feature

When pharmaceutical companies publish large datasets: an abundance of riches or fool's gold?

Sean Ekins[1,2,3,4,*] and Antony J. Williams[5]

The recent announcement that GlaxoSmithKline have released a huge tranche of whole-cell malaria screening data to the public domain, accompanied by a corresponding publication, raises some issues for consideration before this exemplar instance becomes a trend. We have examined the data from a high level, by studying the molecular properties, and consider the various alerts presently in use by major pharma companies. We not only acknowledge the potential value of such data but also raise the issue of the actual value of such datasets released into the public domain. We also suggest approaches that could enhance the value of such datasets to the community and theoretically offer an immediate benefit to the search for leads for other neglected diseases.

Introduction

We are currently witnessing considerable shifts in the ways that pharmaceutical research can be accelerated. These approaches include decentralizing research and engagement with external research communities through crowdsourcing. There is a marked trend toward collaboration of all kinds [1–5]. In parallel, there is a renewed interest in neglected disease research (on malaria, tuberculosis [TB], kinetoplastids and so on [6]), owing to the notable influence of the US National Institutes of Health (NIH), foundations such as the Bill and Melinda Gates Foundation, The European Commission, and increasing investment from pharmaceutical companies and others [6,7]. In the past, drug companies generally published only small chunks of data in the form of molecular structures and biological (pharmacology) data when it was convenient, as needed to influence investors, the public and/or FDA approval, or after a compound or project was terminated. With the combinatorial chemistry and high-throughput screening that have seen explosive growth over the past decade, each large pharmaceutical company has created massive proprietary databases and invested enormous resources in the purchase or development of complex informatic platforms. These software systems probably contain more data than can be realistically mined because of the quest for blockbuster-targeted therapies. In addition, the quality of the data and the applicability of the assays can add a lot of noise into the system. Although it is not unusual for academics (and occasionally pharmaceutical companies) to publish and collate relatively large datasets (from several hundred to thousands of compounds), primarily for quantitative structure–activity analysis (http://www.cheminformatics.org/datasets/index.shtml; http://www.qsar-world.com/qsar-datasets.php) or compile datasets from NIH-funded screening programs (http://www.pdsp.med.unc.edu/indexR.html) [8–12], pharmaceutical companies have been less willing to make larger screening datasets available to the public, and even with such huge efforts, there has been little productivity in screening for antibiotics [13]. This has now changed with GlaxoSmithKline (GSK)'s unprecedented release of approximately 13,500 *in vitro* screening hits against malaria using *Plasmodium falciparum* along with their associated cytotoxicity (in HepG2 cells) data from an initial screen of more than two million compounds (see Ref. [14], which follows an earlier press release from GSK: http://www.nature.com/news/2010/100120/full/news.2010.20.html).

Hosted GSK malaria data

Three databases initially all hosted the data: the European Bioinformatics Institute–European

1359-6446/06/$ - see front matter © 2010 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2010.08.010

Molecular Biology Laboratory (ChEMBL, http://www.ebi.ac.uk/chembl/), PubChem (http://www.pubchem.ncbi.nlm.nih.gov/) and Collaborative Drug Discovery (CDD, http://www.collaborativedrug.com). What happens now the data are hosted and announced to the community is open to prediction: the malaria community might ignore it, which is highly unlikely, or the malaria community might be excited by the availability of the data, use the data in their research and, potentially, find a new drug. There is also, perhaps, a low probability of this, which will realistically take many years unless we find ways to accelerate the process. The question, therefore, is whether such data depositions are an abundance of riches or fool's gold. They might actually be neither.

This massive contribution of data to the community creates a precedent, and it is expected that other companies will follow suit (http://www.nature.com/news/2010/100120/full/news.2010.20.html). It also raises many other questions, which we can only begin to pose responses for. Who should get to host such public data? If the data are truly 'open' then anyone can download the data and reuse, repurpose and host the data. It will be interesting to see what 'licensing' is applied to the different types of data when they are exposed by various companies. The GSK malaria screening hits data were added to the ChEMBL dataset and the chemical structures are available for download. In the CDD database, data associated with public datasets are generally made available for similarity, substructure and Boolean searching (http://www.collaborativedrug.com/register). The SDF file of structures and data is also available for download. In the USA, PubChem has established itself as the *de facto* repository for screening data, so deposition here represents a drop in the ocean of more than 27 million unique molecule structures, although admittedly just a fraction are associated with bioactivity data. The deposition and hosting of the data in the three repositories (ChEMBL, PubChem and CDD) does leverage what each database has to offer in terms of integration to existing information on the compounds. We also believe that deposition into other databases with links out to others, including the original hosting organizations, also offers great benefits. We believe that a great model for this approach is also the ChemSpider database from the Royal Society of Chemistry (http://www.chemspider.com), which has already demonstrated the feasibility of such an approach and now also hosts the GSK data. We suggest there should be connectivity between all the databases hosting these data and others released in the future.

Questions and opportunities upon opening up data

We can predict that there might be increased competition to woo pharmaceutical companies to exclusively deposit data in various databases. There might, however, be more utility if the various databases collaborated to convince companies that releasing data could be a powerful force for change, with each group contributing their technologies and own expertise to the task. Instead of having standards for deposition, *a lá* microarray MIAME [15] and so on, there are no such standards for chemical structures and biological data. Who will ensure the quality of the data and act as a host for future annotation and potential structure validation? It is unlikely to be the databases clamoring to host the data. Would any organization be interested in funding future data uploads and data cleansing? Will companies only deposit compounds into the public domain that are of less interest to them and have been demonstrated to be inactive against other screens, while retaining drug-like hits and negative data (which can be valuable for SAR creation)? Although we applaud the pharmaceutical companies for donating data to the community, we question whether these contributions might create more noise and less signal in the public compound databases. Will people really want to mine the data if they are perceived as cast-off data or just represent commercially available screening libraries with little in the way of novel chemical entities? In many cases, we judge that most researchers will treat these data as of minor importance – in this case, because it is malaria-related. This might be a mistake, however, because GSK suggest that many of these compounds are acting on malaria kinases or proteases that could be important for the treatment of other diseases and that could lead to a drug for other blockbuster diseases. Will there be any incentives to pursue such neglected disease data? For the time being, pharmaceutical companies are investing more in such neglected diseases [6], although much less than in other major diseases.

This brings to mind the United States Environmental Agency's ToxCast project [16], which created a database, at major public expense, and then invited academics to build computational models or mine the data. Ultimately, what is derived from the data will be dependent on the quality of the data populated into the database. The only incentive is if something of interest is found; then, there might be some funding to investigate further. If pharmaceutical companies are to put data into the public domain, it might be of value for these or other organizations to consider incentivizing researchers to mine it or to offer challenges and awards.

GSK malaria screening data analysis

Driven purely by curiosity regarding the nature of the malaria data recently deposited into the public domain by GSK [14], we have undertaken a preliminary evaluation using a simple descriptor analysis, as was performed previously for some very large tuberculosis datasets originally funded by the NIH [17]. In addition, we have used some readily available substructure alerts or 'filters' to identify potentially reactive molecules. Pharmaceutical companies use such computational filters to clean up screening sets and remove undesirable molecules from vendor libraries [18]. Examples include filters from GSK [19] and Abbott [20–22]. An academic group also developed an extensive series of more than 400 substructural features for the removal of pan-assay interference compounds from screening libraries [23], which have yet to be integrated into a public resource. Our simplistic analysis compares the GSK dataset [14] to widely available drug-like molecules from the MicroSource US drugs dataset (http://www.msdiscovery.com/usdrugs.html).

As we would expect, the percentage of GSK malaria screening hit molecules failing the published 'GSK filters' [19] is close to zero, whereas the percentage failing the Pfizer and Abbott filters [20] is considerably higher (~57% and 76%, respectively) because these seem to be more conservative (Table 1). This could be interpreted as representing different business rule decisions instituted according to their own criteria, which we are not in a position to critically judge. The percentage of failures for the set of US FDA drugs is lower for both Pfizer and Abbott filters (Table 1), suggesting that these compounds are by no means perfect but perhaps setting a threshold. A recent study filtered a set of >1000 marketed drugs, and at least 26% failed filters for molecular features undesirable for high-throughput screening [24]. We have also recently used the same rules to filter sets of compounds with activity against TB [11,12], with 81–92% failing the Abbott filters [25], which might be related to mechanism of action. In the GSK paper, the authors suggest they did not find any non-specific inhibitors of lactate dehydrogenase, although cytotoxicity was seen in 1982 compounds [14]. A detailed analysis of our calculated molecular descriptors for the GSK malaria hits [14] shows that most are normally distributed, apart from the skewed Lipinski violations data and the bimodal molecular weight. Table 2 shows the means and standard deviations for each descriptor. Interest-

TABLE 1

Summary of SMARTS filter failures for various datasets[a]

Dataset (N)	Number failing the Abbott ALARM filters [20] (%)	Number failing Pfizer LINT filters[b] (%)	Number that failed Glaxo filters [19] (%)
GSK Malaria hits (13,355)	10,124 (75.8)	7683 (57.5)	129 (0.01)
MicroSource US FDA drugs (1041)	688 (66.1)	516 (49.6)	143 (13.7)

[a] The Abbott ALARM [20], Glaxo [19] and Blake SMARTS filter calculations were performed through the Smartsfilter web application, Division of Biocomputing, Dept. of Biochem & Mol Biology, University of New Mexico, Albuquerque, NM (http://www.pangolin.health.unm.edu/tomcat/biocomp/smartsfilter). The GSK malaria screening data were obtained [14] from the CDD database. We also used the MicroSource US Drugs dataset as a reference set of 'drug-like' molecules. Large datasets >1000 molecules were fragmented into smaller SDF files before running through this website.

[b] Originally provided as a Sybyl script to Tripos by James Blake (Array Biopharma) while at Pfizer.

TABLE 2

Mean (SD) of molecular descriptors from the CDD database for the GSK dataset[a]

Dataset	MW	log *P*	HBD	HBA	Lipinski rule of five alerts	pK_a	PSA	RBN
GSK data (*N* = 13,471)	478.16 (114.34)	4.53 (1.58)	1.83 (1.04)	5.60 (1.99)	0.82 (0.83)	6.67 (3.72)	76.85 (30.05)	7.17 (3.37)

Abbreviations: MW, molecular weight; HBD, hydrogen bond donor; HBA, hydrogen bond acceptor; PSA, polar surface area; RBN, rotatable bond number.

[a] Molecular properties were calculated using the Marvin plug-in (ChemAxon, Budapest, Hungary) within the CDD database.

ingly 3269 (24.3%) of the compounds fail more than one of the Lipinski rules of five (MW $\leq$ 500, log $P \leq 5$, HBD $\leq$ 5, HBA $\leq$ 10) [26] using the descriptors calculated in the CDD database. It should be noted that the performance of log *P* calculators differs between various software vendors and this should be taken into account when considering the calculated values. The GSK screening hits are generally large and very hydrophobic, as is also suggested in their publication [14], and although they suggested this might be important to reach intracellular targets, there is no discussion of the limitations of such compounds. These molecular properties might also present considerable solubility challenges [27]. These molecular properties can also be compared with the published lead-like rules (MW < 350, log $P < 3$, affinity $\sim$0.1 μM) [28,29], the natural product lead-like rules (MW < 460, log $P < 4.2$, Log Sol -5, RBN $\leq$ 10, rings $\leq$ 4, HBD $\leq$ 5, HBA $\leq$ 9) [30] and the mean molecular properties for the respiratory drugs that are delivered by either inhalation or intranasal routes (MW $\sim$370, PSA $\sim$89.2, log P $\sim$1.7) [31]. In summary, the malaria screening hits in total [14] might not be 'lead-like' and are closest to 'natural product lead-like'. Although the malaria paper [14] suggests that the compounds are 'drug-like' the evidence for this is weak (and is a statement open to wide interpretation) in the absence of comparison with drugs or even *in vivo* data for any of their hits. These antimalarial hits as a group are also vastly different to the mean molecular properties of compounds that have shown activity against TB, which – generally – are of a lower molecular weight, are less hydrophobic and have lower pK_a and fewer rotatable bond numbers [17]. The GSK compounds may be used with computational models to find active compounds that inhibit this disease *in vitro* [17]. We have taken Bayesian models generated from previously published very large screening datasets for TB and used them to filter the GSK malaria hits. The two separate models retrieved 4841 and 6030 compounds for the single-point and dose-response models, respectively (Supplementary data; also available at http://www.collaborativedrug.com), that would be predicted as active in the Mtb phenotypic screen from which the data were derived [11,12]. This might provide a starting point for cherry picking compounds for follow-up *in vitro*. It remains to be seen whether GSK will publish Mtb screening data for the same compounds, but we would encourage this.

Awareness of breaking the rules

Some companies try to avoid compounds that have reactive groups and fail related alerts or simply fail 'the rule of five' as a starting point before screening, although many chemists believe even this is too stringent; there are several successful drugs and development candidates that fall outside these requirements and 'break the rules'. Depending on the stringency of computational filtering, this would suggest a minimum of 24–76% of the GSK malaria screening hits would be filtered out or flagged as problematic using these simple methods. Structural alerts can be difficult to assess because the presence of a potentially problematic functionality must be assessed in the context of the molecule under consideration (e.g. is it required for the activity pharmacophore?). Our results would simply indicate that the user should be aware of the properties and features in the compounds in this dataset [14] before embarking upon lead optimization. It remains to be seen whether datasets deposited by other research groups will also follow this trend, and this is something we are currently assessing. Certainly, such computational approaches could be readily used without too much effort to assess large datasets deposited in the future.

Concluding remarks

Although attitudes on the quality of a compound vary between medicinal chemists, as a community we need to be vigilant regarding the data in any public or commercial database and this is ultimately the responsibility of the user. At one level, the structure fidelity should be the responsibility of the depositor but, as this study suggests, some readily available tools can be used to evaluate such datasets for chemical properties and can point to issues that need to be considered before a great deal of time is spent further mining the data. There are many who might be unaware of such technologies or their limitations, however, so how do they start processing such freely available data? There might be needles in this set of malaria hits and they might require computational approaches to pull them out of the haystack, but at the same time the user will need to be mindful of compound quality at the start to avoid blind alleys caused by aggregation [32], false positive issues [33–39] or artifacts [40]. Whoever funds such high-throughput screening – whether industry, government, not-for-profit or academia – needs to seriously consider which compounds are screened (and use appropriate filtering criteria beforehand) *in vitro* and then, ultimately, where and how the data are deposited. We sincerely hope that this provides a starting point for discussion of the many important issues raised above as pharmaceutical companies start depositing their datasets (whether compound screening, solubility, metabolism, toxicity and so on) in databases funded by non-profits or private concerns. It might then have accomplished something more than was originally envisaged by the company, with considerable positive implications. We have waited a long time for this type of landmark data contribution to the community and would like those pharmaceutical companies that follow GSK's exemplary lead to ensure they deliver even higher value and actionable data to the scientific community.

Potential conflicts of interest statement

S.E. consults for Collaborative Drug Discovery, Inc. on a Bill and Melinda Gates Foundation Grant #49852, 'Collaborative drug discovery for TB through a novel database of SAR data optimized to promote data archiving and sharing'. He is also on the advisory board for ChemSpider. A.J.W. is employed by the Royal Society of Chemistry, which owns ChemSpider and associated technologies.

Acknowledgements

The authors thank Jeremy Yang and colleagues (University of New Mexico) for kindly providing access to the Smartsfilter web application and our colleagues for inspiration and support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.drudis.2010.08.010.

References

1 Bingham, A. and Ekins, S. (2009) Competitive collaboration in the pharmaceutical and biotechnology industry. *Drug Discov. Today* 14, 1079–1081

2 Hunter, A.J. (2008) The innovative medicines initiative: a pre-competitive initiative to enhance the biomedical science base of Europe to expedite the development of new medicines for patients. *Drug Discov. Today* 13, 371–373

3 Barnes, M.R. *et al.* (2009) Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discov.* 8, 701–708

4 Bailey, D.S. and Zanders, E.D. (2008) Drug discovery in the era of Facebook – new tools for scientific networking. *Drug Discov. Today* 13, 863–868

5 Ekins, S. and Williams, A.J. (2010) Reaching out to collaborators: crowdsourcing for pharmaceutical research. *Pharm. Res.* 27, 393–395

6 Moran, M. *et al.* (2009) Neglected disease research and development: how much are we really spending? *PLoS Med.* 6, e30

7 Morel, C.M. *et al.* (2005) Health innovation networks to help developing countries address neglected diseases. *Science* 309, 401–404

8 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181

9 O'Connor, K.A. and Roth, B.L. (2005) Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.* 4, 1005–1014

10 Roth, B.L. *et al.* (2004) Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. *Pharmacol. Ther.* 102, 99–110

11 Maddry, J.A. *et al.* (2009) Antituberculosis activity of the molecular libraries screening center network library. *Tuberculosis (Edinb.)* 89, 354–363

12 Ananthan, S. *et al.* (2009) High-throughput screening for inhibitors of *Mycobacterium tuberculosis* H37Rv. *Tuberculosis (Edinb.)* 89, 334–353

13 Payne, D.J. *et al.* (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* 6, 29–40

14 Gamo, F.-J. *et al.* (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature* 465, 305–310

15 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371

16 Dix, D.J. *et al.* (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* 95, 5–12

17 Ekins, S. *et al.* (2010) A collaborative database and computational models for tuberculosis drug discovery. *Mol. BioSyst.* 6, 840–851

18 Williams, A.J. *et al.* (2009) Free online resources enabling crowdsourced drug discovery. *Drug Discovery World, Winter* 10, 33–39

19 Hann, M. *et al.* (1999) Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* 39, 897–902

20 Huth, J.R. *et al.* (2005) ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* 127, 217–224

21 Huth, J.R. *et al.* (2007) Toxicological evaluation of thiol-reactive compounds identified using a La assay to detect reactive molecules by nuclear magnetic resonance. *Chem. Res. Toxicol.* 20, 1752–1759

22 Metz, J.T. *et al.* (2007) Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput. Aided Mol. Des.* 21, 139–144

23 Baell, J.B. and Holloway, G.A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 53, 2719–2740

24 Axerio-Cilies, P. *et al.* (2009) Investigation of the incidence of "undesirable" molecular moieties for high-throughput screening compound libraries in marketed drug compounds. *Eur. J. Med. Chem.* 44, 1128–1134

25 Ekins, S. *et al.* (2010) Analysis and hit filtering of a very large library of compounds screened against *Mycobacterium tuberculosis*, *Mol. BioSyst.* In Press, doi:10.1039/c0mb00104j

26 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26

27 Lipinski, C.A. (2000) Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Method* 44, 235–249

28 Oprea, T.I. (2002) Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput. Aided Mol. Des.* 16, 325–334

29 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315

30 Rosen, J. *et al.* (2009) Novel chemical space exploration via natural products. *J. Med. Chem.* 52, 1953–1962

31 Ritchie, T.J. *et al.* (2009) Analysis of the calculated physicochemical properties of respiratory drugs: can we design for inhaled drugs yet? *J. Chem. Inf. Model.* 49, 1025–1032

32 Seidler, J. *et al.* (2003) Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* 46, 4477–4486

33 Rishton, G.M. (2008) Molecular diversity in the context of leadlikeness: compound properties that enable effective biochemical screening. *Curr. Opin. Chem. Biol.* 12, 340–351

34 Rishton, G.M. (2005) Failure and success in modern drug discovery: guiding principles in the establishment of high probability of success drug discovery organizations. *Med. Chem.* 1, 519–527

35 Oprea, T.I. *et al.* (2009) A crowdsourcing evaluation of the NIH chemical probes. *Nat. Chem. Biol.* 5, 441–447

36 Coan, K.E. and Shoichet, B.K. (2008) Stoichiometry and physical chemistry of promiscuous aggregate-based inhibitors. *J. Am. Chem. Soc.* 130, 9606–9612

37 Feng, B.Y. *et al.* (2007) A high-throughput screen for aggregation-based inhibition in a large compound library. *J. Med. Chem.* 50, 2385–2390

38 Jadhav, A. *et al.* (2010) Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* 53, 37–51

39 Doak, A.K. *et al.* (2010) Colloid formation by drugs in simulated intestinal fluid. *J. Med. Chem.* 53, 4259–4265

40 Schmidt, C. (2010) GSK/Sirtris compounds dogged by assay artifacts. *Nat. Biotechnol.* 28, 185–186

***Sean Ekins*[1,2,3,4,*]**
***Antony J. Williams*[5]**
[1]Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, PA 19046, USA
[2]Collaborative Drug Discovery,
1633 Bayshore Highway, Suite 342,
Burlingame, CA 94403, USA
[3]Department of Pharmaceutical Sciences,
University of Maryland, MD 21201, USA
[4]Department of Pharmacology,
University of Medicine & Dentistry of
New Jersey–Robert Wood Johnson
Medical School, 675 Hoes Lane, Piscataway,
NJ 08854, USA
[5]Royal Society of Chemistry,
904 Tamaras Circle, Wake Forest, NC 27587, USA

**Corresponding author*
ekinssean@yahoo.com,
sekins@collaborativedrug.com (S. Ekins)