To Kill a Mycobacterium:

Novel Antibiotic Discovery in *Mycobacterium tuberculosis H37Rv*

Marena Trinidad

INTRODUCTION

Pathogenic evolution is an arms race. Humans deploy a myriad of antibiotics to combat infection, whilst bacteria are forced to cope with the onslaught, do or die, relying solely on the grace of genetic mutation and population dynamics for survival. Be that as it may, modern medicine is falling short. Evolution has, of course, risen to the task and hyper-impervious, antibiotic-resistant strains continuously emerge. As such, today's post-penicillin World is facing a medical crisis. Humanity's current library of antibiotics is no longer panaceaic, and new antibiotics are required to secure the public's health.

One-third of Earth's population is a carrier of M. tuberculosis, culminating in 2 million deaths a year[1]. However, humanity faces an even greater challenge, for within The US alone, 10,042[2] patients develop treatment-resistant tuberculosis, with infections impervious to all clinically-available antibiotics. This poses a major threat to public health, as natural selection further increases the global presence of antibiotic-resistant tuberculosis. Subsequent effects could be catastrophic and citizen scientists have aligned with The WHO and other organizations[3] to fulfill the gaping demand for novel antibiotics.

This experiment is one such attempt, geared against one of the most notorious strains of M. tuberculosis, H37Rv. Hereby, publicly available datasets were curated from high-throughput, H37Rv chemical-screens and scrutinized in hopes of developing a pharmacophore-driven model capable of predicting bactericidal activity in uncharacterized compounds.

METHODS

DATA COLLECTION

For the purposes of this study, PubChem BioAssay AID1332[4] was web-scraped to substantiate a training dataset. The amalgamated information comprises a suite of 10,466 compounds, all of which were screened *in vitro* with H37Rv. Compound activity was classified by observing chemical lethality over a range of concentrations and timepoints. Chemicals that incurred >90% fatality amongst its initial bacterial-population were labeled "Active", whereas the

remaining compounds were annotated as “Inactive.” All activity assays were executed robotically in a highly controlled experiment as previously described [4].

For the drug-discovery purposes of this experiment, a natural-product database was stockpiled through granted access to repositories from The International Bio Screens Ltd.[6] and The Collaborative Drug Discovery Vault[3] . The assembled dataset contained itemized string representations of 64,566 compounds distilled from naturally occurring sources, varying from plant, marine and microbiological communities. Due to intellectual property constraints, this data is not provided in the project repository.

MOL-TO-VEC FEATURE REPRESENTATION

Molecules are collections of atomic building blocks, and compounds can be divided into arbitrary fragments and represented as a collection of contiguous-atomic motifs or “blocks.” Chemical “ingredients” can then be fashioned into an ordered list to represent a compound, which is easily converted to a numeric array by designating each position in the list to a specific fragment and then using a 0 (i.e. absent) or 1 (i.e.. present) at that index to detail a specific fragment’s presence. Though this approach may seem convoluted, it’s a form of chemical legos--a 1 at the first position in the list demarcates a green lego here and a 1 elsewhere in the list calls for a purple lego there.

Each activity screen was annotated with molecular-structure data encoded as SMILES strings--a streamlined system for representing chemicals in a single line of text. Each compound was then vectorized using the cheminformatic software RDKit[5] (*Release_2016.03.1*), according to the Morgan Fingerprint convention, in a parsimonious computation leveraging binomial trees and hash-function compression. In this fashion, all molecules were divided into small fragments, and each unique fragment was assigned a label, whereby all molecules can be wholly represented as a binary array detailing its fragment constituents. This form of molecular representation is highly fidelic, capable of resolving chemical characteristics down to their enantiomeric and chiral levels. Extensive details about the vectorization are made available by the software[5] providers, but the principle of Circular Fingerprint generation is highly tried and displayed in ***Figure 1.***

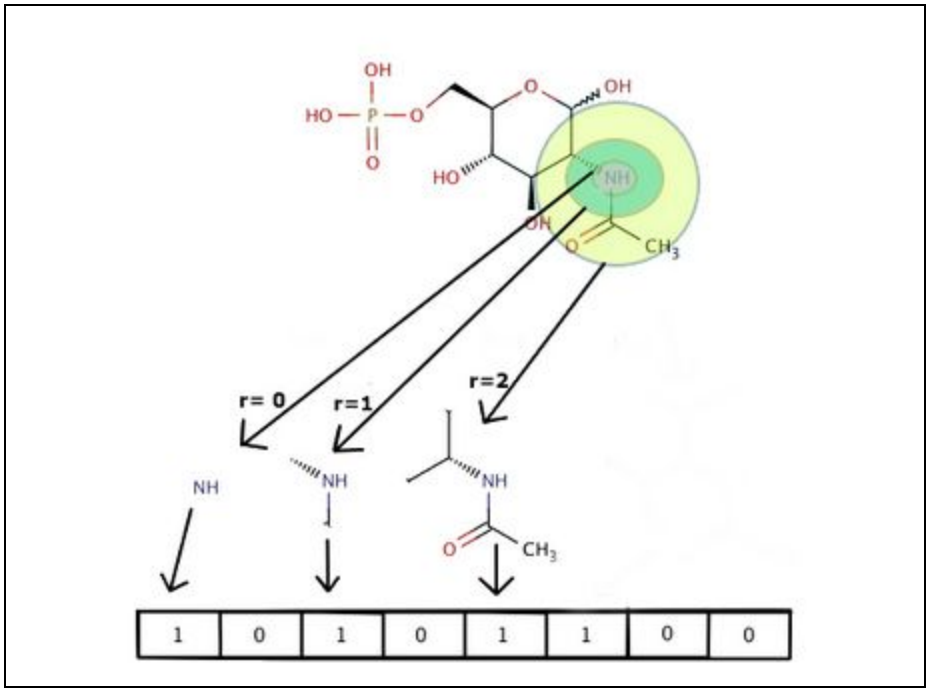


FIGURE 1: A simplified graphic detailing the Morgan Fingerprint process, in which a molecule is tokenized into fragments. Here, fragments are defined as any novel sequence of connected atoms, over a radius of 2, with the highlighted Nitrogen serving as a reference point. Unique fragment motifs are mapped to an array where each index serves as a binary feature detailing the inclusion of a fragment within the corporal lexicon

EXAMINATION OF THE MOLECULAR-FEATURE SPACE

Due to the complex nature of the chemical vectorization process, a comprehensive examination of the molecular-fingerprint landscape is essential in understanding the implications and validity of any model built upon it. Chief among this task is uncovering the distribution of compound activities within the training corpus. That is, 241 active compounds exist along with 10,225 inactive compounds(***Figure 2***). This drastic class imbalance will merit significant stringency in model recall during validation.

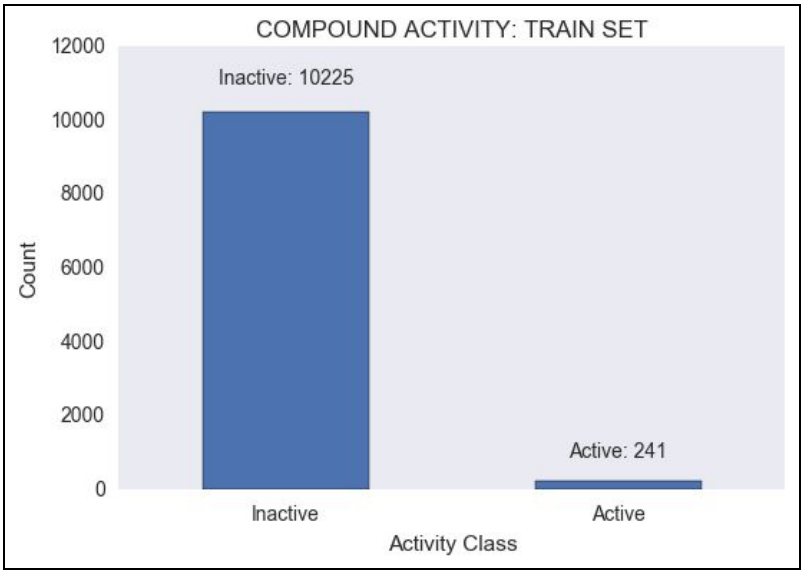


FIGURE 2: Bar chart demonstrating the activity-class distribution of the training set.

Deciphering each bit, and identifying its corresponding atomic-motif, offers a deeper insight into the feature space. Below, bit identities were unpacked and their respective frequencies were tabulated to isolate the most prevalent fragments within the corpus. Bit-specific frequencies are plotted in ***Figure 3***, along with summary statistics in ***Figure 4***, reporting fragment frequencies to vary widely (from 0 to 10,301), with the majority of fragments occurring less often than the median frequency and only a small proportion of bits existing at frequencies above the 75th percentile. To be specific, the corpus' most common fragments are depicted in ***Figure 5***. However, fragments frequencies across the corpus alone offer little help in discriminating between active and inactive compounds. Deeper analysis is required to quantify bit/feature importance and will be addressed in the next section.

FIGURE 3: Bit Frequency

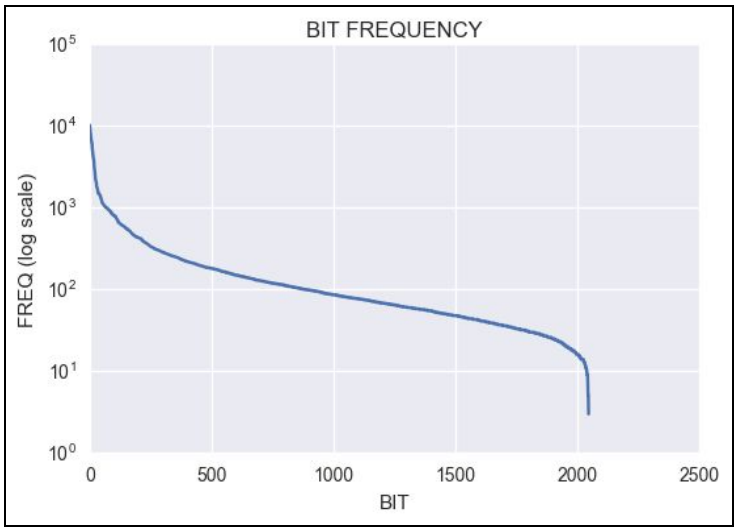
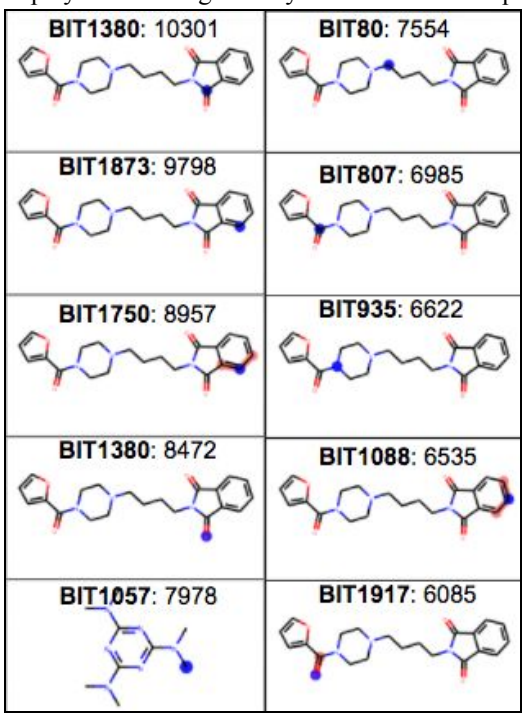


FIGURE 4: Bit Frequency Statistics

Mean	231.520
Standard Deviation	681.523
Minimum	3.000
Maximum	10301.000
25th Percentile	46.000
Median	84.000
75th Percentile	178.000

FIGURE 5: Most Frequent Bits

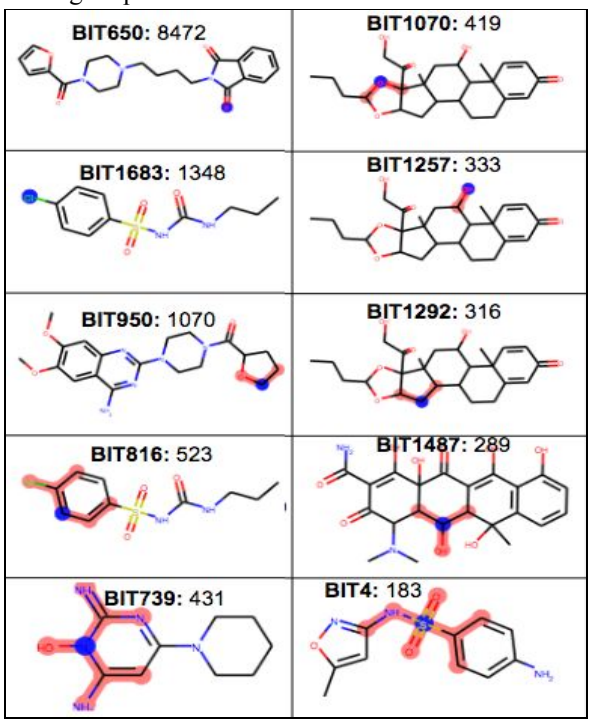
Fragments are highlighted in red, with the anchor molecule in blue, in the context of fragment molecules from the training corpus. The majority of high-frequency bits have a radius of zero and indicate single atoms with the displayed VSEPR geometry. Individual bit-frequencies are listed following each Id.



FEATURE ENGINEERING

To better understand the molecular features that drive activity, the theory of Information Gain was used to identify which atomic motifs most indicative of antituberculars. Though this computation is complicated, the premise is simple--information gain works to "weed out" redundant fragments (e.g. those ubiquitous across both activity classes) and identify fragments most unique to the active class. With this in mind, information gain was used to ascertain the top-ten fragments that best correspond with compound activity(***Figure 7***).

FIGURE 7: Top 10 Bits by Information Gain
Fragments are highlighted in red, with the anchor molecule in blue, in the context of exemplary molecules from the training corpus.



MODEL DEVELOPMENT

Antitubercular drug-discovery is, in essence, a binary classification problem--either a compound is "Active," killing over 90% of bacteria in the reported screen, or "Inactive." Due to the subsequent nature of this problem, a variety of binary classification algorithms were explored to derive a superlative predictor. The findings are outlined in ***Figure 8***.

Though the following table is hampered with jargon, the characteristics of a successful model are intuitive. For instance, the preeminent model should minimize the number of false positives and inaccurately predict only a few molecules as "active." This outcome is best described by precision and recall metrics, whose ideal values are closest to 1. By optimizing for negligible false positive rates like so, companies and researchers can ensure little to no time, and money, is wasted validating spurious candidates.

FIGURE 8: Model Statistics. Summarized model statistics, as derived from a stratified, train-test-split with 33.333% of the training corpus reserved for testing. See "ModelValidation" notebook for complete details.

Model	Recall	Precision	F1	Accuracy	ROC_AUC
Logistic Classifier	0.5732	0.7917	0.7287	0.9900	0.9970
Random Forrest	0.8516	0.9205	0.9139	0.9963	1.0000
Naive Bayes	1.0000	0.5691	0.2428	0.8569	0.9310
Gradient-Boosted LGC	0.2333	0.6256	0.3784	0.9802	0.9462

MODEL OPTIMIZATION

In overview, the best model--a Random Forest classifier--was obtained via SciKit-Learn software implementation[8]. The model details are enumerated in ***Figure 10***, for the curious reader, but do not feel obligated to hash out the details. Again, these details may seem dizzying, but the final model was successively tuned until it attained peerless precision and recall statistics, whose values approximated the ideal of 1. In this fashion, the model boasts the lowest false positives rates, such that future research and production needn't be stymied by the investigation of erroneous candidates.

FIGURE 10: Optimized Parmeters Random Forest Model.

RESULTS

CHEMICAL SCREENING OUTCOMES

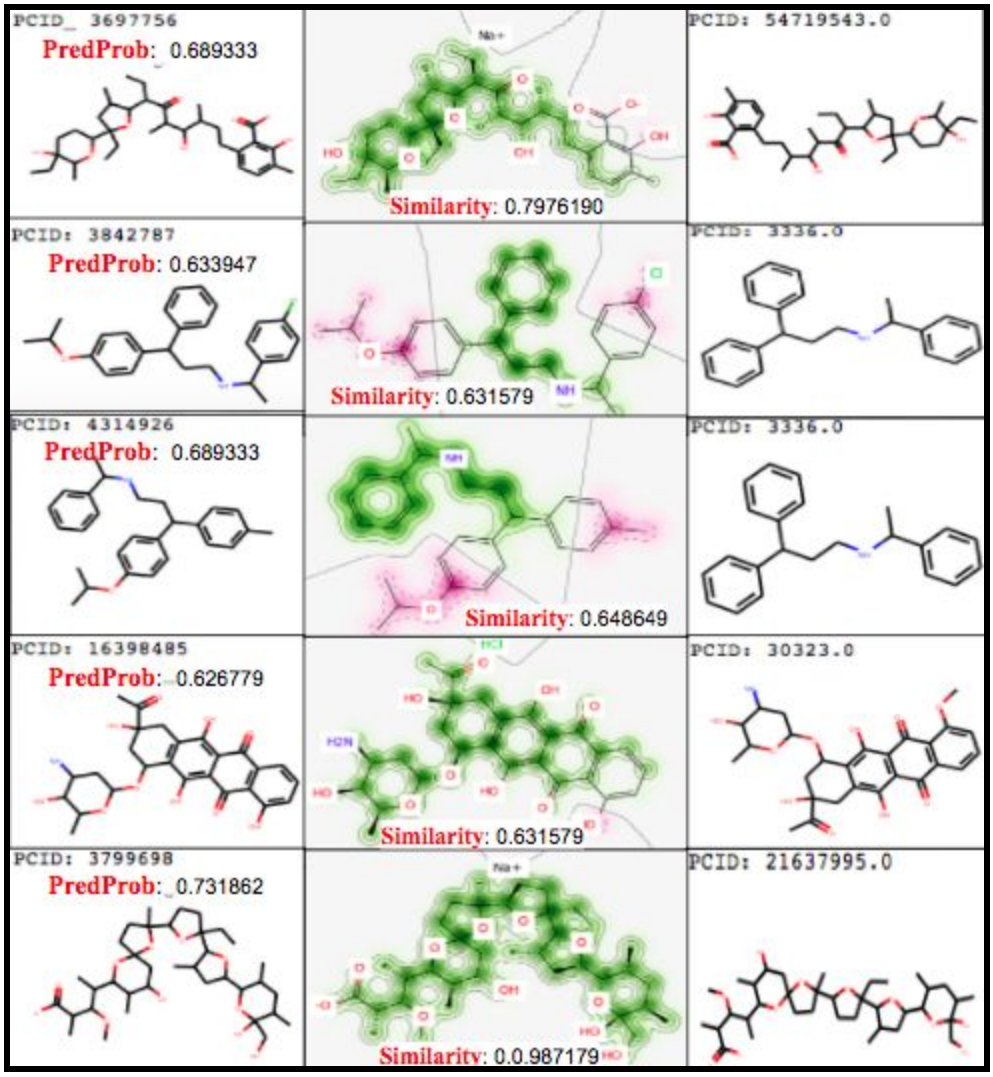


FIGURE 11: Predicted Antitubercular Candidates. Of the 29 predicted, 5 example compounds are listed above. Kekule structures for each are depicted in the leftmost column, then superimposed with respective top matches from the training-set (rightmost column) to generate electron similarity maps (center column).

The finalized model was used to screen the uncharacterized natural-product library, successfully returning 29 drug candidates. Each hit was validated by hand, comparing similarities between the candidate and it's top matching molecule from the training corpus. These figures implicate specific features that weighted toward the probabilistic activity-prediction made by the model. Fragments that contribute to similarity with the active class are highlighted in green and the fragments that detract from similarity are accentuated in magenta. Molecular fragments that had no effect on similarity are ubiquitous atomic motifs, frequent across all classes in the corpus,

and are left unaltered, encased within grey topological markings. Further details for deciphering compound similarity maps are elaborated in Riniker et al[9].

Though the high similarities between candidate and active training molecules bodes well for the predictions, raw Tanimoto Similarity (aka. Jaccard Distance between molecular bit/fragment vectors) scores are notoriously hard to interpret, especially as they are uncomparable across separate corpuses. However, the electron-density similarity maps above, allow for unequivocal demonstration of homologous structures between molecules. One can thereby leverage the fact that form-fits-function, such that the cumulative electrostatic surface of the molecules, from which chemical activity and binding proclivities are defined, must also be similar.

However provocative, these compounds require further validation to discern pharmaceutical viability. This shortcoming is mainly due to experimental design, for the model is fundamentally biased, having been trained solely on *in vitro* data. While the majority of presented candidates likely confer antitubercular activity, there is no ensuring their pharmacokinetic applications *in vivo*, that is, within the context of the human metabolic backdrop. For instance, within a host, a drug may very well become ineffective--perhaps becoming sequestered by off-target human proteins with superior binding affinity, or metabolised into a new, inactive form by ambient conditions. Notwithstanding, mouse-modeled *in vivo* screening data was recently acquired from The CDD and will be ensembled with the *in vitro* predictor into a pipeline that accounts for *in vivo* efficacy. To further increase the model's robustness, it will next be scaled to a training set of 250,000 compounds, followed by further hyperparameter optimization. This improved model will incur less bias, as the training corpus better captures the true complexity of the chemical universe through observation of a more vast collection of molecular fragments in the context of antitubercular activity. Though version 2.0 of this model has been implemented, its findings and details are withheld for intellectual property concerns.

REFERENCES

1. WHO. (2016, October). WHO | Tuberculosis. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
2. Center for Disease Control. (2016, September 8). Biggest Threats| Antibiotic/Antimicrobial Resistance | CDC. Retrieved from http://www.cdc.gov/drugresistance/biggest_threats.html
3. Collaborative Drug Discovery, Inc. (n.d.). CDD Vault: Modern Drug Data Management. Retrieved from https://www.collaborativedrug.com/cdd-vault
4. NIH: Southern Research Institute. (2008, June 26). AID 1332 - High Throughput Screen to Identify Inhibitors of Mycobacterium tuberculosis H37Rv - PubChem. Retrieved from https://pubchem.ncbi.nlm.nih.gov/bioassay/1332
5. Landrum, G. (2016). Module rdMolDescriptors. Retrieved October 11, from http://www.rdkit.org/docs/api/rdkit.Chem.rdMolDescriptors-module.html#GetMorganFingerprintAsBitVect
6. InterBioScreen Ltd. (n.d.). DOWNLOAD DATABASES. Retrieved October 30, 2016, from http://www.ibscreen.com/search.shtml
7. CAYUELA, L., GOTELLI, N. J., & COLWELL, R. K. (n.d.). Ecological and biogeographic null hypotheses for comparing rarefaction curves. Retrieved from https://www.researchgate.net/publication/282744514_Ecological_and_biogeographic_null_hypotheses_for_comparing_rarefaction_curves

8. Buitinik, L. (2013, September 1). API design for machine learning software: experiences from the scikit-learn project. Retrieved November 7, 2016, from https://arxiv.org/abs/1309.0238
9. Riniker, S., & Landrum, G. (2013, May). Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3852750/pdf/1758-2946-5-43.pdf