The danger of antibiotic resistance and deep learning solutions to drug resistance identification
Presentation agenda

Why is it important to talk about antibiotics resistance?

Why does antibiotic resistance occur?

How deep learning and we as computer scientists can help?

Example: predicting drug resistance for tuberculosis with explainability
The first antibiotic (penicillin) was discovered in 1928 by Alexander Fleming.

The advertisement from 1940s in South Carolina. Previously virtually untreatable gonorrhea can now be cured in 4 hours. The same applies to many other infections.
The time may come when penicillin can be bought by anyone in the shops. Here is a hypothetical illustration. Mr. X. has a sore throat. He buys some penicillin and gives himself, not enough to kill the streptococci but enough to educate them to resist penicillin. He then infects his wife. Mrs. X gets pneumonia and is treated with penicillin. As the streptococci are now resistant to penicillin the treatment fails. Mrs. X dies. Who is primarily responsible for Mrs. X’s death? Why Mr. X whose negligent use of penicillin changed the nature of the microbe.

Moral: If you use penicillin, use enough

Nobel Lecture, December 11, 1945
The rate of discovery of new antibiotics

LOW APPROVAL RATINGS
In the United States, the number of new antibiotics approved for use declined between 1980 and 2014, but approvals for cancer drugs rose.

- **Antibiotics**
- **Cancer drugs**
The rate of discovery of new antibiotics

- 1928: Penicillins
- 1932: Sulfonamides
- 1943: Aminoglycosides, Bacitracin (topical)
- 1945: Tetracyclines
- 1948: Cephalosporins
- 1947: Polymyxins, Phenolics
- 1946: Nitrofurans
- 1950: Pleuromutilins
- 1952: Macrolides
- 1953: Glycopeptides, Nitroimidazoles, Streptogramins
- 1955: Cycloserine, Novobiocin
- 1957: Rifamycins
- 1961: Trimethoprim
- 1962: Quinolones, Lincosamides, Fusidic acid
- 1969: Fosfomycin
- 1971: Mupirocin
- 1976: Carbapenems
- 1978: Oxazolidinones
- 1979: Monobactams
- 1987: Lipopeptides

© ReAct Group 2015
The antibiotic resistance catastrophe

- Resistance increasing to all know antibiotic drugs
- No new class of antibiotic drug discovered since 80s
- Significant time gap (10-15 years) from project inception to time on market
- 1.27 million deaths globally attributable to antibiotic resistance occurred in 2019

Antimicrobial resistance is a widespread serious threat that is no longer a prediction for the future, it is happening right now in every region of the world and has the potential to affect anyone, of any age, in any country

The WHO, spring 2014
The dangers of pre-antibiotic era

- Death from common infections: gardening scratches, strep throat (angina)
- Death from hospital-acquired infections after surgery
- Death following immunosuppression (HIV, transplantation, cancer treatment)
- Death from tuberculosis, pneumonia, sexually transmitted infections etc
Some bacteria are more resistant

Gram staining is a common technique used to differentiate two large groups of bacteria based on their different cell wall constituents.

Gram-negative bacteria are more antibiotic resistant because of the presence of outer membrane.

- Gram-negative includes enterococci, salmonella species and pseudomonas species.
- Gram-positive includes all staphylococci, all streptococci and some listeria species.
Parts of bacteria responsible for resistance

Circular DNA - responsible for DNA synthesis

Plasmid is a small, often circular DNA molecule found in bacteria and other cells. They generally carry only a small number of genes, notably some antibiotic resistance genes.

mRNA are synthesised here.

mRNA are read by the ribosomes to make polypeptides. Those proteins are the structures that help bacteria become antibiotic resistant.
Antibiotic resistance genes

**Antibiotic degrading enzymes**

β-Lactamases represent one of the most common causes of bacterial resistance to β-lactam antibiotics, particularly in Gram-negative bacteria. These enzymes can inactivate almost all β-lactam antibiotics by binding covalently to their carbonyl moiety and hydrolyzing the β-lactam ring.

![β-Lactam ring](image)

**Production of efflux pump**

Efflux pumps are transport proteins involved in the extrusion of toxic substrates (including virtually all classes of clinically relevant antibiotics) from within cells into the external environment. These proteins are found in both Gram-positive and -negative bacteria.

![Efflux pump](image)

**Modifying binding target**

Introducing modifications to the target site is one of the most common mechanisms of antibiotic resistance in bacterial pathogens affecting almost all families of antimicrobial compounds. One example of bacteria with very well developed mechanism is MRSA.
When bacterial cells die, they frequently lyse (burst) releasing their intracellular contents, including fragments of DNA, to the environment. These fragments can be taken up and incorporated into the chromosome of a living bacterium to provide the recipient with new characteristics.

Genetic information can also be carried from one bacterium to another by a virus. Bacteriophages are small viruses that infect bacteria and use their cellular components to make bacteriophage replicates. During the infection and replication, it is possible for bacterial genes to get incorporated into the viral genome. One of the viral replicates carrying the bacterial allele may then subsequently infect another bacterium and pass the new allele on.

Bacterial plasmids can carry genes that provide resistance to antibiotics, and bacteria that contain plasmids are able to conjugate with other bacteria and pass a replicate to recipient bacteria.
The resistance problem can be managed

- In the long run, we need to find alternative ways to prevent and treat bacterial infections and not rely solely on antibiotics. It is important to understand that the problem of antibiotic resistance cannot be "solved".

- However, we will still need antibiotics in the short- to medium-term time frame. It is, therefore important to learn from past mistakes to preserve any new antibiotic that reaches the market and to maintain and possibly enhance what power is left in the old ones.

- Continuous research supply of new antibiotics and a continuous needs analysis by the public health sector

- Alternative methods to treat bacterial infections, like using bacteriophages or antimicrobial peptides should be explored.

- Reviving older antibiotics that are not used for different reasons
How DL community can help?

1. Antibiotics discovery speed-up with deep learning models

"Using a machine-learning algorithm, MIT researchers have identified a powerful new antibiotic compound. In laboratory tests, the drug killed many of the world's most problematic disease-causing bacteria, including some strains that are resistant to all known antibiotics."


2. Identifying specific drug-resistance in bacteria

Most common ways to identify drug resistance is to take a sample of liquid and expose resident bacteria to different drugs. If the bacterial colony continues to divide and thrive despite the presence of a normally effective drug, it indicates the microbes are drug-resistant, but it takes weeks to get the result. We can utilize deep learning to predict drug resistance based on bacteria genetic information.
Deep learning predicts tuberculosis drug resistance status from genome sequencing data

Michael L. Chen et al, Harvard Medical School 2018
Tuberculosis

10th

10th tuberculosis place in the top of death causes worldwide 2022

4.1%

4.1% the percentage of new tuberculosis isolates that are multidrug resistant

48%

48% the percentage of patients that have unfavorable treatment outcomes attributable to drug resistance
Diagnosing resistance is a barrier

- Due to insufficient resources for building diagnostic laboratories, fewer than half of the countries with a high MDR-TB burden have modern diagnostic capabilities.

- Even in the best equipped laboratories, conventional culture and culture based antimicrobial susceptibility testing constitutes a considerable biohazard and requires weeks to months before results are reported due to the slow in vitro growth of Mycobacterium tuberculosis.

- Molecular diagnostics are now an increasingly common alternative to conventional cultures. The WHO has endorsed three such molecular tests, but they have a limited sensibility (mostly ranging from 63.7% to 94.4%).

- Molecular tests do not detect most rare gene variants of the targeted loci and only detect resistance to five anti-tuberculosis drugs, notably missing several key first line agents.
Main contributions of the paper

• Improved predictive tool to evaluate drug resistance for 10 antituberculosis drugs using a multidrug wide and deep neural network (MD-WDNN) framework.

• Presented multidrug framework predicts the full resistance profile, instead of predicting resistance of a single drug at a time. It simultaneously allows a drug to share resistance pathway information from the phenotypes of other drugs and incorporates prior knowledge that drug resistance can be caused by both direct genotype-phenotype relationships as well as epistatic effects.

• Analysed t-distributed Stochastic Neighbor Embedding (t-SNE) visualization and feature importance to examine inter-drug similarities.
Genotypic/phenotypic biomarkers as input

**Training data**

- **2,222** whole genome sequences of Mycobacterium Tuberculosis curated by the ReSeqTB knowledge base curating genotypic and phenotypic data of in vitro diagnostic assays for MTB

- **1,379** isolates that underwent sequencing using molecular inversion probes that targeted 28 preselected antibiotic resistance genes

**Testing data**

- **792** MTB isolates was obtained by pooling additional data from ReSeqTB, without overlap with the training set. Phenotype data *curated manually* according to protocols.

* some of the genomic regions that are known to be associated with resistance to particular antibiotics were not sequenced for those strains, so value of 0.5 is assigned instead
## Examples of the genes sequenced

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Drug resistance association</th>
<th>ID (H37Rv)</th>
<th>Strand</th>
<th>Start</th>
<th>End</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>promoter ahpC</td>
<td></td>
<td>Isoniazid</td>
<td>-</td>
<td>+</td>
<td>2726088</td>
<td>2726192</td>
<td>105</td>
</tr>
<tr>
<td>ahpC</td>
<td>alkyl hydroperoxide reductase C protein</td>
<td>Isoniazid</td>
<td>RV2428</td>
<td>+</td>
<td>2726193</td>
<td>2726780</td>
<td>588</td>
</tr>
<tr>
<td>air</td>
<td>alanine racemase</td>
<td>Cycloserine</td>
<td>RV3423c</td>
<td>-</td>
<td>3840194</td>
<td>3841420</td>
<td>1227</td>
</tr>
<tr>
<td>ddl</td>
<td>D-alanine-D-alanine ligase ddlA</td>
<td>Cycloserine</td>
<td>RV2981c</td>
<td>-</td>
<td>3336796</td>
<td>3337917</td>
<td>1122</td>
</tr>
<tr>
<td>embA</td>
<td>membrane indolylacetyllositol arabinosyltransferase A</td>
<td>Ethambutol</td>
<td>RV3794</td>
<td>+</td>
<td>4243233</td>
<td>4246517</td>
<td>3285</td>
</tr>
<tr>
<td>embB</td>
<td>membrane indolylacetyllositol arabinosyltransferase B</td>
<td>Ethambutol, Isoniazid, Rifampicin</td>
<td>RV3795</td>
<td>+</td>
<td>4246514</td>
<td>4249810</td>
<td>3297</td>
</tr>
<tr>
<td>embC</td>
<td>membrane indolylacetyllositol arabinosyltransferase C</td>
<td>Ethambutol</td>
<td>RV3793</td>
<td>+</td>
<td>4239863</td>
<td>4243147</td>
<td>3285</td>
</tr>
<tr>
<td>ethA</td>
<td>monoxygenase</td>
<td>Ethionamide</td>
<td>RV3854c</td>
<td>-</td>
<td>4326004</td>
<td>4327473</td>
<td>1470</td>
</tr>
<tr>
<td>gidB</td>
<td>glucose-inhibited division protein B</td>
<td>Streptomycin</td>
<td>RV3919c</td>
<td>-</td>
<td>4407528</td>
<td>4408202</td>
<td>675</td>
</tr>
<tr>
<td>gyrA</td>
<td>DNA gyrase subunit A</td>
<td>Fluoroquinolones</td>
<td>RV0006</td>
<td>+</td>
<td>7302</td>
<td>9818</td>
<td>2517</td>
</tr>
<tr>
<td>gyrB</td>
<td>DNA gyrase subunit B</td>
<td>Fluoroquinolones</td>
<td>RV0005</td>
<td>+</td>
<td>5123</td>
<td>7267</td>
<td>2145</td>
</tr>
<tr>
<td>inhA</td>
<td>NADH-dependent enoyl-[acyl-carrier-protein] reductase</td>
<td>Ethionamide, Isoniazid</td>
<td>RV1484</td>
<td>+</td>
<td>1674202</td>
<td>1675011</td>
<td>810</td>
</tr>
<tr>
<td>iniA</td>
<td>isoniazid inducible gene protein A</td>
<td>Ethambutol, Isoniazid</td>
<td>RV0342</td>
<td>+</td>
<td>410838</td>
<td>412760</td>
<td>1923</td>
</tr>
<tr>
<td>iniB</td>
<td>isoniazid inducible gene protein B</td>
<td>Ethambutol, Isoniazid</td>
<td>RV0341</td>
<td>+</td>
<td>409362</td>
<td>410801</td>
<td>1440</td>
</tr>
<tr>
<td>iniC</td>
<td>isoniazid inducible gene protein C</td>
<td>Ethambutol, Isoniazid</td>
<td>RV0343</td>
<td>+</td>
<td>412757</td>
<td>414238</td>
<td>1482</td>
</tr>
<tr>
<td>kasA (fabF1)</td>
<td>3-oxoacyl-[acyl-carrier protein] synthase I</td>
<td>Isoniazid</td>
<td>RV2245</td>
<td>+</td>
<td>2518115</td>
<td>2519365</td>
<td>1251</td>
</tr>
<tr>
<td>katG</td>
<td>catalase-peroxidase-peroxynitrite T</td>
<td>Isoniazid</td>
<td>RV1908c</td>
<td>-</td>
<td>2153889</td>
<td>2156111</td>
<td>2223</td>
</tr>
<tr>
<td>promoter mabA</td>
<td></td>
<td>Isoniazid</td>
<td>-</td>
<td>+</td>
<td>1673300</td>
<td>1673439</td>
<td>140</td>
</tr>
</tbody>
</table>
Distribution of target values

All isolates underwent culture based antibiotic susceptibility testing to two or more drugs at approved critical concentrations. Classified as resistant, susceptible, or not available.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Susceptible Isolates</th>
<th>Resistant Isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIF</td>
<td>2257</td>
<td>1285</td>
</tr>
<tr>
<td>INH</td>
<td>2011</td>
<td>1553</td>
</tr>
<tr>
<td>PZA</td>
<td>2445</td>
<td>702</td>
</tr>
<tr>
<td>EMB</td>
<td>2551</td>
<td>975</td>
</tr>
<tr>
<td>STR</td>
<td>1155</td>
<td>1025</td>
</tr>
<tr>
<td>CAP</td>
<td>799</td>
<td>589</td>
</tr>
<tr>
<td>AMK</td>
<td>1174</td>
<td>235</td>
</tr>
<tr>
<td>MOXI</td>
<td>1118</td>
<td>268</td>
</tr>
<tr>
<td>OFLX</td>
<td>651</td>
<td>88</td>
</tr>
<tr>
<td>KAN</td>
<td>1060</td>
<td>272</td>
</tr>
</tbody>
</table>
Multidrug deep and wide NN architecture

‘Wide’ logistic regression model is trained in tandem with a ‘deep’ MLP and the two models are merged in a final classification layer, allowing the network to learn useful rules directly from the input data and higher level nonlinear features.

For genomic data, the logistic regression portion of network can be thought of as modeling the additive portion genotype-phenotype relationship, while the MLP models the nonlinear or epistatic portion.

**Loss.** Weighted cross-entropy without penalization when true label $y_{i,j}$ is “unknown”.

---

**Figure 1.** A schematic of the multidrug wide and deep neural network architecture. Data flows from bottom to top through the wide (left) and deep (right) paths of the neural network. Nonlinear transformations, where applied, are depicted on the corresponding nodes. Each of the 11 nodes in the output layer represents resistance status predictions in all MTB isolates for one of the 11 anti-tuberculosis drugs.
Results

<table>
<thead>
<tr>
<th>Drug</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rifampicin</td>
<td>0.982</td>
<td>95.4%</td>
<td>97.8%</td>
<td>0.35</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>0.959</td>
<td>90.3%</td>
<td>96.4%</td>
<td>0.09</td>
</tr>
<tr>
<td>Pyrazinamide</td>
<td>0.883</td>
<td>75.2%</td>
<td>91.2%</td>
<td>0.32</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>0.922</td>
<td>90.6%</td>
<td>85.6%</td>
<td>0.40</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>0.942</td>
<td>90.1%</td>
<td>89.6%</td>
<td>0.26</td>
</tr>
<tr>
<td>Capreomycin</td>
<td>0.808</td>
<td>71.9%</td>
<td>85.7%</td>
<td>0.23</td>
</tr>
<tr>
<td>Amikacin</td>
<td>0.950</td>
<td>89.5%</td>
<td>90.8%</td>
<td>0.20</td>
</tr>
<tr>
<td>Moxifloxacin</td>
<td>0.902</td>
<td>90.0%</td>
<td>91.0%</td>
<td>0.39</td>
</tr>
<tr>
<td>Ofloxacin</td>
<td>0.866</td>
<td>69.6%</td>
<td>93.7%</td>
<td>0.57</td>
</tr>
<tr>
<td>Kanamycin</td>
<td>0.879</td>
<td>81.1%</td>
<td>89.6%</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Ofloxacin

Fig 5. Tuberculosis drug resistance ROC performance curve of the MD-WDNN.

Table 1. Tuberculosis drug resistance predictive performance of the MD-WDNN on the independent validation set. We also report sensitivity and specificity performance with the probability threshold chosen to maximize the sum of sensitivity and specificity for all anti-tuberculosis drugs.
Single vs Multi-Drug model

The multidrug architecture of the MD-WDNN resulted in a higher average AUC for both first-line and second-line drugs.

It suggests the concept of cross-training is useful when working with genotype/phenotype data.

Fig 4. Comparison of tuberculosis drug resistance predictive performance between single drug and multidrug models.
Visualization of embeddings

- clear separation of network output representation between resistant and susceptible isolates
- ability to classify resistance across multiple drugs, separating them into nested groups of pan-susceptible isolates, followed by mono-INH resistant isolates, multidrug resistant isolates, pre-XDR isolates, and XDR isolates, which is consistent with the order of administration of the drugs clinically as well as the usual order of MTB drug resistance acquisition.
- second-line injectable drugs, AMI, CAP, and KAN show similarly-classified clusters, highlighting the well-known moderate level of cross resistance between them
Genetic lineage of MTB

Sequence data from 33 genetic lineage markers were available in all 3,601 isolates and were used to measure genetic distance between isolates. The isolates fell into five well-defined clusters that corresponded to MTB's known genetic lineage. Overlying t-SNE coordinates for the MD-WDNN's probabilistic representation confirmed that the MD-WDNN's prediction of phenotype was not biased by lineage related variation.
Discussion

- the paper suggest an accurate algorithm for predicting Mycobacterium tuberculosis multi-drug resistance that has the potential to improve the treatment outcomes

- the authors show that multidrug NN achieves better accuracy than single drug one; also, the proposed NN architecture reaches better accuracy than classical ML algorithms (e.g. random forest, linear regression)

- the authors argue that combining precise *a priori* knowledge of causative relationship between genetic loci and drugs with 'permissive' approach for other biomarkers accounts for considerable performance gains, especially for second-line drugs

- model explainability approaches are important in case of usage in clinical settings as it allows to detect bias

- as more routine sequencing increases the amount of isolate data the model can be rapidly updated as the datasets become accessible

- the usage of raw genetic sequences may boost up the model performance even further
References

• Microbiology - Bacteria Antibiotic Resistance [https://www.youtube.com/watch?v=057phDG4mKU]

• The antibiotic discovery void | Professor Jeffery Errington FRS FMedSci [https://www.youtube.com/watch?v=db_MiwOBFIU&t=1352s]
