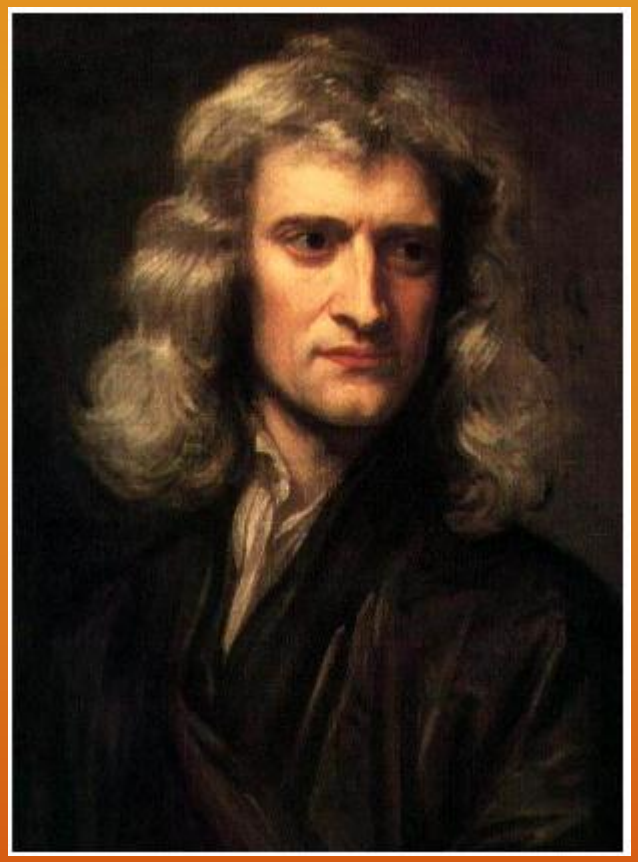
Data Quality Issues That Can Impact Drug Discovery

Sean Ekins[1], Joe Olechno[2] Antony J. Williams[3]

[1] Collaborations in Chemistry, Fuquay Varina, NC.
[2] Labcyte Inc, Sunnyvale, CA.
[3] Royal Society of Chemistry, Wake Forest, NC.

Disclaimer: SE and AJE have no affiliation with Labcyte and have not been engaged as consultants



"If I have seen further than others, it is by standing upon the shoulders of giants."

Isaac Newton

Example e.g. GPCR drug discovery

Pictures: Wikipedia

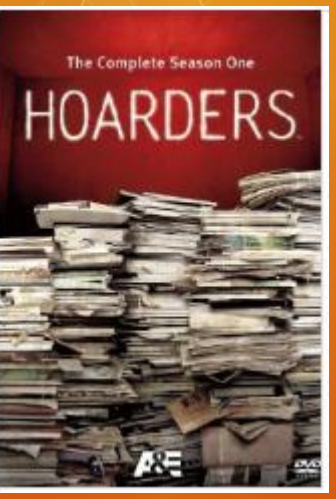
- Clone gene for β-receptor
- Cloning of more genes for other receptors
- Understanding of similarities
- Identify family GPCRs
- Now 1000s of receptors belong to this family
- Understand how to target receptors - identify structure
- Results in drugs tailored to fit these receptors
- "GPCR and drug discovery" – 661 hits in PubMed since 1997

Where do scientists get chemistry/ biology data?

- Databases
- Patents
- Papers
- Your own lab
- Collaborators

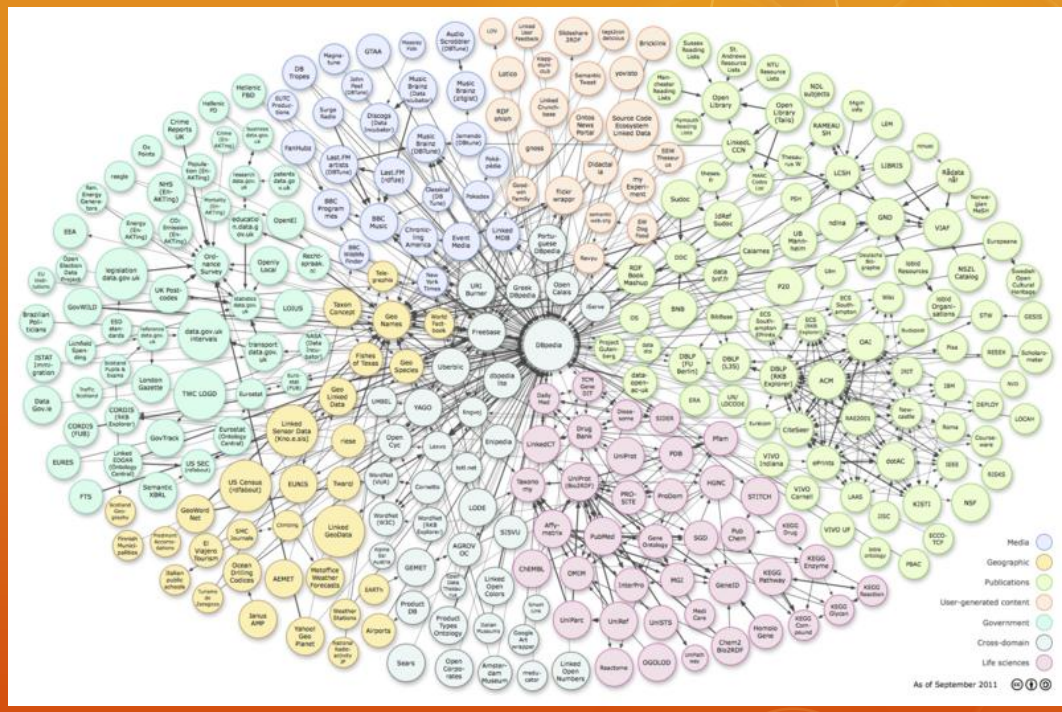
- Some or all of the above?
- What is common to all? – quality issues

From data hoarding to open data



(IMDB)

Me



Linked Open data cloud 2011 (Wikipedia)

Pharma company data hoarding - to open data

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

News & Comment › News › 2012 › November › Article

NATURE | NEWS

Data sharing aids the fight against malaria

'Open innovation' fosters good science but yields few promising drug candidates.

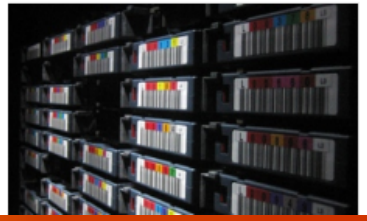
Daniel Cressey

14 February 2012

print
email
rights & permissions
share/bookmark

In the hunt for drugs that target diseases in the developing world, 'open innovation' is creating a buzz. Pharmaceutical companies are making entire libraries of chemical compounds publicly available, allowing researchers to rifle through them for promising drug candidates.

The latest push for open innovation, unveiled last month as part of a World Health Organization road map to control neglected tropical diseases, will see 11 companies sharing their intellectual property to give researchers around the world a head start on investigating drug leads (see 'Road map unveiled to tackle neglected diseases'). It makes for good press, and investors are not worried about giving away potentially blockbuster drugs because the diseases in question are not commercial priorities.



But is it good science? The answer, from the first large-scale initative of this kind, is a cautious 'yes'.

Free for all

Two years ago, GlaxoSmithKline (GSK) announced that it would release details of about 13,500 molecules that had already been shown to

gsk GlaxoSmithKline

Careers | Media | Investors | Consumers | H

Home | Explore GSK | Research | Partnerships | Responsibility

- Press releases
- From our CEO
- Speeches & presentations
- Social media
- Resource centre
- Media contact details

GSK announces further initiatives to advance openness and collaboration to help tackle global health challenges

Issued: Thursday 11 October 2012, London UK

- Tuberculosis (TB) 'compound library' to be made available to help stimulate research into TB
- Investment in GSK's Tres Cantos Open Lab to be doubled with an additional £5m funding awarded
- Detailed data from GSK clinical trials to be made available to researchers to further scientific understanding and knowledge

GSK today announced new measures to further advance its commitment towards

g hosted by
line new
'open
r the world's

FierceBiotechIT

NEWS TOPICS ANALYSIS FEATURE

Topics: Data Management

GSK to open clinical data vaults to scientists via website

October 11, 2012 | By Ryan McBride

SHARE

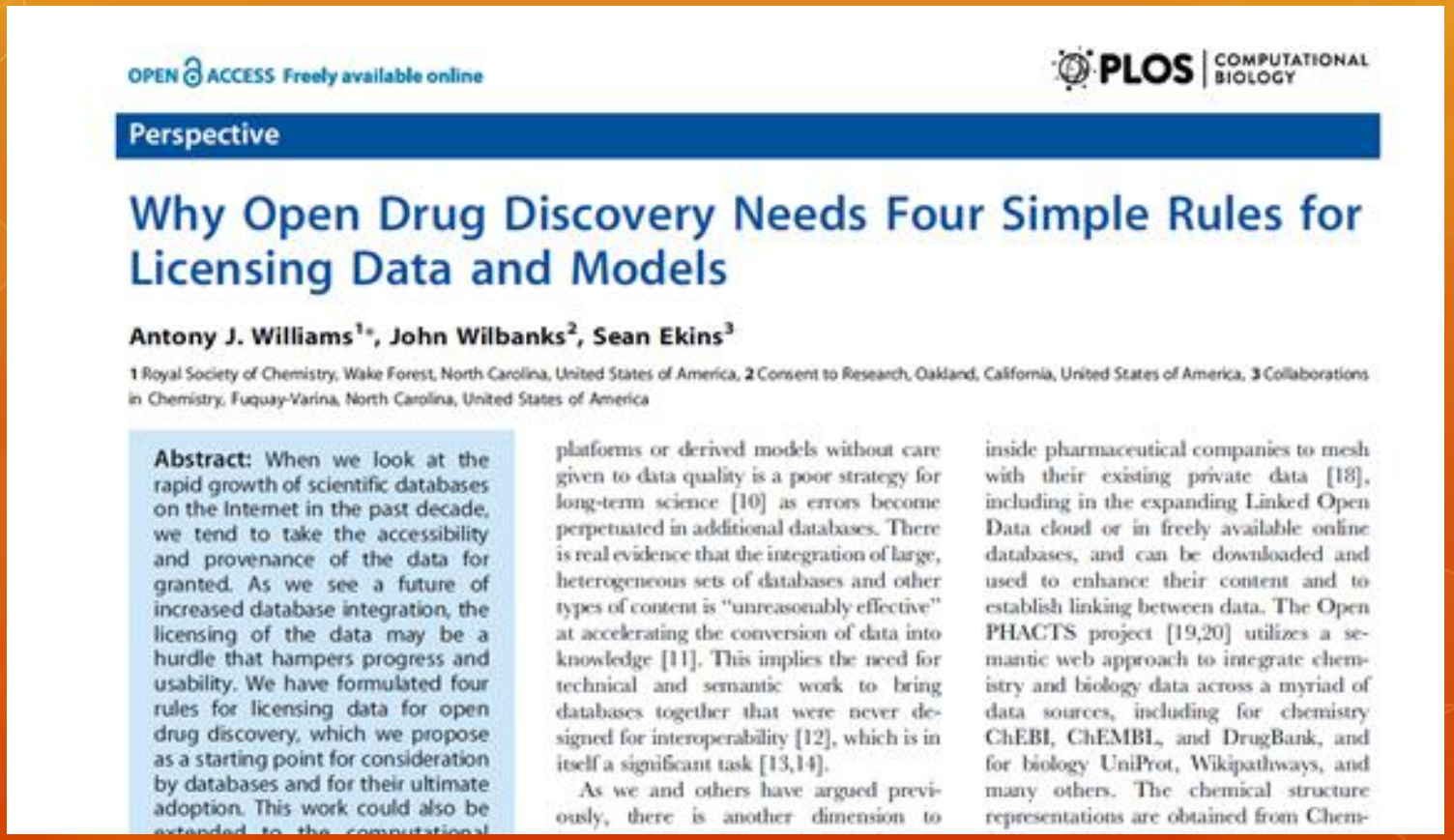
Email

7

GlaxoSmithKline ($GSK) committed to provide researchers with unmatched access to patient-level data from its clinical trials--including studies that failed. The move is among several steps the London-based drug giant announced today to promote open innovation and collaboration with external groups. Yet commentators are skeptical about whether fellow drugmakers will be as bold in opening their data vaults to outsiders.

Simple Rules for licensing "open" data

As we see a future of increased database integration the licensing of the data may be a hurdle that hampers progress and usability.



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Perspective

Why Open Drug Discovery Needs Four Simple Rules for Licensing Data and Models

Antony J. Williams[1]*, John Wilbanks[2], Sean Ekins[3]

1 Royal Society of Chemistry, Wake Forest, North Carolina, United States of America, 2 Consent to Research, Oakland, California, United States of America, 3 Collaborations in Chemistry, Fuquay-Varina, North Carolina, United States of America

Abstract: When we look at the rapid growth of scientific databases on the Internet in the past decade, we tend to take the accessibility and provenance of the data for granted. As we see a future of increased database integration, the licensing of the data may be a hurdle that hampers progress and usability. We have formulated four rules for licensing data for open drug discovery, which we propose as a starting point for consideration by databases and for their ultimate adoption. This work could also be extended to the computational

platforms or derived models without care given to data quality is a poor strategy for long-term science [10] as errors become perpetuated in additional databases. There is real evidence that the integration of large, heterogeneous sets of databases and other types of content is "unreasonably effective" at accelerating the conversion of data into knowledge [11]. This implies the need for technical and semantic work to bring databases together that were never designed for interoperability [12], which is in itself a significant task [13,14].

As we and others have argued previously, there is another dimension to

inside pharmaceutical companies to mesh with their existing private data [18], including in the expanding Linked Open Data cloud or in freely available online databases, and can be downloaded and used to enhance their content and to establish linking between data. The Open PHACTS project [19,20] utilizes a semantic web approach to integrate chemistry and biology data across a myriad of data sources, including for chemistry ChEBI, ChEMBL, and DrugBank, and for biology UniProt, Wikipathways, and many others. The chemical structure representations are obtained from Chem-

Williams, Wilbanks and Ekins. PLoS Comput Biol 8(9): e1002706, 2012

Could open accessibility = Disruption

Drug Discovery Today • Volume 00, Number 00 • October 2012 REVIEWS

ELSEVIER

Reviews • INFORMATICS

Four disruptive strategies for removing drug discovery bottlenecks

Sean Ekins[1], Chris L. Waller[2,3], Mary P. Bradley[4], Alex M. Clark[5] and Antony J. Williams[6]

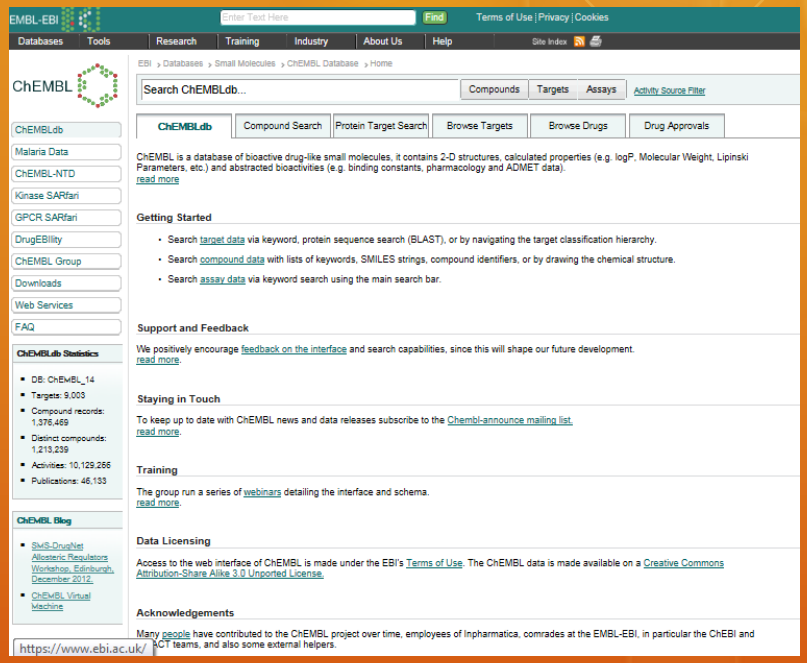
[1] Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, USA
[2] Pfizer Inc., Eastern Point Road, Groton, CT 06340, USA
[3] School of Pharmacy, University of North Carolina, Chapel Hill, NC 27514, USA
[4] CollaborationFinder, Harvard Square, One Mifflin Place, Suite 400, Cambridge, MA 02138, USA
[5] Molecular Materials Informatics, 1900 St. Jacques #302, Montreal, Quebec, Canada H3J 2S1
[6] Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC 27587, USA

Drug discovery is shifting focus from industry to outside partners and, in the process, creating new bottlenecks. Technologies like high throughput screening (HTS) have moved to a larger number of academic and institutional laboratories in the USA, with little coordination or consideration of the outputs and creating a translational gap. Although there have been collaborative public–private partnerships in Europe to share pharmaceutical data, the USA has seemingly lagged behind and this may hold it back. Sharing precompetitive data and models may accelerate discovery across the board, while finding the best collaborators, mining social media and mobile approaches to open drug discovery should be evaluated in our efforts to remove drug discovery bottlenecks. We describe four strategies to rectify the current unsustainable situation.

1: NIH and other international scientific funding bodies should mandate …open accessibility for all data generated by publicly funded research

Ekins, Waller, Bradley, Clark and Williams. DDT In press, 2012

Data can be found – but what about quality?



Why drug structure quality is important?

- More groups doing in silico repositioning
- Target-based or ligand-based
- Network and systems biology

- They are all integrating or using sets of FDA drugs..if the structures are incorrect predictions will be too..

- What is need is a definitive set of FDA approved drugs with correct structures and tools for in silico screening

- Also linkage between in vitro data & clinical data

Structure Quality Issues

Database released and within days 100's of errors found in structures

NPC Browser http://tripod.nih.gov/npc/

PERSPECTIVE

PHARMACOLOGY

Science Translational Medicine 2011

The NCGC Pharmaceutical Collection: A Comprehensive Resource of Clinically Approved Drugs Enabling Repurposing an Chemical Genomics

Ruili Huang,* Noel Southall,* Yuhong Wang, Adam Yasgar, Paul Shinn, Ajit Jadhav, Dac-Trung Nguyen, Christopher P. Austin†

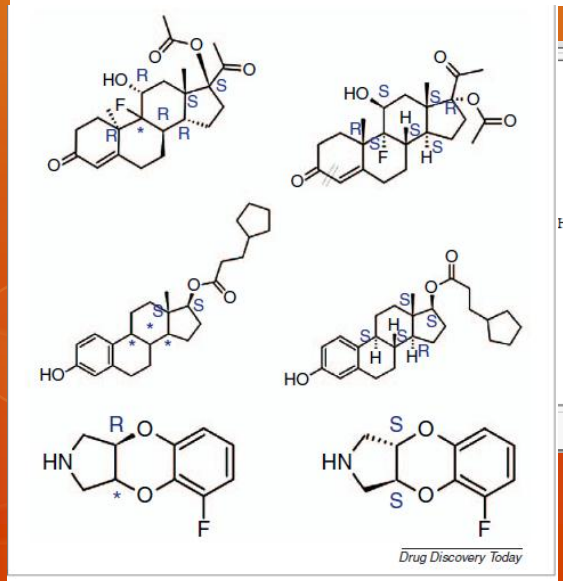
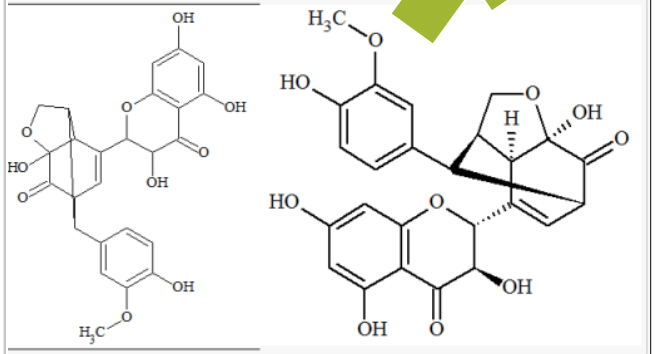


FIGURE 2
Stereochemistry issues with the content of the NPC browser database: left hand side 'NCGC Structures' and right hand side 'Correct Structures'.



Silidianin: Left Structure is Cleaned and Right Structure is with Stereocenters Added

editorial

Antony J. Williams

Sean Ekins

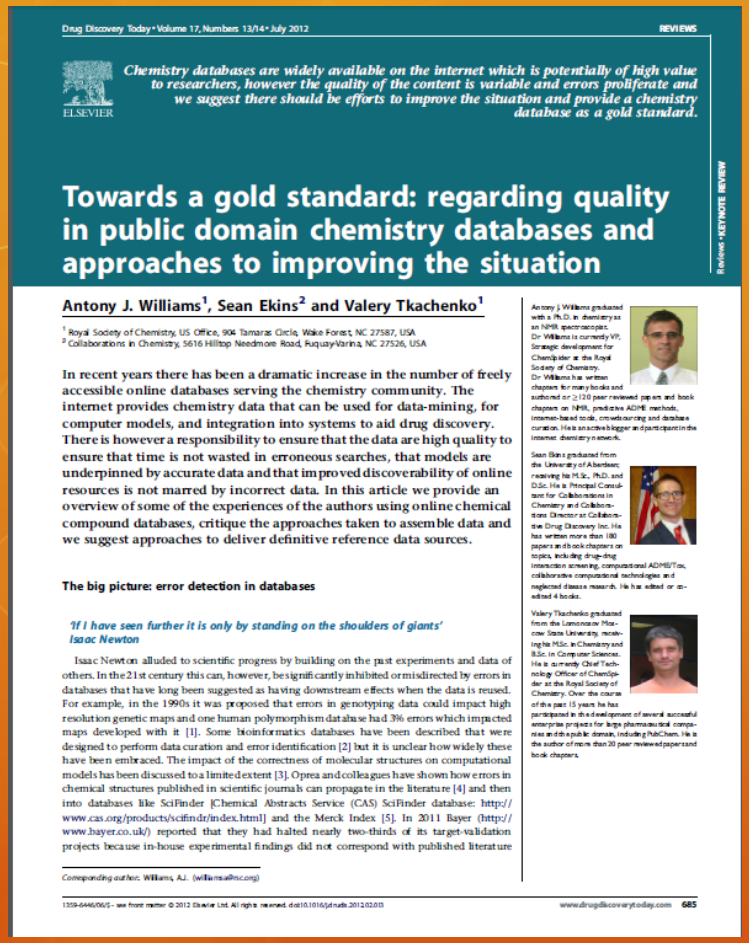
A quality alert and call for improved curation of public chemistry databases

In the last ten years, public online databases have rapidly become trusted valuable resources upon which researchers rely for their chemical structures and data for use in cheminformatics, bioinformatics, systems biology, translational medicine and now drug repositioning or repurposing efforts. Their utility depends on the quality of the underlying molecular structures used. Unfortunately, the quality of much of the chemical structure-based data introduced to the public domain is poor. As an example we describe some of the errors found in the recently released NIH Chemical Genomics Center 'NPC browser' database as an example. There is an urgent need for government funded data curation to improve the quality of internet chemistry and to limit the proliferation of errors and wasted efforts.

US funding agencies have been investing in the development of public domain chemistry platforms with the primary attention being given to the informatics platform itself rather the quality of the data content. This is clearly exemplified by the recently released NPC browser from the NIH Chemical Genomics Center (NCGC) [1]. Public online databases such as PubChem, ChemID-Plus [2] and the EPA's ACToR [3], to name just a few, have rapidly become trusted valuable resources which researchers rely on for downloadable chemical structures and associated data. While online chemistry databases can certainly be of value, we feel the reader should be immediately alerted to consider issues of data quality when using these resources and we call into question both their status and the trust we place in them. To our knowledge the issues we raise, using the example of a recently released database, have not been described elsewhere and the user community, and funding agencies, should not ignore them any longer. The development of cheminformatics platforms without due care given to the data quality they contain, is a poor strategy for long term science.

In the last decade numerous attempts have been made to expand our understanding of biological mechanisms by producing vast ligand and protein–protein interaction databases and by the application of computational methods to mine the data and, where possible, develop computational models. These approaches have enabled: theclustering of biological activity spectra similarity

Williams and Ekins, DDT, 16: 747-750 (2011)



Drug Discovery Today • Volume 17, Numbers 13/14 • July 2012 REVIEWS

Chemistry databases are widely available on the internet which is potentially of high value to researchers, however the quality of the content is variable and errors proliferate and we suggest there should be efforts to improve the situation and provide a chemistry database as a gold standard.

Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation

Antony J. Williams[1], Sean Ekins[2] and Valery Tkachenko[1]

[1] Royal Society of Chemistry, US Office, 904 Tamaras Circle, Wake Forest, NC 27587, USA
[2] Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, USA

In recent years there has been a dramatic increase in the number of freely accessible online databases serving the chemistry community. The internet provides chemistry data that can be used for data-mining, for computer models, and integration into systems to aid drug discovery. There is however a responsibility to ensure that the data are high quality to ensure that time is not wasted in erroneous searches, that models are underpinned by accurate data and that improved discoverability of online resources is not marred by incorrect data. In this article we provide an overview of some of the experiences of the authors using online chemical compound databases, critique the approaches taken to assemble data and we suggest approaches to deliver definitive reference data sources.

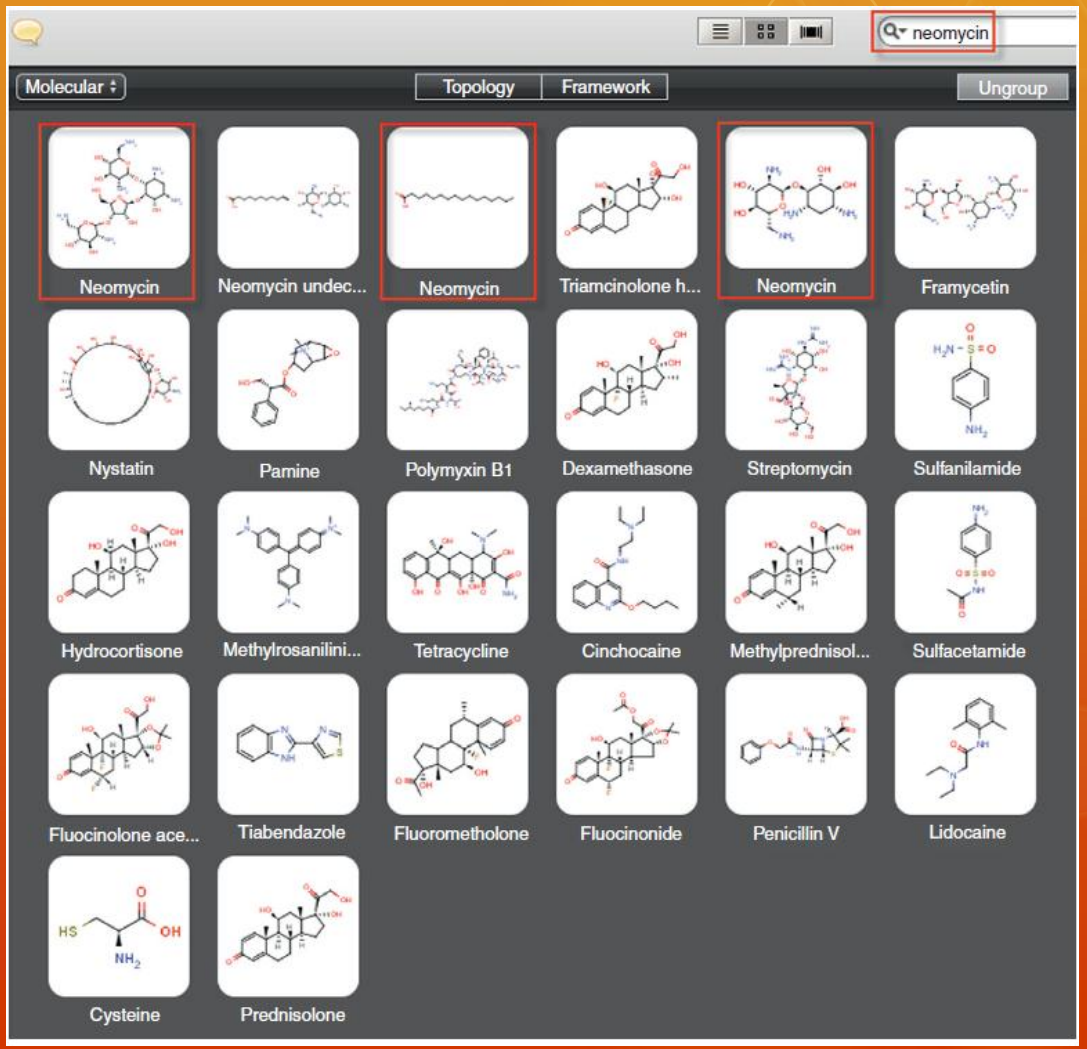
The big picture: error detection in databases

'If I have seen further it is only by standing on the shoulders of giants'
Isaac Newton

Isaac Newton alluded to scientific progress by building on the past experiments and data of others. In the 21st century this can, however, be significantly inhibited or misdirected by errors in databases that have long been suggested as having downstream effects when the data is reused. For example, in the 1990s it was proposed that errors in genotyping data could impact high resolution genetic maps and one human polymorphism database had 3% errors which impacted maps developed with it [1]. Some bioinformatics databases have been described that were designed to perform data curation and error identification [2] but it is unclear how widely these have been embraced. The impact of the correctness of molecular structures on computational models has been discussed to a limited extent [3]. Oprea and colleagues have shown how errors in chemical structures published in scientific journals can propagate in the literature [4] and then into databases like SciFinder [Chemical Abstracts Service (CAS) SciFinder database: http://www.cas.org/products/scifindr/index.html] and the Merck Index [5]. In 2011 Bayer (http://www.bayer.co.uk/) reported that they had halted nearly two-thirds of its target-validation projects because in-house experimental findings did not correspond with published literature

Corresponding author: Williams, A.J. (williamsa@rsc.org)

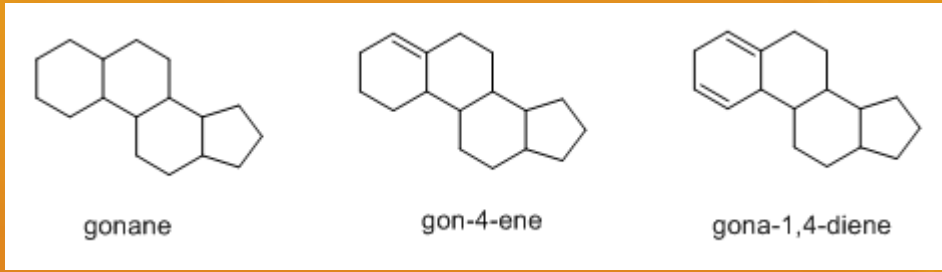
Which is Neomycin?



Describes errors in other types of databases too!!

Williams, Ekins and Tkachenko Drug Disc Today 17: 685-701 (2012)

Data Errors in the NPC Browser: Analysis of Steroids



Substructure	# of Hits	# of Correct Hits	No stereochemistry	Incomplete Stereochemistry	Complete but incorrect stereochemistry
Gonane	34	5	8	21	0
Gon-4-ene	55	12	3	33	7
Gon-1,4-diene	60	17	10	23	10

Williams, Ekins and Tkachenko Drug Disc Today 17: 685-701 (2012)

DDT editorial Dec 2011

Webinars
Downloads/ White Papers
News
Key Research Articles
Biotherapeutics
Companies and People
Drug Metabolism
Drug Trials
Informatics
Medicinal Chemistry
Novel Technologies
Pharmacology/ Therapeutics
Regulatory
Target Identification/ Validation
The View From Here
Events & Training

DRUG DISCOVERY TODAY
Subscribe today for your free Digital Edition!

slas 2013
2nd annual CONFERENCE & EXHIBITION
Short Courses January 12-13

News

The long term cost of inferior database quality

09 December 2011

This month's editorial is written by Antony J. Williams and Sean Ekins. "If I have seen further than others, it is by standing upon the shoulders of giants."Isaac Newton

Without question science has progressed by building on the theories, past experiments and data of others. Historically, scientists have guarded their data carefully if not hoarded it. Others have had to look on in envy as scientists were enabled to generate data that was inaccessible to most. Times have changed and citizen scientists are now 'asked'to review data in a crowdsourced approach to data review and validation (e.g. Galaxy Zoo [http://www.galaxyzoo.org/story/], ChemSpider [1]) and public engagement in data modelling [e.g. foldit (http://fold.it/portal/)]. In parallel, data sets are now made available to the community for download and integration to in-house systems, and the possibility exists where these data could be used by someone sitting in their basement to identify a new drug. Nowadays, immense quantities of scientific information are contained in the thousands of databases that exist on computers worldwide. Progress can however be inhibited by errors in these databases, and it has long been suggested as having downstream effects when the data is reused. There are few publications in this domain but we have collected some that should be of general interest to drug discovery scientists.

In the 1990s it was indicated that errors in genotyping data could impact high resolution genetic maps [2]. Some bioinformatics databases have been described that were designed to perform data curation and error identification [3] although it is unclear how widely these have been embraced. The impact of correctness of molecule structure on computational models has been discussed to a limited extent [4]. Oprea and colleagues have shown how errors in molecule structures published in scientific journals can propagate in the literature [5] and then into databases like SciFinder (http://www.cas.org/products/scifindr/index.html) and the Merck Index [6]. Even manually curated databases, such as the MDL drug data report [MDDR (http://accelrys.com/products/databases/bioactivity/mddr.html)] has been proposed to have errors [7]. It has been suggested that automatic classification of molecules based on SMILES strings might be useful for error detection and aiding biochemical pathway database construction [8]. Even new molecule databases appear to have not learnt from some of these earlier studies. For example, an NIH database was recently reported to have errors in 5% to over 10% of the molecules [9]. A recent review of data governance in predictive toxicology analyzing several public databases mentioned a lack of systematic and standard measures of data quality but did not address error rates or address molecule structure quality [10].

DRUG DISCOVERY TODAY

Sign up for your **FREE** digital edition of Drug Discovery Today

Antony J. Williams (Right) and Sean Ekins (Left).

Related Stories

Drug discovery database launches with 500,000 compounds
ChEMBLdb, an online database of information on the properties of drugs and drug-like small molecules and their targets, has launched with information on more than half a million compounds.

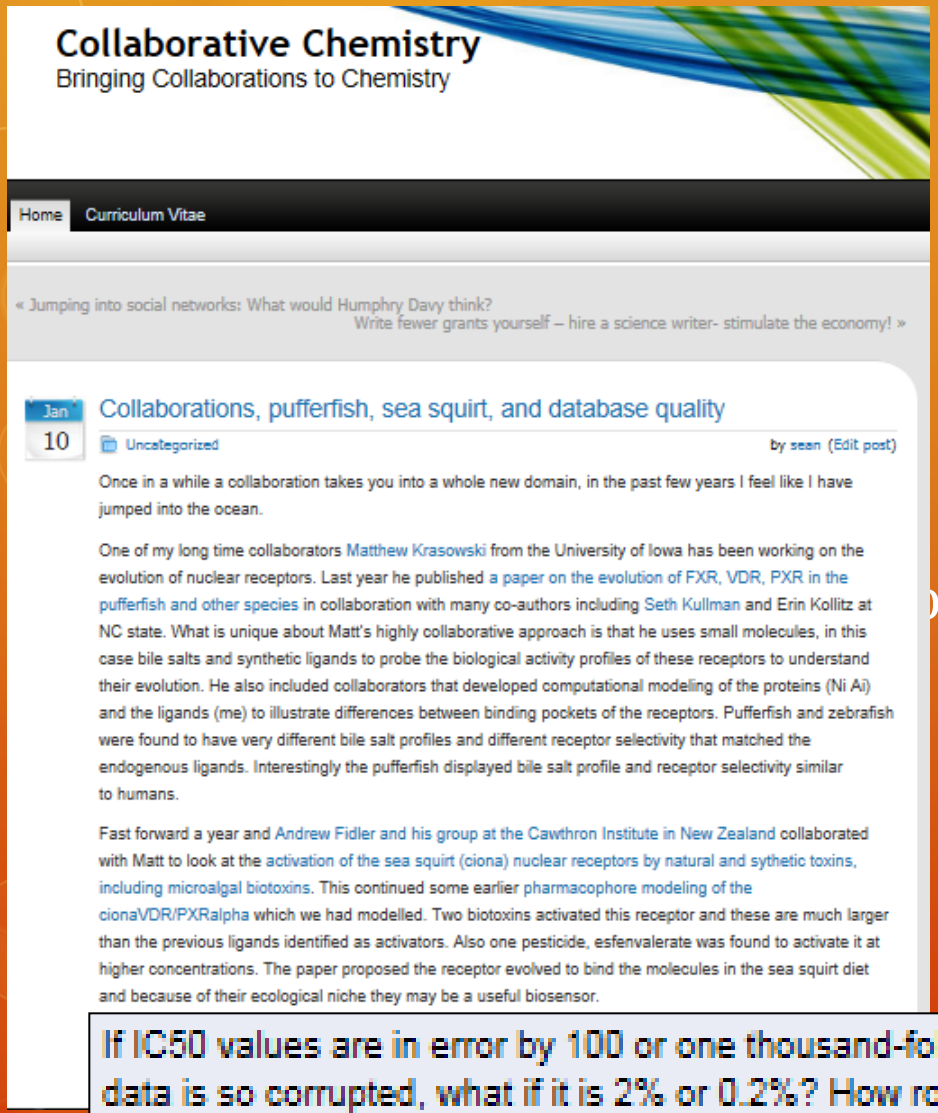
Cell culture in the third dimension
Sustaining life outside the body through the use of cell culture for research and discovery purposes has always been challenging. However, despite these challenges, cell culture has resulted in the discovery of most of today's knowledge about disease processes and their cures. There has always been the desire to increase the relevance of cell culture by reproducing the 3D context of the body and thus the rate of peer reviewed publications employing 3D cell culture has increased dramatically over the past several years [1]. Furthermore, many new 3D cell culture products are emerging in the bioscience market that are focused on making the process simpler and more affordable.

Email Address
Password
LOGIN
Forgotten login?
Not a member?

slas 2013
2nd annual CONFERENCE & EXHIBITION
Short Courses January 12-13
Conference January 14-16
Exhibition January 13-15
Gaylord Palms Resort and Convention Center Orlando, FL, USA
January 12-16 2013
SLAS2013.org

http://goo.gl/dIqhU

Its not just structure quality we need to worry about



Collaborative Chemistry
Bringing Collaborations to Chemistry

Home | Curriculum Vitae

« Jumping into social networks: What would Humphry Davy think?
Write fewer grants yourself – hire a science writer- stimulate the economy! »

Jan 10

Collaborations, pufferfish, sea squirt, and database quality

Uncategorized — by sean (Edit post)

Once in a while a collaboration takes you into a whole new domain, in the past few years I feel like I have jumped into the ocean.

One of my long time collaborators Matthew Krasowski from the University of Iowa has been working on the evolution of nuclear receptors. Last year he published a paper on the evolution of FXR, VDR, PXR in the pufferfish and other species in collaboration with many co-authors including Seth Kullman and Erin Kollitz at NC state. What is unique about Matt's highly collaborative approach is that he uses small molecules, in this case bile salts and synthetic ligands to probe the biological activity profiles of these receptors to understand their evolution. He also included collaborators that developed computational modeling of the proteins (Ni Ai) and the ligands (me) to illustrate differences between binding pockets of the receptors. Pufferfish and zebrafish were found to have very different bile salt profiles and different receptor selectivity that matched the endogenous ligands. Interestingly the pufferfish displayed bile salt profile and receptor selectivity similar to humans.

Fast forward a year and Andrew Fidler and his group at the Cawthron Institute in New Zealand collaborated with Matt to look at the activation of the sea squirt (ciona) nuclear receptors by natural and sythetic toxins, including microalgal biotoxins. This continued some earlier pharmacophore modeling of the cionaVDR/PXRalpha which we had modelled. Two biotoxins activated this receptor and these are much larger than the previous ligands identified as activators. Also one pesticide, esfenvalerate was found to activate it at higher concentrations. The paper proposed the receptor evolved to bind the molecules in the sea squirt diet and because of their ecological niche they may be a useful biosensor.

2 comments No ping yet

Joe Olechno says:
January 13, 2012 at 8:40 pm (UTC -5) | Edit comment Reply

Great blog and great editorial in "Drug Discovery Today"! What is the impact on database quality when the dose-response (IC50) values are in error due to the mechanism of transfer?

It appears that an unintended and unforeseen consequence of assay miniaturization is that standard liquid handling steps are more likely to sow errors than when assay volumes were relatively large. However, cost and time guarantee that miniaturization must continue.

Are SAR experiments being confounded because errors of three orders of magnitude are going into the databases? It seems that IC50 values can be easily corrupted by the mechanism of transfer.
For examples in errors in IC50 values driven by liquid handling see: US Patent, 7,718,653; or American Drug Discovery 3(3), 24-30.
For the impact of leachates from plastic tips see: Science 2008 7 November; 322(5903): 917; Clinical Chem. (2009) 55:1883.
For the impact of carry-over of active compounds see: Matson et al., 2009, 14:476, Table 2.

All of these papers suggest that data going into the databases is wrong and may be compromising the quality of those databases. Since the strength of the database affects the strength of the SAR program and the company, one process addresses all of these problems.

The problems that arise from underestimated potency, plastic leachates and carry-over can be dramatically reduced or eliminated with acoustic liquid handling. This technology eliminates physical contact with the liquid being transferred and that gets rid of leachates and carry-over. The researchers in the US patent cited above used acoustic liquid handling to obtain improved measures of potency—apparently the drug-like compounds were being pulled from the solution by the plastic pipette tips leading to erroneously high IC50 values.

If IC50 values are in error by 100 or one thousand-fold, how useful is the database? What if only 20% of the data is so corrupted, what if it is 2% or 0.2%? How robust will the database be if large errors are incorporated into the data?

sean says:
January 17, 2012 at 11:45 am (UTC -5) | Edit comment Reply

Joe
Many thanks for the comment, few would be aware of this so very much appreciate the summary. Do others have additional comments like this?

If IC50 values are in error by 100 or one thousand-fold, how useful is the database? What if only 20% of the data is so corrupted, what if it is 2% or 0.2%? How robust will the database be if large errors are incorporated into the data?

How do you move a liquid?

Low throughput

High throughput

Images courtesy of Bing

Plastic leaching

Bioactive Contaminants Leach from Disposable Laboratory Plasticware

G. Reid McDonald,[1] Alan L. Hudson,[1] Susan M. J. Dunn,[1] Haitao You,[1] Glen B. Baker,[2] Randy M. Whittal,[3] Jonathan W. Martin,[4] Amitabh Jha,[5] Dale E. Edmondson,[6] Andrew Holt[2]*

Recent reports of leaching of bisphenol A and antimony into foods and beverages from polycarbonate and polyethylene terephthalate containers, respectively, have drawn attention to plastics as potential sources of bioactive environmental contaminants (*1*, *2*). However, numerous other processing additives coat, or intercalate within, polymeric structures, and these also migrate into foods stored in plastic containers (*3*).

Disposable plasticware is used in life science laboratories worldwide. Although labeling of plastics as "sterile" appears to offer researchers some assurance that products are free of bioactive contaminants, the presence of processing additives is unavoidable. Herein, we report identification of two additives leaching from disposable plasticware and demonstrate potent effects on enzyme and receptor proteins.

Observations of anomalous kinetics with human monoamine oxidase-B (hMAO-B), which recognizes numerous xenobiotic substrates, led us to examine disposable tubes used in our assays as a potential source of interferences. Water rinsed through several brands and sizes of plastic tubes adopted inhibitory potency versus hMAO-B; when dimethyl sulfoxide (DMSO) (10%, v/v) was used instead, inhibition was more pronounced, and marked enzyme activation was observed in one case (Fig. 1A). Samples of two tubes (indicated) were rinsed with water (W) or methanol (M), and leachates were dried and subjected to mass spectrometry. Fragmentation spectra for major species present (Fig. 1B and fig. S1A) identified a biocide, di(2-hydroxyethyl)methyldodecylammonium (DiHEMDA), and a slip agent, 9-octadecenamide (oleamide), in W and M leachates, respectively (Fig. 1C, inset). Pure samples of DiHEMDA (*4*) and oleamide were confirmed as hMAO inhibitors, showing selectivity for hMAO-B (Fig. 1C). Inhibition by leachates from oleamide-positive tubes increased markedly over a 10-day plastic exposure period at 20°C (fig. S1B).

Repeated (10 times) pipetting of 10% DMSO with pipette tips from several suppliers resulted in extraction of species that had significant effects on

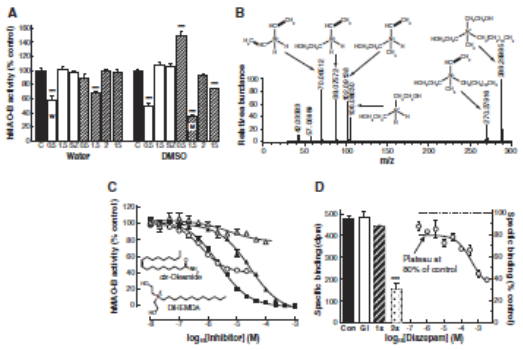


Fig. 1. (**A**) Effects on hMAO-B of water ($N = 3$) or 10% DMSO ($N = 7$) (40% of tube volume, 1 hour, 20°C) from Fisherbrand (Fisher Scientific, Ottawa, Canada) (clear bars) or Sarstedt (Sarstedt AG, Nümbrecht, Germany) (hatched bars) plastic tubes or glass vials (O). *x* axis labels indicate tube volumes (ml). C indicates

26 June 2008; accepted 2 September 2008
10.1126/science.1162395

[1]Department of Pharmacology, University of Alberta, Edmonton,

Clinical Chemistry 55:10
1883–1890 (2009)

Letters to the Editor

Nonylphenol Ethoxylate Plastic Additives Inhibit Mitochondrial Respiratory Chain Complex I

To the Editor:

The first and major entry point of electrons into the mitochondrial respiratory chain (MRC)[1] occurs through NADH-coenzyme Q reductase (complex I). Decreased complex I activity is associated with a wide range of conditions, including inherited mitochondrial diseases and neurodegenerative conditions such as Parkinson disease (*1–2*). More than 60 natural and commercial com-

mined that inhibition resulted from a single brief period of contact between assay reagents and blue polypropylene 1-mL pipette tips ("blue tips") manufactured by Scientific Specialties (Table 1).

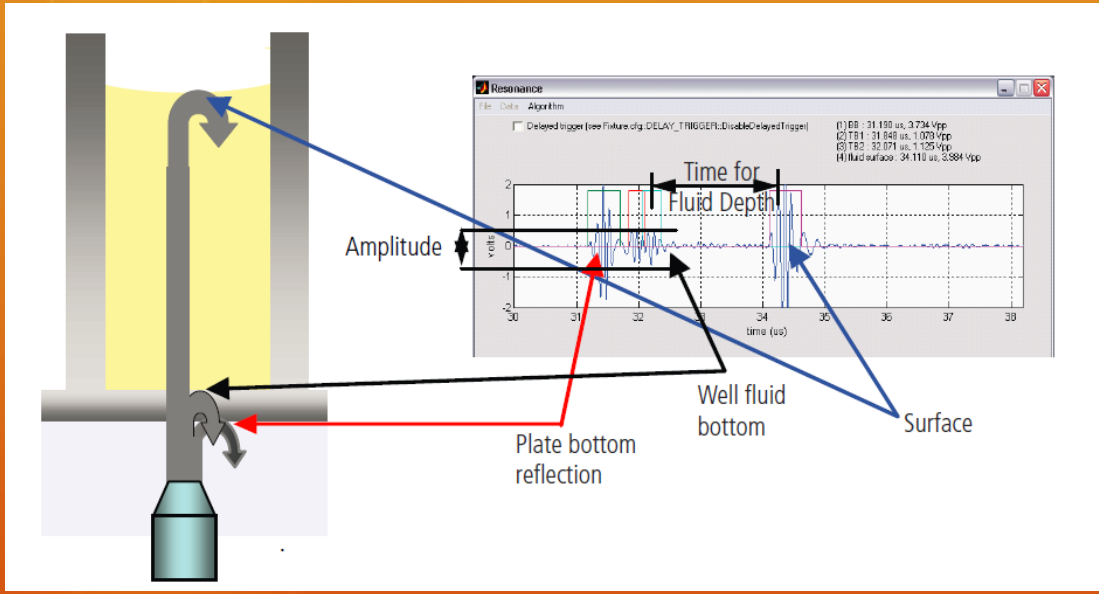
Methanol washes of blue tips were subjected to positive ion–mode electrospray mass spectrometry (Waters Mariner BioSpectrometry Workstation oa-TOF); these analyses confirmed the presence of NP compounds, marketed as surfactants from the Tergitol NP series. The prevalent ion at *m/z* 683.5 was consistent with the use of NP-10 as a dye-solubilizing agent in the plastic manufacturing process, whereas

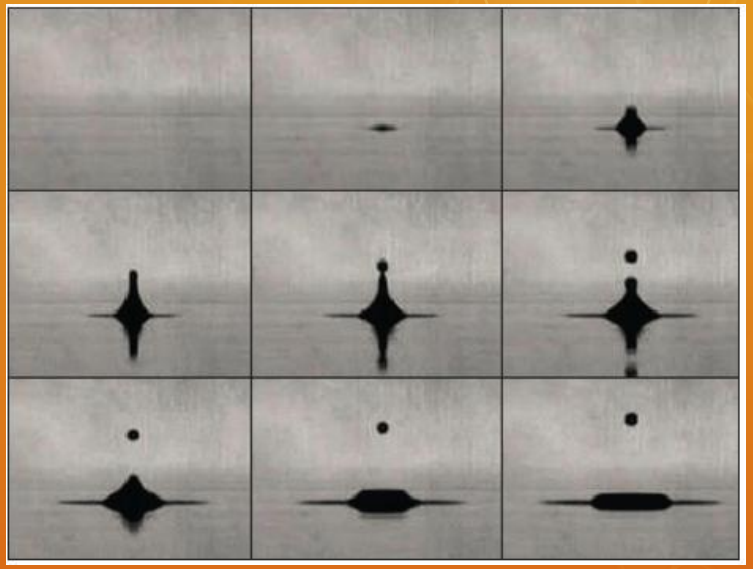
containing medium was minimally affected. The use of blue tips to pipette culture medium was deleterious for cell growth in restrictive medium, but affected growth in permissive medium to a lesser extent (Table 1). Other nonionic detergents—dodecyl maltoside and Tween 20—and clear colorless tips from a different manufacturer had no effects under either set of conditions.

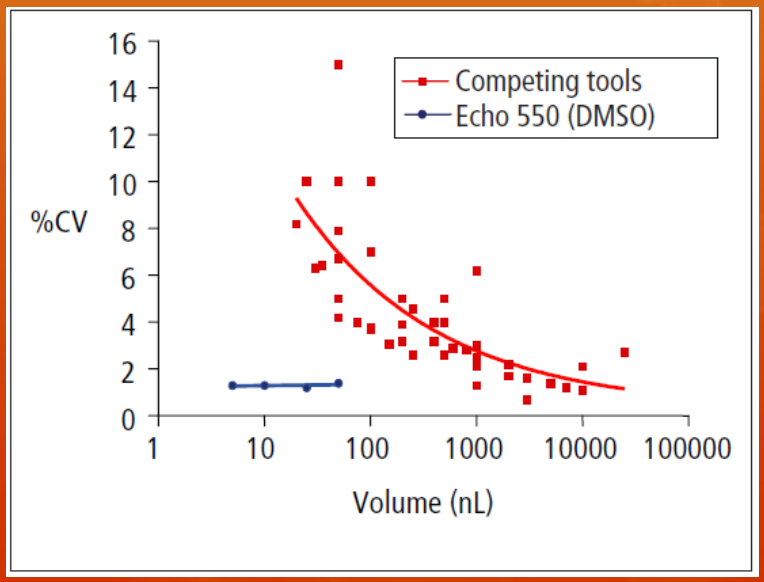
Our data clearly show that blue tips leached NP-10 into assay buffers and interfered with MRC complex I while leaving other complexes unaffected. Previously, McDonald et al. reported that

McDonald et al., *Science* **2008**, *322*, 917. Belaiche et al., *Clin Chem* **2009**, *55*, 1883-1884

Moving liquids with sound

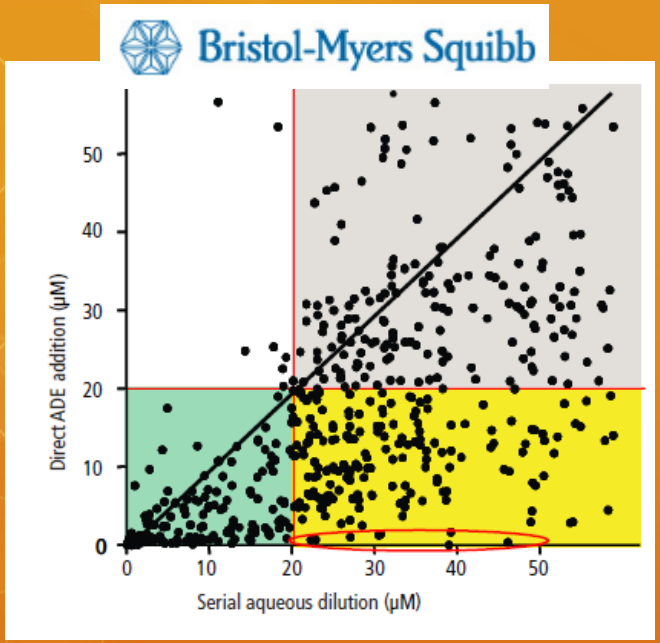


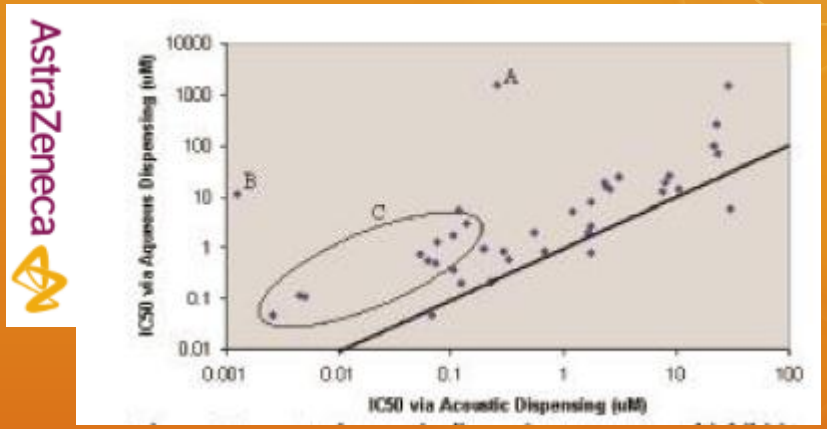




Images courtesy of Labcyte Inc.
http://goo.gl/K0Fjz

Tips vs. Acoustic analysis





Data appears randomly scattered. 24% had IC_{50} values > 3 fold weaker using tip-based dispensing. 8% produced no value using tip-based dispensing.

No analysis of molecule properties.

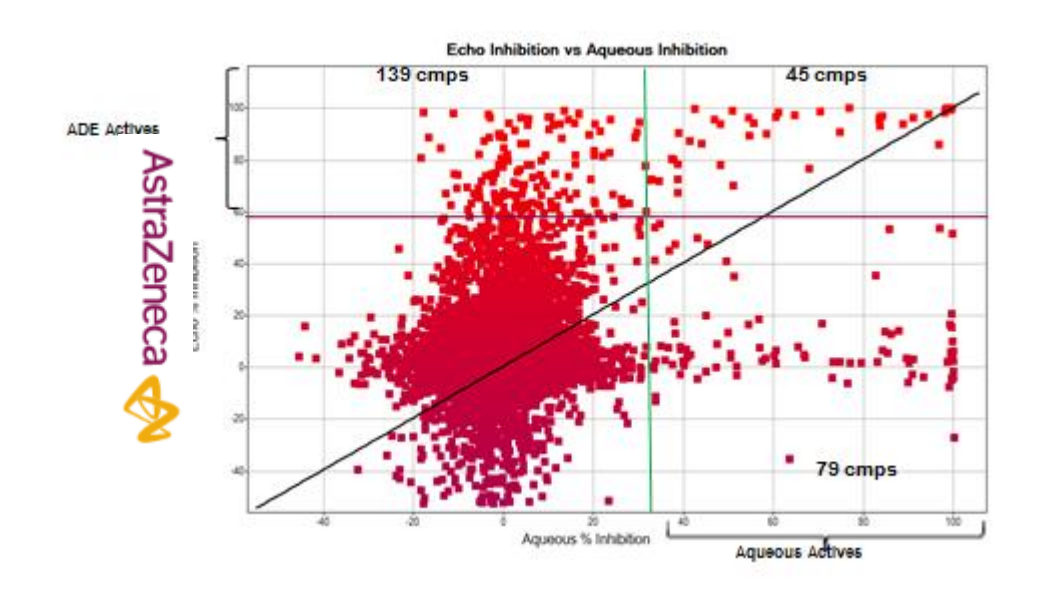
Spicer et al., In *Drug Discovery Technology*: Boston, 2005.

~ 40 12 point IC_{50} values. Compounds more active when using acoustic dispensing. Correlation in data is poor with many compounds showing >10 fold shift in potency depending on dispensing method.

No analysis of molecule properties.

Wingfield et al., *American Drug Discovery* **2008**, *3*, 24-30.

Tips vs. Acoustic analysis – large scale



Inhibition of tyrosine kinases at 10 μM for ~10,000 compounds.
False +ve from acoustic transfer (as measured by subsequent IC_{50} analyses) = 19% of hits.

False +ve from tip-based transfers = 55% of all hits.

60 more compounds were identified as active with acoustic transfer.

No analysis of molecule properties.

Wingfield, J. *Drug Discovery 2012*: Manchester, UK, 2012

Problems with biological data: **how** you dispense matters

Few structures and corresponding data are public

Using data from 2 AstraZeneca patents –

Tyrosine kinase EphB4 pharmacophores (Accelrys Discovery Studio) were developed using data for 14 compounds

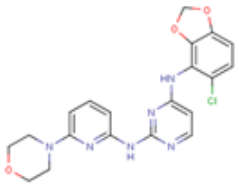
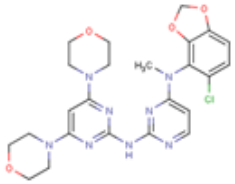
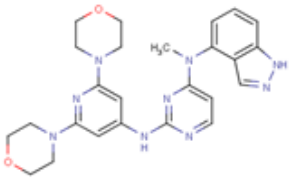
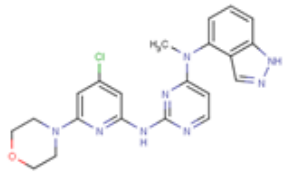
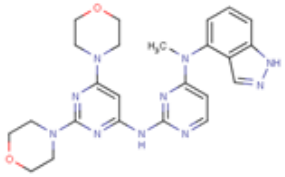
IC_{50} determined using different dispensing methods

Analyzed correlation with simple descriptors (SAS JMP)

Calculated LogP correlation with log IC_{50} data for acoustic dispensing ($r^2 = 0.34$, $p < 0.05$, $N = 14$)

Barlaam, B. C.; Ducray, R., WO 2009/010794 A1, **2009**
Barlaam, B. C.; Ducray, R.; Kettle, J. G., US 7,718,653 B2, **2010**

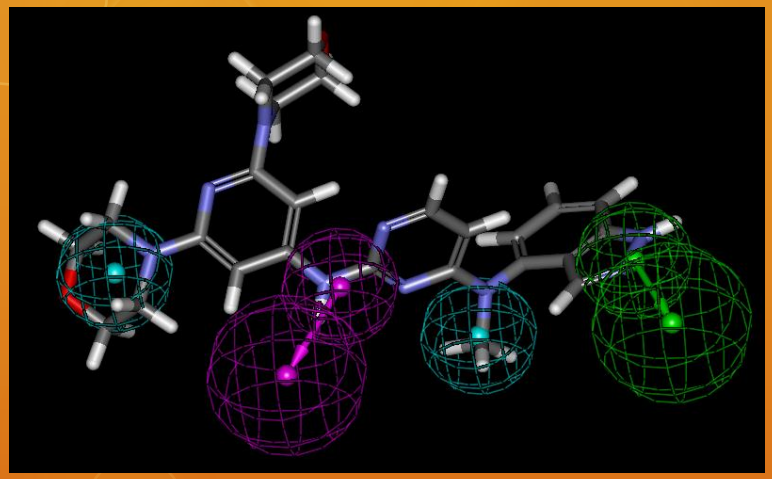
Examples of IC_{50} values produced via acoustic transfer with direct dilution vs those generated with tip-based transfer and serial dilutions

Compound number	Structure	IC_{50} Acoustic Dispensing (µM)	IC_{50} Tip-based Dispensing (µM)	Ratio Tip-Based / Acoustic dispensing
10		0.064	0.817	12.76
11		0.486	3.03	6.23
5		0.002	0.553	276.5
6		0.007	0.973	139
7		0.003	0.778	259.33

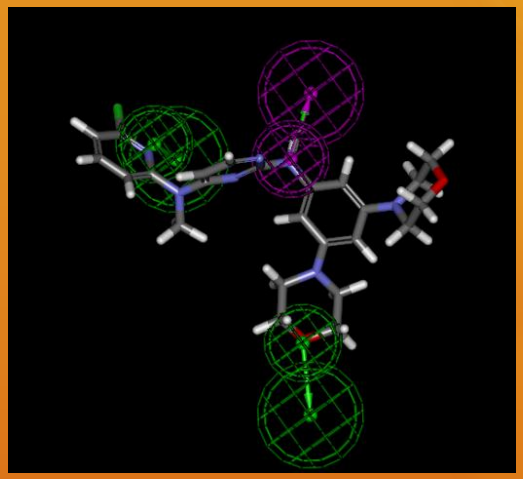
acoustically-derived IC_{50} values were 1.5 to 276.5-fold lower than for tip-based dispensing

Barlaam, B. C.; Ducray, R., WO 2009/010794 A1, **2009**
Barlaam, B. C.; Ducray, R.; Kettle, J. G., US 7,718,653 B2, **2010**

Tyrosine kinase EphB4 Pharmacophores



Acoustic



Disposable tip

Cyan = hydrophobic

Green = hydrogen bond acceptor

Purple = hydrogen bond donor

Each model shows most potent molecule mapping

	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
Disposable tip mediated process	0	2	1	0.80

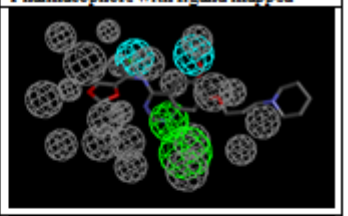
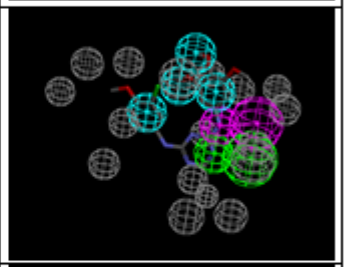
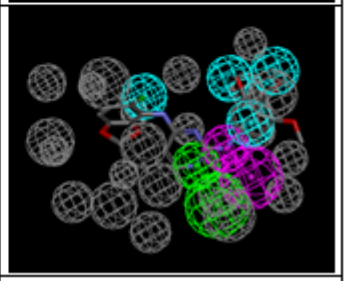
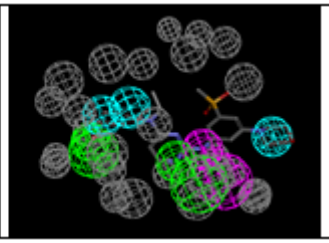
Ekins, Olechno and Williams, Submitted 2012

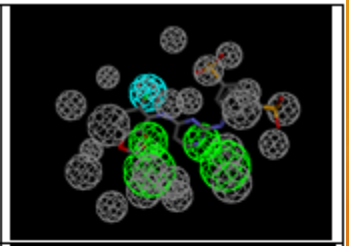
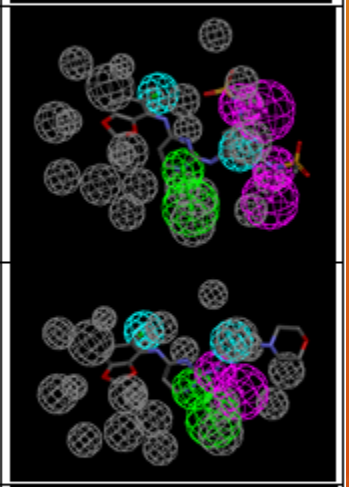
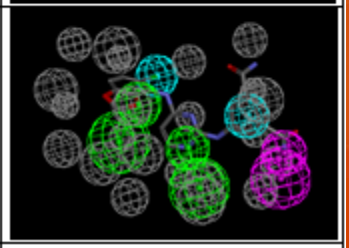
Test set evaluation of pharmacophores

- An additional 12 compounds from AstraZeneca
  Barlaam, B. C.; Ducray, R., WO 2008/132505 A1, **2008**
- 10 of these compounds had data for tip-based dispensing and 2 for acoustic dispensing
- Calculated LogP and logD showed low but statistically significant correlations with tip-based dispensing (r^2= 0.39 $p < 0.05$ and 0.24 $p < 0.05$, N = 36)
- Used as a test set for pharmacophores
- The two compounds analyzed with acoustic liquid handling were predicted in the top 3 using the acoustic pharmacophore
- The tip-based pharmacophore failed to rank the retrieved compounds correctly

Automated receptor-ligand pharmacophore generation method

Pharmacophores for the tyrosine kinase EphB4 generated from crystal structures in the protein data bank PDB using Discovery Studio version 3.5.5

PDB	Reference	Pharmacophore features	Pharmacophore with ligand mapped
2VWU	[2]	2H, 1HBA	
2VWV	[2]	4H, 1HBA, HBD	
2VWW	[2]	4H, 1HBA, 1HBD	
2VWX	[2]	No pharmacophore	
2XVD	[3]	3H, 2HBA, 1HBD	

2VWY	[2]	1H, 2HBA (2 molecules in structure)	
2VWZ	[2]	2H, 1HBA, 2HBD (2 molecules in structure – each maps to a HBD)	
2VXO	[2]	2H, 1HBA, 1HBD	
2VX1	[2]	2H, 2HBA, 1HBD	
2X9F	[4]	No Pharmacophore	

Cyan = hydrophobic

Green = hydrogen bond acceptor

Purple = hydrogen bond donor

Grey = excluded volumes

Each model shows most potent molecule mapping

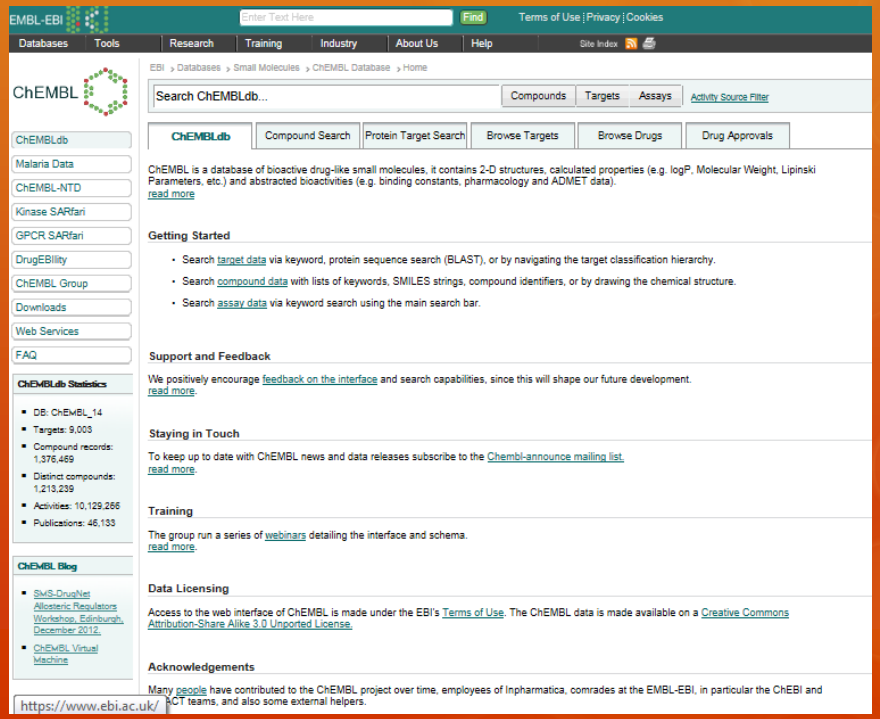
Bioorg Med Chem Lett **2010**, *20*, 6242-6245.
Bioorg Med Chem Lett **2008**, *18*, 5717-5721.
Bioorg Med Chem Lett **2008**, *18*, 2776-2780.
Bioorg Med Chem Lett **2011**, *21*, 2207-2211.

Summary

- In the absence of structural data, pharmacophores and other computational and statistical models are used to guide medicinal chemistry in early drug discovery.
- Our findings suggest non tip-based methods could improve HTS results and avoid the development of misleading computational models and statistical relationships.
- Automated pharmacophores are closer to pharmacophore generated with acoustic data – all have hydrophobic features – missing from tip based model
- **Importance of hydrophobicity seen with logP correlation and crystal structure interactions**
- Public databases should annotate this meta-data alongside biological data points, to create larger datasets for comparing different computational methods.

The stuff of nightmares?

- How much of the data in databases is generated by tip-based methods
- How much is erroneous
- Do we have to start again?
- How does it affect all subsequent science – data mining etc
- Does it impact Pharmas productivity?



Strengths and Weaknesses

Small dataset size – focused on one compounds series

No previous publication describing how data quality can be impacted by dispensing and how this in turn affects computational models and downstream decision making.

No comparison of pharmacophores generated from acoustic dispensing and tip-based dispensing.

No previous comparison of pharmacophores generated from *in vitro* data with pharmacophores automatically generated from X-ray crystal conformations of inhibitors.

Severely limited by number of structures in public domain with data in both systems