5-2019

HIV Resistance Prediction using Feed Forward Neural Networks and Sequence Expansion Methodologies

Christopher S. Luikart
Trinity University, cluikart@trinity.edu

HIV Resistance Prediction using Feed Forward Neural Networks and Sequence Expansion Methodologies

by
Christopher Shea Luikart

A departmental senior thesis submitted to the Department of Computer Science at Trinity University in partial fulfillment of the requirements for graduation with departmental honors.

4/22/19

Matthew A. Hibbs	Yu Zhang
Thesis Advisor	Department Chair

Michael Soto, AVPAA

Student Agreement

I grant Trinity University (“Institution”), my academic department (“Department”), and the Texas Digital Library ("TDL") the non-exclusive rights to copy, display, perform, distribute and publish the content I submit to this repository (hereafter called "Work") and to make the Work available in any format in perpetuity as part of a TDL, Digital Preservation Network (“DPN”), Institution or Department repository communication or distribution effort.

I understand that once the Work is submitted, a bibliographic citation to the Work can remain visible in perpetuity, even if the Work is updated or removed.

I understand that the Work's copyright owner(s) will continue to own copyright outside these non-exclusive granted rights.

I warrant that:

1) I am the copyright owner of the Work, or
2) I am one of the copyright owners and have permission from the other owners to submit the Work, or
3) My Institution or Department is the copyright owner and I have permission to submit the Work, or
4) Another party is the copyright owner and I have permission to submit the Work.

Based on this, I further warrant to my knowledge:

1) The Work does not infringe any copyright, patent, or trade secrets of any third party,
2) The Work does not contain any libelous matter, nor invade the privacy of any person or third party, and
3) That no right in the Work has been sold, mortgaged, or otherwise disposed of, and is free from all claims.

I agree to hold TDL, DPN, Institution, Department, and their agents harmless for any liability arising from any breach of the above warranties or any claim of intellectual property infringement arising from the exercise of these non-exclusive granted rights.”

I choose the following option for sharing my thesis (required):

[X] Open Access (full-text discoverable via search engines)
[ ] Restricted to campus viewing only (allow access only on the Trinity University campus via digitalcommons.trinity.edu)

I choose to append the following Creative Commons license (optional):

HIV Resistance Prediction using Feed Forward Neural Networks and Sequence Expansion Methodologies

Abstract

HIV is a chronic and debilitating disease affecting the lives of millions of people globally. While therapies to treat HIV are available, drug resistance is a consistent problem. For this reason, an effective means of determining drug resistance for a given isolate is needed. In this experiment, we use a simple Artificial Neural Network (ANN) model trained on phenotypically labeled sequences from HIVdb for resistance classifications. We also observe an interesting data processing method, and determine train and test set division before such data processing is optimal for network performance.

Contents

Chapter 1: Introduction

HIV is a pandemic disease which affects the lives of approximately 36 million people worldwide, annually killing 940,000 as of 2017. About 1.8 million of these cases are new. The Human Immunodeficiency Virus (HIV) targets the host immune system, causing a decrease in defense against infections and some cancers. If allowed to progress, Acquired Immune Deficiency Syndrome (AIDS) can develop, with symptoms including severe infections like tuberculosis, weight loss, fever, and lymphomas. There is currently no known cure for HIV/AIDS, however a number of treatment options do exist (WHO 2018). A recent case though was reported cured through bone marrow transplants, and although the incident is not fully understood nor yet fully applicable to general HIV patients, it has sparked interest in new potentials for managing the disease.

1. The HIV Virion and a History of Antiretrovirals

The HIV virion itself is composed if 15 proteins categorized as either structural, enzymatic, gene regulatory, or accessory. They are: MA , CA , NC , p6, SU , and TM , PR , RT , IN, Tat, Rev, Nef, Vif, Vpr, and Vpu. All structural proteins include MA (matrix), CA (capsid), NC (nucleocapsid), p6, SU (surface), and TM (transmembrane). All enzymatic proteins include PR (protease), RT (reverse transcriptase), and IN (integrase), and are of greatest interest here, as they are crucial in the HIV life cycle, unique to the virus, and so are ideal for drug therapy. Most such treatments target specific stages in the HIV life cycle driven by these enzymatic proteins, namely reverse transcription, genome integration, and protein assembly (NIH 2018). The four main classes of drugs used in these treatments are nucleoside reverse transcriptase inhibitors (NRTI's), non-nucleoside reverse transcriptase inhibitors (NNRTI's), protease inhibitors (PI's), and integrase inhibitors. The first of such drugs developed was an NRTI called azidothymidine (AZT), discovered in 1964 by the National Cancer Institute and originally used in cancer therapy.

Since then, a multitude of antiretrovirals (ARV's) have been created, with more than 30 drugs currently available, making treatment of HIV more manageable and effective. However, HIV mutates quickly, allowing some people to become resistant within a matter of days. To counter drug resistance, a triple drug therapy called highly active antiretroviral therapy (HAART) was developed, which combines multiple drug classes in a single treatment, acting as a high genetic barrier to resistance (NIH 2019). Still, due to the lack of proofreading by reverse transcriptase during viral replication, high mutation rates can still generate resistant strains (Ji 1992). It is estimated that 8-20% of untreated carriers in North America are to some degree drug resistant (WHO 2018). And so an efficient means of resistance testing is needed for prescribing effective antiretrovirals, given that current testing is both expensive and time consuming.

2. Resistance Classifiers and Related Works

A solution to efficient resistance testing is a genotypic interpretation algorithm, such as Stanford's HIVdb Program. Most such algorithms though are rules based, using known mutation sites along the HIV genome to determine likelihood of resistance. For instance, a large class of mutations used in the HIVdb Program for resistance in RT are Thymidine Analog Mutations (TAM's). TAM's typically confer AZT and D4T resistance, and are classified into two main types. Type 1 TAM's include the mutations M41L, L210W, and T215Y, while type 2 TAM's include D67N, K70R, T215F, and K219Q/E, with the main difference between these types being that type 1 confers greater resistance to ABC, DDI, and TDF compared to type 2. Another class called multi-nucleoside mutations include the Q151M mutation which occurs in concert with a number of accessory mutations: A62V, V75I, F77L, and F116Y. By itself, this mutation can confer high level resistance to AZT, D4T, DDI and ABC, but with accessories can confer additional mid level resistance to 3TC, FTC and TDF (NRTI Resistance Notes).

Other algorithms use classification methods such as Artificial Neural Networks (ANN's), Sparse Dictionary Classification, Random Forest, and Support Vector Machines to make qualitative predictions (Khalid 2018; Shen 2016; Yu 2014; Yu 2013) . A number of these experiments have unique data processing methods, some of which incorporate protein structural data. For instance, Yu *et al.* 2014 reported using Delaunay triangulation to encode 3D residue position. Khalid *et al.* incorporated hydrophobicity data and protein secondary structure into support vector machines. In this experiment we apply a simple ANN model, while also implementing a data processing method, which uses sequence expansion to handle data ambiguity. This data processing method has been used in a number of studies with reportedly good results, and for this reason we wanted to explore certain characteristics (Amamudy 2018; Yu 2014; Yu 2013). In Yu 2014, initial sequences for HIV Reverse Transcriptase (RT), and HIV Protease (PR) were acquired from Stanford's HIVdb and then expanded using the above mentioned algorithm, producing approximately 60,000 total sequences. Each sequence had a set of drug resistant labels, which were encoded such that for a given PR sequence, a label value < 3.0

was denoted as 0, and a value > 3.0 was denoted as 1, while RT cut-off values varied according to the drug. Amino acids for a particular sequence were then encoded into an adjacency matrix using Delaunay triangulation, where each entry is the average distance between two types of amino acids based on the typical protein structure. While reports from this work indicate good performance, we note that the Delaunay triangulation may overly compress protein structure data and so format things in a way that is less biologically significant.

Amammudy also implements the expansion algorithm for RT and PR data, though with some slight modifications. Before expansion, they removed non B subtypes from the data set, and then performed the expansion with set cut-off values, meaning that sequences were limited in the number of expanded sequences they could create. They then encoded the amino acids with an integer value of 1 to 22. Neural networks with various architectures were made to train on different drug labels from the expanded data, and networks with the best performance were chosen. We note a number of potential issues here as well. Firstly, the amino acid encoding method used implies higher similarity between

certain residues when compared to others, which is not necessarily true. Secondly, choosing of arbitrary and variant architures for each drug label is not best practice, especially when the specific architectures chosen for each label are not indicated.

Due to these exceptional results, despite our skepticism about several algorithmic choices, we attempted to replicate the methods used in Amammudy and evaluate the impact of some of their analysis choices. Particularly, we tested the effects of expansion generally on network performances, and attempted to characterize the importance of dividing training and test sets either before or after applying the sequence expansion method.

Chapter 2: Materials and Methods

1. Sequence Acquisition and Processing

1916 PhenoSense RT sequences were acquired from Stanford's HIVdb. Each sequence entry contained the list of mutated amino acid positions for an isolate, as well as the degree of drug resistance for a set drug list. The degree of resistance for each drug was experimentally determined using Viralogic's Phenosense

assay, with a value greater than 1 indicating above standard resistance. The drug set was a combination of NRTI's and NNRTI's: 3TC, ABC, AZT, D4T, DDI, TDF, EFV, NVP, ETR, and RPV. In order to better understand the data, we performed an analysis of subtype distribution, and found that subtype B was most prevalent. Accordingly, we subtracted non subtype B entries from the data set to eliminate possible noise. The data was then parsed, with the sequence data placed inside of our input tensor, and the resistance data inside of our target. The data sets were then encoded for proper input into the network. For each entry in the sequence data, the amino acids were converted to a sequence of integers, where each amino acid is represented by a number from 1 to 22, and the total length of the sequence is set to the length of HIVdb's RT consensus sequence. The target is converted to a sequence of 0's and 1's, where 1 indicates the degree of resistivity reported for a drug is greater than 1, and 0 indicates a resistivity <= 1.

2. Sequence Expansion Algorithm

Next, we implemented the sequence expansion algorithm. Due to the method of sequence recording and natural diversity of HIV in samples, there are ambiguous sites along a sequence, usually represented as either an X or a combination of amino acids (AA's) separated by a comma. An 'X' here means that the site was not recorded, and multiple AA's mean that several residues were recorded. For instance, if we consider 'X' to be any amino acid, a single ambiguous sequence with an X can represent 22 possible real sequences. For each sequence then, we calculated the number of possible sequences which could be derived from ambiguous sites. Then, if the number of possible sequences for an expansion exceeded a cut-off value, we removed that sequence from the data. To understand the effects of the cut-off on model performance, we attempted three different cut-off values of 1000, 300, and 50, generating three unique data sets. After expansion, we recorded 64,000, 37,000 , and 11,000 sequences total in each set, and also counted negative to positive labels for each drug in each set to monitor effects of expansion on data balance.

3. Network Structure, Training, and Testing

Next, we constructed 11 simple feed forward neural networks, each with 1 hidden layer of 20 nodes using Pytorch. We chose the built in Multi Label Margin Loss as our loss function, and set the learning rate at 1e-4. We then trained each of the 11 networks differently. One of the networks was trained using the entire input and target, which we termed the MultiDrug network. The other 10 networks were trained using only a subset of the target, so that the target was a single value representing resistivity for a single drug. Each network then specialized in the prediction of a single drug, which we labeled Single Drug networks. Before training, we divided the input data into train and test sets, with the division being 70% to 30%. We then trained and tested all networks for 10 epochs, recording accuracy, precision, fallout, and recall, with a cutoff for true predictions at 0.7. After training and testing, AUC (Area Under the ROC Curve) curves were generated for each network.

Training and testing sets were then generated using two different methods, specifically pre-expansion and post expansion.

Out of curiosity in exploring the methods used in Amamudy 2018, we wished to observe the effects of dividing our data set at these distinct points. 70% to 30% train and test sets were then generated before expansions, and then expanded independently, while a separate train and test set were derived from the initial data after expansion. We then encoded, and passed these sequences into 11 networks for training and testing in the same way as described previously for our post expansion set, and then again for the pre-expansion set for each of the three expanded sets.

We also trained and tested in the absence of expansion using “one-hot” encoding. We did this in an attempt to determine the potential effects of this encoding technique on network performance.

Chapter 3: Results

We created AUC curves for each drug, or rather for each Single Drug network at three distinct values for expansion cut off and at two points before and after we expanded. We did this to measure network performance in conjunction with the expansion, and to provide insight into when expansion should be occurring.

1. 3TC

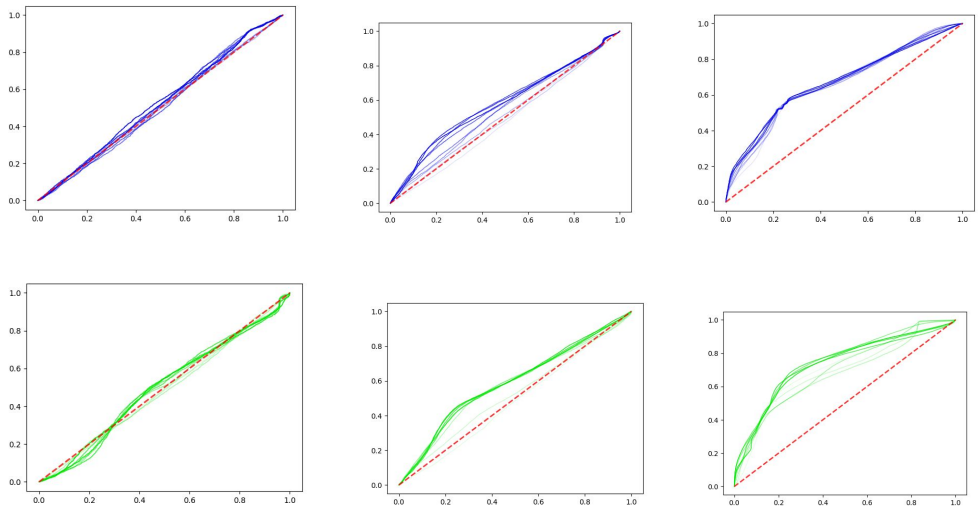


Figure 1: 3TC Single Drug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

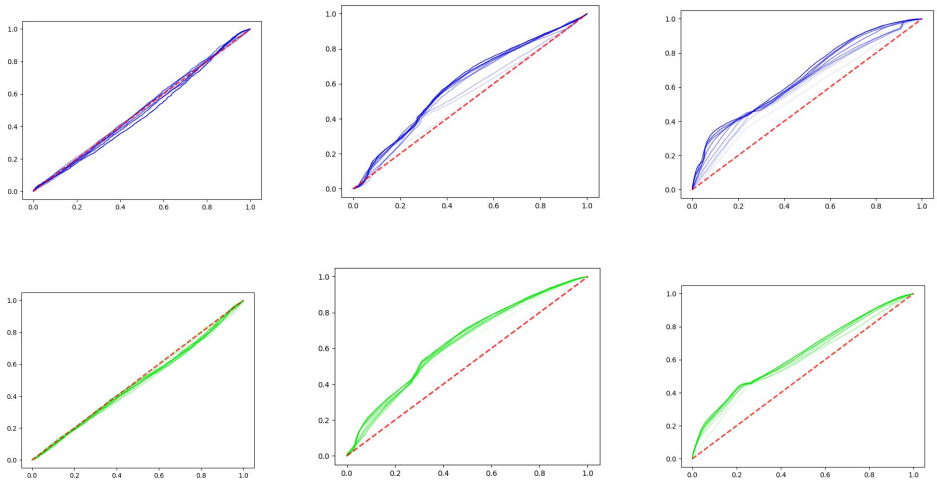


Figure 2: 3TC Single Drug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

2. ABC

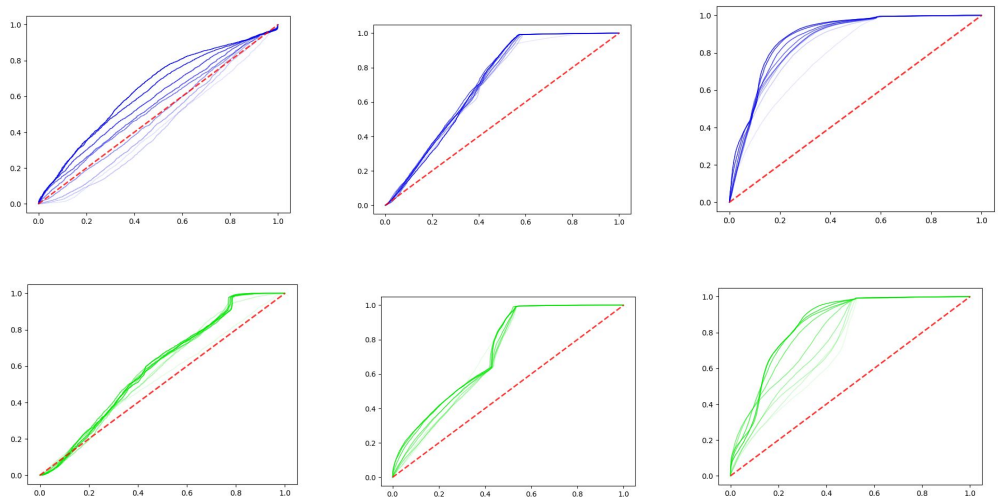


Figure 3: ABC SingleDrug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

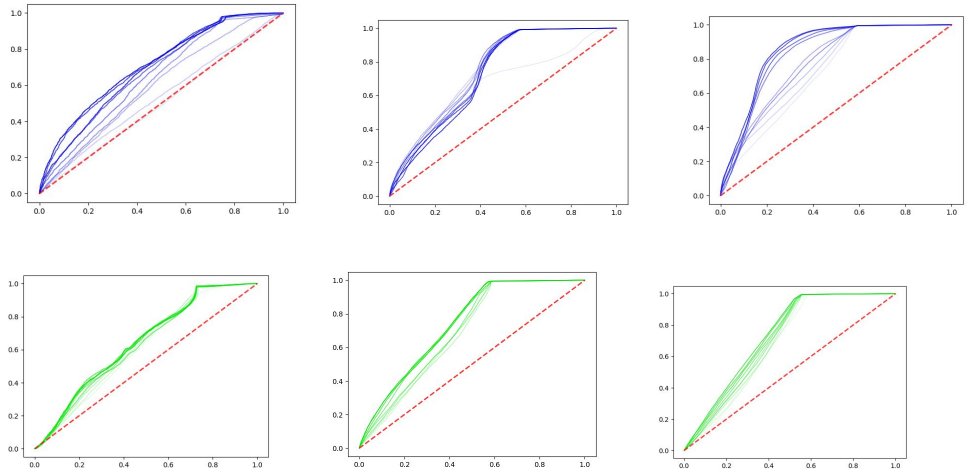


Figure 4: ABC SingleDrug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

3. AZT

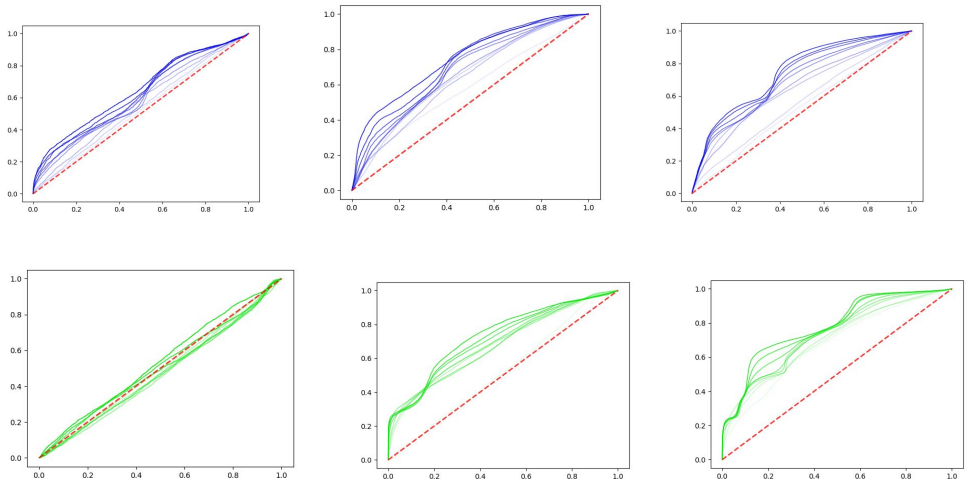


Figure 5: AZT SingleDrug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

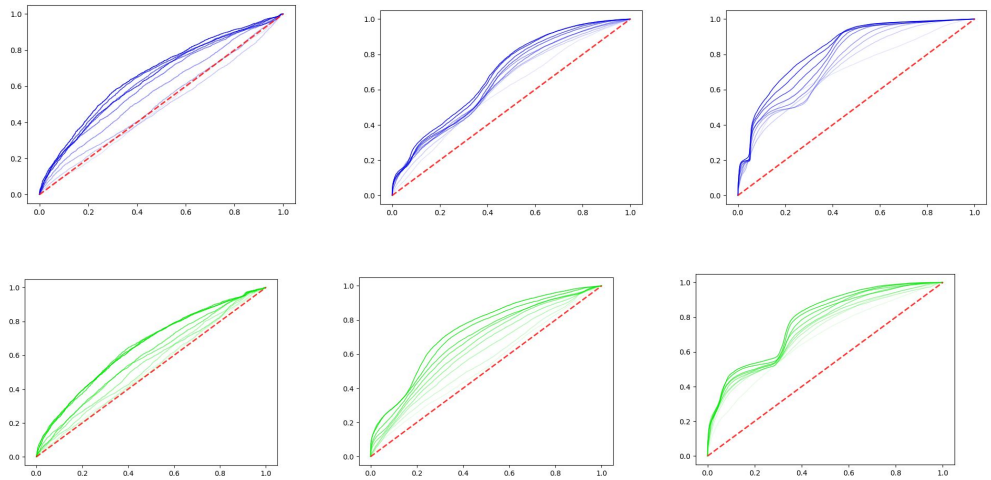


Figure 6: AZT SingleDrug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

4. D4T

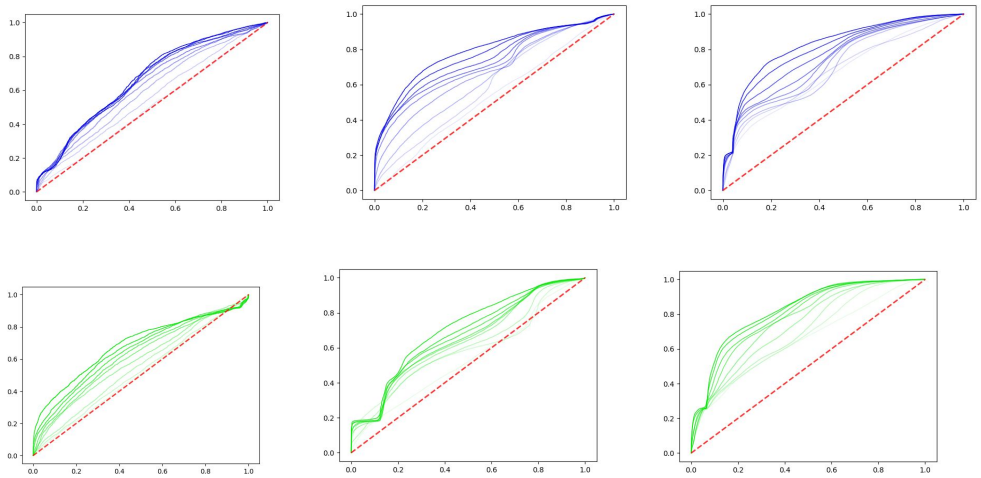


Figure 7: D4T SingleDrug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

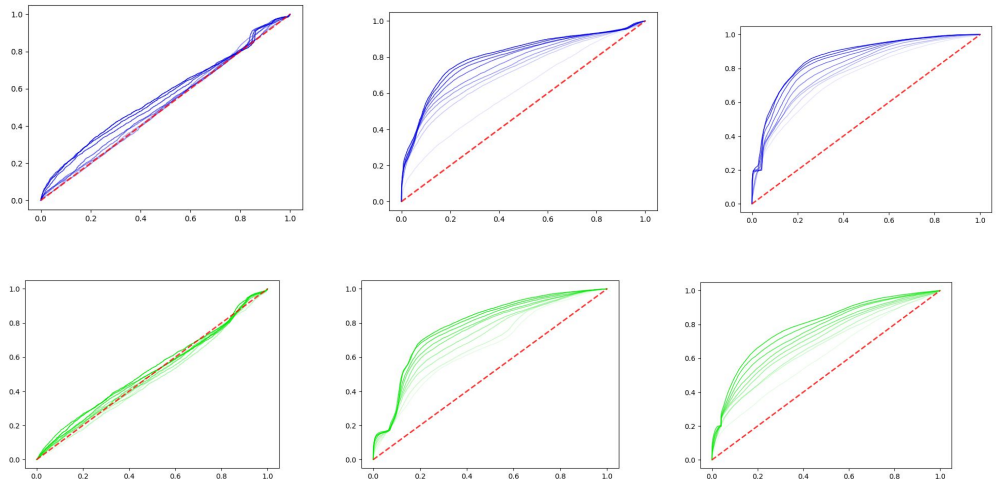


Figure 8: D4T SingleDrug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

5. DDI

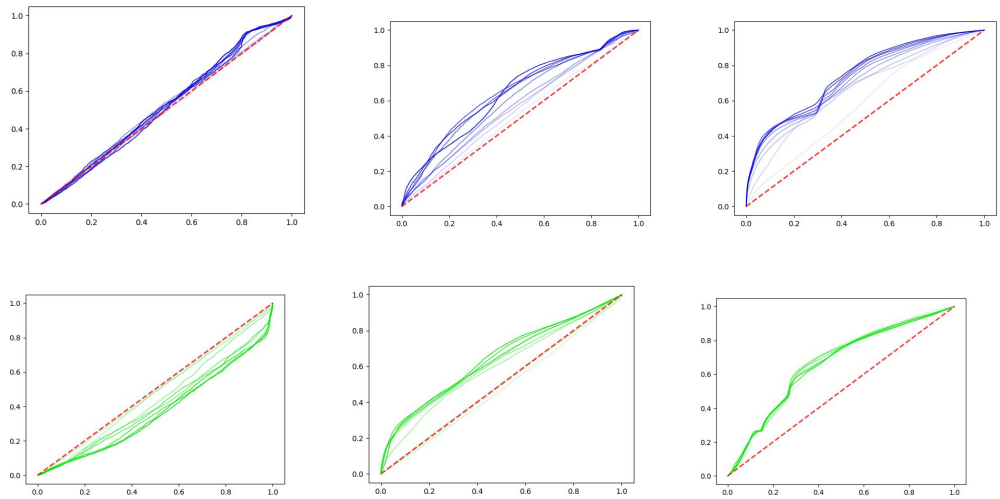


Figure 9: DDI SingleDrug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

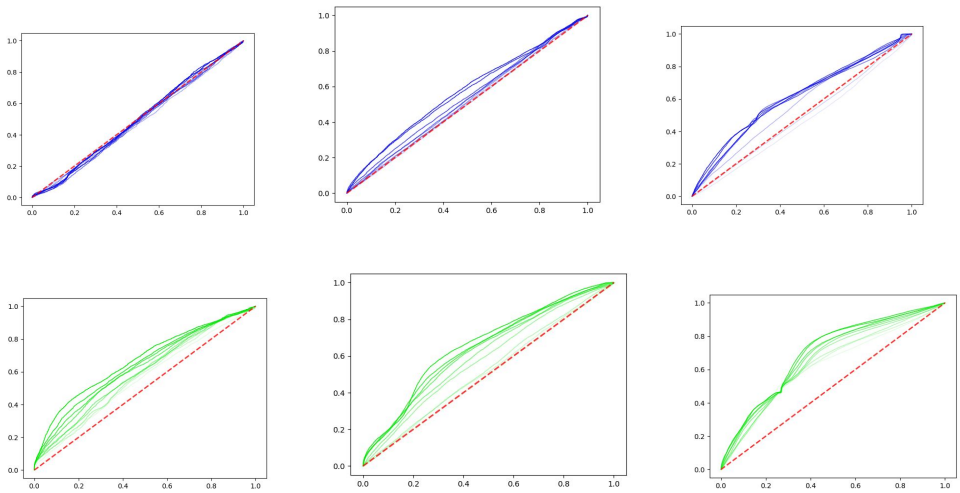


Figure 10: DDI SingleDrug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

6. EFV

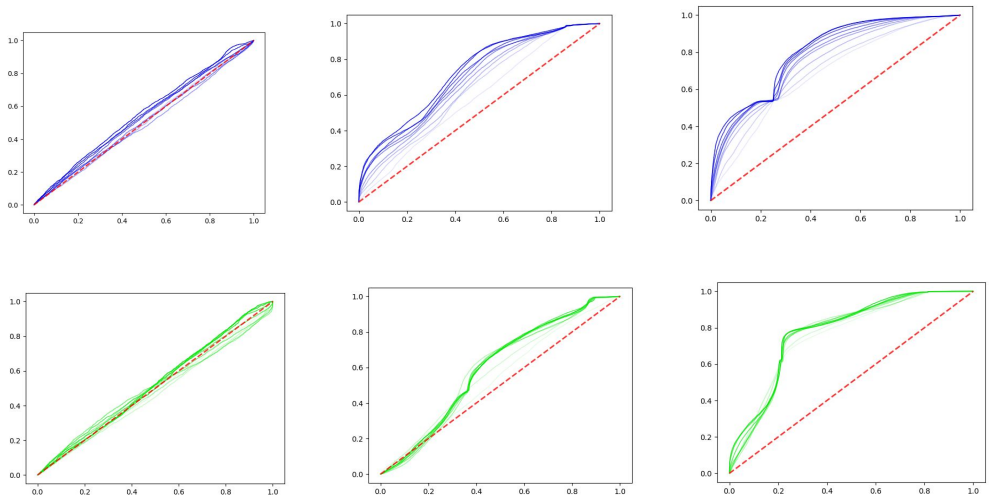


Figure 11: EFV SingleDrug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

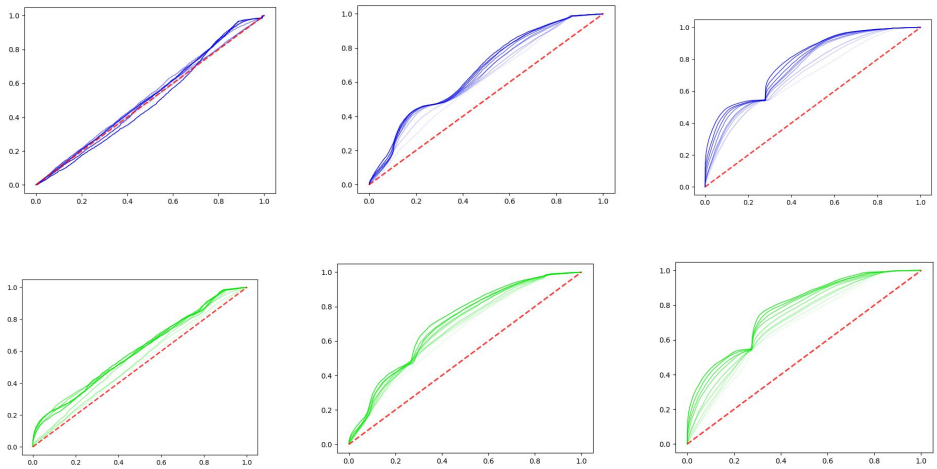


Figure 12: EFV SingleDrug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

7. ETR

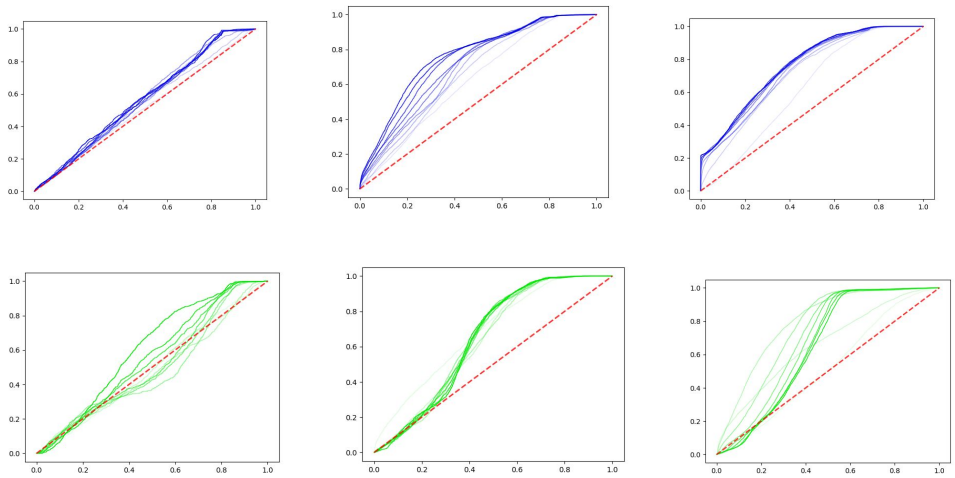


Figure 13: ETR Single Drug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

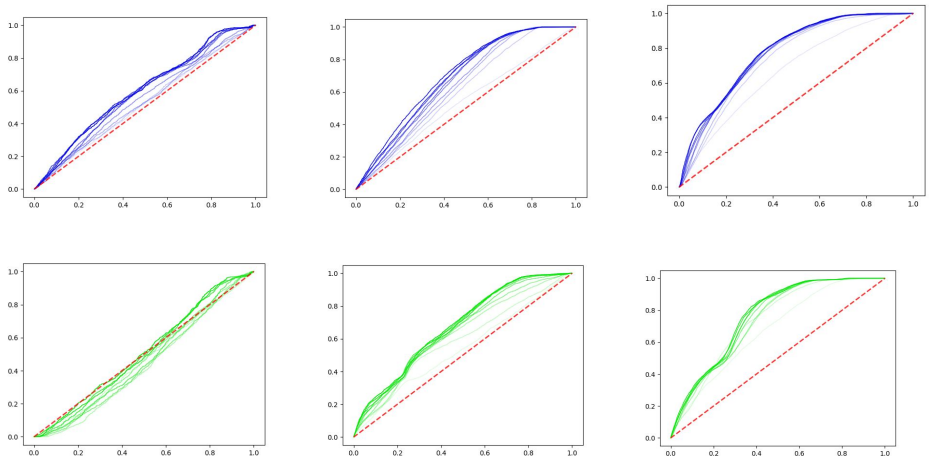


Figure 14: ETR Single Drug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

8. NVP

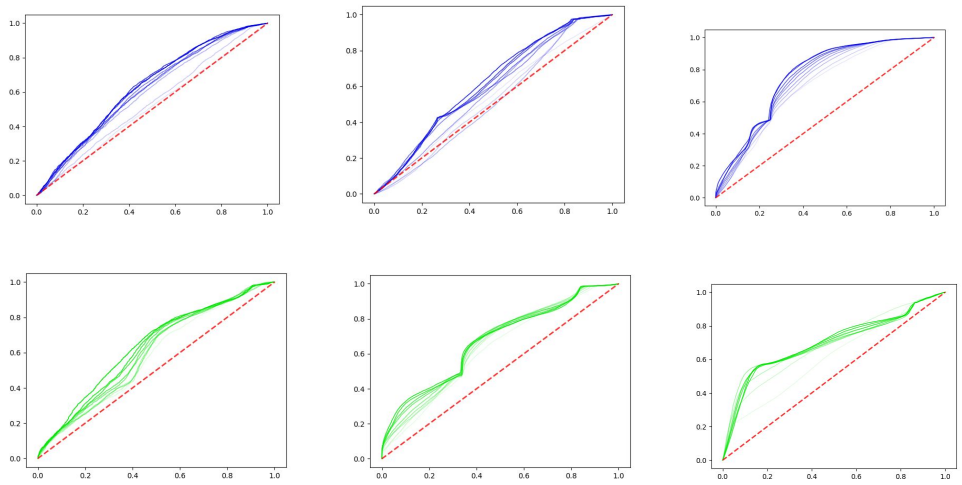


Figure 15: NVP Single Drug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

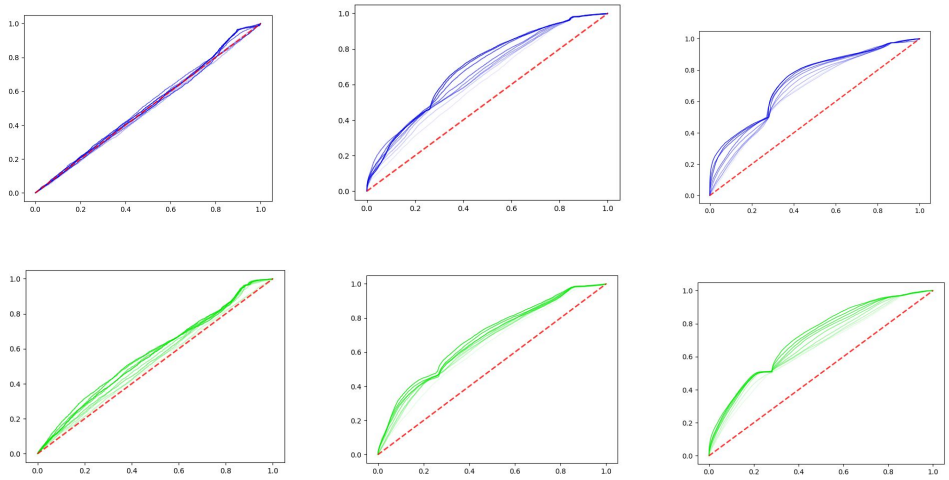


Figure 16: NVP Single Drug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

9. RPV

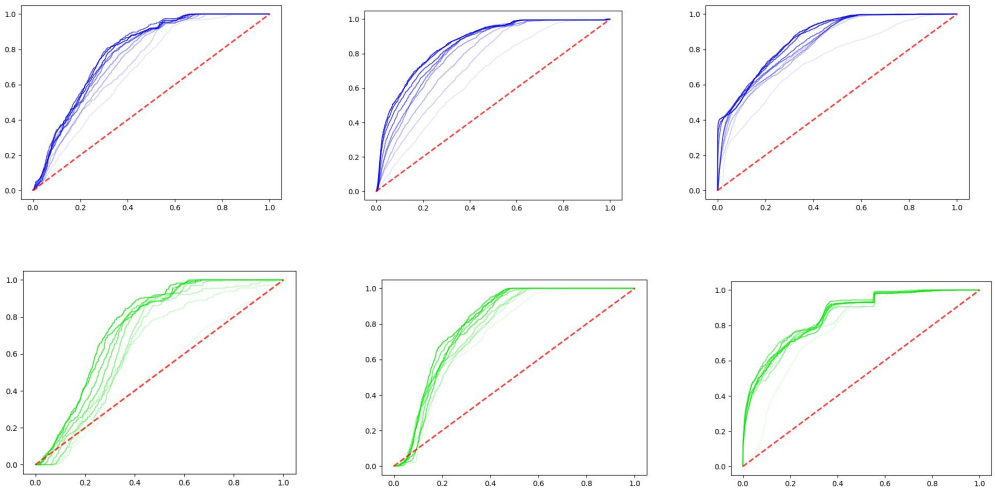


Figure 17: RPV Single Drug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

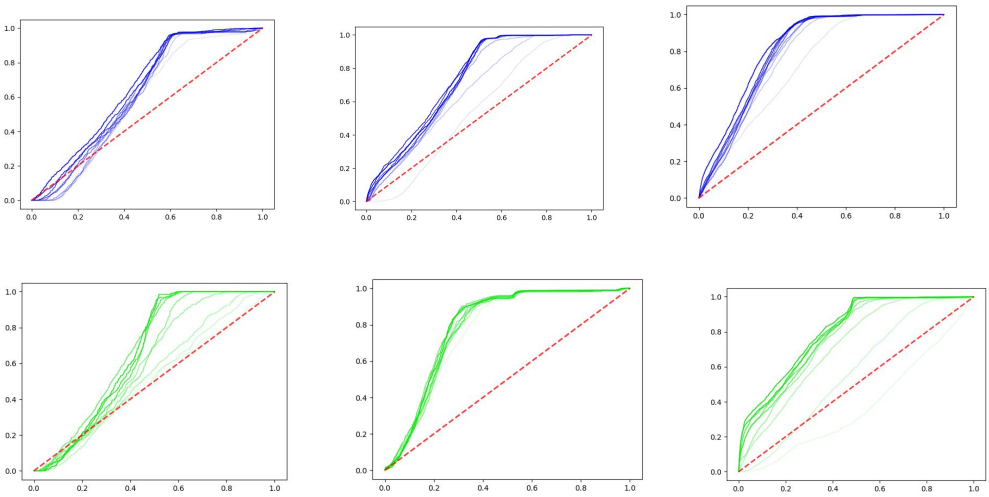


Figure 18: RPV Single Drug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

10. TDF

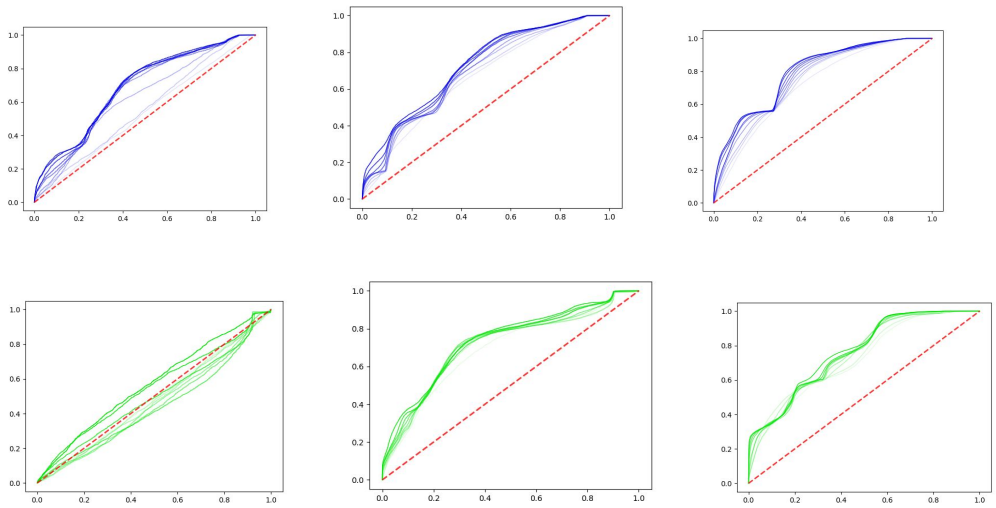


Figure 19: TDF Single Drug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

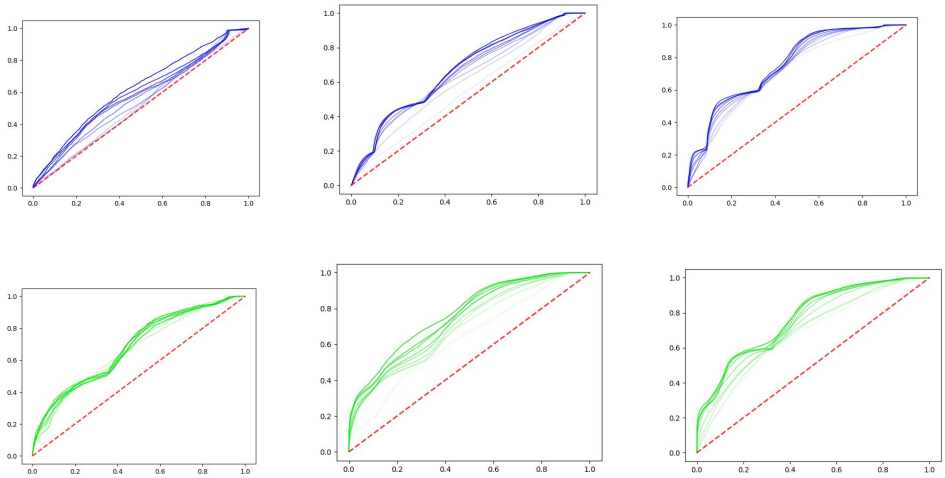


Figure 20: TDF Single Drug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

11. MultiDrug

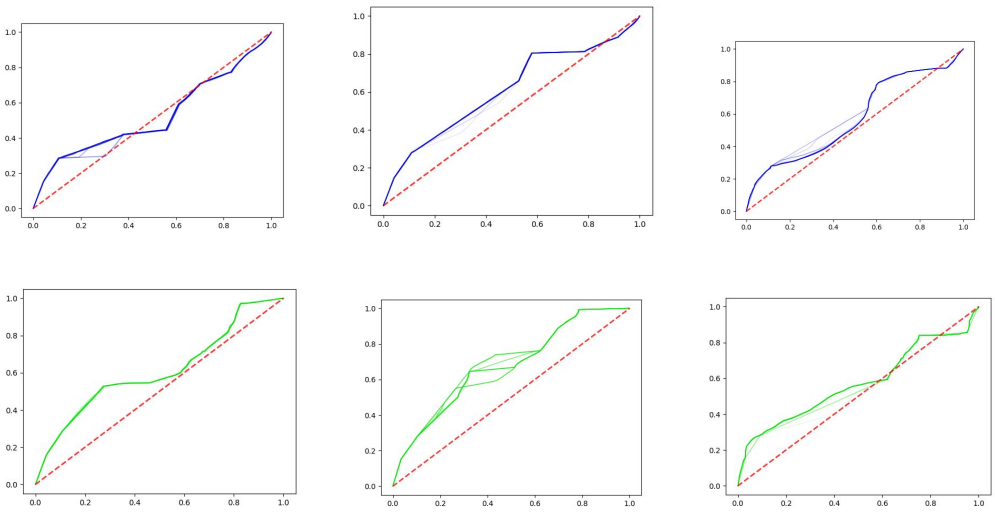


Figure 21: MultiDrug Networks Pre Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

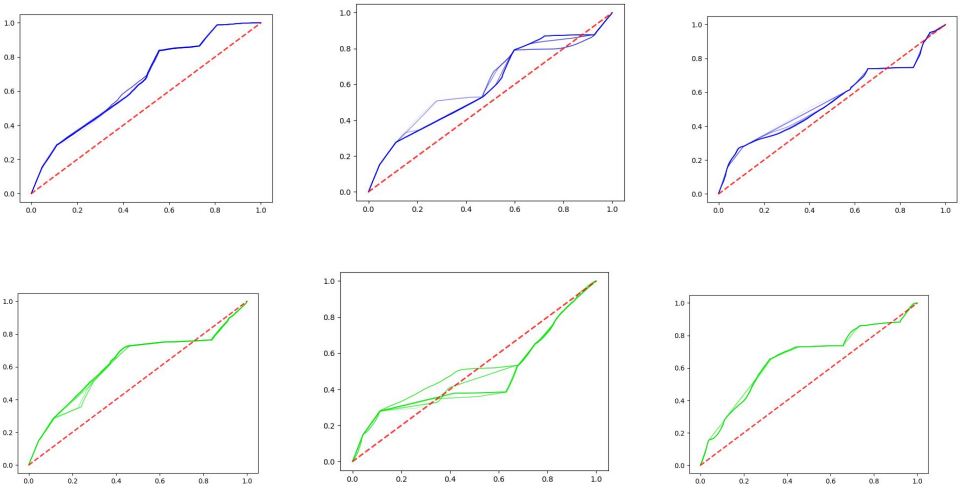


Figure 22: MultiDrug Networks Post Expansion for Train (Blue), and Test (Green) at cut-off values of 50, 300, and 1000 going from left to right.

12. One-Hot

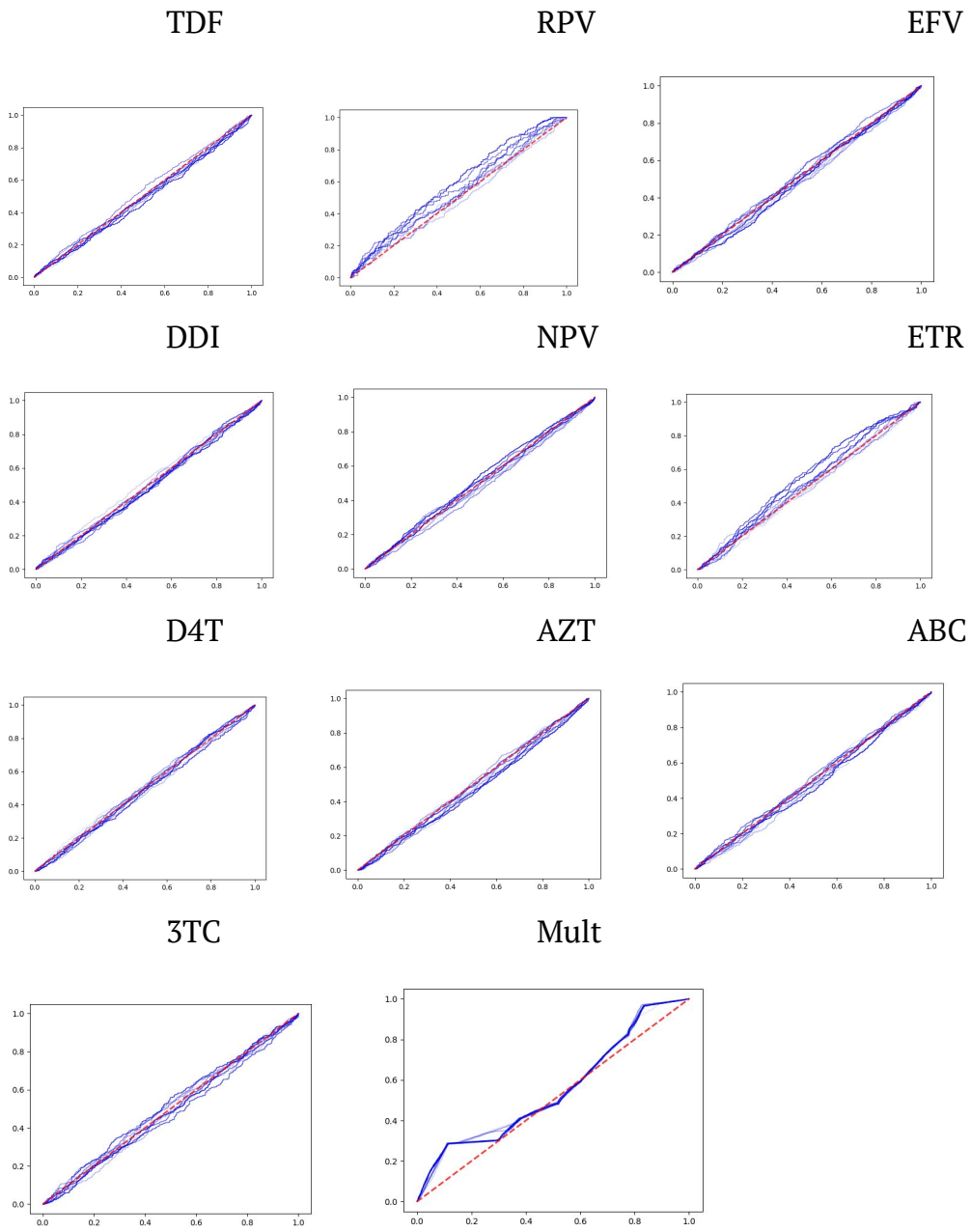
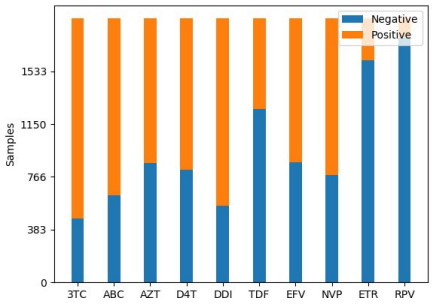


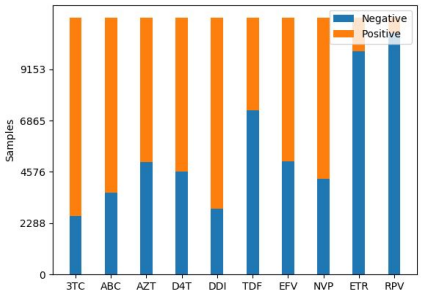
Figure 23: AUC's for Single Drug and Multi Drug networks using one-hot encoding without expansion during training.

13. Label Distribution

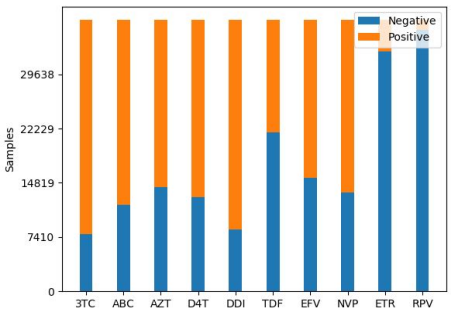
A.



B.



C.



D.

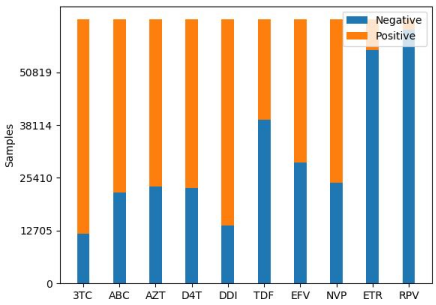


Figure 24: The number of negative and positive labeled samples for each drug at (A). Expansion cut-off=0, (B)Expansion cut-off=50, (C) Expansion cut-off=300, (D) Expansion cut-off=1000.

Chapter 4: Discussion

We wished to elucidate qualities of this expansion technique as it relates to classification algorithms for antiretroviral resistance generally. Given that the technique has become more widely used, it is import to further explore characteristics peculiar to the method.

1. Effects of Expansion and Comparing Pre to Post Methods

Firstly, we note that across all Single Drug networks, an increase in expansion cut-off value seems to optimize network performance. This holds true for training, test, post expansion, and pre expansion sets, indicating that the technique successfully improves unsupervised learning. Importantly, between most train and test AUC's, there was minimal change in curve structure, indicating that overfitting is minimized. However, in Single Drug groups DDI, RPV, TDF, NPV, ETR, and EFV we observe a noticeable shift and

occasional flattening of AUC's in pre expansion test sets when compared to training set counterparts (Figures 9, 11, 13, 15, 17, 19). The corresponding AUC's in post expansion sets though maintained their structure between train and tests sets (Figures 10, 12, 14, 16, 18, 20). This initially suggests that only pre expansion sets observed overfitting. However given that both pre and post expansion groups are trained and tested on the same data, we think this result is indicative of a fault in post expansion methods. Specifically, in high expansion cut-off groups, up to 1000 samples are derived from a single entry, producing high sequence similarity in the data set. The subsequent post expansion train and test sets derived then likely have high similarity, producing tests that are then not distinct from training, appearing to perform well, but are considerably invalid. Pre expansion train and test set division is then the optimal approach when using this technique.

Despite potentially exaggerated test performance in post expansion sets, pre expansions sets performed as well or better than post expansion sets for all drugs except ETR (Figures 13, 14). While we are uncertain why ETR's performance was notably lower, it may be attributable to a lack of ETR resistant samples in the data

set (Figure 24). Though, RPV has a similar deficit in resistant samples and yet performs well, suggesting another factor is involved. RPV actually performs well regardless of expansion cut-off, and appears to outperform RPV post expansion across all expansion cut-offs. We are unsure what may be causing RPV to perform so well, but a lack of resistant samples (Figure 24) may cause the model to always label negative, creating the appearance of learning. Though, this still does not explain RPV post expansion's relatively poorer performance, and in light of ETR's similar condition, it may be another factor altogether.

2. How Our Work Compares

In comparison to our results with those of Amammudy, we see that we were not able to fully recreate their results, and that in general, our networks performed substantially worse across all drugs. For all NRTI's and NNRTI's, Amamudy reported test r-squared values greater than 0.91, which is better than even the best of our Single Drug networks. While we note that AUC's are not equivalent to r-squared values, we would still expect graphs with areas

approaching 1. Given that we had mostly replicated their design, we can attribute this difference in results to two likely sources. First, we anticipate that Amammudy is likely performing the post expansion technique, as they do not indicate any distinction between pre and post techniques, and so the ordering of their methods implies that the post expansion method was used. Doing this could potentially over inflate test performance and so skew their results. But even our post expansion sets could not generate comparative results, and so we believe there to be another potential cause. This cause may be the use of a variety of unique network architectures for each Single Drug network. The use of these unique architectures likely optimized performance further, though as stated previously doing this is not considered best practice, and is certainly not easily replicable as the final unique architectures were not indicated. In Yu 2014, accuracies greater than 91% were reported using an ANN model. However, it appears that they may also be using the post expansion technique, though their exact neural architecture is not well discussed, making comparison limited.

3. The MultiDrug Network and Ambiguity Distribution

The MultiDrug networks performed poorly across all groups (Figures 21, 22). The change in expansion cut-off has no apparent trend across train or test sets in either post expansion or pre expansion. This result confirms that of Amamudy *et al*., and assures us that our model is comparable.

We also tracked the quantity of negative and positive labels for each drug across expansion cut-off as a way of determining the effects of expansion on data balance (Figure 24). In particular, our concern was that ambiguous sequences were distributed disproportionately in the data, such that there were more ambiguous sequences labeled negative or positive for a drug, resulting in unbalanced data after expansion. However, drug labeling balance appears mostly constant across expansions. This indicates that the initial data set is more diverse and well distributed with ambiguous sequences than originally expected.

Chapter 5: Concluding Remarks

1. Future Considerations

We had here explored a very specific quality of a certain algorithm used in HIV resistance prediction. In future endeavors, however, we would like to explore other routes of intrigue. Of particular interest is the encoding of protein structural and chemical data into classifiers. A number of authors here mentioned and also not mentioned have reported successful use of such data in classifiers, and so research into these classifiers to optimize network performance may be an avenue. We have also discussed other means of amino acid residue encoding which may confer more detailed information to a classifier, especially in the case of hydrophobicity and charge. A means of further characterizing the expansion method is still needed also. Creating a sort of expansion profile for a given data set which can indicate the degree of expansion possible for each individual original sequence may be useful in determining if the algorithm is a good fit for that data set.

2. Acknowledgements

I would like to take a moment to thank my thesis advisor, Dr. Hibbs, for his constant support throughout this project, without which this research would not have been possible. I would also like to thank my thesis committee members Dr. Lewis and Dr. Livingstone for their advice and direction. And of course my family and friends who have been a source of constant support during my time at Trinity

References

Amamuddy, Olivier Sheik, et al. "Improving Fold Resistance Prediction of HIV-1 against Protease and Reverse Transcriptase Inhibitors Using Artificial Neural Networks." *BMC Bioinformatics*, vol. 18, Aug. 2017, pp. 1–7.

Khalid, Z., & Sezerman, O. U. "Prediction of HIV Drug Resistance by Combining Sequence and Structural Properties." *IEEE/ACM Transactions on Computational Biology and Bioinformatics,*

Computational Biology and Bioinformatics, IEEE/ACM Transactions on, IEEE/ACM Trans. Comput. Biol. and Bioinf, no. 3, 2018, p. 966. *EBSCOhost*, doi:10.1109/TCBB.2016.2638821.

NIH. Antiretroviral Drug Discovery and Development. 2019, April 03.Retrieved from https://www.niaid.nih.gov/diseases-conditions/antiretroviral-drug-development

NIH. The HIV Life Cycle Understanding HIV/AIDS. 2018, July 27. Retrieved from https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/73/the-hiv-life-cycle

"NRTI Resistance Notes." *PI Resistance Notes - HIV Drug Resistance Database*, Stanford University, hivdb.stanford.edu/dr-summary/resistance-notes/NRTI/#30Miller2004.

Ji, J., & Loeb, L. A. Fidelity of HIV-1 Reverse Transcriptase Copying RNA in Vitro. Biochemistry, 31(4), 1992. 954–958.

Shen, C et al. "Automated Prediction of HIV Drug Resistance from Genotype Data." *BMC Bioinformatics*, vol. 17, Aug. 2016, pp. 563–569. *EBSCOhost*, doi:10.1186/s12859-016-1114-6.

Yu, Xiaxia et al. “Sparse Representation for Prediction of HIV-1 Protease Drug Resistance.” *Proceedings of the ... SIAM International Conference on Data Mining. SIAM International Conference on Data Mining* vol. 2013 (2013): 342-349. doi:10.1137/1.9781611972832.38

WHO. HIV/AIDS. 2018, July 19. Retrieved from https://www.who.int/news-room/fact-sheets/detail/hiv-aids

Xiaxia Yu, et al. “Prediction of HIV Drug Resistance from Genotype with Encoded Three-Dimensional Protein Structure.” *BMC Genomics*, vol. 15, July 2014, pp. 1–13. *EBSCOhost*, doi:10.1186/1471-2164-15-S5-S1