Machine learning assisted virtual screening architecture in drug discovery

Youlin Liu

Oct. 1st

Contents

1 Project Summary

This document proposes a machine learning architecture for structure based virtual screening in drug discovery. In drug discovery, virtual screening is categorized into structure based and ligand based depending on the availability of structural data of the receptor. This proposal focuses on structure based virtual screening, and proposes a two step improvement that targets the docking step and the screening step in virtual screening as aim I and aim II respectively. Both of the improvement utilize a machine learning algorithm, specifically, support vector machine (SVM), which is considered to be a robust and well-performing algorithm for small datasets. Aim III suggests the application of previously constructed architecture as aim I and aim II combined for the screening of Alzheimer's disease related targets.

2 Specific aims

Aim I: build a machine learned predictor for tuned docking parameters In this aim, docking parameters will be generated and trained with SVM. The generation of docking parameters rely on the CASF (Comparative Assessment of Scoring Functions) protocol that is newly published for the evaluation of scoring functions. A mini-batch gradient descent algorithm will be used to update the docking parameters until it passes a user-define threshold. At the end of this step, docking parameters specific to individual ligand-receptor complex and the complex's molecular descriptor will be used to train a SVM so that, the SVM predictor can generate docking parameters based on an input molecular descriptor.

Aim 2: Active selection of ligand to-be-docked in virtual screening Instead of randomly going through the ligands to be tested in virtual screening, this proposals develops an active selection that involves SVM, in which all of the ligands in ligand library to be docked will be projected onto a feature space, with a hyperplane separating the so-far known active ligands and non-active ligands. As screening processes more ligand will be labeled and the hyperplane will be dynamically updated. The key is to always use the ligand that's furtherest away from the hyperplane in the active ligands side.

Aim 3: applicate the combined virtual docking architecture for Alzheimer's disease related receptors This aim applicated the previously built architecture in the realistic scenario of searching inhibitor (binder) for proteins relevant to Alzheimer's disease.

3 Research Strategy

3.1 Significance

Drug development is a very time-consuming and expensive process. Of the drug development pipeline as indicated in figure 1, while each of these steps are very much time and capital consuming, these are inevitable steps that a putative drug has to go through to eventually make it to the market. Recently with the advent of computer science, much effort has been put into *in silico* to aid the 'Primary Assays high through-put, *in vitro*' step as shown in figure 1. As a main contribution brought by the growth of computing powers, virtual screening is the alternative of actual physical high throughput screening (HTS).

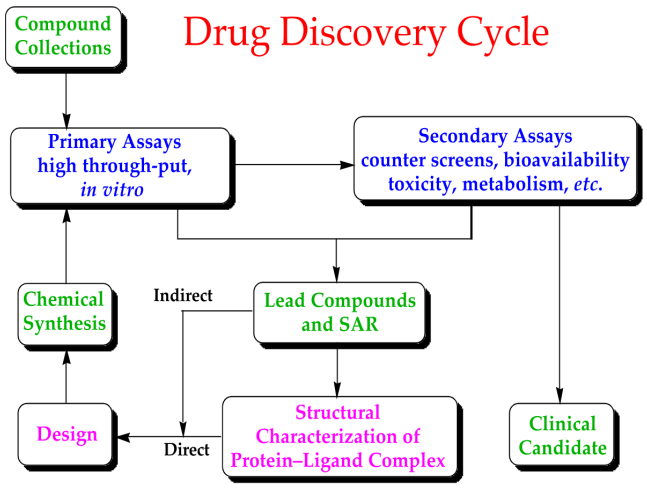

Figure 1: Drug development steps[1]

Virtual screening is defined as 'automatically evaluating very large libraries of compounds using computer programs'. With the ever growing size of databases of proteins (targets) and compounds (drug candidates), A crucial drug identification step can be described as, given a protein that is known to be closely related to a disease that is desired to be cured, of the database of small molecules, which can be a potential drug candidate?

As a solution to the scenario described above, much efforts has been put into virtual screening that aims to identify a reliable drug hit that can be put into the drug development pipeline that eventually get optimized to a commercial drug (also known as hit-to-lead, lead-to-drug).

Among the efforts of developing virtual screening methods, one most popular model is structured-based drug design. In structure based drug design, the structure (3D coordinates) of the ligand and receptor is known via other analyzing methods such as X-ray crystallography, NMR or homology modeling [2]. Structure-based drug design mainly comprise of molecular docking studies. Given the nature of diseases that most involves the malfunction of enzymes and other forms of proteins, protein-ligand docking is of special interest among other docking types (protein-protein, protein-DNA etc.).

This proposal focuses on the improvement of virtual screening via the molecular docking and ligand selection method. Which, if proven to be functional, has the potential of applying to many other receptors that are yet to find the perfect drug.

3.2 Innovation

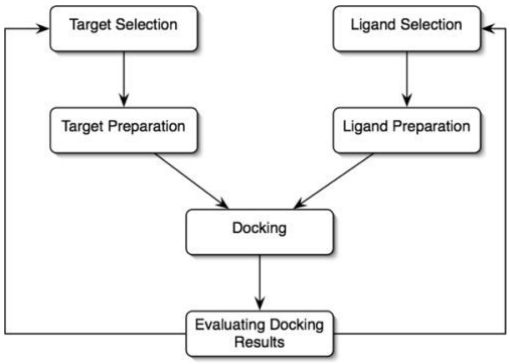

Figure 2: General docking steps[3]

The general process of structure based virtual screening can be briefly described as below:

> Identify the protein of interest as target, prepare a library of ligands that is to be screened, perform molecular docking on each one of the ligand and evaluate the binding affinity of each protein-ligand pair and decide with ligands are to be used for subsequent optimization (hit to lead) and bioassay.

Despite the fact that there are a variety of molecular docking softwares, and much effort has been putting into building more accurate scoring functions for predicting binding affinities, the execution of molecular docking in the virtual screening scenario is often *using the default docking parameters*.

The innovation of aim I will be to build a predictor for generating docking parameters that's tuned to the ligand molecule. The 'predictor' is build via a SVM algorithm, in which the training dataset will be generated from an evaluating protocol.

Aim II focuses on increasing the efficiency of screening, as opposed to 'go through' the ligand library and perform molecular docking one by one, every ligand that has been docked will be labeled as 'binder' or 'non-binder', along with the untested ligands, the ligand library will be projected to a higher space and and the labeled data will serve to help choosing what ligand to dock next. As the screening goes the amount of labeled data will grow as well, hence this is an active learning process in terms of ligand selection.

Aim I and II combined will be implemented in finding a hit for Alzheimer's disease (AD) related targets in aim III. AD is a neurodegenerative disease with no known cure and prevention.

3.3 Approach

3.3.1 General description

In structure based virtual screening, the key step is molecular docking. Molecular docking addresses two problems, the docking problem and the binding affinity prediction problem. The docking problem can be described as the search for the precise ligand conformations and orientations (also referred to as posing) within a given targeted protein where the structure of the protein is known. The binding affinity prediction problem addresses the question of how well the ligands bind to the protein (scoring).

Therefore, the general docking program would include two parts: a searching algorithm and a scoring function. While the searching problem is relatively well resolved with random or stochastic algorithms [4], on going efforts are still being put into developing a more accurate scoring functions. Up until now, many scoring functions based on different methodology have been developed, currently a range of different scoring

functions have incorporated inside of the mainstream docking softwares such as DOCK[5], AutoDock[6], GOLD[7], Glide[8], and Surflex-Dock[9].

Of the various model type of scoring functions out there which are traditionally categorized into force-field-based, empirical, knowledge- based, with each of the categories has it's strengths and weaknesses. Shortly, with the approximation researchers have to adopt in some models, or, limitation of the training set sizes for building the model, it is summarized that no single docking program has dominative advantages than other programs.[10]. While the average drug company has a wide choice of docking software to choose from, there's no agreement of one single scoring functions outperforms the rest, despite of whether it's academically free or commercially expensive.

In fact, depending on what architecture or training data set (for empirical based scoring functions) per scoring function is using, one scoring function tend to perform better for certain protein-ligand docking types. And it is often the case that certain compounds are identified via virtual screening to be a 'binder' while subsequent experimental results show that it's not.

With the advent of computational power, there has been a trending of implementing machine learning algorithm into molecular docking, thus given the forth category of 'machine-learning based' or 'descriptor-based' scoring function. Based on the fact that molecular docking problem can be described as building a model with a known-to-bind data sets that can be treated as the training sets, which suits into the category of supervised-learning in machine-learning concepts, multiple algorithm including Naïve Bayes, k-Nearest Neighbors, Support Vector Machine, Random Forests and Artificial Neural Networks [11].

While machine learning could be a powerful tool, the current architecture of using machine learning in the molecular docking field either be viewed as 'black box feature selection' [12] with the danger of over fitting training data, or has complicated feature parameters[13].

In aim I, machine learning will be used in combination with a traditional scoring function (specifically, smina will be used to illustrate the data flow, but it can be any other scoring function of with enough user defined flexibility). Shortly, after a normal 'docking' performed by smina, docking parameters will be updated iteratively until it passes the CASF protocol on a user defined threshold. Subsequently molecular descriptors will be generated corresponding to the optimized docking configuration, a machine learning algorithm (SVM) will learns the relationship between the descriptors and the docking configurations. The general logic flow is illustrated in figure 3, the specific terms that appear in this flow chart will be explained in later sections.

Aim II steps back to look at virtual screening in a bigger picture. The size of small molecular library is usually on the order of thousands even after pre-filtering. It's natural to dock one by one with random order, but a selective method promises to be more time efficient if the amount of binders are predetermined. Again, a machine learning method is used in this step.

Finally, aim III proposes the application of the architecture to a realistic scenario of Alzheimer's disease. While AD itself is notoriously famous for complicity of cause and different hypothesis for treatment, this proposal focuses on the amyloid cascade hypothesis in which acetylcholinesterase (AChE) is targeted using small molecule inhibitor.

3.3.2 Choice of molecular docking software: Smina

To provide some background information, AutoDock [6] is one of the most cited docking software, and AutoDock Vina [14] is the improved docking software in which the fundamental model of scoring was improved and has since become the popular academical docking software.

In a typical docking event, the docking software would take in the respective structure file of the ligand and the receptor, along with the instruction of docking specific parameters and performs pose searching

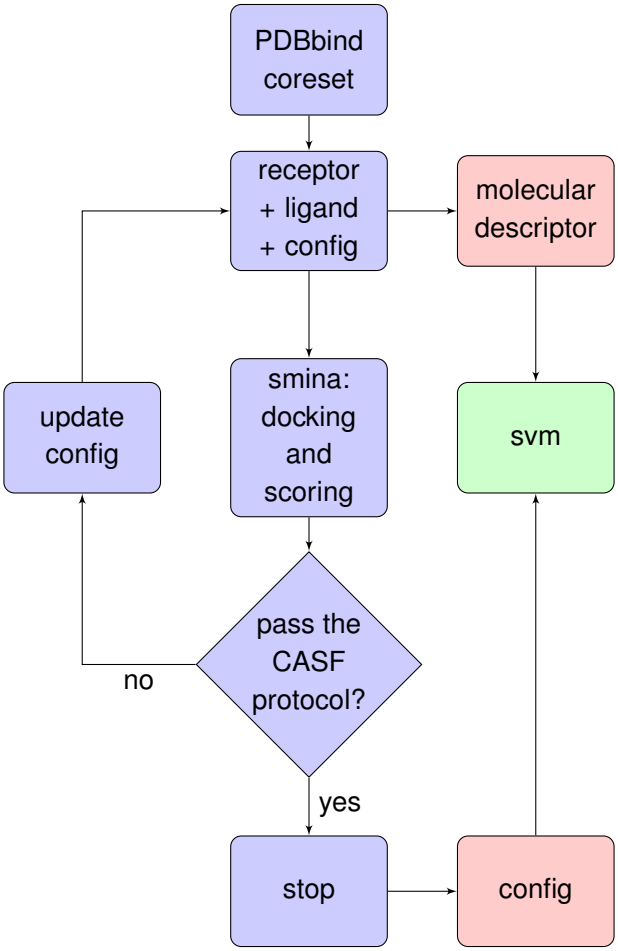

Figure 3: Systematic illustration of the proposed algorithm

and binding affinity calculation subsequently [15]. With any docking software (not just Vina), the parameters must include a minimum of search space, but optional parameters can be specified, take Vina for example, the additional specifications can be:

- seed: random seed for where to start sampling inside of the search box
- exhaustiveness: how much should the sampling algorithm explore into different binding poses
- num_modes: how many most optimized poses the user want Vina to output
- energy_range: maximum energy difference between the best binding mode and the worst one displayed in units of kcal/mol

The search space would ideally be the minimum box that contain the protein's natural co-crystallized ligand, but the larger it is, the more searching Vina will have to do.

Despite the face that Vina offers 'tweaking' of the starting conditions, it doesn't provide much user-defined functionality as in which of the bonds or side chains are flexible.

Note that, there are multiple (more than 50) docking softwares out there for virtual screening, the choice of smina is based on it's versatility for instructing per given protein ligand docking event as described above. Below shows an example of how smina can be 'tweaked' to allow for different docking configurations (which the authors refer to as 'versatile scoring function').

```
1  1.0  ad4_solvation(d-sigma=3.6,_s/q=0.01097,_c=8)  desolvation, q determines
       whether value is charge dependent
2  1.0  ad4_solvation(d-sigma=3.6,_s/q=0.01097,_c=8)  in all terms, c is a
       distance cutoff
3  1.0  electrostatic(i=1,_^=100,_c=8) i is the exponent of the distance
4  1.0  electrostatic(i=2,_^=100,_c=8)
5  1.0  gauss(o=0,_w=0.5,_c=8)   o is offset, w is width of gaussian
6  1.0  gauss(o=3,_w=2,_c=8)
7  1.0  repulsion(o=0,_c=8)  o is offset of squared distance repulsion
8  1.0  hydrophobic(g=0.5,_b=1.5,_c=8)   g is a good distance, b the bad distance
9  1.0  non_hydrophobic(g=0.5,_b=1.5,_c=8) value is linearly interpolated between
       g and b
10 1.0  vdw(i=4,_j=8,_s=0,_^=100,_c=8) i and j are LJ exponents
11 1.0  vdw(i=6,_j=12,_s=1,_^=100,_c=8) s is the smoothing, ^ is the cap
12 1.0  non_dir_h_bond(g=-0.7,_b=0,_c=8) good and bad
13 1.0  non_dir_h_bond_quadratic(o=0.4,_c=8) like repulsion, but for hbond, don't
       use
14 1.0  non_dir_h_bond_lj(o=-0.7,_^=100,_c=8)  LJ 10-12 potential, capped at ^
15 1.0 acceptor_acceptor_quadratic(o=0,_c=8) quadratic potential between hydrogen
       bond acceptors
16 1.0 donor_donor_quadratic(o=0,_c=8) quadratic potential between hydroben bond
       donors
17 1.0  num_tors_div div constant terms are not linearly independent
18 1.0  num_heavy_atoms_div
19 1.0  num_heavy_atoms  these terms are just added
20 1.0  num_tors_add
21 1.0  num_tors_sqr
22 1.0  num_tors_sqrt
```

```
1.0   num_hydrophobic_atoms
1.0   ligand_length
```

Listing 1: smina user defined parameters

Conveniently the `1.0` s are weights to be put into consideration when docking. Taking out the parameters that are not optimizable, but rather, rely on the property of the receptor and the ligand, the optimizable weights will be optimized in an iterative style. For each ligand-receptor complex, the end goal is to construct a vector containing the optimal docking parameters (weights) associated with it.

3.3.3 Usage CASF evaluation protocol for optimizing docking parameters

The default vector in smina, if not user defined, will be a vector of ones. We aim to tune it to a more complex-specific vector which potential gives better binding affinity prediction. This process will be phrased as 'optimizing docking parameters'. To optimize the docking parameters, we propose to use the CASF evaluation protocol.

Generally, scoring functions are evaluated in the context of molecular-docking trials. Specifically, evaluation of the performance of scoring functions in dicking has focused predominantly on two measure.

First, the ability to accurately reproduce the co-crystallized ligand binding poses in crystal structures, it is arbitrarily defined that ligand docking is most accurate if the top ranked pose has a heavy atom root mean square deviation (RMSD, atomic positions as the measure of the average distance between the atoms of superimposed proteins) less than 2Å from the location of the crystallized ligand. Second, the enrichment factor (EF) of the docking and soring algorithm after a virtual screening event. The EF is defined as the accumulated ratio of active ligands found above a certain percentile of the ranked database containing active and inactive ligands ('binders' and 'non-binders'). A higher EF value at the define percentile normally indicates a better scoring function. Another evaluation that is less frequently exercised is the accuracy in prediction of the experimental binding affinity, even though this was the original definition for scoring functions in molecular docking. This is not only because of the inability of the current scoring functions available, but also the data quality deposited in the protein data bank by various researchers not following a universally experimental condition when performing the data acquisition.

To date, several publications has addressed the performance of scoring functions by comparison studies. while the general capabilities described above are covered, there hasn't been a universal benchmark until March, 2018, When Li et al. [16] introduced a protocol in which the capabilities of one given scoring function is summarized into four indicators: scoring power, ranking power, docking power and screening power. Li et al. [16] also introduced the CASF benchmark datasets where the evaluation of the datasets are parameterized and described as 'a common language among the scoring function community'.

There are currently more than 35,000 crystallographic or NMR structures of proteins or nucleic acids available from the Protein Data Bank (PDB) [17]. PDBbind is database is a comprehensive collection of the experimentally measured binding affinity data for all types of biomolecular complexes deposited in the PDB[18]. The CASF benchmark [19] is development upon PDBbind, including a high quality data set and quantitative methods for conducting performance evaluation.

The actual complex dataset that the performance of the scoring functions will be evaluated on will be referred to as 'core set'. The core set consists of 195 protein-ligand complexes in 65 protein clusters. It is designed to cover protein-ligand complexes formed by diverse proteins so that they can be representative of the target-to-be in a real virtual screening situation. The complexes in the core set are chosen such that they have a reliable crystal structure and experimental binding data, and they are also chose to be drug relevant

in the sense that most (82%) of the ligand molecules are evaluated to obey the Lipinski's 'rule of five' as the primary filter for drug-likeness, and the majority of proteins (78%) are validated as potential drug candidate. Most importantly, the 195 complexes can be categorized into 65 clusters with each cluster containing a representative of 3 complexes that span nearly 10 orders of magnitude in binding affinity (Kd = 10 mM – 1 pM).

On top of the high quality data set, the CASF benchmark define the power of a scoring function as scoring power, ranking power, docking power and screening power. While the original intention of a 'scoring' function is to estimate the binding affinity of a given protein and ligand pair, the practical application scenario is far more complicated, in addition to the fact that scoring functions are modules that are incorporated inside of molecular docking softwares, while the software relies on the scoring function on the most part for predicting a binding affinity, it also contains a searching function (pose sampling) functionality prior to determine 1. active site of the target protein 2. binding pose of the given ligand. Therefore, to just evaluate on the scoring function, it is necessary to decouple these variable from the scoring function, in which case the decoy ligand-binding poses for each complex is to be generated in advance and the scoring functions will be instructed to predict a binding affinity based on the decoy poses.

The 'powers' of per given scoring functions are defined as follows according to the CASF protocol.

- Scoring power: the capability to produce binding score in linear correlation with experimental bind data. Quantitatively, by Pearson correlation coefficient (R) and the standard deviation between predicted and experimental.
- Ranking power: the capability of rank the ligands of a given target protein correctly with given binding pose. Quantitatively, define a 'success rate' of $\frac{correct\ ranking\ sets}{overall\ ranking\ sets}$.
- Docking power: the capability to identify the native ligand-binding pose among computer generated decoy poses. Quantitatively, if the RMSD value between the native binding pose from the top-ranked binding pose fell below a predefined cutoff.
- Screening power: capability to identify true binders for a given target protein among a pool of random molecules. Here define enhancement factor (EF) defined as below: (1% can be set to other user defined threshold)

$$EF_{1\%} = \frac{number\ of\ true\ binders\ among\ top\ 1\% candidates}{(total\ number\ of\ true\ binders\ for\ this\ target\ protein) \times 1\%}$$

To fit CASF protocol into the optimizing step, all of the 195 data will be docking using initial docking parameter (the vector of ones) and the initial powers of smina will be evaluated. We propose to set the 'passing threshold' be 10 percent more (can be changed practically in view of computational time) than the initial powers, and the docking parameters will be updated and re-evaluated after each update.

CASF protocol provide bash scripts that runs and output numbers as the representation of each of the powers and the outliers will be updated first. For each 'update of parameters', a simple gradient descent algorithm will be used.

3.3.4 Generating molecular descriptors

Molecular descriptor is a necessity in view of converting molecular to 'computer language', it will serve as the input of any machine learning algorithm. Shortly, molecular descriptors are numerical values assigned to

structures, they can be physiochemical properties (molecular weight, logP), or more sophisticated indicators such as topological indices, maximum common substructures, molecular fields[20].

Usually, the actual 'molecular descriptor' utilized in machine learning scenarios are more complicated than what's described above, and there's are novel algorithms that doesn't not rely on simple linear descriptors (sun as constitutional neural network [21]). To show a illustrative example of how molecular descriptors as numerical matrices can be used describe the molecular itself, consider an example based on elemental atom types as shown below: [22]

For both the protein P and the ligand L:

$$\{P(j)\}_{j=1}^{9} = \{C, N, O, F, P, S, Cl, Br, I\} \quad \{L(i)\}_{i=1}^{9} = \{C, N, O, F, P, S, Cl, Br, I\}$$

The occurrence count for a particular j-i atom type pair is defined as:

$$\varkappa_{Z(P(j)),Z(P(i))} \equiv \sum_{k=1}^{K_j} \sum_{l=1}^{L_i} \Theta(d_{cutoff} - d_{kl})$$

where d_{kl} is the Euclidean distance between $k-th$ protein atom of type j and the $l-th$ ligand atom of type i calculated from the PDBbind structure. K_j is the total number of protein atoms of type j and L_i is the total number of ligand atoms of type i in the considered complex; Z is a function that returns the atomic number of an element and it is used to rename the feature. Θ is the Heaviside step function that counts contacts within a defined cutoff. For example, when $d_{cutoff} = 12$Å it means that it counts the atoms within a 12Å neighborhood of the given atom.

This is an simplified illustration of the most prevailed machine learning descriptors that rely on a pairwise feature selection. Depending on the end users' intension, there are many ready to use commercial softwares that generate numerical descriptors for machine learning input.

In this proposal, Dragon will be used for generating the molecular descriptors. The choice of based on that fact that Dragon is able to calculate 5270 molecular descriptors, covering most of the various theoretical approaches. Of the descriptors that Dragon offers to calculate, choice will be made among the *3D Atom Pairs*, *3D matrix-based descriptors* , *Topological indices* descriptors as these will mostly likely to be relevant.

3.3.5 SVM for learning the relationship between docking parameters and ligand molecular descriptors

The simplicity of this proposal is that it uses well-built machine learning structure instead of building a model that involves the very tedious molecular dynamics model building which is not just time consuming, but also computationally expensive [13].

To this point, we have described the input for SVM, which will be a vector that specifies the docking parameter, and the corresponding molecular descriptor. And we will have this information for all of the 195 complexes.

Below offers a short description of SVM that will be trained will the dataset described above to serve as a predictor that per given molecular descriptor, it outputs the optimal docking parameters.

SVM is chose as the machine learning algorithm based on the advantages below: [23]

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.

- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Before illustration of how the versatility of the kernel functions possible greatly suits the need of this proposal, here is a brief introduction of how SVM works.

In SVM, each data item is plotted as a point in n-dimensional space (where n is the number of features predetermined) with the value of each feature being the value of a particular coordinate. Then classification is performed by finding the hyper-plain that differentiate the two classes. SVM can be used for regression instead of classification with minor differences in which the hyper-plane needs to be individualized. The function that projects the input data into the feature-dimension space is defined as a kernel function. The choice of the suitable kernel function can be crucial to the success of the training as shown in figure 4. It is often suggested that the kernel function be chosen based on prior knowledge of the dataset. As this case is mostly likely a non-linear model, the Radial Basis Function kernel would be used as the default and would be updated if needed.

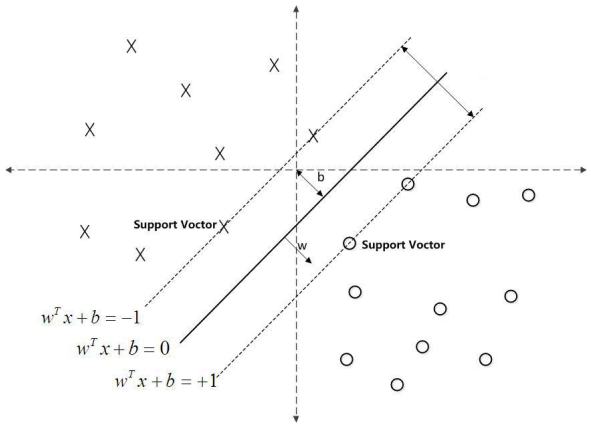

Figure 4: Illustration of how SVM multi-class classification on a data set with different kernel functions[23]

Practically, the 195 datasets will be divided into 3 subsets, with mixed binding affinities in each sets. For training and testing, the leave-one-out cross validation method will be used.

3.3.6 Aim II: active selection of next ligand

In the virtual screening pipeline, the simplest selection strategy is to choose new compounds at random for testing. It's intuitive that with random selection, the number of hits grows linearly with the total number of compounds that have been tested. Since that most of the compounds will likely to be 'non-binder', this random selection will be a very inactive process.

Warmuth et al. [24] suggested an active learning with SVM will be employed in the virtual screening pipeline in this proposal. Refer to figure 4 in section 3.3.5 for the concept of SVM. Practically, all of the ligands to be screened will be projected in the feature space, a hyperplane will be calculated in this feature space to separate the known binders and non-binders with minimized cost. And the ligand to be docked next will be the data point that's furtherest away from this hyperplane on the binder's side. It's intuitive that this data point is most likely to be categorized as a binder. This selection method is refer to as 'active learning' because as the screening goes, the number of labeled data grows and the hyperplane is updated dynamically.

While this active learning selection strategy has be previously described in Warmuth et al. [24], it was never incorporated in the virtual screening event. The SVM algorithm that people utilized nowadays refers to the direct relationship of binders and the compound database without the step of molecular docking, in other words, SVM substitutes the molecular docking as the evaluation for binding affinity.

3.3.7 Validation of the proposed virtual screening architecture: DUD·E

DUD·E is full named as 'a database of useful decoys: enhanced.' [25], as the name suggests, it is an enhanced version of the previously prevailed DUD (directory of useful decoys). It is designed to help benchmark molecular docking programs by providing challenging decoys. It contains 22,886 active compounds and their affinities against 102 targets, an average of 224 ligands per target 50 decoys for each active having similar physico-chemical properties but dissimilar 2-D topology.

We will perform simulated virtual screening using this DUD·E benchmark datasets, and before& after ROC, AUC graphs will be plotted to evaluate the efficiency of the proposed architecture.

3.3.8 Aim III: applicate the combined docking software in virtual screening of drug lead for Alzheimer's Disease

This section aims to applicate the combined new scoring function to the classical virtual screening (VS) scenario. Specifically, a small molecular library is run through with the combined virtual screening architecture.

Alzheimer's disease (AD) is the most common form of dementia in the elderly with progressive cognitive decline and memory loss, it affects 44 million people worldwide as of 2016 [26]. Unfortunately, AD and other neurological diseases are notoriously difficult to treat. The AD drugs currently available only alleviate symptoms[27] and it is imperative to keep searching for a cure as in the most recent AD related drug to pass clinical trials (Memantine) happened in 2003.

Although there are many mechanisms for AD, it is generally agreed that there are only two classes of drugs currently available for AD treatment. This proposal will focus on acetylcholineterase inhibitors (AChEI), which increase acetylcholine concentration in cholinergic gynaptic clefts.

Primarily, BACE1, the M1 subtype of mAChR, APP, CDK5, and GSK-3β are the potential targets [11]. And ZINC [28] small compound library will be screened for potential drug hits.

Several recently studies have been done in search of AChEI [29, 30, 31, 32, 33], while only [29] used a machine learning based virtual screening. It would be mostly interesting to compare the filtered ligand results with the studies shown above.

3.3.9 Concluding remarks

This proposal builds upon currently available resources and construct machine learning architecture in a simple design, and applicates the architecture the ongoing search of AChEI.

The methodology is straightforward yet involved a traditionally robust machine learning algorithm, in future research can be applicate to the search of hits for other proteins as well.

References

[1] Drug development, 2018. URL `https://en.wikipedia.org/wiki/Drug{_}development`.

[2] Tom L Blundell. Structure-based drug design. *Nature*, 384(6604):23, 1996.

[3] Garrett M. Morris and Marguerita Lim-Wilby. Molecular Docking. pages 365–382. Humana Press, 2008. doi: 10.1007/978-1-59745-177-2_19. URL `http://link.springer.com/10.1007/978-1-59745-177-2{_}19`.

[4] Pavlína Rezácová, Dominika Borek, Shiu F Moy, Andrzej Joachimiak, and Zbyszek Otwinowski. Crystal structure and putative function of small Toprim domain-containing protein from Bacillus stearothermophilus. *Proteins*, 70(2):311–319, 2008. ISSN 08873585. doi: $10.1002/\mathrm{prot}$.

[5] Todd J A Ewing, Shingo Makino, A Geoffrey Skillman, and Irwin D Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided. Mol. Des.*, 15(5): 411–428, 2001.

[6] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 30(16):2785–2791, 2009.

[7] Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.

[8] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, and Others. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, 2004.

[9] Ajay N Jain. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, 46(4):499–511, 2003.

[10] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: The prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys.*, 18(18):12964–12975, may 2016. ISSN 14639076. doi: $10.1039/\mathrm{c6cp01555g}$. URL `http://xlink.rsc.org/?DOI=C6CP01555G`.

[11] Kristy A Carpenter and Xudong Huang. Machine Learning-based Virtual Screening and Its Applications to Alzheimer ' s Drug Discovery : A Review. pages 1–12, 2018. doi: $10.2174/138161282466618060712403$8.

[12] Joffrey Gabel, Jérémy Desaphy, and Didier Rognan. Beware of machine learning-based scoring functions-on the danger of developing black boxes. *J. Chem. Inf. Model.*, 54(10):2807–2815, oct 2014. ISSN 15205142. doi: $10.1021/\mathrm{ci500406k}$. URL `http://pubs.acs.org/doi/10.1021/ci500406k`.

[13] Yu Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today*, 23(8):1538–1546, 2018. ISSN 18785832. doi: $10.1016/\mathrm{j.drudis.2018.05.010}$. URL `https://doi.org/10.1016/j.drudis.2018.05.010`.

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

[15] Kay Diederichs, Stefano Forli, Ruth Huey, Michael E Pique, Michel Sanner, David S Goodsell, and J Arthur. HHS Public Access. *curr opin struct biol.*, 11(5):60–68, 2016. doi: $10.1016/\mathrm{j.sbi.2015.07.003}$. Assessing.

[16] Yan Li, Minyi Su, Zhihai Liu, Jie Li, Jie Liu, Li Han, and Renxiao Wang. Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.*, 13(4):666, 2018.

[17] F C Bernstein, T F Koetzle, G J Williams, E F Meyer, M D Brice, J R Rodgers, O Kennard, T Shimanouchi, and M Tasumi. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.*, 80(2):319–24, nov 1977. ISSN 0014-2956. URL http://www.ncbi.nlm.nih.gov/pubmed/923582.

[18] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

[19] Yan Li, Zhihai Liu, Jie Li, Li Han, Jie Liu, Zhixiong Zhao, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *Journal of chemical information and modeling*, 54(6):1700–1716, 2014.

[20] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*, volume 11. John Wiley & Sons, 2008.

[21] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.*, 57(4):942–957, apr 2017. ISSN 1549-9596. doi: 10.1021/acs.jcim.6b00740. URL http://pubs.acs.org/doi/10.1021/acs.jcim.6b00740.

[22] Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26 (9):1169–1175, 2012. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq112.A. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3524828/pdf/emss-50853.pdfhttps://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq112.

[23] 1.4. support vector machines¶. URL http://scikit-learn.org/stable/modules/svm.html#svm-regression.

[24] Manfred K Warmuth, Jun Liao, Gunnar Rätsch, Michael Mathieson, Santosh Putta, and Christian Lemmen. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43(2):667–673, 2003.

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

[26] Bengt Winblad, Philippe Amouyel, Sandrine Andrieu, Clive Ballard, Carol Brayne, Henry Brodaty, Angel Cedazo-Minguez, Bruno Dubois, David Edvardsson, Howard Feldman, et al. Defeating alzheimer's disease and other dementias: a priority for european science and society. *The Lancet Neurology*, 15 (5):455–532, 2016.

[27] Lon S Schneider, Karen S Dagerman, Julian PT Higgins, and Rupert McShane. Lack of evidence for the efficacy of memantine in mild alzheimer disease. *Archives of neurology*, 68(8):991–998, 2011.

[28] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[29] Jiansong Fang, Yongjie Li, Rui Liu, Xiaocong Pang, Chao Li, Ranyao Yang, Yangyang He, Wenwen Lian, Ai-Lin Liu, and Guan-Hua Du. Discovery of multitarget-directed ligands against alzheimer's disease through systematic prediction of chemical–protein interactions. *Journal of chemical information and modeling*, 55(1):149–164, 2015.

[30] Yao Chen, Zong-liang Liu, Ting-ming Fu, Wei Li, Xiao-li Xu, and Hao-peng Sun. Discovery of new acetylcholinesterase inhibitors with small core structures through shape-based virtual screening. *Bioorganic & medicinal chemistry letters*, 25(17):3442–3446, 2015.

[31] Akhil Kumar, Gaurava Srivastava, and Ashok Sharma. A physicochemical descriptor based method for effective and rapid screening of dual inhibitors against bace-1 and gsk-3β as targets for alzheimer's disease. *Computational biology and chemistry*, 71:1–9, 2017.

[32] Hongbo Xie, Haixia Wen, Denan Zhang, Lei Liu, Bo Liu, Qiuqi Liu, Qing Jin, Kehui Ke, Ming Hu, and Xiujie Chen. Designing of dual inhibitors for gsk-3β and cdk5: virtual screening and in vitro biological activities study. *Oncotarget*, 8(11):18118, 2017.

[33] Xiao Hua Ma, Zhe Shi, Chunyan Tan, Yuyang Jiang, Mei Lin Go, Boon Chuan Low, and Yu Zong Chen. In-silico approaches to multi-target drug discovery. *Pharmaceutical research*, 27(5):739–749, 2010.