Bigger data, collaborative tools and the future of predictive drug discovery

ARTICLE *in* JOURNAL OF COMPUTER-AIDED MOLECULAR DESIGN · JUNE 2014

Impact Factor: 2.99 · DOI: 10.1007/s10822-014-9762-y · Source: PubMed

Sean Ekins
Collaborations In Chemistry
289 PUBLICATIONS **8,116** CITATIONS

SEE PROFILE

Alex Michael Clark
Molecular Materials Informatics
40 PUBLICATIONS **486** CITATIONS

SEE PROFILE

Nadia Litterman
Harvard University
15 PUBLICATIONS **268** CITATIONS

SEE PROFILE

Antony John Williams
United States Environmental Protection Age…
370 PUBLICATIONS **3,130** CITATIONS

SEE PROFILE

Bigger Data, Collaborative Tools and the Future of Predictive Drug Discovery

Sean Ekins[†,‡*], Alex M. Clark[‖], S. Joshua Swamidass^, Nadia Litterman[‡] and Antony J. Williams[⊥]

[†] Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC 27526, USA.

[‡] Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA 94010, USA.

[‖] Molecular Materials Informatics, 1900 St. Jacques #302, Montreal, Quebec, Canada H3J 2S1

^ Division of Laboratory and Genomic Medicine, Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA.

[⊥] Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC-27587, U.S.A.

Running title: Bigger data, collaborations and predictions

Keywords: Cloud, Collaboration, Cheminformatics, Drug Discovery, Mobile Apps

ABSTRACT

Over the past decade we have seen a growth in the provision of chemistry data and cheminformatics tools as either free websites or software as a service (SaaS) commercial offerings. These have transformed how we find molecule-related data and use such tools in our research. There have also been efforts to improve collaboration between researchers either openly or through secure transactions using commercial tools. A major challenge in the future will be how such databases and software approaches handle larger data as it accumulates from high throughput screening and enable the user to draw insights, enable predictions and move projects forward. We now discuss how information from some datasets can be made more accessible and how privacy of data should not overwhelm the desire to share it at an appropriate time. We also discuss additional software tools that could be made available and our thoughts on the future of predictive drug discovery in this age of big data. We use some examples from our own research on neglected diseases, collaborations, mobile apps and algorithm development to illustrate these ideas.

INTRODUCTION

A lot can happen in a decade. We have gone from having few if any free resources such as databases of small molecules or software for drug discovery on the web, to literally thousands [1]. For example databases like ChemSpider [2; 3] have grown to house not just tens of millions of molecules but have become repositories for reactions and a vast treasure trove of data. Commercial offerings running software as a service (SaaS) ventures is nothing new, and being on a vendor-hosted cloud or an internal data center is not new either, even though companies continue to define their products by referring to it. What the pioneers of this approach do next will define the types of products we see created for drug discovery for years to come. As an example from our own experiences GeneGo (Thomson Reuters) was one of the earliest providers of drug discovery technologies (MetaDrug, MetaCore [4-10]) related to systems biology and integrated cheminformatics tools as SaaS. Collaborative Drug Discovery, Inc. (CDD) was likely the first to offer a private vault for storing chemistry and biology data as a multitenant SaaS [11]. Meanwhile, other larger software companies have acquired similar companies [12] to give them a presence ‘on the cloud’ and a collaborative software offering. Companies like CDD have built a business around the software and grants focused on using their technologies alongside other tools to advance research on neglected and other diseases [11; 13; 14].

Research collaborations are increasingly seen as being key to accelerating biomedical research and these are likely to be facilitated by computational methods [15]. However, we have suggested earlier that scientists are rarely collaborative or open with their data until publication or patenting, due to intellectual property (IP) concerns [13; 16]. Emerging collaborative software technologies allow researchers to specifically draw the line between pre-competitive and competitive areas and data more than previously possible. Scientific collaborations increasingly are becoming more important for the pharmaceutical industry. The industry has had to adapt by acquiring or partnering to bring in innovation or products [17] as well as outsourcing many aspects of R&D. At some point companies have to share their data whether this is during a collaboration, licensing, due diligence, pre-purchase or post purchase. Each of these processes has challenges when it comes to sharing molecular structures and associated data representing the IP for the companies or research groups involved. Increasingly such groups are involved in multi-organization collaborations such as public-private partnerships (PPP).

As an example of PPPs, CDD is involved in several collaborations such as More Medicines for Tuberculosis (MM4TB), Bill and Melinda Gates Foundation (BMGF) TB Accelerator and the NIH Blueprint for Neuroscience Research (BPN). In all of these initiatives, molecules and screening data with IP are securely shared between collaborators. Other similar software is likely required and available to share genomic and proteomic data but is outside the scope of this discussion. Currently data and associated molecules are selectively shared as a function of complex negotiations.

Often important information is then missing for the other groups involved which could help the global goals of the project. Ideally this could be shared too in a way that did not interfere with other projects, IP or relationships outside the scope of the current project of interest. There is also a growing need for public collaborations through initiatives that require open data [3; 18] though some of these may not truly be open themselves e.g. Open Source Drug Discovery [19]. Other European IMI funded PPP initiatives such as the European Lead Factory [20] is focused on high throughput screening (analogous to what the NIH has funded at its many screening centers) and another initiative, Elixir is a pan European infrastructure for biological information [21]. What cannot be denied is the growing mountain of data in the public domain, the growth in the need for collaboration to move projects rapidly and make sense of the accumulated information.

Bigger data and neglected diseases

A decade ago the amount of HTS data available was just a fraction of what it is now. The arrival of PubChem [22] and the mandate of publishing NIH funded experimental data into this database, has obviously had a big impact putting thousands of assays and millions of data points onto the internet. But for such data to be valuable it requires the underlying data be consistent, reliable and well-linked. The data also has to be of high quality as errors in structure can multiply from database to database [23; 24]. Then we can apply or build algorithms to mine the data, find patterns in it and help make well-informed decisions. Can we really call this ‘Big data’ though? It is all relative as one scientists’ big data is another’s small data. Relative to many nonscientific fields, what cheminformatics data lacks in size, it makes up for in terms of inconvenience and difficulty of handling. Perhaps we can just call this biomedical related data ‘bigger data’ compared to what we had access to in the past (e.g. tens to hundreds of compounds for quantitative structure activity relationships).

One area we are seeing larger amounts of screening data being useful and more accessible is in neglected disease research. These are a group of biologically unrelated diseases that are grouped together because they disproportionately affect marginalized populations, lack effective treatments or vaccines, or existing products are not accessible to the populations affected [25]. While the definition of a neglected disease varies, the category generally includes: tuberculosis, malaria, Chagas disease, African sleeping sickness, schistosomiasis, leishmaniasis and others for which there is a lack of economic incentives or “market” to provide motivation for product development [26-28]. Many of these pathogens, whether bacterial, parasitic, or viral, have complex life cycles and diverse approaches for evading the host immune system, rendering the development of new drugs and vaccines all the more challenging. Furthermore, these neglected diseases receive a relatively small amount of research investment ($80M to approximately $500M [29]) from governments and pharmaceutical companies in the developed world when we know it costs over a $1 billion to bring a drug to market [30]. The scientific challenges and limited funding available for neglected disease drug discovery and development highlight the importance of doing as much as possible with the data. These diseases are not seen as commercially viable next to major diseases, so many companies donate IP, patents and fund some limited research efforts and participate in PPPs. Currently available data relevant to Neglected Disease drug discovery is extremely diffuse, existing in an array of public or private databases (ChemSpider, PubChem, CDD, ChEMBL). One example is *Mycobacterium tuberculosis* (*Mtb*) which is the causative agent of tuberculosis (TB) that has infected approximately 2 billion people, and continues to kill 1.3 million people annually. We are seeing more companies making increasing quantities of data publically accessible, as well as the need to collaborate and share data as GlaxoSmithKline have made 177 compounds with *Mtb* activity [31] and 14,000 compounds with antimalarial activity [32] available. Surprisingly, we are making very slow progress in finding new therapeutics for TB as reviewed in recent years. Ideally we should be learning from the past efforts in TB drug discovery and yet we do not appear to be doing something that is simple yet effective, learning from the data that already exists. The current predominant method for identifying compounds active against *Mtb* is to use phenotypic high throughput screening (HTS) [33-36] and the hit rate of these screens tends to be in the low single digits. We can estimate that upwards of 5 million compounds have been screened against *Mtb* to date over the last 5-10 years [37]. There are around 1500 *Mtb* hits of interest from one laboratory alone [35; 36; 38; 39]. Leveraging this prior knowledge (by curating the data) to produce validated computational models is an approach that can be taken to improve screening efficiency both in terms of cost and relative hit rates. Machine learning and classification methods have been used in TB drug discovery [40], and have enabled rapid virtual screening of compound libraries for novel chemotypes [41; 42]. The use of cheminformatics for tuberculosis drug discovery has been summarized [43-45] and can be readily implemented early in the process as a means to limit the number of compounds needing to be screened and therefore saving time and money [46-50]. Recent publications in this area have hit rates >20% and focus on favorable compounds with low or no cytotoxicity [49; 50]. More recently combining datasets to use all 350,000 molecules with *in vitro* data from a single laboratory for computational models has been attempted. Interestingly our recent data suggests that smaller models with thousands of compounds may perform just as well as these "bigger data' models (manuscript in preparation). Throughout all of this work using the *Mtb* datasets over 5 years, we have shown how additional value can be generated from such published data. Similar cheminformatics approaches have also been applied to other diseases [51-54]. Computational methods result in cost savings by eliminating the need for some experiments or testing many hypotheses which would not normally be possible without such models. While there has been considerable screening and identification of hits, a possible bottleneck is the progression of compounds and expansion of structure activity relationships that could result in viable leads. To date we estimate ca. 2000 *in vitro Mtb* hits that need progressing. The *in vitro, in vivo* and clinical data for TB do not exist in a single database. Our own efforts to collate mouse *in vivo* data for modeling took many months and were recently described [37]. We see this lack of data coordination as a major limitation to progress. There is also no centralized organization for project management and minimal collaboration or coordination in the field. This suggests that even though we are drowning in data, actually a bigger challenge is the integration and analysis of it before ultimately being able to use it for predictive models and prospective testing. These observations may also be broadly applicable beyond *Mtb*, but illustrate what can be achieved with larger datasets.

Collaborative sharing of molecules and data

Do we take the importance of privacy concerns for our data too far? Should we think more carefully about what is the real high value data and perhaps loosen our belts and share more than we hoard data? Should we just find new ways to share data? For example we have already seen several companies compare their compound libraries to each other e.g. Bayer and Schering [55], Bayer and AstraZeneca [56] or in the case of Pfizer to the literature [57] using fingerprints, physicochemical properties and matching/similar compounds to show minimal overlap. While this is not the same as sharing molecules and their proprietary data on assays, many companies are involved in PPPs like those described earlier. What steps could be taken to increase the amount of data sharing?

Finding new ways to share relevant chemical information about screening data that leaves structures blinded could open the door for increased collaboration. These methods include better strategies for identifying active molecules from primary screens, which leverages information from fingerprints [58], scaffold groupings [59; 60], economic modeling [61-63], and improved processing of raw data [64-66]. They also include automatic methods of organizing screening data into workflows [67] and a series of approaches for visualizing how biological activity maps to chemical space [68-71]. Second, secure methods of sharing molecules and data could make outsourcing of chemical analysis possible (without sharing the structure itself). Outsourcing is increasingly important in drug discovery because it reduces the cost of many R&D

efforts and enables centralization of expertise [72-74]. Third, as more data is made available through these efforts it is possible unexpected connections and patterns in data can be identified that could have an impact on research. This is certainly impossible to predict, such as unexpected signals in screening data that indicate either specific molecules or mechanisms by which to treat human disease, or indications that might relate to adverse effects. As an example sharing large collections of proprietary assay data, with structures blinded, would enable researchers not part of the original data collection process to potentially improve how we do drug discovery. For example, a recent study used a small dataset published in patents from AstraZeneca, to show how different liquid dispensing methods can severely impact the IC_{50} data generated in high throughput screening and in turn impact the computational models that are built and decisions based on them [75]. A collaboration across multiple pharma's and academia could potentially address this on a much larger and more convincing scale, but it likely awaits secure sharing methods that do not reveal structures.

Nearly a decade ago there were attempts at securely sharing molecule related structure activity relationship data but these stalled when it was suggested that the proposed encryption methods were all fallible. For example a 2005 American Chemical Society meeting, co-chaired by Dr. Christopher Lipinski and Dr. Tudor Oprea included a session on securely sharing chemical information to support collaborative development of absorption, distribution, metabolism and excretion (ADME) predictors [76-86]. Swamidass *et al.,* recently proposed several approaches to the problem of sharing molecules securely [87] that may overcome the previous failings. First, they propose a totally new, secure method of sharing useful chemical information from small-molecule screens, without revealing the structures of the molecules [87]. The method generates scaffold networks for compounds, enabling sharing of: molecule IDs with assay data; how molecules in a screen are connected to one another in a screening network; how molecules are grouped together into scaffold groups; how these groups are connected into trees; how these groups are connected into networks and how molecules are connected together into R-group networks. Statistical analysis using the PubChem data also clearly demonstrated that scaffold networks do not convey enough information to reliably reveal chemical structure [87].

A second proposed approach from the same group uses a new, secure way of measuring the overlap between two private datasets. This method uses an algorithm to construct a private dataset's shareable summary, which is called a "cryptoset" [88]. The overlap between two private datasets can be estimated by comparing their cryptosets. At the same time, it is not possible to determine which specific items are in a private dataset from its cryptoset. Unlike other approaches to this problem [89-91], the item-level security arises from statistical properties of cryptosets rather than the secrecy of the algorithm or computational difficulty, so cryptosets can be shared in public, untrusted environments.

We are aware of at least one other company, MedChemica which has successfully developed a business model around technology closely related (but not identical) to what Swamidass *et al.,* are proposing above. They successfully negotiated agreements with three big pharma companies (AstraZeneca, Hoffman La Roche, Genentech) to share anonymized match-pair [92] data for the purpose of improving ADME optimization of lead compounds [93]. Their partners pay them to provide software to process the structures in internal ADME data into an anonymized form, very similar to the R-group networks described earlier. This anonymized data is then transferred to Medchemica, where it is analyzed, and specific rules to guide ADME optimization are extracted. These rules are then offered back to their clients to aid in lead optimization.

Such approaches like these for secure data sharing need to be integrated into software tools that are used by scientists to store their data to provide confidence when they do decide to share subsets of their data with different collaborators. This is becoming even more apparent as drug companies reach out increasingly to academics to fill the internal research gaps by externalizing their fundamental chemistry, biology and screening research efforts.

Predictive drug discovery

One of the challenges after high throughput screening is to learn as much as possible about the hits or potential probe compounds being developed. Are they cytotoxic? What liabilities do they have? What off-targets do they have? Could we predict as much as possible about the molecules before we invest more time and efforts in them? This obviously assumes that the computational models for absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) we use for particular properties are predictive and cover enough chemistry space. A major parameter to understand is drug metabolism which we now address in more detail.

The key issues in drug metabolism include identifying: the enzyme/s involved, the site/s of metabolism, the resulting metabolite/s and the rate of metabolism. Methods for predicting human drug metabolism from *in vitro* and computational methodologies and determining relationships between the structure and metabolic activity of molecules are also critically important for understanding potential drug interactions and toxicity. The cytochrome P450 (P450) enzymes are of considerable interest both in terms of metabolism and drug-drug interactions. Computational methodologies can be used for prioritization and uncovering the relationships between the structure and metabolic activity of novel molecules. A recent approach describes a method called XenoSite [94] for building models that predict CYP-mediated sites of metabolism (SOM) for drug-like molecules. The predictive accuracies of these models (average 87%) surpass the accuracies reported for other methods for nine distinct CYP substrate sets. This is achieved by (1) incorporating two types of molecule-level descriptors and (2) building predictive models using neural networks rather than SVMs. Neural networks were found to build models 4 times faster than SVMs when applied to the same set of SOMs encoded with the same set of descriptors. While this approach focused on phase I metabolism it is possible such approaches could be applied to phase II enzymes also.

Introducing such predictive approaches into software that store screening data or at least integrating with such tools may be important for creating a pipeline process, such that the likely enzymes involved in metabolism can be predicted for a compound. This may be very important for avoiding specific patient populations that are perhaps poor or extensive metabolizers of a drug which could present problems such as toxicity or lack of efficacy. Being able to provide information on this level for metabolism and other properties like toxicity [95] in software used for storing and sharing chemistry and biology data is likely to be of value in overall decision making.

For example there are already efforts like qsardb.org and ochem.eu which enable public model sharing and development [96; 97]. In addition websites such as Chembench provide models and tools for modeling to registered users [98]. Our earlier work proposed that open source descriptors and algorithms may be comparable with some commercial software, and that this might facilitate more sharing of computational models [99]. There have also been developments such as QSAR-ML which was developed to enable standards for interoperability of QSAR models [100; 101]. One could imagine that software for secure sharing of models could be carried out similarly to that described earlier for data, such they can be accessed by selected users. The current websites described above do not appear to offer this level of selectivity, and many companies may be wary of accessing them without some idea of security. Vendors that can guarantee that a companies' IP will be secure are likely to be more successful in getting big pharma and biotech to use and share models in this way. Some advantages of sharing models may be that a collaborator can benefit from models developed with your proprietary data, which in turn benefits your shared goals.

Sharing models openly with a community may foster addition of a groups own data to update them and make them more relevant to internal projects if indeed the data were generated under similar conditions. If one was sharing a model and you want to ensure that the user could not identify compounds in the training set, you might disable any features that would measure the distance, similarity etc. to compounds in the training set. It is likely that more work and discussion on model sharing and development will happen in future.

Future vision

We have previously suggested some of the needs and opportunities for cheminformatics which we termed the ‘missing pieces’ [102]. A decade ago commercial tools and academic tools were pretty much the only choices. In recent years we have seen a bigger effort towards open source cheminformatics software [101; 103-105]. Also a decade ago systems biology was piecing together small biology experiments such as protein-protein interactions to understand the “big picture” [106]. Now the amount of data available in some areas of biology (for many diseases or specific targets) is overwhelming. The challenges are to know where to look for the data you need in the first place. It may be feasible to turn this around and say that databases or datasources should be more proactive about making their data more accessible (or telling you what may be of interest). One way to do this is to use different avenues to create more value from the data.

Recently we have taken the approach we have called ‘appification’, that is to make a discrete molecule dataset available as a mobile app. This has become a common theme in the world of software, but is relatively new to structure-centric chemistry data. To our knowledge this was first achieved with the Green Solvents mobile app which used the ACS GCI Pharmaceutical Roundtable Solvent Selection Guide (a PDF). This document lists the 60 solvents by chemical name and rates the solvents against safety, health, air, water and waste categories with scores from 1 (few issues) to 10 (most concern) with additional color coding (green, yellow and red). This appification involved curation of the public data and development of a novel interface [107]. The limitations in access and utility of the original PDF encouraged us to recast the content in a novel manner to greatly enhance visibility and availability to practicing chemists. The data was also used to enable predictions for solvents outside the guide. A similar approach has also been taken with data on 800 molecules with known targets in TB [108; 109] to create TB Mobile. This data originated from a dataset in CDD public [14] but it was felt that the impact could be extended by creating a tool that could be useful for scientists and educators. The resultant app enables the user to view the molecules and known targets alongside other data related to the biology of the target. This represents one relatively simple way to bring cheminformatics and bioinformatics together. We have recently also implemented naïve Bayesian models using our own implementation of open source ECFP_6 descriptors in the app, to enable an alternative approach to target prediction as well as clustering molecules [110].

A further novel approach to creating open chemistry and biology databases can be achieved by building on tools we take for granted like Twitter and RSS feeds. A mobile app called open drug discovery teams (ODDT) [111] harvests Twitter feeds on several hashtags (e.g. #malaria, #tuberculosis, #huntingtons, #hivaids, #greenchemistry, #chagas, #leishmaniasis and #sanfilipposyndrome (and additional rare disease and other topics). This enables open data and molecules to be collected in an app. One could also think of this as a database with each topic being a subsection (e.g. a database on tuberculosis and a database on malaria etc.). The architecture of the currently deployed Open Drug Discovery Teams project is shown in (Figure 1). The cheminformatics framework that powers the molsync.com web service has been extended to include continual querying of Twitter and RSS feeds for relevant content, and collecting them in a database. We and others have tweeted in to these topics, links to molecules, data and papers. We then added the ability to endorse or reject tweets. In addition the ability to visualize a fingernail image for each tweet was added, as well as recognition of molecule images and a summary ticker tape. The ODDT app can now be used to manage multiple twitter accounts for the user too (Figure 2). The entry screen to the app displays the topics ranked by use. Tapping an image opens a topic on the incoming page and the content is listed on the right. Each tweet can be endorsed and the hyperlinks followed. The recent page shows entries with at least one endorsement while the content section shows the most popular voted content in rank order. Molecules can be tapped to open in other apps and could be the start of a workflow [112; 113]. If one imagines one of the hurdles to putting data in public databases is the upload of data files, ODDT represents a simplistic approach enabling true one click upload of molecules and data via a tweet! Perhaps this is an approach that could be used for secure upload via other messaging systems or direct messaging. It could also be an approach the bigger web-based databases could learn from.

From our experiences in neglected disease research we think there is an opportunity to bring together a range of data and tools (Figure 3) that would facilitate and catalyze the identification of novel therapeutic candidates by combining bioinformatics, cheminformatics data, publications, models and data visualization tools and curated *in vitro* and *in vivo* data. This would enable novel algorithms to be developed to infer candidate drug molecules, targets and mechanisms of drug action. This may in turn allow scientists to generate hypotheses in a single interface. The scientific challenges and limited funding available for neglected disease drug discovery and development highlight the importance of exploring alternative, lower cost approaches to advance drug discovery using cheminformatics and maximizing the data in the public domain.

Other challenges we see as opportunities are how to turn the databases and tools into assistants that make you aware of what data you might want to know about. For example, how can you find collaborators who might have interesting molecules or data? Methods like those described earlier for encrypting or sharing data securely might be valuable in this regard to help you find the data or alert you to its availability. Designing algorithms that can discern the most useful data for connecting researchers could reduce the serendipity involved in building collaborations [114]. Creating a tool that uses social networking features for serious applications such that the software users can “like” a molecule rather than a person might be appealing in some cases for finding researchers with orthogonal preliminary results. Such a system could hasten the pace of research and allow for the sharing of negative data, which is often not published.

Our laboratories (if we still have them) may be like our homes, that is an ‘internet of things’. Our databases and software tools should be able to talk wirelessly to devices such as analytical tools and auto upload data (which we term "no click upload"). Perhaps more likely all of our science will be outside our office. We can leverage CRO’s as well as other contractors via sources like Assay Depot [115] and Science Exchange [116], our personal connections and networks of collaborators can all do the science we need following our extensive mining of published data and predictions, perhaps even using virtual screening to decide which compounds to test.

How can we use the published data available to help tailor medicines to overcome our own genetic variability and side effects? For example, variability in metabolism is one issue, but what about variability in transporters and regulation of different proteins that can impact drug disposition? We are at a stage where there is increasing interest in computational models for human drug transporters which could be used proactively in the same way that we use models for P450s [117]. Such metabolism and transporter models should probably be used in parallel to profile compounds and predict liabilities, drug-drug and drug-transporter interactions.

Thinking about what is feasible by integrating data on diseases or at least making it available alongside tools to facilitate collaboration and drug discovery, one thinks of how non-scientists or non-specialists can leverage them also. For example can we bring non-scientists in to help us develop ‘outside the box’ thinking to tackle tough problems, whether in design of molecules, or biological problems to help cure rare diseases [118; 119]. We need to think about developing new tools that leverage the crowd (Box 1).

In summary, the role of collaboration and tools to enable data sharing in drug discovery are likely to continue in their importance. And therefore some of the developments we propose in enabling secure or encrypted sharing methods may be important to consider. As databases are integrated or linked together how we handle and license the data will be key, and some simple rules have already been proposed [120]. The mountain of data available across databases that are either public or private will undoubtedly continue to grow, and this will present challenges we will need to overcome in order to manipulate, mine and model it. We will need some creativity to develop new visualization paradigms that enable insights and lead to the next experiment. On the other hand, as mobile devices continue to expand their utility, useful tools and abilities to interact with data are possible as are extended workflows. While such devices may not be able to handle massive datasets within them just yet, they do present an access point to databases and more powerful tools on the cloud. The utility of being able to take your data with you and explore it on a tablet has some advantages. As we have shown, mobile devices also represent a way to prototype how we can use published data and cheminformatics tools in new ways. The future may not look like the past, we may actually now be able to make cheminformatics more accessible to the masses as it is essential to turn our accumulated data into something of real value that leads to biomedical advances. Our efforts in applying these approaches to neglected diseases, being just one example. That impact of of cheminformatics in itself is an accomplishment that is worthy of more support whether governmental or otherwise.

ACKNOWLEDGMENTS

S.E. gratefully acknowledges colleagues at CDD, Dr. Joel S. Freundlich (Rutgers), Dr. Malabika Sarker (SRI) and Dr. Katalin Nadassy (Accelrys) for valuable discussions and assistance in developing some of the projects discussed. S.E. acknowledges that the Bayesian models were developed with support from Award Number R43 LM011152-01 "Biocomputation across distributed private datasets to enhance drug discovery" from the National Library of Medicine. TB Mobile and the associated dataset was developed with funding from Award Number 2R42AI088893-02 "Identification of novel therapeutics for tuberculosis combining cheminformatics, diverse databases and logic based pathway analysis" from the National Institutes of Allergy and Infectious Diseases.

Competing Financial Interests

NL is an employee and SE is a consultant for CDD Inc. SE is on the advisory board for Assay Depot. AJW is an employee of the Royal Society of Chemistry. AMC is an employee of Molecular Materials Informatics and a consultant for CDD.

References

1. Villoutreix BO, Lagorce D, Labbe CM, Sperandio O, Miteva MA, (2013) One hundred thousand mouse clicks down the road: selected online resources supporting drug discovery collected over a decade. Drug Discov Today, 18: 1081-1089.

2. Pence HE, Williams AJ, (2010) ChemSpider: An Online Chemical Information Resource. J Chem Educ, 87: 1123-1124.

3. Ekins S, Williams AJ, (2010) Precompetitive Preclinical ADME/Tox Data: Set It Free on The Web to Facilitate Computational Model Building to Assist Drug Development. Lab on a Chip, 10: 13-22.

4. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Bugrim A, Nikolskaya T, (2005) Computational prediction of human drug metabolism. Expert Opin Drug Metab Toxicol, 1: 303-324.

5. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Sorokina S, Bugrim A, Nikolskaya T, (2006) A Combined Approach to Drug Metabolism and Toxicity Assessment. Drug Metab Dispos, 34: 495-503.

6. Ekins S, Bugrim A, Brovold L, Kirillov E, Nikolsky Y, Rakhmatulin EA, Sorokina S, Ryabov A, Serebryiskaya T, Melnikov A, Metz J, Nikolskaya T, (2006) Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. Xenobiotica, 36: 877-901.

7. Ekins S, Kirillov E, Rakhmatulin EA, Nikolskaya T, (2005) A Novel Method for Visualizing Nuclear Hormone Receptor Networks Relevant to Drug Metabolism. Drug Metab Dispos, 33: 474-481.

8. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T, (2007) Pathway mapping tools for analysis of high content data. Methods Mol Biol, 356: 319-350.

9. Embrechts MJ, Ekins S, (2007) Classification of Metabolites with Kernel-Partial Least Squares (K-PLS). Drug Metab Dispos, 35: 325-327.

10. Stranz DD, Miao S, Campbell S, Maydwell G, Ekins S, (2008) Combined computational metabolite prediction and automated structure-based analysis of mass spectrometric data. Toxicol Mech Methods, 18: 243-250.

11. Hohman M, Gregory K, Chibale K, Smith PJ, Ekins S, Bunin B, (2009) Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. Drug Disc Today, 14: 261-270.

12. Bost F, Jacobs RT, Kowalczyk P, (2010) Informatics for neglected diseases collaborations. Curr Opin Drug Discov Devel, 13: 286-296.

13. Bunin BA, Ekins S, (2011) Alternative business models for drug discovery. Drug Disc Today, 16: 643-645.

14. Sarker M, Talcott C, Madrid P, Chopra S, Bunin BA, Lamichhane G, Freundlich JS, Ekins S, (2012) Combining cheminformatics methods and pathway analysis to identify molecules with whole-cell activity against Mycobacterium tuberculosis. Pharm Res, 29: 2115-2127.

15. Ekins S, Hupcey MAZ, Williams AJ, 2011 Collaborative computational technologies for biomedical research, Wiley, Hoboken, NJ.

16. Ekins S, Hohman M, Bunin BA, In S. Ekins, M.A.Z. Hupcey and A.J. Williams, (Eds.),2011 Collaborative Computational Technologies for Biomedical Research, Wiley and Sons, Hoboken.

17. Burrill GS, In, 2010 4th Annual CDD Community Meeting, San Francisco.

18. Todd MH, (2007) Open access and open source in chemistry. Chem Cent J, 1: 3.

19. Ardal C, Rottingen JA, (2012) Open source drug discovery in practice: a case study. PLoS Negl Trop Dis, 6: e1827.

20. Anon, In.

21. Anon, (Elixir.

22. Li Q, Cheng T, Wang Y, Bryant SH, (2010) PubChem as a public resource for drug discovery. Drug Discov Today, 15: 1052-1057.

23. Williams AJ, Ekins S, (2011) A quality alert and call for improved curation of public chemistry databases. Drug Disc Today, 16: 747-750.

24. Williams AJ, Ekins S, Tkachenko V, (2012) Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. Drug Disc Today, 17: 685-701.

25. Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, Savioli L, (2007) Control of neglected tropical diseases. N Engl J Med, 357: 1018-1027.

26. Guiguemde WA, Shelat AA, Bouck D, Duffy S, Crowther GJ, Davis PH, Smithson DC, Connelly M, Clark J, Zhu F, Jimenez-Diaz MB, Martinez MS, Wilson EB, Tripathi AK, Gut J, Sharlow ER, Bathurst I, El Mazouni F, Fowble JW, Forquer I, McGinley PL, Castro S, Angulo-Barturen I, Ferrer S, Rosenthal PJ, Derisi JL, Sullivan DJ, Lazo JS, Roos DS, Riscoe MK, Phillips MA, Rathod PK, Van Voorhis WC, Avery VM, Guy RK, (2010) Chemical genetics of Plasmodium falciparum. Nature, 465: 311-315.

27. Ribeiro I, Sevcsik AM, Alves F, Diap G, Don R, Harhay MO, Chang S, Pecoul B, (2009) New, improved treatments for Chagas disease: from the R&D pipeline to the patients. PLoS Negl Trop Dis, 3: e484.

28. Bettiol E, Samanovic M, Murkin AS, Raper J, Buckner F, Rodriguez A, (2009) Identification of three classes of heteroaromatic compounds with activity against intracellular Trypanosoma cruzi by chemical library screening. PLoS Negl Trop Dis, 3: e384.

29. Ponder EL, Freundlich JS, Sarker M, Ekins S, (2014) Computational Models for Neglected Diseases: Gaps and Opportunities. Pharm Res, 31: 271-277.

30. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL, (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov, 9: 203-214.

31. Ballell L, Bates RH, Young RJ, Alvarez-Gomez D, Alvarez-Ruiz E, Barroso V, Blanco D, Crespo B, Escribano J, Gonzalez R, Lozano S, Huss S, Santos-Villarejo A, Martin-Plaza JJ, Mendoza A, Rebollo-Lopez MJ, Remuinan-Blanco M, Lavandera JL, Perez-Herran E, Gamo-Benito FJ, Garcia-Bustos JF, Barros D, Castro JP, Cammack N, (2013) Fueling Open-Source Drug Discovery: 177 Small-Molecule Leads against Tuberculosis. ChemMedChem, 8: 313-321.

32. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF, (2010) Thousands of chemical starting points for antimalarial lead identification. Nature, 465: 305-310.

33. Ballell L, Field RA, Duncan K, Young RJ, (2005) New small-molecule synthetic antimycobacterials. Antimicrob Agents Chemother, 49: 2153-2163.

34. Reynolds RC, Ananthan S, Faaleolea E, Hobrath JV, Kwong CD, Maddox C, Rasmussen L, Sosa MI, Thammasuvimol E, White EL, Zhang W, Secrist JA, 3rd, (2011) High throughput screening of a library based on kinase inhibitor scaffolds against Mycobacterium tuberculosis H37Rv. Tuberculosis (Edinb)

35. Maddry JA, Ananthan S, Goldman RC, Hobrath JV, Kwong CD, Maddox C, Rasmussen L, Reynolds RC, Secrist JA, 3rd, Sosa MI, White EL, Zhang W, (2009)

Antituberculosis activity of the molecular libraries screening center network library. Tuberculosis (Edinb), 89: 354-363.

36. Ananthan S, Faaleolea ER, Goldman RC, Hobrath JV, Kwong CD, Laughon BE, Maddry JA, Mehta A, Rasmussen L, Reynolds RC, Secrist JA, 3rd, Shindo N, Showe DN, Sosa MI, Suling WJ, White EL, (2009) High-throughput screening for inhibitors of Mycobacterium tuberculosis H37Rv. Tuberculosis (Edinb), 89: 334-353.

37. Ekins S, Pottorf R, Reynolds RC, Williams AJ, Clark AM, Freundlich JS, (2014) Looking Back To The Future: Predicting In vivo Efficacy of Small Molecules Versus Mycobacterium tuberculosis. J Chem Inf Model, 54: 1070-1082.

38. Ekins S, Freundlich JS, Hobrath JV, White EL, Reynolds RC, (2014) Combining Computational Methods for Hit to Lead Optimization in Mycobacterium tuberculosis Drug Discovery. Pharm Res, 31: 414-435.

39. Reynolds RC, Ananthan S, Faaleolea E, Hobrath JV, Kwong CD, Maddox C, Rasmussen L, Sosa MI, Thammasuvimol E, White EL, Zhang W, Secrist JA, 3rd, (2012) High throughput screening of a library based on kinase inhibitor scaffolds against Mycobacterium tuberculosis H37Rv. Tuberculosis (Edinb), 92: 72-83.

40. Prakash O, Ghosh I, (2006) Developing an antituberculosis compounds database and data mining in the search of a motif responsible for the activity of a diverse class of antituberculosis agents. J Chem Inf Model, 46: 17-23.

41. Garcia-Garcia A, Galvez J, de Julian-Ortiz JV, Garcia-Domenech R, Munoz C, Guna R, Borras R, (2005) Search of chemical scaffolds for novel antituberculosis agents. J Biomol Screen, 10: 206-214.

42. Planche AS, Scotti MT, Lopez AG, de Paulo Emerenciano V, Perez EM, Uriarte E, (2009) Design of novel antituberculosis compounds using graph-theoretical and substructural approaches. Mol Divers, 13: 445-458.

43. Sundaramurthi JC, Brindha S, Reddy TB, Hanna LE, (2012) Informatics resources for tuberculosis--towards drug discovery. Tuberculosis (Edinb), 92: 133-138.

44. Ekins S, Freundlich JS, Choi I, Sarker M, Talcott C, (2011) Computational Databases, Pathway and Cheminformatics Tools for Tuberculosis Drug Discovery. Trends in Microbiology, 19: 65-74.

45. Ekins S, Freundlich JS, (2013) Computational models for tuberculosis drug discovery. Methods Mol Biol, 993: 245-262.

46. Ekins S, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Hohman M, Bunin B, (2010) A Collaborative Database And Computational Models For Tuberculosis Drug Discovery. Mol BioSystems, 6: 840-851.

47. Ekins S, Freundlich JS, (2011) Validating new tuberculosis computational models with public whole cell screening aerobic activity datasets Pharm Res 28: 1859-1869.

48. Ekins S, Kaneko T, Lipinksi CA, Bradford J, Dole K, Spektor A, Gregory K, Blondeau D, Ernst S, Yang J, Goncharoff N, Hohman M, Bunin B, (2010) Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis Mol BioSyst, 6: 2316-2324.

49. Ekins S, Reynolds RC, Franzblau SG, Wan B, Freundlich JS, Bunin BA, (2013) Enhancing Hit Identification in Mycobacterium tuberculosis Drug Discovery Using Validated Dual-Event Bayesian Models PLOSONE, 8: e63240.

50. Ekins S, Reynolds R, Kim H, Koo M-S, Ekonomidis M, Talaue M, Paget SD, Woolhiser LK, Lenaerts AJ, Bunin BA, Connell N, Freundlich JS, (2013) Bayesian Models Leveraging Bioactivity and Cytotoxicity Information for Drug Discovery. Chem Biol, 20: 370-378.

51. Anderson JW, Sarantakis D, Terpinski J, Kumar TR, Tsai HC, Kuo M, Ager AL, Jacobs WR, Jr., Schiehser GA, Ekins S, Sacchettini JC, Jacobus DP, Fidock DA, Freundlich JS, (2012) Novel diaryl ureas with efficacy in a mouse model of malaria. Bioorg Med Chem Lett, 23: 1022-1025.

52. Alvarez G, Martinez J, Aguirre-Lopez B, Cabrera N, Perez-Diaz L, Gomez-Puyou MT, Gomez-Puyou A, Perez-Montfort R, Garat B, Merlino A, Gonzalez M, Cerecetto H, (2012) New chemotypes as Trypanosoma cruzi triosephosphate isomerase inhibitors: a deeper insight into the mechanism of inhibition. J Enzyme Inhib Med Chem

53. Pires DE, de Melo-Minardi RC, da Silveira CH, Campos FF, Meira W, Jr., (2013) aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics, 29: 855-861.

54. Gunatilleke SS, Calvet CM, Johnston JB, Chen CK, Erenburg G, Gut J, Engel JC, Ang KK, Mulvaney J, Chen S, Arkin MR, McKerrow JH, Podust LM, (2012) Diverse inhibitor chemotypes targeting Trypanosoma cruzi CYP51. PLoS Negl Trop Dis, 6: e1736.

55. Schamberger J, Grimm M, Steinmeyer A, Hillisch A, (2011) Rendezvous in chemical space? Comparing the small molecule compound libraries of Bayer and Schering. Drug Discov Today, 16: 636-641.

56. Kogej T, Blomberg N, Greasley PJ, Mundt S, Vainio MJ, Schamberger J, Schmidt G, Huser J, (2012) Big pharma screening collections: more of the same or unique libraries? The AstraZeneca-Bayer Pharma AG case. Drug Discov Today

57. Tu M, Rai BK, Mathiowetz AM, Didiuk M, Pfefferkorn JA, Guzman-Perez A, Benbow J, Guimaraes CR, Mente S, Hayward MM, Liras S, (2012) Exploring aromatic

chemical space with NEAT: novel and electronically equivalent aromatic template. J Chem Inf Model, 52: 1114-1123.

58. Posner BA, Xi H, Mills JE, (2009) Enhanced HTS hit selection via a local hit rate analysis. J Chem Inf Model, 49: 2202-2210.

59. Gunter B, Brideau C, Pikounis B, Liaw A, (2003) Statistical and graphical methods for quality control determination of high-throughput screening data. J Biomol Screen, 8: 624-633.

60. Varin T, Gubler H, Parker CN, Zhang JH, Raman P, Ertl P, Schuffenhauer A, (2010) Compound set enrichment: a novel approach to analysis of primary HTS data. J Chem Inf Model, 50: 2067-2078.

61. Swamidass SJ, Calhoun BT, Bittker JA, Bodycombe NE, Clemons PA, (2011) Enhancing the rate of scaffold discovery with diversity-oriented prioritization. Bioinformatics, 27: 2271-2278.

62. Swamidass SJ, Calhoun BT, Bittker JA, Bodycombe NE, Clemons PA, (2011) Utility-aware screening with clique-oriented prioritization. J Chem Inf Model, 52: 29-37.

63. Swamidass SJ, (2013) Using economic optimization to design high-throughput screens. Future Med Chem, 5: 9-11.

64. Makarenkov V, Kevorkov D, Zentilli P, Gagarin A, Malo N, Nadon R, (2006) HTS-Corrector: software for the statistical analysis and correction of experimental high-throughput screening data. Bioinformatics, 22: 1408-1409.

65. Makarenkov V, Zentilli P, Kevorkov D, Gagarin A, Malo N, Nadon R, (2007) An efficient method for the detection and elimination of systematic error in high-throughput screening. Bioinformatics, 23: 1648-1657.

66. Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinski HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons PA, (2008) ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res, 36: D351-359.

67. Calhoun BT, Browning MR, Chen BR, Bittker JA, Swamidass SJ, (2012) Automatically detecting workflows in PubChem. J Biomol Screen, 17: 1071-1079.

68. Browning MR, Calhoun BT, Swamidass SJ, (2013) Managing missing measurements in small-molecule screens. J Comput Aided Mol Des

69. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H, (2007) The scaffold tree--visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model, 47: 47-58.

70. Dimova D, Wawer M, Wassermann AM, Bajorath J, (2011) Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. J Chem Inf Model, 51: 258-266.

71. Wassermann AM, Bajorath J, (2012) Directed R-group combination graph: a methodology to uncover structure-activity relationship patterns in a series of analogues. J Med Chem, 55: 1215-1226.

72. Howells J, Gagliardi D, Malik K, (2012) Sourcing knowledge: R&D outsourcing in UK pharmaceuticals. Int J Tech Man, 59: 139-161.

73. Fox S, Farr-Jones S, Sopchak L, Boggs A, Nicely HW, Khoury R, Biros M, (2006) High-throughput screening: update on practices and success. J Biomol Screen, 11: 864-869.

74. McGee, (2012) Outsourcing and contract services. J Biomol Screen, 17: 1379-1381.

75. Ekins S, Olechno J, Williams AJ, (2013) Dispensing processes impact apparent biological activity as determined by computational and statistical analyses. PLoS One, 8: e62325.

76. Bradley D, (2005) Share and share alike. Nat Rev Drug Discov, 4: 180.

77. Masek BB, Shen L, Smith KM, Pearlman RS, (2008) Sharing chemical information without sharing chemical structure. J Chem Inf Model, 48: 256-261.

78. Balaban A, (2005) Can topological indices transmit information on properties but not on structures? J Comp Aided Mol Des, 19: 651-660.

79. Bologa C, Allu TK, Olah M, Kappler MA, Oprea TI, (2005) Descriptor collision and confusion: toward the design of descriptors to mask chemical structures. J Comput Aided Mol Des, 19: 625-635.

80. Clement OO, Guner OF, (2005) Possibilities for transfer of relevant data without revealing structural information. J Comput Aided Mol Des, 19: 731-738.

81. Filimonov D, Poroikov V, (2005) Why relevant chemical information cannot be exchanged without disclosing structures. J Comput Aided Mol Des, 19: 705-713.

82. Kaiser D, Zdrazil B, Ecker GF, (2005) Similarity-based descriptors (SIBAR)--a tool for safe exchange of chemical information? J Comput Aided Mol Des, 19: 687-692.

83. Trepalin S, Osadchiy N, (2005) The centroidal algorithm in molecular similarity and diversity calculations on confidential datasets. J Comput Aided Mol Des, 19: 715-729.

84. Tetko IV, Abagyan R, Oprea TI, (2005) Surrogate data--a secure way to share corporate data. J Comput Aided Mol Des, 19: 749-764.

85. Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP, (2005) Secure analysis of distributed chemical databases without data integration. J Comput Aided Mol Des, 19: 739-747.

86. Faulon JL, Brown WM, Martin S, (2005) Reverse engineering chemical structures from molecular descriptors: how many solutions? J Comput Aided Mol Des, 19: 637-650.

87. Matlock M, Swamidass SJ, (2014) Sharing Chemical Relationships Does Not Reveal Structures. J Chem Inf Model, 54: 37–48.

88. Swamidass SJ, Matlock M, Rozenblit L, (2013) When should we share? Securely measuring he overlap between private datasets. Submitted

89. Johnson SB, Whitney G, McAuliffe M, Wang H, McCreedy E, Rozenblit L, Evans CC, (2010) Using global unique identifiers to link autism collections J Am Med Inform Assoc, 17: 689-695.

90. Kuzu M, Kantacioglu M, Durham EA, Toth C, Malin B, (2013) A practical approach to achieve private medical record linkage in light of public resources J Am Med Inform Assoc, 20: 285-292.

91. Huang Y, Shen C, Evans D, Katz J, Shelat A, 2011 Information Systems Security, Springer.

92. Warner DJ, Griffen EJ, St-Gallay SA, (2010) WizePairZ: a novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. J Chem Inf Model, 50: 1350-1357.

93. Anon, In, 2013.

94. Zaretzki J, Matlock M, Swamidass SJ, (2013) XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. J Chem Inf Model, 53: 3373-3383.

95. Ekins S, (2014) Progress in computational toxicology. J Pharmacol Toxicol Methods, 69: 115-140.

96. Aruoja V, Moosus M, Kahru A, Sihtmae M, Maran U, (2014) Measurement of baseline toxicity and QSAR analysis of 50 non-polar and 58 polar narcotic chemicals for the alga Pseudokirchneriella subcapitata. Chemosphere, 96: 23-32.

97. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin, II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV, (2011) Online chemical modeling environment (OCHEM): web platform for data

storage, model development and publishing of chemical information. J Comput Aided Mol Des, 25: 533-554.

98. Walker T, Grulke CM, Pozefsky D, Tropsha A, (2010) Chembench: a cheminformatics workbench. Bioinformatics, 26: 3000-3001.

99. Gupta RR, Gifford EM, Liston T, Waller CL, Bunin B, Ekins S, (2010) Using open source computational tools for predicting human metabolic stability and additional ADME/TOX properties. Drug Metab Dispos, 38: 2083-2090.

100. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JE, (2010) Towards interoperable and reproducible QSAR analyses: Exchange of datasets. J Cheminform, 2: 5.

101. Williams AJ, Ekins S, Spjuth O, Willighagen EL, (2012) Accessing, using, and creating chemical property databases for computational toxicology modeling. Methods Mol Biol, 929: 221-241.

102. Ekins S, Gupta RR, Gifford E, Bunin BA, Waller CL, (2010) Chemical space: missing pieces in cheminformatics. Pharm Res, 27: 2035-2039.

103. Guha R, Spjuth O, Willighagen EL, In S. Ekins, M.A.Z. Hupcey and A.J. Williams, (Eds.),2011 Collaborative computational technologies for biomedical research, Wiley and Sons, Hoboken, pp. 399-422.

104. Spjuth O, Carlsson L, Alvarsson J, Georgiev V, Willighagen E, Eklund M, (2012) Open source drug discovery with bioclipse. Curr Top Med Chem, 12: 1980-1986.

105. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B, (2012) Open PHACTS: Semantic interoperability for drug discovery. Drug Disc Today, In press:

106. Ekins S, Bugrim A, Nikolsky Y, Nikolskaya T, In S. Gad, (Ed.),2005 Drug discovery handbook, Wiley, New York, pp. 123-183.

107. Ekins S, Clark AM, Williams AJ, (2013) Incorporating Green Chemistry Concepts into Mobile Chemistry Applications and Their Potential Uses. ACS Sustain Chem Eng 1: 8-13.

108. Ekins S, Casey AC, Roberts D, Parish T, Bunin BA, (2013) Bayesian Models for Screening and TB Mobile for Target Inference with Mycobacterium tuberculosis Tuberculosis (Edinb), In press:

109. Ekins S, Clark AM, Sarker M, (2013) TB Mobile: A Mobile App for Anti-tuberculosis Molecules with Known Targets. J Cheminform, 5: 13.

110. Clark AM, Sarker M, Ekins S, (2014) New target predictions and visualization tools incorporating open source molecular fingerprints for TB Mobile 2.0. submitted

111. Ekins S, Clark AM, Williams AJ, (2012) Open Drug Discovery Teams: A Chemistry Mobile App for Collaboration. Mol Inform, 31: 585-597.

112. Clark AM, Ekins S, Williams AJ, (2012) Redefining cheminformatics with intuitive collaborative mobile apps. Molecular Informatics, 31: 569-584.

113. Clark AM, Williams AJ, Ekins S, (2013) Cheminformatics workflows using mobile apps. Chem-Bio Informatics J, 13: 1-18.

114. Ekins S, Waller CL, Bradley MP, Clark AM, Williams AJ, (2013) Four Disruptive Strategies for Removing Drug Discovery Bottlenecks Drug Disc Today, 18: 265-271.

115. Anon, In.

116. Anon, In.

117. Ekins S, Polli JE, Swaan PW, Wright SH, (2012) Computational Modeling to Accelerate the Identification of Substrates and Inhibitors For Transporters That Affect Drug Disposition. Clin Pharmacol Ther, 92: 661-665.

118. Beaulieu CL, Ekins S, Samuels M, Boycott KM, MacKenzie A, (2012) Towards the development of a generalizable pre-clinical research pathway for orphan disease therapy. Orphanet J Rare Dis, 7: 39.

119. Wood J, Sames L, Moore A, Ekins S, (2013) Multifaceted roles of ultra-rare and rare disease patients/parents in drug discovery. Drug Discov Today, 18: 1043–1051.

120. Williams AJ, Wilbanks J, Ekins S, (2012) Why Open Drug Discovery Needs Four Simple Rules for Licensing Data and Models. PLoS Comput Biol, 8: e1002706.

Box 1. Tools for facilitating drug discovery in the future may want to consider integrating multiple features that enable access to non-specialists

Funding research: This enables scientists to post project ideas they want funded. Individuals, foundations (the crowd) could then select and fund this research. Alternatively individuals could post their own ideas for projects they want to see done. The scientists and disease foundations could then engage in dialog. This approach would increase the efficiency of funding research.

Crowdsourcing research: Scientists or disease foundations propose work they cannot do and they ask for help. This may be a request for pro bono or paid help. It may be that people with time and flexibility in their careers could simply volunteer their time to a *project.*

Externalizing to companies: This would provide links to CROs and other companies that could assist with various aspects of R&D.

Sharing research openly: This could merge efforts like ODDT with a database element, which enables the searching by compounds, by text, storage of molecules etc. It would also bring in open data from external sources from the internet.

Precompetitive collaboration: This would be a location that could stimulate such collaborations and provide a location for discussion or to propose projects. Project teams could then self-organize and provide a means for delivering content/projects to be shared.

Finding collaborators: This could use tools to enable foundations and parents to search by topic, disease, search grants, and find scientists that can do the research and enable them to connect with them.

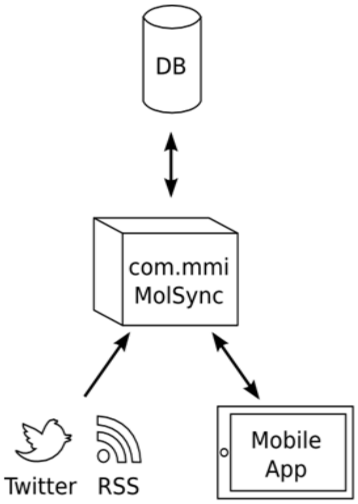
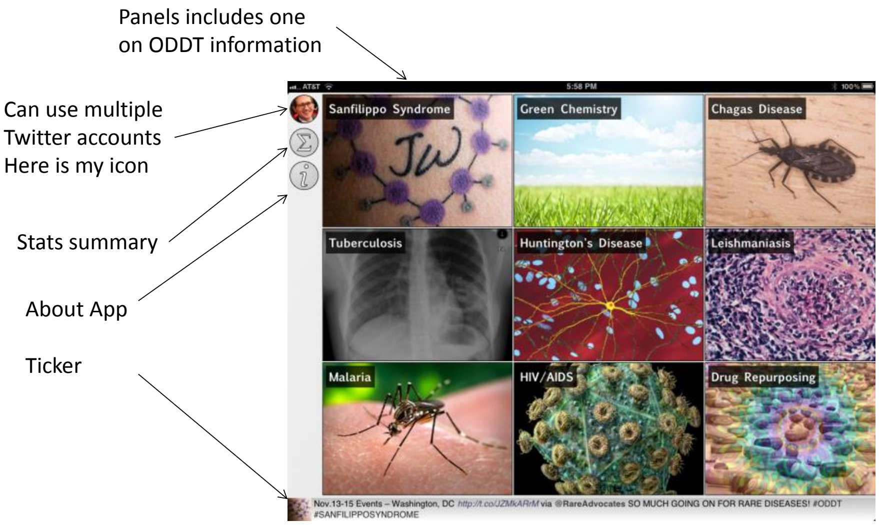


Figure 1.ODDT framework



Nov.13-15 Events – Washington, DC http://t.co/JZMkARrM via @RareAdvocates SO MUCH GOING ON FOR RARE DISEASES! #ODDT #SANFILIPPOSYNDROME

Figure 2. Schematic of ODDT mobile app functions.

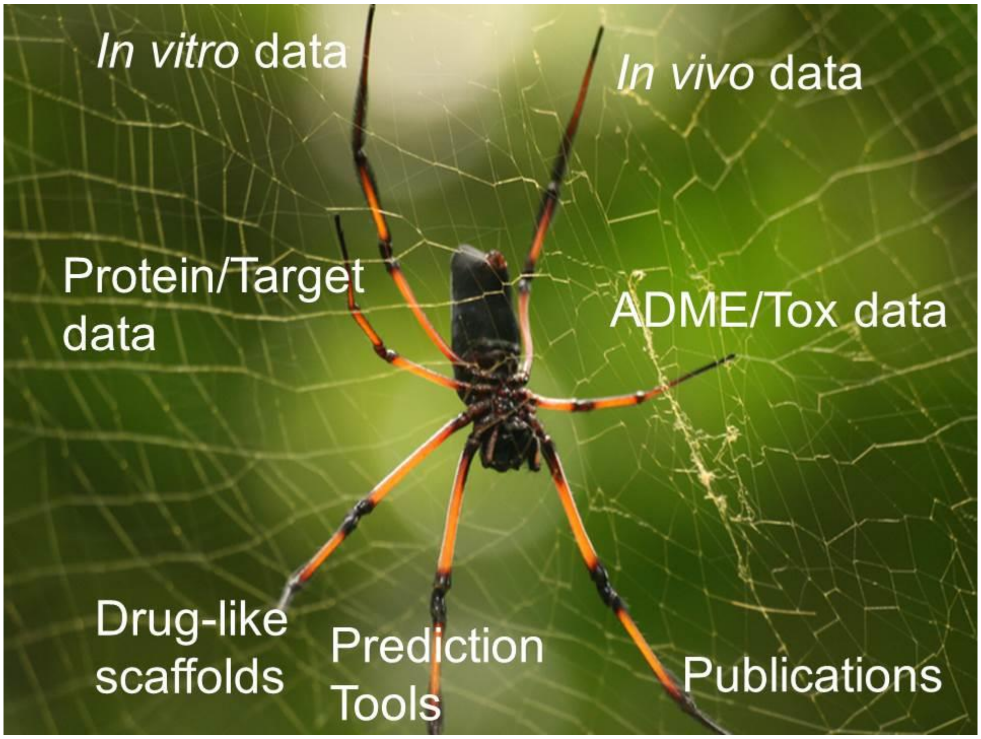


Figure 3. Schematic of the neglected disease related computational tools and information that could be integrated. This could be applicable for any disease or class of diseases.