Applications of Cheminformatics for the Analysis of Proteolysis Targeting Chimeras and the Development of Natural Product Computational Target Fishing Models

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

in the Graduate School of The Ohio State University

By

Nicholas T. Cockroft

Graduate Program in Pharmaceutical Sciences

The Ohio State University

2019

Dissertation Committee

James R. Fuchs, Advisor

Xioalin Cheng, Co-Advisor

Karl A. Werbovetz

Lara E. Sucheston-Campbell

Copyrighted by

Nicholas Thady Cockroft

2019

Abstract

The use of data-driven methods and machine learning has become increasingly pervasive in many industries, including drug discovery and design, as computing power and large amounts of data become increasingly available. In an effort to efficiently leverage this data, cheminformatics has emerged as a data-driven, interdisciplinary field that focuses on storing, accessing, and applying chemical information. Cheminformatics methods and tools facilitate the management and analysis of large annotated chemical datasets that would be difficult or impossible to do manually. A famous application of leveraging large amounts of chemical data was performed by Christopher A. Lipinski in 1997. Lipinski analyzed a large set of bioavailable synthetic drug molecules and identified trends in their molecular properties, which has since been referred to as the "Lipinski's Rule of 5". While these rules are far from absolute, Lipinski's analysis demonstrates the utility of leveraging large amounts of chemical data to gain important insights. This thesis describes the application of cheminformatics methods to tackle two very different research problems: 1) the analysis and binding of a class of protein degraders called proteolysis targeting chimeras (PROTACs) and 2) the development of a target fishing application for the prediction of mechanism of action of natural products.

PROTACs are a novel class of small molecule therapeutics that are garnering significant interest. Unlike traditional small molecule therapeutics, PROTACs simultaneously bind to both their protein target and an E3 ligase to induce degradation. The requirement to simultaneously bind two proteins necessitates a high molecular weight as PROTACs must contain two unique binding moieties that are connected by a linker. As a result, PROTAC molecules are expected to lie outside of the traditional drug-like chemical space described by Lipinski. To gain a better understanding of the physicochemical properties of PROTACs currently in development, the patent literature was searched and PROTAC compounds targeting either the Von Hippel-Lindau (VHL) or cereblon (CRNB) ligases were retrieved. This analysis identified that the physicochemical properties of PROTACs were indeed different from those of drug-like small molecules. However, the importance of each property for activity and permeability cannot yet be addressed without additional annotated biological endpoints. While the physicochemical properties of a PROTAC compound are expected to be important for its pharmacokinetics, the formation of a ternary complex is crucial for its pharmacodynamics. Using the currently available crystallographic data of ternary complexes with resolved PROTACs, a method for prediction of the ternary complex structure was developed and benchmarked. The results of this method were promising with ternary structures predicted correctly for up to 60% of the final predicted complexes. However, the identification of the correct complexes from among the incorrect complexes *a priori* was shown to be a difficult task.

Another class of small molecule therapeutics which do not adhere to traditional drug-like properties is natural products. Natural products have been a tremendous source of new drugs over the past three decades with unaltered natural products and natural product derivatives making up over one-third of FDA approved small molecule drugs.

These natural products have made up a substantial portion of first-in-class drugs identified through phenotypic screening methods. A limitation of phenotypic screening methods is a lack of understanding of the target and molecular mechanism of action, which is desirable for the progression of a chemical entity to the clinic. Cheminformatics methods can be applied to aid in the identification of the molecular mechanism of action of small molecules in a process termed computational target fishing. The current methods for computational target fishing have been trained and tested on datasets containing exclusively synthetic compounds. Based on their inherent structural differences, the relative ability and accuracy of a model trained on synthetic data to predict targets for natural products remains unknown. To address this, a natural product benchmark set containing 5,589 compound-target pairs for 1,943 unique compounds and 1,023 unique targets was collected by cross-referencing 20 publicly available natural product databases with the bioactivity database ChEMBL. A dataset of synthetic compounds from ChEMBL containing 107,190 compound-target pairs for 88,728 unique compounds and 1,907 unique targets was used to train k-nearest neighbors (KNN), random forest (RF), and multi-layer perceptron (MLP) models. Additionally, a model stacking approach was also investigated, which uses logistic regression as a meta-classifier to combine the individual model predictions. A model stacking approach using KNN and RF as the base classifiers showed the best performance on the natural product benchmark set with an area under the receiver operating characteristic (AUROC) score of 0.94 and a Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) score of 0.73. A similarly performing and more computationally efficient model stacking approach using KNN as the base classifier was deployed as a web application, called STarFish, and has been made available for use to aid in the target identification of natural products.

Dedication

To my mother and father for their love, endless support, and encouragement.

Acknowledgments

First, I wish to express my deepest gratitude to my advisor, Dr. James R. Fuchs, for accepting me into his lab and allowing me to continue pursuing my interest in computational medicinal chemistry despite his lab’s focus on synthetic medicinal chemistry. Under his mentorship, I have greatly improved my skills as a researcher and presenter. His personality and demeanor cultivate a unique dynamic in his lab, and I will forever treasure the memories of my time there. I would also like to thank my co-advisor, Dr. Xiaolin Cheng, whose guidance and kind words over the last couple of years has significantly improved my confidence, knowledge, and skills in computational chemistry. Additionally, I would like to thank my former advisor, Dr. Chenglong Li, for his mentorship at the beginning of my graduate career and my dissertation committee members, Dr. Karl A. Werbovetz and Dr. Lara E. Sucheston-Campbell, for their advice and support throughout my graduate studies.

I would be remiss to not mention my former and current lab members whose camaraderie was fundamentally important for persevering through the numerous challenges that are encountered when pursuing a Ph.D. I would like to thank my former lab members, Dr. Guqin Shi and Mohammad Rezaei, for their assistance and discussions at the beginning of my graduate career. Additionally, I would like to thank Cheng lab members, Dr. Kevin Chan and Dr. Sijin Wu, for similar assistance and discussions towards

the end of my graduate career. I would also like to thank my current lab members, Andrew Huntsman, Robert “Bo” Tokarski, Tyler Wilson, Brittney Mize, and Dr. Daniel Adu-Ampratwum, for their friendship and support. I will miss spending time with them in the office and the frequent banter and jokes we shared. I would like to make special mention of Andrew, Bo, and Tyler who I have spent the most time with. Andrew, I have enjoyed our morning chats over coffee about our shared pop-culture interests. Bo, I appreciate your self-deprecating humor and sociability. You also made our shared T.A. appointments much more enjoyable. Tyler, I am grateful for your friendship and that you were my constant lunch companion. I will miss our daily lunchtime scientific, political, and/or philosophical discussion.

Finally, I would like to thank my family, Greg, Nola, and Amanda Cockroft, for their unwavering love and support. Additionally, I would like to thank my fiancée, Kelly Usakoski, who has always been willing to listen to me talk about my research, help me practice for presentations, and proof-read my work. I love you all very much!

Vita

2014	B.A. Chemistry, Kalamazoo College
2014 – 2015	Chemistry-Biology Interface Program T32 AT00753 Fellow
2015 – 2018	Graduate Teaching Associate, Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, The Ohio State University
2017	M.S. Pharmaceutical Sciences, The Ohio State University
2018 – 2019	Graduate Research Associate, Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, The Ohio State University

Publications

1. **Nicholas T. Cockroft**, Xiaolin Cheng, James R. Fuchs. STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products, *J. Chem. Inf. Model.*, **2019** (manuscript submitted, Manuscript ID: ci-2019-00489s)

2. Tyler A. Wilson, Pratibha C. Koneru, Stephanie V. Rebensburg, Jared J. Lindenberger, Matthew J. Kobe, **Nicholas T. Cockroft**, Daniel Adu-Ampratwum, Ross C. Larue, Mamuka Kvaratskhelia, James R. Fuchs. An Isoquinoline Scaffold as a Novel Class of Allosteric HIV-1 Integrase Inhibitors, *ACS Med. Chem. Lett.*, **2019**, *10 (2), pp 215–220*

Fields of Study

Major Field: Pharmaceutical Sciences

Specialized in Computational Medicinal Chemistry

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Dissertation Organization

This dissertation focuses on the development or application of cheminformatics methods for drug discovery and design. Two drug discovery and design projects are described herein.

Chapter 1 gives background information for the methods used in each project. First, a brief overview of cheminformatics methods, including how chemical structures are represented computationally, is provided. Next, an introduction to machine learning, including model training techniques and machine learning algorithms, is given. Finally, structure-based drug-design methods are described with a focus on molecular docking.

Chapter 2 describes the analysis of proteolysis targeting chimeras (PROTACs) and the rational design of cyclin dependent kinase 9 (CDK9) degraders. The first half of the chapter details the analysis of physicochemical properties of PROTAC molecules extracted from the patent literature. The physicochemical properties calculated for the patent PROTACs are analyzed and compared to traditional drug-like properties. In the second half of the chapter, the rational design of CDK9 targeting PROTAC compounds is described. The physicochemical properties of the designed and synthesized CDK9 PROTACs are calculated and compared with the collected patent PROTAC compounds. Furthermore, a method to predict the PROTAC mediated ternary complex is described, benchmarked, and applied to a potent developed CDK9 PROTAC.

Chapter 3 describes the development of a stacked ensemble target fishing method, STarFish, and its performance on a collected set of natural product compounds. This chapter describes how the synthetic compound dataset and the natural product benchmark set were collected and used for model training and validation. Additionally, the machine learning algorithms and model stacking approach used are described. Performance results for each machine learning algorithm and model stacking combination employed during cross-validation and on the natural product benchmark are presented and discussed. The impact of training dataset size, protein target diversity, and training compound similarity is also presented and discussed. Finally, an example application of using the deployed STarFish web application for the target identification of a natural product is shown.

1.2 Cheminformatics

Cheminformatics is an interdisciplinary field in computational molecular science which combines knowledge from many fields including Physics, Chemistry, Biology, Biochemistry, Mathematics, Statistics, and Computer Science.[1] The scope of cheminformatics is not well defined due to the interdisciplinary nature of the field and often has much overlap with the field of bioinformatics, especially for drug discovery and design problems.[2] An important aspect of cheminformatics is the collection, annotation, and storage of chemical information into databases. The use of such databases is fundamental to modern day chemical research. For synthetic chemists, the accessibility to a large amount of chemical information allows for the rapid retrieval of chemical reactions and experimental procedures that can be used to plan synthetic routes. For computational chemists, the information contained in these databases can be retrieved and analyzed to develop predictive tools, such as the prediction of physical properties, chemical reactivity, and biological activity.[3] Due to the broad nature of the field, not all of the methods, applications, or aspects of cheminformatics will be discussed here. Instead, concepts relevant to the work discussed here will be described, such as how chemical structures are represented in a machine-readable way.

1.2.1 Representation of Chemical Compounds

A 2D drawing of chemical structures (e.g. a typical structure generated in ChemDraw, MarvinSketch, or ChemSketch) is the common way chemical compounds are represented when discussing concepts with colleagues or preparing figures for scientific manuscripts, posters, and grant applications. Despite the frequent use of 2D images to represent chemical information in day-to-day discussions, they are not the preferred way to represent and store chemical information computationally. Instead a line notion representing a linear string of characters is generally preferred. The Simplified Molecular Input Line Entry System (SMILES) is one of the most common line notations for representing chemical compounds and can be used for: accessing databases as a key, sharing chemical information, entering chemical data, and artificial intelligence or expert system languages. SMILES are a popular way to represent chemical information as they are compact, human readable, and can be easily converted into a 2D image of a compound.[4] An example of a 2D chemical structure and it's corresponding SMILES string is shown in Figure 1.1.

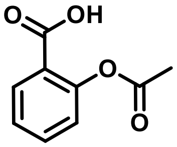
	Aspirin
2D Representation:	
SMILES String:	CC(=O)Oc1ccccc1C(=O)O
Chemical Fingerprint:	11101101101101011110001101001100

Figure 1.1 Compound Representation. An example using Aspirin of a few different ways chemical compound information can be represented. The 2D chemical structure, SMILES string, and chemical fingerprint bitstring are shown.

Another way to represent chemical information is with chemical fingerprints. Chemical fingerprints are an abstract representation commonly used to perform database screening and similarity calculations. A chemical fingerprint is a Boolean array where each bit in the array represents chemical patterns. A fingerprint is generated by examining a molecule and exhaustively identifying every pattern in the molecule up to a given pathlength limit. For example, with a pathlength limit of 3 a pattern would be identified for: every atom, every atom with its nearest neighbor and joining bond, every atom with all atoms and bonds up to 2 bonds away, and every atom with all atoms and bonds up to 3 bonds away. With a larger value for the pathlength limit, patterns covering an increasingly larger number of bonds are identified. The identified patterns are ultimately used to assign bits in the Boolean array. Each pattern is used as a seed for a pseudo-random number generated, a hashing function, which sets bits on the array. Due to the large number of possible patterns, different patterns can set bits in common resulting in a collision. The bits in a fingerprint can be thought of as being shared by many unknown patterns.[5] The frequency of collisions can be reduced by increasing the number of bits in the fingerprint, but as the number of bits in a fingerprint increases so does the computational cost of working with it. An example of a chemical fingerprint and the patterns corresponding to each bit are shown in Figure 1.2. It should be noted that while it appears there are no collisions for this fingerprint this is not the case as only the first pattern for each bit is shown. For example, there are a total of 5 patterns which set the 6th bit.

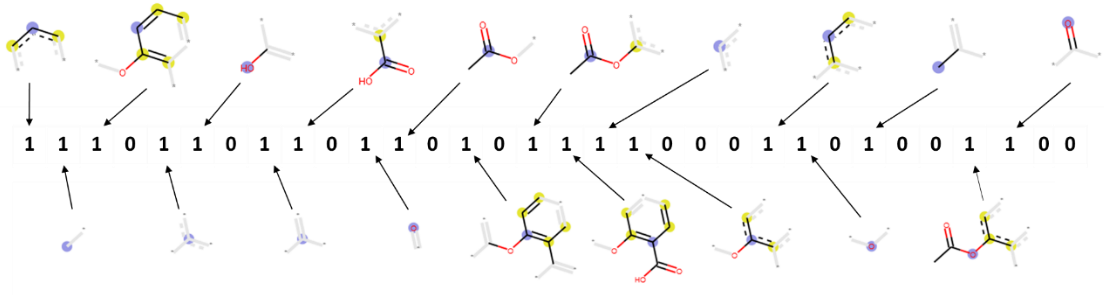


Figure 1.2 Fingerprint Bits and Corresponding Chemical Patterns. Continuing with the Aspirin example, the chemical fingerprint shown earlier is further explained. A 32-bit fingerprint calculated with a pathlength limit of 2 is shown. Each bit in the fingerprint corresponds to a chemical pattern. If the pattern is present, the corresponding bit is turned on and a 1 is shown. If a bit contains 0, it lacks the pattern. For each pattern shown, the central atom for the pattern is highlighted with a blue circle, aromatic atoms are highlighted with a yellow sphere, bonds connecting pattern atoms to non-pattern atoms are shown in light grey, bonds between atoms in the pattern are given a corresponding color based on the atoms they connect (carbon=black, oxygen=red, nitrogen=blue, etc.), and a black arrow shows which bit is turned on by the pattern. Only the first pattern for each bit is shown.

1.2.2 Chemical Similarity

The calculation of chemical similarity is a useful way to assess relationships between compounds. Broadly, a similarity calculation involves the measurement of distance or similarity between sets of molecular descriptors. As there are many choices for descriptors and ways to measure distance between them, it is necessary to select the most appropriate ones for a given application. Molecular descriptors generally fit into the broad categories of physicochemical properties, 2D properties, and 3D properties. The previously described fingerprints are an example of a 2D property. Examples of distance and similarity metrics include: Manhattan distance, Euclidean distance, Tanimoto coefficient, and Dice coefficient.[6] The Tanimoto coefficient is generally preferred for calculating the similarity of molecules which are described using fingerprints.[7] The formula for calculating Tanimoto similarity of chemical fingerprints is shown in eq. 1.1, where the similarity between molecules *A* and *B*, $S_{A,B}$, is calculated from the number of bits in molecule *A, a*, the number of bits in molecule ***B,*** *b*, and the number of bits in common between molecules *A* and *B*, *c*. An example performing this calculation is shown in Figure 1.3.

$$S_{A,B} = \frac{c}{a+b-c} \tag{1.1}$$

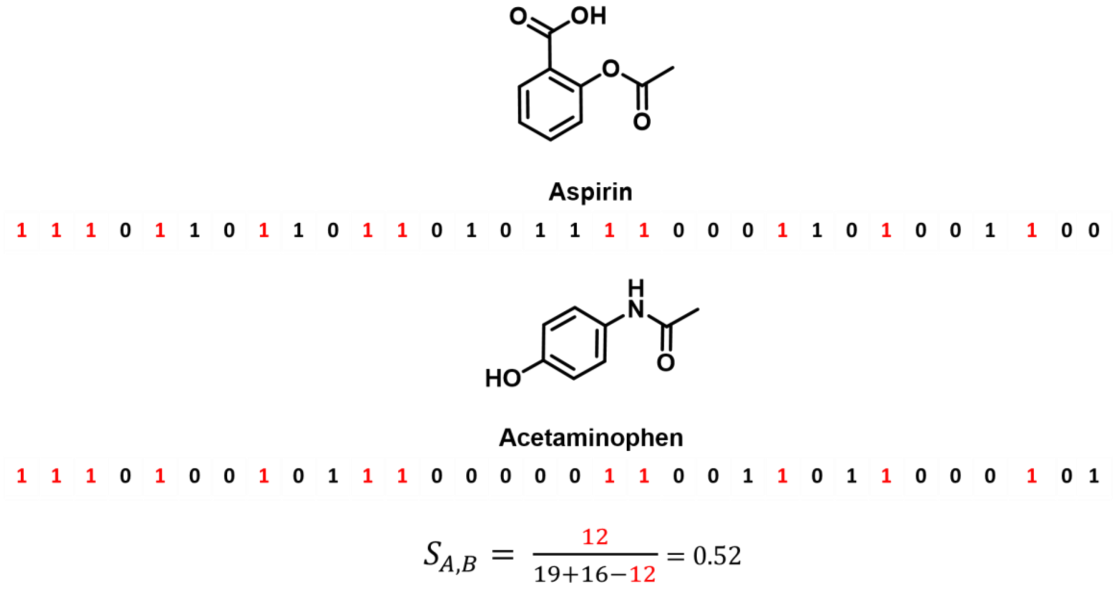


Figure 1.3 Calculation of Tanimoto Similarity Coefficient. An example of Tanimoto similarity calculated for Aspirin and Acetominophen is shown. Each molecule is described using a 32-bit fingerprint. It should be noted that the similarity coefficient calculated here is unexpectedly high due the low number of bits used, which resulted in a greater number collisions than would be present in fingerprints using a higher number of bits.

1.2.3 Physicochemical Properties

Another useful way to describe molecules is with physicochemical properties. While chemical fingerprints are commonly used in screening, calculation and examination of physicochemical properties can be a useful way to establish trends for a set of related molecules. In medicinal chemistry, the most famous example of using physicochemical properties to establish trends is Lipinski’s Rule of 5. Lipinski’s rules were generated by examining the physicochemical properties of a subset of compounds in World Drug Index, which had proceeded to Phase II clinical trials. From this analysis, Lipinski concluded that poor permeation and poor absorption are more likely when: the molecule weight is greater than 500, the LogP is greater than 5, the number of hydrogen bond donors is greater than 5, or the number of hydrogen acceptors is greater than 10.[8] Many of the properties in Lipinski’s Rule of 5 can be calculated fairly easily due to their simplicity. For example, molecular weight is an easy descriptor to calculate as it is just the sum of the individual atomic weights. Calculations of hydrogen bond acceptors and donors are similarly trivial as it just requires the identification and count of particular atom types. While the technical details of how molecular weight, the number of hydrogen bond donors, and the number of hydrogen bond acceptors are calculated may differ between software packages due to how each prefers to represent molecules, each package would ultimately be expected to return the same values.

On the other hand, values like LogP, which is a measured physical property, are not calculated theoretically as simply. LogP is a measure of lipophilicity and specifically is the logarithm of the partition coefficient between octanol and water. The partition coefficient is the ratio between the concentrations of a neutral compound in organic and aqueous solutions at equilibrium. Therefore, a method to theoretically estimate this value is required. These methods fit into two broad categories: substructure-based and property-based. Substructure-based methods cut a molecule down into fragments or atoms and sum the individual contributions of each. Property-based methods rely on calculation of other molecular descriptors, such as molecular volume and atomic charges, which are then combined. In general, each method will have to be benchmarked against known experimental LogP values, and in some cases, weight each contribution differently to improve the predicted values.[9] Therefore, unlike the calculation of molecular weight, the number of hydrogen bond donors, and the number of hydrogen bond acceptors, different software packages will give different values for LogP depending on the method used to calculate it.

A physicochemical property closely related to LogP is LogD. LogD is nearly identical to LogP except it also considers charge. This further complicates the calculation as it requires the estimation of the pK_a for each ionizable species in a molecule. LogP and LogD values are equivalent when there are no ionizable groups in a molecule. However, if a molecule carries a formal charge, the LogD will be significantly lower than the LogP value as it’s aqueous solubility will significantly increase. Therefore, when considering the absorption of a drug, a LogD value will be more informative than the LogP value if the molecule is expected to be charged at a physiologically relevant pH.[10]

1.3 Machine Learning

Machine learning is a subfield of artificial intelligence that has been growing rapidly in recent years due to advances in computing power and the increased availability of large amounts of data. Techniques using machine learning are concerned with building a model to make predictions.[11] These machine learning techniques broadly fit into two categories: supervised or unsupervised learning. Supervised learning requires target output values in the training data and trained models can be used to predict output values for new input data points. Therefore, the datasets for supervised learning require sets of input features along with their corresponding output values. Unsupervised learning uses unlabeled data and learns underlying patterns from the input features. Unlike supervised learning, the datasets for unsupervised learning do not require defined output values.[12] As with many fields, the use of machine learning techniques and software has been adopted by pharmaceutical sciences. Despite a slower uptake compared to fields like the consumer service industry, all stages of drug discovery and development have begun to use machine learning to improve discovery and decision making.[13] The work described here uses a variety of supervised machine learning techniques to make predictions about the bioactivity of small molecules. An overview of how machine learning models are trained, an introduction to the type of learning technique used, and a description of the specific machine learning algorithms used are provided in the subsequent sections.

1.3.1 Cross-Validation

Despite the utility and predictive power of machine learning models, there is significant concern that prospective use of a trained model will perform much worse than indicated during training. Overfitting is the term used to describe the case when a model performs well on the training data but predicts poorly on new data. The ability of machine learning algorithms to learn rules that underlie the data can be detrimental if the learned rules are specific to only that dataset and not generalizable. To combat overfitting, the total amount of data available is split into a training set and a test set. Machine learning models are then built using only the training data split. The trained models are then used to predict the output values of the test data split. A comparison of these predicted values to the known output values for the test set gives a more realistic estimate of model performance as this data had not been seen during training. However, splitting the original data into training and test sets has drawbacks. First, it reduces the number of examples the model can learn from. Second, the performance on the test set may be highly variable depending on how examples are split between the training and test sets.[14]

A related approach called k-fold cross-validation addresses some of these drawbacks. In k-fold cross-validation, the total amount of available data is split into k sets of approximately equal size. One of these subsets is used as the test set while the remaining k-1 subsets are used for training. This collection of subsets is referred to as cross-validation fold. A total of k cross-validation folds are generated, each using a different subset of the data as the test set. A diagram demonstrating how a dataset is split for k-fold cross-validation is shown in Figure 1.4. Since the performance for each fold is averaged, the variance in the performance metric is reduced for k-fold cross-validation compared to the previously described train-test split validation procedure.[14] However, this variance reduction comes at an increased computational expense as a total of k models need to be trained, one for each cross-validation fold.

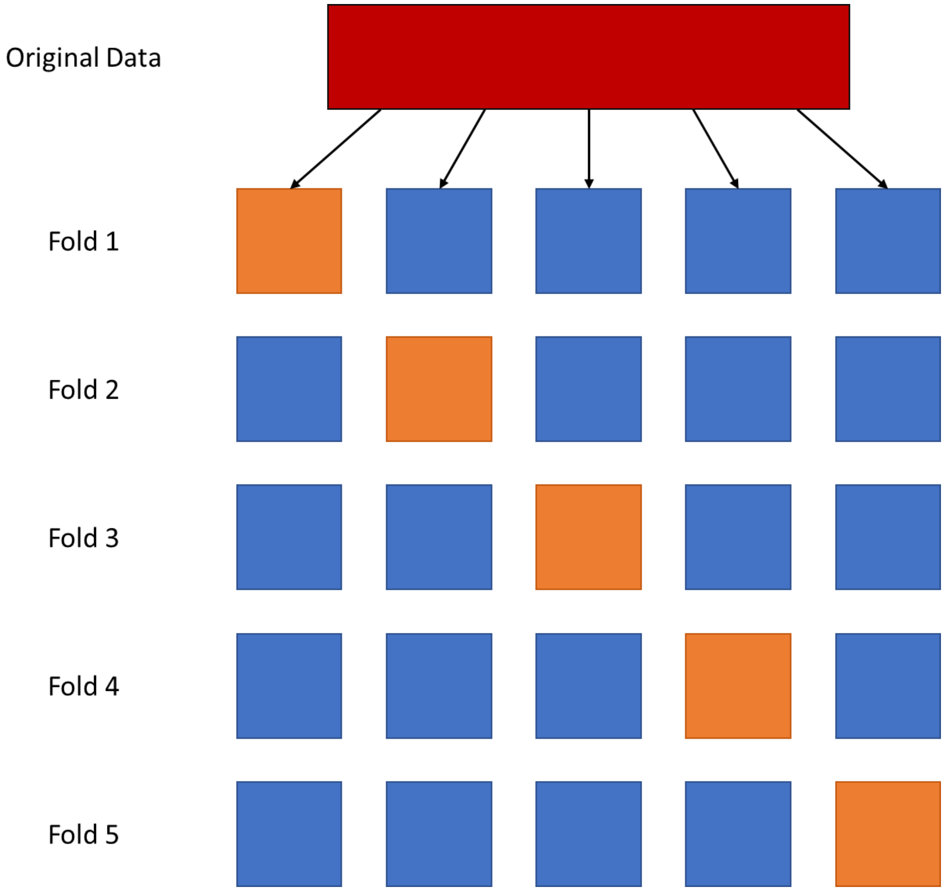


Figure 1.4 k-Fold Cross-Validation Diagram. A diagram which demonstrates how a data set is split in k-fold cross-validation for k=5. The original dataset (red) is split into 5 different subsets (squares). For each fold, a different subset is used as the test set (orange) while the remaining subsets are used as the training set (blue).

1.3.2 Classification

Different supervised learning techniques are used depending on the type of problem to be solved. Problems concerned with the prediction of continuous numerical output values are termed regression problems, while those concerned with the prediction of nominal output values are termed classification problems.[15] Housing prices are a traditional example of a regression problem. An example of a regression problem relating to cheminformatics is the prediction of an IC_{50} value for a small molecule against a protein target. The identification of emails as either “spam” or “not spam” is a traditional example of a classification problem. An example of a cheminformatics related classification problem is the identification of a small molecule as “toxic” or “non-toxic”. In Chapter 3, a machine learning model is trained to predict the protein target of a small molecule and is formulated as a classification problem with protein targets as the classes.

Classification problems may contain multiple classes which may or may not be mutually exclusive. Problems which contain only a single class are termed binary classification problems. Those which contain more than one mutually exclusive class are termed multi-class classification problems. Lastly, classification problems which contain multiple, non-exclusive classes are termed multi-label classification problems.[16] While all machine learning classification algorithms are applicable to binary classification problems, not all machine learning algorithms are applicable to multi-class and multi-label problems.

If the use of a binary classification algorithm for a non-applicable multi-class or multi-label problem is desired, a problem transformation strategy can be used to make it applicable. An example of a problem transformation strategy is binary relevance, which can also be referred to as one-vs-rest. However, binary relevance is commonly used to describe a one-vs-rest type strategy for multi-label problems, while one-vs-rest can also be applied to multi-class problems. The idea of a one-vs-rest type strategy is to decompose a multi-label or multi-class type problem into multiple binary classification problems.[17] Instead of a single trained model that outputs predicted class labels of each example, a model is trained for each individual label and the predictions from each model are aggregated. Therefore, binary classification algorithms can be applied to multi-class or multi-label problems by using a problem transformation strategy.

1.3.3 k-Nearest Neighbors

The k-nearest neighbors (KNN) algorithm is a type of instanced-based learning and therefore stores instances of the training to be referenced during prediction. KNN can be used for either classification or regression problems. When a trained KNN model is passed a query data point, it computes the distance between the query point and the training instances to determine the closest k points. The distance computations can either be done in a brute force fashion or using tree-based approaches to limit the number of distance calculations required. For KNN classification, the classes for which the k closest points belong to are used to assign the class of the query point. Probability of class membership is the simple average of the class label count over the nearest k points. These probabilities can also be weighted by the distance of each training instance to the query point. A diagram illustrating the KNN algorithm is shown in Figure 1.5.

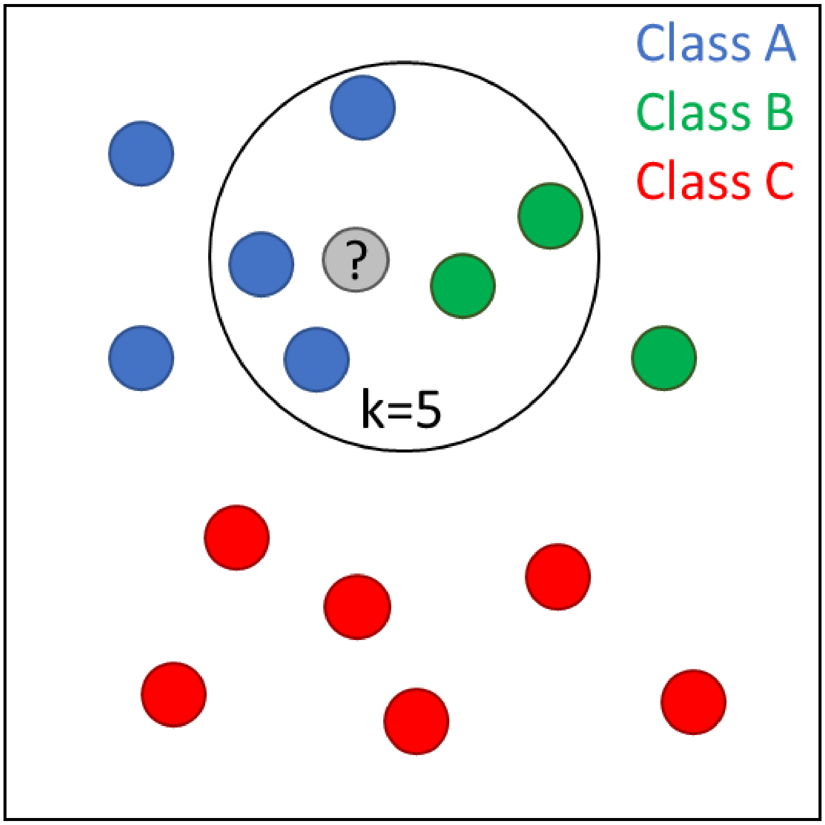


Figure 1.5 k-Nearest Neighbors Diagram. A simple example of the k-Nearest Neighbors classifier for classification of the query point in grey. A black circle encomposses the nearest k points. In this case, there is a probability of 60% of class A membership, 40% of class B membership, and 0% of class C membership. Therefore, the point would be assigned to class A.

1.3.4 Multi-Layer Perceptron

A multi-layer perceptron (MLP) is a feedforward artificial neural network that consists of at least three layers: an input layer, a hidden layer, and an output layer. A schematic of this architecture is shown in Figure 1.6. Each layer consists of a set of neurons. In the input layer, the number of neurons is set to the number of features for a record in the training data. When used for classification, the number of neurons in the output layer corresponds to the number of class labels. The number of hidden layers and the number of neurons in each is a tunable hyperparameter of the model. Each neuron in a hidden layer takes the values from each neuron in the previous layer and combines them by a weighted linear summation. This summed value is passed to a non-linear activation function that yields the output from the hidden neuron. Neurons in the output layer take values from the last hidden layer and transform them into the output probabilities for each label. During model training, the output probabilities are assessed by a loss function. In scikit-learn, the cross-entropy loss function is used. The formula for cross-entropy loss is shown in eq. 1.2.

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \qquad 1.2$$

Where the loss, $L(\theta)$, is a measure of how much the predicted probability, p_i, diverges from the true label, y_i, on average for the n number of i labels for the model parameters θ. The MLP is trained iteratively and at the end of each iteration the aforementioned weights between layers are adjusted by calculating the gradient of the loss function in a process called backpropagation. Training continues until the loss doesn’t change by a specified amount between several iterations or the maximum number of specified iterations is reached.

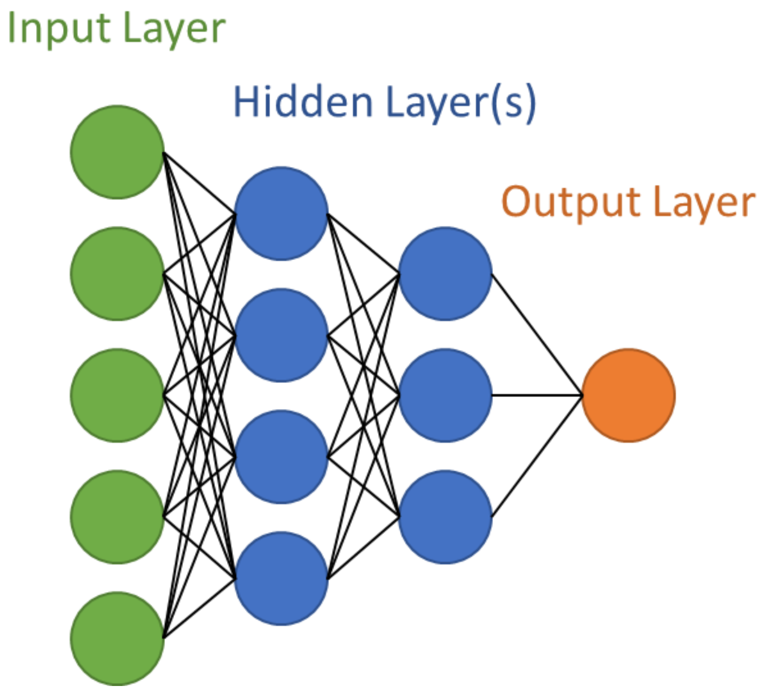
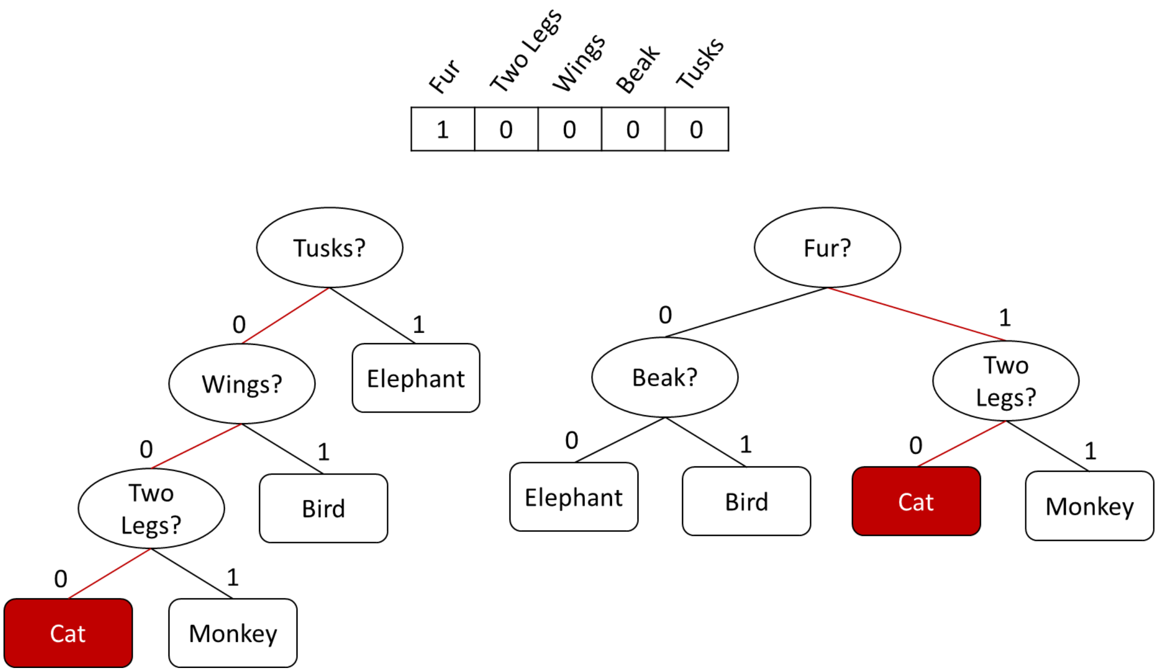


Figure 1.6 Multi-Layer Perceptron Schematic. The architecture of a simple feedforward artificial neural network with two hidden layers is shown. Values are passed from one layer to the next from left to right as represented by the black lines, which connect the nuerons depicted as circles. Each neuron contains an activation function, which combines the values passed to it from each neuron in the previous layer. Each of these can be weighted differently, and those weights are adjusted during model training.

1.3.5 Random Forest

Random forests are an ensemble of decision trees that can be used for either classification or regression. A decision tree is built through a top-down approach where a feature in the training data is selected for its ability to best split the training data. In the case of classification, the best split is the one that yields the purest daughter nodes. Purity is the fraction of each class present in each daughter node. For example, the least pure split would be for a daughter node to contain equivalent portions of each class, while the purest split would be a daughter node with a single class label. Pure nodes are leaf nodes as no further splitting is required. A fully-grown tree contains all pure leaf nodes, but large and complex trees are prone to overfitting. To combat overfitting, the random forest algorithm generates many small trees, a “forest”, and uses a bootstrapped sample of the training data with a random subset of features to split a node; instead of exhaustively checking for the best possible split among all features. The ensemble of predictions made by the individual trees are averaged to give a final prediction. The most important hyperparameters for random forest are the number of trees in the forest and the number of features tried at each split. A simple example of a “trained” forest being used for the prediction of a given input feature vector’s class is shown in Figure 1.7. The logic shown in the simple example can be applied to chemical fingerprints, where specific bits corresponding to particular chemical features may strongly indicate class membership.



Fur	Two Legs	Wings	Beak	Tusks
1	0	0	0	0

Figure 1.7 Random Forest Example. A simple example for the classification of an animal into one of four classes (Cat, Monkey, Bird, or Elephant) given a vector of 5 features (Fur, Two Legs, Wings, Beak, Tusks). Two example decision trees in the "forest" are shown and each returns a class assignment based on the information in the input feature vector.

1.3.6 *Logistic Regression*

Despite the name, logistic regression is used for classification and can be applied to binary, multinomial, and ordinal classification problems. Logistic regression is a linear method, however, the output of the linear combination of features, shown in eq.1.3, is bounded between 0 and 1 as shown in eq. 1.4.

$$l = c + \beta_1 x_1 + \beta_2 x_2 + \cdots + \ \beta_n x_n \qquad 1.3$$

$$p = \frac{1}{1 + e^{-l}} \qquad 1.4$$

The log odds, *l*, is the linear combination of features, *x*, with their weights, *β*, and intercept, *c*, is transformed to probability, *p*, by the logistic function as shown in eq 1.4. For the implementation used here from scikit-learn, the weights, *β*, are adjusted during training through the minimization of the cost function shown in eq. 1.5.

$$\min_{w,c} w^T w + \ C \sum_{i=1}^{n} log(1 + \ e^{-y_i(X^T w \, + \, c)}) \qquad 1.5$$

Where $X^T w$ corresponds to sum of all *βx*, and y_i is a value in the set {-1, 1} for negative and positive observations of a label respectively. Model complexity is penalized by L2 regularization, $w^T w$, and tuned by *C*, which is the inverse of regularization strength. An example logistic regression plot for binary classification problem is shown in Figure 1.8.

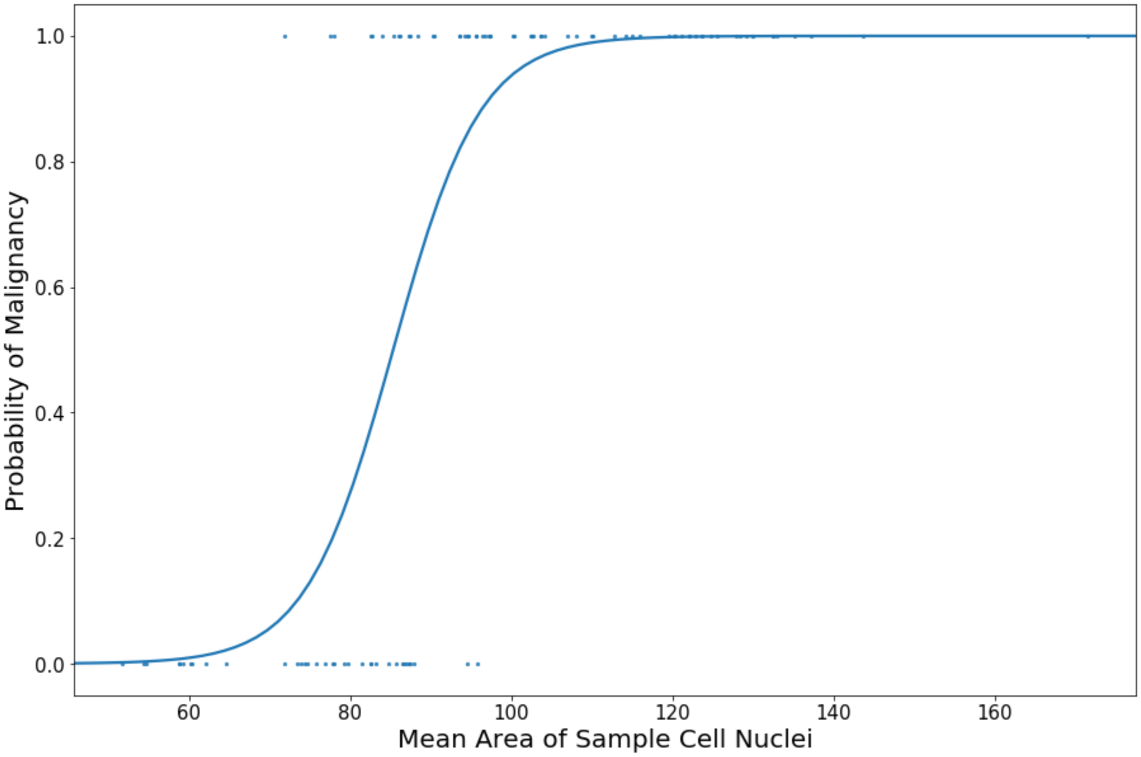


Figure 1.8 Logistic Regression Plot. An example plot of logistic regression applied to a binary classification problem with a single input feature. Toy data was obtained from the UCI ML Breast Cancer Wisconsin (Diagnostic) dataset in scikit-learn.

1.4 Structure-Based Drug Design Methods

Structure-based drug design leverages knowledge of a target protein structure for bioactive small molecule discovery and design. These methods are complementary to the cheminformatics methods discussed previously, which are commonly referred to as ligand-based drug design when they are applied to drug discovery and design projects. Structure-based methods are used when high quality structural data, protein crystal structures, are available for a target protein of interest. This structural data can be used to calculate interaction energies between a small-molecule and the target protein structure. In general, structure-based drug design methods have an emphasis on physical interactions with a protein structure whereas ligand-based methods focus on comparisons between chemical structures. Some examples of structure-based drug design methods include: comparative modeling, binding site prediction, pharmacophore modeling, molecular docking, molecular dynamics, and free energy calculations.[18,19] Many of these methods are complementary and can be used sequentially. For example, in the absence of a co-crystal structure, molecular docking can be used to place a small molecule in a binding site as a starting point for molecular dynamics simulations. Additionally, the trajectories from a molecular dynamics simulation can be subsequently analyzed to compute the free energy of the small molecule ligand binding to the protein target. In the following sections, a brief introduction is given for the structure-based drug design methods relevant to this work. Specifically, molecular docking and a related method, protein-protein docking, are discussed.

1.4.1 Molecular Docking

Molecular docking predicts the pose and free energy of a small-molecule binding to a protein target. It is one of the most frequently used structure-based drug design methods due to its speed and relative ease of use. The process of molecular docking occurs in two-steps: conformational sampling and scoring. The algorithm used for conformational sampling and the function used for scoring are specific to the docking software used. In general, conformational sampling algorithms fall into the broad categories of systematic or stochastic search methods. Systematic methods specifically vary the structural parameters, which influence conformation, whereas stochastic methods vary these parameters randomly. Regardless of the conformational search algorithm utilized, a balance between exhaustive sampling and computational expense is necessary. The other step in molecular docking is the scoring function, which falls into three major categories: force-field-based, empirical, and knowledge-based. Force-field-based scoring functions sum atomic intermolecular and intramolecular interactions to estimate the binding energy. Empirical scoring functions sum weighted, pre-defined terms that describe physical events in the formation of the ligand-protein complex. Knowledge-based scoring functions sum pairwise energy potentials, which are constructed and weighted according to the frequency of specified atom-atom distance occurrences in a training dataset. The scoring functions are ultimately used to evaluate small molecule conformations produced by the conformational sampling algorithm. A cyclical process of sampling and scoring is performed until an energy minimum is reached. After convergence, the small molecule conformation and predicted binding energy are returned as the docking result.[20,21]

1.4.2 Protein-Protein Docking

Protein-protein docking predicts the pose and free energy of a protein binding to another protein. Protein-protein interactions are very important for molecular recognition and signaling.[22] Therefore, the ability to model and predict such interactions is desirable for bioinformatics related fields, including drug discovery and design. While molecular docking typically allows for small-molecule flexibility during docking, proteins are usually treated as a rigid body due to the computational expense of including additional flexibility. Therefore, protein-protein docking uses rigid-body proteins, which are typically further transformed into simplified, coarse-grained representations. Conformations of these rigid, simplified proteins are sampled and evaluated for steric complementarity.[23] This process is commonly performed using a geometry-based algorithm, which matches 3D digitized representations of each protein's molecular surface using a fast Fourier transform.[24] After the coarse-grained matching of molecular surfaces, structural resolution is returned, and the predicted protein-protein conformation may be scored. Similar to the scoring functions described for molecular docking, a variety of protein-protein scoring functions exist that predict binding energy using either physical interactions or knowledge-based functions. However, correct conformations are difficult to predict for proteins, which undergo significant conformational changes upon complex formation[25]. Additionally, there is poor correlation between predicted and known protein-protein binding affinities.[26] Nonetheless, protein-protein docking is an active area of research, and the re-ranking of predictions using machine learning-based scoring functions, the use of experimental information as restraints, and the use of template-based docking can further improve predictive power.[27]

Chapter 2. Analysis of Proteolysis Targeting Chimeras and Rational Design of Cyclin Dependent Kinase 9 Degraders

2.1 Abstract

Proteolysis targeting chimeras (PROTACs) are an exciting new class of small molecule therapeutics with a novel mechanism of action compared to traditional small molecule inhibitors. PROTACs induce the targeted degradation of a protein target by taking advantage of a cell’s protein recycling machinery. By necessity, current PROTACs are extremely large, non-drug-like molecules, which raises concerns regarding their ability to transition from a chemical probe to a clinical therapeutic. However, traditional knowledge of drug-like properties may not be applicable to PROTACs since they function as catalytic degraders instead of stoichiometric inhibitors. To investigate the physicochemical properties of PROTACs currently in development, PROTAC compounds targeting either the Von Hippel-Lindau (VHL) or cereblon (CRBN) ligase were collected from the patent literature and analyzed. Analysis of the collection of patent literature compounds shows that PROTAC compounds indeed lie beyond traditional drug-like chemical space. Therefore, a new paradigm of drug-like physicochemical properties for PROTACs will likely be needed if currently developed compounds are able to be administered orally. The novel mechanism of action for PROTACs, which necessitates their large size and thereby non-drug-like properties, relies on the formation of a ternary

complex containing an E3 ligase, a PROTAC, and the target protein. A method to computationally predict this complex was developed and benchmarked on currently available crystallographic data of ternary complexes containing a resolved PROTAC molecule. For the VHL-MZ1-BRD4, CRBN-dBET6-BRD4, and CRBN-dBET23-BRD4 ternary complexes, 15%, 42%, and 60% of the predicted complexes were correct respectively. While the successful prediction rates are encouraging, definitive identification of correct predictions from the incorrect ones using ZRANK was unsuccessful.

2.2 Introduction

Proteolysis targeting chimeras (PROTACs) are an emerging class of small molecule therapeutics. While traditional small molecule inhibitors, such as competitive inhibitors, bind to a protein active site and block activity, PROTAC molecules promote degradation of a protein target. A comparison of a traditional competitive inhibitor's activity compared to a PROTAC is illustrated in Figure 2.1. PROTACs are bi-functional molecules that contain both an E3 ligase and a target protein binding moiety that are connected by a linker. The bi-functional nature of these molecules allows the PROTAC to bind the E3 ligase and the target protein simultaneously. This simultaneous binding induces targeted degradation as it brings the E3 ligase into proximity with the target protein. When in close proximity, the E3 ligase can ubiquitinate the target protein thereby marking it for degradation.[28]

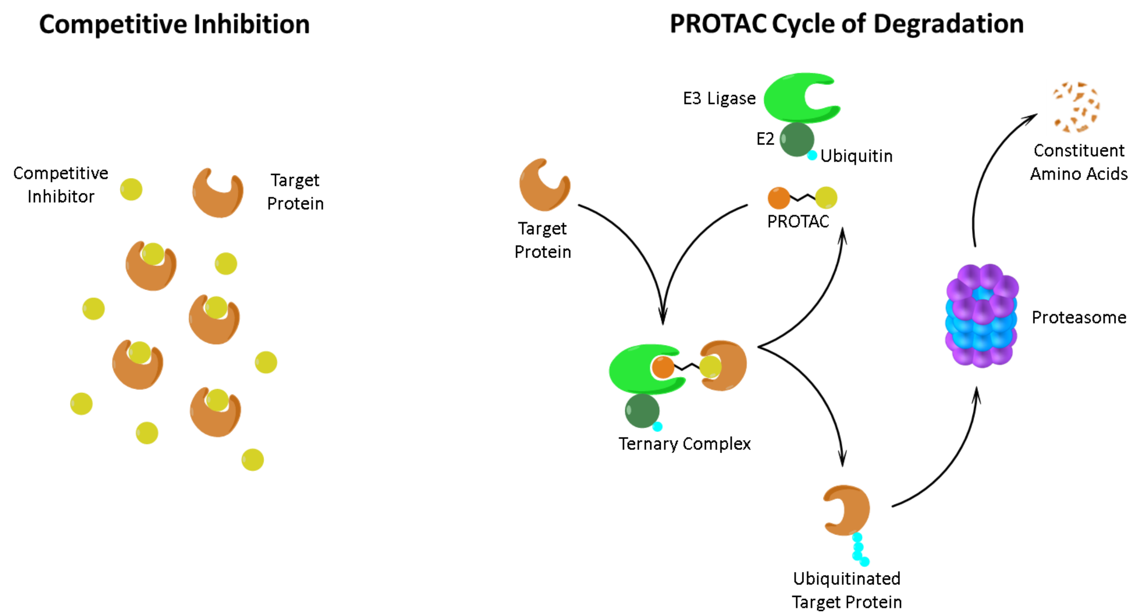


Figure 2.1 Comparison of Competitive Inhibition and PROTAC Mechanism of Action. Competitive inhibition requires a small molecule inhibitor to bind and reside in the binding site of a target protein to block activity. Comparatively, a PROTAC molecule can catalyze the degradation of many target protein copies.

By necessity, PROTACs are very large molecules compared to traditional small molecule drugs. PROTAC compounds are about twice as large as small molecule inhibitors since they contain two compounds linked together. Furthermore, the linkers popularly used to connect the E3 ligase binding moiety and the target protein binding moiety are very long and flexible chains.[29] Orally bioavailable synthetic drug molecules typically have physicochemical properties, such as molecule weight and number of rotatable bonds, which obey the Lipinski's Rule of 5 (Ro5)[8] and Veber's rule[30] guidelines.[31] Therefore, there are valid concerns regarding whether such large and flexible molecules can be effectively developed into an oral therapeutic drug.

However, not all orally bioavailable drugs lie in Ro5 chemical space. For example, a miniperspective by DeGoey et al. from AbbVie in 2018 examined beyond rule of 5 (bRo5) compounds in the AbbVie drug and compound collection and identified the physicochemical properties that were most import for the oral bioavailability of bRo5 compounds.[32] The authors identified that a LogD close to 3 along with a limited number of aromatic rings and rotatable bonds were correlated with increased probability of oral bioavailability even with Ro5 violations.

Despite a seemingly bleak outlook for developing an orally bioavailable PROTAC compound, it is important to note that PROTACs work very differently than the small molecule drugs for which the traditional Ro5 guidelines were devised. For traditional small molecule inhibitors, a sufficiently high concentration of the small molecule is required to stoichiometrically occupy the binding site of a target protein, but such a requirement does not apply to PROTACs.[28] Due to a PROTAC's catalytic mechanism of action, the induced degradation is sub-stoichiometric.[33] Therefore, there is potential for an orally administered PROTAC molecule to have the desired efficacy despite the expected poor permeability for such large and non-drug-like molecule.

Due to the novel mechanism of action of PROTACs, there is an absence of directly applicable knowledge about what physicochemical properties are important for their oral bioavailability. In order to begin investigation into what properties may be important for successful development, PROTAC compounds were extracted from the patent literature and their physicochemical properties were examined. PROTAC structures with E3 ligase binding moieties that targeted either the Von Hippel-Lindau (VHL) or cereblon (CRBN) ligases were collected from the patent literature database SureChEMBL[34]. For each patent PROTAC, the molecular weight, LogP, LogD at pH 7.4, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, topological polar surface area, number of aromatic rings, and the AbbVie multi-parametric score (AB-MPS) for bRo5 compounds were calculated. These collected properties permitted comparisons between VHL and CRBN targeting PROTACs and established a frame of reference of the physicochemical properties of PROTACs currently in development. A similar and complementary analysis was recently reported by Edmondson, Yang, and Fallan at AstraZeneca in 2019, which examined the physicochemical properties of 38 PROTAC compounds published in the academic literature.[35] A major difference between that analysis and the one described here is that significantly more compounds are examined here due to the patent literature containing hundreds instead of tens of PROTAC molecules.

However, the PROTAC compounds in patent literature lack the bioactivity annotations reported for the compounds in the academic literature.

The analysis of the patent literature PROTACs was subsequently used as a frame of reference during the initiation of a project for the development of PROTACs targeting cyclin-dependent kinase 9 (CDK9) in the Fuchs lab. While ideal properties of PROTAC molecules are unknown, comparison to the patent literature compounds facilitated the identification of properties which significantly deviated from the other PROTAC molecules currently in development. Properties of proposed and synthesized CDK9 PROTACs, which were found to greatly deviate from the patent literature PROTACs, could be further examined and given consideration as to whether the deviation was of concern. Additionally, the tracking of physicochemical properties during CDK9 PROTAC development lays the foundation for a better understanding of which properties are ultimately important for success.

The development of a small molecule targeting CDK9 as an anti-cancer therapeutic has been an active area of research. Despite the discovery of many potent small molecule CDK9 inhibitors, their clinical use has been stymied due to a lack of selectivity.[36] A PROTAC approach offers a potential solution for this lack of selectivity. Early reports of CDK9 PROTACs indicate that selectivity can be gained by taking a small molecule CDK9 inhibitor and adapting it into a PROTAC.[37] This selectivity is thought to arise through favorable interactions that can occur at the PROTAC mediated protein-protein interface in the ternary complex.

The formation of the ternary complex is a fundamental aspect of the PROTAC mechanism of action. Formation of the ternary complex and cooperative binding contributions between the E3 ligase and target protein have been shown to influence degradation efficacy[38], target selectivity[39], and degradation rate[40]. Visual analysis of the first ternary complex crystal structure (PDB: 5T35) enabled optimization of the PROTAC linker, which resulted in improved binding affinity, ternary complex stability, and target selectivity.[41] Due the utility offered by visualization of the ternary complex, the ability to computationally generate reliable ternary structure models would be of great assistance to PROTAC discovery and design.

However, the development of reliable computational ternary structure models is a non-trivial task. First, the prediction of a protein-protein interface is an extremely difficult problem and generally relies on protein-protein docking. Success of protein-protein docking is highly variable with good results being more frequently obtained when experimental data can be leveraged as a pose filter or docking restraint.[42] Furthermore, accurate scoring of small-molecule binding energy is still a difficult task[43], and the protein-protein interfaces considered here have many more interactions to consider compared to ligand-protein systems. Additionally, there are currently only six PROTAC ternary crystal structures published and only half have sufficient resolution to observe the PROTAC compound[41,44], which makes the benchmarking and assessment of any predictive method difficult.

As a first step toward addressing these challenges a method was developed for generating ternary complex predictions. This method was benchmarked on the PROTAC ternary crystal structures with a resolved PROTAC compound and was subsequently applied to the prediction of the ternary complex for a novel CDK9 PROTAC. To aid future iterations and improvements on this method, the source code has been made freely available at: https://github.com/ntcockroft/gen_ternary and in Appendix B. Datasets and Code.

2.3 Methods

2.3.1 Dataset

PROTAC molecules were obtained from the patent literature by querying the SureChEMBL database.[34] SureChEMBL is a large publicly available database that contains 17 million compounds extracted from full text, images and attachments of 14 million patent documents. To obtain PROTAC molecules targeting cereblon from this database, a substructure search was performed using the maximum common substructure of the immunomodulatory imide drugs (IMiDs), thalidomide and pomalidomide. This substructure search needs to be modified to return the desired PROTAC molecules and not derivatives and analogues of the IMiDs. This was done by adding the keyword "PROTAC" to the search query and filtering to only include the results with a molecular weight in the range of 650 to 3000. The lower bound was initially set at 500, however, approximately 11% of results appeared to be intermediates. To restrict results to final molecules, the lower bound was increased to 650. The upper bound was set to an arbitrarily high value to ultimately include all molecules with a molecular weight greater than 650. This query yielded 983 unique PROTAC molecules targeting the cereblon E3 ligase, which appeared to be final molecules. To obtain VHL PROTACS, this query was repeated using the hydroxyproline moiety containing small molecule that targets the VHL E3 ligase. This query yielded 1,287 unique PROTAC molecules targeting the VHL E3 ligase.

2.3.2 *Calculation and Representation of Physicochemical Properties*

The SMILES obtained from the SureChEMBL queries were used to calculate the physicochemical properties of interest. The software package, ACD/Percepta[45], was used to calculate the physicochemical properties: molecular weight, LogP, LogD, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, topological polar surface area, and number of aromatic rings. The AbbVie multi-parametric score (AB-MPS)[32] was calculated from the calculated LogD, number of aromatic rings (NAR), and number of rotatable bonds (NROT) as shown in eq 2.1.

$$AB\text{-}MPS = abs(LogD - 3) + NAR + NROT \qquad 2.1$$

The distributions of each property were visualized using the Python library Seaborn.[46] The distributions of physicochemical properties containing integer values were visualized as histograms, while those containing values with decimals were visualized using kernel density estimation plots. Visualization of the distributions using histograms is sensitive to the number and width of the bins chosen. This is not a problem for physicochemical properties containing integer values, as a single bin can be used for each value. Additionally, those values span a relatively narrow range, which allows each value in the range to be easily represented. The physicochemical properties containing decimal values can be better represented using kernel density estimation plots. For kernel density estimation, a smooth function is placed at the center of each data point instead of counting the number of points that fall into each bin.[47] The contributions of each smooth function (kernel) are summed to yield the final curve as given by eq. 2.2.

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2.2}$$

The kernel density estimation function *f(x)* for a number, *n*, of points, x_i, is obtained by summing the kernel, *K*, at each point. The extent to which each kernel contributes is scaled by its bandwidth, *h*. Scott's normal reference rule[48] was used to determine the optimal bandwidth and is shown in eq. 2.3, where *σ* is the sample standard deviation.

$$h = \frac{3.5\sigma}{n^{1/3}} \tag{2.3}$$

While the kernel can be any smooth, peaked, normalized function, the Gaussian kernel is a popular choice, and is the kernel used here as shown in eq. 2.4.

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.4}$$

2.3.3 Property Comparisons with Z-Scores

To put the calculated properties of designed and synthesized PROTAC molecules into the broader context of the patent compound property distributions the calculated values were converted to z-scores. A z-score is simply the number of standard deviations away from the mean that a data point is and is calculated as shown in eq 2.5.

$$z = \frac{x_i - \bar{x}}{\sigma} \qquad 2.5$$

The z-score, *z*, is the difference between a given value, x_i, and the sample mean, $\bar{x}$, divided by the sample standard deviation, σ. Z-scores can give information about the area under the standard normal curve, however, such an interpretation is not appropriate for the empirical sample distributions of physicochemical properties described here as they are not described by a standard normal curve. In this context, the z-scores give a quick assessment of how the properties of designed PROTACs compare to one another and to the collected patent literature compounds.

2.3.4 Generation of CDK9 Ternary Complex Structure Predictions

Crystal structures for cereblon (PDB: 4TZ4) and CDK9 (PDB: 4BCG) were downloaded from the Protein Data Bank.[49] The relevant chains from each protein structure were extracted, which included “Chain C” from the cereblon structure and all chains present in the CDK9 structure. Missing loops and side-chains were filled using MODELLER[50] using the UCSF Chimera[51] GUI. 100 models were generated for each structure, and the top model was selected according to the DOPE-HR score. Final loop-filled models were inspected for any issues and residues which contained errors were manually fixed using Maestro Academic from Schrödinger Suite 2017-2.[52] The lenalidomide ligand that was removed from the cereblon structure during loop filling was replaced and the AT7519 ligand from a CDK2 crystal structure (PDB: 2VU3) was inserted into the loop filled CDK9 structure. These structures were then prepared for protein-protein docking with ZDOCK 3.0.2.[53] The cereblon structure was prepared as the "receptor" and the CD9 structure as the "ligand". ZDOCK was run with default sampling and produced 2000 structures. Prior to generation of these structures from the ZDOCK output file, the input structures had hydrogen atoms added to them using Maestro Academic. In the ZDOCK output file, the filenames of the original input structures were replaced with the filename of the newly hydrogenated structures. This was done so the protein-protein docking results produced could eventually be scored by ZRANK[54], which requires the hydrogen atoms that are purposefully not used in ZDOCK. The hydrogenated ZDOCK output structures were filtered with a python script using the PyMOL 2.3.0b API.[55] Only docked protein-protein complexes that had a center of mass distance between the ligands less than a specified distance and at least 100 Å^2 of protein hydrophobic solvent-accessible surface area buried were kept. The ligand distance was selected on a case-by-case basis and was based on the approximate length in angstroms of the fully elongated PROTAC structure of interest. Hydrophobic solvent-accessible surface was approximated by calculating the solvent exposed surface area of hydrophobic residues for each protein structure. The difference between the total non-complexed protein surface area and complexed protein surface area was used to calculate hydrophobic surface burial. Ligands were extracted from the protein-protein complexes that passed these filters and used as positional restraints when generating possible PROTAC conformers with a Python script using the RDKit[56] 2019.3.2 and Pybel[57] for OpenBabel 2.4.1 APIs. Positional restraints were selected by the identification of common substructure atoms shared between the PROTAC molecules and binding site ligands. Complexes for which a conformer could be successfully generated were scored using ZRANK.

2.3.5 Benchmarking of Ternary Complex Structure Predictions

All currently available crystal structures of PROTAC ternary complexes for which the PROTAC compound was resolved were obtained. This includes the VHL-MZ1-BRD4 complex (PDB: 5T35), the CRBN-dBET6- BRD4 complex (PDB: 6BOY), and the CRBN-dBET23-BRD4 complex (PDB: 6BN7). The linker region in each PROTAC was removed so that the only the corresponding ligands were present in the ligase and target protein pockets. Ternary structures were predicted for each complex following the procedure detailed in Section 2.3.4 for the CDK9 ternary complex predictions. Predicted ternary complex structures which had a Cα RMSD $\leq$ 10 Å to the known crystal structure were deemed successful predictions. The PDB module in Biopython was used for calculation of Cα RMSD.[58] Final predicted complex structures were locally minimized using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm implemented in OpenMM 7.3.1.[59] Prior to complex minimization, PROTAC molecules were locally optimized for 1000 steps using Pybel with the Merck Molecular Forcefield 94 (MMF94). The optimized molecules were parameterized using the general forcefield

SMIRNOFF99Frosst with the Open Force Field Tool Kit in Python.[60] The Cα RMSD to the known crystal structure and ZRANK scores were calculated for both the minimized and non-minimized complexes for comparison.

2.3.6 Acknowledgement of Experimental Contributors

The PROTAC compounds described herein were synthesized by Andrew C. Huntsman and Robert J. Tokarski II from the James R. Fuchs group in the Division of Medicinal Chemistry and Pharmacognosy at The Ohio State University College of Pharmacy. Biological evaluation of the PROTAC compounds was performed by Bridget Carmichael from the John C. Byrd group at The Ohio State University Comprehensive Cancer Center.

2.4 Results and Discussion

2.4.1 Comparison of CRBN and VHL PROTACs Physicochemical Properties

Traditional medicinal chemistry small molecule drug design principles may not apply to PROTAC molecules due to their inherently high molecular weight and novel mechanism of action. As the interest in PROTAC strategies increase, the knowledge of what properties are important for the development a successful PROTAC molecule becomes paramount. As a first step in pursuit of this knowledge, the physicochemical properties of cereblon and VHL patent PROTACs were examined to give context about the physicochemical properties associated with PROTAC molecules and also compared to explore any notable differences in properties.

The VHL targeting PROTACs are a great deal larger than the cereblon targeting PROTACs (Figure 2.2). The cereblon targeting PROTACS have an average molecular weight of 813.20 g/mol, while for VHL they have an average weight of 1007.81 g/mol for an absolute difference of 194.61 g/mol. This difference is most likely explained by the size difference of the E3 ligase targeting ligands. The IMiD maximum common substructure has a molecular weight of 258.23 g/mol compared to the VHL targeting ligand's molecular weight of 430.56 g/mol. The VHL targeting ligand is larger by 172.33 g/mol, which closely matches the mean difference of the two datasets. For both sets the molecular weight of the patent PROTACs lie in a fairly narrow range. For example, the interquartile region spans from 770.90 to 839.89 g/mol and 1053.75 to 1162.77 g/mol for the cereblon and VHL sets respectively. Thus, the development of a PROTAC appears to increase the molecular weight by about 500 to 600 g/mol compared to the E3 ligase targeting ligand. This increase in molecular weight is consistent with what would be expected from the attachment of a linker and small molecule drug to the E3 ligase targeting ligand.

.

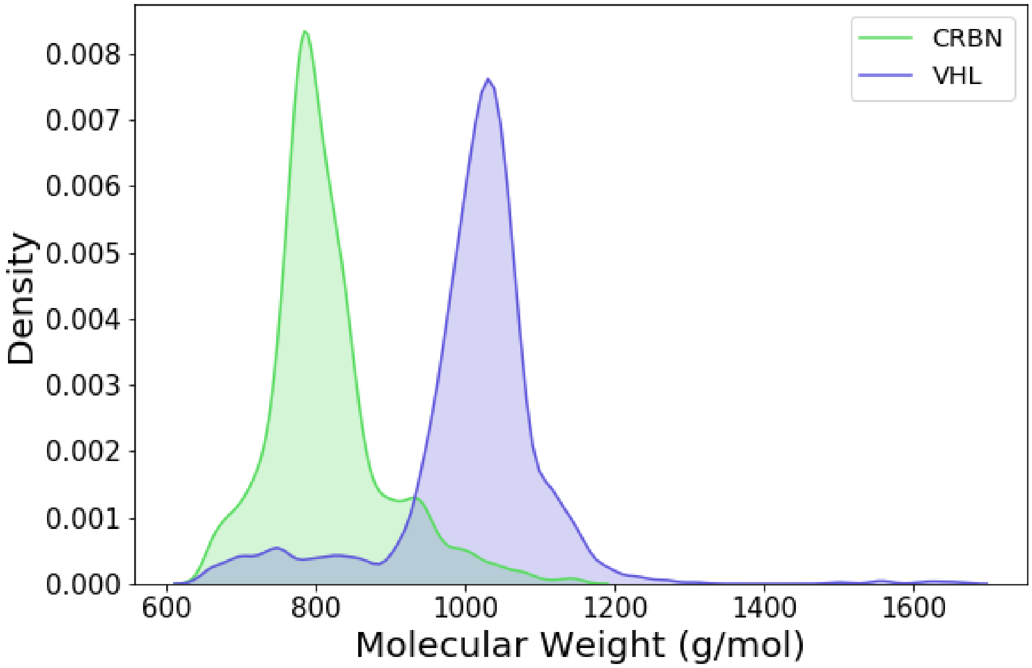


Figure 2.2 Comparison of Patent PROTAC Molecular Weight. The kernel density estimated distribution of molecular weight for patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

The VHL patent PROTACs also have a large topological polar surface area (TPSA) compared to the cereblon PROTACs (Figure 2.3). The cereblon targeting PROTACs have an average TPSA of 172.96 Å^2 compared to an average TPSA of 239.09 Å^2 for the VHL targeting PROTACs. As observed for the molecular weight, this difference most likely arises from the differences in properties between the E3 ligase targeting ligands. The IMiD maximum common substructure has a TPSA of 83.55 Å^2 compared to the VHL targeting ligand's TPSA of 136.79 Å^2. The VHL targeting ligand is larger by 53.24 Å^2, which closely matches the mean difference of the two datasets. The interquartile range of both distributions is also quite narrow and spans from 154.21 to 187.89 Å^2 and 223.25 to 255.59 Å^2 for the cereblon and VHL sets respectively. Overall, the trends observed for TPSA were very similar to the trends observed for molecular weight.

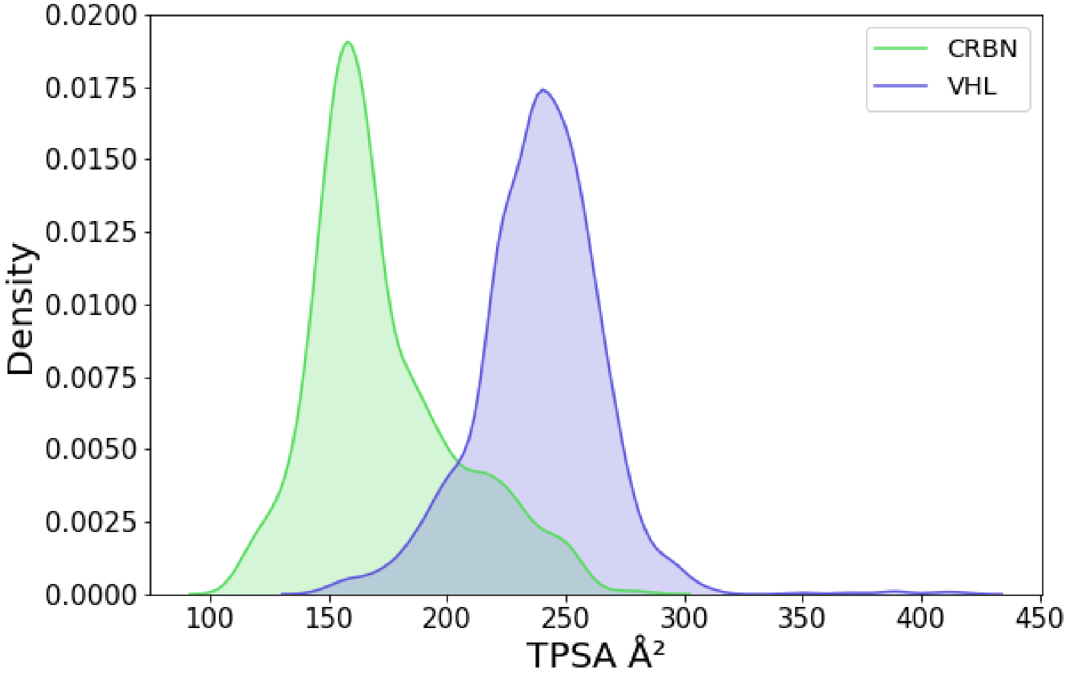


Figure 2.3 Comparison of Patent PROTAC TPSA. The kernel density estimated distribution of topological polar surface area for patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

Despite a larger TPSA, the VHL targeting PROTACs are still more lipophilic than the cereblon targeting PROTACs according to calculated LogP (Figure 2.4) and LogD (Figure 2.5) values. The cereblon targeting PROTACs have an average LogP of 4.38 and LogD of 3.54, while the VHL set has an average LogP of 5.52 and LogD 5.31. Consistent with the previous observations, the difference in LogP closely matches the difference in LogP between the two E3 ligase targeting ligands. The LogP of the VHL target ligand is greater by 0.8 compared to the common IMiD substructure. However, the difference in LogD cannot be explained with the same logic. Firstly, the common IMiD substructure has a greater calculated LogD value than the VHL targeting ligand. However, the calculated LogD value for the VHL targeting ligand is misleading as it contains an amino group that is commonly used as the synthetic handle to form an amide bond to the linker in the final PROTAC compound. Therefore, the contribution to the calculated LogD should be equivalent to the contribution of the calculated LogP value. In general, the LogD distribution closely matches the LogD distribution for the VHL set. The increased relative difference of LogD appears to be due to the shift in the distribution of the cereblon set towards lower LogD values. This suggests that the IMiDs have been linked to more inhibitors that are ionizable at pH 7.4 or have had more ionizable groups included in the linker compared to the VHL targeting ligand.

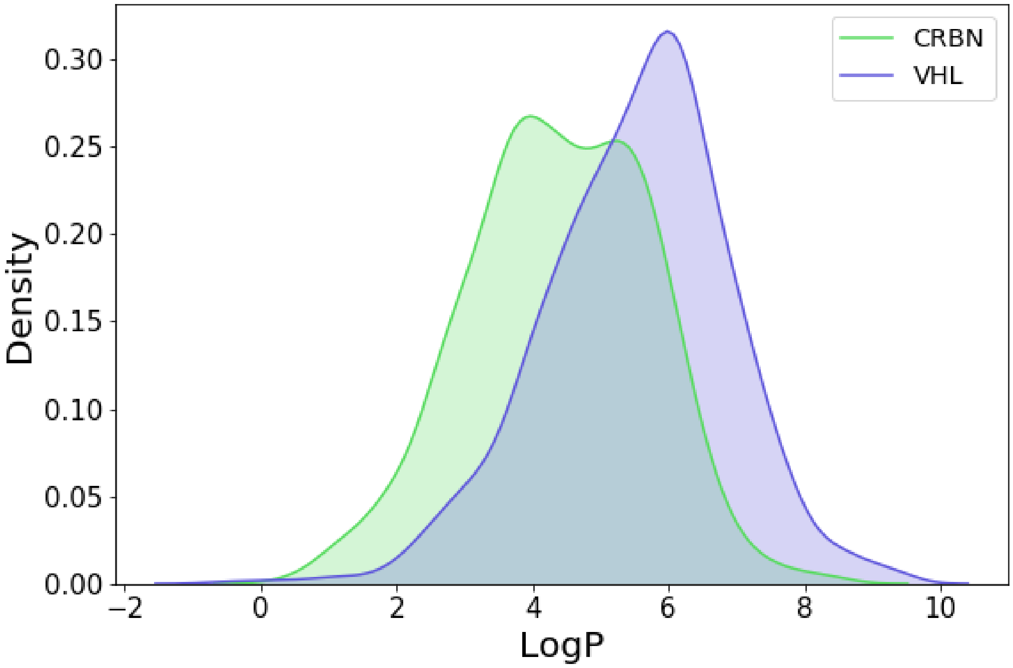


Figure 2.4 Comparison of Patent PROTAC LogP. The kernel density estimated distribution of LogP for patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

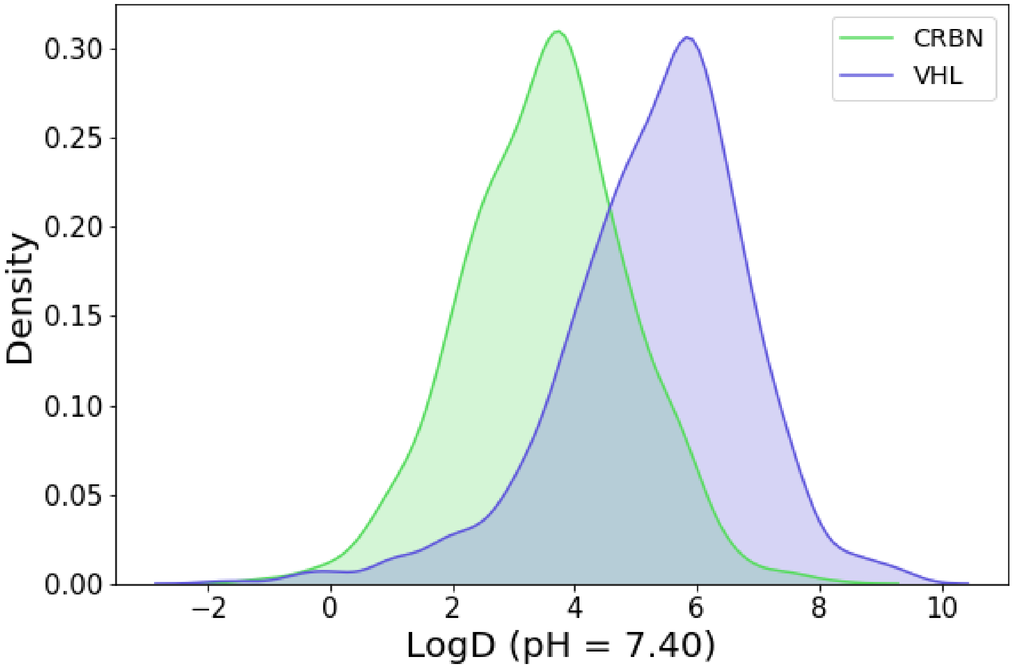


Figure 2.5 Comparison of Patent PROTAC LogD. The kernel density estimated distribution of LogD at pH 7.4 for patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

The VHL targeting PROTACs have a larger number of rotatable bonds compared to the cereblon targeting PROTACs (Figure 2.6). PROTACs in the cereblon set have on average 15 rotatable bonds compared to the 20 for the VHL set, which matches the difference of 5 rotatable bonds between the two E3 ligase targeting ligands. The interquartile region spans a moderately large range of 6 and 5 rotatable bonds for the cereblon and VHL sets respectively. A large range of rotatable bonds is not unexpected for PROTAC molecules as an important part of PROTAC design is determining the optimal linker length. Therefore, several similar PROTAC molecules are usually generated with varying linker lengths to identify what length affords optimal activity.

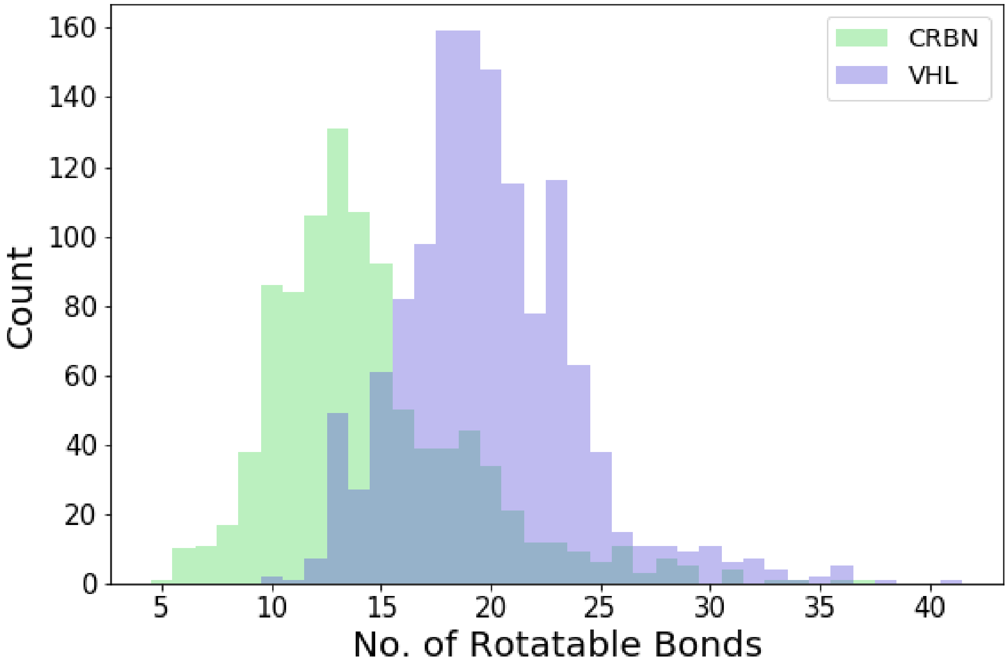


Figure 2.6 Comparison of the Number of Rotatable Bonds for Patent PROTACs. A histogram of the number of rotatable bonds in patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

The VHL targeting PROTACs also have a couple more hydrogen bond donors (Figure 2.7) and acceptors (Figure 2.8) than cereblon targeting PROTACs. The cereblon targeting PROTACs have 2 hydrogen bond donors and 14 hydrogen bond acceptors on average, while the VHL set has 4 hydrogen bond donors and 16 hydrogen bond acceptors. Again, these differences are most easily explained by the differences in the E3 ligase ligands, which have 1 hydrogen bond donor and 6 acceptors for the IMiD common substructure, and 4 donors and 7 acceptors for the VHL targeting ligand. Conversion of the E3 targeting ligand to a PROTAC does not appear to change the number of hydrogen bond donors much, but dramatically increases the number of hydrogen bond acceptors. A likely explanation for the increase in acceptors is the popularity of using polyethylene glycol in the linker region of the PROTAC molecules.

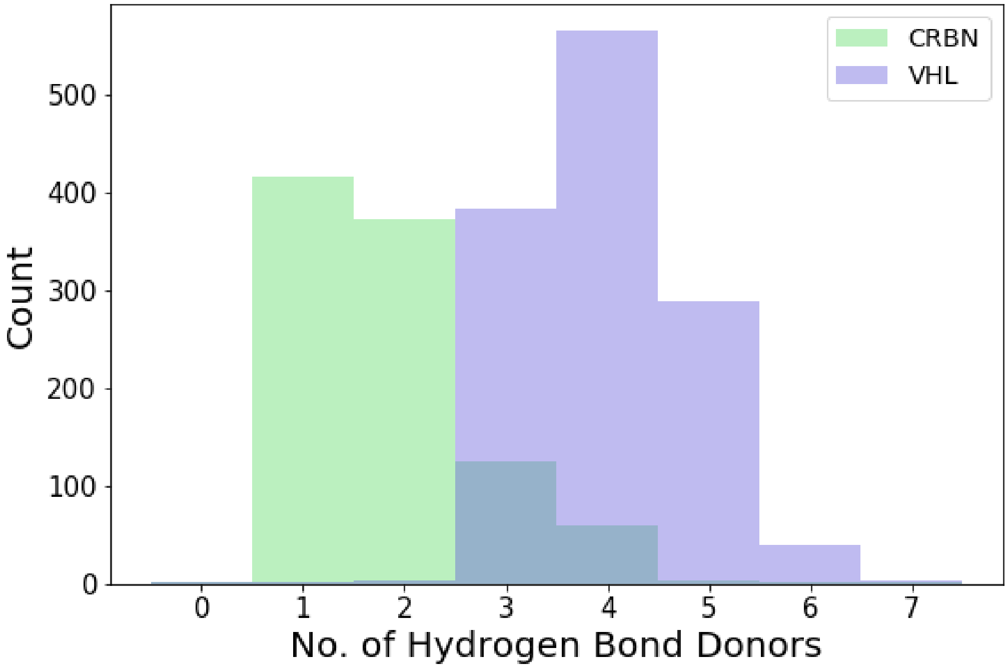


Figure 2.7 Comparison of the Number of Hydrogen Bond Donors for Patent PROTACs. A histogram of the number of hydrogen bond donors in patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

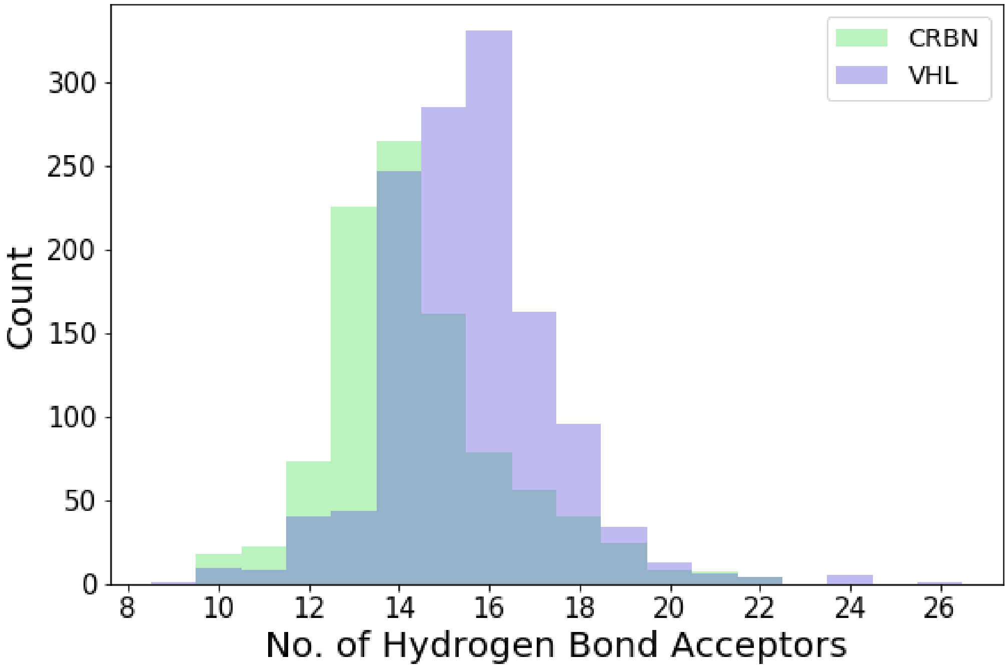


Figure 2.8 Comparison of the Number of Hydrogen Bond Acceptors for Patent PROTACs. A histogram of the number of hydrogen bond acceptors in patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

The VHL and cereblon targeting PROTACs have the same number of aromatic rings (Figure 2.9). Both the PROTACs in the cereblon set and the VHL set have on average 5 aromatic rings. The VHL targeting ligand has only 1 more aromatic ring than IMiD common substructure. In general, the patent PROTACs appear to mainly contain 4 or 5 aromatic rings.

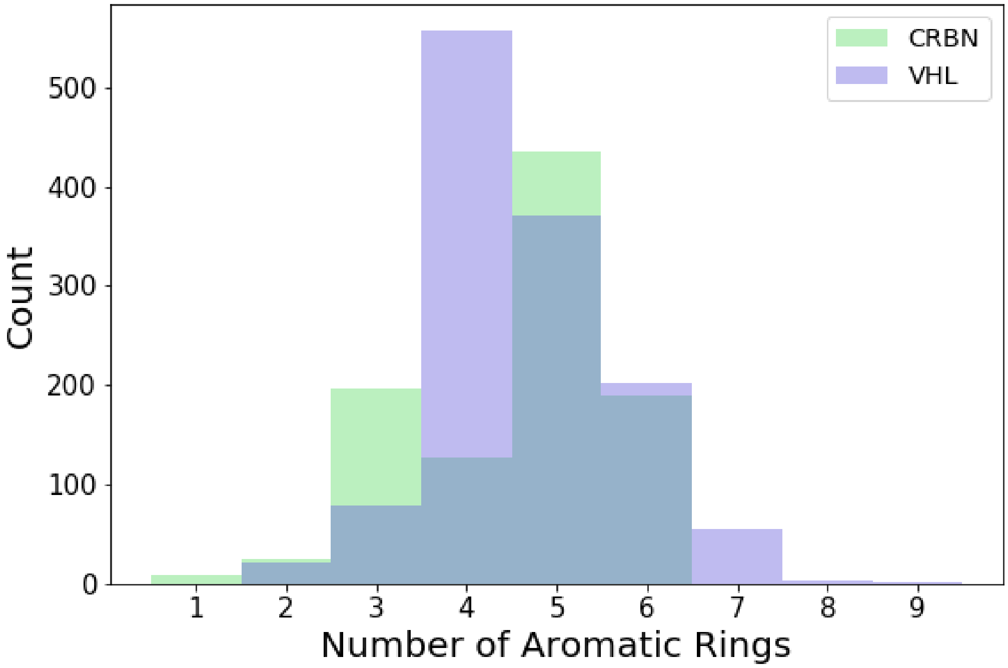


Figure 2.9 Comparison of the Number of Aromatic Rings in Patent PROTACs. A histogram of the number of aromatic rings in patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

Analysis of the physicochemical properties of the patent PROTACs shows that they contain many rule of 5 violations and are generally beyond the traditional rule of 5 chemical space. The patent PROTACs are high molecular weight molecules with many rotatable bonds and many hydrogen bond acceptors. The AbbVie multi-parametric score (AB-MPS) was devised to predict the likelihood of a compound having acceptable oral absorption even with rule of 5 violations. AB-MPS is dependent on the calculated LogD, the number of aromatic rings, and the number of rotatable bonds. This score was calculated for the patent PROTACs and is shown in Figure 2.10. A score less than or equal to 14 is associated with higher probability of acceptable oral absorption. Only 6% of the patent PROTAC compounds from the cereblon set, 62 compounds, and a single compound from the VHL set have an AB-MPS of 14 or less. Therefore, the cereblon targeting patent PROTACs appear to have a higher likelihood of being orally bioavailable according to the AB-MPS, however, the vast majority of PROTAC compounds in the patent literature seem unlikely to be orally bioavailable according to traditional drug design principles.

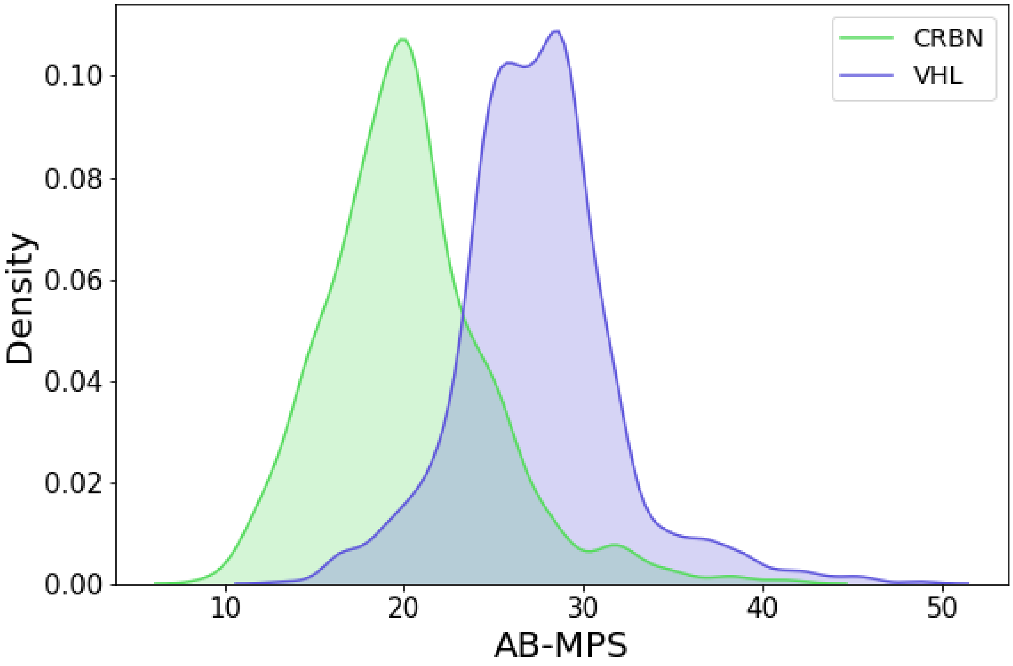


Figure 2.10 Comparison of the AbbVie Multi-Parametric Score for Patent PROTACs. The kernel density estimated distribution of the AbbVie multi-parametric score for patent literature PROTAC compounds targeting either the cereblon (green) or von Hippel-Lindau (blue) E3 Ligase.

However, traditional drug design principles and the AB-MPS assume classical mechanisms of action that are not applicable to PROTAC compounds. Due to the catalytic nature of a PROTAC's mechanism of action, even a small amount of absorbed compound has the potential for efficacy. The first PROTAC, ARV-110, entered Phase I clinical trials in March 2019.[61] The exact structure of the compound has not been disclosed by Arvinas, but is thought to be the one depicted in Figure 2.11 or a closely related compound based on recent patent filings.[62,63] This compound was included in the patent VHL dataset and has a molecular weight of 1020.19 g/mol, a TPSA of 246.99 Å^2 a LogP and LogD of 6.2, 24 rotatable bonds, 3 hydrogen bond donors and 15 acceptors, 4 aromatic rings, and an AB-MPS of 31.2. This compound would not be expected to have good bioavailability according to the rule of 5 guidelines and the AB-MPS. However, the clinical trial is administering the PROTAC as an oral tablet[64], suggesting confidence in either the molecule's oral bioavailability or efficacy even with poor oral absorption. As the clinical trial progresses, the measured bioavailability and efficacy of this new class of compounds will be of great interest.

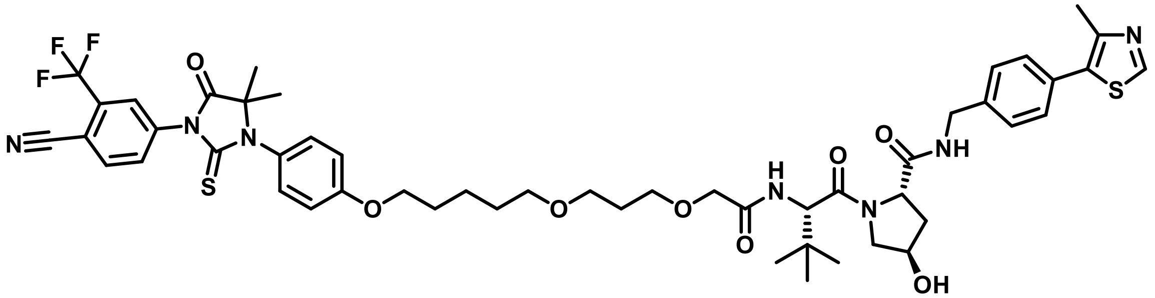


Figure 2.11 Potential Structure of ARV-110. The presumed structure of ARV-110 or a closely related structure to the compound in Phase I clinical trials from Arvinas.

2.4.2 Rational Design of CDK9 PROTACs

A PROTAC molecule consists of three major pieces: the ligase binding ligand, the linker, and the target protein binding ligand. The attachment point of the linker to each of the ligands is very important as an improper point of attachment can interfere with ligand-protein binding. The point of attachment for the ligase binding ligand is well studied, but this is not always the case for the protein binding ligand. In the absence of a crystal structure of the ligand-protein complex, known structure-activity relationships can be leveraged to identify potentially successful sites for linker attachment. Crystallographic data greatly aids the identification of linker attachment sites as binding site exit vectors can be visually identified. CDK9 is a promising target for a PROTAC approach due to significant amount of crystallographic data available for CDK proteins.

The CDK9 protein binding ligand selected for PROTAC generation was AT7519 (Figure 2.12). AT7519 is an extremely potent inhibitor of CDK9, $IC_{50} < 10$ nM, and has a published crystal structure with CDK2.[65] Due to the structural similarity of CDK9 and CDK2 (Figure 2.13), the crystal structure of AT7519 with CDK2 can still provide the desired insights into ideal sites of linker attachment. The PROTAC approach has yet to be explored with AT7519 and its high potency and crystallographic data make it a prime candidate.

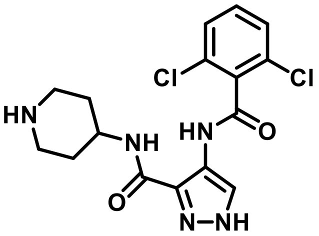


Figure 2.12 Structure of AT7519. The CDK9 inhibitor selected for PROTAC development.

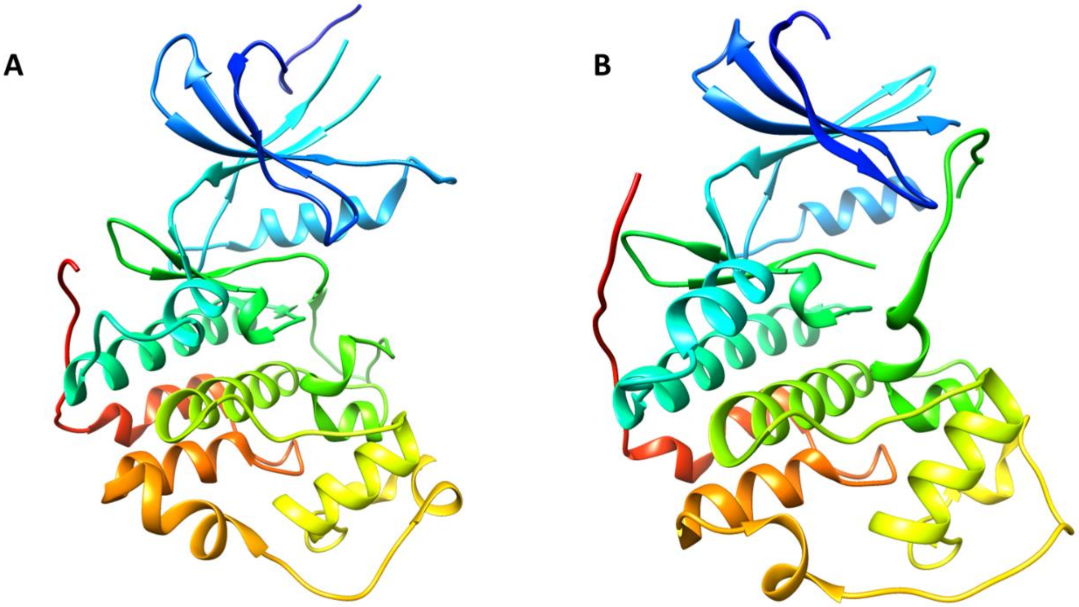


Figure 2.13 Comparison of CDK9 and CDK2. The secondary structures of **(A)** CDK9 (PDB: 4BCG) and **(B)** CDK2 (PDB:2VU3) rainbow colored from N to C terminus for Chain A. The CDK2 structure has more unresolved loop regions compared to the CDK9 structure.

The crystal structure of AT7519 bound to CDK2 (PDB: 2VU3) was inspected to identify sites for linker attachment. Examination of AT7519 bound to the ATP binding site of CDK2 reveals two potential sites for linker attachment (Figure 2.14). Both the dichlorophenyl and piperidine moieties project toward solvent. Therefore, a linker could be potentially attached at either the *para* position on the dichlorophenyl ring or to the nitrogen atom in the piperidine without causing a steric clash in the binding pocket. While the dichlorophenyl projects towards the solvent, it sits more deeply in the pocket than the piperidine ring. A site of attachment that is deeper in the pocket would require a longer linker relative to a site of attachment that is more solvent exposed, which would increase the molecular weight and likely the number of rotatable bonds in the PROTAC molecule. While PROTAC molecules may tolerate larger molecular weights and a higher number of rotatable bonds compared to traditional inhibitors, reducing the amount of non-drug like properties is still desirable when possible. Therefore, the nitrogen on the piperidine ring was selected as the point for linker attachment.

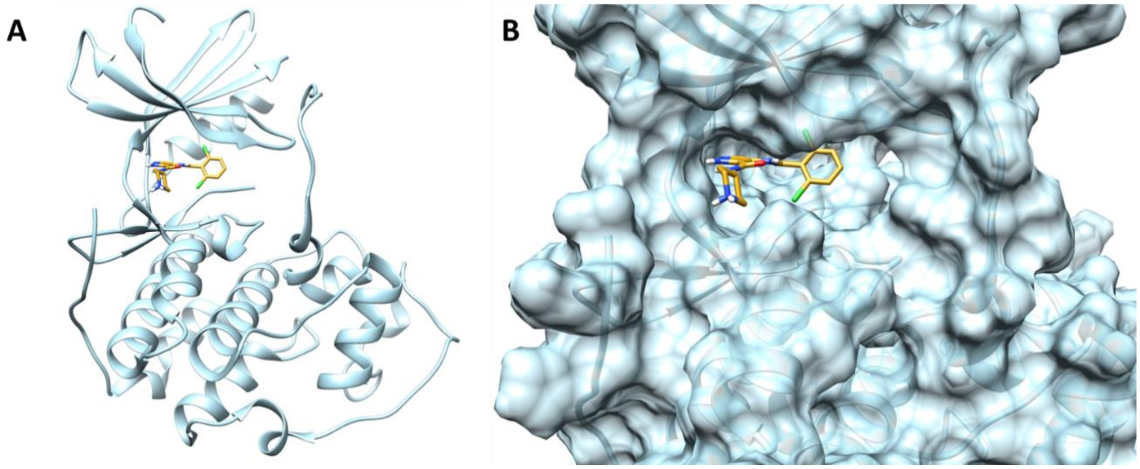
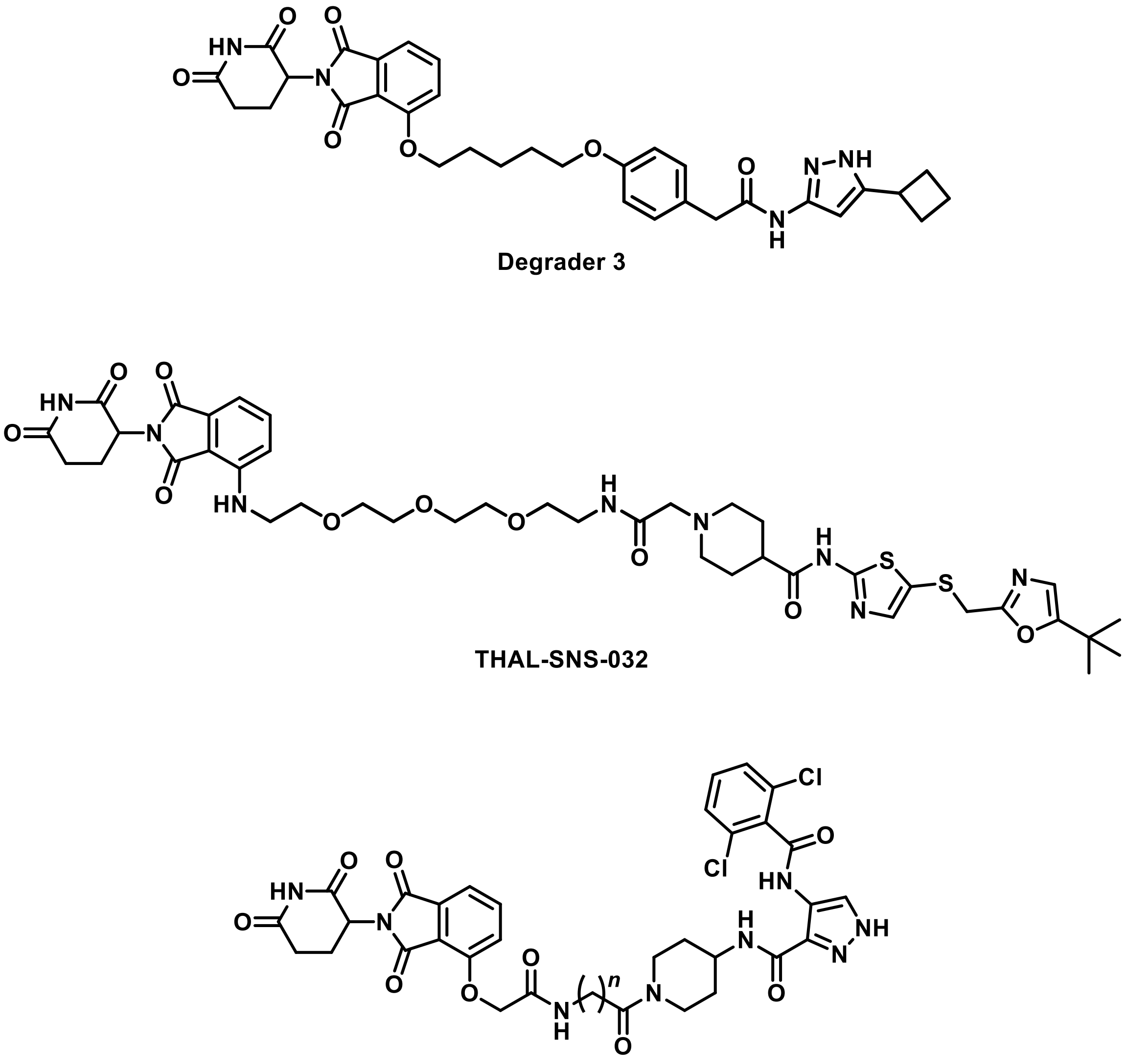


Figure 2.14 Crystal Structure of AT7519 Binding to CDK2. (A) The secondary structure of CDK2 (light blue, PDB:2VU3) and AT7519 (gold) in ATP binding site. **(B)** A closer view of AT7519 in the ATP binding site of CDK2 with the protein surface shown.

The design strategy described is further supported by literature precedence. Currently, two PROTAC molecules have been published that target CDK9 (Figure 2.15). These PROTAC molecules utilize either a previously unpublished 3-aminopyrazole scaffold or SNS-32 for the CDK9 targeting ligand.[37,66] The activity of the 3-aminopyrazole compound against CDK9 has not yet been reported, while SNS-32 has been shown to inhibit CDK9 with an IC_{50} value of 4 nM.[67] Furthermore, crystal structures of a similar 3-aminopyrazole compound and SNS-32 bound to CDK2 have been published. The overlay of these two compounds with AT7519 (Figure 2.16) show that the selected points of linker attachment for the 3-aminopyrazole PROTAC "Degrader 3" and THAL-SNS-32 are in close proximity with the selected attachment point for AT7519. Both "Degrader 3" and THAL-SNS-32 have been shown to successfully degrade CDK9. However, CDK9 is still present at concentrations as high as 20 µM for "Degrader 3", while complete degradation is observed at concentrations as low as 250 nM for THAL-SNS-32. Therefore, a PROTAC approach using AT7519 with a linker attached to the nitrogen piperidine appears to be a promising strategy.



Degrader 3

THAL-SNS-032

AT7519-PROTAC

Figure 2.15 Literature CDK9 PROTACs Compared to AT7519 PROTAC Design. The published CDK9 PROTAC compounds, Degrader 3 and THAL-SNS-032 are shown in comparison with the proposed design of an AT7519-PROTAC.

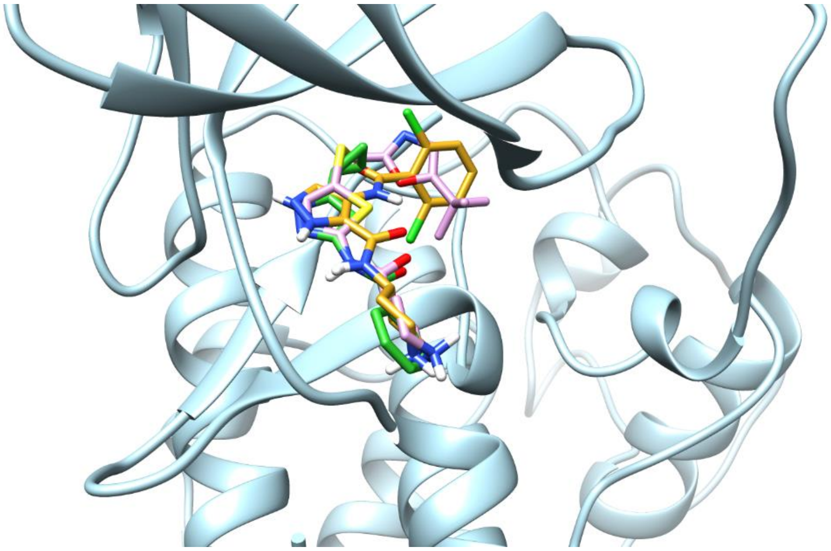


Figure 2.16 Overlay of the PROTAC CDK9 Targeting Ligands Bound to CDK2. The CDK9 targeting ligands: 3-amino pyrazole (green), SNS-32 (pink), and AT7519(gold) are shown overlayed in the CDK2 (light blue) ATP binding site.

2.4.3 Properties of CDK9 PROTACs

To test the PROTAC design strategy, several PROTAC compounds were synthesized. These compounds contained either relatively short (Figure 2.17) or long (Figure 2.18) linker chains. The physicochemical properties of each PROTAC were calculated and converted to z-scores using the patent literature physicochemical property distributions for cereblon targeting PROTACs. The z-scores enable quick comparisons between the properties of each compound and also put the properties in the context of the patent literature. Furthermore, it allows multiple properties to be plotted together even if the original values had greatly different scales. A given z-score represents how many standard deviations a given physicochemical property values is away from the mean of that property distribution of the patent literature compounds. For reference, the mean values (standard deviation) for each property in the cereblon patent data set are shown in Table 2.1. Currently, the ideal physicochemical properties values for PROTAC molecules are unknown. Therefore, high or low z-scores are not necessarily better or worse. However, the lower the z-score, the closer the property is to obeying Lipinski's Rule of 5 (Ro5) and traditional drug design principles. The goal is to gain a better understanding of which properties are important for the desired permeability and efficacy for this new compound class by closely tracking these physicochemical properties and observing any potential correlations with biological endpoints.

In general, the PROCDK9 series has higher values for the number of hydrogen bond donors (HBD), the number of hydrogen bond acceptors (HBA), and topological polar surface area (TPSA) compared to the patent literature of cereblon targeting PROTACs. The compound with the shortest linker, PROCDK9-27, has z-scores of 2.30, 0.30, and 0.88 for HBD, HBA, and TPSA respectively. For all the other PROCDK9 compounds, these z-scores increase to 3.37, 1.29, and 1.84 for HBD, HBA, and TPSA respectively.

Predominantly, the shorter linker PROCDK9 compounds have lower values for molecular weight (MW), number of rotatable bonds (NROT), and the AbbVie multi-parametric score (AB-MPS) compared to the patent literature while the longer linker compounds have higher values. PROCDK9-27 has z-scores of -1.52, -1.42, and -1.59 and PROCDK9-21 has z-scores of -0.23, -0.33, and -0.55 for MW, NROT, and AB-MPS respectively. On the contrary, PROCDK-24 has z-scores of 0.32, 0.33, and -0.07 and PROCDK-25 has z-scores of 0.86, 0.98, and 0.44 for MW, NROT, and AB-MPS respectively.

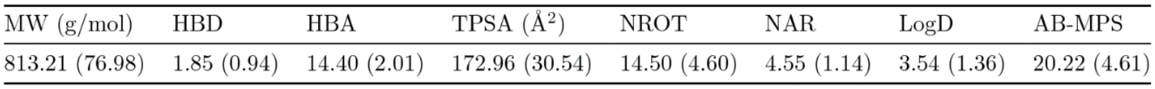
Furthermore, all the PROCDK9 compounds have lower values for LogD and the number of aromatic rings (NAR) compared to the patent literature. For LogD, PROCDK9-27 has a z-score of -1.79, PROCDK9-21 a z-score of -1.64, PROCDK9-24 a z-score of -1.04, and PROCDK9-25 a z-score of -0.20. Since all the PROCDK9 compounds have the same number of aromatics rings, they all have the same z-score of -1.36.

The values for HBD, HBA, NAR, and TPSA essentially remain constant as the linker length increases while the values for MW, LogD, NROT, and AB-MPS increase with linker length. From the shortest linker PROTAC, PROCDK9-27, to the next shortest, PROCDK9-21, an additional amide functional group is present in the linker. This accounts for the additional hydrogen bond donor, additional hydrogen bond acceptor, and increase in polar surface area. All the other longer PROTACs also include this functional group and the only difference between them is the length of the alkyl chain. Thus, the number of hydrogen bond donors, hydrogen bond acceptors, and topological polar surface areas are identical for all but the shortest PROTAC. Additionally, the number of aromatic rings remains constant as no aromatic rings are included in the linker. However, the increasing length of the alkyl chain directly increases MW, LogD, and NROT. The AB-MPS also increases as the score is dependent on LogD, NROT, and NAR.

For the properties which are linker length independent, the high values of HBA and TPSA appear as potentially problematic while HBD and NAR are in an acceptable range. On first examination the HBD appears very high as these PROTACs have either 4 or 5 hydrogen bonds, which is more than was observed on average for the cereblon patent compounds. However, this is not particularly concerning as this is still a traditionally drug-like number of hydrogen bonds. Conversely, the number of hydrogen bond acceptors for the PROCDK9 series, which contain 15 or 17 HBA, is relatively close to the average for the cereblon patent PROTACs. Despite relatively closely matching the average in the patent literature, this is a property of potential concern due to these values greatly exceeding the traditional Lipinski guideline value of 10 and even extends past the maximum value considered in the AbbVie beyond rule of 5 (bRo5) dataset which was 15. Additionally, the topological polar surface area is a fair bit higher for the PROCDK9 series compared to the patent compounds. These compounds have a TPSA of 199.97 Å^2 and 229.07 Å^2, which is on the border of the $\leq$ 229 Å^2 maximum acceptable TPSA threshold described by AbbVie for bRo5 oral bioavailability. The PROCDK9 compounds each have 3 aromatic rings, which is lower than the average value for the patent compound and appears to be in an acceptable range. AbbVie noted that a large number of aromatic rings was associated with poor oral bioavailability and the number of aromatic rings in the PROCDK9 compounds is lower than the average value in the AbbVie dataset for compounds with good oral bioavailability.[32]

For the properties which are linker length dependent, the values of LogD, MW, and NROT for the PROCDK9 series appear to be in acceptable range according to the bRo5 properties described by AbbVie with some compounds exceeding the recommended soft threshold for AB-MPS. PROCDK9-27, -21, -24, and -25 have LogD/LogP values of 1.10, 1.30, 2.12, and 3.27 respectively, which conform to Lipinski's rule of a LogP $\leq$ 5. However, it is possible that the shorter PROTACs are too polar as AbbVie's bRo5 data indicates that a LogP close to 3 is important. PROCDK9-27, -21, -24, and -25 have molecular weights of 696.49, 795.63, 837.70, and 879.78 g/mol respectively, which violate Lipinski's rule of MW $\leq$ 500 g/mol. These molecular weights are deemed acceptable despite this violation as the AbbVie bRo5 dataset indicates that compounds with molecular weights as high as 1132 g/mol can be orally bioavailable. Furthermore, compared to the other physicochemical properties, MW will be most difficult to significantly reduce due to the necessity of linking two molecules together. Compounds PROCDK9-27, -21, -24, and -25 have 8, 13, 16, and 19 rotatable bonds respectively. PROCDK9-27 has a traditionally acceptable number of rotatable bonds and the other compounds have an acceptable number of rotatable bonds according to the bRo5 dataset, which observed oral bioavailability at NROT values of up to 19. Compounds PROCDK9-27, -21, -24, and -25 have AB-MPS scores of 12.90, 17.70, 19.88, and 22.27 respectively. The AbbVie bRo5 dataset suggests that AB-MPS value $\leq$ 14 predict higher probability of success. The shortest PROTAC, PROCDK9-27, has an acceptable value and the other compounds extend beyond the ideal range.

Table 2.1 Mean and Standard Deviation of Physicochemical Properties for Cereblon Patent PROTACs



MW (g/mol)	HBD	HBA	TPSA (Å^2)	NROT	NAR	LogD	AB-MPS
813.21 (76.98)	1.85 (0.94)	14.40 (2.01)	172.96 (30.54)	14.50 (4.60)	4.55 (1.14)	3.54 (1.36)	20.22 (4.61)

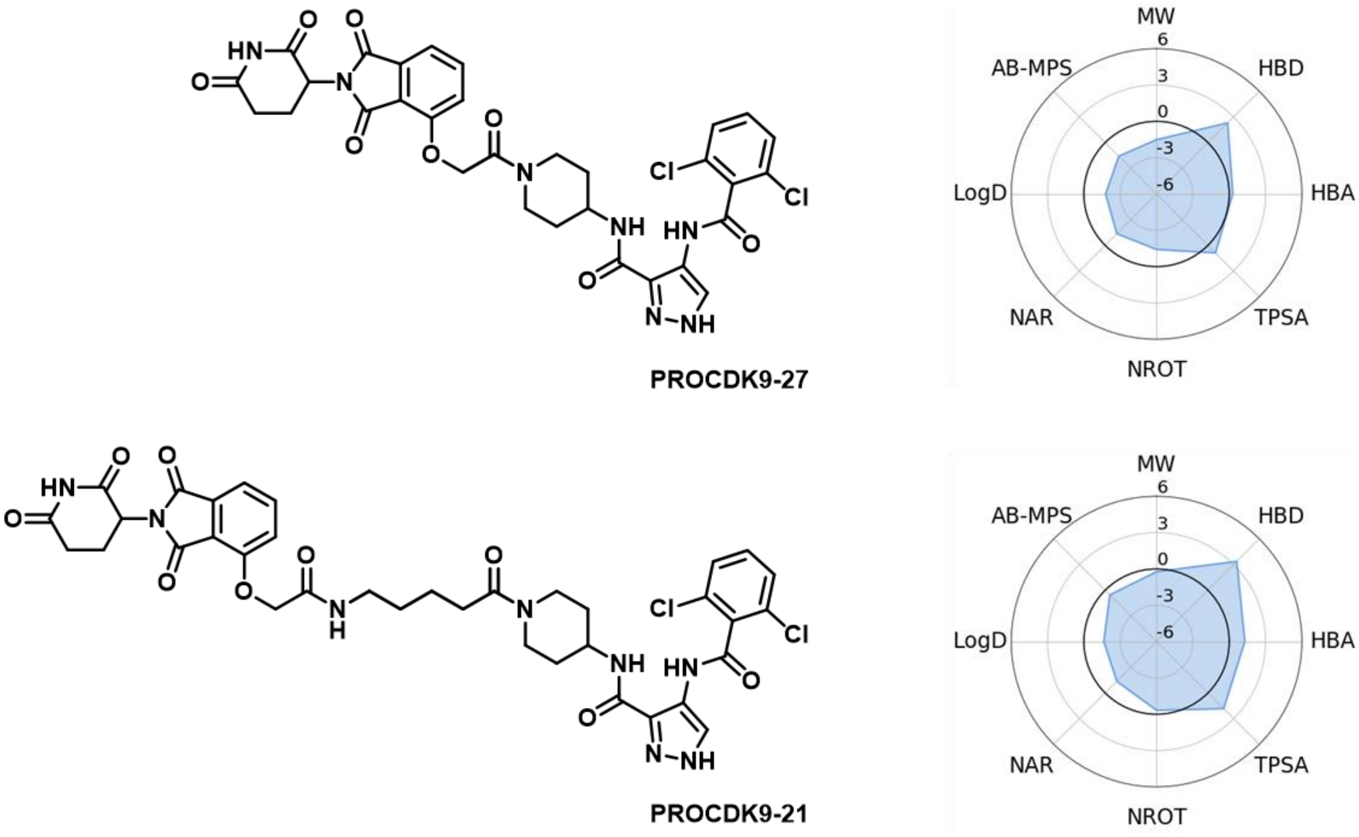


Figure 2.17 Structure and Properties of Short Linker CDK9 PROTACs. The structure and physicochemical properties of CD9 PROTACs with a short linker chain. To the right of each structure, a spider plot is shown which contains the z-scores of each physiochemical property: Molecular Weight (MW), Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Topological Polar Surface Area (TPSA), Number of Rotatable Bonds (NROT), Number of Aromatic Rings (NAR), LogD, and the AbbVie Multi-Parametric Score (AB-MPS).

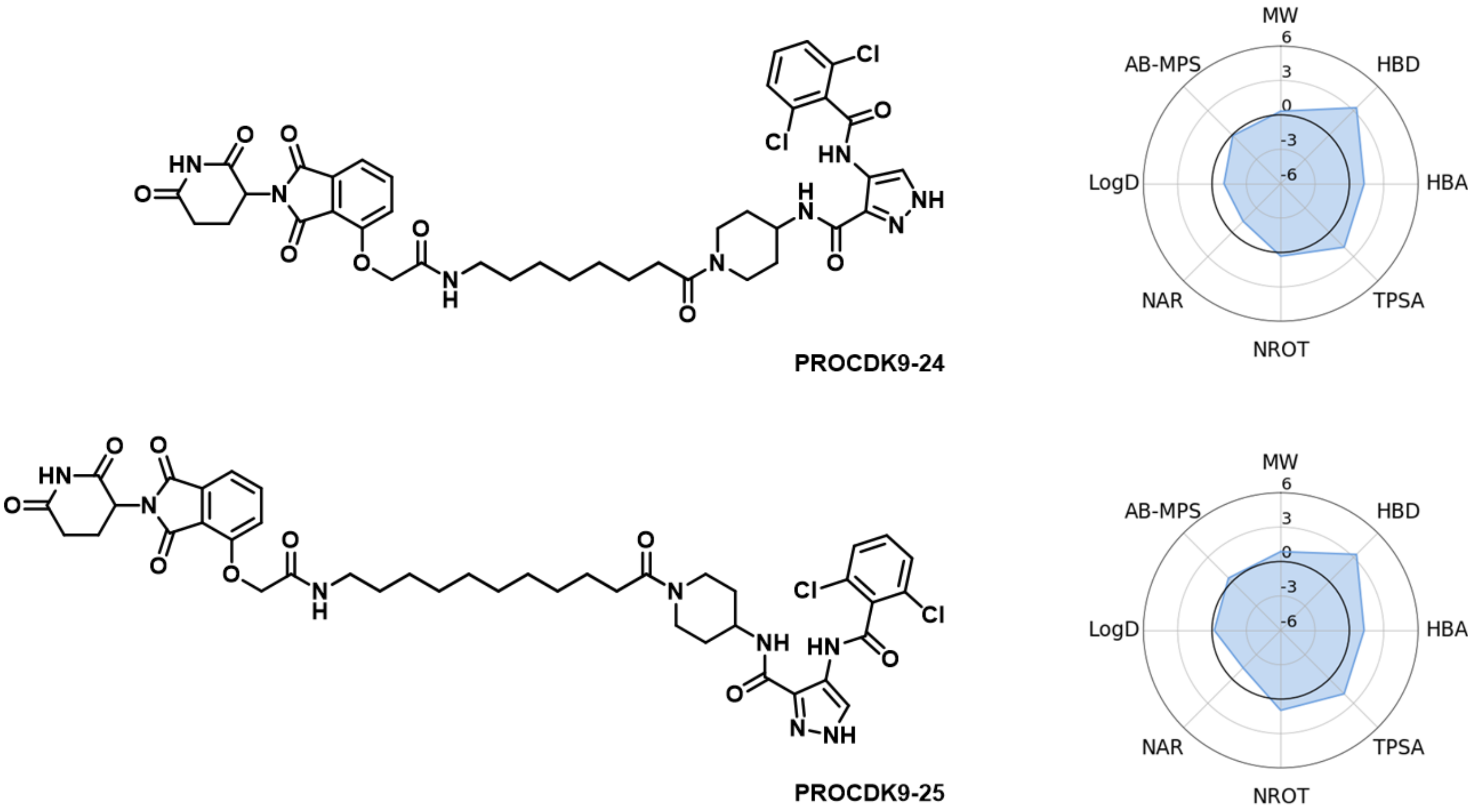


Figure 2.18 Structure and Properties of Long Linker CDK9 PROTACs. The structure and physicochemical properties of CD9 PROTACs with longer linker chains. To the right of each structure, a spider plot is shown which contains the z-scores of each physiochemical property: Molecular Weight (MW), Hydrogen Bond Donors (HBD), Hydrogen Bond Acceptors (HBA), Topological Polar Surface Area (TPSA), Number of Rotatable Bonds (NROT), Number of Aromatic Rings (NAR), LogD, and the AbbVie Multi-Parametric Score (AB-MPS).

2.4.4 Degradation Activity of CDK9 PROTACs

The degradation of both the 55 kDa and 42 kDa isoforms of cyclin dependent kinase (CDK) CDK9 were assessed. The 42 kDa isoform is more abundant than the 55 kDa and the isoforms are identical except for an additional 117 terminal residues on the 55 kDa isoform. The isoforms have been shown to phosphorylate all of the same peptide substrates tested so far, but have differing expression patterns and localization.[68] While the expression and localization differences suggest some functional differences, this has yet to be fully characterized. [36]

In addition to monitoring CDK9 degradation, expression levels of the induced myeloid leukemia cell differentiation protein (MCL-1) were also monitored. MCL-1 is an anti-apoptotic factor with a short half-life. CDK9 phosphorylation of RNA polymerase II facilitates transcription and expression of MCL-1. When CDK9 activity is inhibited, this transcription of the MCL-1 gene is prevented and MCL-1 protein levels are rapidly reduced.[69] Therefore, monitoring MCL-1 levels allows the monitoring of the downstream effect of CDK9 degradation.

The degradation activity of the PROCDK9 compounds were initially assessed at relatively high concentrations (Figure 2.19). PROCKD9-27 shows the expected loss of MCL-1 from CDK9 inhibition, but little to no CDK9 degradation. This indicates that PROTCDK9-27 is binding to CDK9 and inhibiting its activity, but not forming the ternary complex with the cereblon E3 ligase to cause degradation. PROCDK9-21 shows little to no effect on both the expression levels of MCL-1 and CDK9. Therefore, it is likely not binding and inhibiting CDK9 nor forming the ternary complex. PROCDK9-24 shows complete absence of MCL-1 expression at all concentrations tested and some reduction in CDK9 expression. Thus, it is likely binding CDK9 and successfully forming the ternary complex resulting in degradation. However, it is also possible that the reduction in MCL-1 is also from direct inhibition of CDK9 as expression of CDK9 is still observed. Interestingly, higher levels of CDK9 expression are observed as the dose increases from 5 µM to 10 µM to 20 µM. This is most likely due to a phenomenon referred to as the "hook effect". The hook effect occurs at high doses of PROTAC compounds as the compounds fully saturate the binding sites of the E3 ligase and the target protein. If the binding sites are saturated, a single PROTAC molecule will have difficulty forming the ternary complex as it will have to compete with another PROTAC molecule already bound to the other protein partner. Assessing PROCDK9-24 at lower concentrations will be important for determining if this is indeed occurring. PROCDK9-25 shows reduction in MCL-1 and the complete absence of CDK9 at all doses. Therefore, PROCDK9-25 appears to be successfully forming the ternary complex resulting in CDK9 degradation and in turn reduced expression of MCL-1.

The degradation activity of the same PROCDK9 compounds were also assessed at lower concentrations (Figure 2.20). Again, both PROCDK9-27 and -21 show no degradation of CDK9 and now have no impact on MCL-1 expression levels at this lower concentration. PROCDK9-24 now demonstrates the expected dose-dependent reduction in CDK9 expression and a corresponding loss in MCL-1, which supports that the hook effect was indeed occurring at the higher doses. PROCDK9-25 appears to be a very potent PROTAC compound as minimal CDK9 expression is observed at concentrations as low as

250 nM. Overall, PROCDK9-24 and -25 appear to be functioning as degraders while PROCK9-27 and -21 do not.

While great attention has been paid to the physicochemical properties of these compounds, the most important feature for activity at this point appears to be linker length. The number of atoms in the linker for PROCDK9-27, -21, -24, and -25 are 2, 8, 11, and 14 atoms respectively. This count is including all the atoms between the 4-hydroxy oxygen on thalidomide and the nitrogen piperidine. A linker length of 8 atoms or less appears inadequate for ternary complex formation between cereblon and CDK9 for this series of PROTAC compounds. While 11 atoms appear to allow for ternary complex formation and degradation, it does not appear to be as effective as 14 atoms. Further exploration of linker length is of interest. Specifically, lengths of 12 and 13 atoms and the determination of the upper limit of tolerated linker length.

Analysis of physicochemical properties is expected to be of greater value after degradation activity has been demonstrated and optimal linker length has been determined. Once an optimal length has been determined, a series of PROTAC molecules can be synthesized with that consistent length but containing various functionalities and thereby different physicochemical properties. As additional biological endpoints relating to permeability and bioavailability are collected for this series, any correlations between the calculated physicochemical properties and these endpoints will be examined. Any observations can be used to guide further design iterations and this knowledge ultimately transferred to PROTAC design efforts for other protein targets.

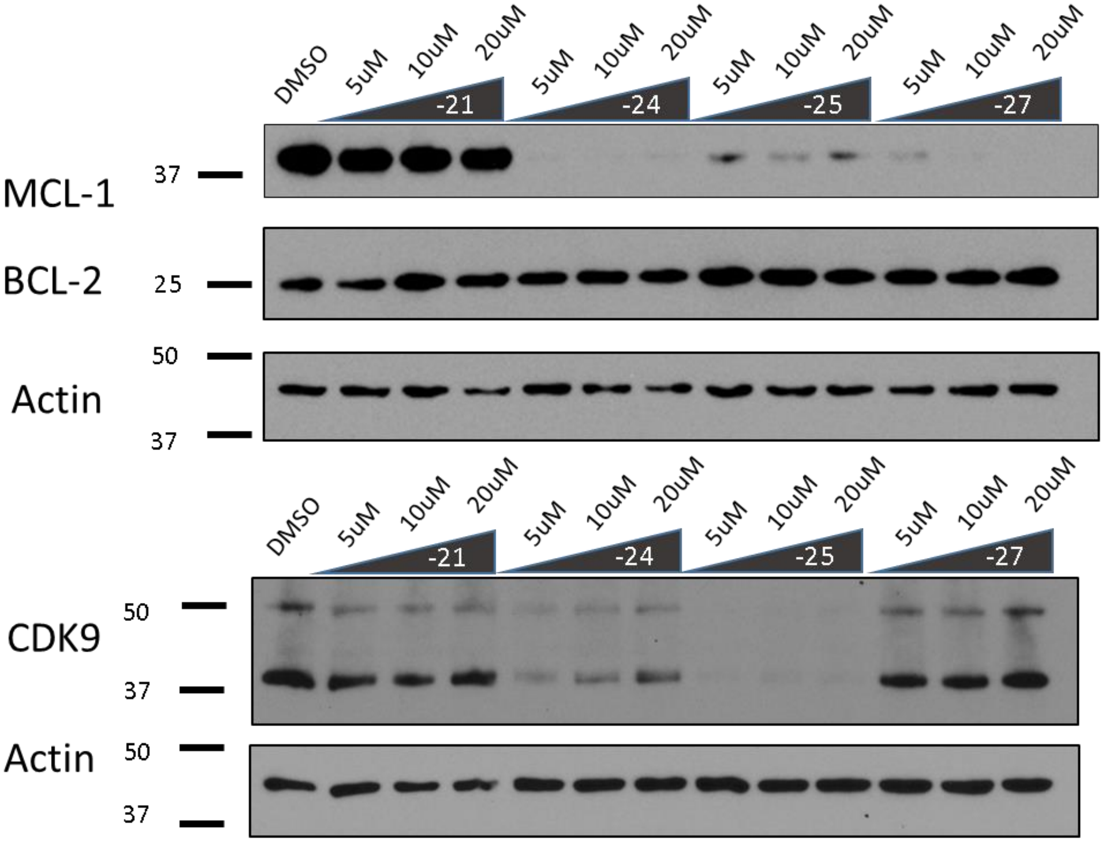


Figure 2.19 Activity of PROCDK9 Compounds at High Concentrations. A Western blot showing the degradation activity of PROTACs targeting CDK9 after 6 hours of exposure to compound in the MV4-11 cell line. MCL-1 is a rapidly turned over protein that is not replenished without CDK9 activity, thus CDK9 inhibition or degradation reduces MCL-1 expression levels. BCL-2 has a longer half-life than MCL-1 and BCL-2 expression levels are monitored to control for long-term aberrations in transcription. Expression level of actin is shown as a control for the background protein expression level. Degradation of CDK9(55) and CDK9(42) isoforms by PROCDK-21, -24, 25, and -27 at 5 µM, 10 µM, and 20 µM are shown.

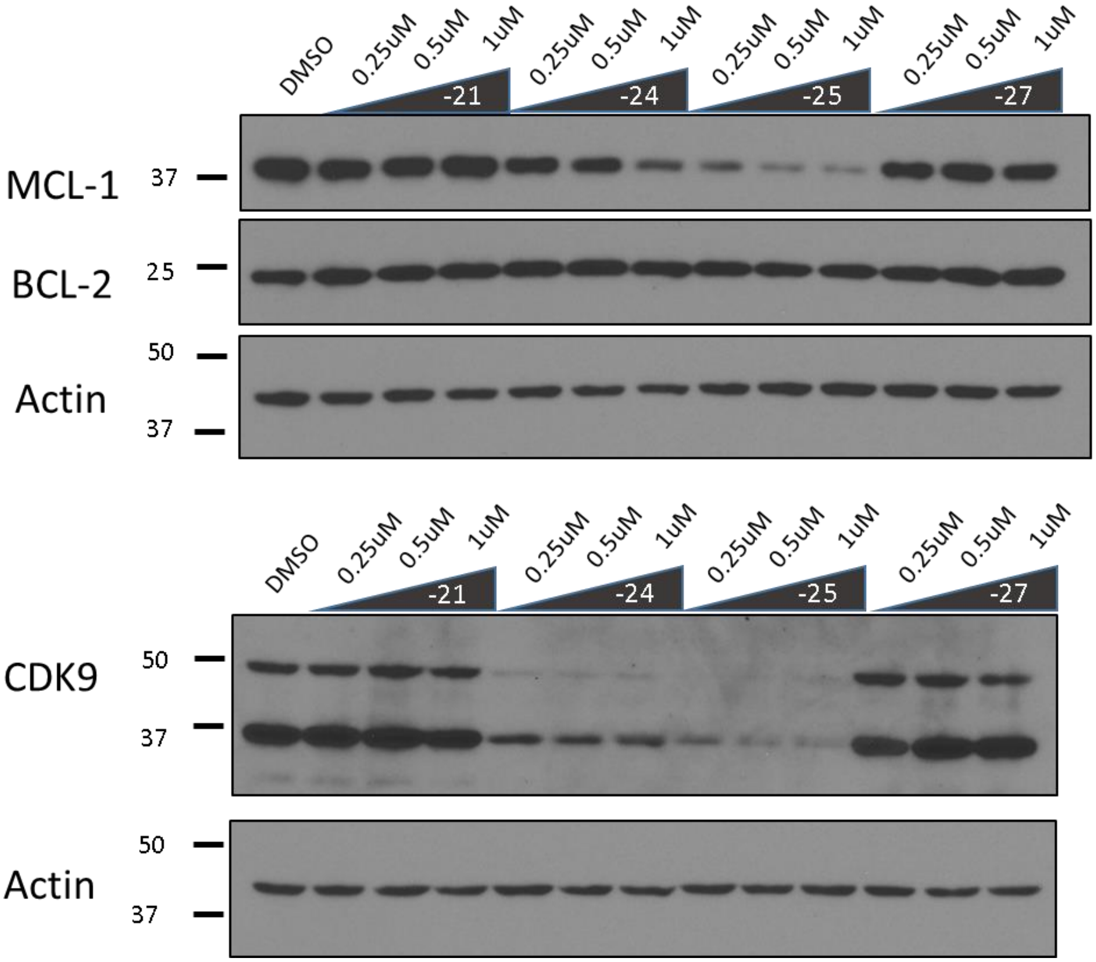


Figure 2.20 Activity of PROCDK9 Compounds at Low Concentrations. A Western blot showing the degradation activity of PROTACs targeting CDK9 after 6 hours of exposure to compound in the MV4-11 cell line. MCL-1 is a rapidly turned over protein that is not replenished without CDK9 activity, thus CDK9 inhibition or degradation reduces MCL-1 expression levels. BCL-2 has a longer half-life than MCL-1 and BCL-2 expression levels are monitored to control for long-term aberrations in transcription. Expression level of actin is shown as a control for the background protein expression level. Degradation of CDK9(55) and CDK9(42) isoforms by PROCDK-21, -24, 25, and -27 at 0.25 µM, 0.5 µM, and 1 µM are shown.

2.4.5 Proof of Concept for Generation of Ternary Structure Predictions

In 2019, Drummond and Williams published a series of methods for modeling PROTAC-mediated ternary complexes.[70] These methods and corresponding filtering processes were benchmarked on the available crystal structure data using the proprietary software package Molecular Operating Environment (MOE)[71]. Their most successful method, “Method 4”, produced a set of ternary complex poses of which ~40% had a Cα RMSD $\leq$10 Å for the VHL-MZ1-BRD4 ternary complex. However, this method had more difficulty in predicting the ternary complexes which contained cereblon as the E3 ligase instead of VHL.

Inspired by their approach, a similar protocol is described here using freely accessible software packages. A schematic of this method is shown in Figure 2.21 and was described in detail in Section 2.3.4. The method consists of three main steps: protein-protein docking, filtering, and conformer generation. The first and last steps are the most crucial. The first step generates all the poses to be considered and the last step applies the strictest filter. The second step, which filters by ligand distance and hydrophobic surface burial, is a quick way to remove many irrelevant poses. The final conformer generation step should also remove those poses but would take significantly more computational resources to do so.

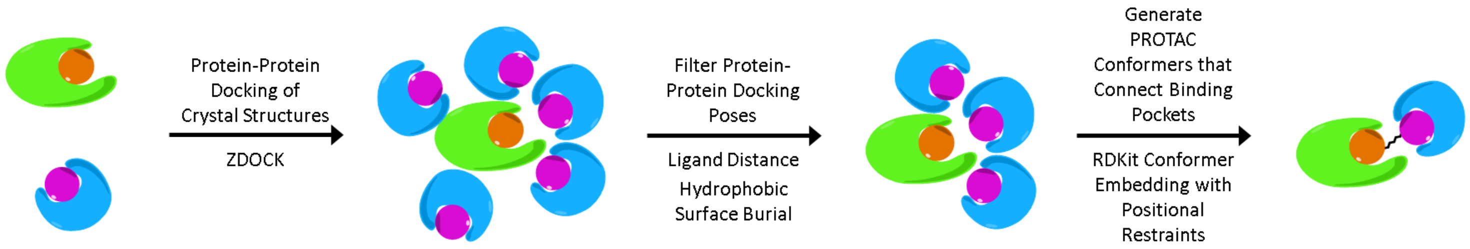


Figure 2.21 Schematic of Ternary Structure Prediction. A cartoon schematic showing the steps in the PROTAC ternary structure prediction procedure. First, prepared crystal structures with ligands in the binding site are docked using ZDOCK. These predicted protein-protein interfaces are filtered by both: the center of mass distance between the ligands that reflects the approximate legth of the fully extended PROTAC and if at least 100 Å^2 of protein hydrophobic solvent-accessible surface is buried. For structures which pass the filtering process, PROTAC conformers are generated with restraints on the atomic positions of substructure atoms in common with the crystallographic ligands.

To determine if this method was capable of producing correct ternary complex structures, it was benchmarked using the VHL-MZ1-BRD4 complex (PDB: 5T35), the CRBN-dBET6-BRD4 complex (PDB: 6BOY), and the CRBN-dBET23-BRD4 complex (PDB: 6BN7).The results for VHL-MZ1-BRD4, CRBN-dBET6-BRD4, and CRBN-dBET23-BRD4 complexes are shown in Figure 2.22, Figure 2.23, and Figure 2.24 respectively. Of the filtered ternary complex structures for which a PROTAC conformer could be generated, 15%, 42%, and 60% of the predicted complexes were correct for VHL-MZ1-BRD4, CRBN-dBET6-BRD4, and CRBN-dBET23-BRD4 respectively. For comparison, the best results obtained by Drummond and Williams for these same complexes were 40%, 0%, and 10% when only considering the top results regardless of any methodological differences. Interestingly, the methodology used by Drummond and Williams performs significantly better for the VHL ligase containing complex compared to the method described here, while the converse is true for the CRBN ligase containing complex. As the method here was inspired by the work done by Drummond and Williams the general workflow is very similar. Therefore, the performance difference is most likely due to the different protein-protein docking software used.

While the correct structure was able to be predicted for each benchmark complex, delineating the correct complex predictions from the incorrect predictions *a priori* is difficult. Each predicted complex structure was scored using ZRANK and the scores were compared to the calculated Cα RMSD. Interestingly, the top ranked complex by ZRANK was a successful prediction for each benchmark complex. However, ZRANK scores did not generally separate correct predictions from incorrect ones well.

Additionally, the final predicted structures were locally minimized with the PROTAC molecule present to investigate whether this might improve the ability of ZRANK scores to distinguish between correct and incorrect predictions. For example, approximately 19% of the correct predictions for the CRBN-dBET6-BRD4 complex have ZRANK scores greater than 0. These high scores are due to steric clashes at the protein-protein interface that results after adding the hydrogen atoms that are neglected by ZDOCK but required by ZRANK. Therefore, resolving these clashes through local minimization could greatly improve the ZRANK score and permit correct structures to be better distinguished from incorrect by the ZRANK score. Local minimization did indeed resolve these clashes and improve ZRANK score, but it did so for the incorrect predictions as well. Therefore, the local minimization procedure did not improve the ability of ZRANK scores to differentiate between correct and incorrect predictions.

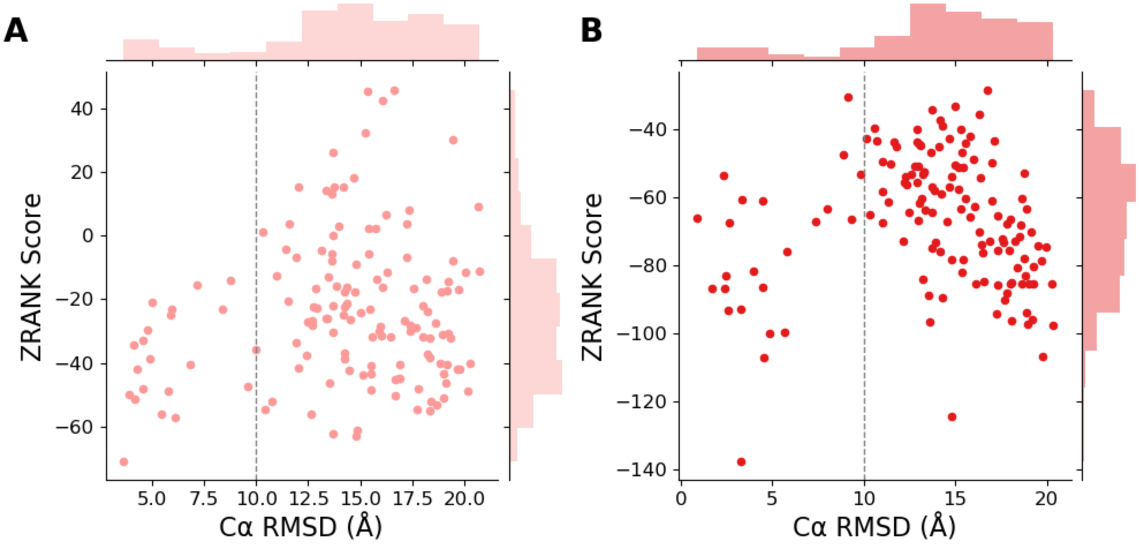


Figure 2.22 Ternary Structure Predictions Results for VHL-MZ1-BRD4. The Cα RMSD of the predicted ternary complex to the known crystal structure vs. the predicted complex's ZRANK score is shown for the final complexes (A) output directly by ZDOCK and the (B) locally minimized structures. A dotted line is place at Cα RMSD 10 Å which shows the threshold for which given structure was deemed correct.

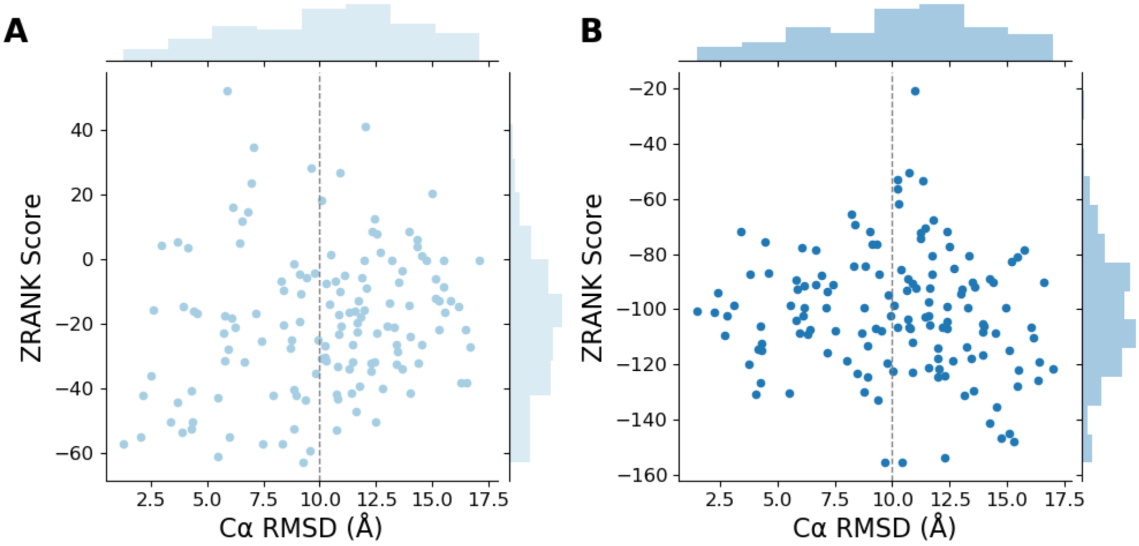


Figure 2.23 Ternary Structure Predictions Results for CRBN-dBET6-BRD4. The Cα RMSD of the predicted ternary complex to the known crystal structure vs. the predicted complex's ZRANK score is shown for the final complexes (A) output directly by ZDOCK and the (B) locally minimized structures. A dotted line is place at Cα RMSD 10 Å which shows the threshold for which given structure was deemed correct.

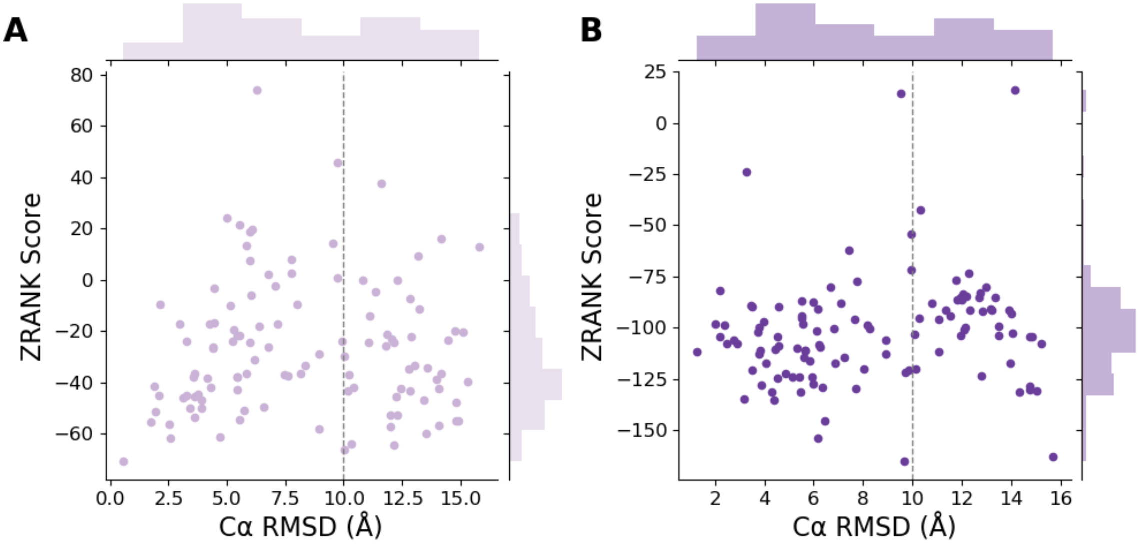


Figure 2.24 Ternary Structure Predictions Results for CRBN-dBET23-BRD4. The Cα RMSD of the predicted ternary complex to the known crystal structure vs. the predicted complex's ZRANK score is shown for the final complexes (A) output directly by ZDOCK and the (B) locally minimized structures. A dotted line is place at Cα RMSD 10 Å which shows the threshold for which given structure was deemed correct.

As the ternary structure prediction method was relatively successful in the prediction of ternary complexes containing the cereblon ligase, the method was applied for the prediction of the CRBN-PROCDK9_25-CDK9 complex. In addition to PROCDK9-25, predictions were also made for PROCDK9-21 and -24. Since the only difference between these compounds is the length of the linker, it is expected that the increased potency will be due to positive cooperativity from interactions at the protein-protein interface. As PROCDK9-25 is the longest and most potent, it may form a complex interface that is inaccessible by the shorter linker containing PROTACS. Therefore, complexes that were predicted to be formed by PROCDK9-25 that were also predicted to be formed by -21 or -24 were removed. Of the 22 predicted ternary complexes for PROCDK9-25, 11 were not shared by either -21 or -24. The complex which had the best ZRANK score, -70.63, among all 22 predicted complexes was also unique to PROCDK9-25 and is shown in Figure 2.25. A closer look at PROCDK9-25 spanning the binding pockets of cereblon and CDK9 is shown in Figure 2.26. In this prediction, PROCDK9-25 appears fully elongated, and it is apparent that the shorter linkers of -21 and -24 would be unable to span the same gap between the pockets.

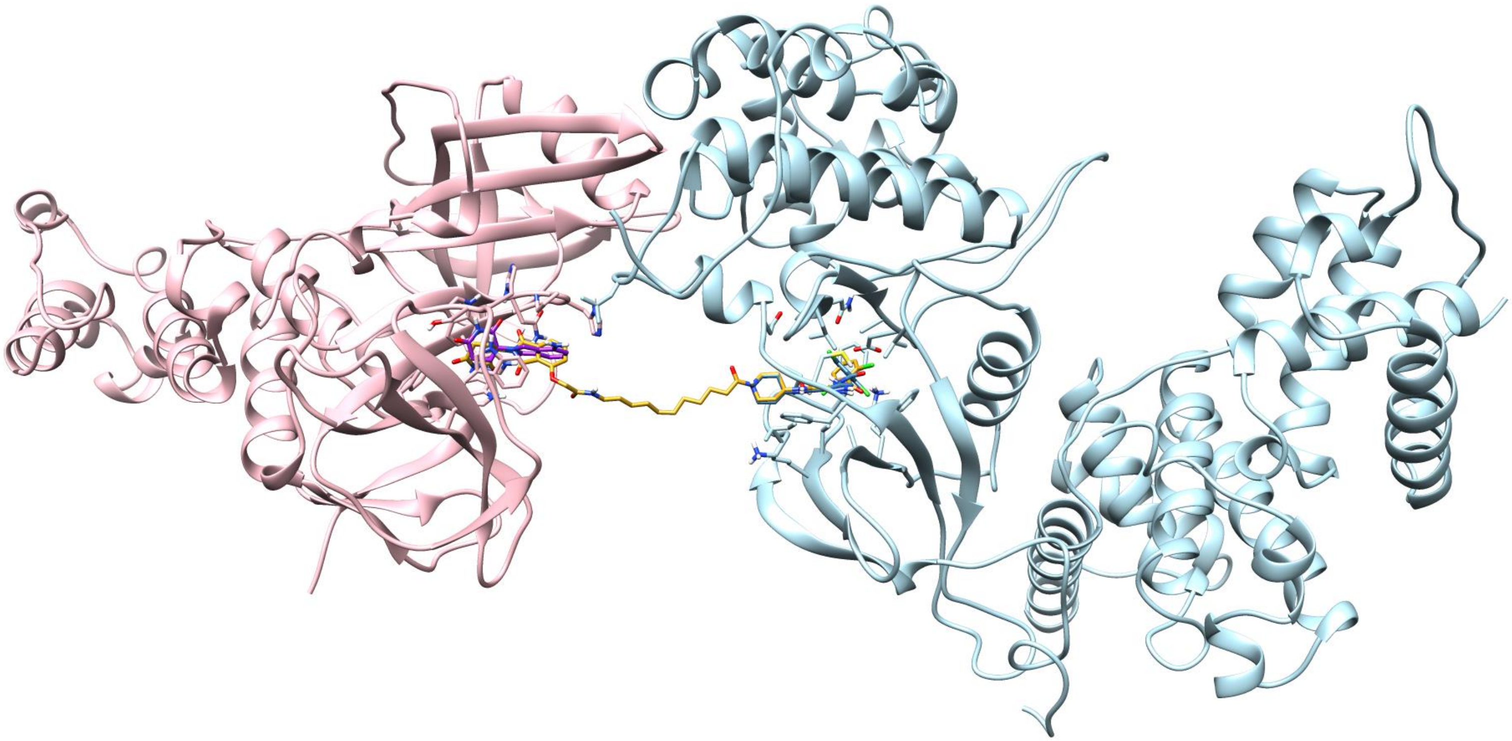


Figure 2.25 Ternary Structure Prediction for CRBN-PROCDK9_25-CDK9. The PROTAC PROCDK9-25 (gold) connects the binding sites of cereblon (light pink) and CDK9 (light blue). The crystallographic ligands for each, lenalidomide (purple) and AT7519 (teal), which were used as restraints for generating the PROTAC conformer, are also shown.

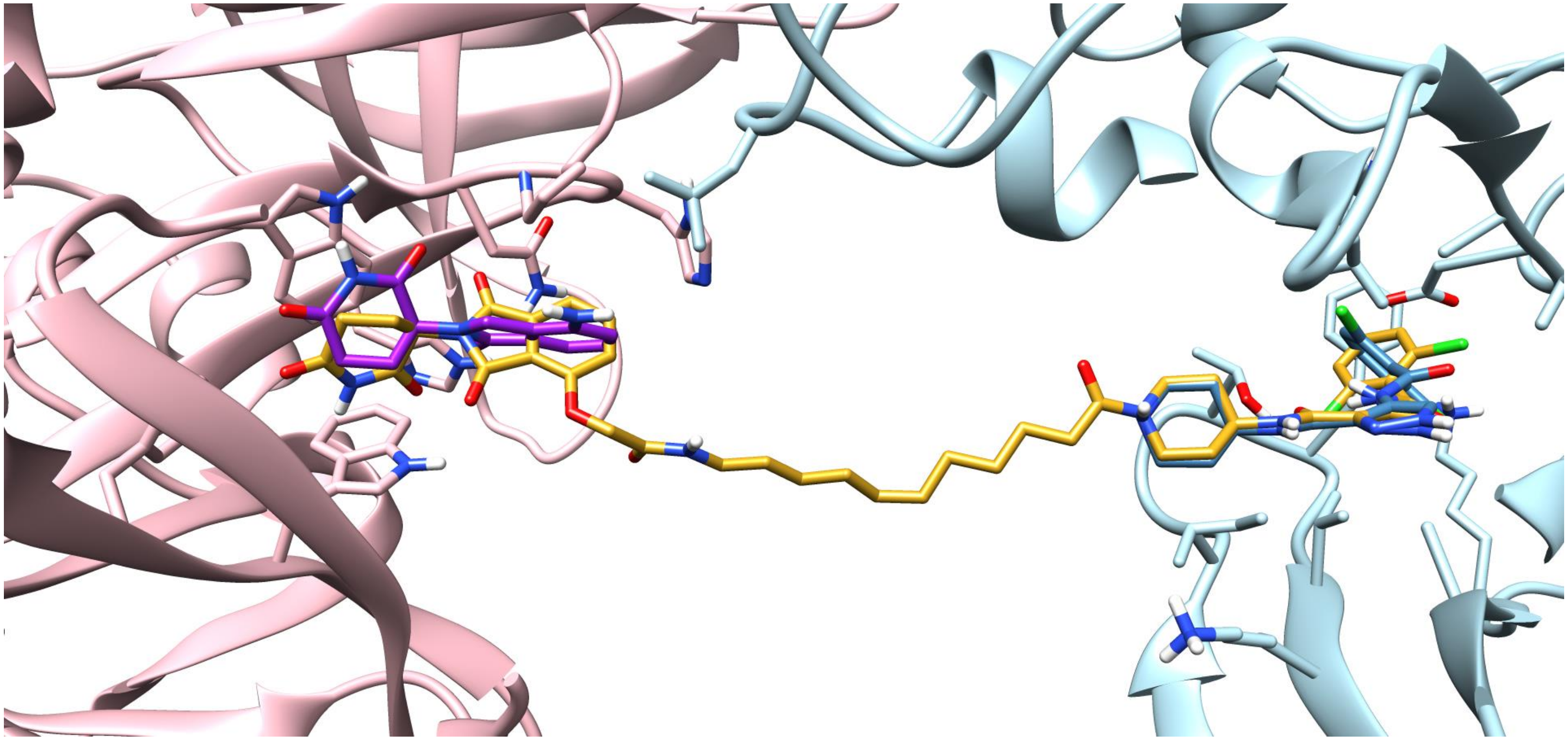


Figure 2.26 Binding Pockets in the Ternary Structure Prediction for CRBN-PROCDK9_25-CDK9. A closer look at the binding site of PROTAC PROCDK9-25 (gold) that connects the binding sites of cereblon (light pink) and CDK9 (light blue). The crystallographic ligands for each, lenalidomide (purple) and AT7519 (teal), which were used as restraints for generating the PROTAC conformer, are also shown. Some protein ribbons and side-chains have been hidden for easier visualization of each binding site.

2.5 Conclusions

PROTAC compounds were collected from the patent literature as a first step in understanding the chemical space of PROTACs. The molecular weight, LogP, LogD at pH 7.4, number of rotatable bonds, number of hydrogen bond donors, number of hydrogen bond acceptors, topological polar surface area, number of aromatic rings, and the AbbVie multi-parametric score of each patent PROTAC were calculated. As expected, analysis of these physicochemical properties demonstrated that PROTAC molecules generally lie well beyond rule of 5 chemical space. A comparison of PROTACs binding to the Von Hippel-Lindau (VHL) or cereblon (CRBN) ligases showed that PROTACs targeting CRBN are more drug-like than those that target VHL mainly due to smaller and more favorable properties of the CRBN binding moiety compared to the VHL binding moiety. The physicochemical properties calculated for patent CRBN PROTACs were used to give context to the development of CDK9 PROTACs. Comparisons between developed CDK9 PROTACs and patent PROTACs allowed for the identification of properties of CDK9 PROTACs which deviated significantly from the patent literature to highlight potentially problematic properties.

Furthermore, a method was developed to predict the PROTAC mediated ternary complex due to its value for rational PROTAC design. This method successfully produced correct predictions for three benchmark crystal structures of ternary complexes with a resolved PROTAC molecule. Ternary structure prediction with this method appeared to work better for ternary structures containing CRBN over VHL. However, while this method produced correct predictions, many incorrect predictions were also made. Attempts to definitively delineate between correct and incorrect predictions using the scoring function ZRANK were unsuccessful. For confident ternary structure predictions, further work is needed to improve the scoring of the ternary complexes and will also require a larger amount of crystallographic data for benchmarking.

.

Chapter 3. STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products

In accordance with the American Chemical Society's Policy on Theses and Dissertations, this chapter contains work reproduced with permission from:

Nicholas T. Cockroft, Xiaolin Cheng, James R. Fuchs. STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products, *J. Chem. Inf. Model.*, **2019** (manuscript submitted, Manuscript ID: ci-2019-00489s).

Copyright **2019** American Chemical Society.

3.1 Abstract

Target fishing is the process of identifying the protein target of a bioactive small molecule. To do so experimentally requires a significant investment of time and resources, which can be expedited with a reliable computational target fishing model. The development of computational target fishing models using machine learning has become very popular over the last several years due to the increased availability of large amounts of public bioactivity data. Unfortunately, the applicability and performance of such models for natural products has not yet been reported. This is in part due to the relative lack of bioactivity data available for natural products compared to synthetic compounds. Moreover, the databases commonly used to train such models do not annotate which compounds are natural products, which makes the collection of a benchmarking set

difficult. To address this knowledge gap, a dataset comprised of natural product structures and their associated protein targets was generated by cross-referencing 20 publicly available natural product databases with the bioactivity database ChEMBL. This dataset contains 5,589 compound-target pairs for 1,943 unique compounds and 1,023 unique targets. A synthetic dataset comprised of 107,190 compound-target pairs for 88,728 unique compounds and 1,907 unique targets was used to train k-nearest neighbors, random forest, and multi-layer perceptron models. The predictive performance of each model was assessed by stratified 10-fold cross-validation and benchmarking on the newly collected natural product dataset. Strong performance was observed for each model during cross-validation with area under the receiver operating characteristic (AUROC) scores ranging from 0.94 to 0.99 and Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) scores from 0.89 to 0.94. When tested on the natural product dataset, performance dramatically decreased with AUROC scores ranging from 0.70 to 0.85 and BEDROC scores from 0.43 to 0.59. However, the implementation of a model stacking approach, which uses logistic regression as a meta-classifier to combine model predictions, dramatically improved the ability to correctly predict the protein targets of natural products and increased the AUROC score to 0.94 and BEDROC score to 0.73. This stacked model was deployed as a web application, called STarFish, and has been made available for use to aid in the target identification of natural products.

3.2 Introduction

Experimental approaches for identifying small molecule hits in a drug discovery project typically include target-based screening or phenotypic screening. Target-based approaches involve selecting a protein target believed to be relevant to the disease state of interest and then measuring, directly or indirectly, a compound's ability to bind the target. Phenotypic approaches are target agnostic and instead measure a compound's effect on a biologically relevant system, such as cell cytotoxicity or tumor growth inhibition.[72] Both approaches are widely used in drug discovery and development. While traditionally viewed as opposing alternatives, target-based and phenotypic assays can also be complementary approaches.[73] An important limitation of the phenotypic approach is the inherent lack of understanding of the target and molecular mechanism of action. While a known target and molecular mechanism of action are not required to progress a new chemical entity to the clinic, it is considered a significant risk factor by most large pharmaceutical companies for the clinical development and regulatory approval process.[73] Due to the importance of target identification, both experimental and computational target fishing methods have been developed.

The process of experimental target fishing requires a significant investment of time and resources. One method commonly used to directly identify the target protein of a small molecule is biochemical affinity purification. This process involves immobilization of a compound on a column, exposure to cell extracts, stringent washing to remove non-specific binding, proteomic profiling to determine the identity of bound proteins, and ultimately a confirmatory binding assay.[74] While this process has been very successful, it is not without limitations.[75] For example, it requires the bioactive small molecule to be modified in order to be immobilized on the column. Points of modification can be difficult to determine as they require a synthetic handle in a region where a bulky linker can be attached without interfering with target binding. Overall, experimental target fishing requires a great deal of biological and synthetic expertise and effort.

In an effort to aid and accelerate the target identification process, a variety of computational target fishing methods have been developed. Computational target fishing methods generally fit into one of three broad categories: ligand-based, structure-based, or network-based. A recent review by Sydow et al. gives a good overview on the specifics and current methods applied for each category.[76] Ligand-based methods rely on the assumption that proteins will bind similar small molecules. A simple ligand-based target fishing approach typically involves computing Tanimoto similarities between a compound of interest and compounds with known targets in a bioactivity database. The protein targets of compounds with high similarity to a query compound are then predicted as potential protein targets. An early and successful ligand-based target fishing method, similarity ensemble approach (SEA), builds upon this approach by comparing a query molecule's similarity to a group of compounds of a potential target and assessing the statistical significance of the resulting similarity score.[77] The growing amount of publicly available bioactivity data in databases such as ChEMBL and PubChem has made the application of machine learning methods to computational target fishing popular.[78,79] Methods such as Random Forest (RF), Support Vector Machines (SVM), and Naïve Bayes (NB) have long been used in this regard, but deep learning methods have recently garnered significant attention due to their impressive performance.[80–84] The majority of the data that these models were trained with is synthetic compound bioactivity data and despite the impressive performance observed for these machine learning target fishing models, little is known about how they might perform when applied to natural products

Natural products have been a tremendous source of new drugs over the past three decades. Unaltered natural products and natural product derivatives comprise over one-third of the FDA approved small molecule drugs.[85] Taken even more broadly with the inclusion of "natural product mimics", natural products account for or have inspired in some way up to 60% of all of these approved drugs. Historically, natural products have made up a substantial portion of first-in-class drugs identified through phenotypic methods.[86] Therefore, the process of target identification is very important for natural products, but this area is currently underexplored. In 2017, Fang et al. used a network-based target fishing approach for natural product target prediction. A balanced substructure-drug-target network-based inference (bSDTNBI) model was trained and tested on 2,388 unique natural products and 751 targets.[87] However, most available binding data is for synthetic compounds and almost all target fishing models are trained using synthetic data. A study by Keum et al. in 2016 developed a target fishing model using the bipartite local model and support vector machines (SVM) trained on 3,612 compounds and 831 targets.[88] The trained model was used to predict the targets of 6,320 natural product compounds. Unfortunately, the protein targets for these natural products were unknown. Model predictions were examined based on whether a predicted target was implicated in the disease state for which a given herb, containing the natural product, was associated with. Ultimately, how well a model trained on synthetic data can predict targets for natural products remains unknown.

To address this, a stacked ensemble target fishing (STarFish) approach has been developed and was benchmarked on a newly collected natural product set that considers the largest number of protein targets for a natural product dataset so far. Model stacking is a popular and successful approach in Kaggle competitions and has also been recently applied to other areas of cheminformatics.[89–92] This model stacking approach expands upon the idea that the combination of model predictions can produce better predictions than individual models alone. In this study, different combinations of stacked classifiers are trained on a large synthetic data training set comprised of 107,190 compound-target pairs for 88,728 unique compounds and 1,907 unique targets. The trained stacked classifiers are subsequently evaluated through cross-validation on the synthetic compound dataset and through benchmarking on the newly collected natural product dataset comprised of 5,589 compounds-target pairs for 1,943 unique compounds and 1,023 unique targets. Furthermore, a multi-label classification approach is taken. Historically, computational target-fishing models have been trained under the assumption that a single molecule binds a single protein, but in recent years more emphasis has been placed on the consideration of polypharmacology during training.[93] Therefore, the individual models which comprise the stacked model are trained on a multi-label classification problem to account for this polypharmacology. Overall, a web application, STarFish, was developed, which makes predictions on natural targets based on small molecule binding to 1,907 targets. The datasets, source code, and API are freely available for download and use at: https://github.com/ntcockroft/STarFish and in Appendix B. Datasets and Code.

3.3 Methods

3.3.1 Dataset

Natural product compound records were extracted from the following freely accessible datasets/databases: AfroCancer[94], AfroDb[95], AfroMalaria[96], AnalytiCon[97], Carotenoids[98], ConMedNP[99], InterBioScreen(IBS) natural product collection[100], Mitishamba[101], NANPDB[102], Natural Product Atlas[103], NPACT[104], NPASS[105], NuBBE[106], p-ANAPL[107], SANCDB[108] Super Natural II[109], TCM[110], TIPdb[111], UNPD[112], and ZINC natural product subset[113]. The compounds from each database were retrieved in various chemical formats and were ultimately converted to simplified molecular-input line-entry system (SMILES) strings if not provided. All the provided or generated SMILES strings were cleaned and standardized using MolVS.[114] The resulting combined set contained 438,258 unique natural products in total. Since the majority of the natural product databases listed do not have bioactivity annotations, the compound set was cross-referenced with the ChEMBL database (version 23)[78] to identify natural product compounds with known protein targets.

The ChEMBL database was queried to retrieve compound activity records which had reported activities (IC50, Ki, Kd, EC50) of 1 µM or better in assays with a confidence score of 9 and had a known target with a corresponding UniProt ID. A confidence score of 9 was selected so that only assay data that resulted in a single protein target being assigned with a high degree of confidence were used. This query yielded a dataset of 485,813 compound-target activity pairs with many redundant activity pairs. The SMILES strings in this dataset were cleaned and the natural product dataset was standardized using MolVS. Following standardization, the ChEMBL dataset was used to determine protein targets for the natural product dataset by identifying InChIKeys or SMILES that were present in both datasets. After this comparison, any redundant compound-target pairs were removed, which yielded two datasets: a "synthetic" set consisting of 395,590 unique compound-target pairs and a natural product set consisting of 6,339 unique compound-target pairs.

The synthetic set was pruned further prior to model training. Only compound-target pairs containing targets with at least 10 compounds were kept. Furthermore, the number of compounds per target class was capped at 100 through random sampling to limit the imbalance between protein target classes. This pruning resulted in 107,190 compound-target pairs for 88,728 unique compounds and 1,907 unique targets. A breakdown of the target protein classes present in these 1,907 unique targets is shown in Figure 3.1. For the natural product dataset, compound-target pairs that contained protein targets in common with the pruned synthetic set (Figure 3.2) were retained resulting in 5,589 compound-target pairs for 1,943 unique compounds and 1,023 unique targets. The synthetic set was used for model training and cross-validation while the natural product set was used to benchmark model performance on a more realistic but difficult test case.

All compound-target activity pairs were converted to a multi-label format. For each compound record, a binary label vector was constructed that annotated the protein targets to which these compounds are known to bind. On average, compounds in the synthetic compound dataset have 1.2 annotated protein targets per compound and compounds in the natural product dataset have 2.9 annotated protein targets per compound. Therefore, while 1,907 possible target associations are considered for each compound, these associations have not all been tested experimentally, and most are unknown. It was assumed that for these unknown cases that the compound did not bind to the protein and this unknown data was treated as negative data. According to a recent estimate of drug polypharmacology, drugs have on average 11.5 targets below 10 µM.[115] Applying this estimate to the unknown small molecule compound-target associations implies a negative label would be correct for 99% of the labels. However, it is likely that many compound records will be assigned a negative protein target label for a protein that they actually interact with. This will influence training as strong classifier predictions for such labels would be penalized. Additionally, during performance evaluation such labels are considered false positives and negatively impact performance when ranked highly. While the assumption of negative labels for unknown compound-protein target records is reasonable, there are indeed drawbacks.

The use of a multi-label format does not fully capture the ways in which compounds can interact with protein targets. While a set of ligands may all be reported to bind to a common protein, the compounds may bind at different sites or have different effects on the protein. For example, one compound may bind to a catalytic site on the protein while another binds to an allosteric site. Additionally, two compounds may bind to the same site on the protein, but one may be an agonist while the other an antagonist. In the multi-label format described here, these pharmacological differences are ignored, and all compound-protein interactions are treated as equivalent. Therefore, a classifier trained using ligands that bind to the catalytic site would be expected to perform poorly when used to predict the target of compounds which bind to the allosteric site of the same protein.

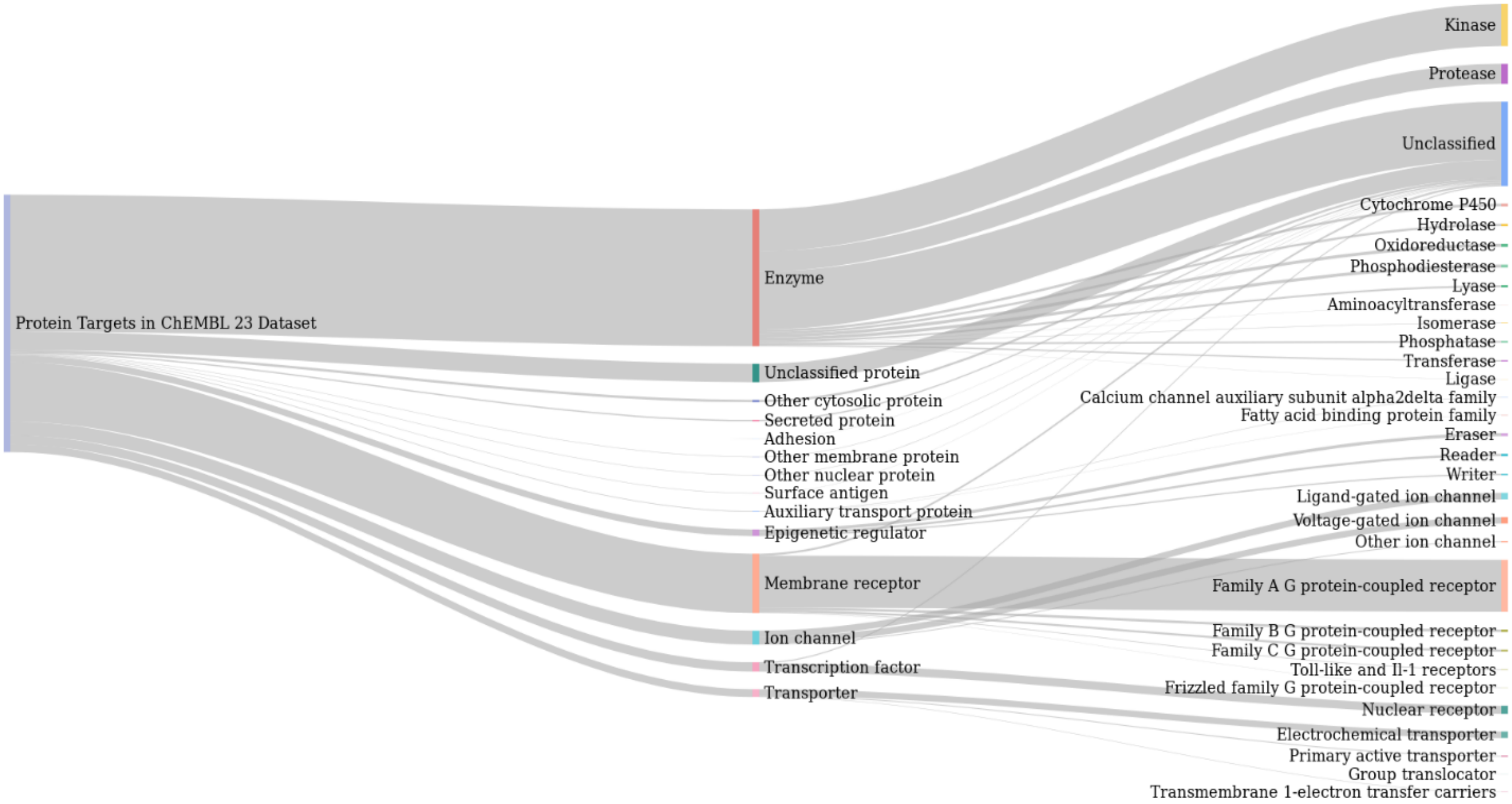


Figure 3.1 Protein Class Labels. Sankey diagram of the protein classes present in the ChEMBL23 activity data used in model training. The proportion of protein targets belonging to L1 and L2 protein classes as defined by ChEMBL is represented by line thickness.

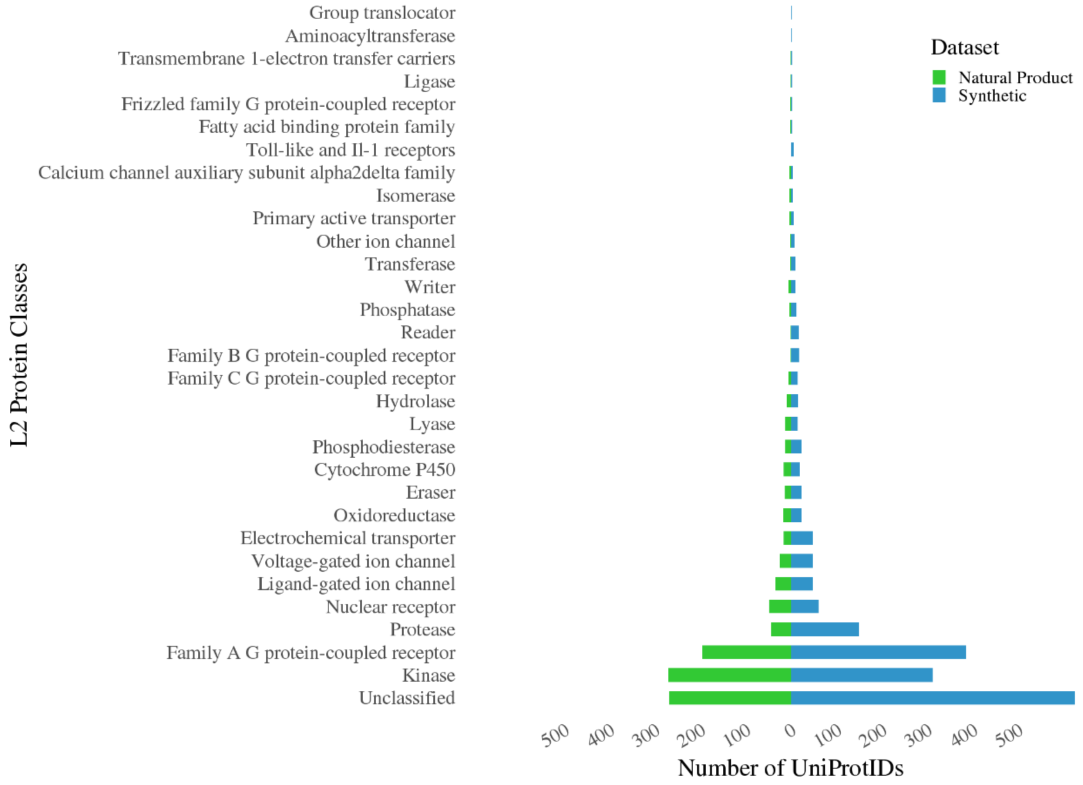


Figure 3.2 Synthetic and Natural Product Dataset Protein Classes. Comparison of the ChEMBL L2 protein classes between the synthetic compound dataset and the natural product dataset.

3.3.2 Compound Descriptors

RDKit was used to generate molecular fingerprints for each compound.[56] Molecular fingerprints are bit vector representations of a compound. A kernel is applied to a molecule to extract chemical features, hash them, and set bits based on the hash. If two compounds contain the same functional group, they will both set a bit for that functional group. However, more than one functional group can set the same bit resulting in collisions. Increasing the number of bits used to represent molecules reduces collisions but increases the computational cost of working with the fingerprint. The SMILES string for each compound was converted to a 2048-bit Morgan Functional-Class Fingerprint (FCFP) using a radius of 2. FCFP was selected over Extended-connectivity fingerprints (ECFP) to generate a more abstract and pharmacophoric representation of each compound.[116]

3.3.3 Machine Learning Models

All models were built using Scikit-Learn 0.19.1 in Python 3.6.5.[117] Since compounds can bind to more than one target protein, compound-target identification was formulated as a multi-label classification problem. Different classification models handle multi-label classification problems differently and therefore how each handle multi-label problems are addressed specifically for each classifier. Additionally, each classifier was asked to predict label probabilities instead of assigning labels directly.

3.3.4 k-Nearest Neighbors

The k-nearest neighbors (KNN) algorithm is a type of instance-based learning and computes the distance between the query point and the training instances to determine the closest k points. The KNN classification scheme is easily applied to a multi-label format. In a multi-label case, the query point is assigned the class labels of the closest k points with the probability of each label corresponding to the simple average of label counts over k points. These probabilities can also be weighted by the distance of each training instance to the query point. The KNN model used herein was trained using 10 neighbors, brute force distance calculations with the Jaccard metric, and uniform weights.

3.3.5 Multi-layer Perceptron

A multi-layer perceptron (MLP) is a class of feedforward artificial neural networks that consists of at least three layers: an input layer, a hidden layer, and an output layer. Each layer consists of a set of neurons. In the input layer, the number of neurons is set to the number of features for a record in the training data. In the case of the 2048 bit Morgan fingerprint, the number of neurons in the input layer is 2048; one neuron for each bit. When used for classification, the number of neurons in the output layer corresponds to the number of class labels, in this case one neuron per protein target, and is inherently applicable to multi-label problems. The MLP classifier used herein consists of a single hidden layer with 1000 neurons and ReLU activation function. A stochastic gradient-based optimizer referred to as “Adam” was the solver used for weight optimization with an initial learning rate of 0.001, an exponential decay rate of 0.9 and 0.999 for the first and second moment vectors respectively, and the constant for numerical stability set to 1e-8. The maximum number of iterations was set to 200 with a convergence tolerance of 1e-4 after 2 consecutive iterations.

3.3.6 Random Forest

Random forests are an ensemble of decision trees that can be used for either classification or regression. While inherently applicable to multi-label problems, there are technical limitations, such as memory consumption, when training with a large amount of high-dimensional data and trying to predict a large number of class labels. To circumvent this issue, the multi-label problem was re-cast as many individual binary classification problems. In the multi-label learning literature, this strategy is referred to as one-vs-the-rest or binary relevance. Therefore, a random forest model was trained for each label and to predict whether that label should be assigned or not. A total of 1,907 random forest models were trained, one for each protein target, using 1,000 trees and 45 features were considered when looking for the best split.

3.3.7 Logistic Regression

Despite the name, logistic regression is used for classification and can be applied to binary, multinomial, and ordinal classification problems. Logistic regression is a linear method, however, the output of the linear combination of features is bounded between 0 and 1 by using a logistic function. To apply logistic regression to a multi-label classification problem, the one-vs-the-rest strategy described above must also be applied here. Logistic regression models were trained using the "liblinear" solver and L2 regularization. A total of 1,907 logistic regression models were trained, one for each protein target, with C =1.0.

3.3.8 Model Stacking

Model stacking, also referred to as stacked generalization or meta ensembling, is a method which combines information from base models to generate a new model. A stacking approach takes advantage of the fact that individual models may have different strengths in label prediction compared to others and attempts to improve prediction through their combination. During stacking, the input features, in this case the level 0 data, is passed to all individual base models, the level 0 classifiers, which yield predicted probabilities for each individual label. These predicted label probabilities, the level 1 data, are then used as the input features for the next model, the level 1 classifiers. Although this process can continually be repeated, only two levels were used for the stacked model described here as shown in Figure 3.3.

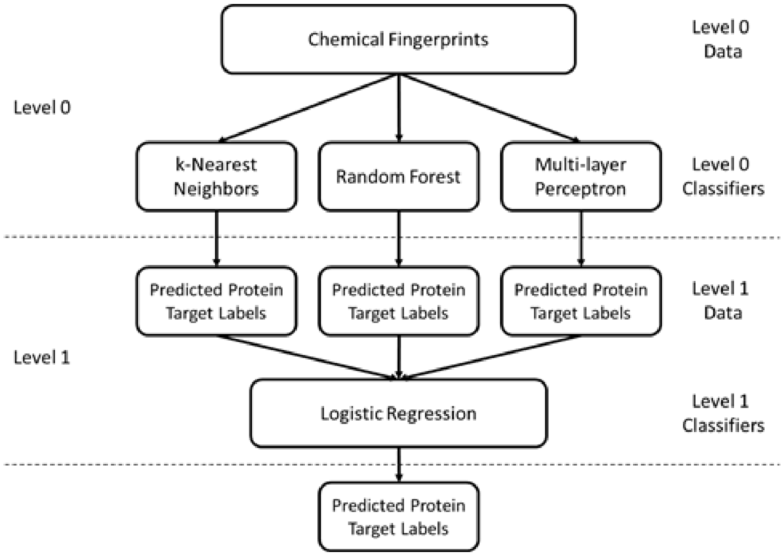


Figure 3.3 Model Stacking Schematic. Diagram of the model stacking approach used to predict protein target labels from chemical fingerprints. Chemical fingerprints are used as input features for the level 0 classifiers: k-nearest neighbors, random forest, and multi-layer perceptron. The predicted probabilities of each protein label from each level 0 classifier are concatenated and used as input features for the level 1 classifier: logistic regression. Final predicted label probabilities are output by the logistic regression.

3.3.9 Model Tuning, Training, and Validation

The synthetic dataset was used for model tuning, training, and testing whereas the natural product set was used as an external test set. A stratified 10-fold cross-validation was performed on the synthetic dataset resulting in 10 folds of 90/10 split training/testing sets. The stratification process guaranteed that examples for each label were present in both the training and test cross-validation datasets. Parameters for k-nearest neighbors, random forest, and multi-layer perceptron models were tuned using the training sets for each fold. A stratified random split was used to further subdivide the training data portion of each cross-validation fold into 90/10 training/test sets for tuning. Parameters were chosen based on performance on the test tuning set (Table 3.1-Table 3.3) and were then used to train all subsequent models. Following evaluation by cross-validation, the entire synthetic set was used to train models which were evaluated on the natural product dataset.

Models were trained and tested using High Performance Computing resources from the Ohio Supercomputer Center.[118] Cross-validation and model combination calculations were run in parallel on the Owens cluster dense compute nodes (Dell PowerEdge C6320 two-socket servers with Intel Xeon E5-2680 v4 Broadwell, 14 cores, 2.40GHz processors, 128GB memory).

Table 3.1 Results of KNN base classifier tuning on a stratified random 90/10 train/test split of the training dataset for each cross-validation fold.

metric	n_neighbors	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage
jaccard	1	0.796388 (0.00176)	0.780529 (0.002284)	0.617222 (0.00322)	0.594705 (0.00355)	0.612085 (0.003563)	0.583117 (0.004359)	744.301234 (6.726699)
jaccard	5	0.923237 (0.001405)	0.909096 (0.002784)	0.868757 (0.00295)	0.844148 (0.002442)	0.853131 (0.002763)	0.826756 (0.005218)	259.071313 (5.575081)
jaccard	10	0.940646 (0.000933)	0.924678 (0.001974)	0.899521 (0.002092)	0.875196 (0.002173)	0.885127 (0.001885)	0.855073 (0.003874)	194.402763 (2.762402)
minkowski	1	0.794396 (0.002119)	0.779087 (0.002493)	0.613079 (0.00411)	0.590713 (0.00434)	0.608348 (0.004196)	0.580374 (0.00478)	752.279818 (8.264903)
minkowski	5	0.92039 (0.001419)	0.906489 (0.002227)	0.86295 (0.002556)	0.838582 (0.002263)	0.847581 (0.002752)	0.821665 (0.004188)	270.140751 (5.016876)
minkowski	10	0.937781 (0.001126)	0.922548 (0.00218)	0.89363 (0.002057)	0.869434 (0.002147)	0.87948 (0.002298)	0.85074 (0.004148)	205.32594 (3.692552)

Table 3.2 Results of MLP base classifier tuning on a stratified random 90/10 train/test split of the training dataset for each cross-validation fold.

hidden_layer_sizes	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage
(100,)	0.983686 (0.001387)	0.984229 (0.000853)	0.871111 (0.004238)	0.847315 (0.004101)	0.919742 (0.004053)	0.917158 (0.002717)	35.779642 (2.875651)
(1000, 100)	0.985217 (0.001098)	0.982851 (0.000956)	0.85854 (0.003482)	0.834136 (0.003386)	0.914784 (0.003445)	0.911042 (0.002612)	30.289004 (1.233247)
(1000, 1000)	0.980335 (0.001652)	0.978336 (0.001368)	0.862274 (0.003771)	0.838016 (0.004102)	0.901066 (0.005523)	0.896809 (0.004918)	30.565095 (1.180071)
(1000, 1000, 100)	0.982636 (0.001496)	0.979935 (0.001754)	0.846172 (0.003339)	0.822105 (0.003833)	0.900011 (0.00702)	0.895426 (0.00801)	30.411428 (0.740376)
(1000, 1000, 1000)	0.982331 (0.00317)	0.981026 (0.002729)	0.848929 (0.005684)	0.824962 (0.005778)	0.896086 (0.014113)	0.890992 (0.015034)	29.929988 (2.233989)
(1000,)	0.978462 (0.002088)	0.97966 (0.001521)	0.884938 (0.002679)	0.860374 (0.002652)	0.916133 (0.003451)	0.914374 (0.003722)	38.572175 (3.145195)

Table 3.3 Results of RF base classifier tuning on a stratified random 90/10 train/test split of the training dataset for each cross-validation fold.

max_features	n_estimators	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage
0.333	10	0.889216 (0.001681)	0.876576 (0.001867)	0.80111 (0.003377)	0.776924 (0.003388)	0.787611 (0.003383)	0.764548 (0.003436)	388.936712 (6.855073)
0.333	100	0.930415 (0.001785)	0.914428 (0.002373)	0.872884 (0.003036)	0.848404 (0.003051)	0.863443 (0.003669)	0.833557 (0.004559)	231.333976 (6.515776)
auto	10	0.895672 (0.001981)	0.881752 (0.002302)	0.813003 (0.00384)	0.788617 (0.003898)	0.798712 (0.003663)	0.773639 (0.00409)	364.072808 (7.762124)
auto	100	0.9454 (0.001129)	0.927314 (0.002217)	0.893945 (0.002532)	0.869258 (0.002393)	0.889633 (0.002503)	0.855677 (0.004295)	173.219298 (3.652797)
auto	1000	0.965165 (0.000616)	0.946285 (0.002496)	0.911903 (0.001612)	0.887386 (0.001935)	0.917467 (0.001513)	0.88161 (0.004596)	99.560729 (2.017088)

3.3.10 Area Under Receiver Operating Characteristic Curve (AUROC)

A common metric used to assess the performance of a classifier is the receiver operating characteristic (ROC) curve. Classifier predicted class probabilities, confidence values, or binary decisions are compared to the known labels. The fraction of true positives correctly recovered, the true positive rate, is plotted against the fraction of true negatives that were incorrectly identified as positive, the false positive rate. The true positive and false positive rates vary with the threshold used to split records by their probability or confidence scores into the positive and negative classes. Therefore, the true positive and false positive rates are plotted at various thresholds. The ROC curve can be summarized by a single value by calculating the area under the ROC curve. An AUROC score is represented by a value between 0 and 1, where a score of 1 denotes perfect classification, a score of 0.5 denotes random classification, and a score of 0 denotes completely incorrect classification. In general, the AUROC value can be interpreted as the probability of an active being ranked before an inactive. The AUROC score is designed for binary classification problems but can be easily extended to multi-label classification problems by averaging over the labels. This averaging can be done through either micro- or macro-averaging. In micro-averaging, each record-label pair contributes equally to the overall score and essentially treats all labels as a single combined binary classification problem. In macro-averaging, the binary AUROC is calculated for each label and then averaged. Therefore, each label contributes equally regardless of the number of records contained.

3.3.11 Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC)

While the AUROC score is a widely used and intuitive metric, it is not sensitive to early recognition. Early recognition is particularly important for target fishing problems as it is only feasible to run confirmatory experimental tests for a relatively small number of protein targets. In 2007, Truchon and Bayly proposed a metric called the Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) to address this early recognition problem, and it has become a popular metric for assessing virtual screening performance.[119] Similar to AUROC scores, a BEDROC score is between 0 and 1 and it has a probabilistic interpretation. However, while AUROC relates to a uniform distribution, BEDROC relates to an exponential distribution. These distributions can be considered as reference ranked lists. When a trained classifier makes predictions for a protein target label, it ultimately produces a sorted list of compounds ranked by the classifier's confidence in a compound binding to the protein target. The AUROC or BEDROC score that this classifier sorted list receives is the probability that a known active compound randomly selected from the classifier sorted list would be ranked higher than an "active" compound randomly selected from the reference list. For the AUROC score, this reference list is random and contains "active" and "inactive" compounds uniformly distributed throughout the list. For the BEDROC score, this reference list contains a large portion of "active" compounds at the beginning of the list. When calculating the BEDROC score a parameter α is required, which controls how highly "active" compounds are ranked in the reference list. For BEDROC scores to be comparable, they must use the same α value. The commonly used value is α=20 and was also used here. This α value indicates that 80% of actives are present in the first 8% of the list.

3.3.12 Fraction of Compounds with a True Target in the Top 10 Predictions

Because target fishing is concerned with the identification of a protein target for a given compound record, the fraction of compounds for which at least a single true target was identified in the top 10 of the ranked list was calculated. As with the BEDROC score, this score is concerned with early retrieval, however, an arbitrary cutoff of 10 predictions is used and differences in classifier performance after this cutoff will be missed. For example, a correct prediction at rank 11 is no better than a correct prediction at rank 1000 according to this metric since only correct predictions from ranks 1-10 are rewarded. Additionally, this differs from the other metrics described as both AUROC and BEDROC scores were calculated from the target protein label perspective while this is calculated from the compound perspective. A cutoff of 10 targets was selected as a being a feasible number of protein targets that could be screened. This score is relatively harsh as it requires a classifier to have placed a correct target for a compound in the top 0.5% of the list in order to be rewarded but gives an indication for the practical utility of a model for target fishing.

3.3.13 Coverage Error

The coverage error is a metric that is also calculated from the compound record perspective and determines on average how far down the classifier sorted list one would need to look in order to recover all true labels. The best possible value for this metric is the average number of labels for each compound record.

3.4 Results and Discussion

3.4.1 Natural Product Databases

There are many natural product databases or datasets that are published and available online. These databases range in size from a few hundred compounds to hundreds of thousands of compounds. A review from 2017 by Chen, Kops and Kirchmair gives a good overview of both virtual and physical natural product compound libraries.[120] Many databases have a particular bioactivity focus, such as anticancer or antimalarial activities, and a focus on the geographical region from which the natural products were obtained. The smaller databases tend to have a narrow focus while the large databases attempt to aggregate and organize all known natural products, leading to significant overlap. The size and overlap of the natural product databases are shown in Figure 3.4. Prior to comparison, SMILES strings were standardized for each database and only unique compounds were retained, which accounts for any discrepancies between the number of compounds shown here and the published database sizes. No single database contains all of the 438,258 unique natural products that were collected. The Super Natural II database is the largest and contains 52.7% of the collected natural products. The top 5 largest databases, which include Super Natural II, Universal Natural Product Database (UNPD), ZINC Natural Products Subset, InterBioScreen (IBS) Natural Compounds, and Traditional Chinese Medicine (TCM) Database@Taiwan comprise 86.4% of the collected natural products.

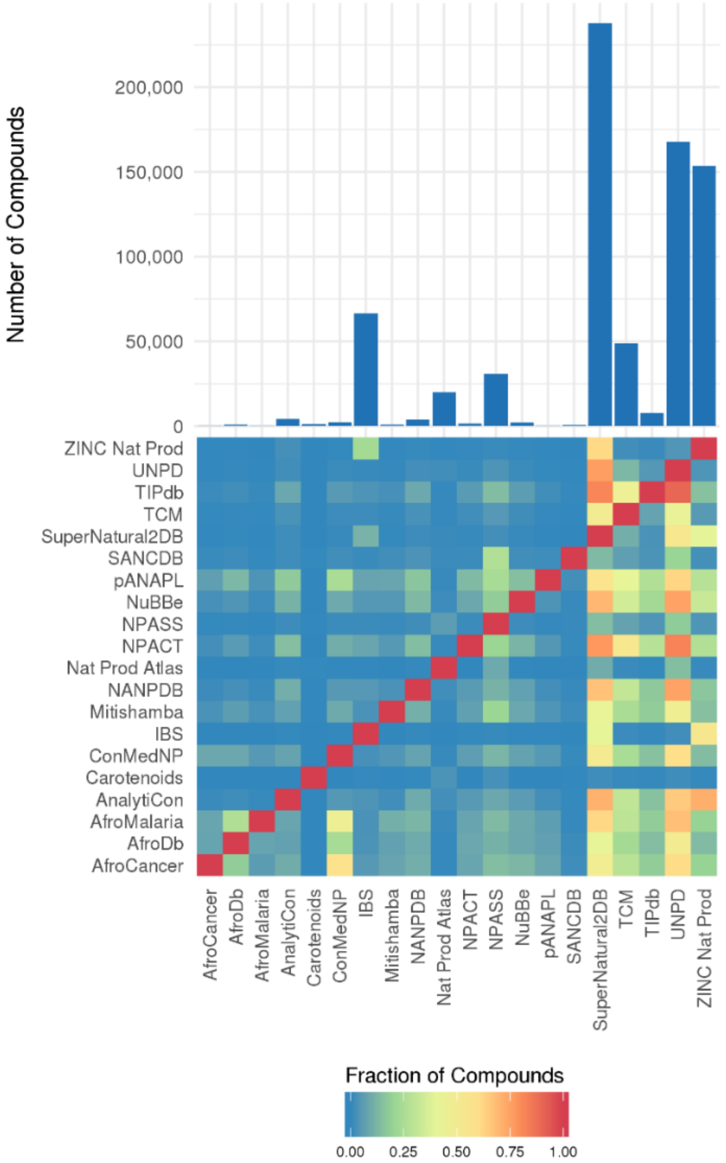


Figure 3.4 Natural Product Databases. Size and overlap of collected natural product databases. The bar graph on the top shows the number of unique compounds in each database. The heat map shows the fraction of compounds from a database on the y-axis present in a database on the x-axis.

3.4.2 Synthetic Cross-Validation

Prior to benchmarking on the collected natural product data, models were trained and evaluated with the synthetic dataset using stratified 10-fold cross-validation. Overall, all trained models performed extremely well (Figure 3.5). Without stacking micro-averaged AUROC values ranged from 0.94 to 0.99, micro-averaged BEDROC values ranged from 0.89 to 0.94, and 89% to 92% of compounds had a true target identified in the top 10 predictions. In general, performance slightly improved when stacked. With stacking micro-averaged AUROC values ranged from 0.97 to 0.99, micro-averaged BEDROC values ranged from 0.89 to 0.97, and 85% to 95% of compounds had a true target identified in the top 10 predictions. Coverage error showed more distinct differences between different models and how stacking impacted performance. Without stacking coverage error ranged from 187 to 29 labels. Unlike the other described metrics, a lower value is better for coverage error as it represents the average number of labels that need to be considered to recover all the true labels. With stacking, this generally improved to 55 to 14 labels. The only machine learning model that did not benefit from stacking was the multilayer-perceptron (MLP).

For each metric, the unstacked MLP performs better than the stacked MLP. The performance degradation is likely due to overfitting. To assess this, the penultimate layer activations of the MLP were collected and normalized. The 1,000 activations were passed to the logistic regression as features in the place of the predicted target labels. Essentially, the output layer is being removed after training the MLP. As shown in Figure 3.6, this process mitigates the performance degradation observed when using MLP classifier in the stacked model, supporting that the degradation was a result of overfitting to the training data.

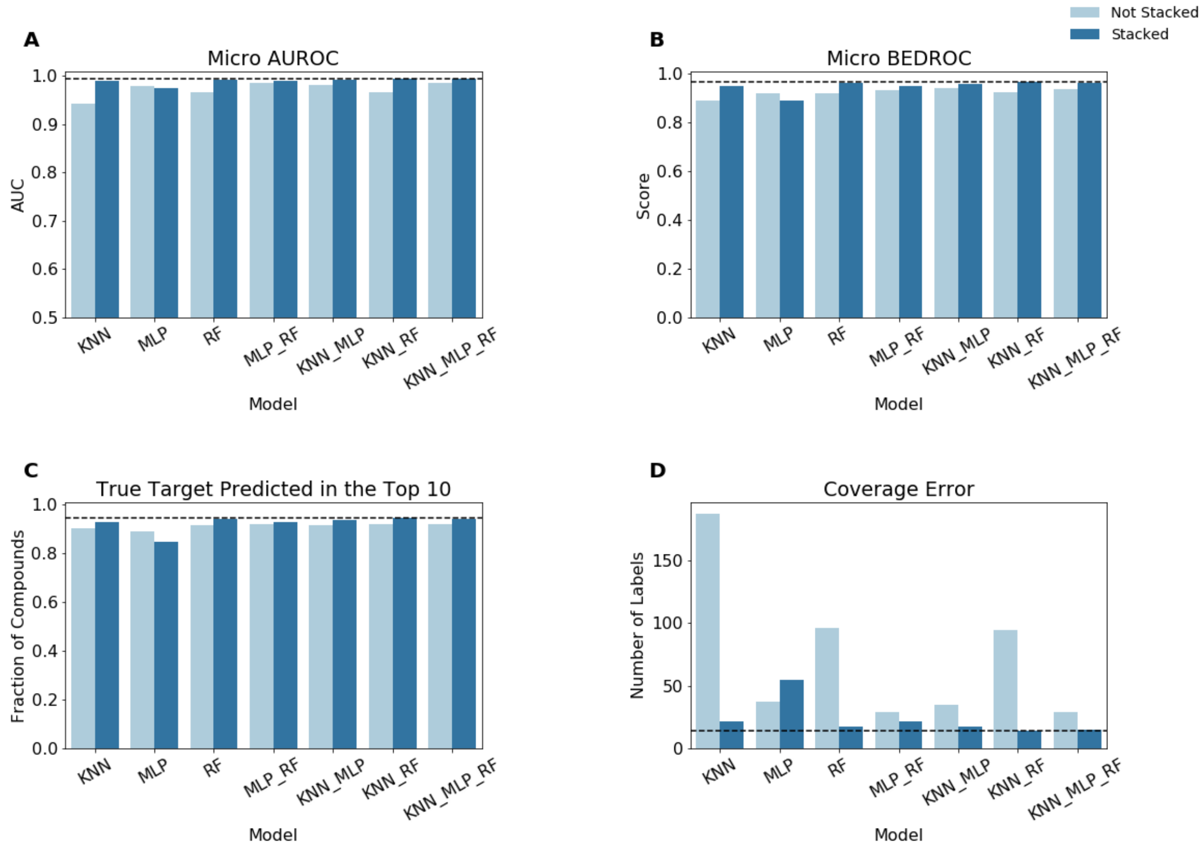


Figure 3.5 Synthetic Compound Cross-Validation Performance. Model performance for stratified 10-fold cross-validation on the synthetic compound dataset. For a single model, “Not Stacked” indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which have at least one true target among the top 10 predictions, and (D) coverage error are shown.

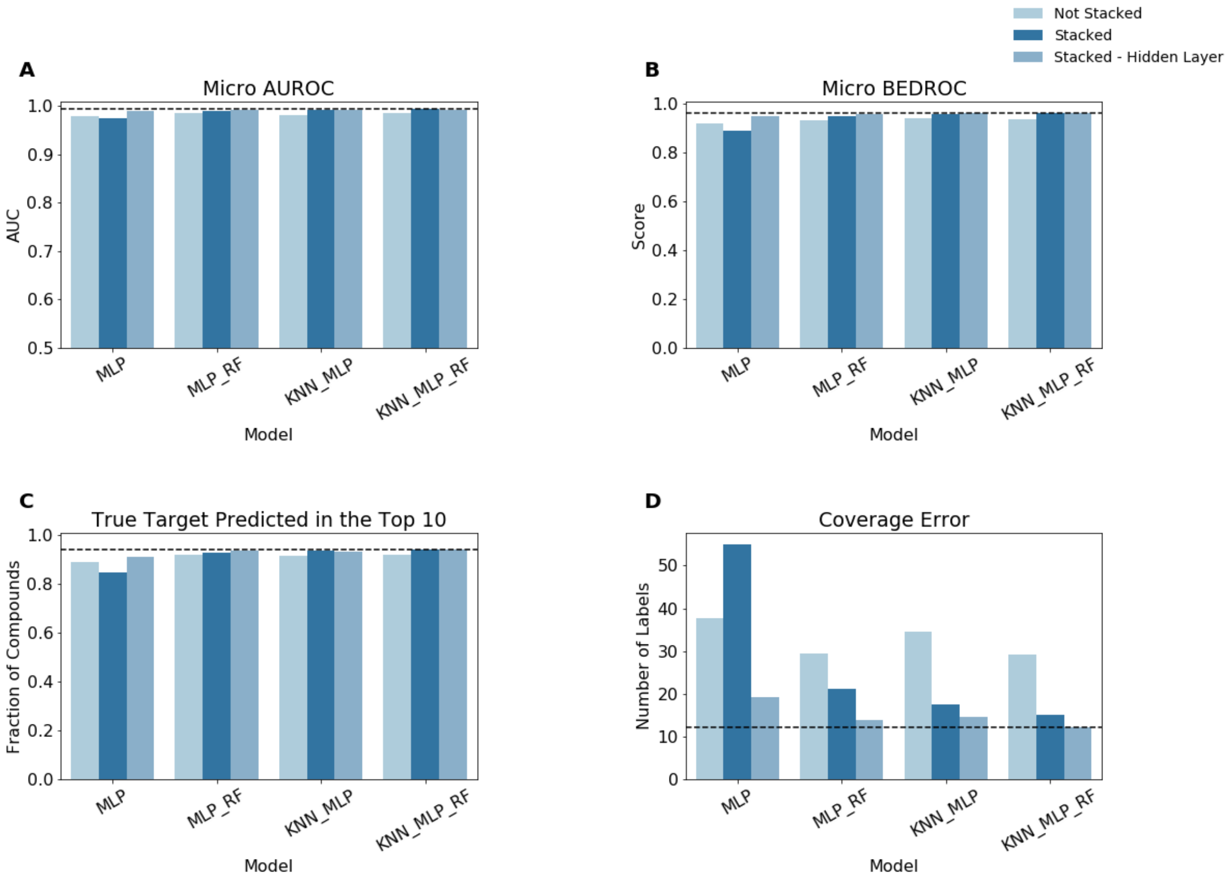


Figure 3.6 Synthetic Dataset Performance with MLP Hidden Layer Features. Model performance for stratified 10-fold cross-validation on the synthetic compound dataset. Classifier combinations using the MLP classifier are shown. “For a single model, “Not Stacked” indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. “Hidden Layer” indicates that the normalized penultimate layer activations were used as input features for the logistic regression instead of the predicted labels during model stacking. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which have at least one true target among the top 10 predictions, and (D) coverage error are shown.

While the performance measured for cross-validation is exemplary, it is undoubtedly an overly optimistic estimate of model performance for a prospective application. When using a random split cross-validation approach there is often redundancies between compounds present in the training and test folds. Therefore, predictions may be made on analogues of compounds that the model was trained on, which makes the prediction of the correct target for that compound a very easy problem. Methods such as temporal split validation or clustering techniques can be used to generate more dissimilar training and testing splits to offer more realistic performance estimates.[76,121,122] However, doing so requires removing activity data points and ultimately reducing the number of targets that can be considered. Consideration of a large number of targets is important to the utility of a computational target fishing method, because the method can only predict for targets it has been trained on. Despite the limitations of random splitting, other splitting techniques were not used in order to include as many target protein labels as possible.

Assessment on the natural product benchmark is expected to give a less optimistic and more realistic performance estimate. To demonstrate the difference between synthetic and natural product compounds, similarities between cross-validation training and test sets, in addition to natural product compounds, were assessed. For each protein target label, pairwise Tanimoto similarities were calculated between the training compounds themselves, training compounds with test compounds, and training compounds with the natural product benchmark compounds. The cumulative density function (CDF) plotted for each pairwise similarity distribution is shown in Figure 3.7. The CDFs for the synthetic training and test sets are nearly identical. Overall, the test compounds are very similar to the training compounds and thus the good model performance observed is expected. On the other hand, the natural products are less similar and performance on this benchmark is expected to be a better indicator of realistic performance.

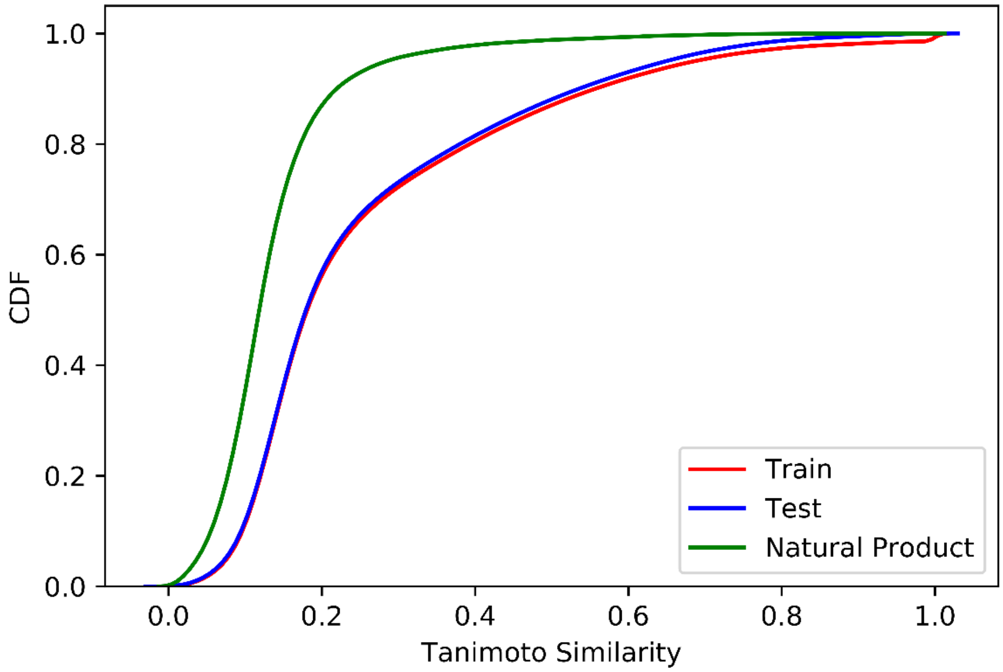


Figure 3.7 Cumulative Density Function (CDF) of Intra-Target Compound Similarities. All pairwise compound similarities were calculated between the training compounds and a given set for each protein target label. “Training” and “Test” sets are from a single cross-validation fold and “Natural Product” is the natural product benchmark set.

3.4.3 Natural Product Benchmark

Following cross-validation, new models were trained using the entirety of the synthetic compound dataset and predictive performance was assessed for the natural product benchmark. As expected, predictive performance decreased for the natural product benchmark, especially for unstacked models (Figure 3.8). Without stacking micro-averaged AUROC values ranged from 0.70 to 0.85, micro-averaged BEDROC values ranged from 0.43 to 0.59, 55% to 60% of compounds had a true target identified in the top 10 predictions, and coverage error ranged from 1286 to 416. In general, model performance greatly improved when stacked. With stacking micro-averaged AUROC values ranged from 0.82 to 0.94, micro-averaged BEDROC values ranged from 0.45 to 0.73, 43% to 63% of compounds had a true target identified in the top 10 predictions, and coverage error ranged from 426 to 190. As observed in cross-validation, MLP stacked models appeared to suffer from overfitting resulting in performance degradation. While the micro-averaged AUROC value slightly increased for the MLP stacked model, all other metrics showed a performance decrease. Modifying the stacked MLP model to use the normalized penultimate layer activations as inputs to the logistic regression did not rescue performance to the same extent as with cross-validation (Figure 3.9).

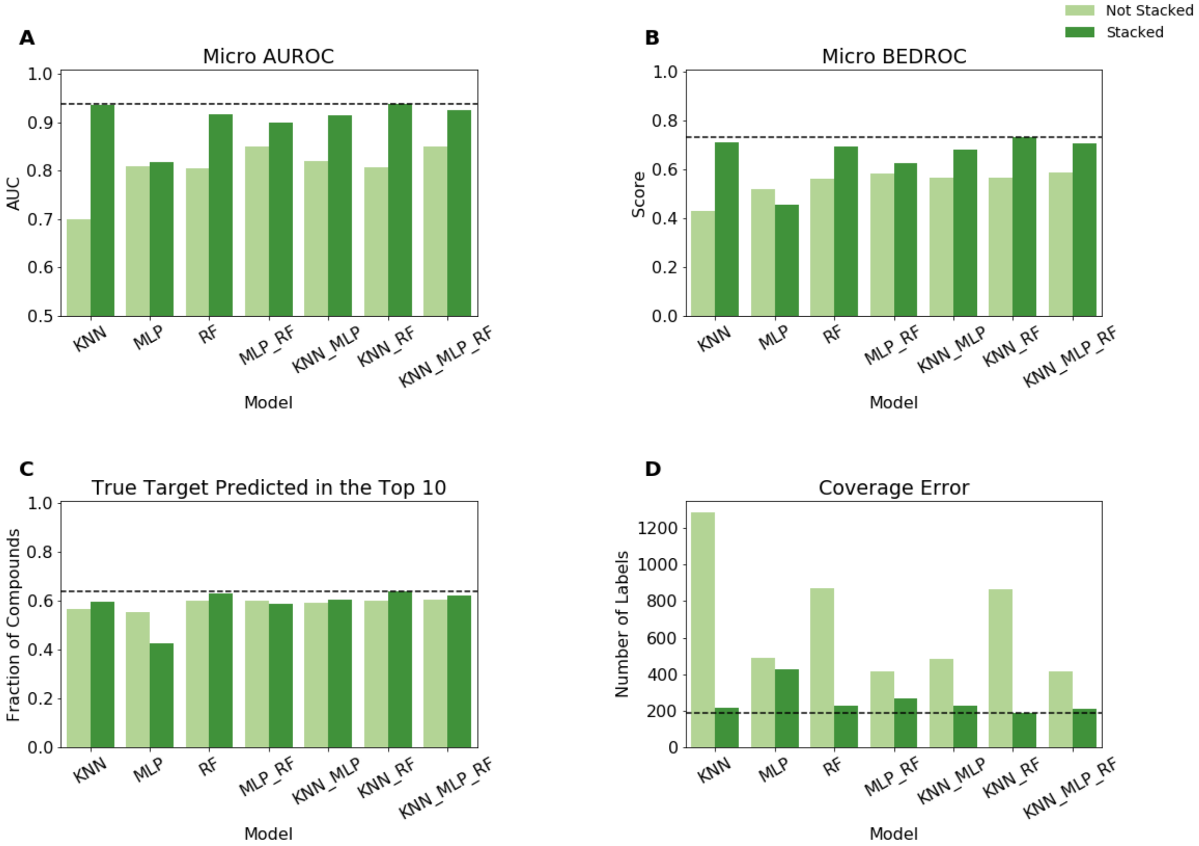


Figure 3.8 Natural Product Benchmark Performance. Model performance for benchmarking on the natural product dataset. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which have at least one true target among the top 10 predictions, and (D) coverage error are shown.

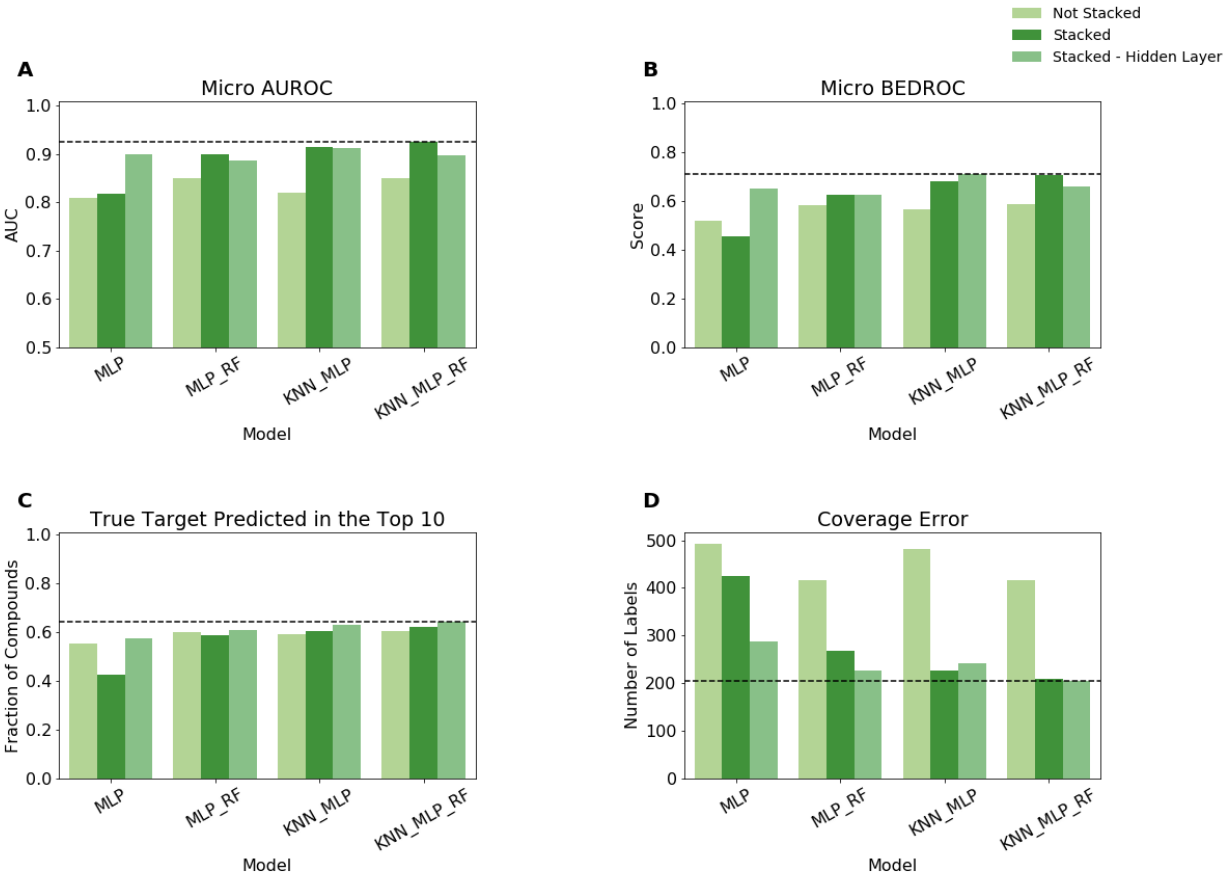


Figure 3.9 Natural Product Benchmark Performance with MLP Hidden Layer Features. Model performance for benchmarking on the natural product dataset. For a single model, "Not Stacked" indicates that the probability predictions of the listed model were used directly. If more than one model is listed, the mean probabilities for each label were used. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. "Hidden Layer" indicates that the normalized penultimate layer activations were used as input features for the logistic regression instead of the predicted labels during model stacking. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which have at least one true target among the top 10 predictions, and (D) coverage error are shown.

Interestingly, the use of a single level 0 classifier, with the exception of MLP, saw performance improvements with model stacking. This phenomenon is particularly apparent when comparing unstacked and stacked KNN models on the natural product benchmarking set. For example, the unstacked KNN model shows the worst micro-averaged AUROC score among the unstacked classifiers but stacking improves the score from 0.70 to 0.94. Such a dramatic increase in performance is unexpected, when the power of stacking is cited as being a result of combining level 0 classifiers. However, this assumes each model is passing singular values to be combined; either 1, 2, or 3 total values for each label depending on the number of base classifiers considered. In the models described here, all 1,907 predicted probabilities are passed from each level 0 classifier to the level 1 logistic regression classifier. Since the logistic regression is trained in a one-vs-rest fashion for this multi-label classification problem, each protein target label is predicted using all predicted probabilities; either 1,907, 3,814, or 5,721 total values for each label depending on the number of base classifiers considered. In the example of KNN, many of these predicted probabilities are 0. However, information of the non-zero values can be used to influence the prediction of a given protein target label.

The predicted target protein label information being used to give final predictions can be examined through the extraction of model coefficients from each trained logistic regression classifier. For example, the logistic regression model for predicting the protein target label “Q12884” (Prolyl endopeptidase FAP) has coefficients greater than 1 for the predicted probabilities of target labels “P48147”, “P97321”, “P27487”, “Q86TI2”, “Q9UHL4”, and “Q6V1X1'”. Inspecting the UniProt records for each reveal that these proteins share a common function, which is the cleavage of proline-containing peptide bonds. Since these proteins share a similar function and substrate preference it would be unsurprising if a given compound was able to bind to more than one of these related proteins. However, direct binding data is difficult to obtain and will be unavailable for many compound-protein target combinations. Therefore, while level 0 model predictions may strongly and reasonably predict for one of these related proteins, this prediction would ultimately be treated as a false positive due to the unknown binding relationship. Through stacking, the level 1 classifier can learn from this information and ultimately make better predictions for the known protein target labels.

To demonstrate that the logistic regression is using probabilities of functionally related proteins to improve predictions, semantic similarities were calculated. Gene ontology (GO) is a widely used basis for the measurement of functional similarity.[123–126] GO terms from the molecular function ontology were able to be obtained for 1,878 of the 1,907 UniProt protein target labels through programmatic access to QuickGO via the provided API.[127] Semantic similarities for each pairwise combination of protein target label GO terms were then computed according to the Lin expression of term similarity with the best-match product method using the OntologyX package suite in R.[128–131] For each predicted label, the corresponding UniProt IDs for logistic regression coefficients with values greater than one were obtained, which resulted in 1,595 label groups of the possible 1,907. The average semantic similarity of a query group of labels was calculated from the pairwise similarity matrix. Significance of group similarity for each query group of labels was assessed by a permutation test. Subsets containing the same number of labels as the query group of labels are sampled from the calculated pairwise similarity matrix. The proportion of these samples that have at least as high of an average similarity value as the query group of labels yields an unbiased estimate of the p-value for the group.[132] From these calculations it is observed that 80% of the label groups had scores with associated p-values <0.05 (Figure 3.10). Therefore, 80% of the label groups had a similarity score higher than at least 95% of the permutated groups. Overall, the semantic similarity calculation indicates that most of protein target labels predicted by logistic regression were obtained by combination of the probabilities from functionally related proteins. This relationship was not given explicitly as an input feature during model training but was inferred from the similarity between the training ligands for which each of the proteins were known to bind.

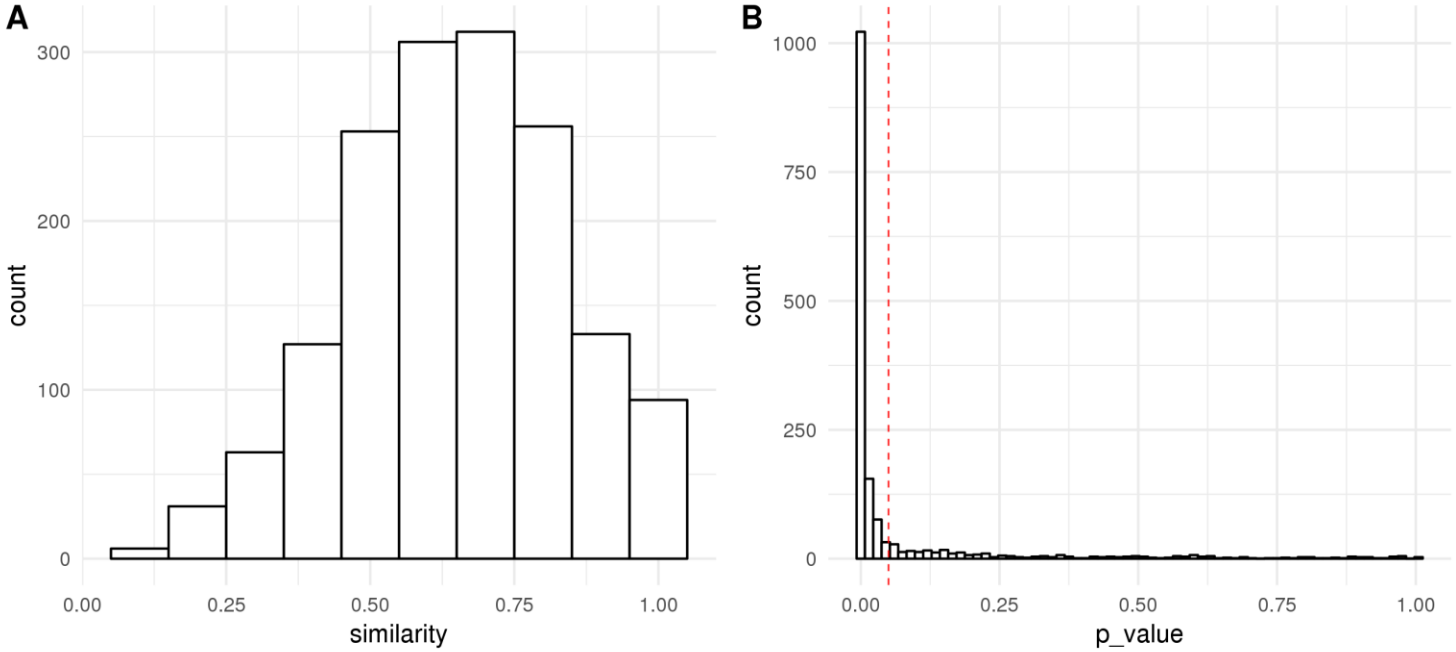


Figure 3.10 Protein Semantic Similarity. Comparison of protein functional similarity measured by semantic similarity of molecular function gene ontology (GO) ID annotations for each protein UniProt ID. Sets of protein target labels, as UniProt IDs, were obtained from the coefficients of the logistic regression models that were trained to predict each protein target label in the KNN stacked model. (A) Distribution of the average semantic similarities of each protein target label set. (B) Distribution of p-values for the average similarity values of each group of protein target labels. The dashed red line is placed at the p-value 0.05. 80% of the protein target label sets had a p-value < 0.05. Significance of group similarity for each query group of labels was assessed by a permutation test.

3.4.4 *Cross-Validation on Subsets of the Synthetic Compound Dataset*

Additional stratified cross-validations were performed to understand how the size of the training data, the number of targets considered, and the use of a dissimilar set of compounds impacted model performance during cross-validation for the synthetic compound dataset. A total of six modified datasets were generated from the original synthetic compound set (Figure 3.11). First, all target labels with 100 compound records were used, which resulted in 635 protein targets and 63,500 compound-target activity pairs. This 100 compound record set ("100") was one of the final datasets used for cross-validation, but further sets were also generated from the set for comparison. To assess how the number of targets considered impacted performance, half of the protein target labels were randomly selected to yield another set ("100_ht") with 100 compounds per target label, but now with 318 of the possible 635 protein target labels for a total of 31,800 activity pairs. Further subsets were made to examine the effect of reducing the number of compounds per target label. Therefore, both the 318 and the 635 protein target label sets were further subset by sampling 10 compounds from the 100 possible compound records for each protein target label. The 10 compounds were sampled through either random sampling ("10" and "10_ht) or through the selection of the most dissimilar 10 compounds ("10_ds" and "10_ht_ds"). This yielded two of the smaller 10 compound per label sets from each of the 100 compound per label sets, which were used to assess how intra-target class compound similarity influenced performance. Finally, stratified 10-fold cross-validation was performed on each of the six subsets.

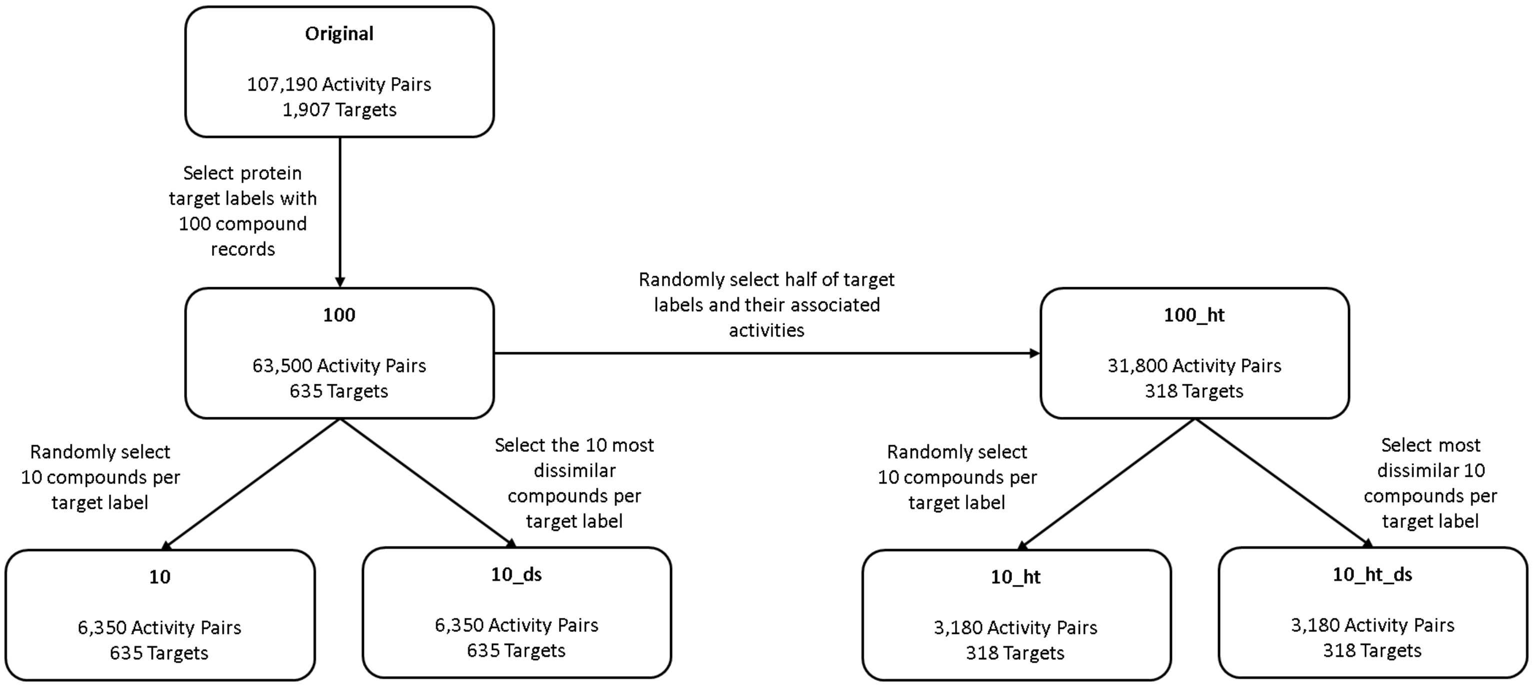

Original
107,190 Activity Pairs
1,907 Targets
Select protein target labels with 100 compound records
100
63,500 Activity Pairs
635 Targets
Randomly select half of target labels and their associated activities
100_ht
31,800 Activity Pairs
318 Targets
Randomly select 10 compounds per target label
Select the 10 most dissimilar compounds per target label
Randomly select 10 compounds per target label
Select most dissimilar 10 compounds per target label
10
6,350 Activity Pairs
635 Targets
10_ds
6,350 Activity Pairs
635 Targets
10_ht
3,180 Activity Pairs
318 Targets
10_ht_ds
3,180 Activity Pairs
318 Targets

Figure 3.11 Synthetic Dataset Subset Schematic. Schematic that shows how the original synthetic compound dataset was split into different sets for subsequent cross-validations. The values "10" and "100" refer to the number of compounds per protein target label. The abbreviations "ht" and "ds" refer to the number of protein target labels being halved and that the most dissimilar compounds were selected respectively.

Performance was assessed for the six subsets using the same metrics as the original cross-validation and natural product benchmark, with the exception of coverage error. Coverage error was instead converted to fractional coverage error by dividing the measured coverage error by the number of labels considered. This was necessary for a fair comparison due to differences in the number of target labels among the datasets being compared. The results of the KNN_RF classifier for each subset are shown in Figure 3.12.

Reducing the number of compounds had a negative effect on the classifier's performance. When comparing the "100" and "10" datasets for the unstacked model, the micro-averaged AUROC score decreased by 0.06, micro-averaged BEDROC score decreased by 0.15, compounds with a true target identified in the top 10 predictions decreased by 18%, and 9% more labels were required in order to recover all true labels. Model stacking slightly mitigated the performance decrease that resulted from the reduced number of compounds. When comparing the "100" and "10" datasets for the stacked model, the micro-averaged AUROC score decreased by 0.04, micro-averaged BEDROC score decreased by 0.14, compounds with a true target identified in the top 10 predictions decreased by 18%, and 5% more labels were required in order to recover all true labels.

The use of dissimilar compound sets dramatically reduced the classifier's performance. When comparing the "10" and "10_ds" datasets for the unstacked model, the micro-averaged AUROC score decreased by 0.2, the micro-averaged BEDROC score decreased by 0.44, compounds with a true target identified in the top 10 predictions decreased by 47%, and 30% more labels were required in order to recover all true labels. Model stacking mitigated the performance decrease that resulted from the use of dissimilar compound sets. When comparing the "10" and "10_ds" datasets for the stacked model, the micro-averaged AUROC score decreased by 0.09, the micro-averaged BEDROC score decreased by 0.33, compounds with a true target identified in the top 10 predictions decreased by 40%, and 9% more labels were required in order to recover all true labels.

Halving the number of protein target labels showed little effect on classifier performance. When comparing the difference between “100” and “100_ht”, “10” and “10_ht”, and “10_ds” and “10_ht_ds” for the unstacked model, on average the micro-averaged AUROC score increased by 0.007 (0.003), the micro-averaged BEDROC score increased by 0.006 (0.005), compounds with a true target identified in the top 10 predictions increased by 3.6% (0.79%), and 3.3 % (2.4%) fewer labels were required in order to recover all true labels. While halving the number of targets slightly increased unstacked model performance, it slightly decreased stacked model performance. When comparing the difference between “100” and “100_ht”, “10” and “10_ht”, and “10_ds” and “10_ht_ds” for the stacked model, on average the micro-averaged AUROC score decreased by 0.007(0.009), the micro-averaged BEDROC score decreased by 0.025 (0.028), compounds with a true target identified in the top 10 predictions decreased by 2.0% (0.66%), and 0.84% (0.80%) more labels were required in order to recover all true labels.

Overall, reducing the number of training compounds and using dissimilar compound sets had the greatest impact on performance, while halving the number of targets had a very minor impact on performance. Furthermore, the trends observed here prompted further investigation on the impact of training dataset size, target diversity, and training-test compound similarity.

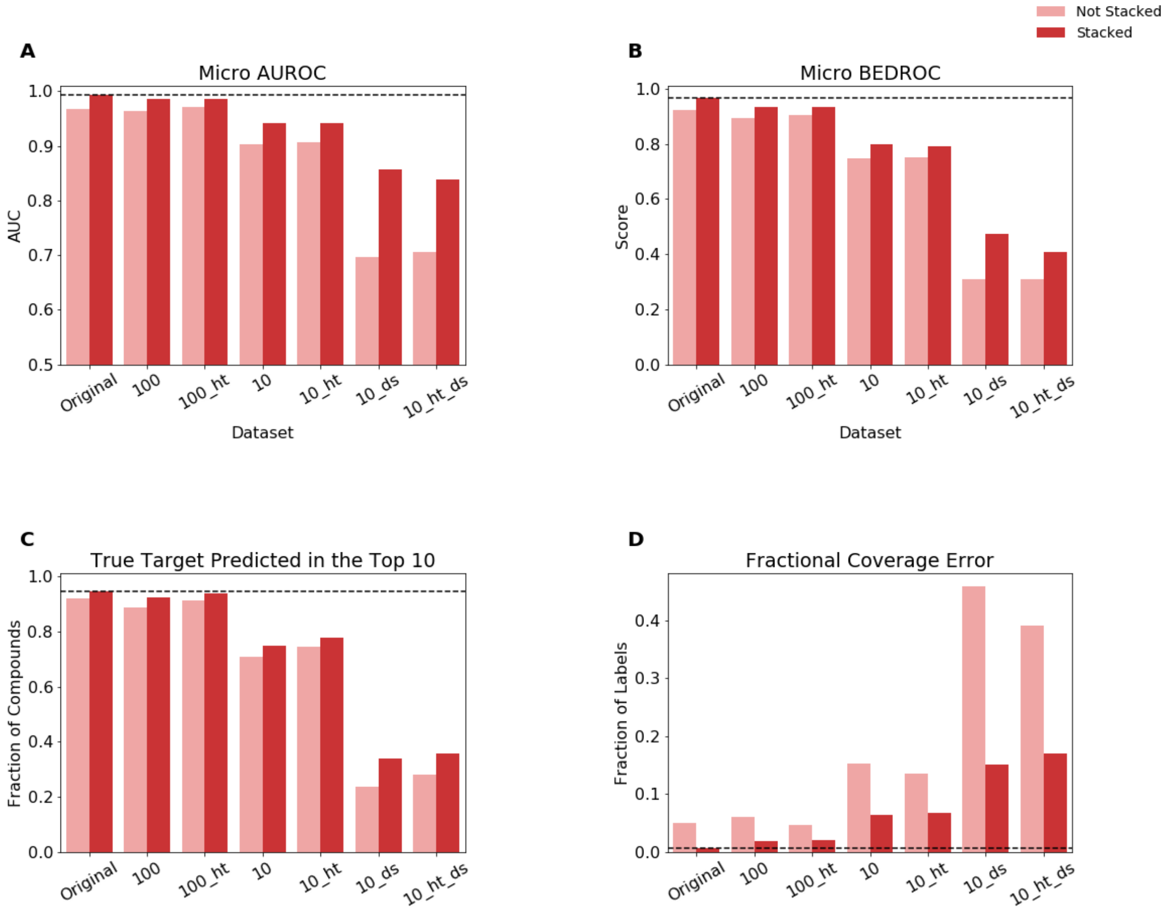


Figure 3.12 KNN_RF Model Performance on Synthetic Dataset Subsets. "Original" refers to the original cross-validation performance on the unmodified synthetic compound dataset. The values "10" and "100" refer to the number of compounds per protein target label. The abbreviations "ht" and "ds" refer to the number of protein target labels being halved and that the most dissimilar compounds were selected respectively. "Not Stacked" refers to the mean probabilities of model predictions when more than a single model is listed. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which have at least one true target among the top 10 predictions, and (D) fractional coverage error are shown.

3.4.5 Impact of Training Dataset Size on Cross-Validation Performance

It is expected that the number of training records for each protein target label influences classifier performance. To assess this in a systematic way, protein class labels with many compound records were collected and assessed through 10-fold cross-validation. The top 5 largest sets were selected, which included the D2 dopamine receptor (UniProtID: P14416), beta-secretase 1 (UniProtID: P56817), melanin-concentrating hormone receptor 1 (UniProtID: Q99705), cannabinoid receptor 2 (UniProtID: P34972), and vascular endothelial growth factor receptor 2 (UniProtID: P35968). A total of 2,500 compound records were randomly sampled for each protein target label and further randomly subsampled into sets of 2,000, 1,500, 1,000, 500, 100, and 10 compound records. The stratified 10-fold cross-validation procedure was then performed on each of these seven sets (Figure 3.13). Performance was assessed for the seven subsets using the same metrics as the original cross-validation and natural product benchmark, with the exception of true targets predicted in the top 10 results. This metric was modified to instead assess the fraction of compounds with a true target predicted as the top result.

The number of training records for each protein target label indeed had an impact on classifier performance. As the number of training compound records increases, a corresponding increase is observed in performance. However, this effect begins to plateau at 500 compound records with a micro-averaged AUROC score of 0.999, a micro-averaged BEDROC score of 0.998, 97% of compounds had a true target identified as the top results, and coverage error of 1.04 for the KNN_RF classifier. Additionally, “Not Stacked” and “Stacked” classifier performance converged at this point due to both achieving essentially perfect classification for the subset. The trends observed for the KNN_RF classifier were also observed for the other classifier combinations (Figure A.1-Figure A.6).

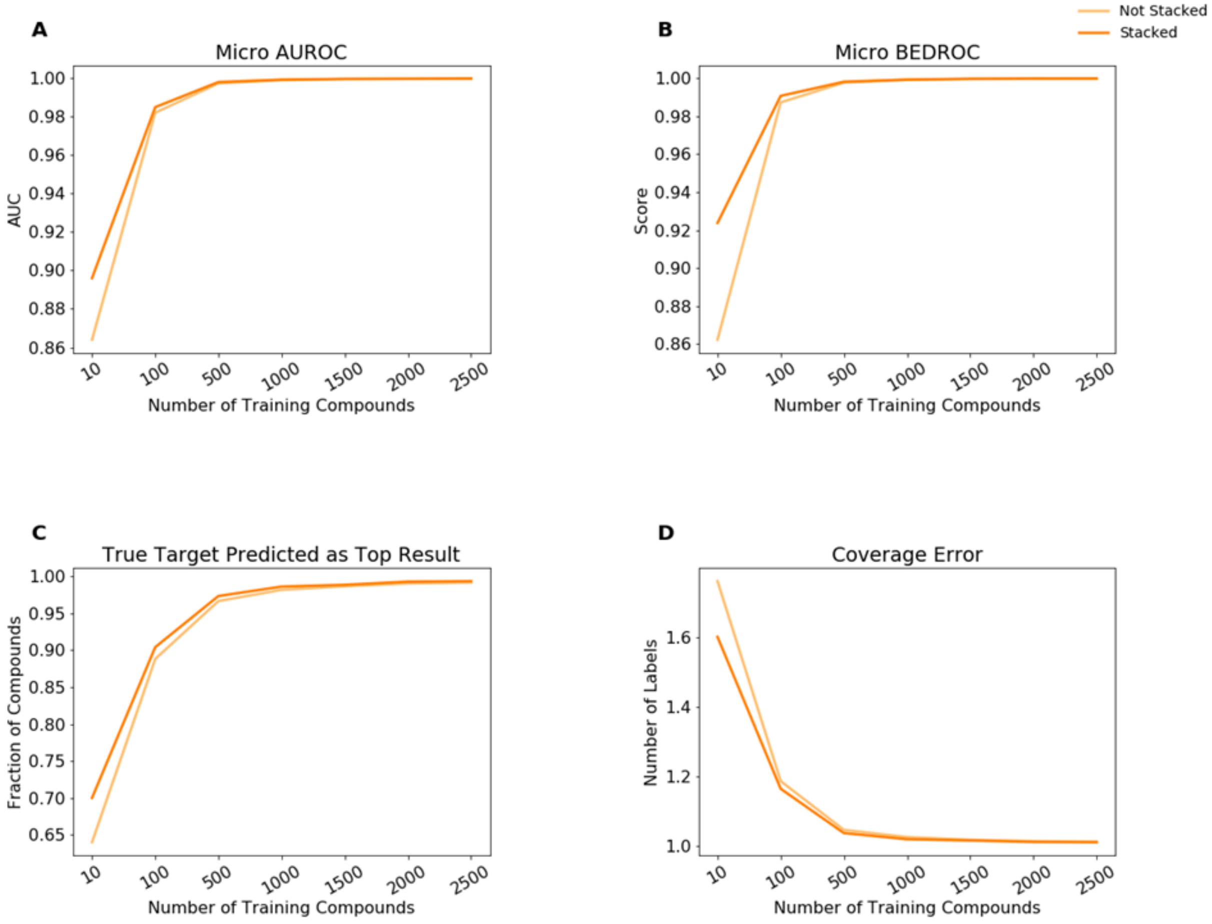


Figure 3.13 KNN_RF Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN_RF classifier. “Not Stacked” refers to the mean probabilities of model predictions when more than a single model is listed. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

3.4.6 Impact of Protein Target Diversity on Cross-Validation Performance

Another factor that is expected to influence performance is the diversity of protein targets in the dataset. Related protein targets are more likely to bind to similar small molecule compounds than diverse protein targets. As previously mentioned, any compound-protein target associations that were unknown were treated as negative data. This assumption has a negative impact on performance when a compound is assigned a negative label for a protein target that it may likely bind to but has never been tested against. A classifier may reasonably predict this protein target strongly and be penalized for doing so in the performance evaluation as it is ultimately treated as a false positive prediction.

To illustrate this effect, a diverse set of protein target labels were selected from the synthetic dataset used in full model training based on their L2 protein class as defined in ChEMBL 23. A single UniProtID was selected for each L2 protein class with priority given to the protein target labels with the largest number of compound records. This resulted in a dataset containing 2,825 compound-target records for 31 diverse protein targets. Another set was obtained for comparison that contained only kinases. A total of 31 UniProtIDs were selected that belonged to the kinase L2 protein class. During UniProtID selection, labels that contained a similar number of compounds records to those selected for the diverse protein target sets were selected. This resulted in a dataset containing 2,824 compound-target records for 31 kinase protein targets.

Classifier performance was assessed by stratified 10-fold cross-validation for the two datasets using the same metrics as described for the assessment of training compound set size. The expected performance degradation when considering related targets is observed (Figure 3.14). Without stacking, micro-averaged AUROC decreased by 0.10, micro-averaged BEDROC decreased by 0.26, 32% less compounds had a true target identified as the top prediction, and coverage error increased by 3.1 for the kinase set compared to the diverse target set. Stacking slightly improved the relative performance for micro-averaged AUROC and coverage error and had almost no effect on micro-averaged BEDROC and the number of compounds with a true target predicted as the top result. With stacking, micro-averaged AUROC decreased by 0.07, micro-averaged BEDROC decreased by 0.27, 33% less compounds had a true target identified as the top prediction, and coverage error increased by 2.4 for the kinase set compared to the diverse target set. This trend observed for the KNN_RF stacked classifier was also observed for the other classifier combinations (Figure A.7). Overall, the consideration of similar targets reduces performance since the classifier more frequently predicts that a compound binds to a target for which no interaction had yet been reported.

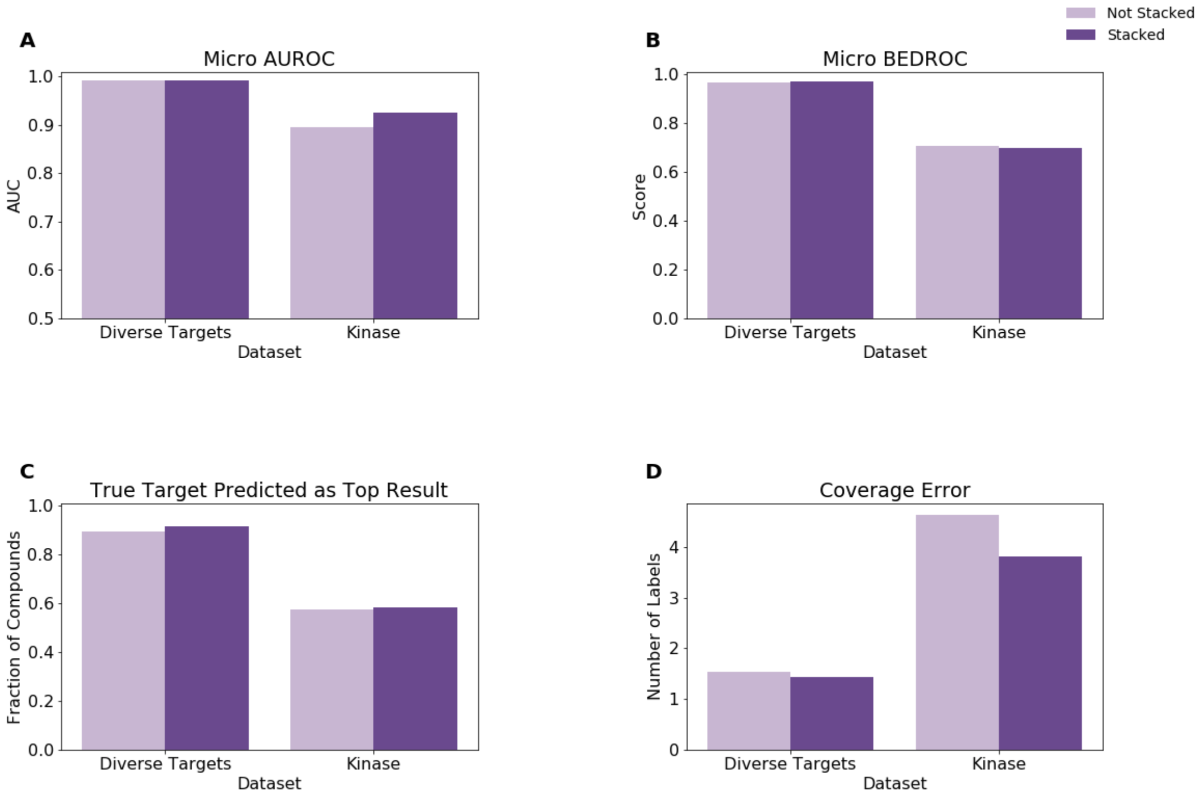


Figure 3.14 KNN_RF Model Performance with Diverse Target Labels. Model performance for stratified 10-fold cross-validation on the diverse target and kinase datasets for the KNN_RF classifier. “Not Stacked” refers to the mean probabilities of model predictions when more than a single model is listed. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

3.4.7 Impact of Intra-label Training-Test Compound Similarity on Predicted Probability Scores

Despite the use of machine learning and model stacking, this classification model is inherently dependent on ligand similarity. The underlying assumption for all ligand-based computational fishing methods is that proteins bind similar compounds. Therefore, if a query compound is dramatically dissimilar from compounds used in training the classification model for a protein target label, then low probability scores for that label are expected. Conversely, higher probability scores are expected as similarity between a query compound and the training compounds increases. However, a high degree of similarity to the training compounds is not always the case, as shown above for the natural product set, which has ramifications for the magnitude of the predicted probability scores.

To demonstrate the impact of training compound set similarity to the query compound predicted label probabilities, pairwise similarities were calculated and then compared to predicted label probability values. For each compound in the natural product benchmark set, pairwise similarities were calculated between the natural product and the training compounds belonging to the natural product's known target label classes. This yielded a similarity distribution for each known natural product-protein target activity pair. Additionally, the predicted probabilities output by the stacked classification model for each known natural product-protein target activity pair were collected.

The similarity distributions for each activity pair were aggregated and binned according to the probability predicted for the known labels. The aggregated similarity distributions for each probability range are compared and shown in Figure 3.15 for the KNN_RF stacked classifier. For each predicted probability range bin, (0.0, 0.25], (0.25, 0.5], (0.5, 0.75], and (0.75, 1.0] the interquartile ranges span from 0.08 to 0.15, 0.10 to 0.26, 0.09 to 0.27, and 0.09 to 0.34 respectively. The lower quartile values are all very close, and more distinct differences are observed between the upper quartiles especially for the lowest and highest probabilities ranges. In general, each probability range has a large proportion of low similarity values, and the letter-value plots[133] for each range look very similar below the median value. The major differences between distributions are observed above the median value. Comparison of the same portions of each distribution above the median, the boxes with the same width, shows an increase in average Tanimoto similarity as probability scores increase. This trend is also observed for the synthetic compound cross-validation (Figure A.8) and is also more strongly observed for the non-stacked base classifiers (Figure A.9-Figure A.12).

The number of predicted probabilities is not equally distributed among the four described ranges. There is a much larger number of probabilities predicted in the (0.0, 0.25] range, especially for the natural product set. Of the probabilities predicted by the KNN_RF stacked classifier for the natural product set, 93.8% of predicted probabilities are in the (0.0, 0.25] range, 2.8% in the (0.25, 0.5] range, 1.9% in the (0.5, 0.75] range, and 1.4 % in the (0.75, 1.0] range. For a synthetic compound cross-validation fold, 50.5% of the predicted probabilities are in the (0.0, 25] range, 9.0% are in the (0.25, 0.5] range, 9.1% are in the (0.5, 0.75] range, and 31.6% are in the (0.75, 1.0] range. Consistent with the knowledge that the synthetic compound cross-validation sets have higher intra-target similarity between training and test sets than for the natural products, the proportion of compounds receiving high probability predictions is far greater for the synthetic compounds than for the natural product set.

The observation that query compounds dissimilar from the training data yield low predicted probability scores for correct predictions has implications for model usage and interpretation. As demonstrated, the stacked classifiers had good predictive power on the natural product benchmark. Therefore, correct targets are generally ranked before incorrect targets despite the low probability scores given to correct targets. While top ranking predictions should not be taken as an absolute truth, users are also encouraged to not immediately dismiss top ranked hits based purely on a low score. No matter the score received, top ranked hits should be critically evaluated in the context of the available experimental data regarding the compound's bioactivity.

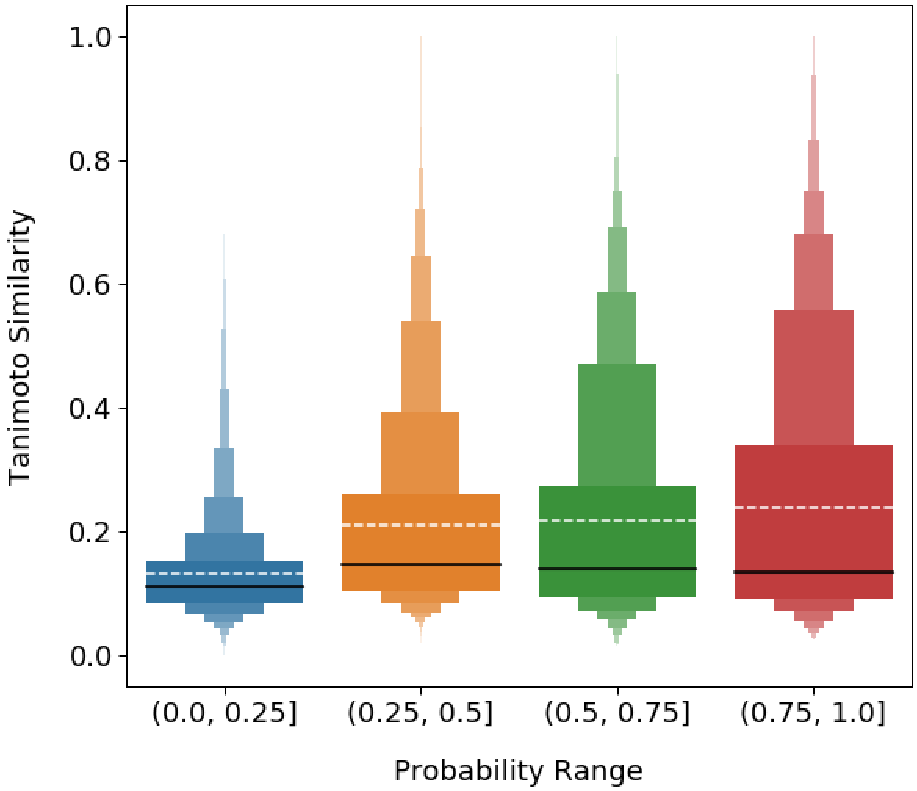
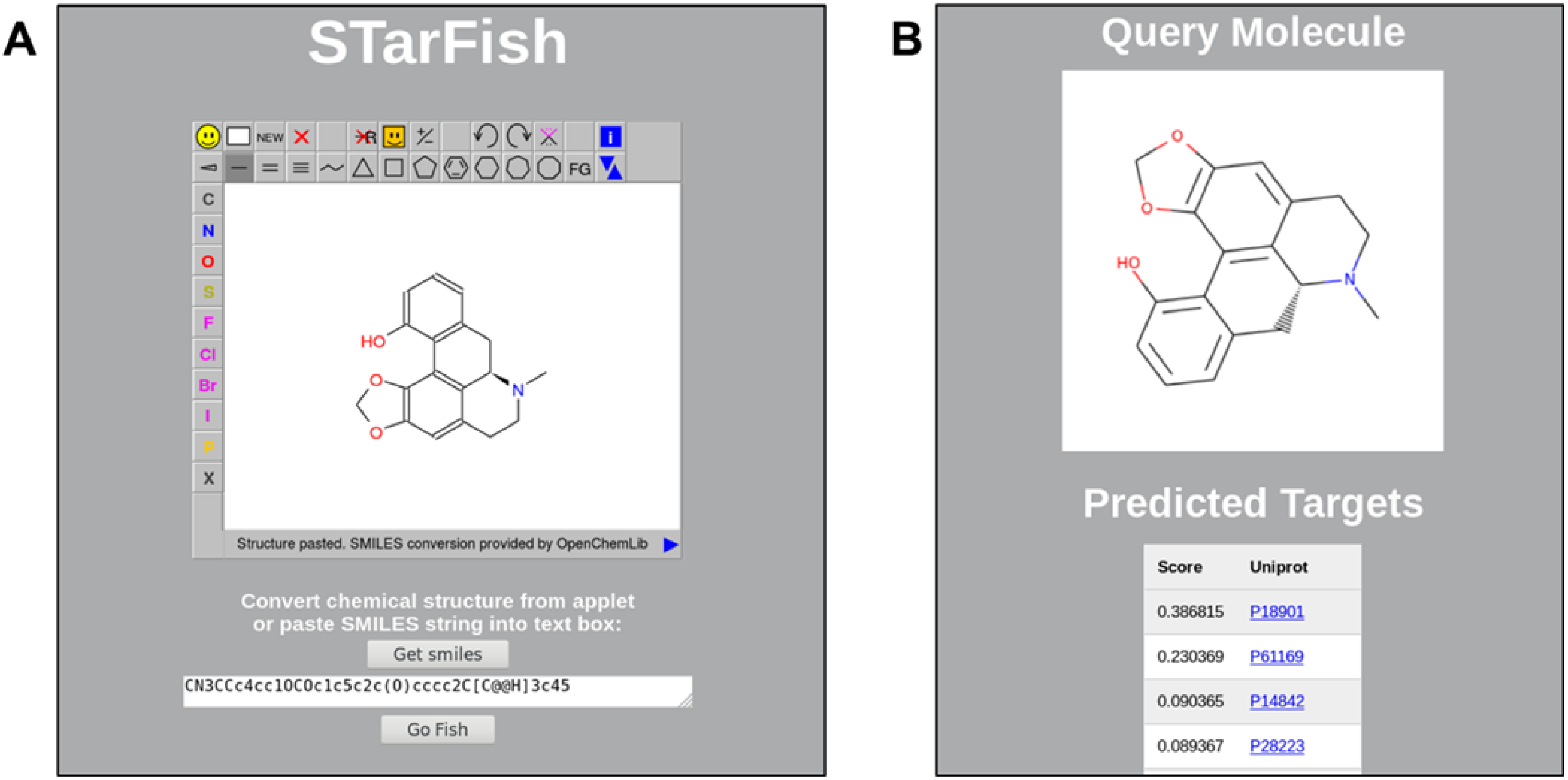


Figure 3.15 Letter-Value Plot of Aggregated Pairwise Similarity Distributions. Letter-value plot showing the aggregated pairwise similarity distributions for benchmark natural product compounds and synthetic training compounds for known positive protein target labels. Similarity distributions were aggregated based on the predicted probability from the KNN_RF stacked classifier for the known protein targets of each natural product. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

3.4.8 Deployment of Stacked Model as a Web Application

The trained model was deployed via an application programing interface (API) using Flask 0.12.2. The use of an API allows target predictions for molecules of interest to be made with an application run in a web browser. An example query for the natural product pukateine is shown in Figure 10. Pukateine is an aporphine alkaloid from the bark of the pukatea tree, *Laurelia novae-zelandiae*. Alkaloids extracted from the pukatea tree are thought to be the constituents responsible for the analgesic properties traditionally associated with the tree.[134] Pukateine is reported to bind to dopamine D_1 and D_2 receptors.[135] When pukateine is input into the STarFish web application, dopamine D_1 (UniProtID: P18901) and D_2 (UniProtID: P61169) receptors are the top two predicted targets. The next two predicted targets are the 5-hydroxytryptamine receptor 2A (5-HT2A) for rat (UniProtID: P14842) and human (UniProtID: P28223). No binding data for pukateine has been reported for this receptor, however, other aporphine alkaloids have been reported to have 5-HT2A activity.[136,137] Therefore, in addition to predicting two correct protein targets, STarFish, has also predicted another likely target.

While the KNN_RF stacked model demonstrated the best performance during cross-validation and on the natural product benchmark, the KNN stacked model was selected for use in the STarFish web application. Predictions using the RF models are significantly more computationally expensive, and the use of the KNN stacked model is computationally efficient with only a slight loss in relative performance. The use of a computationally efficient model allows for end users to easily run the STarFish web application on their own computers with minimal hardware requirements. However, experienced users can modify the API to include other model combinations if desired.


A
STarFish
NEW
C
N
O
S
F
Cl
Br
I
P
X
FG
HO
O
O
N
Structure pasted. SMILES conversion provided by OpenChemLib
Convert chemical structure from applet or paste SMILES string into text box:
Get smiles
Go Fish
B
Query Molecule
Predicted Targets

```
CN3CCc4cc1OCOc1c5c2c(O)cccc2C[C@@H]3c45
```

Score	Uniprot
0.386815	P18901
0.230369	P61169
0.090365	P14842
0.089367	P28223

Figure 3.16 STarFish Web Application. Example query using the STarFish web application. (A) Query SMILES obtained by sketching a compound or directly pasting a SMILES string into the text box. (B) The query molecule and a list of predicted protein targets along with a probability score for each.

3.5 Conclusions

To predict protein targets for natural products, a computational target fishing model, STarFish, was constructed using a model stacking approach and evaluated on a collected natural product benchmarking set. The collected natural product benchmark set consisted of 5,589 compound-target pairs for 1,943 unique compounds and 1,023 unique targets. All models were trained using potent synthetic compounds collected from ChEMBL and accounted for 1,907 protein targets. Model stacking combinations using k-nearest neighbors, random forest, and a multi-layer perceptron as level 0 classifiers and a logistic regression as a level 1 meta-classifier were examined. In general, model stacking approaches outperformed unstacked approaches, especially for the natural product benchmark. The stacked model comprised of KNN and RF as the level 0 classifiers showed the best performance with an AUROC score of 0.94 and a BEDROC score of 0.73. The stacked model comprised of KNN as the level 0 classifier had similar performance with an AUROC score of 0.94 and a BEDROC score of 0.71, but with significantly less computational expense. By default, STarFish uses the stacked KNN model to allow for use even with limited computing resources and has been deployed as an API, which can be downloaded and run in a web browser.

Bibliography

(1) Begam, B. F.; Kumar, J. S. A Study on Cheminformatics and Its Applications on Modern Drug Discovery. *Procedia Engineering* **2012**, *38*, 1264–1275. https://doi.org/10.1016/j.proeng.2012.06.156.

(2) Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46* (6), 2267–2277. https://doi.org/10.1021/ci600234z.

(3) Engel, T.; Gasteiger, J. *Applied Chemoinformatics: Achievements and Future Opportunities*; John Wiley & Sons, 2018.

(4) Daylight Theory: SMILES https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed Jun 27, 2019).

(5) Daylight Theory: Fingerprints https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed Jun 27, 2019).

(6) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996. https://doi.org/10.1021/ci9800211.

(7) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *Journal of Cheminformatics* **2015**, *7* (1), 20. https://doi.org/10.1186/s13321-015-0069-3.

(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings1PII of Original Article: S0169-409X(96)00423-1. The Article Was Originally Published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46* (1), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

(9) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on More than 96,000 Compounds. *Journal of Pharmaceutical Sciences* **2009**, *98* (3), 861–893. https://doi.org/10.1002/jps.21494.

(10) Xing, L.; Glen, R. C. Novel Methods for the Prediction of LogP, PKa, and LogD. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (4), 796–805. https://doi.org/10.1021/ci010315d.

(11) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547. https://doi.org/10.1038/s41586-018-0337-2.

(12) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discovery Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

(13) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nature Reviews Drug Discovery* **2019**, 1. https://doi.org/10.1038/s41573-019-0024-5.

(14) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer Texts in Statistics; Springer New York: New York, NY, 2013; Vol. 103. https://doi.org/10.1007/978-1-4614-7138-7.

(15) Kotsiantis, S. B. Supervised Machine Learning: A Review of Classification Techniques. 20.

(16) Tsoumakas, G.; Katakis, I. Multi-Label Classification: An Overview. 17.

(17) Trajdos, P.; Kurzynski, M. An Extension of Multi-Label Binary Relevance Models Based on Randomized Reference Classifier and Local Fuzzy Confusion Matrix. In *Intelligent Data Engineering and Automated Learning – IDEAL 2015*; Jackowski, K., Burduk, R., Walkowiak, K., Wozniak, M., Yin, H., Eds.; Lecture Notes in Computer Science; Springer International Publishing, 2015; pp 69–76.

(18) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol Rev* **2014**, *66* (1), 334–395. https://doi.org/10.1124/pr.112.007336.

(19) Anderson, A. C. The Process of Structure-Based Drug Design. *Chemistry & Biology* **2003**, *10* (9), 787–797. https://doi.org/10.1016/j.chembiol.2003.09.002.

(20) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys Rev* **2017**, *9* (2), 91–102. https://doi.org/10.1007/s12551-016-0247-1.

(21) Ferreira, L. G.; dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20* (7), 13384–13421. https://doi.org/10.3390/molecules200713384.

(22) Kabir, M. H.; Patrick, R.; Ho, J. W. K.; O'Connor, M. D. Identification of Active Signaling Pathways by Integrating Gene Expression and Protein Interaction Data. *BMC Syst Biol* **2018**, *12* (Suppl 9). https://doi.org/10.1186/s12918-018-0655-x.

(23) Vakser, I. A. Protein-Protein Docking: From Interaction to Interactome. *Biophys J* **2014**, *107* (8), 1785–1793. https://doi.org/10.1016/j.bpj.2014.08.033.

(24) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89* (6), 2195–2199. https://doi.org/10.1073/pnas.89.6.2195.

(25) Lensink, M. F.; Velankar, S.; Wodak, S. J. Modeling Protein–Protein and Protein–Peptide Complexes: CAPRI 6th Edition. *Proteins: Structure, Function, and Bioinformatics* **2017**, *85* (3), 359–377. https://doi.org/10.1002/prot.25215.

(26) Kastritis, P. L.; Bonvin, A. M. J. J. Are Scoring Functions in Protein−Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *J. Proteome Res.* **2010**, *9* (5), 2216–2225. https://doi.org/10.1021/pr9009854.

(27) Porter, K. A.; Desta, I.; Kozakov, D.; Vajda, S. What Method to Use for Protein–Protein Docking? *Current Opinion in Structural Biology* **2019**, *55*, 1–7. https://doi.org/10.1016/j.sbi.2018.12.010.

(28) An, S.; Fu, L. Small-Molecule PROTACs: An Emerging and Promising Approach for the Development of Targeted Therapy Drugs. *EBioMedicine* **2018**, *36*, 553–562. https://doi.org/10.1016/j.ebiom.2018.09.005.

(29) Scheepstra, M.; Hekking, K. F. W.; van Hijfte, L.; Folmer, R. H. A. Bivalent Ligands for Protein Degradation in Drug Discovery. *Computational and Structural Biotechnology Journal* **2019**, *17*, 160–176. https://doi.org/10.1016/j.csbj.2019.01.006.

(30) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/jm020017n.

(31) Benet, L. Z.; Hosey, C. M.; Ursu, O.; Oprea, T. I. BDDCS, the Rule of 5 and Drugability. *Adv Drug Deliv Rev* **2016**, *101*, 89–98. https://doi.org/10.1016/j.addr.2016.05.007.

(32) DeGoey, D. A.; Chen, H.-J.; Cox, P. B.; Wendt, M. D. Beyond the Rule of 5: Lessons Learned from AbbVie’s Drugs and Compound Collection. *J. Med. Chem.* **2018**, *61* (7), 2636–2651. https://doi.org/10.1021/acs.jmedchem.7b00717.

(33) Bondeson, D. P.; Mares, A.; Smith, I. E. D.; Ko, E.; Campos, S.; Miah, A. H.; Mulholland, K. E.; Routly, N.; Buckley, D. L.; Gustafson, J. L.; et al. Catalytic in Vivo Protein Knockdown by Small-Molecule PROTACs. *Nat Chem Biol* **2015**, *11* (8), 611–617. https://doi.org/10.1038/nchembio.1858.

(34) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; et al. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res* **2016**, *44* (Database issue), D1220–D1228. https://doi.org/10.1093/nar/gkv1253.

(35) Edmondson, S. D.; Yang, B.; Fallan, C. Proteolysis Targeting Chimeras (PROTACs) in ‘beyond Rule-of-Five’ Chemical Space: Recent Progress and Future Challenges. *Bioorganic & Medicinal Chemistry Letters* **2019**, *29* (13), 1555–1564. https://doi.org/10.1016/j.bmcl.2019.04.030.

(36) Morales, F.; Giordano, A. Overview of CDK9 as a Target in Cancer Research. *Cell Cycle* **2016**, *15* (4), 519–527. https://doi.org/10.1080/15384101.2016.1138186.

(37) Robb, C. M.; Contreras, J. I.; Kour, S.; Taylor, M. A.; Abid, M.; Sonawane, Y. A.; Zahid, M.; Murry, D. J.; Natarajan, A.; Rana, S. Chemically Induced Degradation of CDK9 by a Proteolysis Targeting Chimera (PROTAC). *Chem. Commun.* **2017**, *53* (54), 7577–7580. https://doi.org/10.1039/C7CC03879H.

(38) Bondeson, D. P.; Smith, B. E.; Burslem, G. M.; Buhimschi, A. D.; Hines, J.; Jaime-Figueroa, S.; Wang, J.; Hamman, B. D.; Ishchenko, A.; Crews, C. M. Lessons in PROTAC Design from Selective Degradation with a Promiscuous Warhead. *Cell Chemical Biology* **2018**, *25* (1), 78-87.e5. https://doi.org/10.1016/j.chembiol.2017.09.010.

(39) Smith, B. E.; Wang, S. L.; Jaime-Figueroa, S.; Harbin, A.; Wang, J.; Hamman, B. D.; Crews, C. M. Differential PROTAC Substrate Specificity Dictated by Orientation of Recruited E3 Ligase. *Nature Communications* **2019**, *10* (1), 131. https://doi.org/10.1038/s41467-018-08027-7.

(40) Roy, M. J.; Winkler, S.; Hughes, S. J.; Whitworth, C.; Galant, M.; Farnaby, W.; Rumpel, K.; Ciulli, A. SPR-Measured Dissociation Kinetics of PROTAC Ternary

Complexes Influence Target Degradation Rate. *ACS Chem Biol* **2019**, *14* (3), 361–368. https://doi.org/10.1021/acschembio.9b00092.

(41) Gadd, M. S.; Testa, A.; Lucas, X.; Chan, K.-H.; Chen, W.; Lamont, D. J.; Zengerle, M.; Ciulli, A. Structural Basis of PROTAC Cooperative Recognition for Selective Protein Degradation. *Nat Chem Biol* **2017**, *13* (5), 514–521. https://doi.org/10.1038/nchembio.2329.

(42) Xue, L. C.; Dobbs, D.; Bonvin, A. M. J. J.; Honavar, V. Computational Prediction of Protein Interfaces: A Review of Data Driven Methods. *FEBS Letters* **2015**, *589* (23), 3516–3526. https://doi.org/10.1016/j.febslet.2015.10.003.

(43) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46* (1), 531–558. https://doi.org/10.1146/annurev-biophys-070816-033654.

(44) Nowak, R. P.; DeAngelo, S. L.; Buckley, D.; He, Z.; Donovan, K. A.; An, J.; Safaee, N.; Jedrychowski, M. P.; Ponthier, C. M.; Ishoey, M.; et al. Plasticity in Binding Confers Selectivity in Ligand-Induced Protein Degradation. *Nature Chemical Biology* **2018**, *14* (7), 706. https://doi.org/10.1038/s41589-018-0055-y.

(45) *ACD/Percepta, Version 2018.1.1, Advanced Chemistry Development, Inc., Toronto, ON, Canada, Www.Acdlabs.Com, 2019.*

(46) Michael Waskom; Olga Botvinnik; Drew O’Kane; Paul Hobson; Joel Ostblom; Saulius Lukauskas; David C Gemperline; Tom Augspurger; Yaroslav Halchenko; John B. Cole; et al. *Mwaskom/Seaborn: V0.9.0 (July 2018)*; Zenodo, 2018. https://doi.org/10.5281/zenodo.1313201.

(47) Janert, P. K. *Gnuplot in Action: Understanding Data with Graphs*; Manning Publications: Greenwich, Conn, 2010.

(48) Scott, D. W. On Optimal and Data-Based Histograms. 6.

(49) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(50) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **1993**, *234* (3), 779–815. https://doi.org/10.1006/jmbi.1993.1626.

(51) *Molecular Graphics Images Were Produced Using the UCSF Chimera Package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (Supported by NIH P41 RR-01081).*

(52) *Schrödinger Release 2017-2: Maestro, Schrödinger, LLC, New York, NY, 2018.*

(53) Pierce, B. G.; Hourai, Y.; Weng, Z. Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library. *PLOS ONE* **2011**, *6* (9), e24657. https://doi.org/10.1371/journal.pone.0024657.

(54) Pierce, B.; Weng, Z. ZRANK: Reranking Protein Docking Predictions with an Optimized Energy Function. *Proteins: Structure, Function, and Bioinformatics* **2007**, *67* (4), 1078–1086. https://doi.org/10.1002/prot.21373.

(55) Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8, 2015.

(56) RDKit: Open-source cheminformatics; https://www.rdkit.org/ (accessed Dec 20, 2018).

(57) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chemistry Central Journal* **2008**, *2* (1), 5. https://doi.org/10.1186/1752-153X-2-5.

(58) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(59) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Computational Biology* **2017**, *13* (7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

(60) Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; et al. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14* (11), 6076–6092. https://doi.org/10.1021/acs.jctc.8b00640.

(61) Mullard, A. First Targeted Protein Degrader Hits the Clinic. *Nature Reviews Drug Discovery* **2019**, *18*, 237. https://doi.org/10.1038/d41573-019-00043-6.

(62) Crew, A. P.; Dong, H.; Wang, J.; Chen, X.; Qian, Y.; Zimmermann, K.; Crews, C. M.; Berlin, M.; Snyder, L. COMPOUNDS AND METHODS FOR THE TARGETED DEGRADATION OF ANDROGEN RECEPTOR. US-20180346461-A1, December 6, 2018.

(63) Swain, C. Protacs https://www.cambridgemedchemconsulting.com//resources//resources/lead_identification//resources/lead_identification/protacs/protacs.html (accessed May 29, 2019).

(64) A Phase 1 Clinical Trial of ARV-110 in Patients With Metastatic Castration-resistant Prostate Cancer. - Full Text View - ClinicalTrials.gov https://clinicaltrials.gov/ct2/show/NCT03888612 (accessed May 29, 2019).

(65) Squires, M. S.; Feltell, R. E.; Wallis, N. G.; Lewis, E. J.; Smith, D.-M.; Cross, D. M.; Lyons, J. F.; Thompson, N. T. Biological Characterization of AT7519, a Small-Molecule Inhibitor of Cyclin-Dependent Kinases, in Human Tumor Cell Lines. *Mol Cancer Ther* **2009**, *8* (2), 324–332. https://doi.org/10.1158/1535-7163.MCT-08-0890.

(66) Olson, C. M.; Jiang, B.; Erb, M. A.; Liang, Y.; Doctor, Z. M.; Zhang, Z.; Zhang, T.; Kwiatkowski, N.; Boukhali, M.; Green, J. L.; et al. Pharmacological Perturbation of CDK9 Using Selective CDK9 Inhibition or Degradation. *Nature Chemical Biology* **2018**, *14* (2), 163–170. https://doi.org/10.1038/nchembio.2538.

(67) Chen, R.; Wierda, W. G.; Chubb, S.; Hawtin, R. E.; Fox, J. A.; Keating, M. J.; Gandhi, V.; Plunkett, W. Mechanism of Action of SNS-032, a Novel Cyclin-Dependent Kinase Inhibitor, in Chronic Lymphocytic Leukemia. *Blood* **2009**, *113* (19), 4637–4645. https://doi.org/10.1182/blood-2008-12-190256.

(68) Liu, H.; Herrmann, C. H. Differential Localization and Expression of the Cdk9 42k and 55k Isoforms. *Journal of Cellular Physiology* **2005**, *203* (1), 251–260. https://doi.org/10.1002/jcp.20224.

(69) Boffo, S.; Damato, A.; Alfano, L.; Giordano, A. CDK9 Inhibitors in Acute Myeloid Leukemia. *Journal of Experimental & Clinical Cancer Research* **2018**, *37* (1), 36. https://doi.org/10.1186/s13046-018-0704-8.

(70) Drummond, M. L.; Williams, C. I. In Silico Modeling of PROTAC-Mediated Ternary Complexes: Validation and Application. *J. Chem. Inf. Model.* **2019**, *59* (4), 1634–1644. https://doi.org/10.1021/acs.jcim.8b00872.

(71) *Molecular Operating Environment (MOE) 2018.0101; Chemical Computing Group, ULC: Montreal, Quebec, Canada, 2018.*

(72) Swinney, D. C.; Anthony, J. How Were New Medicines Discovered? *Nat. Rev. Drug Discov.* **2011**, *10* (7), 507–519. https://doi.org/10.1038/nrd3480.

(73) Moffat, J. G.; Vincent, F.; Lee, J. A.; Eder, J.; Prunotto, M. Opportunities and Challenges in Phenotypic Drug Discovery: An Industry Perspective. *Nat. Rev. Drug Discov.* **2017**, *16* (8), 531–543. https://doi.org/10.1038/nrd.2017.111.

(74) Oda, Y.; Owa, T.; Sato, T.; Boucher, B.; Daniels, S.; Yamanaka, H.; Shinohara, Y.; Yokoi, A.; Kuromitsu, J.; Nagasu, T. Quantitative Chemical Proteomics for Identifying Candidate Drug Targets. *Anal. Chem.* **2003**, *75* (9), 2159–2165. https://doi.org/10.1021/ac026196y.

(75) Schenone, M.; Dančík, V.; Wagner, B. K.; Clemons, P. A. Target Identification and Mechanism of Action in Chemical Biology and Drug Discovery. *Nat. Chem. Biol.* **2013**, *9* (4), 232–240. https://doi.org/10.1038/nchembio.1199.

(76) Sydow, D.; Burggraaff, L.; Szengel, A.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P.; Volkamer, A. Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* **2019**, 1728–1742. https://doi.org/10.1021/acs.jcim.8b00832.

(77) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206. https://doi.org/10.1038/nbt1284.

(78) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids. Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(79) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (Database issue), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(80) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958. https://doi.org/10.1021/ci034160g.

(81) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem.* **2001**, *26* (1), 5–14. https://doi.org/10.1016/S0097-8485(01)00094-8.

(82) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47* (18), 4463–4470. https://doi.org/10.1021/jm0303195.

(83) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/C8SC00148K.

(84) Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent Applications of Deep Learning and Machine Intelligence on in Silico Drug Discovery: Methods, Tools and Databases. *Brief. Bioinform. 44*, 1–36. https://doi.org/10.1093/bib/bby061.

(85) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79* (3), 629–661. https://doi.org/10.1021/acs.jnatprod.5b01055.

(86) Eder, J.; Sedrani, R.; Wiesmann, C. The Discovery of First-in-Class Drugs: Origins and Evolution. *Nat. Rev. Drug Discov.* **2014**, *13* (8), 577–587. https://doi.org/10.1038/nrd4336.

(87) Fang, J.; Wu, Z.; Cai, C.; Wang, Q.; Tang, Y.; Cheng, F. Quantitative and Systems Pharmacology. 1. In Silico Prediction of Drug–Target Interactions of Natural Products Enables New Targeted Cancer Therapy. *J. Chem. Inf. Model.* **2017**, *57* (11), 2657–2671. https://doi.org/10.1021/acs.jcim.7b00216.

(88) Keum, J.; Yoo, S.; Lee, D.; Nam, H. Prediction of Compound-Target Interactions of Natural Products Using Large-Scale Drug and Protein Information. *BMC Bioinformatics* **2016**, *17* (Suppl 6). https://doi.org/10.1186/s12859-016-1081-y.

(89) Grenet, I.; Merlo, K.; Comet, J.-P.; Tertiaux, R.; Rouquié, D.; Dayan, F. Stacked Generalization with Applicability Domain Outperforms Simple QSAR on in Vitro Toxicological Data. *J. Chem. Inf. Model.* **2019**. https://doi.org/10.1021/acs.jcim.8b00553.

(90) Li, W.; Miao, W.; Cui, J.; Fang, C.; Su, S.; Li, H.; Hu, L.; Lu, Y.; Chen, G. Efficient Corrections for DFT Noncovalent Interactions Based on Ensemble Learning Models. *J. Chem. Inf. Model.* **2019**. https://doi.org/10.1021/acs.jcim.8b00878.

(91) Kaggle: Your Home for Data Science https://www.kaggle.com/ (accessed Apr 18, 2019).

(92) Otto Group Product Classification Challenge https://kaggle.com/c/otto-group-product-classification-challenge (accessed Apr 18, 2019).

(93) Afzal, A. M.; Mussa, H. Y.; Turner, R. E.; Bender, A.; Glen, R. C. A Multi-Label Approach to Target Prediction Taking Ligand Promiscuity into Account. *J. Cheminform.* **2015**, *7*. https://doi.org/10.1186/s13321-015-0071-9.

(94) Ntie-Kang, F.; Simoben, C. V.; Karaman, B.; Ngwa, V. F.; Judson, P. N.; Sippl, W.; Mbaze, L. M. Pharmacophore Modeling and in Silico Toxicity Assessment of Potential Anticancer Agents from African Medicinal Plants. *Drug Des. Devel. Ther.* **2016**, *10*, 2137–2154. https://doi.org/10.2147/DDDT.S108118.

(95) Ntie-Kang, F.; Zofou, D.; Babiaka, S. B.; Meudom, R.; Scharfe, M.; Lifongo, L. L.; Mbah, J. A.; Mbaze, L. M.; Sippl, W.; Efange, S. M. N. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLOS ONE* **2013**, *8* (10), e78085. https://doi.org/10.1371/journal.pone.0078085.

(96) Onguéné, P. A.; Ntie-Kang, F.; Mbah, J. A.; Lifongo, L. L.; Ndom, J. C.; Sippl, W.; Mbaze, L. M. The Potential of Anti-Malarial Compounds Derived from African Medicinal Plants, Part III: An in Silico Evaluation of Drug Metabolism and Pharmacokinetics Profiling. *Org. Med. Chem. Lett.* **2014**, *4* (1), 6. https://doi.org/10.1186/s13588-014-0006-x.

(97) Natural Resources and Technologies https://ac-discovery.com/ (accessed Dec 19, 2018).

(98) Yabuzaki, J. Carotenoids Database: Structures, Chemical Fingerprints and Distribution among Organisms. *Database (Oxford)* **2017**, *2017*. https://doi.org/10.1093/database/bax004.

(99) Ntie-Kang, F.; Onguéné, P. A.; Scharfe, M.; Owono, L. C. O.; Megnassan, E.; Mbaze, L. M.; Sippl, W.; Efange, S. M. N. ConMedNP: A Natural Product Library from Central African Medicinal Plants for Drug Discovery. *RSC Adv.* **2013**, *4* (1), 409–419. https://doi.org/10.1039/C3RA43754J.

(100) InterBioScreen ltd. | Compound Libraries https://www.ibscreen.com (accessed Dec 19, 2018).

(101) MITISHAMBA DATABASE http://mitishamba.uonbi.ac.ke/ (accessed Dec 19, 2018).

(102) Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; A. Moumbock, A. F.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S. NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **2017**, *80* (7), 2067–2076. https://doi.org/10.1021/acs.jnatprod.7b00283.

(103) Natural Products Atlas | Home https://www.npatlas.org/joomla/index.php (accessed Dec 19, 2018).

(104) Mangal, M.; Sagar, P.; Singh, H.; Raghava, G. P. S.; Agarwal, S. M. NPACT: Naturally Occurring Plant-Based Anti-Cancer Compound-Activity-Target Database. *Nucleic Acids Res.* **2013**, *41* (Database issue), D1124–D1129. https://doi.org/10.1093/nar/gks1047.

(105) Zeng, X.; Zhang, P.; He, W.; Qin, C.; Chen, S.; Tao, L.; Wang, Y.; Tan, Y.; Gao, D.; Wang, B.; et al. NPASS: Natural Product Activity and Species Source Database for Natural Product Research, Discovery and Tool Development. *Nucleic Acids Res.* **2018**, *46* (D1), D1217–D1222. https://doi.org/10.1093/nar/gkx1026.

(106) Pilon, A. C.; Valli, M.; Dametto, A. C.; Pinto, M. E. F.; Freire, R. T.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. NuBBE DB : An Updated Database to Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep.* **2017**, *7* (1), 7215. https://doi.org/10.1038/s41598-017-07451-x.

(107) Ntie-Kang, F.; Onguéné, P. A.; Fotso, G. W.; Andrae-Marobela, K.; Bezabih, M.; Ndom, J. C.; Ngadjui, B. T.; Ogundaini, A. O.; Abegaz, B. M.; Meva'a, L. M. Virtualizing the P-ANAPL Library: A Step towards Drug Discovery from African Medicinal Plants. *PLOS ONE* **2014**, *9* (3), e90655. https://doi.org/10.1371/journal.pone.0090655.

(108) Hatherley, R.; Brown, D. K.; Musyoka, T. M.; Penkler, D. L.; Faya, N.; Lobb, K. A.; Tastan Bishop, Ö. SANCDB: A South African Natural Compound Database. *J. Cheminform.* **2015**, *7*. https://doi.org/10.1186/s13321-015-0080-8.

(109) Banerjee, P.; Erehman, J.; Gohlke, B.-O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II—a Database of Natural Products. *Nucleic Acids Res.* **2015**, *43* (Database issue), D935–D939. https://doi.org/10.1093/nar/gku886.

(110) Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. *PLOS ONE* **2011**, *6* (1). https://doi.org/10.1371/journal.pone.0015939.

(111) Lin, Y.-C.; Wang, C.-C.; Chen, I.-S.; Jheng, J.-L.; Li, J.-H.; Tung, C.-W. TIPdb: A Database of Anticancer, Antiplatelet, and Antituberculosis Phytochemicals from Indigenous Plants in Taiwan https://www.hindawi.com/journals/tswj/2013/736386/ (accessed Dec 19, 2018). https://doi.org/10.1155/2013/736386.

(112) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLOS ONE* **2013**, *8* (4). https://doi.org/10.1371/journal.pone.0062839.

(113) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(114) MolVS: Molecule Validation and Standardization — MolVS 0.1.1 documentation https://molvs.readthedocs.io/en/latest/ (accessed Dec 20, 2018).

(115) Peón, A.; Naulaerts, S.; Ballester, P. J. Predicting the Reliability of Drug-Target Interaction Predictions with Maximum Coverage of Target Space. *Sci. Rep.* **2017**, *7* (1), 3820. https://doi.org/10.1038/s41598-017-04264-w.

(116) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(117) Pedregosa, F. Scikit-Learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* 6.

(118) Ohio Supercomputer Center. **1987**.

(119) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508. https://doi.org/10.1021/ci600426e.

(120) Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **2017**, *57* (9), 2099–2111. https://doi.org/10.1021/acs.jcim.7b00341.

(121) Lopez-del Rio, A.; Nonell-Canals, A.; Vidal, D.; Perera-Lluna, A. Evaluation of Cross-Validation Strategies in Sequence-Based Binding Prediction Using Deep Learning. *J. Chem. Inf. Model.* **2019**. https://doi.org/10.1021/acs.jcim.8b00663.

(122) Lenselink, E. B.; Dijke, N. ten; Bongers, B.; Papadatos, G.; Vlijmen, H. W. T. van; Kowalczyk, W.; IJzerman, A. P.; Westen, G. J. P. van. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminform.* **2017**, *9* (1), 45. https://doi.org/10.1186/s13321-017-0232-0.

(123) Teng, Z.; Guo, M.; Liu, X.; Dai, Q.; Wang, C.; Xuan, P. Measuring Gene Functional Similarity Based on Group-Wise Comparison of GO Terms. *Bioinformatics* **2013**, *29* (11), 1424–1432. https://doi.org/10.1093/bioinformatics/btt160.

(124) Weichenberger, C. X.; Palermo, A.; Pramstaller, P. P.; Domingues, F. S. Exploring Approaches for Detecting Protein Functional Similarity within an Orthology-Based Framework. *Sci. Rep.* **2017**, *7* (1), 381. https://doi.org/10.1038/s41598-017-00465-5.

(125) Mazandu, G. K.; Mulder, N. J. Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data

Type? *PLOS ONE* **2014**, *9* (12), e113859. https://doi.org/10.1371/journal.pone.0113859.

(126) Liu, M.; Thomas, P. D. GO Functional Similarity Clustering Depends on Similarity Measure, Clustering Method, and Annotation Completeness. *BMC Bioinformatics* **2019**, *20* (1), 155. https://doi.org/10.1186/s12859-019-2752-2.

(127) Binns, D.; Dimmer, E.; Huntley, R.; Barrell, D.; O'Donovan, C.; Apweiler, R. QuickGO: A Web-Based Tool for Gene Ontology Searching. *Bioinformatics* **2009**, *25* (22), 3045–3046. https://doi.org/10.1093/bioinformatics/btp536.

(128) *RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL Http://Www.Rstudio.Com/.*

(129) *R Development Core Team (2008). R: A Language and Environment Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL Http://Www.R-Project.Org.*

(130) Greene, D.; Richardson, S.; Turro, E. OntologyX: A Suite of R Packages for Working with Ontological Data. *Bioinformatics* **2017**, *33* (7), 1104–1106. https://doi.org/10.1093/bioinformatics/btw763.

(131) Lin, D. An Information-Theoretic Definition of Similarity. In *In Proceedings of the 15th International Conference on Machine Learning*; Morgan Kaufmann, 1998; pp 296–304.

(132) Greene, D. J. Methods for Determining the Genetic Causes of Rare Diseases, University of Cambridge, 2018.

(133) Hofmann, H.; Wickham, H.; Kafadar, K. Letter-Value Plots: Boxplots for Large Data. *J. Comput. Graph. Stat.* **2017**, *26* (3), 469–477. https://doi.org/10.1080/10618600.2017.1305277.

(134) Fogg, W. S. The Pharmacological Action of Pukateine. *J. Pharmacol. Exp. Ther.* **1935**, *54* (2), 167–187.

(135) Dajas-Bailador, F. A.; Asencio, M.; Bonilla, C.; Scorza, Ma. C.; Echeverry, C.; Reyes-Parada, M.; Silveira, R.; Protais, P.; Russell, G.; Cassels, B. K.; et al. Dopaminergic Pharmacology and Antioxidant Properties of Pukateine, a Natural Product Lead for the Design of Agents Increasing Dopamine Neurotransmission. *Gen. Pharmacol.* **1999**, *32* (3), 373–379. https://doi.org/10.1016/S0306-3623(98)00210-9.

(136) Munusamy, V.; Yap, B. K.; Buckle, M. J. C.; Doughty, S. W.; Chung, L. Y. Structure-Based Identification of Aporphines with Selective 5-HT2A Receptor-

Binding Activity. *Chem. Biol. Drug Des.* **2013**, *81* (2), 250–256. https://doi.org/10.1111/cbdd.12069.

(137) Ponnala, S.; Gonzales, J.; Kapadia, N.; Navarro, H. A.; Harding, W. W. Evaluation of Structural Effects on 5-HT2A Receptor Antagonism by Aporphines: Identification of a New Aporphine with 5-HT2A Antagonist Activity. *Bioorg. Med. Chem. Lett.* **2014**, *24* (7), 1664–1667. https://doi.org/10.1016/j.bmcl.2014.02.066.

Appendix A. Supplemental Tables and Figures.

Table A. 1 Results of 10-fold cross-validation on the synthetic compound dataset.

Model	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage	Type
KNN	0.942301 (0.00204)	0.926946 (0.002619)	0.903073 (0.003441)	0.878424 (0.004442)	0.887923 (0.003782)	0.858709 (0.004589)	187.432522 (5.965479)	Not Stacked
MLP	0.979047 (0.00154)	0.980881 (0.001518)	0.889626 (0.003988)	0.864936 (0.005078)	0.919088 (0.004065)	0.918307 (0.004824)	37.669495 (2.086298)	Not Stacked
RF	0.966171 (0.001907)	0.947825 (0.002528)	0.915714 (0.002656)	0.890845 (0.004556)	0.919514 (0.003409)	0.884665 (0.00463)	96.05143 (4.288706)	Not Stacked
MLP_RF	0.985001 (0.000851)	0.982912 (0.001303)	0.917994 (0.002563)	0.893223 (0.004476)	0.933013 (0.003042)	0.916546 (0.004095)	29.402247 (1.140787)	Not Stacked
KNN_MLP	0.981905 (0.001228)	0.982763 (0.001351)	0.91647 (0.003034)	0.891882 (0.004705)	0.938412 (0.003113)	0.931853 (0.003854)	34.485059 (1.744688)	Not Stacked
KNN_RF	0.966719 (0.001945)	0.948387 (0.002586)	0.917974 (0.003122)	0.893239 (0.004643)	0.923396 (0.00365)	0.887987 (0.004632)	94.662067 (4.420363)	Not Stacked
KNN_MLP_RF	0.985188 (0.000868)	0.983078 (0.00131)	0.919445 (0.002943)	0.89475 (0.004735)	0.935549 (0.003141)	0.91861 (0.0041)	29.285386 (1.155606)	Not Stacked
KNN	0.989754 (0.001177)	0.989983 (0.001013)	0.926564 (0.002868)	0.902647 (0.005331)	0.950907 (0.003202)	0.950178 (0.002287)	21.891085 (2.285872)	Stacked
MLP	0.97467 (0.001503)	0.985998 (0.000695)	0.849289 (0.004211)	0.827107 (0.005889)	0.890829 (0.002729)	0.945159 (0.002604)	54.948871 (2.853801)	Stacked
RF	0.992117 (0.000812)	0.993382 (0.00081)	0.942336 (0.002908)	0.918346 (0.004259)	0.961141 (0.001832)	0.96196 (0.002096)	17.607684 (1.705223)	Stacked
MLP_RF	0.990756 (0.00075)	0.992894 (0.000855)	0.926472 (0.003249)	0.902666 (0.004413)	0.950749 (0.001785)	0.961727 (0.001978)	21.209976 (1.548678)	Stacked
KNN_MLP	0.99245 (0.000724)	0.992431 (0.000764)	0.935409 (0.002499)	0.911633 (0.003481)	0.958234 (0.001665)	0.958074 (0.002156)	17.41799 (1.516647)	Stacked
KNN_RF	0.994133 (0.000625)	0.99308 (0.000792)	0.945509 (0.002734)	0.921512 (0.004051)	0.967476 (0.001585)	0.961067 (0.002037)	13.722058 (1.306009)	Stacked
KNN_MLP_RF	0.993727 (0.000631)	0.993192 (0.000811)	0.940914 (0.00285)	0.91691 (0.003607)	0.963391 (0.001632)	0.961474 (0.002002)	14.998962 (1.308995)	Stacked

Table A. 2 Model performance results on the natural product benchmark.

Model	micro_AUROC	macro_AUROC	Frac_1_in_top10	Frac_all_in_top10	micro_BEDROC	macro_BEDROC	coverage	Type
KNN	0.700767	0.72812	0.566135	0.339775	0.430205	0.473909	1286.037056	Not Stacked
MLP	0.809043	0.821639	0.555327	0.342816	0.517641	0.513207	492.21719	Not Stacked
RF	0.805814	0.797995	0.601647	0.383253	0.560575	0.55939	870.233145	Not Stacked
MLP_RF	0.850754	0.845468	0.602162	0.386116	0.584146	0.584559	416.970664	Not Stacked
KNN_MLP	0.820458	0.836301	0.592898	0.380032	0.567536	0.587548	482.108595	Not Stacked
KNN_RF	0.806571	0.798452	0.599074	0.386831	0.565894	0.560926	866.966032	Not Stacked
KNN_MLP_RF	0.851126	0.845666	0.602676	0.388978	0.587345	0.585724	416.318065	Not Stacked
KNN	0.935312	0.889168	0.594442	0.425121	0.711535	0.687939	217.677818	Stacked
MLP	0.81811	0.719749	0.427689	0.278046	0.455236	0.485233	425.947504	Stacked
RF	0.917567	0.892882	0.630468	0.44194	0.692452	0.702486	230.226454	Stacked
MLP_RF	0.898711	0.874591	0.587751	0.405618	0.626237	0.68325	267.904786	Stacked
KNN_MLP	0.915233	0.883064	0.60422	0.421363	0.681662	0.676551	225.667524	Stacked
KNN_RF	0.938025	0.899854	0.637159	0.448918	0.732045	0.710923	190.437983	Stacked
KNN_MLP_RF	0.925955	0.890382	0.62069	0.435677	0.704735	0.697931	208.935152	Stacked

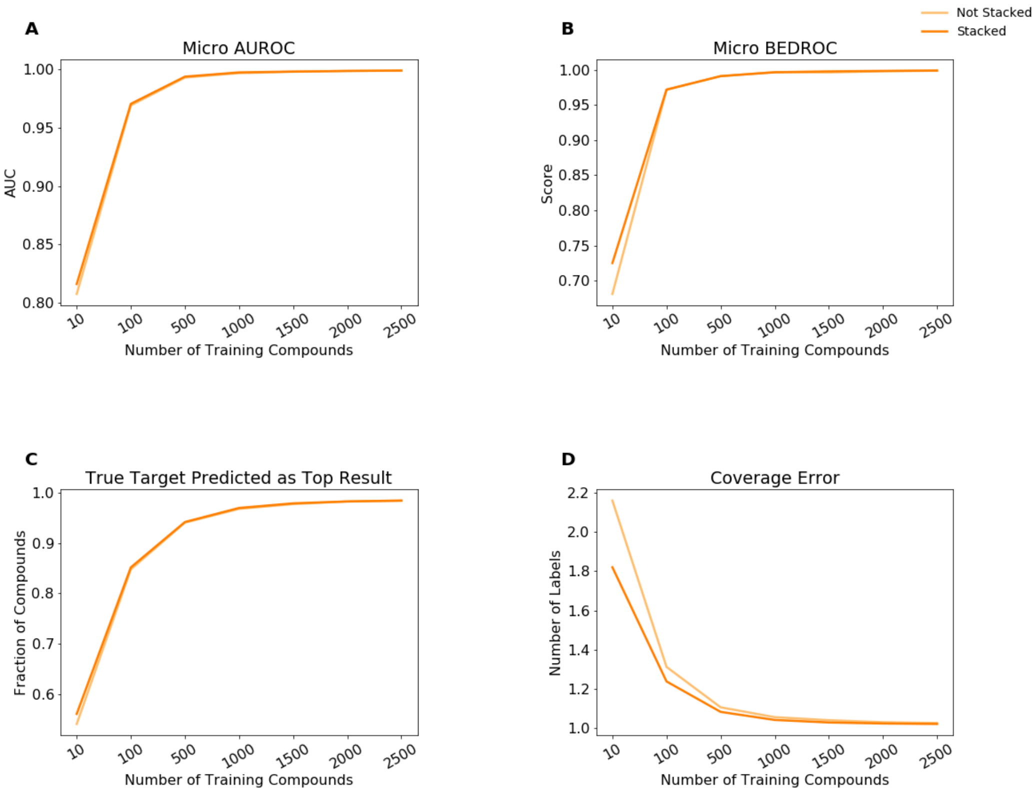


Figure A.1 KNN Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN classifier. “Not Stacked” refers to the mean probabilities of model predictions when more than a single model is listed. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

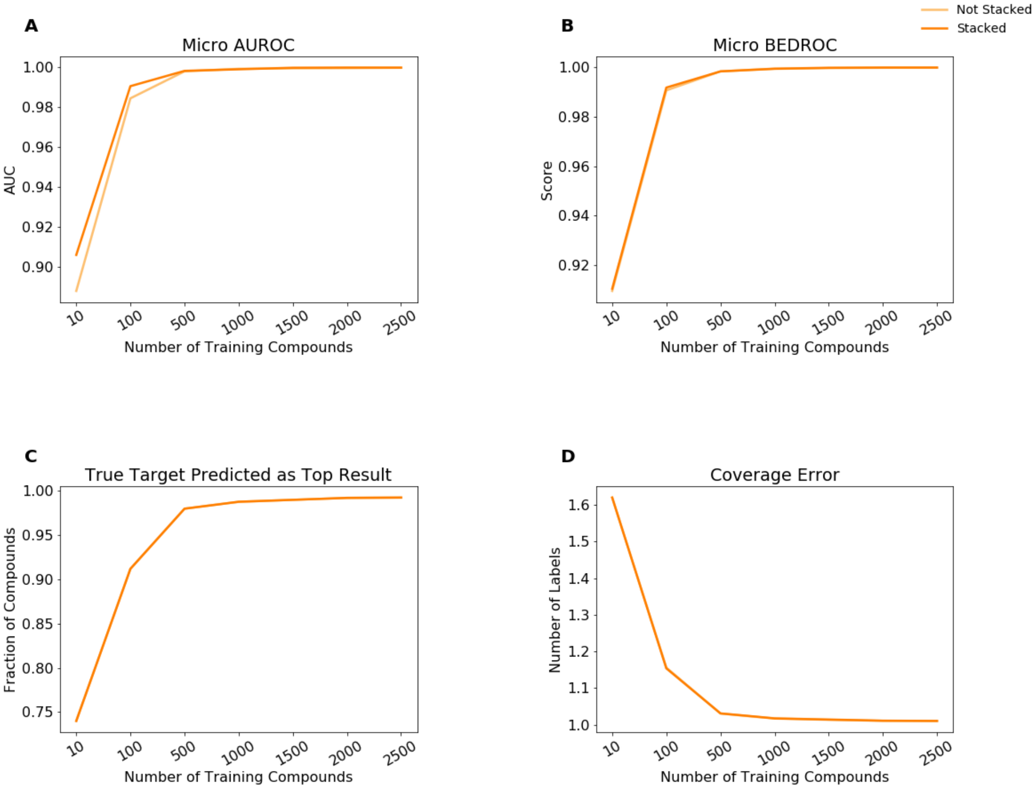


Figure A.2 MLP Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the MLP classifier. "Not Stacked" refers to the mean probabilities of model predictions when more than a single model is listed. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

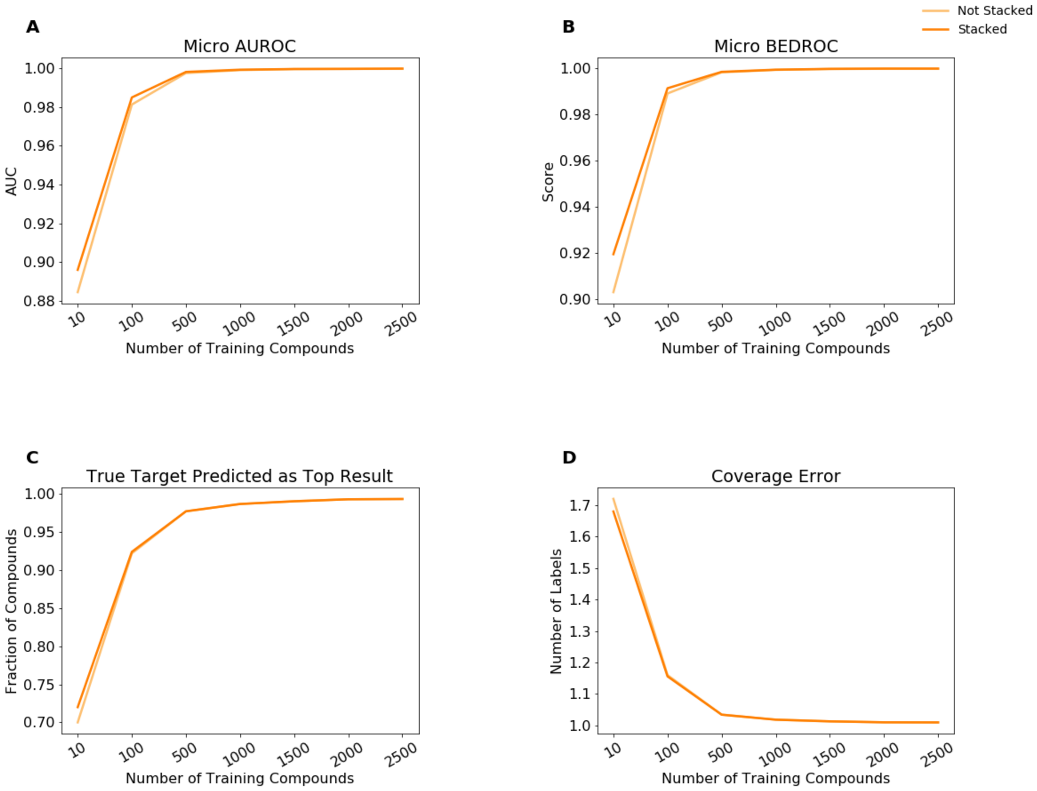


Figure A.3 RF Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the RF classifier. “Not Stacked” refers to the mean probabilities of model predictions when more than a single model is listed. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

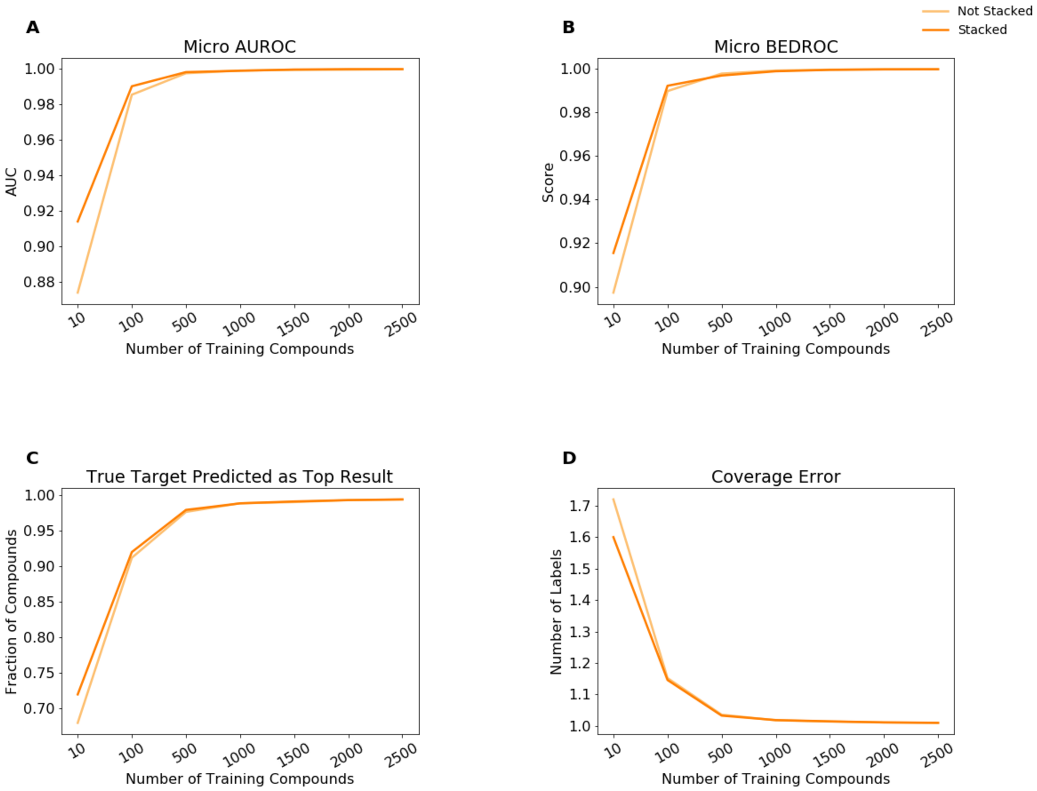


Figure A.4 KNN_MLP Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN_MLP classifier. “Not Stacked” refers to the mean probabilities of model predictions when more than a single model is listed. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

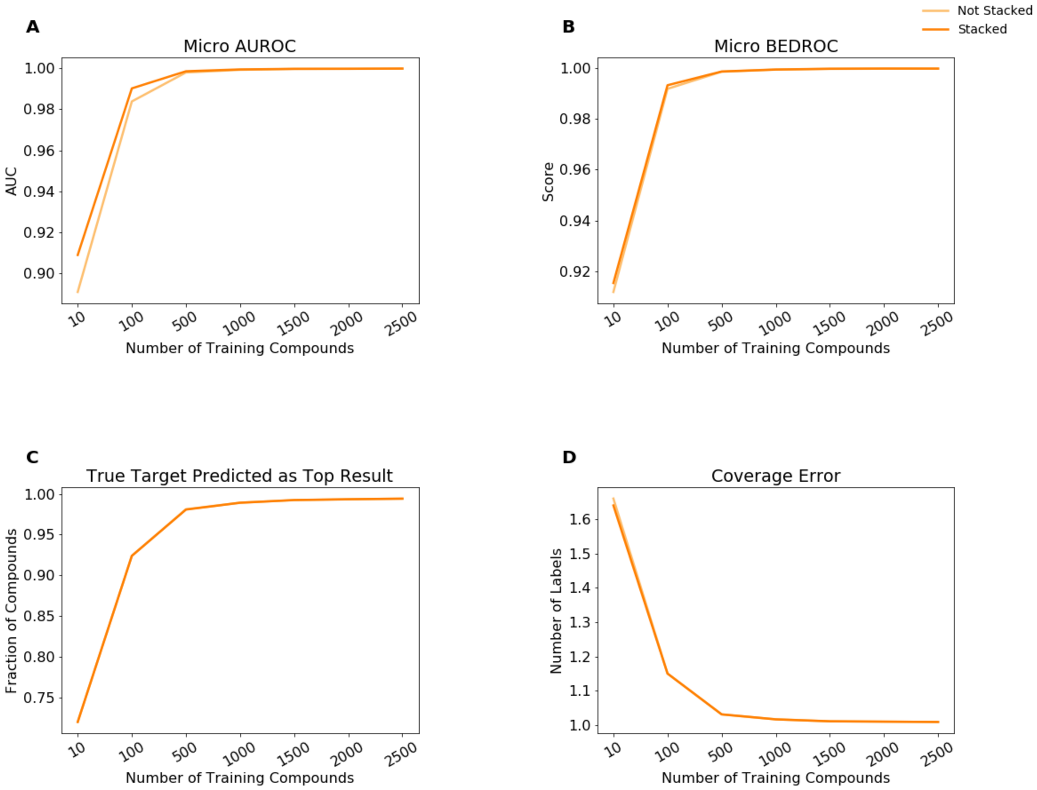


Figure A.5 MLP_RF Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the MLP_RF classifier. “Not Stacked” refers to the mean probabilities of model predictions when more than a single model is listed. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

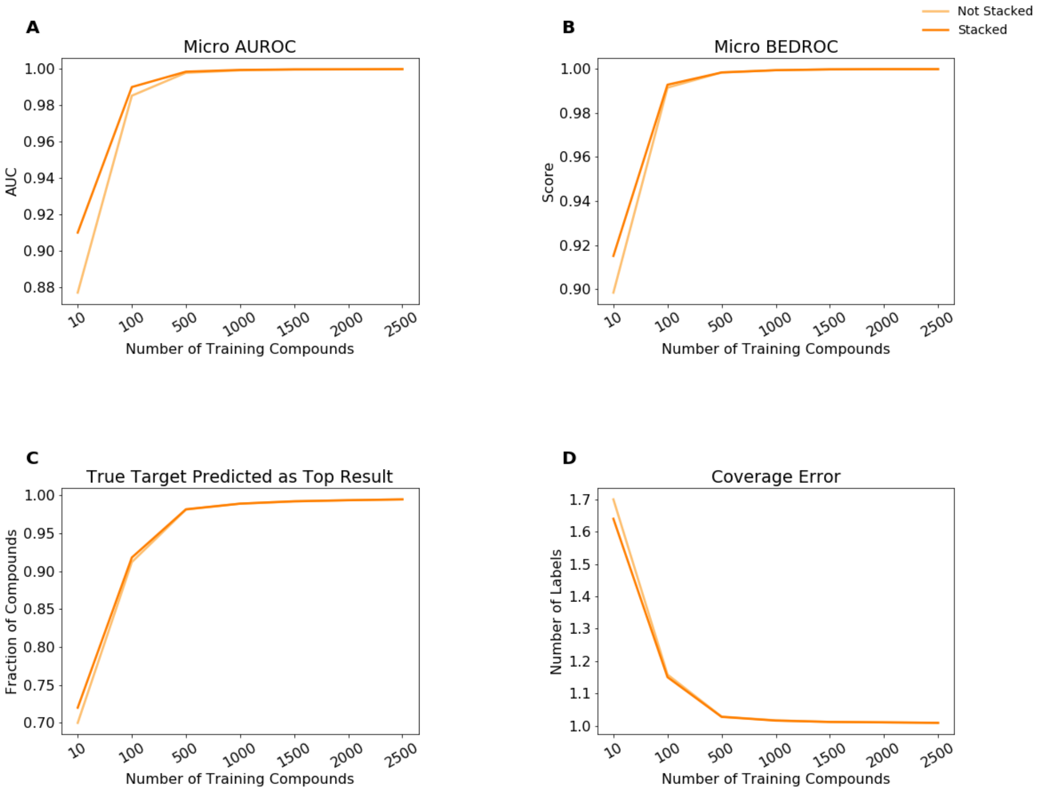


Figure A.6 KNN_MLP_RF Model Performance with Different Training Set Sizes. Model performance for stratified 10-fold cross-validation on datasets containing various numbers of compound training records for each protein target label for the KNN_MLP_RF classifier. "Not Stacked" refers to the mean probabilities of model predictions when more than a single model is listed. "Stacked" indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

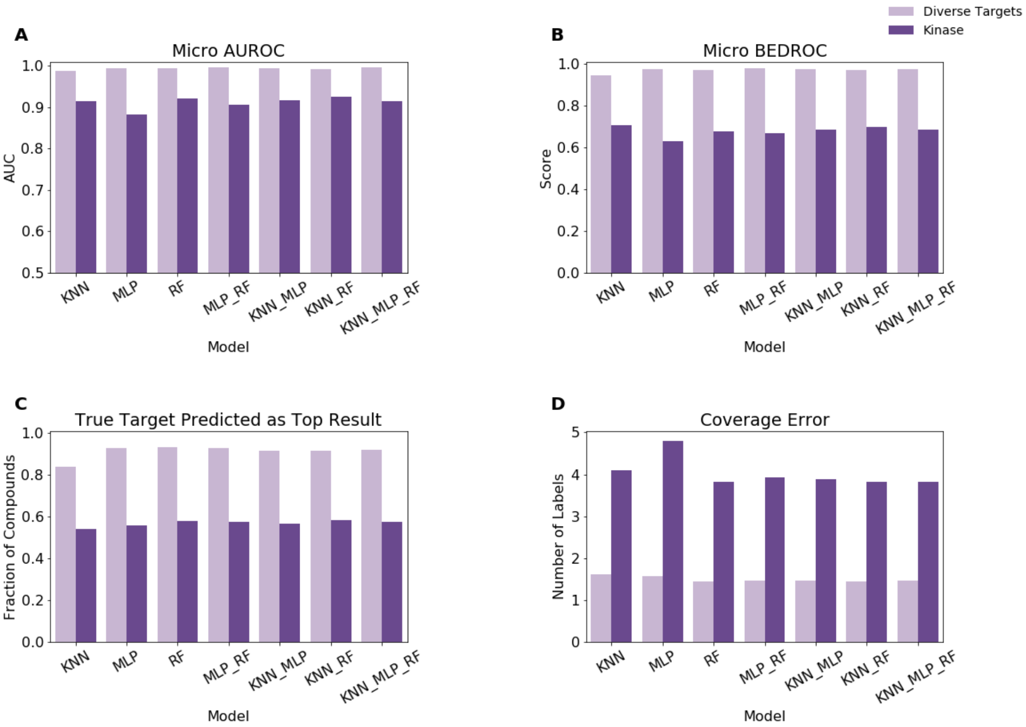


Figure A.7 Stacked Classifier Model Performance with Diverse Target Labels. Model performance for stratified 10-fold cross-validation on the diverse target and kinase datasets for “Stacked” classifiers. “Stacked” indicates that the probability predictions of the listed models were passed to the logistic regression to obtain the final predicted probabilities. Model performance as measured by (A) micro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve, (B) micro-averaged Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), (C) the fraction of compounds which yielded a true target as the top prediction, and (D) coverage error are shown.

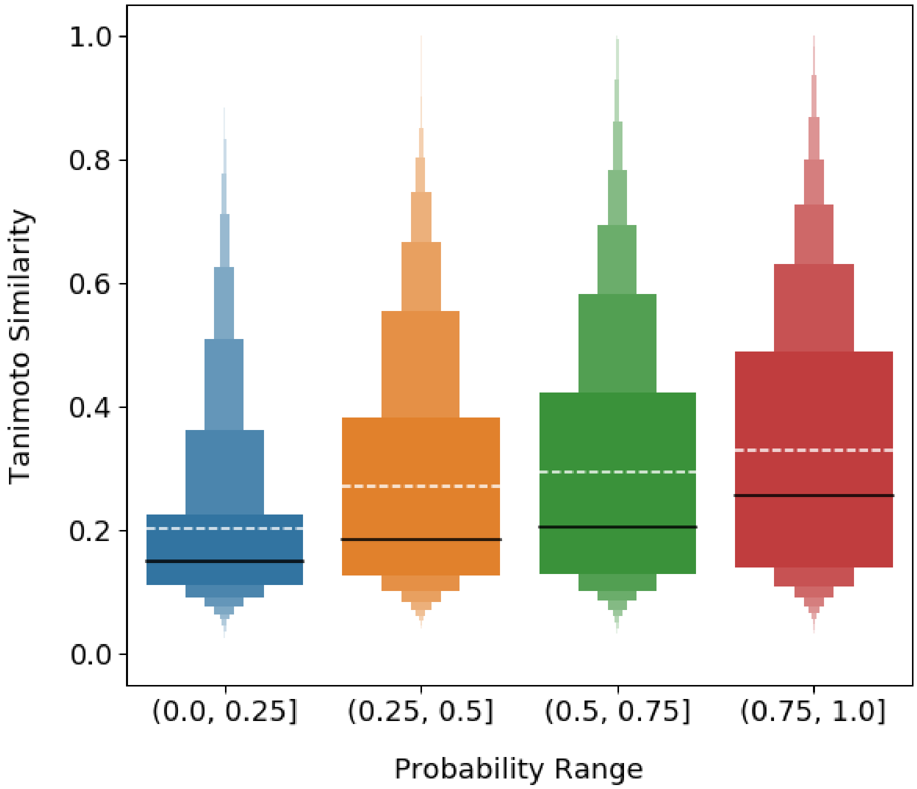


Figure A.8 Letter-Value Plot of Aggregated Pairwise Similarity Distributions for the KNN_RF Stacked Classifier on the Synthetic Compound Test Set. Letter-value plot showing the aggregated pairwise similarity distributions for synthetic test compounds and synthetic training compounds for known positive protein target labels in a cross-validation fold. Similarity distributions were aggregated based on the predicted probability from the KNN_RF stacked classifier for the known protein targets of each synthetic test compound. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

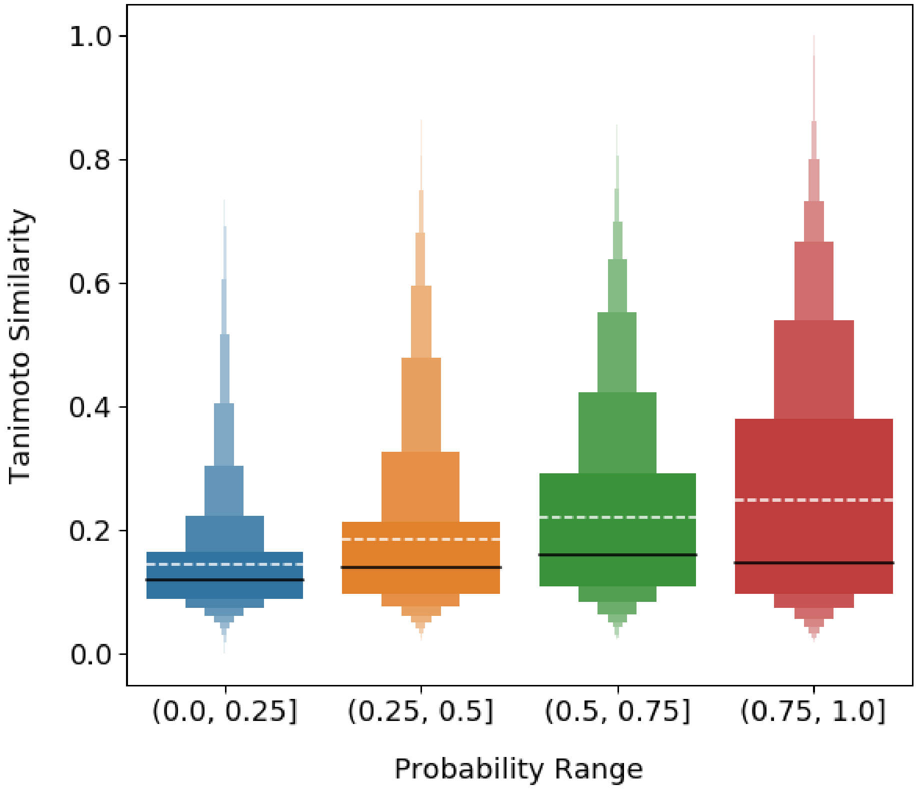


Figure A.9 Letter-Value Plot of Aggregated Pairwise Similarity Distributions for the KNN Base Classifier on the Natural Product Benchmark Set. Letter-value plot showing the aggregated pairwise similarity distributions benchmark natural product compounds and synthetic training compounds for known positive protein target labels. Similarity distributions were aggregated based on the predicted probability from the KNN base classifier for the known protein targets of each natural product. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

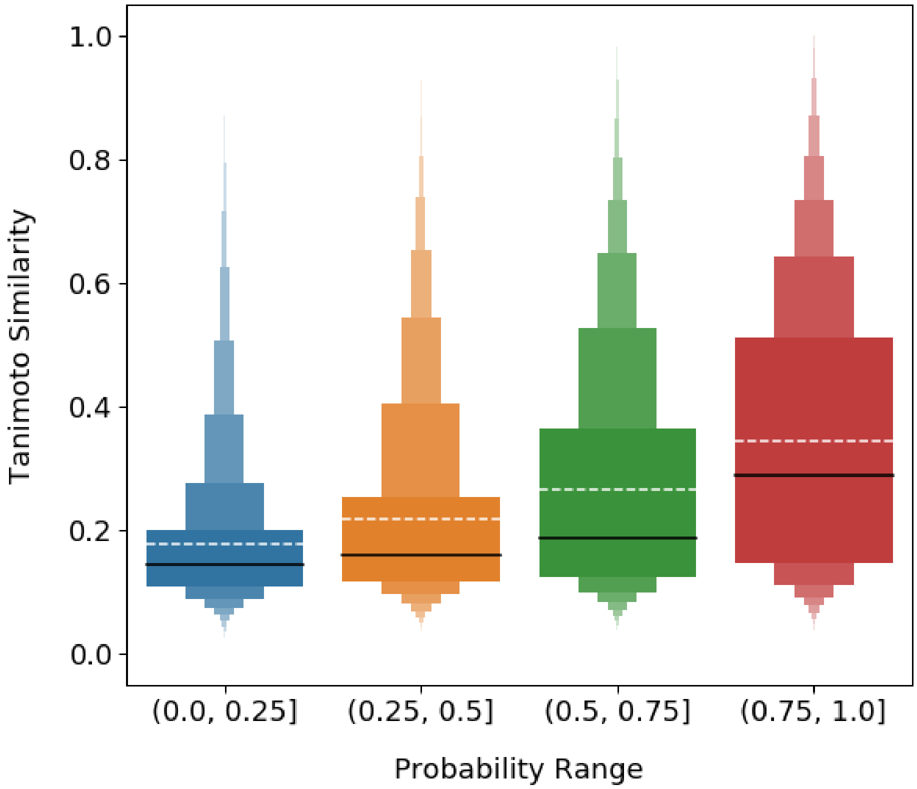


Figure A.10 Letter-Value Plot of Aggregated Pairwise Similarity Distributions for the KNN Base Classifier on the Synthetic Compound Test Set. Letter-value plot showing the aggregated pairwise similarity distributions for synthetic test compounds and synthetic training compounds for known positive protein target labels in a cross-validation fold. Similarity distributions were aggregated based on the predicted probability from the KNN base classifier for the known protein targets of each synthetic test compound. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

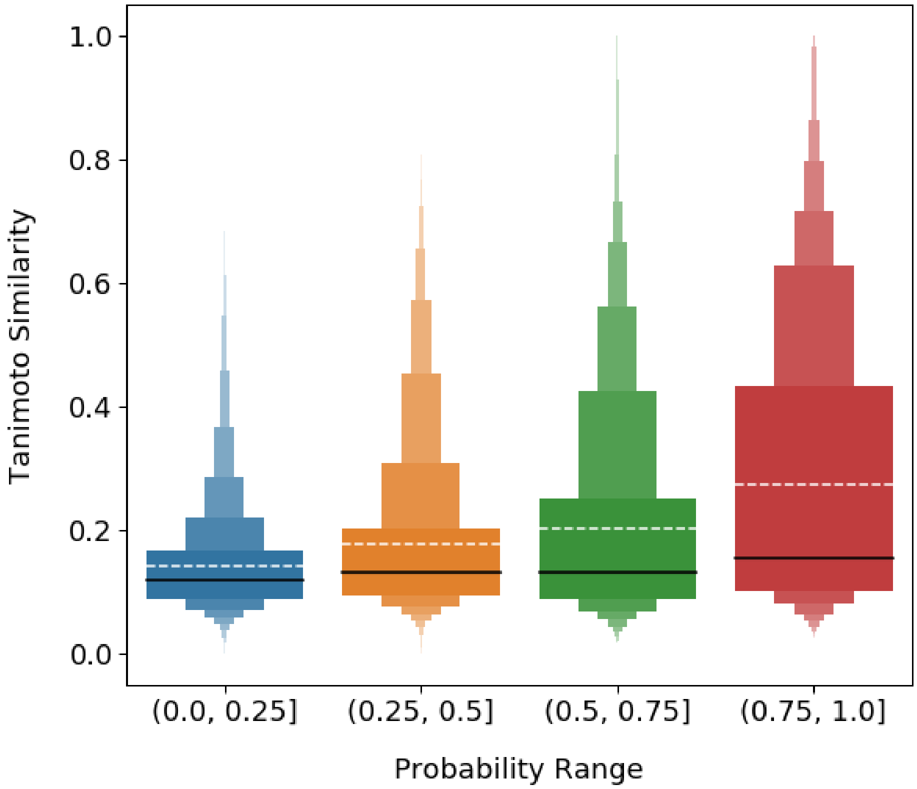


Figure A.11 Letter-Value Plot of Aggregated Pairwise Similarity Distributions for the RF Base Classifier on the Natural Product Benchmark Set. Letter-value plot showing the aggregated pairwise similarity distributions benchmark natural product compounds and synthetic training compounds for known positive protein target labels. Similarity distributions were aggregated based on the predicted probability from the RF base classifier for the known protein targets of each natural product. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

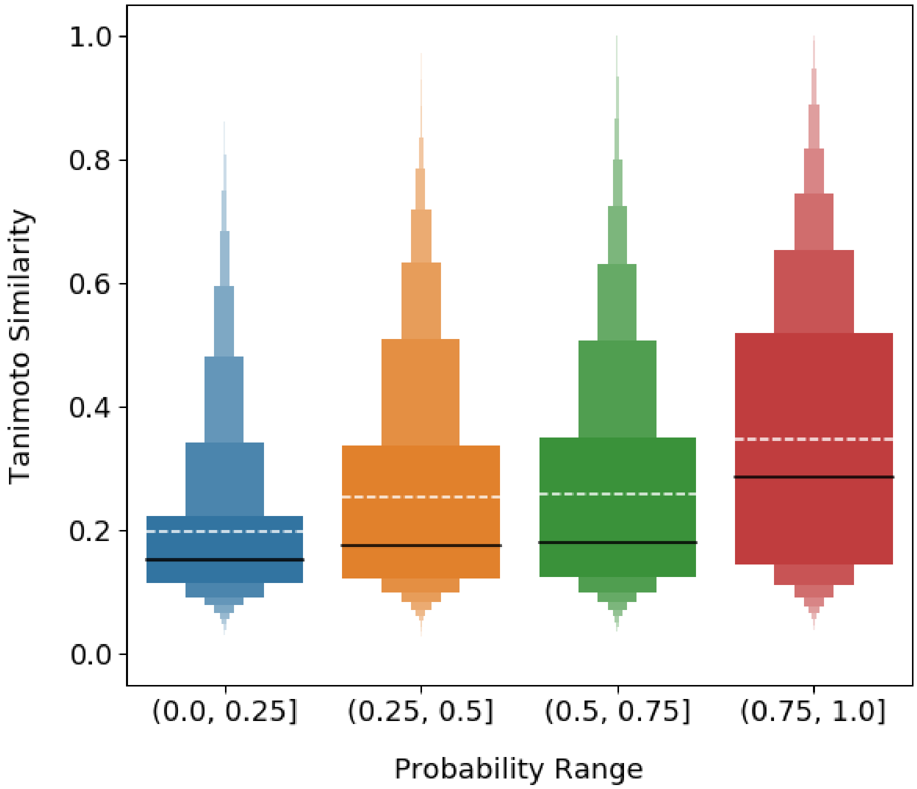


Figure A.12 Letter-Value Plot of Aggregated Pairwise Similarity Distributions for the RF Base Classifier on the Synthetic Compound Test Set. Letter-value plot showing the aggregated pairwise similarity distributions for synthetic test compounds and synthetic training compounds for known positive protein target labels in a cross-validation fold. Similarity distributions were aggregated based on the predicted probability from the RF base classifier for the known protein targets of each synthetic test compound. The solid black line represents the median and the white dashed line the mean. Letter-value plots are similar to box plots but provide more information about the tails of a distribution. Each box represents a portion of a distribution according to its width shown. The widest box is identical to the interquartile range in a box plot and represents 50% of the data. The next widest boxes, as more than one box now has identical width, comprise 25% of the data. Those boxes are present directly above and below the interquartile range. For each successive box width reduction, the amount of data represented is halved.

Appendix B. Datasets and Code.

Snapshots of the mentioned GitHub repositories are also available in the accompanying supplemental data file.