Machine Learning Prediction of Mycobacterial Cell Wall Permeability of Drugs and Drug-like Compounds

E.V. Radchenko$^{1,2}$, G.V. Antonyan$^1$, S.K. Ignatov$^2$, V.A. Palyulin$^{1,2}$

$^1$ Lomonosov Moscow State University, Russia
$^2$ Lobachevsky State University of Nizhny Novgorod, Russia

genie@qsar.chem.msu.ru
Drug-resistant tuberculosis (TB), a major global health challenge

- TB is caused by the pathogenic *Mycobacterium tuberculosis* (*Mt*)
- One of the most widespread and socially significant infections
- Every year, 1.6 million people die worldwide, making TB the leading cause of death from a single infectious agent
- New emerging strains of mycobacteria: multidrug-resistant (MDR) and extensively drug-resistant (XDR) tuberculosis
- HIV-associated TB
Severely lacking tuberculosis therapy options

- About a dozen antibiotic agents belonging to several drug classes are used clinically
- Often limited efficacy, long and inconvenient regimens, combination therapies
- Often toxicity and other adverse effects
- High risk of preexisting or developing drug resistance

- Massive worldwide efforts to identify novel promising anti-TB drug targets and active compounds
- Target-oriented drug development and optimization often unsuccessful
- High attrition rates
- Many compounds are potent against isolated targets but lack activity in whole-cell or in vivo settings
- One of the key causes: low penetration of a drug into Mtb cells
Key factor of *Mtb* resilience is its extremely complicated and persistent cell wall

Thick and dense outer membrane of mycolic acids: long molecules with hydrocarbon chains of ~70-90 carbon atoms

Also contains various porins, efflux pumps, and transporters

“Normal” lipid membrane

Figure from [Dulberger C.L. et al. *Nat Rev Microbiol*, 2020, 18, 47]
Prediction and optimization of *Mtb* “pharmacokinetics”

• Explicit modeling of drug permeation promising but complicated

• Effective complementary approach: use general QSAR methodology to derive predictive machine learning models

• **Key challenge:** lack of direct measurements of permeability

• **Solution:** indirect estimation from comparison of the target and whole-cell activities [originally proposed for the MycPermCheck model, *Merget et al., Bioinformatics* 2013, 29, 62–68]

• Implicitly captures not only membrane permeation but also active transport/efflux and inactivation
**Mtb permeability datasets based on Big Data analysis**

- Extensive anti-TB bioassay data are available in PubChem 2022

### Target-based assays

Total 926,660 compounds

9450 compounds active in at least one assay

<table>
<thead>
<tr>
<th>AID</th>
<th>ID</th>
<th>Type</th>
<th>Activity / Compound Count</th>
<th>Description</th>
<th>Activity condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>375</td>
<td>T01</td>
<td>Target</td>
<td>10011 / 10009</td>
<td>Mycobacterium tuberculosis pantothenate synthetase assay</td>
<td>Outcome</td>
</tr>
<tr>
<td>1376</td>
<td>T02</td>
<td>Target</td>
<td>216162 / 215860</td>
<td>Inhibitors of mycobacterial glucosamine-1-phosphate acetyl transferase (GlnU)</td>
<td>Outcome</td>
</tr>
<tr>
<td>2606</td>
<td>T03</td>
<td>Target</td>
<td>324858 / 324747</td>
<td>Primary biochemical high throughput screening assay to identify inhibitors of the membrane-associated serine protease Rv3671c in M. tuberculosis</td>
<td>Outcome</td>
</tr>
<tr>
<td>540406</td>
<td>T04</td>
<td>Target</td>
<td>324148 / 324048</td>
<td>High throughput screening of inhibitors of Mycobacterium tuberculosis UDP-galactopyranose mutase (UGM) enzyme</td>
<td>Outcome</td>
</tr>
<tr>
<td>540299</td>
<td>T05</td>
<td>Target</td>
<td>103205 / 102628</td>
<td>A screen for compounds that inhibit the MenB enzyme of Mycobacterium tuberculosis</td>
<td>Outcome</td>
</tr>
<tr>
<td>588335</td>
<td>T06</td>
<td>Target</td>
<td>356407 / 356160</td>
<td>Counterscreen for inhibitors of the fructose-bisphosphate aldolase (FBA) of M. tuberculosis</td>
<td>Outcome</td>
</tr>
<tr>
<td>1159583</td>
<td>T07</td>
<td>Target</td>
<td>356486 / 353572</td>
<td>Mycobacterium tuberculosis BioA enzyme inhibitor</td>
<td>Outcome</td>
</tr>
<tr>
<td>1626</td>
<td>C02</td>
<td>Cell</td>
<td>215397</td>
<td>High throughput screen to identify inhibitors of Mycobacterium tuberculosis H37Rv</td>
<td>Inh30</td>
</tr>
<tr>
<td>167178</td>
<td>T10</td>
<td>Target</td>
<td>67199 / 66591</td>
<td>Mycobacterium tuberculosis polyketide synthase 13 thioesterase (PKS13)</td>
<td>Inh30</td>
</tr>
</tbody>
</table>

### Cell-based assays

Total 557,527 compounds

96,040 compounds active in at least one assay

<table>
<thead>
<tr>
<th>AID</th>
<th>ID</th>
<th>Type</th>
<th>Activity / Compound Count</th>
<th>Description</th>
<th>Activity condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1332</td>
<td>C01</td>
<td>Cell</td>
<td>1118</td>
<td>High throughput screen to identify inhibitors of Mycobacterium tuberculosis H37Rv</td>
<td>Inh30</td>
</tr>
<tr>
<td>1949</td>
<td>C03</td>
<td>Cell</td>
<td>100697</td>
<td>High throughput screen of 100,000 compound library to identify inhibitors of Mycobacterium tuberculosis H37Rv</td>
<td>Inh30</td>
</tr>
<tr>
<td>2842</td>
<td>C04</td>
<td>Cell</td>
<td>23823</td>
<td>High throughput screen of a putative kinase compound library to identify inhibitors of Mycobacterium tuberculosis H37Rv</td>
<td>Inh30</td>
</tr>
<tr>
<td>1259343</td>
<td>C06</td>
<td>Cell</td>
<td>6225</td>
<td>High throughput screening of small molecules that kill Mycobacterium tuberculosis</td>
<td>Inh30</td>
</tr>
<tr>
<td>1259417</td>
<td>C07</td>
<td>Cell</td>
<td>1105</td>
<td>High throughput whole cell screen to identify inhibitors of Mycobacterium tuberculosis</td>
<td>Inh30</td>
</tr>
</tbody>
</table>

Total 557,527 compounds

96,040 compounds active in at least one assay
**Mtb permeability datasets based on Big Data analysis**

- Intersection of target-active and cell-tested compounds: 8242 compounds
- Compounds active in at least one cell-based assay are classified as penetrating ($MtbPen = 1$), otherwise as non-penetrating ($MtbPen = 0$)

**Full dataset $MtbPen8242$**
- 8242 compounds
- 2671 penetrating
- 5571 non-penetrating
- Moderately imbalanced data

**Balanced dataset $MtbPen5371ad$**
- 5371 compounds
- 2671 penetrating
- 2700 diverse non-penetrating
QSAR modeling: fragmental (substructural) molecular descriptors

- Occurrence counts (or presence) of fragments
- Thousands of fragments for real datasets
- “Holographic portrait” of a molecule
- Applicable to diverse series of compounds
- Easy prediction for new compounds
- Simple structural interpretation
- Mutual arrangement of structural features is handled indirectly via larger and/or overlapping fragments
- Acceptable for non-specific properties and/or diverse datasets

Up to 8 non-hydrogen atoms
Fragments present at least in 100 compounds

Basic subgraphs
Path (linear)
\[ p_1 \bullet \quad p_2 \bullet • • p_3 \bullet • • • p_4 \bullet • • • p_5 \bullet • • • \]

Cycles
\[ c_3 \bullet \quad c_4 \bullet • • c_5 \bullet • c_6 \bullet \]

Branches
\[ s_4 \bullet \quad s_5 \bullet • \quad s_6 \bullet \]

Hierarchical atom type classification
\[ * \rightarrow C \rightarrow C_{sp3} \rightarrow \text{CH}_3 \quad \text{CH}_2 \quad \text{CH} \quad \text{C} \]
\[ \rightarrow C_{sp2} \rightarrow \text{CH}_2 \quad \text{CH} \quad \text{C} \rightarrow \text{C} \]
\[ \rightarrow C_{sp} \rightarrow \equiv \text{CH} \quad \text{C} \rightarrow \text{C} \]
\[ \rightarrow \text{Car} \rightarrow \text{CH} \quad \text{C} \rightarrow \text{C} \]
\[ \rightarrow N_{sp3} \rightarrow \equiv \text{NH}_2 \quad \text{NH} \quad \text{N} \rightarrow \text{N} \]
Machine learning modeling approach

• Similar to ADMET modeling workflow
• Fragmental descriptors
• (Deep) feed-forward back-propagation neural network (BPNN)
• Repeated randomized double cross-validation (5x4 fold) to prevent overfitting and chance correlations
• Ensemble prediction

Perform endpoint scaling
Perform descriptor scaling
Perform descriptor selection
Repeat NR times
Split dataset into NO subsets
For each of NO subsets
  # Outer loop: use current subset for validation, other subsets for training
  Split outer loop training dataset into NI subsets
  For each of NI subsets
    # Inner loop: use current subset for termination, other subsets for training
    Build individual neural network model using other subsets for training and current subset for termination
    Evaluate model on the outer loop validation subset, collect statistics
    Save individual submodel
  Consolidate validation errors, compute final statistics
  Save complete ensemble model
Parallelized double cross-validation

- Neural network “forest” model
- TensorFlow 2.4.1/Keras 2.4.3
- High-performance NVIDIA RTX3080Ti GPU
- Hyperparameter optimization: fragment size, descriptor count, number and sizes of DNN layers, dropout
Predictive *Mtb* permeability models

**Full dataset *MtbPen8242***

- 500 fragmental descriptors up to 6 atoms
- 2 hidden layers
- $Acc_{cv} = 0.752$
- $BalAcc_{cv} = 0.683$
- $Sens_{cv} = 0.486$
- $Spec_{cv} = 0.880$

Low recognition of penetrating compounds, likely due to imbalance in favor of non-penetrating

**Balanced dataset *MtbPen5371ad***

- 900 fragmental descriptors up to 6 atoms
- 2 hidden layers
- $Acc_{cv} = 0.768$
- $BalAcc_{cv} = 0.768$
- $Sens_{cv} = 0.768$
- $Spec_{cv} = 0.769$

**Model can be used to screen or design likely penetrating compounds**

Acknowledgments

Laboratory of chemoinformatics and molecular modeling
Lobachevsky State University of Nizhny Novgorod

Strategic Academic Leadership Program “Priority 2030”