RESCUING PROGRESSION IN ANTIBIOTIC DISCOVERY BY INCREASING MACHINE LEARNING COMPATIBILITY OF HIGH-THROUGHPUT SCREENING

An Undergraduate Research Scholars Thesis

by

ROBERT TEVONIAN

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Shuiwang Ji, PhD

May 2021

Major: Biochemistry

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Robert Tevonian, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Research Faculty Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

ABSTRACT

Rescuing Progression in Antibiotic Discovery by Increasing Machine Learning Compatibility of High-Throughput Screening

Robert Tevonian
Department of Biochemistry & Biophysics
Texas A&M University

Research Faculty Advisor: Shuiwang Ji, PhD
Department of Computer Science and Engineering
Texas A&M University

Antibiotic discovery has stagnated. To avoid catastrophe, it must speed up. One of the most heavily used methods of drug discovery is high-throughput screening, yet in the 40 years of use of high-throughput screening, zero antibiotics have come to market from this method. Recent advancements in deep learning have provided a potential solution to this problem. It has been demonstrated that with a clean yet relatively small training dataset, meaningful predictions can be made on large chemical libraries. However, relying on cherry-picked data with extremely confident 'hits' or 'misses' fails to represent the uncertainty of large real-world datasets. In this paper, I analyze the current state of HTS and propose and new workflow that is compatible with machine learning. The key to machine learning compatibility is determined to be the aversion of false negatives. More specifically, it is most important to reduce the 'noise' relative to the size of the dataset for maximum compatibility. Furthermore, using the standard tool ChemProp, I discern that the size of matters significantly, and small datasets of strong data will still fail to be compatible with machine learning models.

DEDICATION

I dedicate this thesis to my family who helped me get to A&M in the first place and gave me the support I needed early on to be successful. I am ever grateful.

ACKNOWLEDGEMENTS

Contributors

I would like to thank my faculty advisor, Dr. Shuiwang Ji, and my colleague in Dr. Ji’s lab, Bora Oztekin, for their technical guidance and support throughout the course of this research. I also would like to thank Dr. Dwight Baker in the Sacchettini lab for helping me understand datasets, and Jeremy Wood for creating the most coherent HTS data available from the Sacchettini lab.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my family for their encouragement.

The data analyzed for the dataset size-compatibility analysis was provided by Jeremy Woods. The first machine learning model used was provided open-source by MIT – known as Chemprop. The second machine learning model used was provided by Dr. Ji’s lab – known as MoleculeKit.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

No explicit funding was given for this project. However, the Sacchettini Laboratory at Texas A&M University provided various resources. The data for the initial screening campaign was provided by this laboratory.

NOMENCLATURE

Hit	Prospective active molecule
Miss	Prospective inactive molecule
HTS	High Throughput Screening
Fingerprinting	Generating a ‘fingerprint’ for molecules based on their molecular features
Fingerprint	The unique identity of a molecule, which can be plotted in a chemical space
GNN	Graph Neural Network
IC50	50% inhibitory concentration – the concentration of an antimicrobial agent required to inhibit 50% of the growth of a microbe
MTb	*Mycobacterium Tuberculosis*
AMR	*Antimicrobial Resistant*

1) INTRODUCTION

1.1 The Problem and Historical Context

Continuous antibiotic discovery is undoubtedly a crucial process to maintain stability of society. Despite a significant negative societal impact if this process slows, little progress has been made in antibiotic discovery since the "Golden Age" of the 1940s-1970s. [1] Decreasing economic incentives to those who research antibiotics paired with hastening antibiotic resistance has made this a uniquely tough challenge. [2] On the forefront of this problem is academia – conducting large scale high-throughput screening campaings across the world. However, in nearly 40 years of high-throughput screening, not a single antibiotic has come to market from this method. [3] A solution is needed to reconcile these massive libraries of data and millions of man-hours – not only so that the research will not have been done in vain, but so directed screening campaigns can be much more efficient in the future. The solution proposed in this paper is a refined approach to antibiotic discovery building off of previous deep-learning research.

1.1.1 The Dire Need for New Antibiotics

The immediate societal effects of pandemic-scale mortality are now known by everyone, with the seeming irreversible changes that COVID-19 has brought. This is the scale of the problem that antibiotic resistance presents if a solution is not found. Current projections estimate that deaths from antibiotic resistant infections will reach over 10 million per year by 2050, placing mortality around that of cancer. Among the obvious moral obligation to prevent these deaths if possible, the economic impact is also extreme. This increase in mortality is estimated to drop global GDP by 2-3.5%, just from the drop in the workforce. [4] Economic impacts will be much more extreme if COVID-19-esque prevention measures are taken, as 10 million deaths a year is roughly 4 times COVID's mortality rate.

To make matters worse, the economic incentives to discover antibiotics have been deficient and declining over time. There are too few drugs in development to address the range of antibiotic resistant infections, and it is inevitable that most of the drugs in development will be discontinued. Only 60% of drugs that get to Phase 3 are approved. Also, it generally costs over a billion dollars to get a single drug to market, yet 95% of antibiotics in development are being researched by small pharmaceutical companies. [2] This means that if there are not immediate positive results, and a top pharmaceutical company doesn't invest, even promising antibiotics will have development discontinued.

To address the problem of economic incentives and the FDA price barriers, the initial discovery phase must be made much cheaper, and the confidence of initial screening must be much higher. Using deep-learning and virtual screening is a very cost-effective solution, but historically, has fallen short in terms of accuracy. [5] Recent technological advancements, data science solutions, and a shift in screening methodology may be able to solve these shortcomings.

1.1.2 Historical Approaches to Antibiotic Discovery

There are 3 main historical eras of antibiotic discovery: the Golden Era, the Medicinal Chemistry Era, and the Resistance Era. [5] Despite the extreme success of antibiotic discovery when biochemistry was an emerging field, the process has slowed in recent times.

1.1.2.1 The Golden Era (1929-1960s)

The discovery of penicillin in 1929 ushered in an era of antibiotic discovery in natural products. Specifically, there would be screening of microbes in soil that produce secondary metabolites to prevent bacterial infections. This Golden Era of natural product discovery lead to the discovery of some of the most widely used antibiotics – the Beta-Lactams (penicillin), the Macrolides, Tetracyclines, and more. There was not an absence of synthetic antibiotics from this time, however, characterized by the Sulfa- drugs (1932), Oxazolidinones (1955), and Quinolones (1961). [6] This era was ended for two reasons: 1) These drugs had significant toxicological drawbacks due to the lack of human design, and 2) Resistance due to horizontal gene transfer happened rapidly. This second reason introduces a trend in the history of antibiotic discovery – adaptive resistance beats the discovery method and heralds a new era. [5]

1.1.2.2 The Medicinal Era (1970s-1980s)

The Medicinal Era began in the 1970s, as a response to the inefficiencies that came from discovery without design. So, researchers rode the coattails of the Golden Era discoveries and focused on design. There were very few novel 'scaffolds' discovered in this era, but there was a huge emphasis on using medicinal chemistry to take the natural molecular scaffolds discovered in the golden era and to improve upon them. This allowed for them to fix toxicological parameters, improve delivery and potency, and avoid resistance for another couple decades. However, the inevitable happened, and resistance caught up again and forced researchers to reevaluate their approach.

1.1.2.3 The Resistance Era (1990s – Present)

The Resistance Era marks the mass adoption of high throughput screening. The innovation of high throughput synthesis and chemical handling robots allowed for large chemical libraries to be made and quickly screened. The technological revolution of accessible computers in this era allowed for extremely large datasets to be taken and handled. [5] However, this is where the issue seems to lay. Since the advent of this era, no antibiotics have come to market from high throughput screening. The overall efficiency and success of screening is extremely low, despite unprecedented advances in technology.

The problem becomes clearer when considering the scale of the chemical space in which high throughput screening attempts to discover drugs. Following a restrictive structural rule that tends to describe FDA-approved drugs, Lipinski's Rule of 5, the amount of druggable molecules that can be screened is conservatively estimated to be 10^{60}. [7] This is more than the number of stars in the universe, squared.

Historically, evolution has been able to navigate this space by randomly trying different molecules for millions of years and only keeping what works. Now, we have reached the limits of natural product screening, so we try to navigate this incomprehensibly big space manually. 'Manually', in this context, means 'in a reasonable amount of time'. Given this scale, it is very evident why these irrational brute-force screening efforts have failed. Rational drug discovery must be paired with irrational (brute force HTS) drug discovery if we want a chance at gaining any useful information. The elite tool for this is deep learning.

1.2 The State of Research

Neural networks are the ideal tool for virtual screening for a variety of reasons. The efficacy is best demonstrated in the flagship paper by Stokes et. al in 2019, *A Deep Learning Approach to Antibiotic Discovery*. Specifically, the researchers trained a new generation of deep neural networks called graph neural networks (GNNs), which were released the year prior in a package called ChemProp. [8] They were able to screen the Drug Repurposing Hub and discover a novel antibiotic. Since being released in 2020, it has been cited over 300 times. This paper undoubtedly caused excitement, but the methods and data they used differ from the real-world high throughput screening situation in a few key ways. In the following sections, current GNN models are explained, and key issues are highlighted in regards to the useability of the software on real world data.

1.2.1 Graph Neural Networks

ChemProp uses Message Passing Neural Networks (MPNNs) – a type of GNN. First of all, graphs are a type of data structure in which you have nodes (vertices) and many connections (edges) between individual nodes. GNNs use the graph data structure. Each node integrates data from neighboring nodes until the final layer is reached, and a classification is determined. In ChemProp, MPNNs with graphs that have atoms as vertices and bonds as edges are used. MPNNs work in two phases. First, the ‘message passing phase’ transmits information across the molecule to build a neural representation of the molecule of interest. Second, there is a ‘readout phase’, in which predictions are made for the molecular property of interest using the built neural representation. [8] Creating a neural representation of the graph of the molecule rather than using fixed molecular descriptors (i.e., fingerprints) has proven to be uniquely effective in classifying drugs. [3] Again, however, the MPNNs used still use a type of pseudo-fingerprinting method. Specifically, the models in question will use atom features (i.e. atom type, # of bonds, etc.) or bond features (i.e. bond type, stereochemistry, etc.) as information to pass along the network. The weight of individual features is set manually, and which features are used is artificially determined. This needs to be done in some fashion, as one cannot get a perfect representation of large molecules with a reasonable amount of computing power. However, this is important to take into consideration when considering chemical space, as described in the following section.

1.2.2 Chemical Space

Chemical space is the set of all possible stable molecules. Representations of chemical space are n-dimensional and grouped based on molecular features. When using different molecular features as dimensions in plotting chemical space, different spatial relationships can arise.

An intuitive way to think about how these neural networks classify molecules is thinking about the chemical space that the molecules inhabit. When the neural network is trained, it 'knows' certain structural parameters – the atom or bond features in MPNNs. The different atom or bond features serve as the dimensions of the chemical space that the neural network can see. The neural network eventually discovers a structural motif that is correlated with the desired classification (i.e. antibacterial activity). In this chemical space, since the motif is conserved among hits, the hits would be located near each other spatially.

In order for the neural network to predict activity based on a structural motif, the structural motif must be present (or nearly present) in the training data. Because of this, Stokes et. al [3] made sure that the chemical space of the training data and the library they predicted on were similar. They used t-SNE, an n-dimensional data visualization method, to plot the library they predicted on. Then, they picked an evenly distributed sample of chemicals in the same area of space to order and test. This was their training data. The best antibiotic discovered was found at an arbitrary location in this space.

This creates a key issue. In the many decades of high-throughput screening, there hasn't been as much special attention to sampling portions of chemical space that are similar to libraries available to predict on. Many labs end up with data that resides in a small, yet nicely distributed, section of chemical space. A neural network will struggle to make prediction on chemicals outside of this space, as it does not have the information required to make good predictions.

Another consideration to keep in mind is that the ways that methods of plotting chemical space yield different groupings. The MIT paper used t-SNE to plot their chemical space. Generally, the programs that plot with t-SNE have presets with a large amount of molecular parameters (dimensions). However, the neural network sees a chemical space based on its own molecular parameters. For best results, the data visualization method that decides the chemicals to train with should use the same fingerprinting method as the neural network.

The solutions I propose to these issues are:

1. When ordering chemicals for HTS, consider large libraries available to predict on, and purchase from a chemical space comparable to larger libraries.
2. Determine the training set's spatial similarity to the prediction libraries based on the molecular fingerprints that your neural network sees – not an arbitrary one.
3. When trying to reconcile data that isn't evenly distributed across a large chemical space, only predict on chemicals that reside in and around a similar chemical space as the training library. Otherwise, the predictions will have little correlation to reality despite a well-trained model.

1.2.3 Data Quality

There are massive issues with the quality of data in most high-throughput screening operations. There are many sources for error in data – from heat inconsistencies based on the well position on a plate that change the growth of a culture without any drug intervention, to experimental error associated with individual assays. Multiple screening campaigns and with the same and different assays have been shown to give wildly different results.

The issue of data quality in existing data is more easily addressed for hits than misses. In a screening campaign, the noisiest data is the initial screen. Then, the most promising candidates from the initial screen are re-tested and validated against possible assay experimental error. The next step for antibiotic discovery is to take the subsequent best candidates and generate IC50 curves. The IC50 curves are very confident, as the binary 'hit' or 'miss' value is determined from many datapoints rather than a single one. So, we can be confident that the existing hits are indeed hits. However, we can not be confident in the misses. The initial screening campaign is very noisy. There could be thousands of hits that came up as misses. If this error is significant enough, it will hurt the accuracy of the model in practice and may decouple the measures of accuracy of the model from its true accuracy. For example, precision recall curves (PRC) are used to judge the accuracy of a model. The model will predict on itself, and if the prediction aligns with the training data, it will be a 'true positive' or 'true negative'. If the prediction doesn't align with the training data, it will be a 'false positive' or 'false negative'. The area under the PRC (the PRC-AUC) then can give you a score based on the ratios of true/false positives/negatives to give you a measure of the accuracy of the model. Since certain misses may really be hits, even if the model is trained well, it will create more ‘true negatives’ than there are in reality. This means that even the scores of accuracy themselves will become inaccurate.

The solution to this is to create IC50 curves for both types of molecules – misses and hits – and to only train the neural network on molecules that you have IC50 data on. If you rely on any data from an initial screening campaign, the noise will significantly hurt the model. This makes screening harder and more expensive in the short term but gives you the power to predict on much larger groups of molecules than otherwise possible.

1.3 The New Workflow

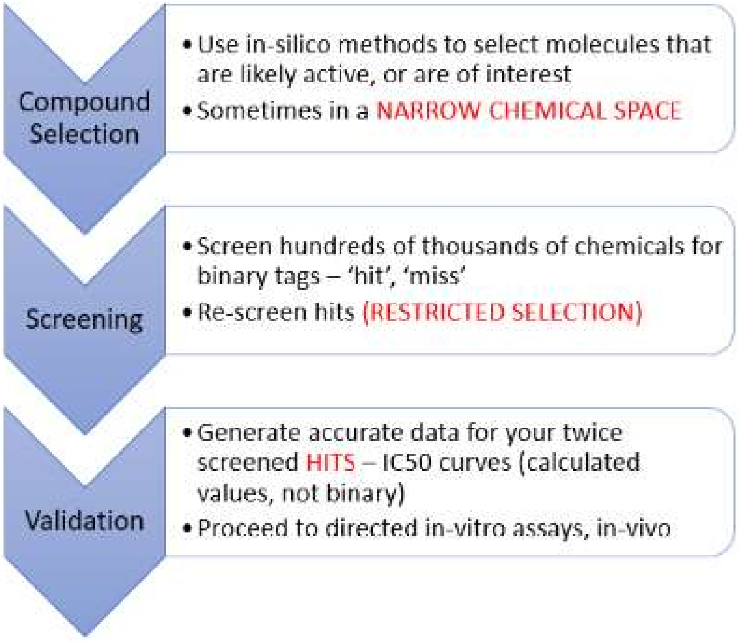
Based on the previous arguments about the current state of high-throughput screening and data quality, a new workflow can be constructed that is compatible with in-silico tools.

1.3.1 The Traditional HTS Workflow

The traditional high-throughput screening workflow is summarized in Figure 1 below. It begins with some type of virtual screening for compound selection. This essentially comes from a guess by the scientist. They will determine an interesting region of chemical space that may have certain molecular properties they think may be active, or simply choose an unexplored region of chemical space arbitrarily. They end up choosing a very *narrow region of chemical space*, which does not give useful information for the machine learning model. The machine learning model can only predict accurately on the chemical space neighbors of the tested molecules – the less similar the testing library is to the screening library, the lower the efficacy. So, from the first step, it is incompatible with machine learning.

The second step is the actual screening. In the traditional workflow, researchers do an initial screening and then select the 'hits' from the screening. Then, they do a secondary screening and rescreen the 'hits' to preliminarily validate the screening and select the best among the hits for further testing. This leads to a *restricted, hit-biased selection* of chemicals represented in the data. A machine learning compatible workflow must have equal confidence in hits and misses.

The final step in the traditional workflow is validation. Generally, the researcher validates the hit data with a calculated value. For example, an IC50 curve gathers datapoints of inhibition at different concentrations, and the concentration of 50% inhibition is calculated. This gives less random error. This data is again useless for machine learning models, as you need to have confident misses as well. Hit bias continues to pose a large problem.



Compound Selection

- Use in-silico methods to select molecules that are likely active, or are of interest
- Sometimes in a NARROW CHEMICAL SPACE

Screening

- Screen hundreds of thousands of chemicals for binary tags – 'hit', 'miss'
- Re-screen hits (RESTRICTED SELECTION)

Validation

- Generate accurate data for your twice screened HITS – IC50 curves (calculated values, not binary)
- Proceed to directed in-vitro assays, in-vivo

Fig 1. Traditional HTS Workflow. This method has a low success rate, a low number of molecules screened, and a high chance for false negatives. This is costly and time consuming.

1.3.2 Machine-Learning Catered HTS Workflow

The machine-learning catered HTS workflow is summarized in Figure 2 below. The first step is to consider the library that the researchers want to predict on, rather that what they want to screen. Once you determine this un-screened library via your data science tools of choice, researchers then can take a small, evenly distrusted sample representative of this chemical space – only a few thousand molecules. The larger the sample of chemical space you choose, however, the more molecules you will need to select for screening. Machine learning models find trends among structural motifs of the molecules and correlates them with a 'hit' or 'miss' tag, so the structural motifs of the chemicals in the prediction space must be somewhat represented in the training (screening) data.

Once your screening set is determined, the researcher screens once to determine hits and misses. Then, the researcher selects a sample rate of 'hits' vs 'misses' in accordance to the historical performance of the machine learning model – i.e., a specific model may perform better with a 4% hit rate than an 8% hit rate. This is learned over multiple trials. This gives the researcher a lot more control over the data. They now have a manual, configurable, non-hit-biased selection criteria that can be adjusted to give better performance if the model underperforms. Another validation screening isn't required, as the data should be accurate enough to move on to the IC50 curves and generate concrete data.

Finally, the researcher validates *both hits and misses* with calculated values such as IC50 curves. This takes longer than the IC50 hit validation in the traditional workflow, since it's testing at more chemicals by this stage. However, since the initial screening was much smaller and a second initial screening to re-test the hits did not need to happen, time is still saved. Then, the researcher trains the model and predicts on the prediction library. Then, they take the most confident hits from the prediction library and use more in-silico methods to further discriminate them (toxicity-predicting neural networks, drugability predictions, etc.). Finally, they order these chemicals and test them empirically.

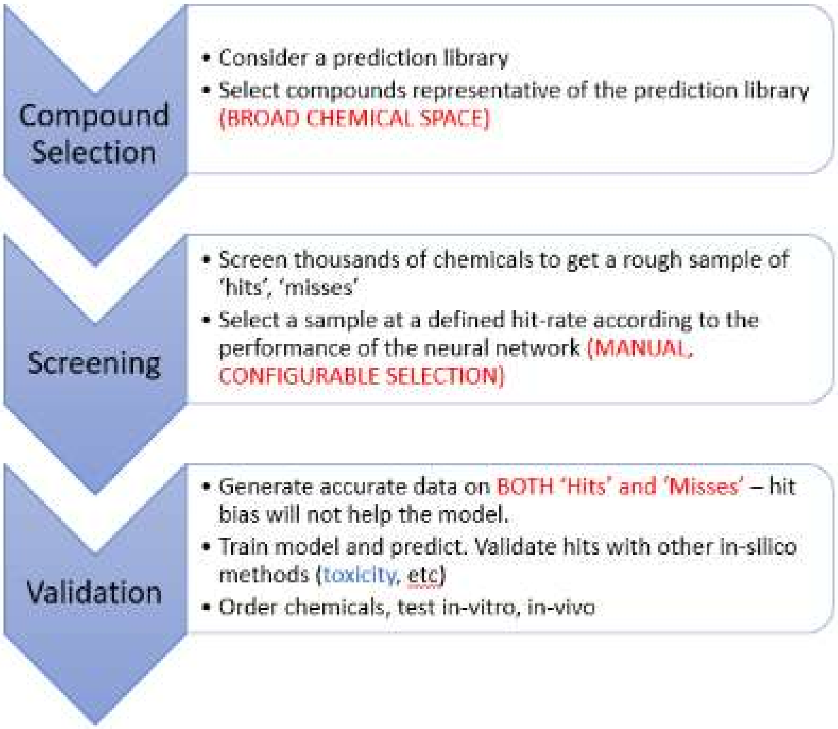

Compound Selection
• Consider a prediction library
• Select compounds representative of the prediction library (BROAD CHEMICAL SPACE)
Screening
• Screen thousands of chemicals to get a rough sample of 'hits', 'misses'
• Select a sample at a defined hit-rate according to the performance of the neural network (MANUAL, CONFIGURABLE SELECTION)
Validation
• Generate accurate data on BOTH 'Hits' and 'Misses' – hit bias will not help the model.
• Train model and predict. Validate hits with other in-silico methods (toxicity, etc)
• Order chemicals, test in-vitro, in-vivo

Fig 2. HTS Workflow with In-Silico screening compatibility. There is a lower chance of false negatives with this workflow. Because of this, millions more compounds can be screened in a cheap and effective manner.

1.3.2.1 How to Reconcile Old Data

There is still 40 years of old HTS data that has not been able to create new antibiotics. The best way to reconcile this data is to do a variation of the new HTS workflow – one that is directed at re-screening the misses that are in the neighborhood in chemical space of the confident 'hits' that were previously discovered. So, the workflow is modified to be:

1. Analyze the chemical space of the 'hits'

2. Choose the desired hit-miss ratio, and take the data for misses evenly distrusted across the space of the validated hits
3. Validate (generate IC50 curves) the misses.
4. Train the model and predict on the chemical space that the screening data resides in.

Without any re-screening, the data is virtually useless. This is the most efficient way to make this data useful again – for target agnostic high-throughput screening. However, if the screening is for docking on a known protein target, there is software that can perform hit-enrichment and reconcile these datasets without father screening. An example of this software is GFScore. [9] Again, if the target is unknown, further screening is required.

1.3.3 Workflow Discussion

Overall, the newly proposed HTS workflow is much more cost and time effective. It is able to achieve an unprecedented throughput. The key component of the new workflow that allows compatibility is the lack of *false negatives*. The other differences are merely consequences of machine learning's efficacy that can save time.

The false negative consideration (and thus the workflow modification) is an analytical argument based on the metrics that are used to appraise the performance of neural networks and is trivial – better quality data will yield better results. Regardless, in the following section, data with a higher chance of having false negatives is compared to confident data in order to assess what makes a 'good quality' dataset.

This is the questions that can be answered in this thesis. With machine learning methods, how much does the size of the dataset matter? There are powerful machine learning tools – if very small training datasets with confident data are used, will they still be accurate? Finally, is using confident data demonstrably better than using noisy data with false negatives? These questions are what will be addressed by using previously made real-world HTS data in the following sections.

2) METHODS

2.1 The Data

A primary high-throughput screening campaign for activity against *Mycobacterium Tuberculosis* was performed by Jeremy Woods in mid to late 2019 in the Sacchettini laboratory at Texas A&M. The initial screening was done via two separate assays: a luciferase-based and a resazurin-based assay. The library screened is 'SAC3', a collection of 9976 molecules selected based on the potential for activity. After the best 960 hits [further used as 957, as 3 structural identifiers are missing] were identified (3 plates, 320 chemicals per plate), dose response assays for both luciferase and resazurin assays were performed and IC50 values were calculated.

Relating to the previous sections, the IC50 values refer to the more confident values and the initial screening is less confident. However, the IC50 dataset is smaller.

2.1.1 Primary Screening – Hit selection

Two assay types were used. The first is the luciferin-luciferase gene reporting complex. The luciferase gene, when transfected into MTb, produces luciferase, an enzyme that acts on the luminescent protein luciferin. If the MTb survives, luminescence continues to occur. Error may come if the chemical inhibits gene expression or if the chemical inhibits luciferase itself.

The second type of assay used was a resazurin based assay. A resazurin assay relies on fluorescence rather than bioluminescence. In live MTb, resazurin is reduced to resorufin. Resorufin is fluorescent and can be detected by fluorescence spectroscopy. If the MTb dies, fluorescence stops. However, if the chemical interferes with the redox process there is error.

Another large source of error is ‘edge effects’. In the chemical plates, heating tends to be uneven based on the position of the plate in the incubator and the heat transfer of the plate itself. So, significant random error is introduced.

Because of all these sources of error, data will be more accurate if the hits are selective comparatively across different assays. A boolean discrimination was used to generate two different datasets. One had a ~5% hit rate (z-score < 1.5 in luciferase and resazurin, 516/9976), and the other had around a ~9.5% hit rate (z-score < 0.5 in luciferase and resazurin, 936/9976). These datasets were both used to train Chemprop, and AUC was used as the metric to determine how well the model can classify the training molecules.

2.1.2 IC50 Data – Hit selection

This data is the ‘most confident’ data that the Sacchettini lab had produced, as multiple doses are tested and the IC50 is calculated. Around half of these molecules had no discernable activity (>40 µM IC50). A Boolean discrimination was used to match the hit rates with the larger datasets. For the ~5% hit rate, the best 50 out of 957 molecules had an IC50 of less than 1.5 µM in luciferase and resazurin. For the ~9.5% hit rate, the best 91 out of 957 molecules had an IC50 of less than 3.5 µM in luciferase and resazurin. These datasets were both used to train Chemprop, and the AUC was recorded.

2.1.3 Generation of Primary Screening Datasets for Comparison with IC50 Data

Finally, two more training datasets were generated with 957 molecules each (50, 91 hits) to directly compare the IC50 data to the primary screening data. These datasets use the ‘hit’ or ‘miss’ criteria of the primary screening rather than the IC50 data, yet all of the structures are the same. So, this represents a small training set with data that has a high probability of having false negatives. This was used to train Chemprop, and the PRC-AUC was recorded. The details for the datasets referenced above are summarized in the appendix, in Table A.1 .

3) RESULTS

3.1 Results

Table 1: Compatibility of Various Datasets with Chemprop

Dataset	**PRC-AUC**
Large Primary Screening ~9.5% Hit rate	0.845
Large Primary Screening ~5% Hit rate	0.807
Small Primary Screening ~9.5% Hit rate	0.706
Small Primary Screening ~5% Hit rate	0.489
Small IC50 Screening ~9.5% Hit rate	0.722
Small IC50 Screening ~5% Hit rate	0.521

Table 1. The AUC measures the area under a PRC curve, which gives a metric of how well the machine learning model was able to classify the training data after being trained (closer to 1 is better). The 'PRC curve' is a plot comparing the 'precision' on the y axis to 'recall' on the x-axis as discrimination threshold is varied. These are defined as: precision = (true positives / true positives + false positives). recall = (true positives / true positives + false negatives). A good classifier will have a PRC-AUC of around 1, while a random classifier will have a PRC-AUC close to 0.5.

3.1.1 The Significance of Simple Data in Table 1

The data in Table 1 is simple, yet extremely significant. The most significant thing to note is that having a higher hit rate resulted in a much better PRC-AUC. First, appears that this does not scale in a linear fashion relative to the size of the dataset, as the difference between the PRC-AUC in the larger datasets is much smaller than the difference in the smaller datasets. In practice, this means that having a higher hit rate may give a more accurate model overall and may lend more confidence in seeing *some* activity in the predicted hits. However, this also means that the hits that are confidently predicted may not be *elite*. Having a lower threshold for hits may mean that you are training the model to create weak antibiotics. A sweet spot can be found by the researcher according to their individual needs.

The second thing to note is the difference between the small IC50 datasets and the small primary screening datasets. There is a clear increase in the PRC-AUC for the IC50 datasets, showing that this specific method of validation aids the model in a significant way. The PRC-AUC can be further enhanced by model-optimizing tools that likewise favor data with less noise, such as hyperparameter optimization.

The final note to make is that there is a huge difference between the PRC-AUC of both small datasets and the larger dataset that has ‘noisy’ data. Despite the IC50 data being more confident and working with the model better, having a larger training dataset simply creates a better model. There is likely an upper limit to this, as you begin to approximate the ‘true’ configuration of the neural network eventually.

3.1.2 Discussion – What does this tell us?

Through consideration of Table 1’s data, it is evident that goal of any modified HTS workflow is to *reduce as much noise as possible relative to the size of the dataset*. A few false negatives will not affect a very large dataset, yet a few may affect a small one significantly. Likewise, having many false negatives or positives will harm any model you try to train. However, the model should not be perfect. Having some noise is natural. If the PRC-AUC is 1, and the model retroactively predicts every training molecule ‘perfectly’, it is likely to be an overfit and may hold less extrapolatory power than a slightly lower PRC-AUC.

4) CONCLUSION

4.1 A Broad Spectrum Era

Contrary to the literature [5], the new era of antibiotic discovery will not be a narrow-spectrum era. Machine learning is the first tool to ever be able to traverse chemical space in a meaningful way. In configuring the workflow and model correctly, researchers are able to virtually screen millions of molecules within a day and tell how elite their antibiotic predictions may be. Because it's now easy to quantify the strength of predicted antibiotics by minimizing the training set hit rate relative to the resulting AUC, this spectrum will be *broad spectrum*. With this new ability to test millions of chemicals in a target-agnostic fashion, there is no reason to believe that the future of antibiotics is target-specific. There is likely to be another 'penicillin' in the future, that can continue to delay the AMR infection doomsday. With strong tools such as machine learning, there is no reason not to be optimistic about the future. We just need to make sure that the paradigm shift happens now that we know what must be done.

REFERENCES

[1] Wohlleben, W., Mast, Y., Stegmann, E., & Ziemert, N. (2016). Antibiotic drug discovery. Microbial biotechnology, 9(5), 541-548.

[2] Pew Charitable Trusts. (2020). Tracking the global pipeline of antibiotics in development, April 2020.

[3] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., ... & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.

[4] Review on Antimicrobial Resistance. (2014). Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Review on Antimicrobial Resistance.

[5] Brown, E. D., & Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. Nature, 529(7586), 336-343.

[6] Kealey, C., Creaven, C. A., Murphy, C. D., & Brady, C. B. (2017). New approaches to antibiotic discovery. Biotechnology letters, 39(6), 805-817.

[7] Reymond, J. L., & Awale, M. (2012). Exploring chemical space for drug discovery using the chemical universe database. ACS chemical neuroscience, 3(9), 649-657.

[8] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., ... & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. Journal of chemical information and modeling, 59(8), 3370-3388.

[9] Betzi, S., Suhre, K., Chétrit, B., Guerlesquin, F., & Morelli, X. (2006). GFscore: a general nonlinear consensus scoring function for high-throughput docking. Journal of chemical information and modeling, 46(4), 1704-1712.

APPENDIX: DATASET INFORMATION

Table A.1: Dataset Information

Dataset	Hits	Misses	Net Molecules
Large Primary Screening ~9.5% Hit rate	936	9040	9976
Large Primary Screening ~5% Hit rate	516	9460	9976
Small Primary Screening ~9.5% Hit rate	91	866	957
Small Primary Screening ~5% Hit rate	50	907	957
Small IC50 Screening ~9.5% Hit rate	91	866	957
Small IC50 Screening ~5% Hit rate	50	907	957

Table A.1: 6 different MTb datasets were used with the above number of molecules in order to simulate different dataset types that may be used for machine learning are shown. 4 datasets were primary screening. Two different assays – resazurin and luciferase – were used, and a boolean discrimination of z-scores determined the hit selection for the large primary screens. Then, the 2 IC50 datasets were processed with a boolean discrimination to get similar hit rates as the large datasets. Finally, the best 957 molecules from the large primary screening were taken to create the small primary screening datasets, and a boolean discrimination matched the hitrate to the IC50 data. This directly shows the advantage of IC50 data over primary screening discriminations. This also shows the advantage of large datasets. The data was collected by Jeremy Woods in the Sacchettini Lab @ TAMU.