Building a chemical library

Generating new molecules

To generate new molecules using genetic algorithms, we have used a package called Data Warrior. Data Warrior was used because it is an open source easy-to-use platform for generating new molecules based on the scaffolds provided. In addition to the creation of compound libraries, Data Warrior can be used to calculate physiochemical properties, create graphs, and visualize data. We used the genetic algorithm functionality of Data Warrior to generate compounds that were similar in structure to the starting 16 scaffolds.

Briefly, the genetic algorithm of Data Warrior works in the following way: (a) Input a user provided structure of a molecule called scaffold; (b) Mutate it randomly by changing fragments on the molecule and select the structures that are most similar to the original structure; (c) Generate a pre-specified number of children for every scaffold; (d) Select the most structurally similar molecules from the population. These molecules become the starting structures for the next generation. The above steps are repeated till we have a desired number of generations. While Data Warrior is a powerful tool to rapidly generate a large number of molecules ($\sim10^5$), it is inflexible in the sense that it does not allow us to choose a different selection criterion than structural similarity. We used Data Warrior to generate 400 generations, 32 children selected per generation out of 4096 per generation. The algorithm selected the 32 molecules most similar to the parent generation. Since we wanted to generate compounds that were more druglike, we selected the compounds that had a high drug score and performed a second round of generation of new molecules. This gave us 12800 compounds per scaffold.

Drug Score is a numerical score given to a molecule to predict how good are its drug-like characteristics. Drug Score is based on the statistics of how frequently the fragments that comprise the molecule appear in known drugs. In addition, Drug Score considers desirable physical properties, such as solubility, molecular weight etc.

Post-Processing

Druglikeness. The Lipinski rules, from which the expression for drug score derived, states that druglike compounds are more likely to have molecular weight less than 500, lipophilicity less than 5, less than 5 hydrogen bond donors, less than 10 hydrogen bond acceptors, molar refractivity should be between 40-130). So, by selecting the molecules which have a high druglikeness (> 0.9) we generated more molecules. This resulted in a larger number of compounds with scores greater than 0.9. We also found that the higher the drug score, the more likely a compound would not contain problematic functional groups.

PAINS. After generating a population of molecules, we screened them for Pan Assay Interference Compounds (PAINS). PAINS are problematic functional groups that are false positives in bioassays in the sense that they appear to be strong binders to a protein targets, but really are not very selective in their binding. The PAINS screen used in this work is the NIH filter in RDKit. For a review of PAINS kindly see (Baell and Walters 2014). We found that 82 percent of marketed drugs did not contain PAINS as defined by the NIH filter. In addition, we also screened for the fraction of sp^3 hybridized carbons since a ratio of greater than .47 is associated with more selective binding (Baell and Walters, 2014). We found that 48 percent of FDA approved drugs have fraction of sp^3 hybridized carbons fsp3 > 0.47 and 65% have scores greater than 0.36. Compounds with a higher fSP3 ratio were more complex and did not contain as many double bonds (Baell and Walters, 2014). We found that by adding an additional carbon or two this would improve this ratio for those scaffolds which had double bonds located in their 5 membered ring. In total, we had 4 scaffolds with double bonds, and 4 without. Since according to Drugbank drugs like cysteine, methimazole, cysteamine, azathioprine, and mercaptopurine that have anti-inflammatory properties also contain a sulfhydryl group (-SH), we added an SH group to create 8 more parent scaffolds (Drug Bank). With increasing drug score, the frequency of PAINLESS compounds increased.

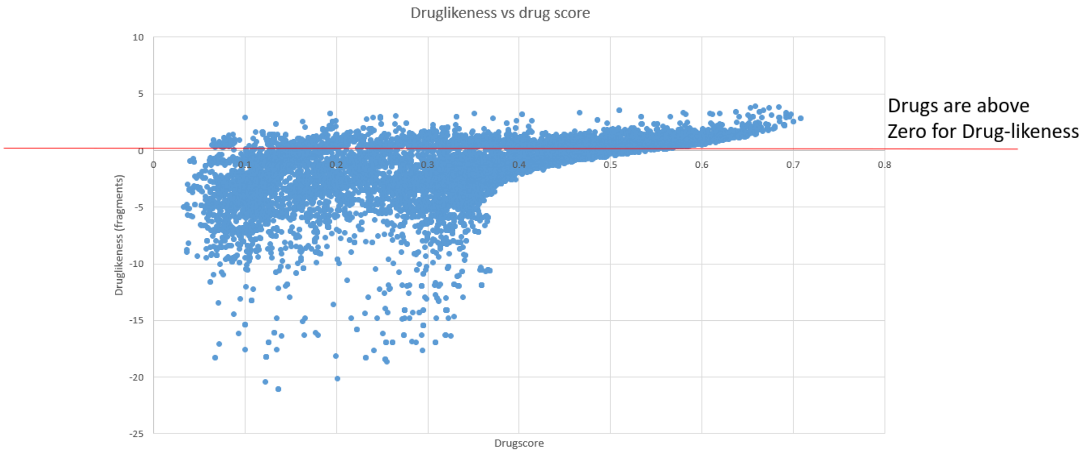
As drugscore increases, probability of a low drug-likeness decreases. Below is the formula used for calculating drug score. Drug score incorporates druglikeness.

Drugscore=(0.5+0.5/(1+exp(cLogP-5)))*(1-0.5/(1+exp(cLogS+5)))*(0.5+0.5/
(1+exp(0.012*Molweight-6)))*(1-0.5 (1+exp(Druglikeness)))*if(Mutagenic=="high",
0.6,if(Mutagenic=="low",0.8,1))*if(Tumorigenic=="high",0.6,if(Tumorigenic=="low",

0.8,1))*if(ReproductiveEffective=="high",0.6,if(ReproductiveEffective=="low",
0.8,1))*if(Irritant=="high",0.6,if(Irritant=="low",0.8,1))   (Sander, 2019)

In the Figure 3, drug score is plotted against druglikeness. The compounds above the red line are potential drug candidates as these compounds have the highest number of druglike fragments. Druglike fragments are derived from FDA approved drugs. The higher the frequency of druglike fragments the higher the drug score.

Figure 3 Drugscore vs. Druglikeness

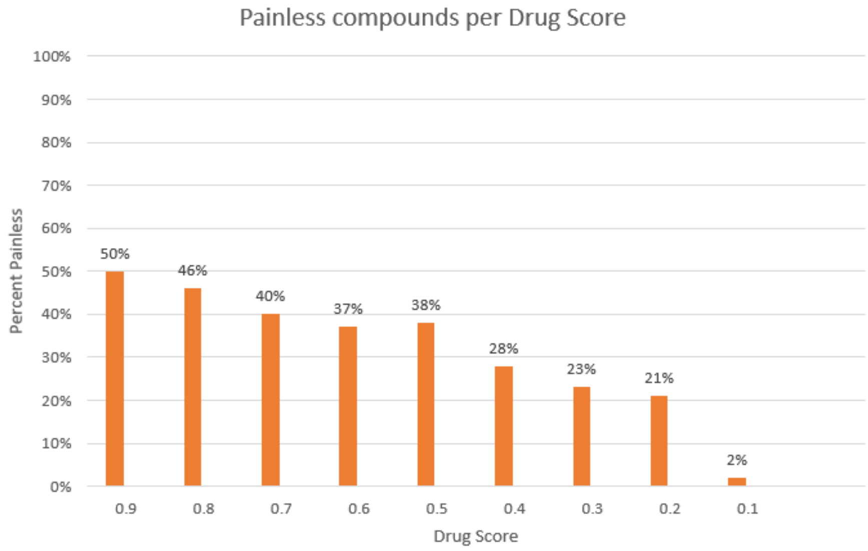


Drugs are above
Zero for Drug-likeness

Most of the compounds have negative drug-likeness (not very druglike fragments)
As drug score increases, probability of a low drug-likeness decreases

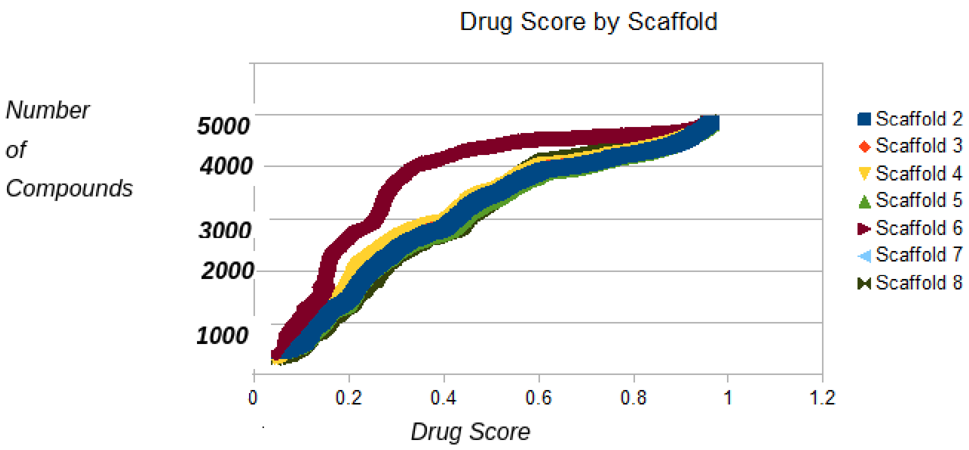
In addition, drug score correlates with a molecule being PAINless, where PAIN stands for Pan Assay Interference compounds. PAIN fragments are promiscuous fragments that may bind to many biological targets. Therefore, when a molecule has PAIN fragments, it may give a false positive result in high-throughput screens. Compounds with high drug scores do not contain functional groups that were PAINful [Figure 4]. This means that the higher the drug score, the more selective the binder a compound will be.

Figure 4. Drug score correlates with PAINLESS



In Figure 5, cumulative distribution of the drug score of molecules that were generated from each of the seven scaffolds. The drug score is on a scale from 0 to 1. There was an even distribution of compounds across the drug scores. It is observed that the scaffold 6 is has higher probability of generating molecules with good drug scores.

Figure 5. Drug Score by Scaffold



Protein ligand docking with Rosetta

After a population of molecules is generated using Data Warrior, one needs to identify the best binders to IRF3 protein. This is accomplished by docking these molecules to protein binding sites and calculating their binding energy. Docking programs predict the best pose for a ligand with its binding site on a protein. They can be used to screen out ligands that are not good binders. We used Rosetta because it is a widely used program for studying protein-ligand docking. The scores it returns are not actual binding energies but are correlated. The compounds that bind the target with the lowest energy are considered the best binders. Docking makes several assumptions. Proteins are treated as rigid or flexible only for residues close to the binding pocket (Varnek, 2017). Moreover, the binding predictions of ligands are limited by the number of conformations in 3D space that are sampled by the program (Varnek, 2017). Furthermore, effect of solvent in the binding process is ignored. In addition, the program does not simulate conformational changes in the protein that may result due to ligand binding. Other considerations, such as dissociation and protonation of protein residues due to environmental conditions may alter the binding capacities (Varnek, 2017).

Combinatorial Approach.

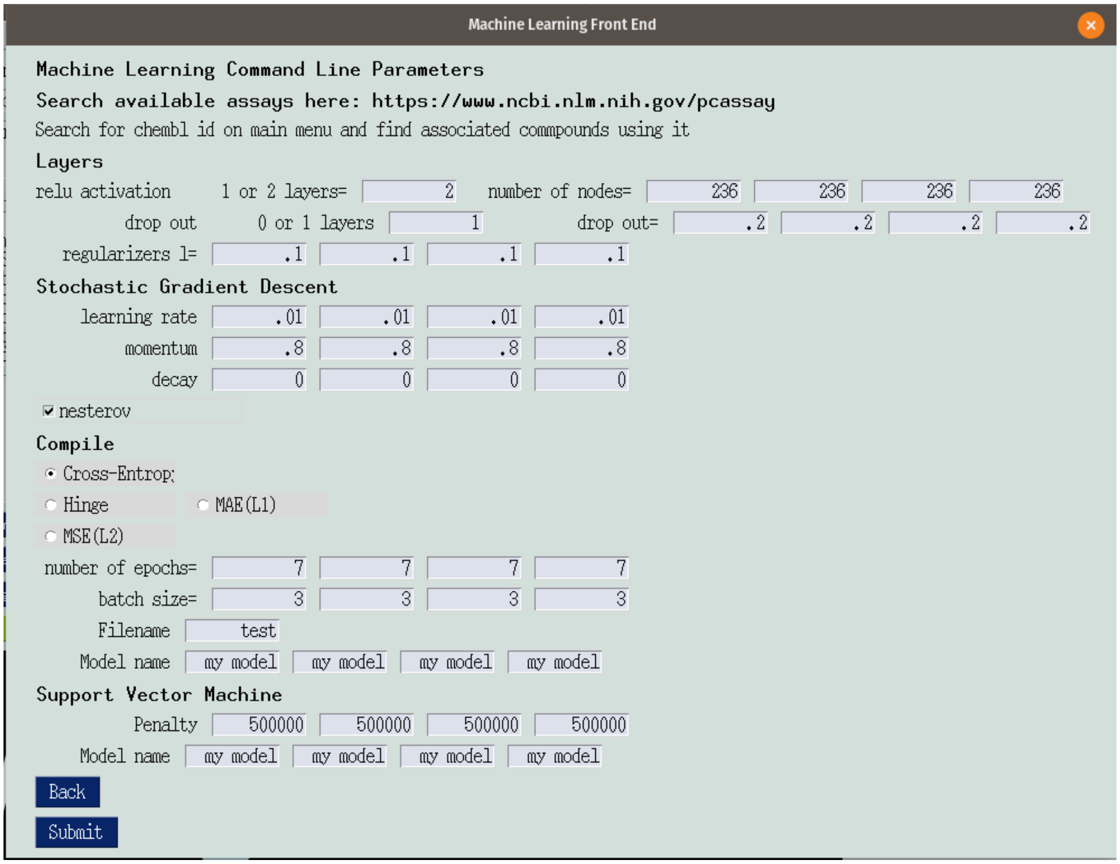
After finding the best binders for each scaffold using Rosetta, we took the top binders and fragmented them and then recombined them to form new molecules. The goal was to produce compounds with lower binding scores/better binders/lower energy required for binding than the original by recombining the best fragments. This produced some compounds with scores for IRF3 pocket 3 as low as -20 (our lowest score). It should be noted that C10 has a Rosetta score of -11.9, and that we almost doubled how efficiently our compounds bind to the pocket responsible for IRF3 dimerization. The fragmenting was performed in RDKit using synthesizable fragments called BRICS. Here were the best binding compounds.

QSAR Models

Docking methods are often combined with ligand-based methods like quantitative structure activity relationships (QSAR). It is useful to use large molecule databases of ligands to predict the targets of our compounds. The SVM and ANN models for 88 targets had sensitivity, specificity, and accuracy greater than 90%. Sensitivity is the ability to correctly identify positive test results. Specificity is the ability to correctly rule out negative test results. Accuracy is the ability to correctly identify both true positives and true negatives. These models are quick to create, over a million targets across many animal species that are available in databases, and no protein structure is needed. Only ligands are required to build a model. Ligands that bind to a target are considered positive and labeled a 1. Ligands that do not bind the target are labeled a 0. The models generalize well to 91 percent accuracy on a set of 243 marketed drugs. We tested this using the test h5 model function in the toolkit. Below is a screenshot of what the toolkit looks like. For the QSAR models, it will output a list of probabilities that a compound binds a target ranked from highest to lowest. We hope to make this software accessible to the non-technical user, so that it is very easy to perform assays with reasonable accuracy prior to performing actual experiments.

Below is a screenshot of how we make our QSAR models using the toolkit.

Figure 8. SAR Model Builder



Machine Learning Command Line Parameters

Search available assays here: https://www.ncbi.nlm.nih.gov/pcassay

Search for chembl id on main menu and find associated compounds using it

Layers

relu activation 1 or 2 layers= 2 number of nodes= 236 236 236 236

drop out 0 or 1 layers 1 drop out= .2 .2 .2 .2

regularizers l= .1 .1 .1 .1

Stochastic Gradient Descent

learning rate .01 .01 .01 .01

momentum .8 .8 .8 .8

decay 0 0 0 0

☑ nesterov

Compile

● Cross-Entrop;

○ Hinge ○ MAE(L1)

○ MSE(L2)

number of epochs= 7 7 7 7

batch size= 3 3 3 3

Filename test

Model name my model my model my model my model

Support Vector Machine

Penalty 500000 500000 500000 500000

Model name my model my model my model my model

The toolkit allows you to build 8 models at once. You can then test the pickles or H5 models it produces by providing smiles.

Generating an SVM model to predict targets of IRF3 protein

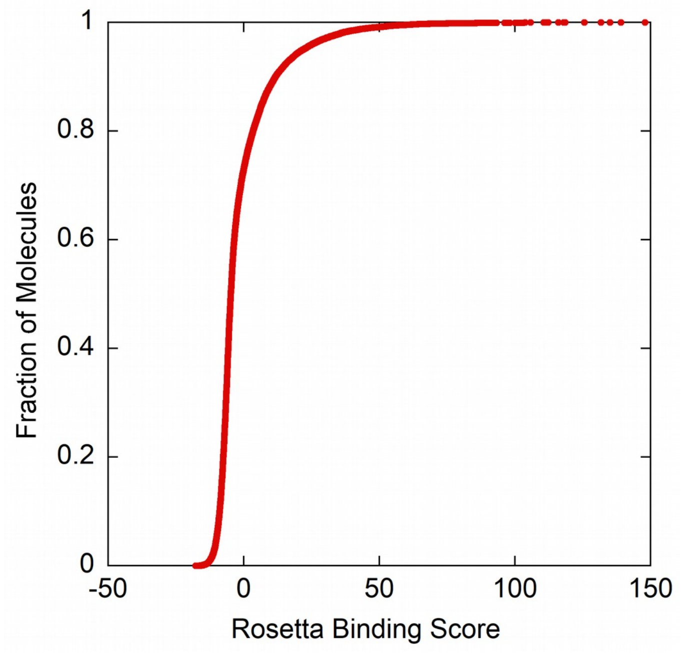
We used the automatic SAR builder function of the toolkit (see **Figure 8**) to build an IRF3 binding predictor. We took the compounds with the most negative binding energy from figure 6 above and used Data Warrior to generate 100 compounds that were PAINLESS, high drug score, and below -15 binding energy. These compounds were labeled with 1's. 100 random compounds were labeled 0. The model that worked the best, as tested by the test pickle menu was a SVM that had a penalty of 5,000,000. We then used a set of 350 compounds known to bind IRF3 and 350 compounds that did not as a test set. To test the model, you type in the name of the model into the first box and enter the smiles into the second box. There was 0 percent error on this independent test data set. The model was trained based on Rosetta binding energy and tested on known compounds that bind or do not bind to IRF3.

The Rosetta Algorithm has a low resolution and a high resolution phase. During low resolution docking, In low resolution docking, Rosetta picks a random starting position based on xyz coordinates. Then the ligand is translated, it moves the ligand up a specified distance in any direction from its starting point. Then it rotates the ligand randomly through all rotational degrees of freedom accepting only those that pass a Lennard Jones attractive/repulsive filter. Then during slide together, move the ligand 2 angstroms closer to the protein at a time until the protein and ligand collide. Following docking there is low resolution/centroid based scoring. Low resolution sampling involves replacement of the backbone with peptide fragments three and nine amino acids in length. It uses the following measures-- hydrophobicity term for each amino acid, steric repulsion between two residues, probability of two residues interacting, radius of gyration, solvation term based on a number of surrounding residues, secondary structure terms, 6-12 Lennard Jones potential, Eef1 solvation term, proline ring closure energy, omega backbone dihedral potential, updated disulfide geometry potential, potential of phi and psi angles for each amino acid, probability of an amino acid given a set of phi and psi angles, rotamer likelihood, combined covalent electrostatic hydrogen bond potentials, tyrosine hydroxyl out of plane penalty. During low resolution sampling, there is a 500 step Monte Carlo search with a 25% acceptance ratio. Once the centroid mode is complete, the lowest energy structure from the low resolution stage is selected for high resolution refinement. Using only the lowest energy ligand protein pair, all atoms are scored representing side chains in atomic detail. This comprises weighted individual terms that are summed to create a total energy for a protein. During high resolution docking, side chains are rotated around a bond one side chain at a time (cycling) and simultaneous sampling of multiple side chain rotations are combined with small movements of the ligands (repacking). Structures are minimized after each cycle using Monte Carlo sampling and Boltzmann probability to accept or reject a new structure. A final minimizer minimizes the structure of the protein ligand complex. During this stage, there are 50 Monte Carlo steps. If the change in score is less than +15, then minimize and if accepted output a decoy. Every 8 cycles do a full repack and Metropolis check. The score function uses van der Waals attractive and repulsive terms, solvation term, explicit hydrogen bonding term, statistical residue-residue pair wise interaction term, internal side chain conformational energy term and an electrostatic term. After scoring the complex the ligand is moved 1000 angstroms away from the protein and then scored again. Interface score equals complex energy minus separated energy.

Before we talk about the model more, let’s talk about how Rosetta binding works. The Rosetta algorithm has a low resolution and a high resolution phase. During low resolution docking, In low resolution docking, Rosetta picks a random starting position based on xyz coordinates. Then the ligand is translated, it moves the ligand up a specified distance in any direction from its starting point. Then it rotates the ligand randomly through all rotational degrees of freedom accepting only those that pass a Lennard Jones attractive/repulsive filter. Then during slide together, move the ligand 2 angstroms closer to the protein at a time until the protein and ligand collide. Following docking there is low resolution/centroid based scoring. Low resolution resolution sampling involves replacement of the backbone with peptide fragments three and nine amino acids in length. It uses the following measures-- hydrophobicity term for each amino acid, steric repulsion between two residues, probability of two residues interating, radius of gyration, solvation term based on a number of surrounding residues, secondary structure terms, 6-12 lennard jones potential, Eef1 solvation term, proline ring closure energy, omega backbone dihedral potential, updated disulfide geometry potential, potential of phi and psi angles for each amino acid, probability of an amino acid given a set of phi and psi angles, rotamer likelihood, combined covalent electrostatic hydrogen bond potentials, tyrosine hydroxyl out of plane penalty. During low resolution sampling, there is a 500 step monte carlo search with a 25% acceptance ratio. Once the centroid mode is complete, the lowest energy structure from the low resolution stage is selected for high resolution refinement. Using only the lowest energy ligand protein pair, all atoms are scored representing side chains in atomic detail. This comprises weighted individual terms that are summed to create a total energy for a protein. During high resolution docking, side chains are rotated around a bond one side chain at a time (cycling) and simultaneous sampling of multiple side chain rotations are combined with small movements of the ligands (repacking). Structures are minimized after each cycle using monte carlo sampling and Boltzman probability to accept or reject a new structure. A final minimizer minimizes the structure of the protein ligand complex. During this stage, there are 50 monte carlo steps. If the change in score is less than +15, then minimize and if accepted output a decoy. Every 8 cycles do a full repack and Metropolis check. The score function uses van der waals attractive and repulsive terms, solvation term, explicit hydrogen bonding term, statistical residue-residue pair wise interaction term, internal side chain conformational energy term and an electrostatic term. After scoring the complex the ligand is moved 1000 angstroms away from the protein and then scored again. Interface score equals complex energy minus separated energy.

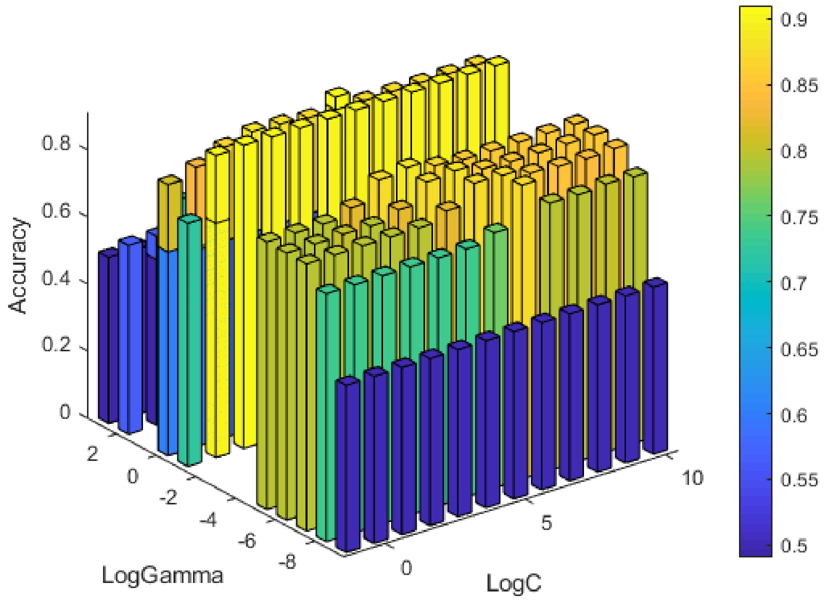
After docking the compounds, we graphed the fraction of compounds vs. Rosetta binding score.

Figure 9. Rosetta Energy Normalized



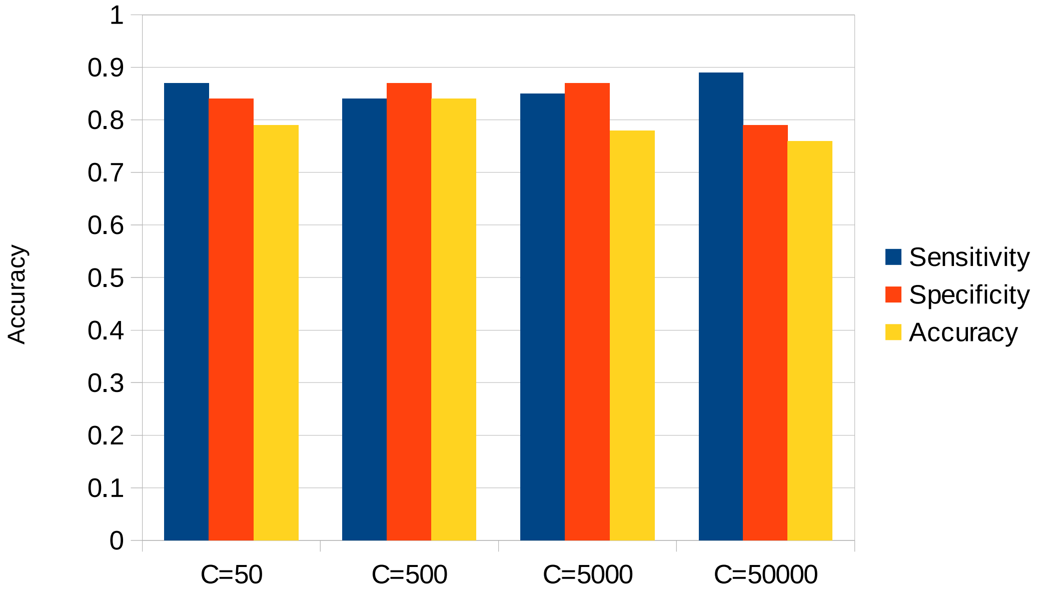
We created a gridsearch program that loops through the various possible parameters for SVM models. The grid search varies the gamma and C. Then we plotted accuracy. The figure below shows how test set results varied with parameters.

Figure 10. Gamma, vs. C, vs. Accuracy for IRF3



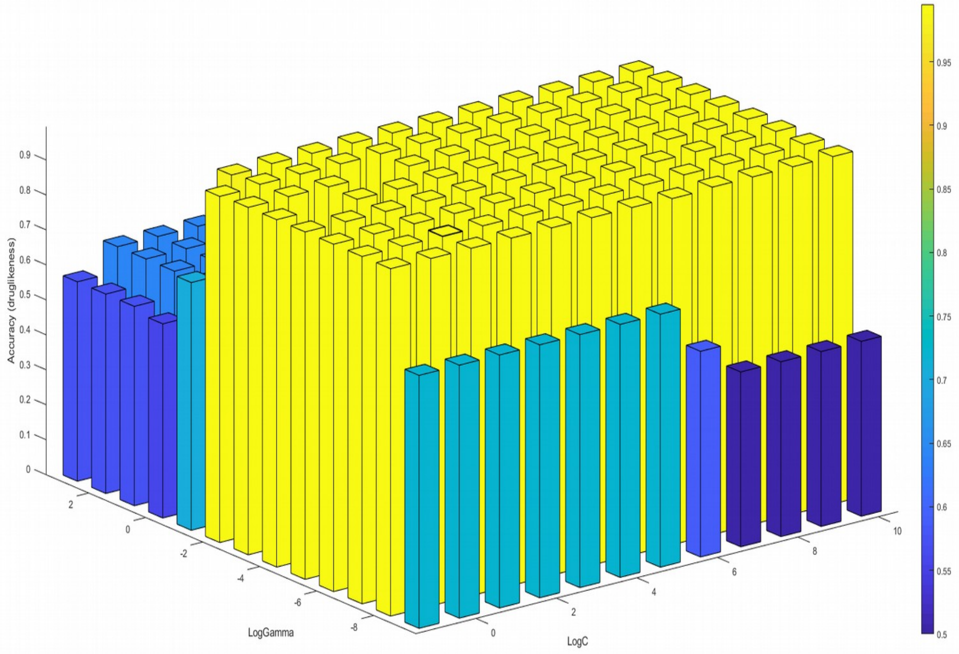
The best model had a loggamma of 1.8, a logC of .5 and an accuracy of 0.86. On an independent dataset we had the following accuracy, sensitivity, and specificity. We tested this on a new set of 500 IRF3 inhibitors that were not in the training set.

Figure 11. Accuracy, Sensitivity, and Specificity of IRF3 SVM Models on an Independent Validation Data Set



Next, I wanted to create a model that would predict whether or not a compound contains druglike compounds. I went to the enamine database and downloaded druglike fragments and a set of natural product fragments. We wanted to find the best model for druglikeness using a gridsearch.

Figure 12. Druglikess SVM Gridsearch



On an independent dataset-- 914 low drug score compounds marked 0 and 914 FDA approved compounds marked 1.

Figure 13. Druglikeness on Independent Dataset

We also created an ANN model for druglikeness.

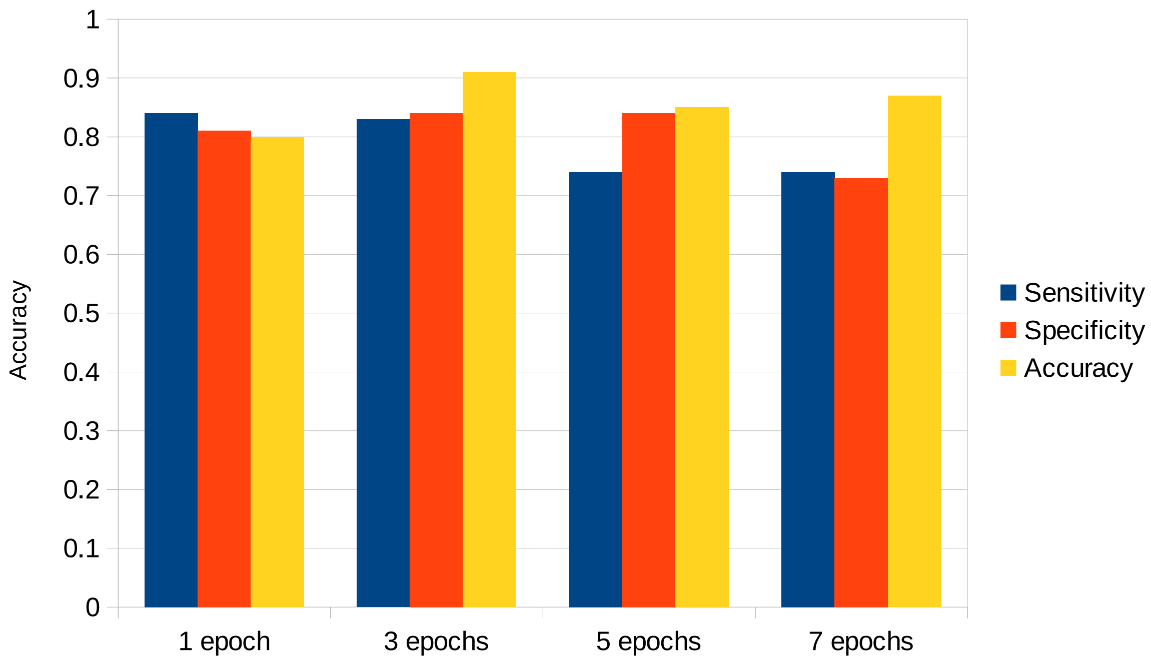
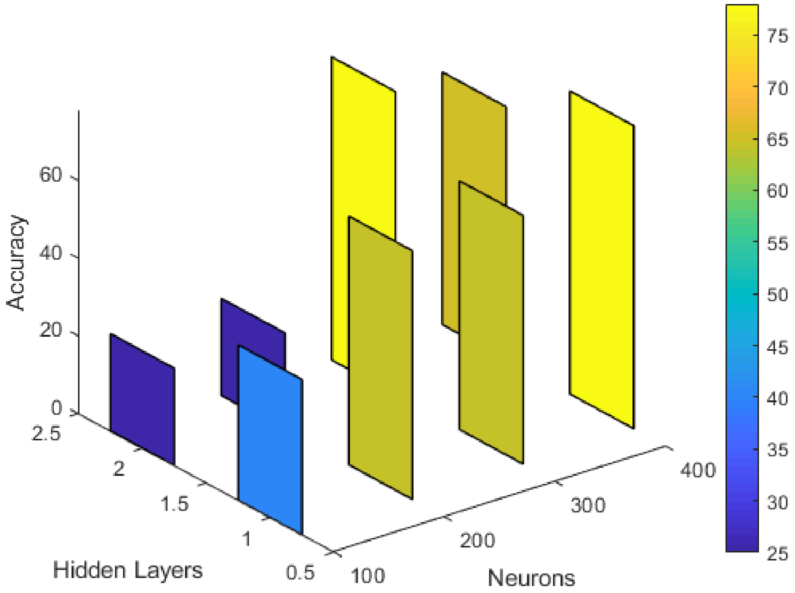
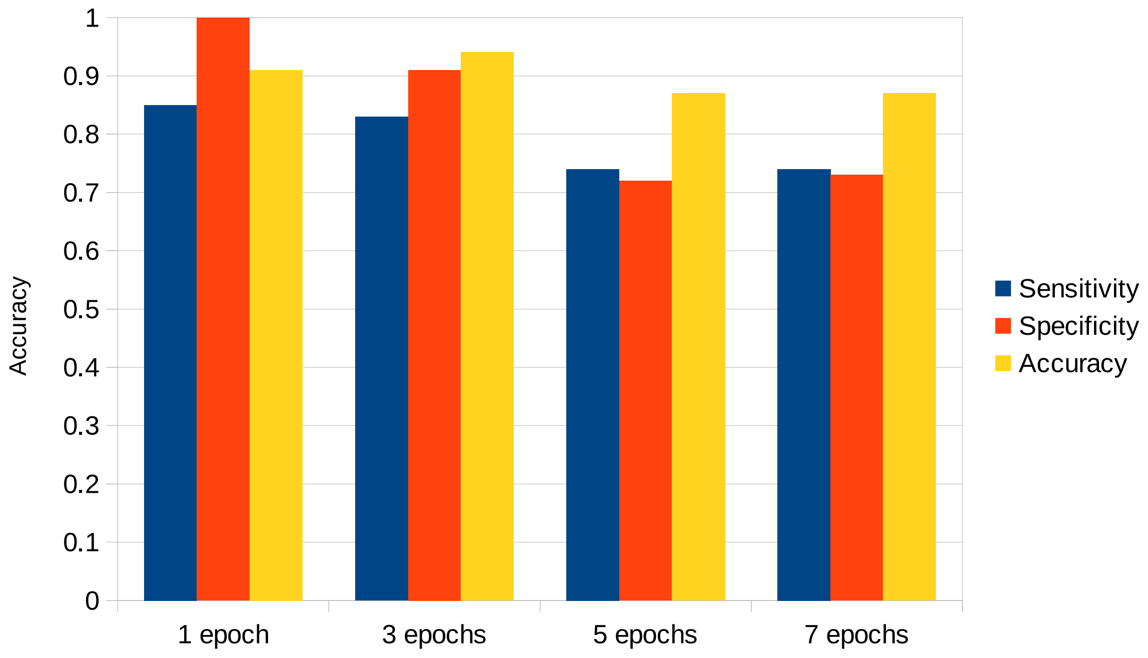


Figure 14. Artificial Neural Network with the number of neurons and layers varied



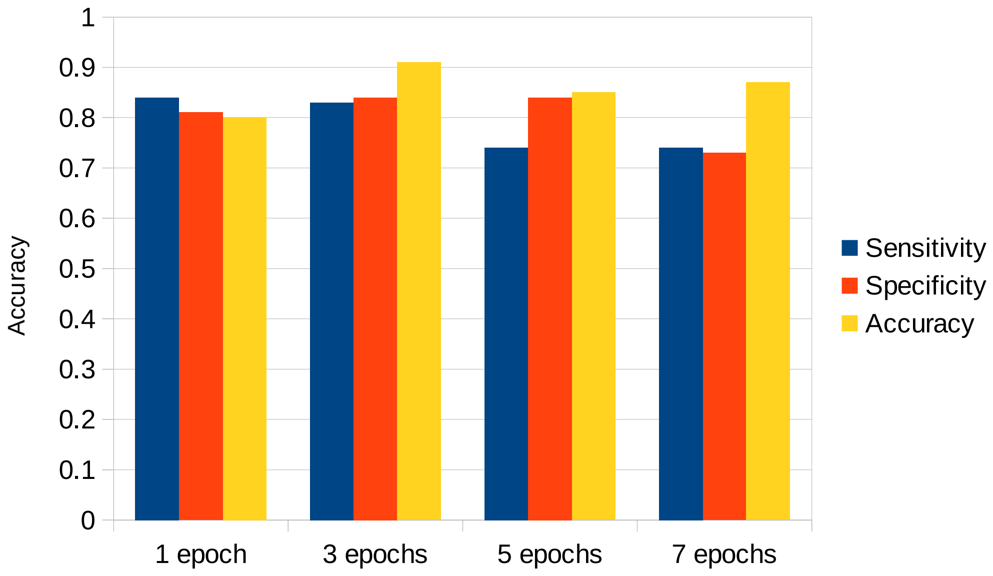
On an independent set with 2 layers and 300 neurons we had an accuracy of 74 percent. Next, we wanted to vary the number of epochs. 236 nodes. Regularizers=.2. learning rate=.01. Momentum=.08. Decay= 1. Nesterov. Batch size 3.

Figure 15. Druglikeness ANN Test Dataset



We tested the ANN on a new set of 500 new IRF3 inhibitors.

Figure 16. Druglikeness ANN Independent Dataset



To make the toolkit accessible to users without having them install a large number of associated software programs, the toolkit and Autodock were added to a Virtualbox. Virtualbox allows you to run lubuntu on your operating system in a portable application. It allows you to save your data in a workspace that can be shared with others. It also allows users to share their desktop virtually using the Chrome Web browser with Chrome Remote Desktop. Additionally, users can share code using google colab. Additional packages that have been installed include Java and Spark.   We will be publishing a journal paper illustrating the functionalities of the toolkit. After graduating, I plan to write a research proposal for a grant to hire students and programmers to keep improving our toolkit and commercialize it.