discuss the changes made in each step of CCORPS below.

In substructure extraction, previous work in CCORPS selects PDB structures based on referenced PDB IDs in a Pfam alignment file [24] for sequences. A majority of the sequences in the alignment file have no references for PDB structures. The original CCORPS pipeline ignores sequences with no referenced PDB structures. We want to obtain homology models for sequences with no known PDB structure. Thus, we now consider all sequences in the alignment file. Each sequence is mapped to PDB structures if there are known PDB structures. For sequences without a PDB structure, we search a set of homology models for structural data for the sequence. The set of homology models is a set of structural data in addition to PDB structures. We have selected homology models from the ModBase database [23]. More details on the specific dataset of homology models used is provided in the dataset section. For each sequence in the alignment file we obtain the UniProtKB ID [18] for the sequence, which is mapped to the UniProtKB accession numbers [18]. Afterwards, we obtain the unique sequence

ID used for homology models from the UniProtKB accession numbers. The homology models refer to protein sequences by a unique sequence ID determined in ModBase. We obtained a file from ModBase containing a mapping of accession numbers to the unique sequence IDs. The process of selecting homology models for a sequence was difficult, as each step in the mappings of sequence IDs could have aliasing. Once all sequences are mapped to structures, substructures are extracted from each protein structure, whether it is a PDB structure or a homology model. Homology model files are in PDB format so parsing and extracting substructures works in the same way as with known PDB structures. Filters are applied such that we consider only valid, high quality homology models. These filters are discussed in the dataset section. After filters, a certain number of homology models are selected dependent on the experiment. The setup for the experiments are discussed in the results section.

In the clustering step of CCORPS, PCA is used to reduce the dimensionality of feature vectors for each substructure. We have to consider how to integrate homology models in the dimensionality reduction. Since homology models are structural data predicted based on templates, experimental structures from PDB are considered more reliable. Thus, we decided to perform PCA on PDB structures only to capture information based on reliable, known structures. We do not want to capture information that might represent the modeling method instead of actual structural information from nature in our principal components. After getting the principal components, the transformation is applied to all structural data: PDB structures and homology models. This allows for consistent principal components as we vary the number of homology models in our experiments so that we can isolate the effect of homology models on the clusterings and subsequent predictions.

In prediction, CCORPS applies labels to all substructures in each subset clustering. With the addition of homology models, we now have to consider how to apply labels to homology models. The default process in CCORPS is to label the substructures with a label of “unknown” if there is no known binding affinity data. However, labeling the homology models as “unknown” restricts the potential impact of homology models on prediction. With unknown labels, homology models can only indirectly impact prediction by changing the clusterings. Initial experiments using unknown labels for homology models confirmed this issue by showing little change in prediction performance. We remedy the problem of homology model labels by having homology models inherit the binding affinity label of the template structure the homology model is based on. While this may introduce some error in binding affinity labels, the additional information from homology models outweighs the possible error as shown later in the results. Labeling homology models in this manner assumes that the model is similar enough to the template structure that they are likely to have similar binding affinity to drugs. By inheriting the label of template structures, homology models could then directly affect HPC voting and predictions.

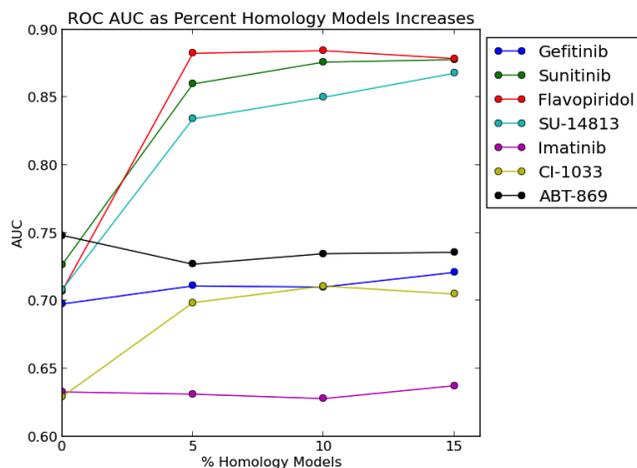
The final step in CCORPS is validation. Our homology model integration was validated by comparing the performance of the modified CCORPS pipeline with previous CCORPS experiments. In order to compare prediction performance across experiments with different homology models, we selected test sets just for PDB structures in cross-fold

validation. We do not make predictions for homology models in validation. This means that homology models are in the training set for every fold. We have to account for the fact that some homology models may be derived from structures in the test set. If the template PDB structure for a homology model is in the test set for a fold, the homology model does not inherit the template’s binding affinity label and, instead, is labeled as “unknown.” Experiments showed that it is important to account for this issue, as the prediction performance is deceptively high when we do not adjust for this issue.

## 4. DATASET

For our experiments in predicting kinase binding affinity, we use sequences in PFAM:Pkinase and PFAM:Pkinase\_Tyr multiple sequence alignments from release 25 of Pfam [24]. The PFAM:Pkinase and PFAM:Pkinase\_Tyr families contain all eukaryotic protein kinase domains covering the full kinase family tree. These are the same alignment files used in previous CCORPS experiments [3]. PDB structures are used that are referenced in the alignment files to obtain 1957 PDB substructures. We select the same 27 binding site residues as previous CCORPS experiments to define the substructures [3]. The binding affinity labels used are the same dataset from Karaman et al. [12] used in previous experiments as well. This dataset from Karaman et al. contains binding affinities for 317 kinases and 38 inhibitors [12].

In addition to PDB structures, we need to obtain homology models. We use homology models from the ModBase database [23] that contains models generated by the ModPipe pipeline [6] using Modeller [7]. Available homology models are collected for sequences indicated in the Pfam alignment files. The dataset of homology models contains over 200,000 homology models. However, to ensure accurate predictions, we need to select high quality models to be used in CCORPS. The dataset requires much filtering and cleaning to be used for our purposes. The first step is to ensure the model data is accurate with regards to the sequence alignment data. There are possible inconsistencies due to the versions of sequences and/or structural templates used. We check the residue numbering to ensure that all the data is consistent. Some models require adjustment of the residue numbering by an offset to match the Pfam sequence alignment. After ensuring consistency amongst our protein data, we have a series of filters for the homology models. The first filter is to make sure we select homology models that contain the 27 binding site residues. Some models may contain only a portion of the sequence data. In the next filter, we select high quality models based on various quality scores reported by ModBase. The ModBase database suggests particular thresholds for “reliable” models stating that at least one of five criteria should be satisfied [23]. We have decided to use the strictest selection by requiring models to satisfy all five criteria at suggested thresholds. For the final filter, we remove models that have a side chain centroid LRMSD distance from PDB:3HEC or PDB:2PL0 greater than 5 Å for the 27 residue binding site substructure. The two PDB structures selected are the same structures used as representative structures for binding site selection for the Pkinase and Pkinase\_Tyr families. Previous work in CCORPS selects these two structures because they contain a bound Imatinib ligand, which was chosen as a reference inhibitor [3]. After cleaning and filtering the homology models for our needs, we



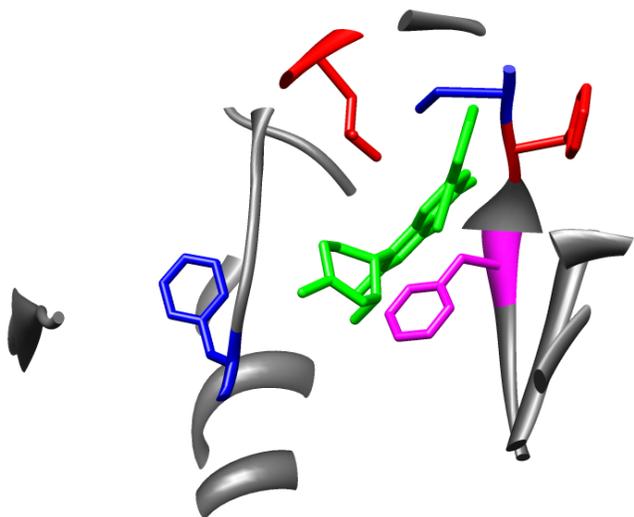
**Figure 1: The baseline experiment is at 0% homology models. We ran 5 experiments for each percentage interval, 5%, 10%, and 15%, selecting different random homology models. The average ROC AUC is calculated over 5 runs.**

get 5442 substructures from homology models. Even after all the filters, the number of homology model substructures is still more than twice the 1957 PDB structures.

## 5. RESULTS

We have many homology model substructures available for experiments. However, because the focus of our method is to use homology models to boost experimental data we do not want to use all homology model substructures. If we use more homology models than experimental structures, then we risk detecting biases in the homology models rather than actual structural patterns in kinases. Thus, we gradually add homology models in our experiments as a percentage of the number of PDB substructures. We ran experiments at 5% intervals from 0% to 20%. The experiment with 0% homology models serves as a baseline for comparison. We also ran experiments at 10% intervals from 30% to 100%, but no significant change in prediction performance was observed in this range relative to the experiments in the range of 0% to 20%. For that reason, we focus our attention on experiments in the range of 0% to 20%. Homology model substructures in each experiment are selected randomly from all filtered homology substructures. To ensure our results are consistent and not by chance, we perform five different runs per percent homology models to statistically confirm the trends that appear. It is important to note that random selection of homology models is likely not an optimal selection for prediction performance. Further work is needed to find possible heuristics for deterministic selection of homology models.

As with experiments in previous work in CCORPS, we evaluate the prediction performance of our experiments by finding the Receiver Operator Characteristic (ROC) curve for predictions [3]. The predictions are made using cross-fold validation, where each fold correspond to a non-redundant sequence cluster. The ROC curve measures the sensitivity at different specificities. The Area Under Curve (AUC) is calculated for each ROC curve to quantify the prediction



**Figure 2:** The binding site of CDK9/cyclinT1 in complex with Flavopiridol (PDB:3BLR) is shown. Flavopiridol inhibitor is highlighted in green. The three residues for subset 1805 (Phe-103, Phe-105, Leu-156) are highlighted in red and the residues for subset 1825 (Phe-103, Cys-106, Phe-168) are highlighted in blue. The subsets share residue Phe-103 which is colored purple.

performance for each label.

Some drugs have a baseline AUC value that is above 0.90 at 0% homology models, which is already quite high. For these drugs, prediction is already accurate and would thus be hard to improve on. In our experiments, we do not see any significant increase or decrease in prediction performance for these drugs as homology models are added. Thus, we focus our attention on drugs that have low prediction performance in the baseline, where there is room for improvement. We identified seven drugs with an AUC of less than 0.75 in the baseline experiment. These drugs are ABT-869, CI-1033, Flavopiridol, Gefitinib, Imatinib, SU-14813, and Sunitinib. As shown in Figure 1, the prediction performance for CI-1033, Flavopiridol, SU-14813, and Sunitinib improves significantly, and the prediction performance for ABT-869, Gefitinib, and Imatinib remains about the same. From the observed AUC values, we see that prediction performance either improves or remains about the same when we add just a few homology models. Experiments with additional homology models show little change in prediction performance indicating that using more than approximately 400 homology models does not provide any additional useful information for prediction.

Predictions are affected by homology models in two ways. First, the addition of homology models changes the clusterings which can indirectly affect HPCs if cluster memberships change for structures. Second, inheriting binding affinity labels from template structures has a more direct impact on HPCs. Due to our definition of HPCs, it only takes a single differently labeled structure to remove an HPC. While this can introduce error, indistinct HPCs that could be contributing poor votes can be removed. When tallying votes for a structure, all HPCs the structure falls in count as exactly one vote regardless of the size of the HPC. Spurious patterns

that can occur in sparse experimental data may be removed when homology models are added.

In order to better understand the effects of the addition of homology models in prediction, we look more closely at a specific example. Each experiment we run contains 2925 different subset clusterings, where 38 inhibitor binding affinity labels can be applied. We focus on the Flavopiridol inhibitor and subsets 1805 and 1825 for this example. Figure 2 shows the binding site for PDB:3BLR which contains the Flavopiridol inhibitor. The three residues for subset 1805 (Phe-103, Phe-105, Leu-156) and subset 1825 (Phe-103, Cys-106, Phe-168), as well as the Flavopiridol inhibitor, are highlighted in Figure 2 with different colors. If we look at the clusterings shown in Figure 3 for 0% and 5% homology models we can see how HPCs are affected in this example. Figure 3(a) and 3(b) shows an example of small, indistinct HPCs being removed in a clustering for subset 1805, as we add 5% homology models. We see two sparse HPC clusters in Figure 3(a) combine to become one non-HPC cluster in Figure 3(b). In addition to removal of weak HPCs, addition of homology models can also affect clusterings to reveal new HPCs. Figure 3(c) and 3(d) shows an example where the addition of 5% homology models results in the emergence of distinct true HPCs. We see clusters get broken up into smaller, more distinct clusters in subset 1825 when we add homology models, as Figure 3(d) has many more HPCs than Figure 3(c). In these ways, the addition of homology models affects the clustering landscape which, in turn, affects overall prediction performance.

## 6. CONCLUSION

We have shown that our method for integrating homology models with CCORPS improves prediction of kinase binding affinity. Tapping into the abundance of structural data in homology models, we are able to better identify structural features and make more accurate predictions from correlations with kinase binding affinity. We observed that the additional information from homology models outweighs any error that may have been introduced. Experiments showed no significant decrease in prediction performance for any drugs as we add homology models, while several drugs increased in prediction performance significantly.

There are several potential modifications that can be made to further analyze the use of homology models to predict binding affinity. One such modification is the selection of homology model substructures. In our experiments homology models are selected randomly. A heuristic can be implemented to select homology model substructures in a more intelligent way. For example, we can identify the non-redundant sequence clusters each homology model belongs to and then select homology models such that we cover all non-redundant sequence clusters.

Another possible modification is to apply labels to homology models in a more complex manner instead of inheriting template labels. It is clear from our experiments that using unknown labels for homology models does not have much impact on prediction performance, indicating that the labels applied are an important factor. Future work could involve labeling homology models based on docking experiments and simulations. Another modification in the prediction step is to use different purity thresholds in determining HPCs. Here, purity is defined as the ratio of points with the majority label to the total number of labeled points. We currently use a strict threshold requiring 100% purity.

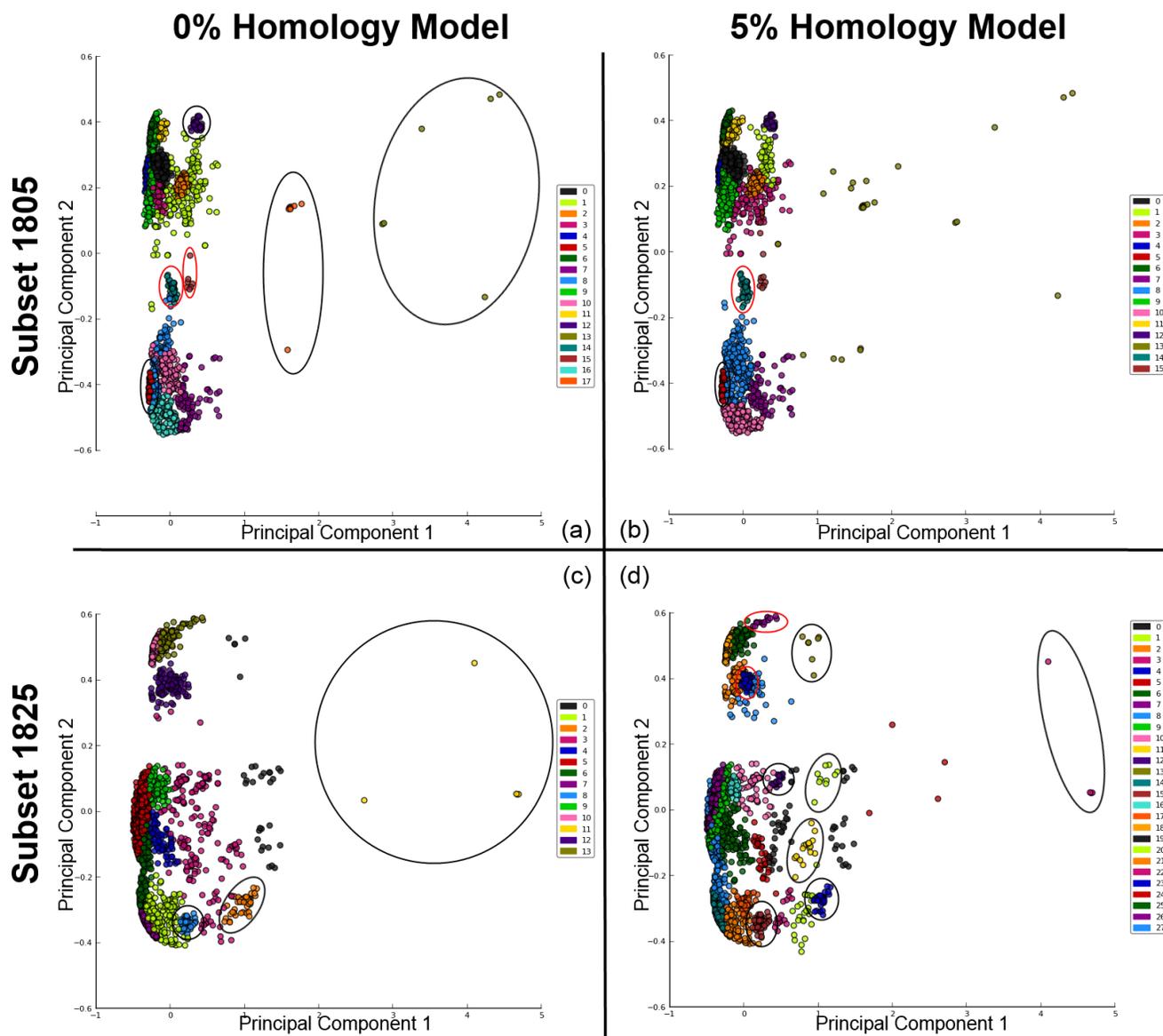


Figure 3: In each experiment there are a total of 2925 subsets clustering triplets of residues. Each plot above shows the structural and chemical variation for one such subset of 3 residues among all substructures. On the left the clusterings are shown for two subsets with only experimental structures, while on the right clusterings for the same subsets are shown with 5% homology models added. The three residues for subset 1805 are Phe-103, Phe-105, and Leu-156. The three residues for subset 1825 are Phe-103, Cys-106, and Phe-168. In each clustering plot, each point represents a single substructure. The points are colored based on cluster membership. The HPCs in each clustering plot are circled based on binding affinity labels for Flavopiridol. True HPCs (clusters predictive of binding Flavopiridol) are the red circles, and false HPCs (clusters predictive of *not* binding Flavopiridol) are the black circles. Changes in cluster memberships and HPCs can be seen for both subsets as 5% homology models are added.

Perhaps loosening the purity threshold could result in better predictions.

While we are specifically analyzing kinase binding affinity in this paper, the method has general input parameters. Any set of computationally predicted protein structures can be used. Homology models can be used from other databases such as the SWISS-MODEL Repository [13]. Other methods of predicting protein structure, such as ab initio modeling [2], can also be used to obtain protein structural data.

There are many avenues available for future work to improve on the predictive power of the method. Furthermore, the method can be expanded to other applications with other proteins and annotations beyond kinase binding affinity due to the generality of the method. There is great potential for the use of homology models for functional annotation of proteins.

## 7. ACKNOWLEDGMENTS

This work has been supported in part by NSF grant ABI-0960612, the John and Ann Doerr Fund for Computational Biomedicine at Rice University, and the Texas Higher Education Coordinating Board NHARP 01907. Equipment used to run the experiments presented in this paper is part of the Shared University Grid at Rice which is funded in part by NSF grant EIA-0216467, and a partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.

## 8. REFERENCES

- [1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, May 2000.
- [2] R. Bonneau and D. Baker. Ab initio protein structure prediction: progress and prospects. *Ann. rev. of biophys. and biomolec. structure*, 30(1):173–189, 2001.
- [3] D. H. Bryant, M. Moll, P. W. Finn, and L. E. Kavraki. Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome. *PLoS Computational Biology*, 9(6):e1003087, June 2013.
- [4] P. Cohen. Protein kinases—the major drug targets of the twenty-first century? *Nat Rev Drug Discov*, 1(4):309–15, Apr 2002.
- [5] P. Cohen and D. R. Alessi. Kinase drug discovery—what’s next in the field? *ACS Chem Biol*, 8(1):96–104, Jan 2013.
- [6] N. Eswar, B. John, N. Mirkovic, A. Fiser, V. A. Ilyin, U. Pieper, A. C. Stuart, M. A. Marti-Renom, M. S. Madhusudhan, B. Yerkovich, et al. Tools for comparative protein structure modeling and analysis. *Nucleic acids research*, 31(13):3375–3380, 2003.
- [7] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using Modeller. In *Curr Protoc Bioinformatics*, chapter 5. Wiley, Oct. 2006.
- [8] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca. MCLUST version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington, June 2012.
- [9] K. Ginalski. Comparative modeling for protein structure prediction. *Current opinion in structural biology*, 16(2):172–177, 2006.
- [10] B. Godshall and B. Chen. Improving accuracy in binding site comparison with homology modeling. In *IEEE Intl. Conf. on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 662–669, 2012.
- [11] D. Huang, T. Zhou, K. Lafleur, C. Nevado, and A. Cafisch. Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis. *Bioinformatics*, 26(2):198–204, Jan 2010.
- [12] M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, and P. P. Zarrinkar. A quantitative analysis of kinase inhibitor selectivity. *Nat Biotechnol*, 26(1):127–32, Jan 2008.
- [13] F. Kiefer, K. Arnold, M. Künzli, L. Bordoli, and T. Schwede. The SWISS-MODEL repository and associated resources. *Nucleic acids research*, 37(suppl 1):D387–D392, 2009.
- [14] S. L. Kinnings and R. M. Jackson. Binding site similarity analysis for the functional classification of the protein kinase family. *J Chem Inf Model*, 49(2):318–29, Feb 2009.
- [15] D. Kuhn, N. Weskamp, E. Hüllermeier, and G. Klebe. Functional classification of protein kinase binding sites using Cavbase. *ChemMedChem*, 2(10):1432–47, 2007.
- [16] T. Liu and R. B. Altman. Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding. *PLoS Comput Biol*, 7(12):e1002326, Dec 2011.
- [17] Y. Liu and N. S. Gray. Rational design of inhibitors that bind to inactive kinase conformations. *Nat Chem Biol*, 2(7):358–64, Jul 2006.
- [18] M. Magrane and Uniprot Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011:bar009, 2011.
- [19] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–34, Dec 2002.
- [20] M. J. McGregor. A pharmacophore map of small molecule protein kinase inhibitors. *J Chem Inf Model*, 47(6):2374–82, 2007.
- [21] F. Milletti and J. C. Hermann. Targeted kinase selectivity from kinase profiling data. *ACS Medicinal Chemistry Letters*, 3(5):383–386, 2012.
- [22] M. Moll, D. Bryant, and L. Kavraki. The labelhash algorithm for substructure matching. *BMC Bioinformatics*, 11(1):555, 2010.
- [23] U. Pieper, B. M. Webb, D. T. Barkan, D. Schneidman-Duhovny, A. Schlessinger, H. Braberg, Z. Yang, E. C. Meng, E. F. Pettersen, C. C. Huang, et al. Modbase, a database of annotated comparative protein structure models, and associated resources. *Nucleic acids research*, 39(suppl 1):D465–D474, 2011.
- [24] M. Punta, P. C. Coghill, R. Y. Eberhardt, J. Mistry,

- J. Tate, C. Bournnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al. The Pfam protein families database. *Nucleic acids research*, 40(D1):D290–D301, 2012.
- [25] W. M. Rockey, A. H. Elcock, et al. Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? *Current protein & peptide science*, 7(5):437, 2006.
- [26] A. Roy, A. Kucukural, and Y. Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–738, 2010.
- [27] E. C. Webb. *Enzyme nomenclature*. Academic Press, San Diego, CA, 1992.