Collaborative Database and Computational Models
For Tuberculosis Drug Discovery Decision Making

Sean Ekins 1,2,3,4, Justin Bradford 1, Krishna Dole 1, Anna Spektor 1, Kellan Gregory 1, David Blondeau 1, Moses Hohman 1 and Barry A. Bunin 1

1 Collaborative Drug Discovery, Burlingame, CA.
2 Collaborations in Chemistry, Jenkintown, PA.
3 School of Pharmacy, Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD.
4 Department of Pharmacology, University of Medicine & Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, NJ.
MASSIVE datasets = Information overload

**Main Entry: mas-sive**

Pronunciation: /ˈma-siv\  
Function: adjective  
Etymology: Middle English mas-siffe, from Anglo-French mascif, alteration of massiz, from Vulgar Latin *massicius*, from Latin massa mass  
Date: 15th century  

1: forming or consisting of a large mass: a: **bulky** b: **weighty, heavy** c: **massive walls** d: **massive volume**  
c: impressively large or ponderous  
d: having no regular form but not necessarily lacking crystalline structure  

2 a: large, solid, or heavy in structure  
b: large in scope or degree  
<the feeling of frustration, of being ineffectual, is massive — David Halberstam>  
c (1): large in comparison to what is typical  
(2): being extensive and severe  
<massive hemorrhage> <massive collapse of a lung>  
(3): imposing in excellence or grandeur  
<massive dose of penicillin> <massive hemorrhage> <massive collapse of a lung>  
(3): imposing in excellence or grandeur  
<massive hemorrhage> <massive dose of penicillin>  
3: having mass  
<massive boson>  
— mas-sive-ly adverb
A history of software, model and database development, increasing drug development costs versus registered compounds in the CAS Registry and ChemSpider

Year

Registered Compounds (millions)
0 10 20 30 40 50 60

- Hansch analysis 1960's
- QSAR and drug metabolism 1972
- CoMFA developed 1988
- CYP CoMFA 1993
- SciFinder 1995
- PubChem 2004
- HMDB 2006
- DrugBank 2005
- eMolecules 2005
- ChemSpider 2009
- RSC purchases ChemSpider 2009
- Docking @home 2006
- NMR ALARM Filters 2005
- In Silico Models for absorption and BBB 1996
- P-gp models 2002
- hERG models 2002
- Systems Biology and ADME/Tox reviews 2000
- Cost of Drug Devt 2006
- Cost of Drug Devt 2003
- Cost of Drug Devt 1991
- Cost of Drug Devt 1979

Williams et al., Drug Discovery World, Winter 2009
Drug Discovery Today. Volume 00, Number 00 - February 2009

Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery

Moses Hohman¹, Kellan Gregory¹, Kelly Chibale², Peter J. Smith³, Sean Ekins⁴,⁵,⁶ and Barry Bunin¹

¹ Collaborative Drug Discovery, Inc. 1811 Gilbreth Road, Suite 220, Burlington, CA 94403, USA
² Institute of Infectious Disease and Molecular Medicine and Department of Chemistry, University of Cape Town, Rondebosch 7701, South Africa
³ Department of Pharmacology, School of Medicine, University of Cape Town, Medical School, 646, OMM, Groote Schuur Hospital, Observatory, 7925, South Africa
⁴ Collaborations in Chemistry, Jenkintown, PA 19046, USA
⁵ University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854, USA
⁶ Department of Pharmaceutical Sciences, University of Maryland, 30 Penn Street, Baltimore, MD 21201, USA

A convergence of different commercial and publicly accessible chemical informatics, databases and

Typical Lab: The Data Explosion Problem & Collaborations

DDT Feb 2009
Collaborative Drug Discovery's web-based software organizes preclinical research data to help scientists advance new drug candidates more effectively.

CDD offers an industrial-strength database at a price affordable to academic laboratories, research foundations, and companies of any size.

Analyze and mine your data intuitively through a web browser. Collaborate securely with other researchers in your own lab ... or across the globe. CDD's software accelerates international R&D projects combatting neglected diseases, as well as traditional commercial drug discovery programs.

Learn more »
Watch a three minute demo »
Register for a free trial account »
CDD Platform

- **CDD Vault** – Secure web-based place for private data – private by default
- **CDD Collaborate** – Selectively share subsets of data
- **CDD Public** – Public data sets - Over 3 Million compounds, with molecular properties, similarity and substructure searching, data plotting etc
  - will host datasets from companies, foundations etc
  - vendor libraries (Asinex, TimTec, Chembridge)
- **Unique to CDD** – simultaneously query your private data, collaborators’ data, & public data, Easy GUI
Funded by Bill & Melinda Gates Foundation for 2 years  Nov 2008- Oct 2010

Provide CDD software to Pilot groups and train them to use to store data

Provide custom cheminformatics support to pilot groups

Accelerate the discovery of new therapies and Advance clinical candidates through pipeline

Capture TB Literature and make accessible

CDD TB is freely available to any groups doing TB research (minimal support from CDD)

Integrate academic, non-profit, and corporate laboratories distributed across the globe

www.collaboratedrug.com
Building a disease community for TB

- Tuberculosis Kills 1.6-1.7m/yr (~1 every 8 seconds)
- 1/3rd of world’s population infected!!!!

- Multi drug resistance in 4.3% of cases and extensively drug resistant have increasing incidence
- No new drugs in over 40 yrs
- Drug-drug interactions and Co-morbidity with HIV

- Collaboration between groups is rare
- These groups may work on existing or new targets
- Use of computational methods with TB is rare
- Literature TB data is not well collated (SAR)

- Pub #17, Aug 23, 1.30-5.05pm Room 208

www.collaborativedrug.com
### Public Datasets for TB

- **15 public datasets for TB**
- **>300,000 cpds**
- **Patents, Papers Annotated by CDD**

### Shared Publicly

<table>
<thead>
<tr>
<th>Compound</th>
<th>Published By</th>
<th>Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB: TAAVF Assay Results</td>
<td>TB Early Phase Drug Discovery Program</td>
<td>812</td>
</tr>
<tr>
<td>TB: Literature Review</td>
<td>TB Literature Data</td>
<td>49</td>
</tr>
<tr>
<td>TB: Mics Prathipati GVKBio</td>
<td>CDD - Sean Elkins</td>
<td>2880</td>
</tr>
<tr>
<td>TB: Mics Prathipati NAID</td>
<td>CDD - Sean Elkins</td>
<td>3746</td>
</tr>
<tr>
<td>Sacchettini et al., review</td>
<td>CDD - Sean Elkins</td>
<td>14</td>
</tr>
<tr>
<td>MLSMR</td>
<td>Southern Research Institute</td>
<td>214307</td>
</tr>
<tr>
<td>TB: Mics Makarov et al., NM4TB consortia</td>
<td>CDD - Sean Elkins</td>
<td>32</td>
</tr>
<tr>
<td>TB Efficacy Data from Published Literature</td>
<td>CDD: TB Curated Literature</td>
<td>6771</td>
</tr>
<tr>
<td>TB Toxicity Data from Published Literature</td>
<td>CDD: TB Curated Literature</td>
<td>638</td>
</tr>
<tr>
<td>TB Pharmacokinetic Data from Published Literature</td>
<td>CDD: TB Curated Literature</td>
<td>28</td>
</tr>
<tr>
<td>TB Absorption Data from Published Literature</td>
<td>CDD: TB Curated Literature</td>
<td>24</td>
</tr>
</tbody>
</table>

**Additional Resources**

- [www.collaborativedrug.com/register](http://www.collaborativedrug.com/register)

---

**Additional Information**

- Antimicrobial activity of a publicly available library of 812 compounds against Mycobacterium tuberculosis (H37Rv) in AlamarBlue whole cell assay.
- Tuberculosis SAR data compiled in a survey of agents active against M. tuberculosis, including those with both known and unknown modes of action (Bottlet et al., “New Small-Molecule Synthetic Antimycobacterial” in antimicrobial agents and chemotherapy, June 2003). Updated 4/17 with TuberoList/1008/other target list and improved references.
- SAR Mics data from a recent publication by Prathipati et al at Novartis (PMD: 19053116). Consists of a dataset culled from the GVKBio database. The dataset was published as supplement information on the journal website.
- Literature TB Mics SAR data from a recent publication by Prathipati et al at Novartis (PMD: 19053116). Consists of a dataset culled from the NAID website. The dataset was published as supplement information on the journal website.
- A diverse collection of over 200,000 compounds collected by the Molecular Libraries Small Molecule Repository (MLSMR) were made available to the Southern Research Molecular Libraries Screening Center in Spring 2008 for primary testing against Mtb H37Rv. The most active compounds from this primary screen were selected and tested at 10 concentrations in both a dose response assay against H37Rv as well as a cytotoxicity excursion screen using vero cells.
- Structure activity relationship data for 1,3-benzothiazin-4-ones (BTZ). Data obtained from the paper “Benzothiazolinones Kill Mycobacterium tuberculosis by Blinding Arabinian Synthesis” published in Science by Makarov et al., 2009 and colleagues at the NM4TB consortia (PMD: 19209584).
### Chemical Properties

<table>
<thead>
<tr>
<th>Molecule</th>
<th>Molecular weight</th>
<th>log P</th>
<th>H-bond donors</th>
<th>H-bond acceptors</th>
<th>Lipinski violations</th>
<th>pKa</th>
<th>Exact mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDD-1730</td>
<td>327.846</td>
<td>2.18267</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>9.65963</td>
<td>327.16</td>
</tr>
<tr>
<td>CDD-1813</td>
<td>337.287</td>
<td>4.15005</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>8.00344</td>
<td>336.116</td>
</tr>
</tbody>
</table>

**(-)-Levobunolol hydrochloride**
TB Early Phase Drug Discovery Program
Visualizing in CDD

Data source: Search results

- CONFIGURE AXES
  - X: log P
  - Y: Mtb (H37Rv) Percent Inhibition (10 μg/mL): %

- DATA POINTS
- STATISTICS
- LABELS & PLOT OPTIONS
  - Title:
  - X-axis label: log P
  - Y-axis label: Mtb (H37Rv) Percent Inhibition
  - Legend: Display legend
  - Grid: Display grid

Mtb (H37Rv) Percent Inhibition (10 μg/mL): % Inhibition (%) statistics
- Mean: 5.25
- Standard Deviation: 17.16
- Min: 4.37
- Median: N/A
- Max: 100

www.collaborativedrug.com
Data source: Search results

5531498

Batch name: 1
Protocol: Mtb (H37Rv) Percent Inhibition (10 μg/mL)
Run: 2009-02-13
log P: 4.84326
% Inhibition 97.68
**Definition**

**Name:** 5531498

**Synonyms:** AB00277809

**Description:**

**Structure:**

```
CC(C):1cc(OCC(O)C=NC2CCN(CC2)c2ccc(F)c2)cc1C
```

**User-defined Fields**


**Owner:** Melinda Sosa
Learning from TB screening data: MLSMR PCA

PCA analysis - MLSMR dose response (blue, N = 2273) and known Mtb drugs (yellow) with simple descriptors. IV dosed Mtb drugs cluster (top right).

83.6% of the variance is explained using 9 descriptors in Discovery Studio

www.collaboratedrug.com

Ekins et al., Mol BioSyst, 6: 840-851, 2010
Laplacian-corrected Bayesian classifier models were generated using FCFP-6 and simple descriptors. 2 models 220K and 2K compounds active compounds with MIC < 5uM

Bayesian Classification Dose response

**Good**

- G1: 173580544
  - 10 out of 10 good
  - Bayesian Score: 1.253

- G2: 1393614142
  - 10 out of 10 good
  - Bayesian Score: 1.253

- G3: -1646949305
  - 9 out of 9 good
  - Bayesian Score: 1.228

- G4: -695845985
  - 9 out of 9 good
  - Bayesian Score: 1.228

- G5: 192299924
  - 8 out of 8 good
  - Bayesian Score: 1.199

**Bad**

- B1: -1699460258
  - 0 out of 56 good
  - Bayesian Score: -2.565

- B2: -1448737182
  - 0 out of 42 good
  - Bayesian Score: -2.303

- B3: 758147873
  - 0 out of 41 good
  - Bayesian Score: -2.281

- B4: 458381731
  - 0 out of 37 good
  - Bayesian Score: -2.189

- B5: -1185638464
  - 0 out of 33 good
  - Bayesian Score: -2.088

Ekins et al., Mol BioSyst, 6: 840-851, 2010

www.collaborativedrug.com
Bayesian Classification

Leave out 50% x 100

<table>
<thead>
<tr>
<th>Dataset</th>
<th>External ROC Score</th>
<th>Internal ROC Score</th>
<th>Concordance</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLSMR</td>
<td>0.86 ± 0</td>
<td>0.86 ± 0</td>
<td>78.56 ± 1.86</td>
<td>78.59 ± 1.94</td>
<td>77.13 ± 2.26</td>
</tr>
<tr>
<td>All single point screen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N = 220463)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLSMR dose response set</td>
<td>0.73 ± 0.01</td>
<td>0.75 ± 0.01</td>
<td>66.85 ± 4.06</td>
<td>67.21 ± 7.05</td>
<td>65.47 ± 7.96</td>
</tr>
<tr>
<td>(N = 2273)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ekins et al., Mol BioSyst, 6: 840-851, 2010

www.collaborativedrug.com
Both Bayesian models were evaluated with two separate test sets (NIAID and GVKbio) to describe how they ranked the most active molecules in each database.

Ekins et al., Mol BioSyst, 6: 840-851, 2010
<table>
<thead>
<tr>
<th>Number of compounds screened</th>
<th>Random hit rate (%)</th>
<th>single point screening (200k) Bayesian model (%)</th>
<th>dose response Bayesian model (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>1.66 (0.10)</td>
<td>23 (1.35)</td>
<td>24 (1.41)</td>
</tr>
<tr>
<td>200</td>
<td>3.32 (0.19)</td>
<td>48 (2.82)</td>
<td>42 (2.47)</td>
</tr>
<tr>
<td>300</td>
<td>4.98 (0.29)</td>
<td>64 (3.76)</td>
<td>54 (3.17)</td>
</tr>
<tr>
<td>400</td>
<td>6.63 (0.39)</td>
<td>77 (4.52)</td>
<td>58 (3.41)</td>
</tr>
<tr>
<td>500</td>
<td>8.29 (0.49)</td>
<td>92 (5.41)</td>
<td>70 (4.11)</td>
</tr>
<tr>
<td>600</td>
<td>9.95 (0.58)</td>
<td>107 (6.29)</td>
<td>82 (4.82)</td>
</tr>
</tbody>
</table>

>10 fold Enrichment with TB Bayesian model
Hits in the top 100 of the 100,000 compound test set

<table>
<thead>
<tr>
<th>Structure</th>
<th>CDD Number</th>
<th>Molecule Name</th>
<th>MLSMR 200k</th>
<th>Mtb (H37Rv) Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Structure" /></td>
<td>CDD-145638</td>
<td>7,734,051</td>
<td>37.649</td>
<td>96.12</td>
</tr>
<tr>
<td><img src="image2.png" alt="Structure" /></td>
<td>CDD-930626</td>
<td>6,607,472</td>
<td>35.4857</td>
<td>97.54</td>
</tr>
<tr>
<td><img src="image3.png" alt="Structure" /></td>
<td>CDD-715749</td>
<td>6,045,157</td>
<td>34.6798</td>
<td>97.87</td>
</tr>
<tr>
<td><img src="image4.png" alt="Structure" /></td>
<td>CDD-829859</td>
<td>7,638,272</td>
<td>34.514</td>
<td>100</td>
</tr>
<tr>
<td><img src="image5.png" alt="Structure" /></td>
<td>CDD-930673</td>
<td>5,768,432</td>
<td>32.8597</td>
<td>99.73</td>
</tr>
<tr>
<td><img src="image6.png" alt="Structure" /></td>
<td>CDD-932210</td>
<td>5,657,509</td>
<td>32.7089</td>
<td>93.07</td>
</tr>
<tr>
<td><img src="image7.png" alt="Structure" /></td>
<td>CDD-805965</td>
<td>7,264,531</td>
<td>32.6038</td>
<td>93.56</td>
</tr>
<tr>
<td><img src="image8.png" alt="Structure" /></td>
<td>CDD-856573</td>
<td>2,871,079</td>
<td>32.4961</td>
<td>90.55</td>
</tr>
<tr>
<td><img src="image9.png" alt="Structure" /></td>
<td>CDD-934423</td>
<td>7,271,481</td>
<td>32.4212</td>
<td>93.6</td>
</tr>
<tr>
<td><img src="image10.png" alt="Structure" /></td>
<td>CDD-640046</td>
<td>7,744,437</td>
<td>32.1454</td>
<td>98.55</td>
</tr>
</tbody>
</table>

Ekins et al., Mol BioSyst, In Press
### Simple descriptors

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MWT</th>
<th>logP</th>
<th>HBD</th>
<th>HBA</th>
<th>RO 5</th>
<th>Atom count</th>
<th>PSA</th>
<th>RBN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MLSMR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Active ≥ 90%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inhibition at 10uM</td>
<td>357.10</td>
<td>3.58</td>
<td>1.16</td>
<td>4.89</td>
<td>0.20</td>
<td>42.99</td>
<td>83.46</td>
<td>4.85</td>
</tr>
<tr>
<td>(N = 4096)</td>
<td>(84.70)</td>
<td>(1.39)</td>
<td>(0.93)</td>
<td>(1.94)</td>
<td>(0.48)</td>
<td>(12.70)</td>
<td>(34.31)</td>
<td>(2.43)</td>
</tr>
<tr>
<td><strong>Inactive &lt; 90%</strong></td>
<td>350.15</td>
<td>2.82</td>
<td>1.14</td>
<td>4.86</td>
<td>0.09</td>
<td>43.38</td>
<td>85.06</td>
<td>4.91</td>
</tr>
<tr>
<td>inhibition at 10uM</td>
<td>(77.98)**</td>
<td>(1.44)**</td>
<td>(0.88)</td>
<td>(1.77)</td>
<td>(0.31)**</td>
<td>(10.73)</td>
<td>(32.08)*</td>
<td>(2.35)</td>
</tr>
<tr>
<td>(N = 216367)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TAACF- NIAID CB2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Active ≥ 90%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inhibition at 10uM</td>
<td>349.58</td>
<td>4.04</td>
<td>0.98</td>
<td>4.18</td>
<td>0.19</td>
<td>41.88</td>
<td>70.28</td>
<td>4.76</td>
</tr>
<tr>
<td>(N = 1702)</td>
<td>(63.82)</td>
<td>(1.02)</td>
<td>(0.84)</td>
<td>(1.66)</td>
<td>(0.40)</td>
<td>(9.44)</td>
<td>(29.55)</td>
<td>(1.99)</td>
</tr>
<tr>
<td><strong>Inactive &lt; 90%</strong></td>
<td>352.59</td>
<td>3.38</td>
<td>1.11</td>
<td>4.24</td>
<td>0.12</td>
<td>42.43</td>
<td>77.75</td>
<td>4.72</td>
</tr>
<tr>
<td>inhibition at 10uM</td>
<td>(70.87)</td>
<td>(1.36)**</td>
<td>(0.82)**</td>
<td>(1.58)</td>
<td>(0.34)**</td>
<td>(8.94)*</td>
<td>(30.17)**</td>
<td>(1.99)</td>
</tr>
</tbody>
</table>
Pharmacophores for TB drugs

1. Select Molecules

<table>
<thead>
<tr>
<th>Molecule</th>
<th>$K_i$ (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU 486</td>
<td>0.24</td>
</tr>
<tr>
<td>RU 43044</td>
<td>0.1</td>
</tr>
<tr>
<td>1 Sa</td>
<td>10000</td>
</tr>
<tr>
<td>1 Sb</td>
<td>10000</td>
</tr>
<tr>
<td>1 Ra</td>
<td>10000</td>
</tr>
<tr>
<td>1 Rb</td>
<td>157</td>
</tr>
<tr>
<td>4</td>
<td>5.96</td>
</tr>
<tr>
<td>CP-394531</td>
<td>0.10</td>
</tr>
<tr>
<td>CP-409069</td>
<td>0.17</td>
</tr>
</tbody>
</table>

2. Alignment of inhibitors

3. Search databases

4. Hits retrieved

5. Map to pharmacophore

6. Order molecules & test in vitro

7. Rebuild
Pharmacophores for TB drugs

We generated common features pharmacophores (Discovery Studio, Accelrys) based on the five current first line Mtb agents (rifampicin, ethambutol, streptomycin)

Searched 3D databases of TB screening sets etc.. NIAID (3748), GVKBio (2880) and MLSMR dose response data (2273)

Could this help us suggest mechanism / targets for whole cell screening data?

Ekins et al., Mol BioSyst, 6: 840-851, 2010
Pharmacophores for TB drugs

<table>
<thead>
<tr>
<th>Pharmacophore</th>
<th>NIAID compounds retrieved (% of dataset)</th>
<th>NIAID actives retrieved MIC &lt; 5(\mu)M (% of total actives)</th>
<th>GVKbio compounds retrieved (% of dataset)</th>
<th>GVKbio actives retrieved MIC &lt; 5(\mu)M (% of total actives)</th>
<th>MLSMR dose response data compounds retrieved (% of dataset)</th>
<th>MLSMR dose response actives retrieved IC(_{50}) &lt; 5(\mu)M (% of total actives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rifampicin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rifampicin (no shape)</td>
<td>3 (0.08)</td>
<td>3 (0.16)</td>
<td>1 (0.03)</td>
<td>1 (0.26)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>1 (0.03)</td>
<td>1 * (0.05)</td>
<td>1 (0.03)</td>
<td>0</td>
<td>1 (0.04)</td>
<td>1 * (0.13)</td>
</tr>
<tr>
<td>Ethambutol (no shape)</td>
<td>19 (0.51)</td>
<td>13 (0.69)</td>
<td>8 (0.28)</td>
<td>0</td>
<td>2 (0.09)</td>
<td>2 (0.27)*</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>10 (0.27)</td>
<td>9 * (0.48)</td>
<td>4 (0.14)</td>
<td>1 * (0.26)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

* Recovered compound used in pharmacophore

Ekins et al., Mol BioSyst, 6: 840-851, 2010
Pharmacophores for TB drugs

We created Catalyst databases with the Kyoto Encyclopedia of Genes and Genomes (KEGG, N = 11,286), Human metabolome database (N = 2,899) and LipidMaps (N = 10,220).

<table>
<thead>
<tr>
<th>Pharmacophore</th>
<th>KEGG hits</th>
<th>LipidMaps database hits</th>
<th>HMDB hits</th>
<th>US Drugs Microsource database hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rifampicin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rifampicin (no shape)</td>
<td>61</td>
<td>1</td>
<td>14</td>
<td>5 (5 antibacterials)</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>1 *</td>
<td>0</td>
<td>0</td>
<td>1 *</td>
</tr>
<tr>
<td>Ethambutol (no shape)</td>
<td>9 *</td>
<td>2</td>
<td>0</td>
<td>5 * (2 antibacterials)</td>
</tr>
<tr>
<td>Streptomycin</td>
<td>203*</td>
<td>247</td>
<td>86</td>
<td>11 * (4 antibacterials)</td>
</tr>
</tbody>
</table>

* Recovered compound used in pharmacophore

See poster on TB molecular mimics
Freundlich et al., Pub # 258 Tues Aug 24 – 5.30-7.30pm Hall C
Ekins et al., Mol BioSyst, 6: 840-851, 2010
Large datasets in Pharma

How could we make ADME or Tox models from pharmas available for neglected disease researchers?

Need open technologies so models can be shared

What can be developed with very large training and test sets?

HLM training 50,000 testing 25,000 molecules
  training 194,000 and testing 39,000
MDCK training 25,000 testing 25,000
MDR training 25,000 testing 18,400

Open molecular descriptors – Pfizer work

(Gupta et al., Pub # 405, Wed 7-9pm Ballroom)

Rishi R. Gupta, Eric M. Gifford, Ted Liston, Chris L. Waller, Moses Hohman, Barry A. Bunin and Sean Ekins, Drug Metab Dispos, In Press 2010
<table>
<thead>
<tr>
<th>CDK descriptors</th>
<th>SVM</th>
<th>RP Forest Uni Class</th>
<th>RP Forest</th>
<th>C5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kappa = 0.14</strong></td>
<td>Kappa = 0.16</td>
<td>Kappa = 0.11</td>
<td>Kappa = 0.39</td>
<td></td>
</tr>
<tr>
<td><strong>Sensitivity = 0.11</strong></td>
<td><strong>Sensitivity = 0.54</strong></td>
<td><strong>Sensitivity = 0.85</strong></td>
<td><strong>Sensitivity = 0.54</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Specificity = 0.96</strong></td>
<td><strong>Specificity = 0.70</strong></td>
<td><strong>Specificity = 0.33</strong></td>
<td><strong>Specificity = 0.91</strong></td>
<td></td>
</tr>
<tr>
<td><strong>PPV = 0.43</strong></td>
<td><strong>PPV = 0.33</strong></td>
<td><strong>PPV = 0.25</strong></td>
<td><strong>PPV = 0.61</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOE2D and SMARTS Keys</th>
<th>SVM</th>
<th>RP Forest Uni Class</th>
<th>RP Forest</th>
<th>C5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not evaluated</td>
<td>Not evaluated</td>
<td>Not evaluated</td>
<td>Not evaluated</td>
<td><strong>Kappa = 0.43</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Sensitivity = 0.58</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Specificity = 0.91</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>PPV = 0.63</strong> (Baseline)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CDK and SMARTS Keys</th>
<th>SVM</th>
<th>RP Forest Uni Class</th>
<th>RP Forest</th>
<th>C5.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not evaluated</td>
<td>Not evaluated</td>
<td>Not evaluated</td>
<td>Not evaluated</td>
<td><strong>Kappa = 0.43</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Sensitivity = 0.58</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Specificity = 0.91</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>PPV = 0.63</strong></td>
</tr>
</tbody>
</table>

(Gupta et al., Pub # 405, Wed 7-9pm Ballroom)
Massive Human liver microsomal stability model

<table>
<thead>
<tr>
<th>HLM Model with CDK and SMARTS Keys:</th>
<th>HLM Model with MOE2D and SMARTS Keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>• # Descriptors: 578 Descriptors</td>
<td>• # Descriptors: 818 Descriptors</td>
</tr>
<tr>
<td>• # Training Set compounds:</td>
<td>• # Training Set compounds:</td>
</tr>
<tr>
<td>193,650</td>
<td>193,930</td>
</tr>
<tr>
<td>• Cross Validation Results:</td>
<td>• Cross Validation Results:</td>
</tr>
<tr>
<td>38,730 compounds</td>
<td>38,786 compounds</td>
</tr>
<tr>
<td>• Training $R^2$: 0.79</td>
<td>• Training $R^2$: 0.77</td>
</tr>
<tr>
<td>• 20% Test Set $R^2$: 0.69</td>
<td>• 20% Test Set $R^2$: 0.69</td>
</tr>
<tr>
<td><strong>Blind Data Set (2310 compounds):</strong></td>
<td><strong>Blind Data Set (2310 compounds):</strong></td>
</tr>
<tr>
<td>• $R^2$ = 0.53</td>
<td>• $R^2$ = 0.53</td>
</tr>
<tr>
<td>• RMSE = 0.367</td>
<td>• RMSE = 0.367</td>
</tr>
<tr>
<td><strong>Continuous $\rightarrow$ Categorical:</strong></td>
<td><strong>Continuous $\rightarrow$ Categorical:</strong></td>
</tr>
<tr>
<td>• $\kappa$ = 0.40</td>
<td>• $\kappa$ = 0.42</td>
</tr>
<tr>
<td>• Sensitivity = 0.16</td>
<td>• Sensitivity = 0.24</td>
</tr>
<tr>
<td>• Specificity = 0.99</td>
<td>• Specificity = 0.987</td>
</tr>
<tr>
<td>• PPV = 0.80</td>
<td>• PPV = 0.823</td>
</tr>
<tr>
<td><strong>Time (sec/compound): 0.252</strong></td>
<td><strong>Time (sec/compound): 0.303</strong></td>
</tr>
</tbody>
</table>

PCA of training (red) and test (blue) compounds

Overlap in Chemistry space

(Gupta et al., Pub # 405, Wed 7-9pm Ballroom)
Open descriptors results almost identical to commercial descriptors
Across many datasets and quantitative and qualitative data
Provides confidence that open models could be viable

Rishi R. Gupta, Eric M. Gifford, Ted Liston, Chris L. Waller, Moses Hohman, Barry A. Bunin and Sean Ekins, Drug Metab Dispos, In Press 2010
The next opportunity for crowdsourcing...

- Open source software for molecular descriptors and algorithms
- Spend only a fraction of the money on QSAR
- Selectively share your models with collaborators and control access
- Have someone else host the models / predictions

Current investments

>$1M/yr

>$10-100’s M/yr
Acknowledgments

Antony Williams (RSC)
Joel Freundlich, (Texas A&M)
Gyanu Lamichhane, Bill Bishai (Johns Hopkins)
Jeremy Yang (UNM)
Nicko Goncharoff (SureChem)
Chris Lipinski
Takushi Kaneko (TB Alliance)
Bob Reynolds (SRI)
Chris Waller, Eric Gifford, Ted Liston, Rishi Gupta (Pfizer)
Carolyn Talcott and Malabika Sarker (SRI International)

Accelrys, ChemAxon
Bill and Melinda Gates Foundation

ekinssean@yahoo.com ;
sekins@collaborativemed.com


CDD is Secure & Simple

- Web based database (log in securely into your account from any computer using any common browser – Firefox, IE, Safari)
- Hosted on remote server (lower cost) dual-Xeon, 4GB RAM server with a RAID-5 SCSI hard drive array with one online spare
- Highly secure, all traffic encrypted, server in a secure professionally hosted environment
- Automatically backed up nightly
- MySQL database
- Uses JChemBase software with Rails via a Ruby-Java bridge, (structure searching and inserting/ modifying structures)
- Marvin applet for structure editing
- Export all data to Excel with SMILES, SDF, SAR, & png images

www.collaborativedrug.com