

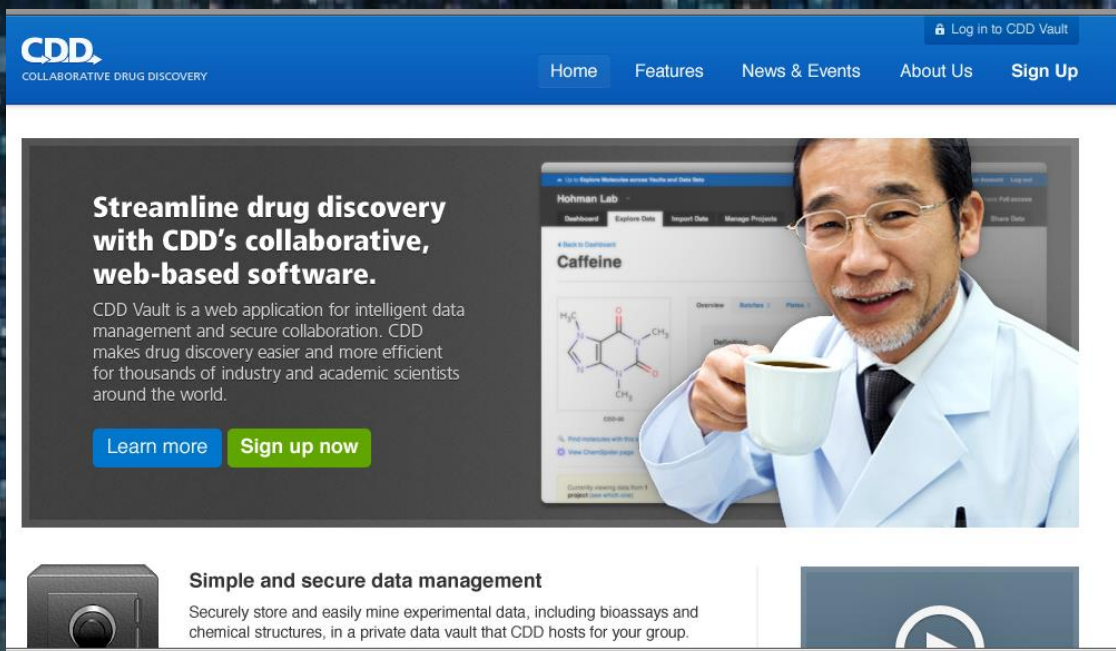
Exploiting Bigger Data and Collaborative Tools for Predictive Drug Discovery

**Sean Ekins, CSO
CDD**

This guy deserves a coffee

CDD website 2010-2013

Perkin Elmer in
Laboratory Informatics Guide 2014



CDD
COLLABORATIVE DRUG DISCOVERY

Log in to CDD Vault

Home Features News & Events About Us Sign Up

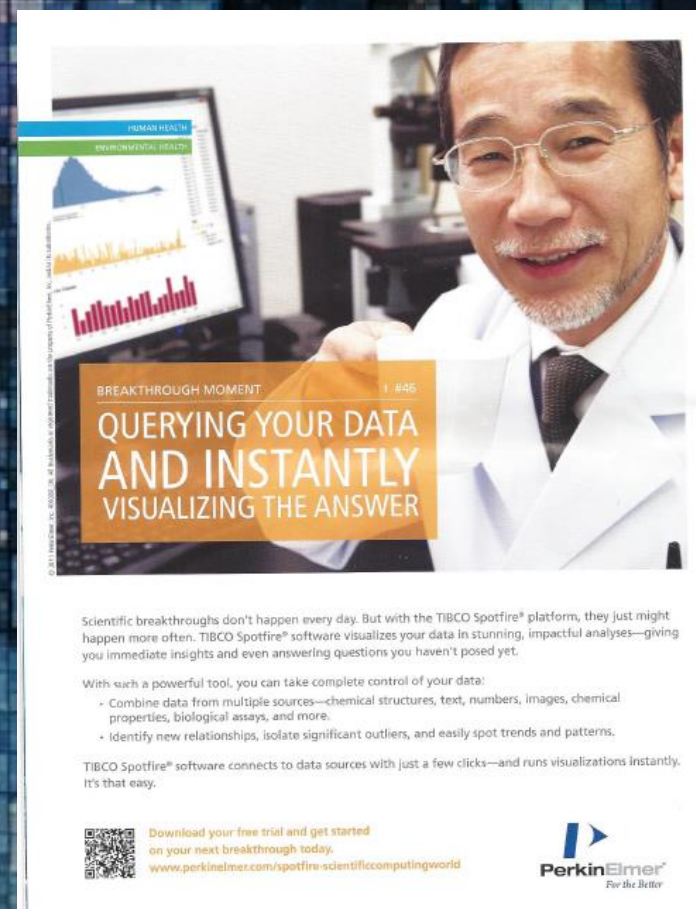
Streamline drug discovery with CDD's collaborative, web-based software.

CDD Vault is a web application for intelligent data management and secure collaboration. CDD makes drug discovery easier and more efficient for thousands of industry and academic scientists around the world.

[Learn more](#) [Sign up now](#)

Simple and secure data management

Securely store and easily mine experimental data, including bioassays and chemical structures, in a private data vault that CDD hosts for your group.



BREAKTHROUGH MOMENT | #45

QUERYING YOUR DATA AND INSTANTLY VISUALIZING THE ANSWER

Scientific breakthroughs don't happen every day. But with the TIBCO Spotfire® platform, they just might happen more often. TIBCO Spotfire® software visualizes your data in stunning, impactful analyses—giving you immediate insights and even answering questions you haven't posed yet.

With such a powerful tool, you can take complete control of your data:

- Combine data from multiple sources—chemical structures, text, numbers, images, chemical properties, biological assays, and more.
- Identify new relationships, isolate significant outliers, and easily spot trends and patterns.

TIBCO Spotfire® software connects to data sources with just a few clicks—and runs visualizations instantly. It's that easy.

Download your free trial and get started on your next breakthrough today.
www.perkinelmer.com/spotfire-scientificcomputingworld

PerkinElmer
For the Better

Cloud Services for the Scientific Community

developerWorks > Technical topics > Cloud computing > Technical library >

Data science in the cloud

Investment analysis with IPython and pandas

Domino: Data Analysis, Accelerated

Easily run R, Python, and Matlab code in the cloud.

Automatic version control and collaboration for data, code, and results.

Sense

Plans & Pricing Sense Enterprise Research Computing About Us

A Collaborative Cloud Platform for Data Science and Big Data Analytics

Collaborate on, scale, and deploy data analysis and advanced analytics projects 10x faster. Use the most powerful tools — R, Python, JavaScript, Redshift, Hive, Impala, Hadoop, and more — scaled and integrated in the cloud.

Cloud

NIMBIX
Accelerated Compute Cloud™

Call Us: +1-866-307-0819

LAUNCH HPC TASK

SIGN UP

CLOUD SUPERCOMPUTING HPC MANAGED SERVICES APPLICATIONS BLOG ABOUT NIMBIX

Next-Gen Bioinformatics

Accelerate sequence analysis pipelines to crunch data faster. Align, map and search genome data in minutes versus hours, all without having to deploy any new equipment.

Learn More >



The Cost of Doing Science on the Cloud: A Real Example

Ewa Deelman¹, Gurmeet Singh¹, Miriam S. H. ... John Good⁴

¹USC Information Sciences Institute, Marina del Rey, CA

²University of Wisconsin Madison, WI

³Infrared Processing and Analysis Center & Michelson Science Center, California Institute of Technology, Pasadena, CA

⁴Infrared Processing and Analysis Center, California Institute of Technology, Pasadena, CA

“by provisioning the right amount of storage and compute resources, cost can be significantly reduced with no significant impact on application performance”

CDD's Influence spreads beyond the cloud

2004 - present

CDD
COLLABORATIVE DRUG DISCOVERY

CDD
COLLABORATIVE DRUG DISCOVERY

Home CDD Vault Community Blog About Us

CDD Vault®

Finally, a modern approach to drug research informatics

CDD Vault is a hosted biological and chemical database that securely manages your private and external data. Accomplish more with an intuitive solution designed by scientists.

A SYSTEM YOUR ENTIRE TEAM CAN ACTUALLY USE

- ✓ Intuitive web interface
- ✓ Economical cloud deployment
- ✓ Biologists and chemists interacting

BEATS JUGGLING SPREADSHEETS OF SCREENING DATA

- ✓ Eliminates the risk of data loss
- ✓ Unified data yields better results
- ✓ Easier to find, analyze, and share data

2014

accelrys®

ScienceCloud by Accelrys Transforms Externalized Drug Discovery

New cloud solution for externalized life science research and development redefines collaboration and creativity

San Diego, CA, Feb. 6, 2014 / PRNewswire / — Accelrys, Inc. (NASDAQ: ACCL), a leading provider of scientific innovation lifecycle management software.

ScienceCloud

Username Password SIGN IN

Forgot Password?

SOLUTIONS CUSTOMERS SUPPORT ABOUT

Transform Scientific Collaboration

Transform the way you run your externalized R&D project: multiple collaborators, single platform, one conversation.

Empower your virtual teams with a new generation of integrated, cloud-based applications built on a world-class scientific platform.



- Global Economics
- Companies & Industries
- Politics & Policy
- Technology
- Markets & Finance
- Innovation & Design
- Lifestyle

Analytics **5** tips to manage, maintain and mobilize your Big Data.

How Big Data Helped Cut Emergency Room Visits by 10 Percent

By Karen Weise March 25, 2014

SCIENTIFIC AMERICAN

Sign In Register

Search ScientificAmerican.com

- Subscribe
- News & Features
- Topics
- Blogs
- Audio & Podcasts
- Education

More Science » Scientific American

Big Data and the Transformation of Society

The forces we learn each day reveal about us than we know. It could be a nightmare—or it could be the foundation of a healthier, more prosperous world.



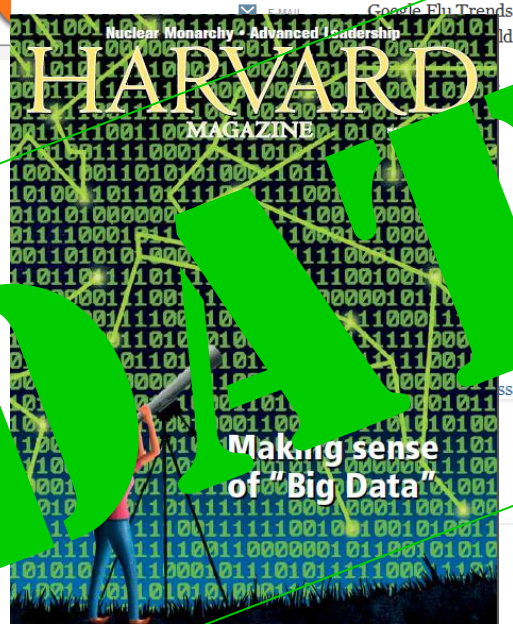
- NEWS
- JOURNALS & MORE
- MEMBERSHIP
- CAREERS
- PROGRAMS
- GIVING
- ABOUT
- SEARCH

View More Events

1 APR

Big Data, Life Sciences, and National Security

- Like
- Tweet
- reddit this
- Email
- Print



BIG DATA

Google Flu Trends: The Limits of Big Data

By STEVE LOHR MARCH 28, 2014, 7:00 AM



...not only wildly overestimated the flu season in the United States in the 2012-13 flu season — but has also consistently overshot in the last

Making sense of "Big Data"

CONTEXT NUMBERS

Why Big Data is bad for science

UC BERKELEY News Center

- Latest News
- Categories »
- Events
- Sports
- Multimedia »
- Media Relations »

New data science institute to help scholars harness 'big data'

By Robert Sanders, Media Relations | November 13, 2013

BERKELEY — In a world awash in data, UC Berkeley is meeting the flood head-on by establishing a new institute to support faculty, researchers and students in their efforts to mine this information in areas as diverse as astronomy and economics, genetics and demography.

C

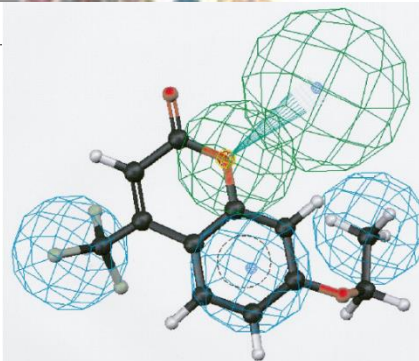
LXR

Small Data circa late 1990's

VDR

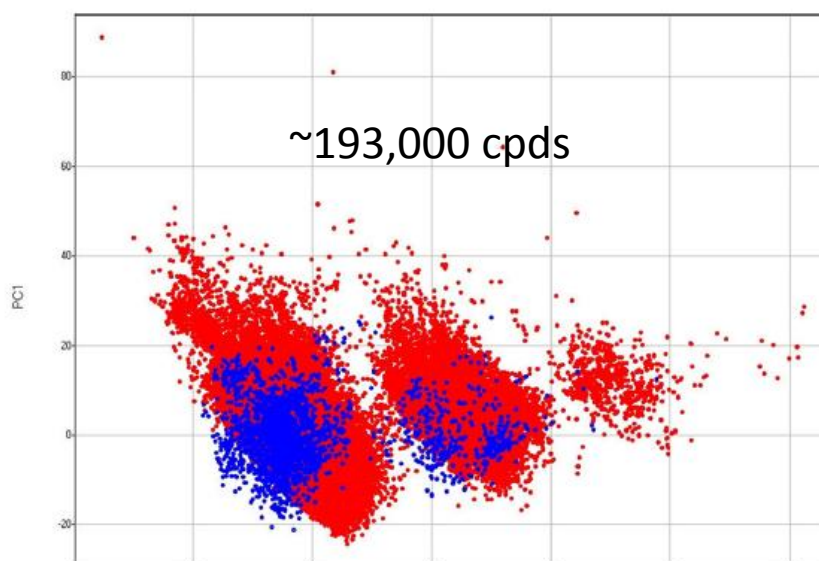
 $K_{m(\text{apparent})}$ for metabolism of substrates for expressed CYP2B6

Substrate	Metabolic Pathway	$K_{m(\text{apparent})}$
Antipyrine	4-Hydroxylation	17.7 mM
Benzphetamine	N-Demethylation	93.4 μM
Benzoyloxyresorufin	N-Demethylation	1.28 μM
Cinnarizine	p-Hydroxylation	17.2 μM
4-Chloromethyl-7-ethoxycoumarin	O-Deethylation	33.7 μM
3-Cyano-7-ethoxycoumarin	O-Deethylation	71.3 μM
Dextromethorphan	N-Demethylation	350 μM
Diazepam	N-Demethylation	113 μM
1,2-dibromoethane	2-Bromoacetaldehyde formation	9.7 mM
7-ethoxycoumarin	O-Deethylation	115 μM
7-Ethoxy-4-trifluoromethylcoumarin	O-Deethylation	1.7 μM
Imipramine	N-Demethylation	383 μM
Midazolam	1'-Hydroxylation	46.1 μM
RP73401	Ring hydroxylation	22.5 μM
S-Mephenytoin	N-Demethylation	564 μM
Testosterone	16 β -Hydroxylation	50.5 μM



HNF-4a

Big Data circa 2010's



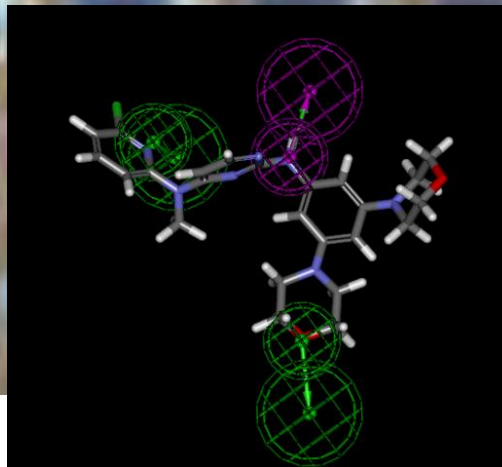
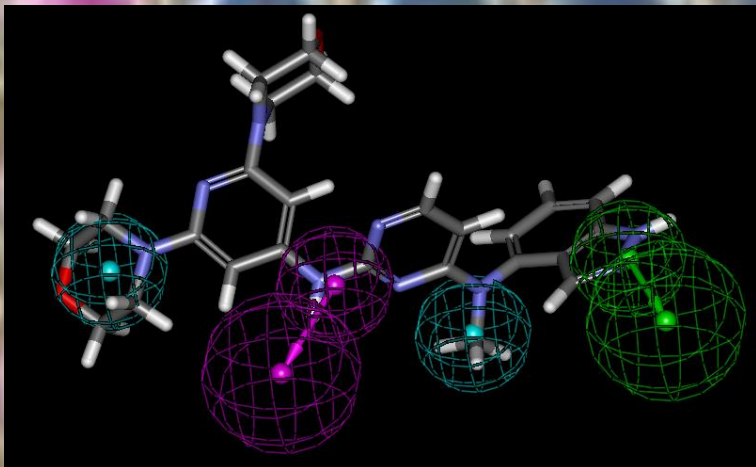
Drug Metab Dispos, 38: 2083-2090, 2010

CAR

ESR

PXR

How you dispense liquids may be important: insights from small data



Generated with Discovery Studio (Accelrys)

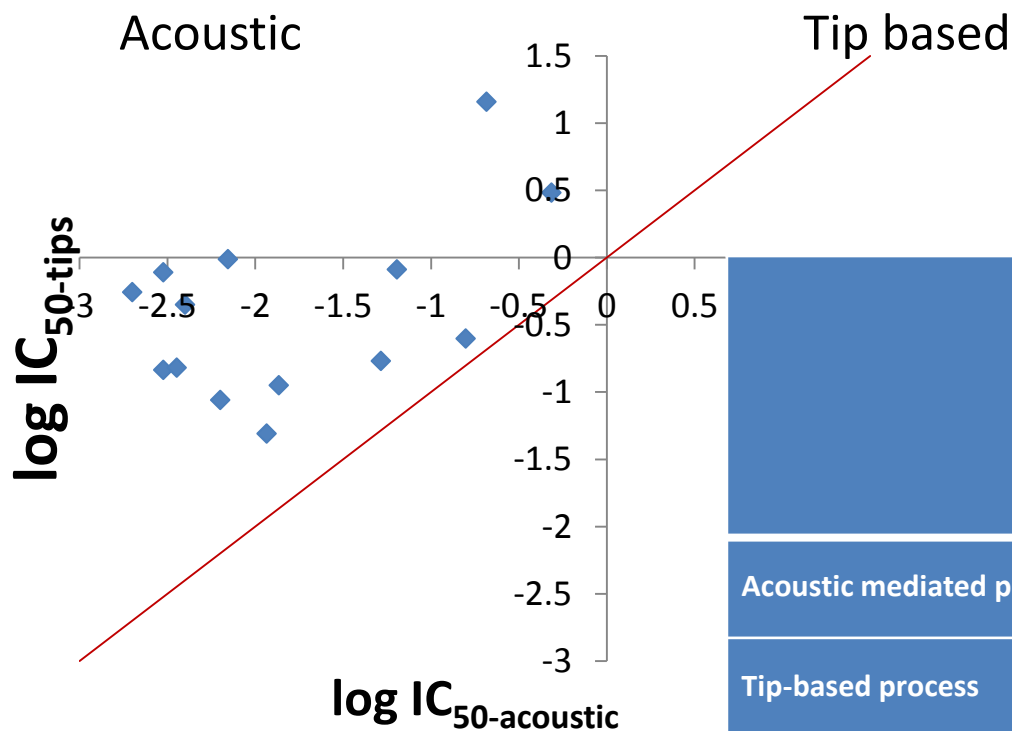
Cyan = hydrophobic

Green = hydrogen bond acceptor

Purple = hydrogen bond donor

Each model shows most potent molecule mapping

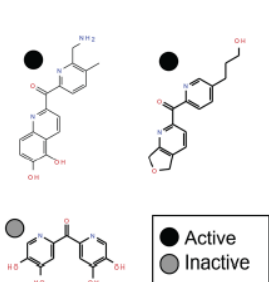
PLoS ONE 8(5): e62325 (2013)



	Hydrophobic features (HPF)	Hydrogen bond acceptor (HBA)	Hydrogen bond donor (HBD)	Observed vs. predicted IC_{50} r
Acoustic mediated process	2	1	1	0.92
Tip-based process	0	2	1	0.80

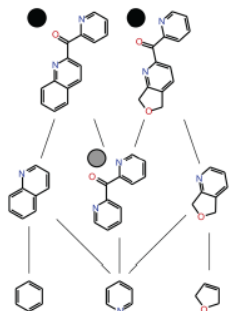
Future: sharing chemical relationships without structures

Private Data

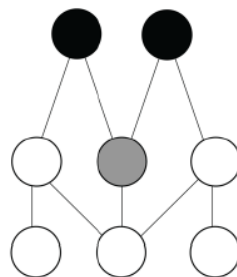


Structures
Activity Data

Shared Data



Structures
Relationships
Activity Data

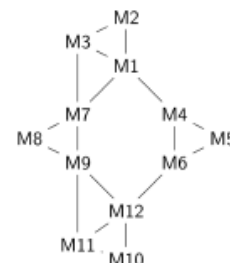


Relationships
Activity Data

Molecule IDs

M2
M3
M1
M7
M4
M8
M9
M6
M5
M12
M11
M10

Neighbors



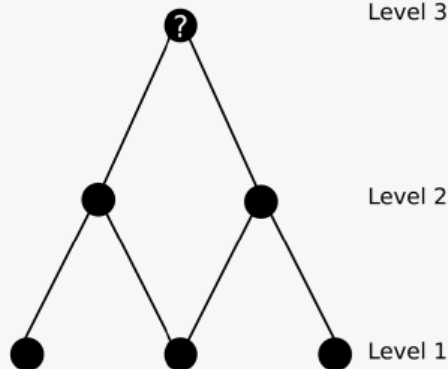
Scaffold Groups



statistical corrections
marginal cost of discovery
workflow inference

local hit rate
nearest neighbor

compound set enrichment
ontology pattern identification
clique-oriented prioritization
diversity-oriented prioritization

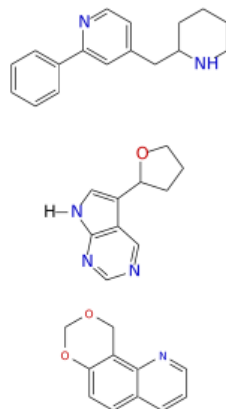


Anonymized
Network

Level 3

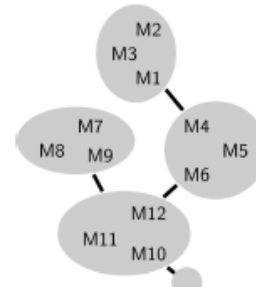
Level 2

Level 1



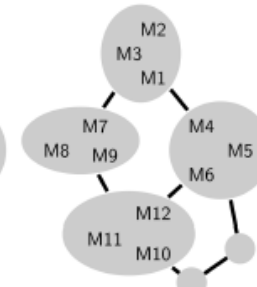
Candidate
Scaffolds

Scaffold Tree



tree visualization
imputed trees

Scaffold Network



network visualization
imputed networks

R-Group Network



matched-pair analysis
R-group analysis
structure-activity relationships

Drug Discovery Archeology

- Still a heavy emphasis on “testing” “doing “ rather than ‘learning’
- Mining data and historic data will increase in value
- Data becomes a repurposing opportunity
- How do we position databases for this?
- What about neglected diseases?

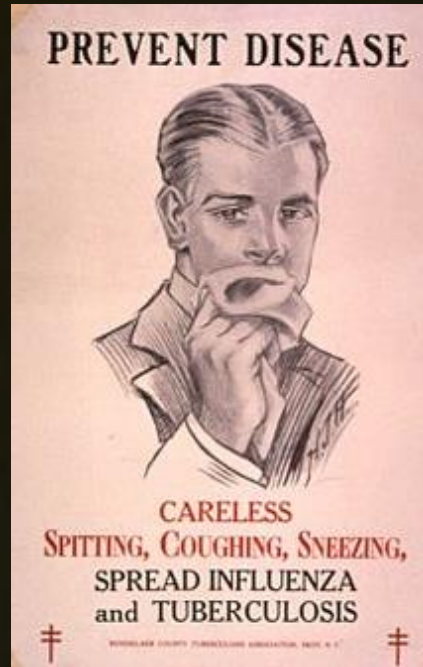


WHAT YOU SHOULD KNOW ABOUT

Multi drug resistance in
4.3% of cases

Extensively drug resistant
increasing incidence

one new drug
(bedaquiline) in 40 yrs



streptomycin (1943)

para-aminosalicylic acid (1949)

isoniazid (1952)

pyrazinamide (1954)

cycloserine (1955)

ethambutol (1962)

rifampicin (1967)

TUBERCULOSIS

Tuberculosis kills 1.6-1.7m/yr (~1 every
8 seconds)

1/3rd of worlds population infected!!!!



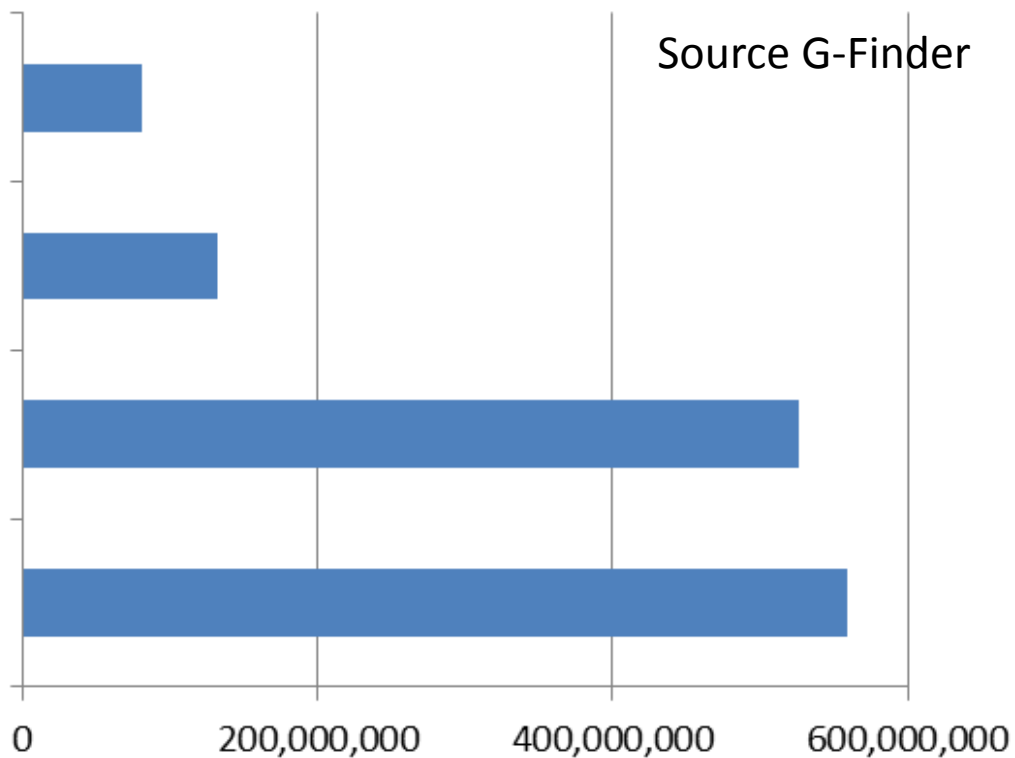
Helminths (worms and flukes)

Kinetoplastids

Tuberculosis

Malaria

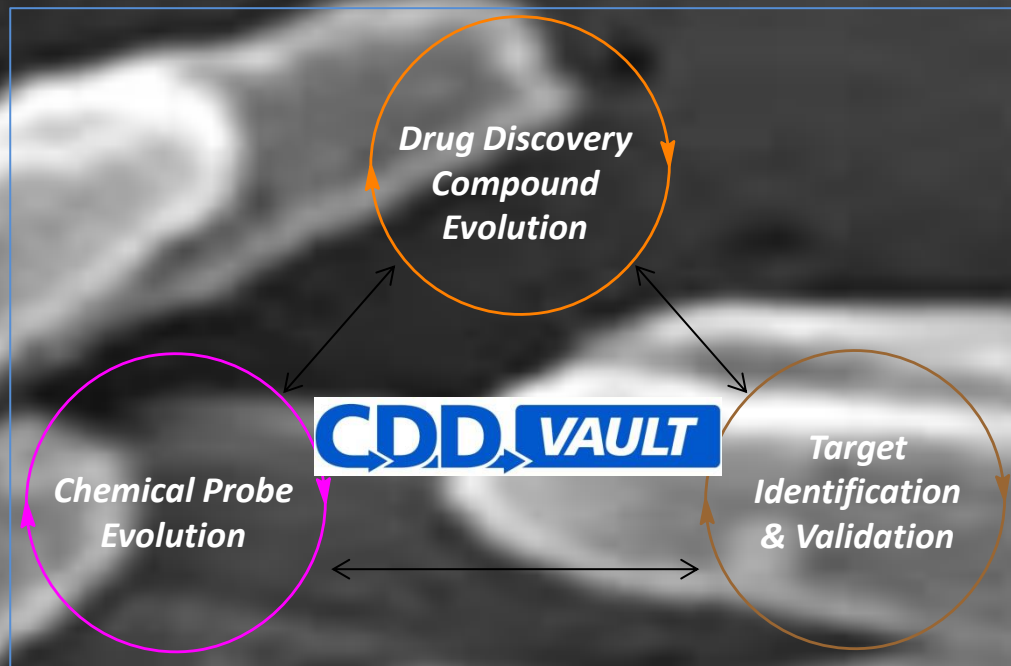
Source G-Finder



Global funding of innovation for neglected diseases (\$)

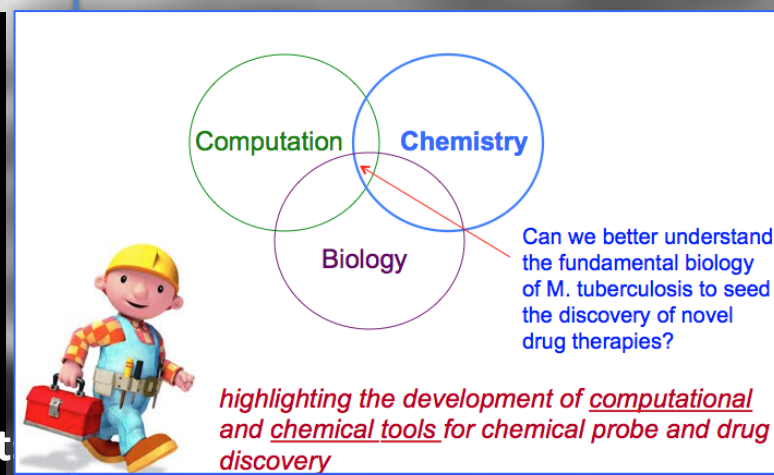
Ponder et al., Pharm Res 31: 271-277, 2014

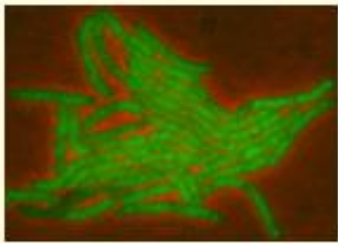
Freundlich Laboratory Collaborations Rely on CDD for Data Tracking!



- Three collaborations within Rutgers–NJMS
- Collaboration with Johns Hopkins, SRI, and CDD
- Collaboration with Johns Hopkins
- Collaboration with CDD

Supported by
7 Active NIH Grants



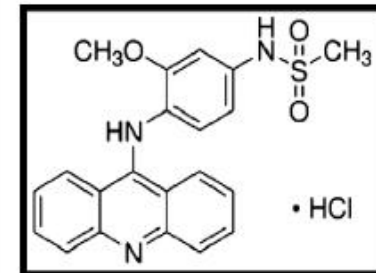
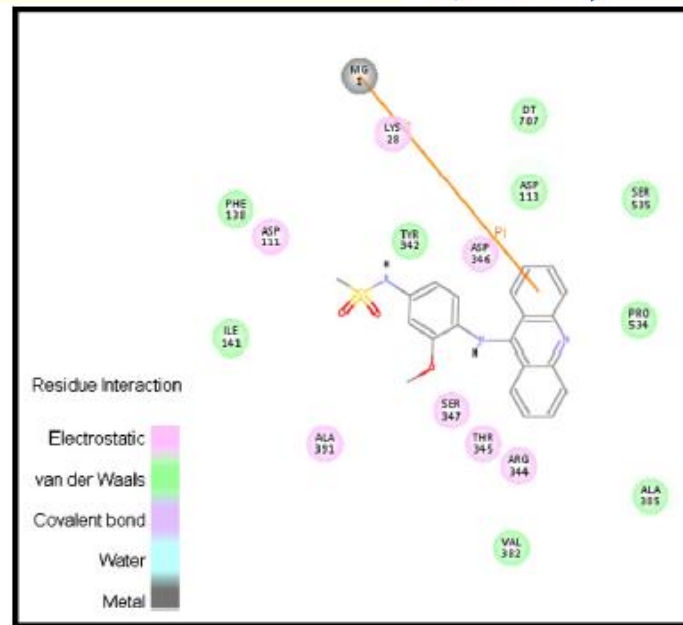
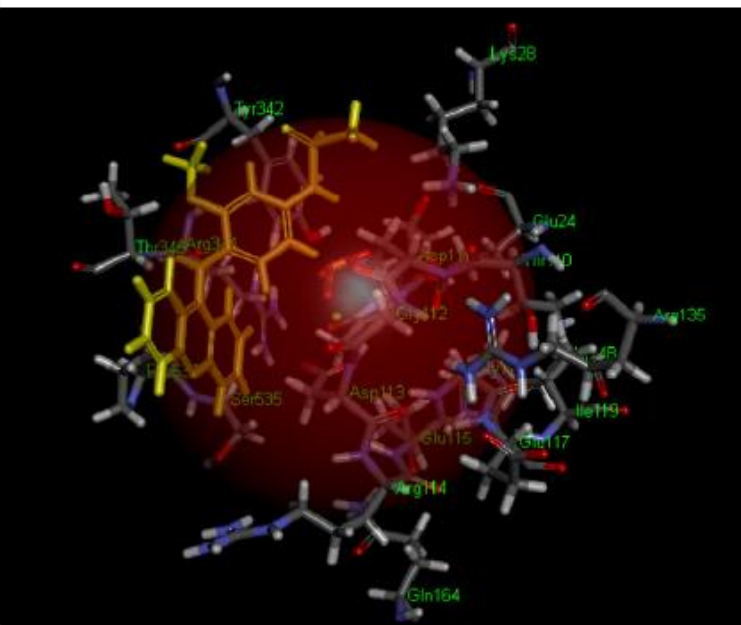


MM4TB



More Medicines for Tuberculosis

24 groups in this project use a single **CDD VAULT**



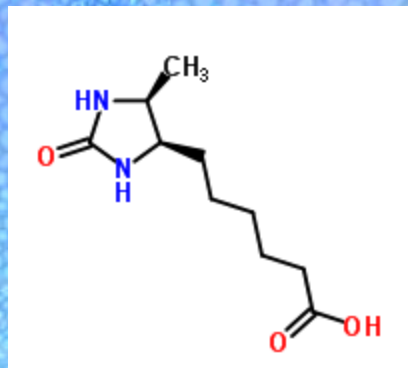
Amsacrine

Godbole et al., Biochem Biophys Res Comm 2014, in press

More Medicines for Tuberculosis (MM4TB)

Fishing: Example of mimic strategy for bioB Rv1589

Biotin biosynthesis

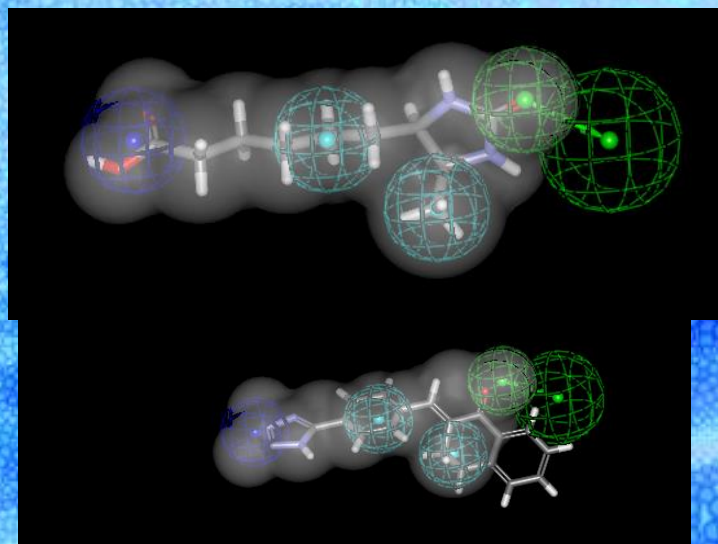


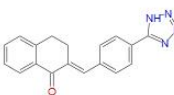
dethiobiotin

Take substrate and generate 3D conformers and build a pharmacophore

Use the pharmacophore to search vendor libraries in 3D

Buy and test compounds



Structure	Index	Cat_No	FitValue	Name
	3	JFD00142SC	3.1892	Compound Numb...

Pharmacophore

Searching Maybridge (57K) gives 72 molecules – many of them hydrophobic so they stand a chance of in vitro activity

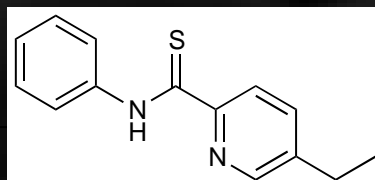
- Sarker et al., Pharm Res 2012, 29:2115-27

Over 5 years analyzed *in vitro* data and built models

High-throughput phenotypic *Mtb* screening



Mtb screening molecule database/s

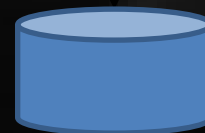


Bayesian Machine Learning classification *Mtb* Model

Descriptors + Bioactivity (+Cytotoxicity)

$$p(h | d) = \frac{P(d | h)P(h)}{P(d)}$$

Molecule Database (e.g. GSK malaria actives) virtually scored using Bayesian Models

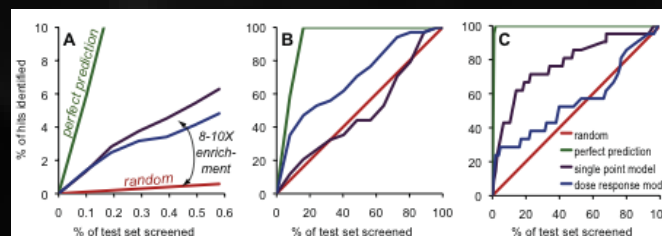


Top scoring molecules assayed for *Mtb* growth inhibition



New bioactivity data may enhance models

Identify *in vitro* hits and test models



- Ekins et al., Pharm Res 31: 414-435, 2014
- Ekins, et al., Tuberculosis 94; 162-169, 2014
- Ekins, et al., PLOS ONE 8; e63240, 2013
- Ekins, et al., Chem Biol 20: 370-378, 2013
- Ekins, et al., JCI, 53: 3054-3063, 2013
- Ekins and Freundlich, Pharm Res, 28, 1859-1869, 2011
- Ekins et al., Mol BioSyst, 6: 840-851, 2010
- Ekins, et al., Mol. Biosyst. 6, 2316-2324, 2010,

3 x published prospective tests >20% hit rate
Multiple retrospective tests 3-10 fold enrichment

A summary of some of the numbers involved – filtering for hits.

>250,000 molecules screened through Bayesian models

~750 molecules were tested *in vitro*

198 actives were identified

>20 % hit rate

Identified several novel potent hit series with good cytotoxicity & selectivity Identified known human kinase inhibitors and FDA approved drugs as hits

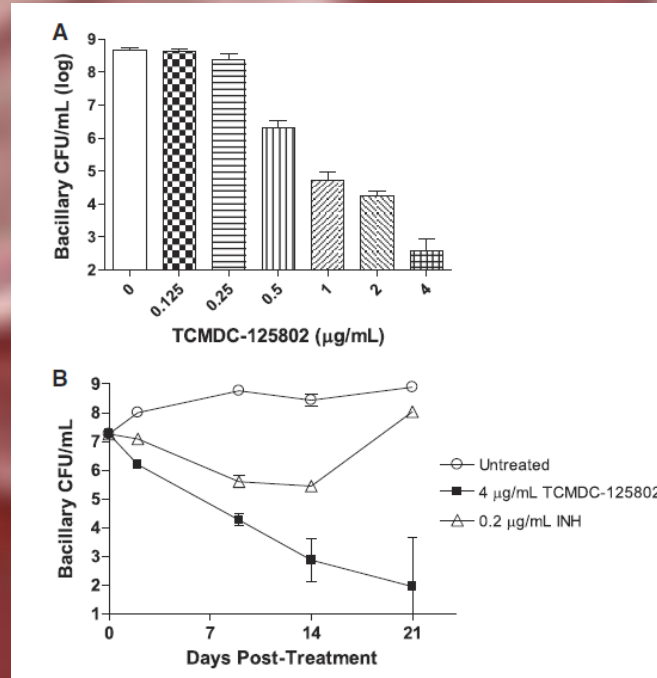
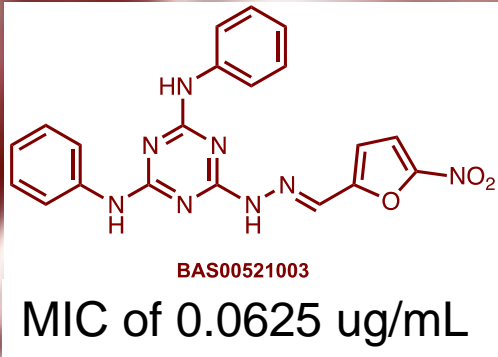
Ekins et al., PLOS ONE 2013 May 7;8(5):e63240;

Ekins et al., Chem Biol 20, 370–378, 2013

Ekins et al., Tuberculosis 94: 162-169

Taking a compound *in vivo* identifies issues

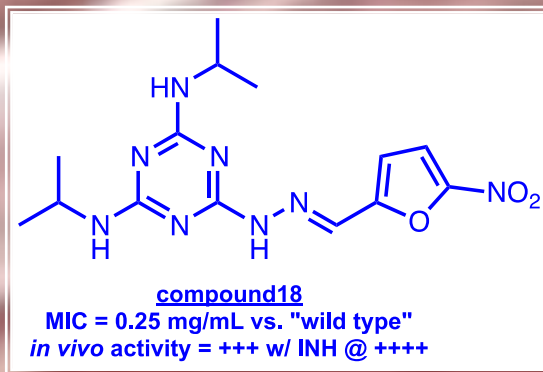
- BAS00521003/ TCMDC-125802 reported to be a *P. falciparum* lactate dehydrogenase inhibitor
- Only one report of antitubercular activity from 1969
 - solid agar MIC = 1 µg/mL (“wild strain”)
 - “no activity” in mouse model up to 400 mg/kg
 - *however, activity was solely judged by extension of survival!*

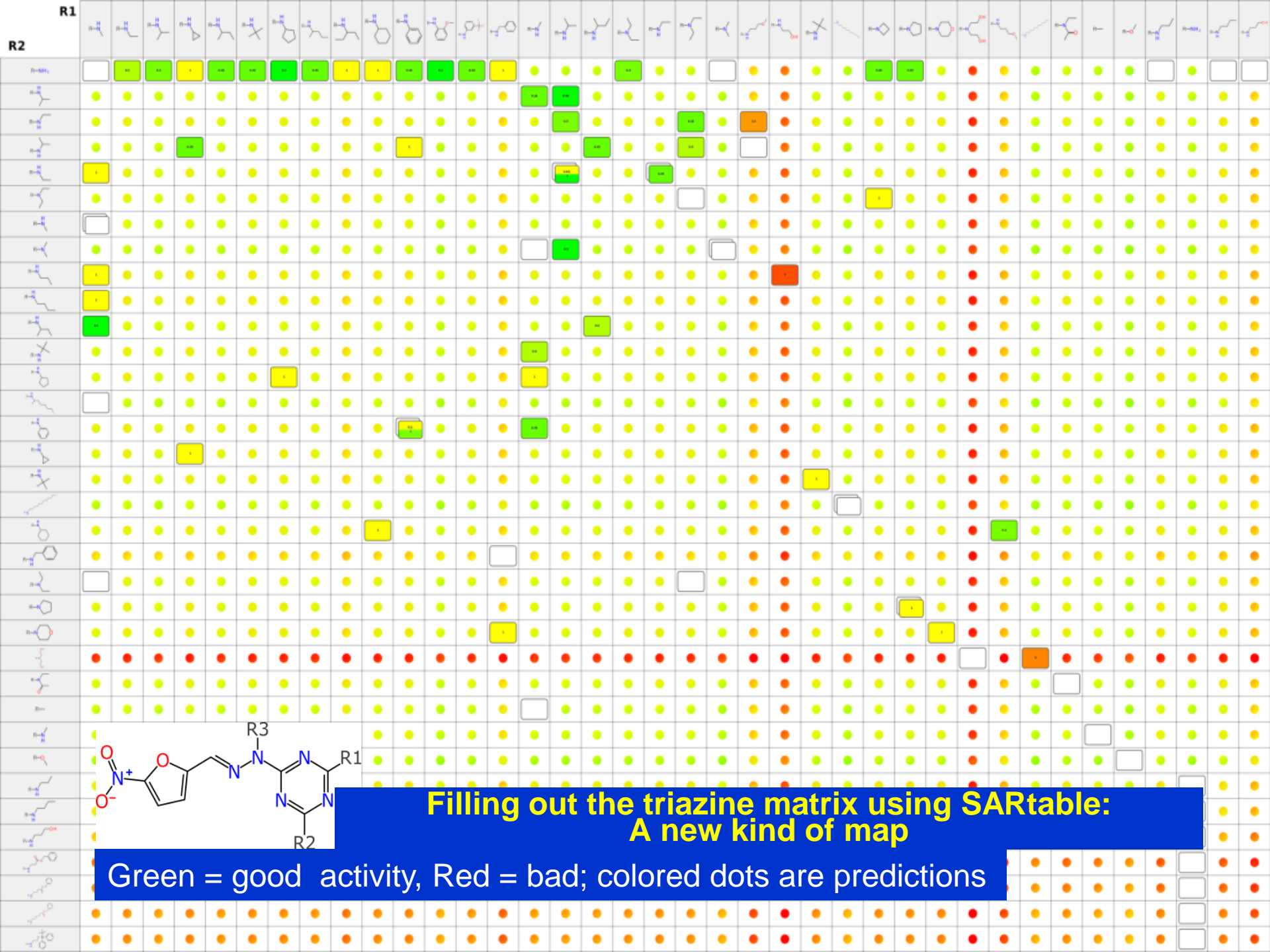


- 64X MIC affords 6 logs of kill
 - Resistance and/or drug instability beyond 14 d
- Vero cells : $CC_{50} = 4.0$ µg/mL

Selectivity Index SI = $CC_{50}/MIC_{Mtb} = 16 - 64$

In mouse no toxicity but also no efficacy in GKO model – probably metabolized.





Big Data: Screening for New Tuberculosis Treatments

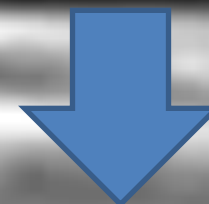
NIH National Institute of Allergy and Infectious Diseases
Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases.

**SOUTHERN RESEARCH
INSTITUTE**



NOVARTIS

**BROAD
INSTITUTE**



Tested >350,000 molecules
>1500 active and non toxic

Tested ~2M
Published 177

2M
100s

>300,000
800



Others have likely screened another 500,000

How many will become a new drug?
How do we learn from this big data?

Hunting High and Low for new molecules to test

We need to search sources..

From the Oceans...

To the ground

To the trees

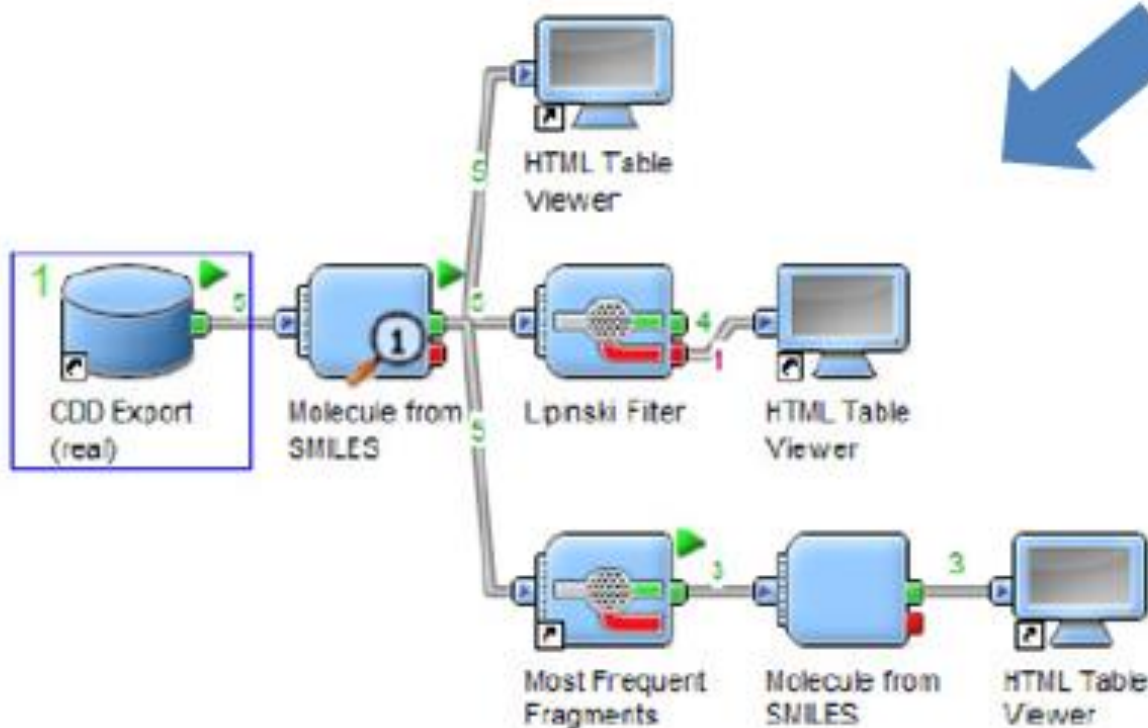
To the air..

And do it virtually

Find new libraries
to screen virtually
and test

Take everything out of CDD public

- Run through TB Bayesian models
- Score
- Test



What is the next bottleneck?



Five-fold increase in the publication of TB mouse model studies from 1997 to 2009
Franco, *PLoS One* 7, e47723 (2012).

**Billions of \$ of your money spent on TB
but no database of mouse *in vivo* data !**



Hunting for the *in vivo* data



It's out there.. be patient

Building the mouse TB database

Manually curated, structures sketched Mobile Molecular DataSheet (MMDS) iOS app or ChemDraw (Perkin Elmer)

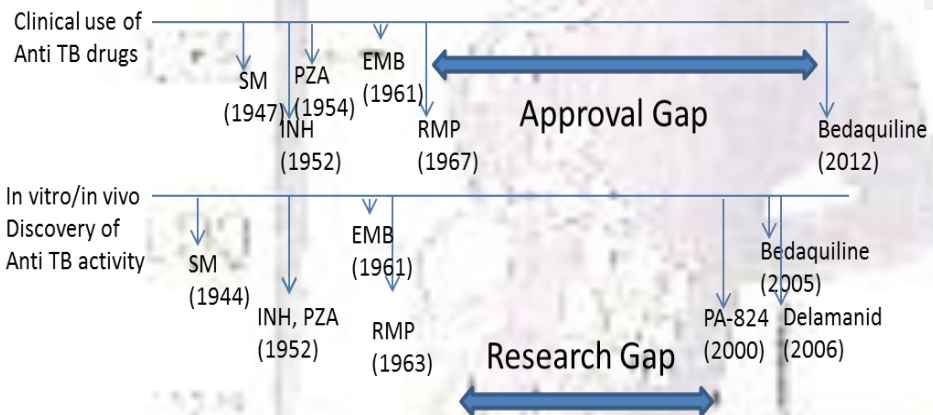
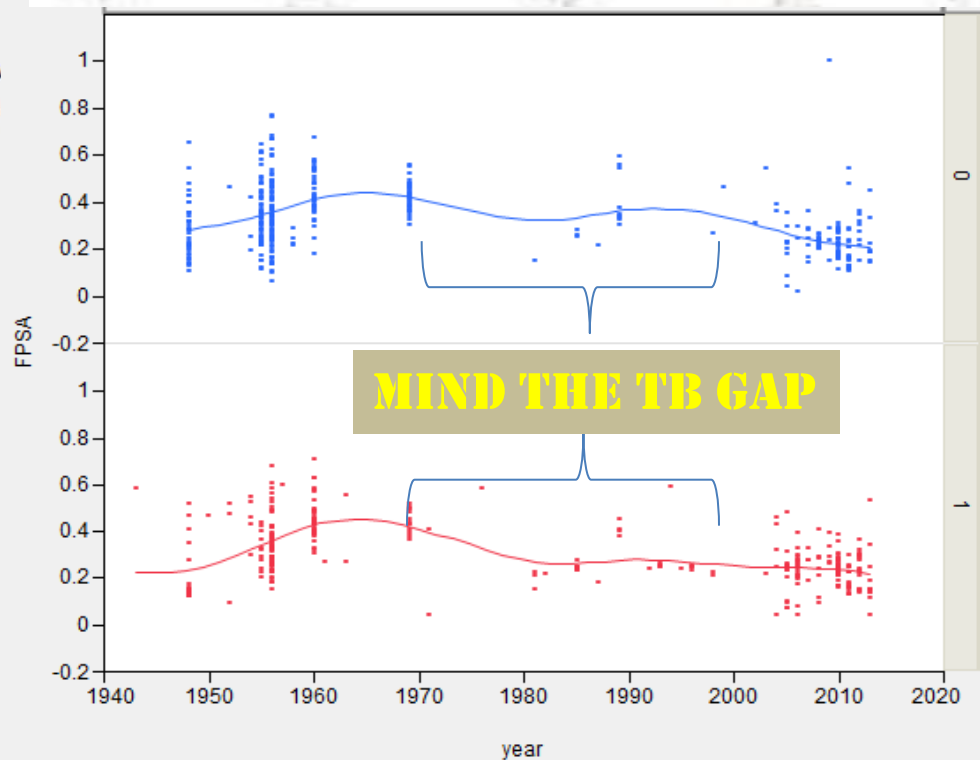
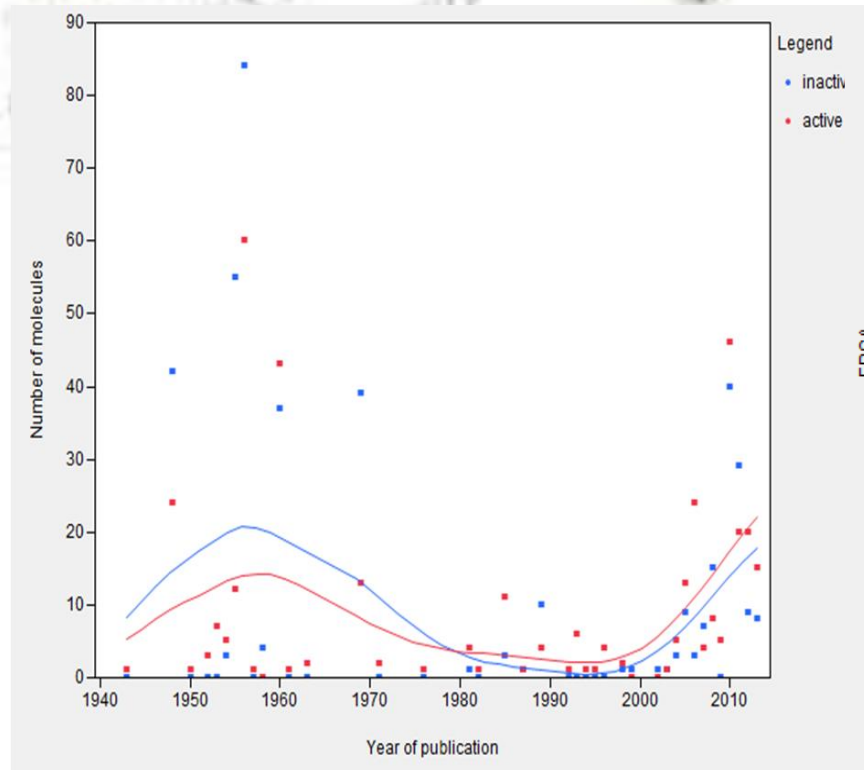
Downloaded from www.chemspider.com

Combined with pertinent data fields

1 \log_{10} reduction in *Mtb* colony-forming units (CFUs) in the lungs

Publically available CDD TB database (In process)

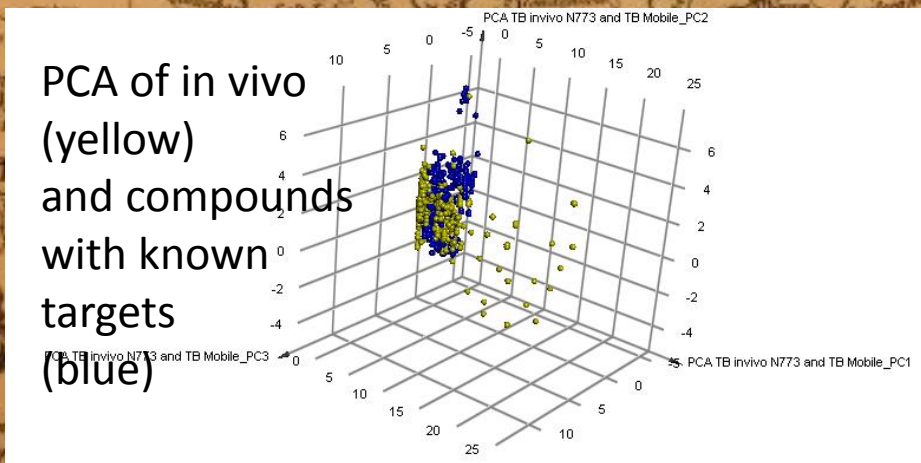
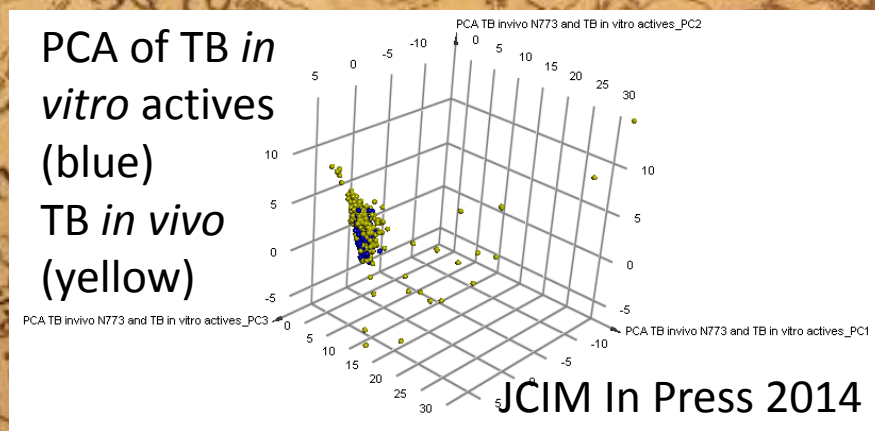
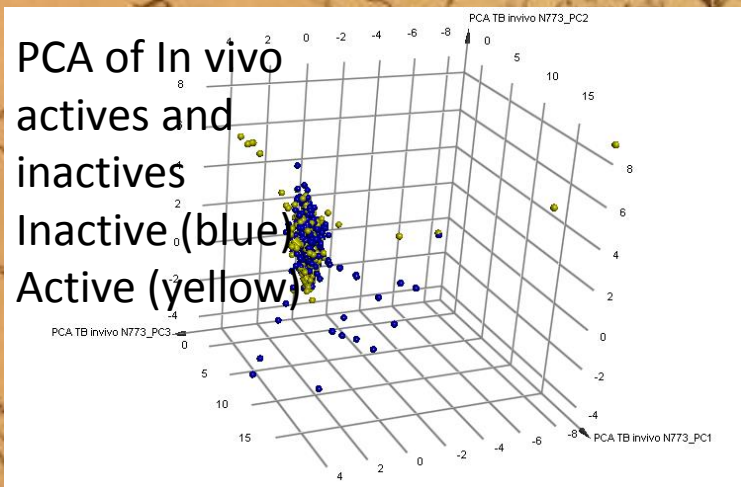
30 years with little TB mouse *in vivo* data



JCIM In Press 2014

Ekins, Nuermberger & Freundlich Submitted

Where are the New TB drugs to be found?



Machine Learning Models

Bayesian

Support Vector Machine

Recursive partitioning (single and multiple trees)

Using Accelrys Discovery Studio and R.

RP Forest

RP Single

SVM

Bayesian

Tree

ROC 5 fold cross validation

0.75

0.71

0.77

0.73

External Test set

11 additional active molecules obtained from 1953-2013

RP Forest

RP Single

SVM

Bayesian

Tree

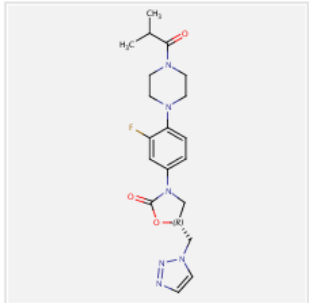
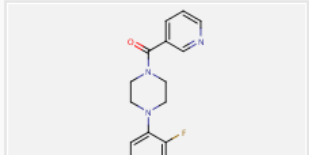
3 / 11

4 / 11

7 / 11

8 / 11

778 Selected: [Plot](#) [Export](#) [Add to collection](#) [Customize your report](#)

Select...	Molecule	Chemical Properties	TB mouse in vivo data from the literature						
all none		Molecular weight (g/mol)	Mouse model	Treatment	Results	In vivo activity	Activity Score	PubMed Id	Year
<input checked="" type="checkbox"/>	<p>flag outliers</p>  <p>1g phillips TB in-vivo data2</p>	416.449	GKO mouse model	22 days	Lung < 1 log CFU reduction, 25 mg/kg	Inactive	0	Phillips et al. 2012	2012
			JCIM In Press 2014						
<input checked="" type="checkbox"/>	<p>flag outliers</p> 	451.454	GKO mouse model	22 days	Lung < 1 log CFU reduction, 25 mg/kg	Inactive	0	Phillips et al. 2012	2012

The Clock is ticking

A much higher ratio of compounds were tested *in vivo* to *in vitro* in the 1940s-1960s rather than now

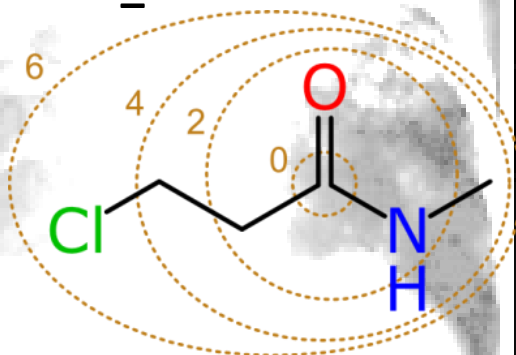
Infrastructure to provide a clear understanding of the position of compounds in the pipeline is essentially lacking

Shortage of new candidates suggest we may lack the commitment and resources we had 60 years ago

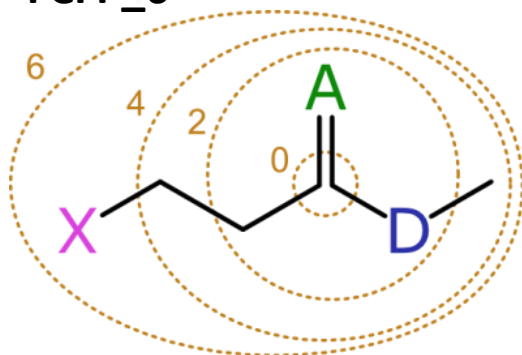
Use machine learning *in vivo* models to prioritize Mouse studies

Open Extended Connectivity Fingerprints

ECFP_6



FCFP_6



- Collected, deduplicated, hashed
- Sparse integers

- Invented for Pipeline Pilot: public method, proprietary details
- Often used with Bayesian models: many published papers
- Built a new implementation: open source, Java, CDK
 - stable: fingerprints don't change with each new toolkit release
 - well defined: easy to document precise steps
 - easy to port: already migrated to iOS (Objective-C) for *TB Mobile* app
- Provides core basis feature for CDD open source model service

Same datasets – Versus published data

Dataset	Leave one out ROC Published	Reference	Leave one out ROC Open fingerprints
In vivo data (773 molecules) FCFP_6 fingerprints	0.77	this study	0.75
Combined model (5304 molecules) FCFP_6 fingerprints	0.71	<i>J Chem Inf Model</i> 53:3054- 3063.	0.77
MLSMR dual event model (2273 molecules) and FCFP_6 fingerprints	0.86	<i>PLOS ONE</i> 8:e63240	0.83

Clark et al., submitted 2014

Open fingerprints and bayesian method used in TB Mobile Vers.2

CDD
COLLABORATIVE DRUG DISCOVERY

Tuberculosis Mobile

ROGERS 12:15 PM

HOCCNCC1=CC=CC=C1 ald (Rv2780): 0.362169
alr (Rv3423c): 0.142056
Rv1885c: 0.068163

CCN(CC)C1=CC=CC=C1 ald (Rv2780): 0.367381
alr (Rv3423c): 0.405582
ftsZ (Rv2150c): 0.412138
dprE1 (Rv3790): 0.142056

CC(=O)OC1=CC=CC=C1 Rv1885c: 0.147751
ftsZ (Rv2150c): 0.0693469

ROGERS 12:28 PM

Cluster molecules

Legend:
inhA (Rv1484)
glf (Rv3809c)
fbpC (Rv0129c)

ROGERS 12:15 PM

Predict targets

HOCCNCC1=CC=CC=C1 ald (Rv2780)
alr (Rv3423c)
Rv1885c

CCN(CC)C1=CC=CC=C1 ftsZ (Rv2150c)
alr (Rv3423c)
ald (Rv2780)
dprE1 (Rv3790)
gyrB (Rv0005)
dprE2 (Rv3791)

CC(=O)OC1=CC=CC=C1 Rv1885c
ftsZ (Rv2150c)

AirDrop
Share instantly with people nearby. If they do not appear automatically, ask them to open Control Center and turn on AirDrop.

Open in MolSync Open in MMDS Open in Yield101 Open in Reaction101

Cancel

iTunes Preview

TB Mobile

By Collaborative Drug Discovery

Open iTunes to buy and download apps.

Description
TB Mobile makes available a set of molecules with activity against Mycobacterium Tuberculosis, and known targets available in CDD. It links to pathways (biocyc.org), genes (tbd.org), literature (PubMed) and essentiality information.

[Collaborative Drug Discovery Web Site](#) [TB Mobile Support](#)

What's New in Version 1.0.1
Minor bug fixes.

Free
Category: Productivity

<http://goo.gl/vPOKS>

Google play

TB Mobile

Drugs with activity against *tuberculosis*

TB Mobile
Collaborative Drug Discovery

Pyrazinoid acid
Isoniazid
Clotrimazole

Description
TB Mobile makes available a set of molecules with activity against Mycobacterium

<http://goo.gl/iDJFR>

Could we add *in vivo* prediction models to this?

Ekins et al., J Cheminform 5:13, 2013

Clark et al., submitted 2014

In vitro data

In vivo data



Target data

ADME/Tox data & Models

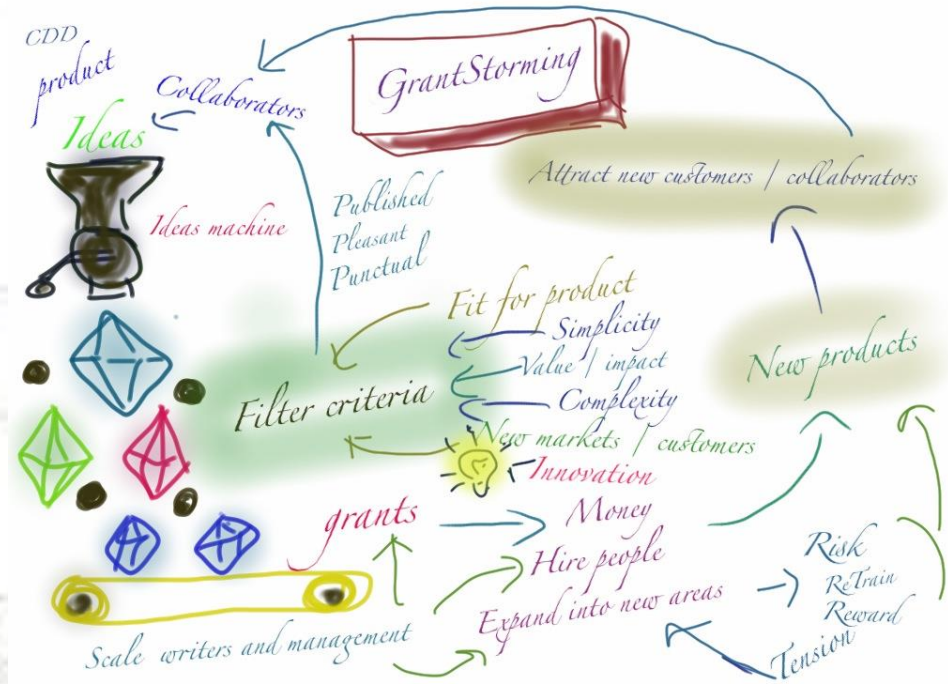
Data sources and tools we could integrate

Drug-like scaffold creation

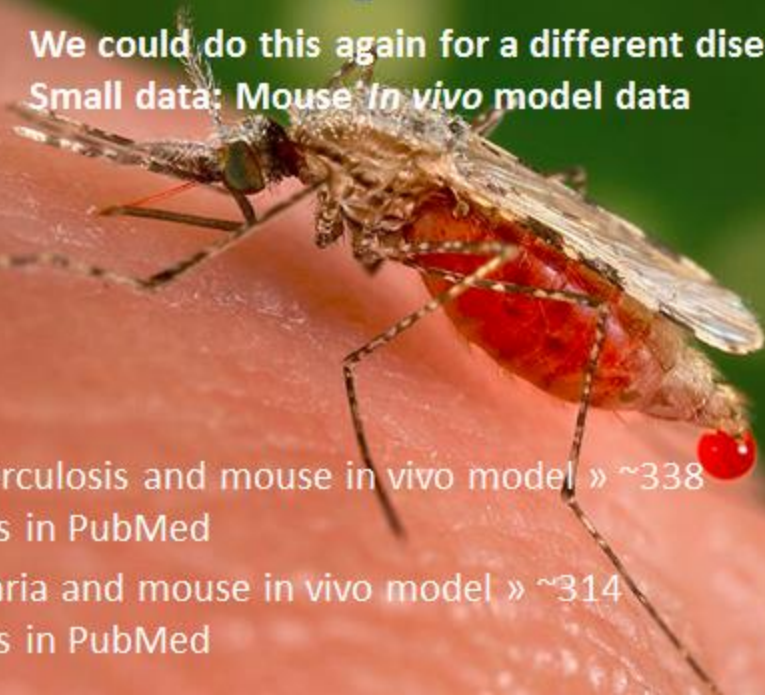
TB Prediction Tools

TB Publications

Future: How can we tackle more diseases?



We could do this again for a different disease
Small data: Mouse *In vivo* model data



«Tuberculosis and mouse in vivo model» ~338 papers in PubMed
«Malaria and mouse in vivo model» ~314 papers in PubMed

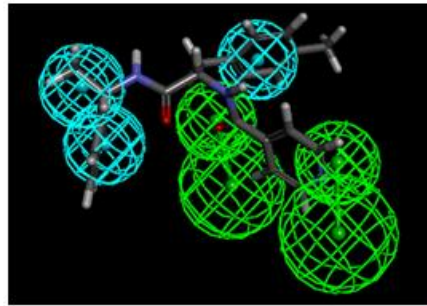


Chagas Disease

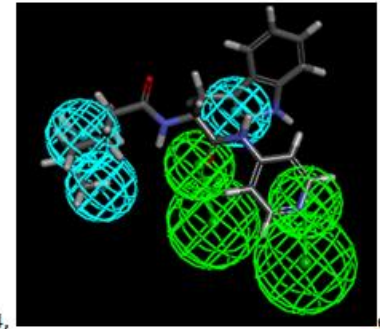
Reverse the mimic approach to predict targets of hits

Use pharmacophores for targets e.g. CYP51

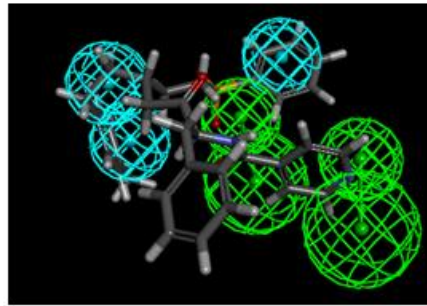
Use machine learning models to identify novel compounds



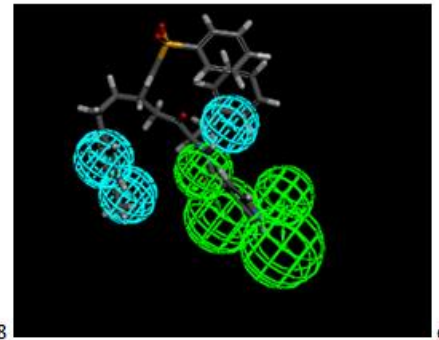
cpd 14,



cpd 5



cpd 8



cpd 4



Jonah's Just Begun

www.jonahsjustbegun.org



**The new faces of
personalized medicine:
children with rare diseases**



The Rare Disease Parent Odyssey

- Diagnosis of child
- Try to find out about disease – papers behind paywall
- Try to connect with scientists
- Form not-for-profit
- Raise funds
- Fund Scientific research on disease
- Advocate for support from NIH, FDA etc
- Start a company
- Try to find a cure before its too late

Could we create a rare disease community for scientists & foundations ?



Future: how do we deliver data

Rare diseases inspired an App that may be a new kind of database

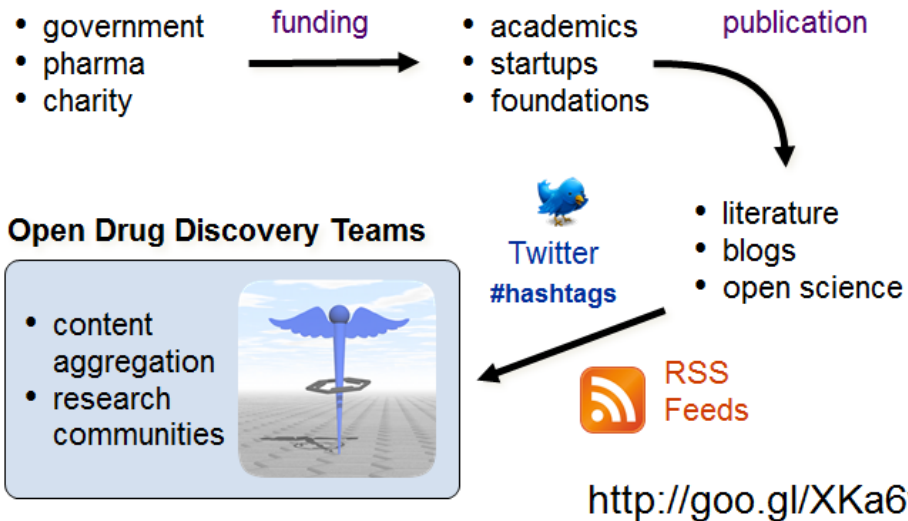
upload molecules by tweeting them-
1 tweet upload

Take our data with us anywhere

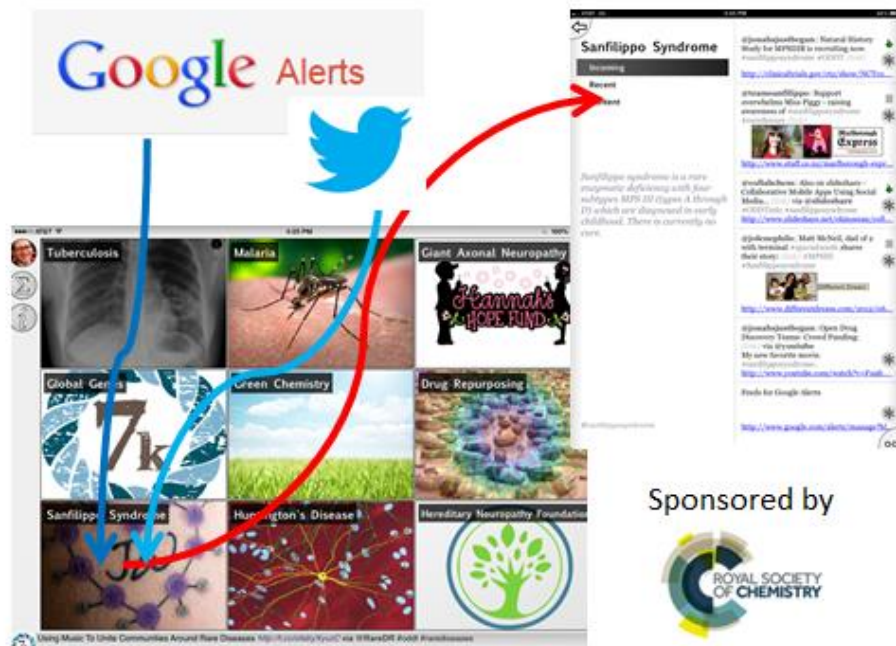
Bring data off the cloud into device

Advantages you get to analyze it in the Cloud on a plane

The Solution Schematic



The Open Drug Discovery Teams (ODDT) iOS App





All at CDD and many others ...Funding: 1R41AI088893-01, 2R42AI088893-02, R43 LM011152-01, 9R44TR000942-02, 1R41AI108003-01, MM4TB, Software: Accelrys

BRAINTRUST