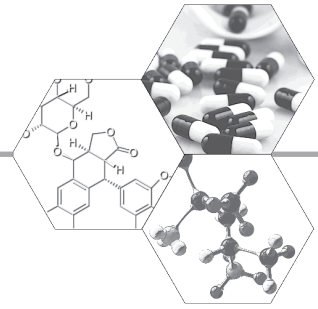
MedChemNet

Powered by fsg

# Challenges and opportunities in collaborative drug discovery

TOP ARTICLE SUPPLEMENTS

## CONTENTS





# The ChEMBL database: a taster for medicinal chemists

"*It has never been easier to access structure–activity relationship data pertinent to a particular chemical series or linked to a specific biological target.*"

**Keywords:** chemogenomics ■ databases ■ open data ■ patents

Contemporary medicinal chemistry has entered an increasingly information-rich era, with increasing focus on simultaneous multi-parameter optimization. The days when medicinal chemists operated behind closed doors using exclusively in-house resources, either in academia or in commercial contract research organizations, biotechnology or pharmaceutical sectors are long gone. It has never been easier to access structure–activity relationship data pertinent to a particular chemical series or linked to a specific biological target. It has never been easier to analyze such data using freely available chemoinformatics analysis or modeling tools.

ChEMBL has arguably transformed the landscape of the available medicinal chemistry data [1,2]. In terms of contents, ChEMBL covers a broad range of curated and annotated data, manually extracted from the primary medicinal chemistry literature. The data include experimental biological readouts, such as binding, functional and absorption, distribution, metabolism, and excretion assay measurements, standardized to common units where possible, and indexed 2D chemical structures, along with linkage to the biological targets and source species. The targets range from single proteins, to protein complexes, then tissues and finally whole organism *in vivo* data. In addition to the literature-extracted information, ChEMBL also integrates deposited screening results from other public databases (e.g., PubChem Bioassay [3]), along with information on approved drugs and their likely efficacy targets. Other sources include kinase screening results from Millipore [4] and the Protein Kinase Inhibitor Set (PKIS) compound collection [5], as well as data from databases such as DrugMatrix [6]. The ChEMBL database is updated on a regular basis and, as of February 2014, the current version (version 17) contains more than 12 million assay measurements for more than 1.3 million distinct compounds tested against more than 9500 biological targets [7].

The data in ChEMBL can be accessed in a number of ways – a dedicated web interface where the user can search, browse, filter and download the results [7]; complete relational database dumps for various database platforms available for download; RESTful web services to facilitate access via programming languages or workflow tools such as Pipeline Pilot and KNIME [8,9]; a Resource Description Framework (RDF) triplestore and associated SPARQL endpoint that allows for federated queries across other EMBL-EBI resources [10], and broader, including integration in the Innovative Medicines Initiative project OpenPHACTS [11]; and finally, myChEMBL [12], a Linux virtual machine, which packages the ChEMBL database, along with a web interface and open access chemoinformatics tools, such as the RDKit toolkit and database cartridge [13]. This latter development is the first time that a complete large-scale, turn-key open cheminformatics system and data has been developed and made freely available.

At each release, in addition to newly extracted data, deposited data [14], and curation and quality assurance of existing data [15], ChEMBL provides new data annotations and analysis tools to the web interface to expand its utility. Examples of this are the widely used ligand efficiency metrics, such as ligand efficiency [16], binding efficiency and surface efficiency indices [17], and lipophilic ligand efficiency [18]. For the binding efficiency and surface efficiency indices, the interface provides interactive scatter plots for each target, where one can select sets of the most ligand efficient molecules that bind to the target of interest. Moreover, drug-likeness metrics such the rule-of-3, the rule-of-5 and QED score [19–21] are also calculated and included in the database.

**George Papadatos**

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, Hinxton, CB10 1SD, UK

**John P Overington**

Author for correspondence: European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, UK
jpo@ebi.ac.uk

FUTURE SCIENCE part of fsg

ChEMBL serves also an open data-sharing hub for the crucial field of neglected tropical diseases (NTD) research. In particular, ChEMBL hosts a series of special data sets related to tropical pathogens in its ChEMBL-NTD pages [22]. The data sets are derived from high-throughput screening campaigns, typically against whole organisms such as *Plasmodia*, *Trypanosoma* and *Mycobacteria* species, from organizations such as GlaxoSmithKline, Novartis-GNF, St. Jude Children's Research Hospital, Harvard, and the Drugs for Neglected Diseases Initiative (DNDi). Moreover, these data sets are fully integrated within the ChEMBL database. Specifically for Malaria research, ChEMBL, in collaboration with the Medicines for Malaria Venture (MMV), provides Malaria-Data [23], a large bioactivity resource focused on compounds, biological targets and assays pertinent to research against this devastating disease. In addition, Malaria-Data hosts screening results from the Open Source Malaria project [24], as well as the MMV Malaria Box [25], a physical set of 400 probe- and drug-like compounds with confirmed antimalarial activity. A number of depositions from academic laboratories around the world has been already submitted and are available in ChEMBL. Needless to say, the data are open and freely shareable by everyone, while we make sure that their provenance is visible and appropriately acknowledged.

> *"Linking and enriching data will allow researchers to better comprehend the bigger picture in drug discovery and leverage the accumulated corpus of knowledge across a multitude of domains of human expertise."*

ChEMBL has been quickly adopted and extensively used by the medicinal chemistry community as indicated by the more than 440 citations to the original ChEMBL publication, within just 2 years since its release in 2011. Indeed, both academic and industrial groups have used ChEMBL as a key corpus of evidence and prior medicinal chemistry knowledge, in order to design, train and optimize data mining algorithms and predictive models and/or as part of integrated tools and drug discovery platforms [26–28]. Recent examples of the applications of the ChEMBL database include target and off-target de-convolution and prediction of mode of action [29,30]; polypharmacology prediction and visualization [31,32]; mining for bioisosteric replacements and activity cliffs [33–35]; matched molecular pair analysis [27]; statistical analyses of assay reproducibility and experimental uncertainty [36,37]; and construction of reference datasets for virtual screening comparative studies [38].

One class of requests we frequently receive from medicinal chemists is the inclusion of patent documents to complement the existing literature data. Patents are a driver for technological innovation and one of the pillars of our modern knowledge economy. Chemical patents, in particular, encompass an unprecedented wealth of knowledge, most of which has never been disclosed in any other source [39].

EMBL-EBI has recently acquired a vast repository of approximately 15 million annotated patented structures from SureChem [40], along with the underlying text-mining and extraction technology, which will be transferred into the public domain and will be rebranded to SureChEMBL [41]. Hardly a static snapshot of chemical patent data, the novelty in SureChem's automated pipeline lies in the fact that it takes regular feeds of new full text patents from three major patent offices, identifies chemical entities from either the in-line text or from images and converts them to the respective 2D chemical structures. The annotated structures are then loaded in a chemistry-aware database. The pipeline takes a day to process the input patent documents when converted from text, and a few days when converted from image sources. Once the newly commissioned SureChEMBL data pipeline and web interface is in place, chemical and keyword searches along with downloads of the results will be freely available for everyone. Checking a structure for novelty against the comprehensive patent chemical space and extracting all patented chemical series for a biological target will be just two popular use cases among medicinal chemists. Chemoinformaticians and other data scientists will no doubt think of more complex workflows; for example to track breakthroughs against historically tough targets, or to significantly enhance competitive position analysis with respect to in-house program status.

Having been established as the largest freely available primary source of literature and patent 2D structure and medicinal chemistry bioactivity data, how would ChEMBL further contribute shaping the future of drug discovery? Going back to the introduction of this editorial, access to bioactivity and biological target data has never been easier and this trend will undoubtedly continue. Given, however, the multifaceted riddle that drug discovery poses, this is certainly not

sufficient on its own, and is likely to shift focus to other challenging areas, such as target validation for a specific disease. For target validation, annotation and integration of data from different resources and domains of chemistry, biology, bioinformatics, proteomics, genomics, clinical and phenotypic studies, pharmacology and toxicology is key. UniChem [42] and the Open PHACTS project [11] are currently focusing on this space, from a chemistry and biology perspective respectively. Linking and enriching data will allow researchers to better comprehend the bigger picture in drug discovery and leverage the accumulated corpus of knowledge across a multitude of domains of human expertise.

**Financial & competing interests disclosure**

*The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.*

*No writing assistance was utilized in the production of this manuscript.*

## References

1 Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(database issue), D1100–D1107 (2011).

2 Bento AP, Gaulton A, Hersey A *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* doi:10.1093/nar/gkt1031 (2013).

3 Wang Y, Suzek T, Zhang J *et al.* PubChem BioAssay: 2014 update. *Nucleic Acids Res.* 42(1), D1075–D1082 (2013).

4 Gao Y, Davies SP, Augustin M *et al.* A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery. *Biochem. J.* 451(2), 313–328 (2013).

5 Dranchak P, Macarthur R, Guha R *et al.* Profile of the GSK published protein kinase inhibitor set across ATP-dependent and-independent luciferases: implications for reporter-gene assays. *PLoS ONE* 8(3), e57888 (2013).

6 DrugMatrix. https://ntp.niehs.nih.gov/drugmatrix/index.html

7 The ChEMBL database. www.ebi.ac.uk/chembl

8 Accelrys Pipeline Pilot. http://accelrys.com/products/pipeline-pilot

9 KNIME. www.knime.org/knime

10 ChEMBL SPARQL endpoint. www.ebi.ac.uk/rdf/services/chembl/sparql

11 Williams AJ, Harland L, Groth P *et al.* Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* 17(21–22), 1188–1198 (2012).

12 Ochoa R, Davies M, Papadatos G, Atkinson F, Overington JP. myChEMBL: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics* 30(2), 298–300 (2013).

13 The RDKit toolkit. www.rdkit.org

14 Hersey A, Senger S, Overington JP. Open data for drug discovery: learning from the biological community. *Future Med. Chem.* 4(15), 1865–1867 (2012).

15 Tiikkainen P, Bellis L, Light Y, Franke L. Estimating error rates in bioactivity databases. *J. Chem. Inf. Model.* 53(10), 2499–2505 (2013).

16 Hopkins AL, Groom CR, Alex A: Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9(10), 430–431 (2004).

17 Abad-Zapatero C, Metz JT. Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* 10(7), 464–469 (2005).

18 Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* 6(11), 881–890 (2007).

19 Congreve M, Carr R, Murray C, Jhoti H. A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* 8(19), 876–877 (2003).

20 Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25 (1997).

21 Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nat. Chem.* 4(2), 90–98 (2012).

22 ChEMBL NTD. www.ebi.ac.uk/chemblntd

23 Malaria-Data. www.ebi.ac.uk/chembl/malaria

24 Open Source Malaria. http://opensourcemalaria.org

25 Spangenberg T, Burrows JN, Kowalczyk P, Mcdonald S, Wells TNC, Willis P. The open access malaria box: a drug discovery catalyst for neglected diseases. *PLoS ONE* 8(6), e62906 (2013).

26 Wang L, Ma C, Wipf P, Liu H, Su W, Xie X-Q. TargetHunter: an *in silico* target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J.* 15(2), 395–406 (2013).

27 Wirth M, Zoete V, Michielin O, Sauer WHB. SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic Acids Res.* 41(D1), D1137–D1143 (2013).

28 Halling-Brown MD, Bulusu KC, Patel M, Tym JE, Al-Lazikani B. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res.* 40(database issue), D947–D956 (2011).

29 Lounkine E, Keiser MJ, Whitebread S *et al.* Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486(7403), 361–367 (2012).

30 Martínez-Jiménez F, Papadatos G, Yang L *et al.* Target prediction for an open access set of compounds active against *Mycobacterium tuberculosis*. *PLoS Comput. Biol.* 9(10), e1003253 (2013).

31 Besnard J, Ruda GF, Setola V *et al.* Automated design of ligands to polypharmacological profiles. *Nature* 492(7428), 215–220 (2012).

32 Fechner N, Papadatos G, Evans D *et al.* ChEMBLSpace – a graphical explorer of the chemogenomic space covered by the ChEMBL database. *Bioinformatics* doi:10.1093/bioinformatics/bts711 (2012).

33 Wassermann AM, Bajorath J. Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.* 3(4), 425–436 (2011).

34 Hu Y, Bajorath J. Extending the activity cliff concept: structural categorization of activity

cliffs and systematic identification of different types of cliffs in the ChEMBL database. *J. Chem. Inf. Model.* 52(7), 1806–1811 (2012).

35 Papadatos G, Brown N. *In silico* applications of bioisosterism in contemporary medicinal chemistry practice. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3(4), 339–354 (2013).

36 Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The experimental uncertainty of heterogeneous public $K_i$ data. *J. Med. Chem.* 55(11), 5165–5173 (2012).

37 Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed $IC_{50}$ data – a statistical analysis. *PLoS ONE* 8(4), e61007 (2013).

38 Riniker S, Landrum G. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* 5(1), 26 (2013).

39 Bregonje M. Patents: a unique source for scientific technical information in chemistry related industry? *World Patent Information* 27, 309–315 (2005).

40 SureChem. https://surechem.com

41 SureChEMBL. www.surechembl.org

42 Chambers J, Davies M, Gaulton A *et al.* UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminf.* 5(1), 3 (2013).

Editorial

*For reprint orders, please contact reprints@future-science.com*

Future Medicinal Chemistry

# The promise of open innovation in drug discovery: an industry perspective

"The last decade has seen a huge increase in ... academic drug discovery..."

**Keywords:** high-throughput screening • lead discovery • open innovation

The last decade has seen a huge increase in the number of academic drug discovery centers with the expertise to discover lead molecules, often termed chemical probes, with the pharmacological properties to test a biological hypothesis in disease models [1]. This has required the creation of academic compound libraries and high-throughput screening (HTS) centers, alongside growth in expertise in techniques including assay development, fragment-based lead generation and computational and medicinal chemistry. The most significant examples of these have been the NIH-sponsored Molecular Libraries initiative, which created a number of HTS centers across universities in the USA [2], and the European Lead Factory (ELF) [3]. Alongside this, universities and charities have established drug discovery groups including Medical Research Council Technology and Cancer Research Technology in the UK, and the EU-OPENSCREEN project, which aims to establish a co-ordinated network of 15 or more screening centers across Europe [4,5]. The common objective of these groups is to source novel target hypotheses from academia, to generate lead molecules and to use these molecules to validate the target in cellular and animal models of disease, with the ultimate objective of validating new targets and creating a drug discovery program. While there have been notable examples of the delivery of candidate drugs in rare and orphan diseases relatively few academic drug discovery projects have progressed into the clinic. Reasons for this include the low levels of target validation associated with academic drug targets, the quality of the compound collections and hence lead molecules identified and the lack of expertise, and funding, within academia to deliver drug candidates and support clinical studies. These gaps in capability can be addressed through collaboration with pharma, while the gap in disease biology in pharma can be addressed through collaboration with academia [6,7]. To bring the strengths of both communities together requires the establishment of new relationships where both parties contribute complementary expertise and share in the resulting success of collaborative projects. This new collaborative model has been termed open innovation (OI).

The concept of OI was first discussed by Henry Chesbrough in the 1960s. Central to the concept of OI is a recognition that any one organization does not possess the knowledge or capability to succeed alone, and through developing partnerships with complimentary organizations is able to bring increased knowledge to a problem to create the conditions to advance innovation [8]. It is only in recent years that the term has been used to describe drug discovery collaborations between industry and academia [9]. Traditionally, drug discovery and development has been performed by pharmaceutical and biotechnology companies, in large research campuses employing many thousand scientists, in a model that has been termed closed innovation. While pharma and academia have always interacted, such collaborations have typically been transactional with academia funded to conduct studies on behalf of industry with little sharing of costs between the partners or sharing of the value resulting from the project. There is extensive literature describing the decline in productivity of

**Steve Rees**
AstraZeneca, Darwin Building, 310 Cambridge Science Park, Milton Road, Cambridge, CB4 OWG, UK
Tel.: +44 7717 800162;
steve.rees@astrazeneca.com

FUTURE SCIENCE part of fsg

pharma that suggests that the traditional model of drug discovery is broken. Reasons for this include increased regulatory hurdles, a move to treating more chronic diseases and significant project attrition due to safety and efficacy issues in the clinic [10]. To address this pharma has adopted a strategy that requires significant evidence of target validation before starting a drug discovery project [11], which requires significant investments in target biology to understand the role of the target in disease. To achieve this pharma brings together pharma expertise in lead discovery and clinical science with academic expertise in disease biology, in which each partner brings complimentary expertise and shares in the success of any resulting project [12]. This model of shared invention and ownership has been successfully applied for developing treatments to neglected tropical diseases and is now being applied more broadly in the areas of clinical research, lead discovery and target identification.

> "...compound collection has little value unless it is screened..."

It is well documented that >90% of clinical development projects fail to reach the market as a result of changes in company strategy, the identification of safety liabilities associated with the molecule or the absence of efficacy when tested in a clinical study [7]. Hence, there exist in pharma many hundreds of pharmacologically-optimized molecules with activity against drug targets. Alongside this there exists many lead molecules, generated in discovery programs, with the appropriate pharmacological properties for use in cellular and animal models of disease. Taken together, there is a huge opportunity to use these molecules to better understand the role of targets in disease. OI provides a mechanism through which this valuable chemical resource can be made available to the scientific community for testing in disease models to potentially generate new medicines. The NIH 'Discovering New Therapeutic Uses for Existing Molecules' program involves eight pharma making available a number of clinical candidates for testing for new therapeutic use.

Similarly, seven pharma have released over 80 compounds to the UK Medical Research Council for use in preclinical or clinical studies. In an extension to this model AstraZeneca, through its OI program, is making available molecules that have failed in clinical development and molecules within active development programs for testing in new disease areas. In these studies pharma makes available optimized chemical equity, with the new research study being funded by academia, with both parties sharing in success in the event of the identification of a new disease indication. While these examples demonstrate the potential to be gained through sharing of optimized molecules, there remain huge opportunities for further knowledge generation as greater numbers of compounds are made available.

Many pharma companies own compound libraries for use in HTS to identify start points for drug discovery. These libraries, consisting of several million molecules, contain compounds generated throughout the history of that company, supplemented with purchased compounds and compounds specifically designed to enrich the chemical space covered by the collection. The collections contain molecules with the physiochemical properties most likely to result in the generation of drug candidates that will be safe and efficacious in the clinic [13]. Alongside this pharma has created the infrastructure to curate and screen these compounds together with the computational and medicinal chemistry expertise to optimize hit molecules. Through OI partnerships, there exists an opportunity to pair academia's expertise in biology with industry expertise in hit discovery to identify quality hit molecules with activity at novel drug targets and, through the expertise that exists in academia, to use these molecules to validate targets within disease model systems with both parties sharing in the success of these projects as they progress to the clinic. Establishing these collaborations requires that pharma realizes that a compound collection has little value unless it is screened, and academia realize that novel drug targets have little value unless paired with a therapeutic agent that can be used to test the disease hypothesis.

Many companies have established such partnerships. GlaxoSmithKline (GSK) has established the Discovery Partnerships with Academia (DPAC) program in which academics gain access to HTS and other expertise in GSK. AstraZeneca has established a portfolio of partnerships in which academic targets are screened within AstraZeneca laboratories, and in which up to 250,000 compounds from the AstraZeneca compound collection are shared with academic screening centers for screening in their facilities. Through these programs AstraZeneca is supporting a portfolio of 25 OI projects. The Bayer Grants4Targets program has similar objectives and other pharma support similar activities. In an extension to these models AstraZeneca has recently announced the creation of the AstraZeneca MRC Centre for Lead Discovery, a unique collaboration in which AstraZeneca, the UK Medical Research Council and Cancer Research UK scientists will share common laboratories within the AstraZeneca research campus in Cambridge, UK. This collaboration will give academic scientists full access to AstraZeneca drug discovery expertise while enabling AstraZeneca to develop new partnerships with academia. The details of these programs can be found on the respective pharma company websites. While the

structure of the partnership varies, the objective of pairing industry expertise in lead discovery with academic expertise in basic biology is consistent. While it is early to judge success, it is reasonable to expect that a significant proportion of the future industry pipeline will consist of projects derived from such collaborations.

This model of OI underpins the strategy of the ELF. Seven pharma have contributed a proportion of their compound libraries to create a 270,000 compound Joint European Compound Library, which will grow to 500,000 compounds through academic design and synthesis [3]. Donation of these compounds leveraged funds to create an industry quality screening center performing 24 HTS annually, with target proposals sourced from academia. The industry partners are able to license projects resulting from the program and gain access to the Joint European Compound Library for screening against internal drug targets. This is a highly effective mechanism to increase the diversity of compounds available for screening. In a similar initiative AstraZeneca and Bayer have established a collaboration whereby each partner screens its compound library against drug targets nominated by the other partner on a *quid pro quo* basis [14]. This enables high value drug targets to be tested against two compound libraries without the requirement to purchase and store the additional molecules. While compound exchanges between pharma are rare, compound sharing offers an opportunity to increase the chemical diversity available for testing against high value drug targets at low cost. It will be interesting to see if the number of such collaborations increases in the years to come.

It has been suggested that target identification and validation is precompetitive, with value being established as a therapeutic agent is identified and the project progresses to the clinic. There are relatively few examples of OI collaborations to identify novel targets. Perhaps the most significant example is the Structural Genomics Consortium, which has the objective of developing chemical probes to a range of targets including kinases and epigenetic enzymes, and making these available to the scientific community [15]. This initiative is supported by a number of pharma companies and has generated wealth of information about these target classes. One area in which collaboration could significantly advance the understanding of disease is the sharing of human biological samples. Academia and industry have established collections of human biological samples from healthy and disease patients. Typically these are not made available to the broader scientific community. This represents a missed opportunity; the broad availability of these unique samples to researchers would enable their use to better understand disease pathways and lead to the identification of new drug targets and the characterization of stratified patient populations to support personalized medicine strategies.

The pharma industry is on a journey from an era of closed innovation, when discovery was largely performed internally, with little sharing of knowledge and research assets, toward an era of OI where pharma exists as part of a broader symbiotic community. The academic community is under increasing pressure from funding bodies to demonstrate translation of academic discovery into new medicines. Taken together, this creates the opportunity for OI to become commonplace. Establishing such collaborations requires an increase in trust and an understanding of the value of target and compound-related intellectual property, accompanied by an appreciation that success can only be achieved through the sharing of knowledge and expertise. It is the contention of this author that the successful pharma companies will be those that transition to a model in which they exist as part of a broad network with many collaborative partners, with a free exchange of information and expertise based on trust and shared objectives, to identify and validate novel drug targets and deliver the next generation of medicines, with all parties sharing in the commercial success resulting from these efforts.

### Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

## References

1 Frye S. US academic drug discovery. *Nat. Rev. Drug Disc.* 10, 409–410 (2011).

2 Austin CP, Brady LS, Insel TR, Collins FS. NIH molecular libraries initiative. *Science* 306, 1138–1139 (2004).

3 Mullarrd A. European Lead Factory opens for business. *Nat. Rev. Drug Disc.* 12, 173–175 (2013).

4 Shanks E, Ketteler R, Ebner D. Academic drug discovery within the United Kingdom: a reassessment. *Nat. Rev. Drug Disc.* 14, 510 (2015).

5 Frank R. EU-OPENSCREEN – a european infrastructure of open screening platforms for chemical biology. *ACS Chem. Biol.* 9, 853–854 (2014).

6 Chung TDY. Collaborative pre-competitive preclinical drug discovery with academics and pharma/biotech partners

at Sanford/Burnham: infrastructure, capabilities and operational models. *Comb. Chem. High Throughput Scr.* 17, 272–289 (2014).

7 Hudson J, Khazzragui HF. Into the valley of death: research to innovation. *Drug Disc. Today* 18, 610–614 (2013).

8 Chesbrough HW. The era of OI. *MIT Sloan Management Review* 44, 35–41 (2003).

9 Schumacher A, Germann PG, Gassmann O. Models for OI in the pharmaceutical industry. *Drug Disc. Today* 18, 1133–1137 (2013).

10 Paul SM, Mytelka DS, Dunwidde CT *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Disc.* 9, 203–214 (2010).

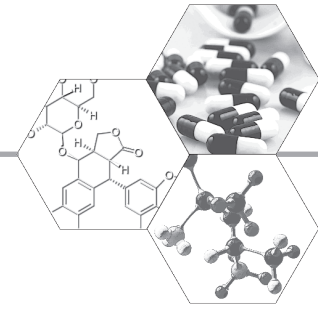
11 Cook D, Brown D, Alexander R *et al.* Lessons learnt from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Disc.* 13, 419–531 (2014).

12 Simpson PB, Reichman M. Opening the lead generation toolbox. *Nat. Rev. Drug Disc.* 13, 3–4 (2014).

13 Cumming J, Davis AM, Muresan S, Haeberlein M, Chen H. Chemical predictive modeling to improve compound quality. *Nat. Rev. Drug Disc.* 12, 948–962 (2013).

14 Kojel T, Blomberg N, Greasley PJ *et al.* Big pharma screening collections: more of the same or unique libraries: the AstraZeneca-Baher Pharma AG case. *Drug Disc. Today* 18, 1014–1024 (2012).

15 Gileadi O, Knapp S, Lee WH *et al.* The scientific impact of the Structural Genomic Cnsortium: a protein family and ligand-centred approach to medically-relevant human proteins. *J. Struct. Funct. Genomics* 8, 107–119 (2007).





# Exploiting open data: a new era in pharmacoinformatics

Within the last decade open data concepts has been gaining increasing interest in the area of drug discovery. With the launch of ChEMBL and PubChem, an enormous amount of bioactivity data was made easily accessible to the public domain. In addition, platforms that semantically integrate those data, such as the Open PHACTS Discovery Platform, permit querying across different domains of open life science data beyond the concept of ligand-target-pharmacology. However, most public databases are compiled from literature sources and are thus heterogeneous in their coverage. In addition, assay descriptions are not uniform and most often lack relevant information in the primary literature and, consequently, in databases. This raises the question how useful large public data sources are for deriving computational models. In this perspective, we highlight selected open-source initiatives and outline the possibilities and also the limitations when exploiting this huge amount of bioactivity data.

## Open data: a rich resource of small-compound bioactivity data

With the launch of publicly available bioactivity databases, such as PubChem [1] in 2004 and ChEMBL [2] in 2009, the chemoinformatics community now has access to millions of data points. Full exploitation of this rich source of data is of high interest to researchers in the fields of chemo- and bio-informatics, medicinal chemistry and drug discovery – especially for those working in an academic environment. In addition, public–private partnerships that have emerged as part of **open data** initiatives, also attracted the pharmaceutical industry to enrich their in-house data with the one freely available in the open domain [3].

The current version of the ChEMBL database (ChEMBL, version 17, prepared in August 2013) comprises data on biological activity of 1,519,640 compound records measured in 734,201 biological assays and acting on 9356 different targets. A third of the bioactivity data present in ChEMBL is manually extracted from peer-reviewed articles and is available to the public online [4]. It is hosted by the European Bioinformatics Institute (EMBL-EBI) under the leadership of John Overington. Since this data bank was launched, it has become an indispensable tool for the scientific community as one of the main reference points to explore the pharmacological space of a target of interest. Collected data include information regarding compounds (chemical structure representations, compound names), their bioactivities tested on certain targets, target information (target sequence, target organism, target name, protein target UniProtID [5]), a description of the biological assay extracted from the reference paper and information on the literature source itself (e.g., PubMedID, journal, year of publication, and so forth). Every compound, target, and **bioassay** is assigned with a ChEMBL identifier. All information is introduced in machine-readable format, a web service allows easy navigation and search results are available for download under secure protocol. Very recently, the whole setup has also been provided as virtual machine for download [6].

The second huge source of publicly available compound-bioactivity data is PubChem, first made available in September 2004, as part of the US NIH Molecular Libraries Programme. The BioAssay section of the database works as a repository of small-molecule screening data generated by the Molecular Library Screening Center Network and the Molecular Library Probe Production Center Network under the Molecular Libraries Programme. Currently, PubChem also receives contributions from many other organizations, including ChEMBL, ChemBank [7], BindingDB [8], pharmaceutical companies, federally funded screening centers and academic research laboratories [9]. PubChem contained, in September 2013, more than 710,000 registered BioAssays and more than 47 million compounds.

More specialized databases are likely to fulfill the needs of a smaller branch of the community, such as the IUPHAR [10] database containing data from G protein-coupled receptors (GPCRs), voltage-gated ion channels, ligand-gated ion channels and nuclear hormone receptors [11].

**Daria Goldmann[1], Floriane Montanari[1], Lars Richter[2], Barbara Zdrazil[1] & Gerhard F Ecker*[1]**

[1]University of Vienna, Department of Pharmaceutical Chemistry, Division of Drug Design & Medicinal Chemistry, Althanstrasse 14, 1090 Vienna, Austria
[2]Medical University of Vienna, Department of Medical Biochemistry, Dr. Bohr-Gasse 9/2, 1030 Vienna, Austria
*Author for correspondence:
Tel.: +43 1 4277 55110
Fax: +43 1 4277 9551
gerhard.f.ecker@univie.ac.at

FUTURE SCIENCE part of fsg

**Key Terms**

**Open data:** Publicly accessible data on various subjects such as chemical structures of compounds, their biological activity, protein structures, scientific publications.

**Bioassay:** Describes a biological testing procedure for determining the biological activity of a substance by measuring its effect on an organism, tissue, cell, enzyme or receptor preparation compared with a standard preparation.

**Ontology:** In computer and information science, an ontology formally represents knowledge concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts. Thus, ontologies are used to organize information within a domain and (especially in the life sciences) to annotate experimental data.

TP-search [12,13], recently integrated to ChEMBL, complements existing literature-retrieved data on transporters from the ABC and SLC gene families.

Another public database of special interest to the drug-discovery community is the GPCR DB [14]. It contains structural and mutation data, but also ligand-binding data on GPCRs. The latter is a collection of data by Seeman [15] and data made available by Organon [16].

The public data source BindingDB [17] focuses primarily on a collection of binding affinity information for proteins that are drug targets or candidate drug targets interacting with drug-like molecules. Data on chemical compounds, protein targets and assay conditions are extracted from literature and several public projects (PubChem BioAssay and ChEMBL). In addition, curated data are interconnected with other open-access databases through UniProt, ChEMBL, PDB [18], PubChem and PubMed identifiers.

Another focused collection of chemicals is DrugBank [19,20]. It contains annotated data of small- and large-molecule drugs, drug–target proteins, related genes and diseases, drug pathways, drug transporter information, metabolite data, information on toxicity, drug–food and drug–drug interactions. Compared with previously mentioned databases DrugBank does not provide precise bioassay and bioactivity data.

ChemBank [21] is one of the successful examples of internally generated and annotated databases of bioactive compounds. For small organic molecules it stores raw screening data, such as high-throughput screening (HTS) and small-molecule microarray data, obtained at the Broad Chemical Biology screening center (Broad Institute of Harvard and Massachusetts Institute of Technology [MA, USA]). At the same time, it provides a rigorous definition of screening experiments and related bioassays.

Last but not least, it is worth mentioning the chemical structure database of organic molecules ChemSpider [22]. Structure and property information is collected from various sources, which comprise published journal articles, open-access databases, academic institutions as well as vendors. However, the bioactivity profiles of small molecules are omitted in the database.

All these open-access databases are rich sources of compound-structure, compound-bioactivity and compound-pharmacology data. However, they are provided in different formats, using different ontologies and standards for reporting chemical structures and biological assays, and the data are of varying quality. To fully exploit these data in drug discovery, integration is a mandatory task. Once integrated, they can become an extremely powerful tool for fully exploring the chemical and pharmacological space of biological targets of interest. Furthermore, linking pathway and disease information to these data will allow to approach research questions of hitherto unmet complexity [23].

## Semantic integration of public data sources

Integration of data sources of different domains could on principle follow two routes:

- Creating a 'super-database' with a quite complex schema;
- Keeping the original sources and performing integrating on the fly, using a schema free technique.

The flagship example for approach is ChEMBL, implemented as an SQL database hosted by the EBI and improved and enhanced towards new areas. For schema free integration of data sources, the recently launched Open PHACTS discovery platform [24] may serve as paradigm example [25]. It started to work in March 2011 as a public–private partnership between the European Community (Innovative Medicines Initiative) and selected European Federation of Pharmaceutical Industries and Associations members. Using semantic web technology, it allows scientists to explore and interrogate these integrated biological and chemical data. By September 2013, it integrates data from ChEMBL, ChEBI [26], DrugBank, ChemSpider, GeneOntology [27], WikiPathways [28], UniProt, ENZYME [29], and ConceptWiki [30]. Apart from the Open PHACTS Explorer [31] – a simple web-based user interface for querying and viewing the integrated data – the Open PHACTS Discovery Platform provides a convenient application programming interface in a form of web service to query across multiple data sources in the field of life sciences. The latter allows addressing complex research queries that step out of the simple target-compound-pharmacology space, but also involve information from pathways and their relation to certain **ontology** tags. Integration of related disease information is ongoing. Still, one should be aware that despite laborious mapping of the varying identifiers from different resources, the limitations of data quality and completeness retrieved from the Open PHACTS discovery platform will

always be those of the underlying data sources themselves.

For both integration approaches identification of duplicates is key. In chemistry, this is starting to get resolved by the growing acceptance of the IUPAC standard international chemical identifier (InChI) [32] as a unique chemical representation format. However, a current drawback in this field is a lack of unified and agreed rules on the crucial step of chemical standardization prior to standard InChI creation [33]. With respect to biological data, the problem is much more complex. Main challenges are the interlaboratory variances as well as the numerous ways of reporting biological activity values. In addition, there is an almost infinite number of possibilities for conducting and describing a biological assay.

## Which type of biological activity should be used by chemoinformaticians?

There are numerous types of bioactivity data reported in open-access databases, which implicates the existence of different numerical end points such as $IC_{50}$ (concentration at half-maximum inhibition), $EC_{50}$ (concentration at half-maximum effect), $K_i$ (inhibitor constant for the protein-inhibitor complex), ratios (e.g., efflux ratio, fluorescence activity ratio), percentage of inhibition, and so on.

Especially for building classification models (predicting active vs inactive compounds), this raises the question whether one should focus on a single type of activity measurement or combine data from different end points with the goal to enlarge the chemical and bioactivity space? Of course, mixing data with different end points is always a compromise between quality of the data and amplitude of a data set.

PubChem BioAssay does not differentiate between different end points and the user has to go through all reported bioassays and select those which could be equal or comparable. In contrast, ChEMBL reports over thousands different end points of bioactivity for its compounds.

It was recently discussed by Hu and Bajorath [34] that, to avoid building models on data originating from diverse pharmacological protocols, one would consider to use only $K_i$ values for computational studies. The main argument for this is that $K_i$ values, in contrast to $IC_{50}$ and $EC_{50}$, are considered to be assay independent. However, this poses the risk to get a quite incomplete picture of the available chemical and pharmacological space. The quantity of the $IC_{50}$ data reported in open-data sources is, for instance, much higher than that for other activity end points (e.g., the amount of $IC_{50}$ data in ChEMBL is approximately three-times higher than the amount of $K_i$ data) [35]. Thus, it is definitely worth exploring their use for computational studies.

To further support this line of argumentation, we compared the variability of $K_i$ and $IC_{50}$ data present in ChEMBL version 17. Using an in-house script we collected compounds reported with several $K_i$ or $IC_{50}$ values against the same biological target. The following target types were included in the analysis: single protein, protein complex, protein complex group and protein family. Only entries that have 'nM' as their unit, have a standard relation of '=' and are not flagged as 'outside typical range' were used in the comparison [35]. These correspond to 99.82% of all non-null $K_i$ entries and 95.47% of all non-null $IC_{50}$ entries. In total, 22,136 compounds are reported in ChEMBL version 17 with multiple $K_i$ values against 1027 biological targets **(FIGURE 1A & B)**. For the vast majority of these compounds (18,202) reported $K_i$ values lie within one log range (marked in dark on **FIGURE 1A**), which might be considered as experimental error and hence might be considered as equivalent. The remaining 3934 compounds

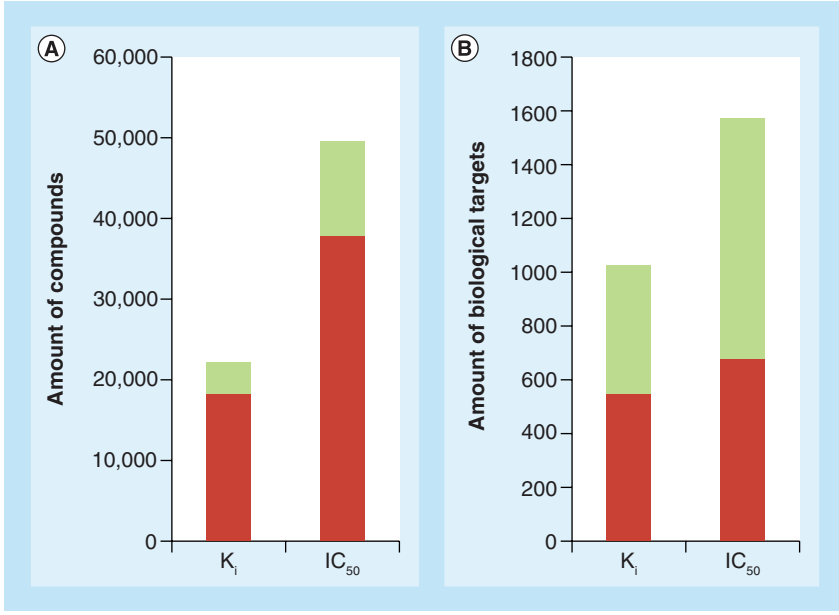



**Figure 1. Overview of repeated measurements in ChEMBL as $K_i$ and $IC_{50}$ values. (A)** Amount of compounds in ChEMBL version 17 that have at least two $K_i$ or $IC_{50}$ values against one biological target. **(B)** Amount of biological targets for which these compounds are reported. Quantity of compounds and biological targets for which reported $K_i$ ($IC_{50}$) values lie within one log range is marked in a dark color. Quantity of compounds and biological targets for which reported $K_i$ ($IC_{50}$) values are beyond one log range is marked in light color.

**Key Term**

**Scaffold:** Main part of the chemical skeleton of a molecule, composed of the rings and linkers, to which substituents are branched.

(about 17.8% of all compounds having multiple $K_i$ values) remain for which the activity values differ by more than tenfold. Such dispersal of values might not be explained by experimental errors and compounds should therefore not be included in data sets prior to manual inspection of the activity values. Projecting these findings on the target space we can state that for almost half of the targets (550) with multiple $K_i$ values, all reported activity values lie within one log range (marked in black on **Figure 1B**) whereas the other half contains compounds that differ more than tenfold.

For $IC_{50}$ data, an even larger amount of compounds (49,518) with multiple activity values against 1575 biological targets are present in ChEMBL version 17 **(Figure 1A)**. However in this case, nearly a quarter (~23.8% of all compounds with multiple $IC_{50}$ values) are reported with $IC_{50}$ values that lie outside one log unit (11,775 compounds dispersed through 903 targets, **Figure 1B**). This is in agreement with findings of Kalliokoski and colleagues [35], who demonstrated that the variability of pairs of $IC_{50}$ data (two independent measurements for the same compound–protein system) is approximately 25% worse than that of $K_i$ data. Thus, enlarging to a $K_i$ data set by $IC_{50}$ data is not recommended because this would decrease the data quality. It still seems feasible to do it the other way around: if the assay conditions are clearly described in the published work, it is possible to use the Cheng–Prusoff equation to convert $K_i$ into $IC_{50}$ values [36]. Even if specific assay knowledge is unavailable, the authors propose considering a conversion factor of 2.0 [35].

Taken together, in our point of view the most efficient way to explore the chemical and pharmacological space of a target of interest is to consider all data available, manually curate it and decide individually, on basis of data quality and size of the chemical and pharmacological space covered, which data to include into the final test set.

## Completeness of the chemical space: enriched scaffolds & singletons

While collecting and curating data sets for one of our in-house projects, we have noticed that bioactivity information downloaded from ChEMBL for a given biological target usually originates from several structure–activity relationship (SAR) studies performed by different groups [37]. Most of the studies provide pharmacological data for chemical series sharing the same **scaffold**. As a result, databases compiled from such studies overrepresent some scaffolds while other areas of the chemical space are poorly explored. In addition, data sets compiled from SAR series tend to be biased towards active compounds, as most journals disfavor publication of hundreds of inactive compounds.

**Figure 2** illustrates the situation for TRPV1 antagonists: the nodes represent Murcko scaffolds [38] occurring in the data set, their size corresponding to the amount of compounds sharing that scaffold. Two nodes are connected if the similarity between the scaffolds is superior to 0.7 (chemical structures are described by ECFP6 fingerprints, similarity is measured by Tanimoto) and the edges are thicker if this similarity is higher.

Thus, in the TRPV1 antagonist data set, three scaffolds seem to be overrepresented, five families of scaffolds are linked by similarity and the rest of the scaffolds are, for most of them, singletons (meaning that their scaffold is unique and they do not have close structural neighbors in the data set; **Figure 2**).

We expect this particular situation to be common for many targets in ChEMBL. This irregular coverage of the chemical space has to be taken into account when attempting to build predictive models using such data sets, as overrepresentation of scaffolds might bias classical evaluation methods. From a pure machine-learning point of view, the ideal data set would contain equally distributed scaffolds, building a complete disconnected graph covering the largest possible chemical space.

There is no such problem in PubChem, as screening libraries are typically designed to be chemically diverse due to the specifications of the NIH Roadmap initiative **(Figure 3)** [39]. In addition, the vast majority of tested compounds are 'drug-like' and inactive compounds are reported along with their active counterparts, allowing researchers to gather negative information aswell.

Obviously the quantity and quality of the available data heavily influences the computational approaches used for creating knowledge from these data. An equally populated chemical space allows the application of machine learning, whereas highly populated scaffold islands with diverse substituents allow quantitative SAR studies within these islands. Finally, several highly active singletons could qualify for performing pharmacophore-based analyses. Therefore, we recommend carefully exploring the compound data sets collected from different sources before deciding which computational method to use.

## Biological & chemical heterogeneity of public data sources

Although public data sources are most of the time the only available resource for academic researchers, their reasonable usage is a challenging task. Challenges mainly arise from issues concerning the uniform and standardized representation of entities, which may be observed both among different data providers and also within the same data source. Since small-molecule bioactivity data are mainly manually compiled from literature and from high-throughput screens, they are highly heterogeneous with respect to the representation of compounds, their target taxonomies, assay descriptions and underlying ontologies used.

As discussed previously, biological data are by far less standardized than the chemical ones. As the textual bioassay description in ChEMBL, for instance, is directly excerpted from the underlying paper, the level of detail given is completely dependent on the way the authors of the publication present their pharmacological data. In addition, the vocabulary used for defining the assay parameters, such as cell lines, radioligands and assay types, most often does not correspond to any standardized identifiers (e.g., the ones defined in biomedical ontologies, such as the bioassay ontology [BAO] [40,41] or the experimental factor ontology [42]). Thus, in ChEMBL, every assay description excerpted from a paper gets a unique ID, in order to avoid annotation errors. Therefore, assays following the same protocol but extracted from different papers cannot be identified as being the same by the user. Consequently, combining such data for identical compound-target systems across different assays requires a thorough study of the underlying experiment [43]. In addition, redundancy in the assay descriptions requires laborious cleaning steps of compiled data sets and also raises the question of which entries to keep and which ones to remove. As it was demonstrated recently for sets of $IC_{50}$ data points [35], there might be multiple entries for the same compound with apparent distinct bioactivity values (targeting the same protein), which in fact proved to be unit or value errors.

Compared with ChEMBL, in PubChem each assay is classified by the contributing organization according to the stage of the assay project. The assay method can be a 'primary high-throughput screening' assay where the activity outcome is based on percentage inhibition from a single dose, a 'confirmatory' assay (typically a low-throughput assay where the activity outcome is based on a dose–response relationship with multiple tested concentrations) or a 'summary' assay, for validated chemical probes or small-molecule leads [1]. PubChem also has limitations that burden data retrieval and analysis. The main reason resides in its semi-structured data representation, which is largely dependent on the submitter: the assay description, being organized by several free text fields, makes it impossible to query PubChem by simple concepts such as 'GPCR agonist assays' [44]. Moreover, the two required compound activity fields are 'assay outcome' and 'assay score' that are attributed subjectively by the submitter, which renders quantitative comparisons between assays almost impossible.

In addition, for HTS data like those available from PubChem, the problem of false positives complicates the process of data retrieval [45–49]. Frequent hitters or promiscuous compounds are known to regularly sully the results of HTS. Thus, Rohrer and Baumann [45] advise to use promiscuity filters to remove suspicious compounds and assay artifacts. Furthermore, it

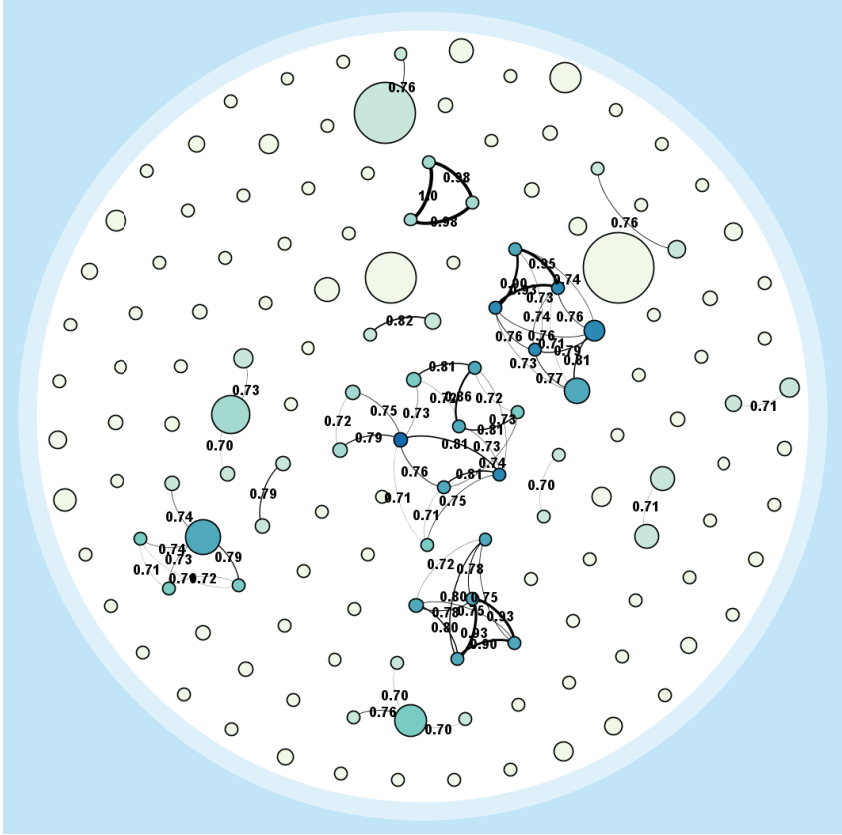


**Figure 2. Graphical representation of a ChEMBL version 13 extracted TRPV1 data set of competitive antagonists.** The nodes represent Murcko scaffolds and their size encodes the amount of compounds sharing this scaffold. The edges represent similarity between scaffolds. The initial data set contained 408 compounds, which resulted in 145 different Murcko scaffolds.

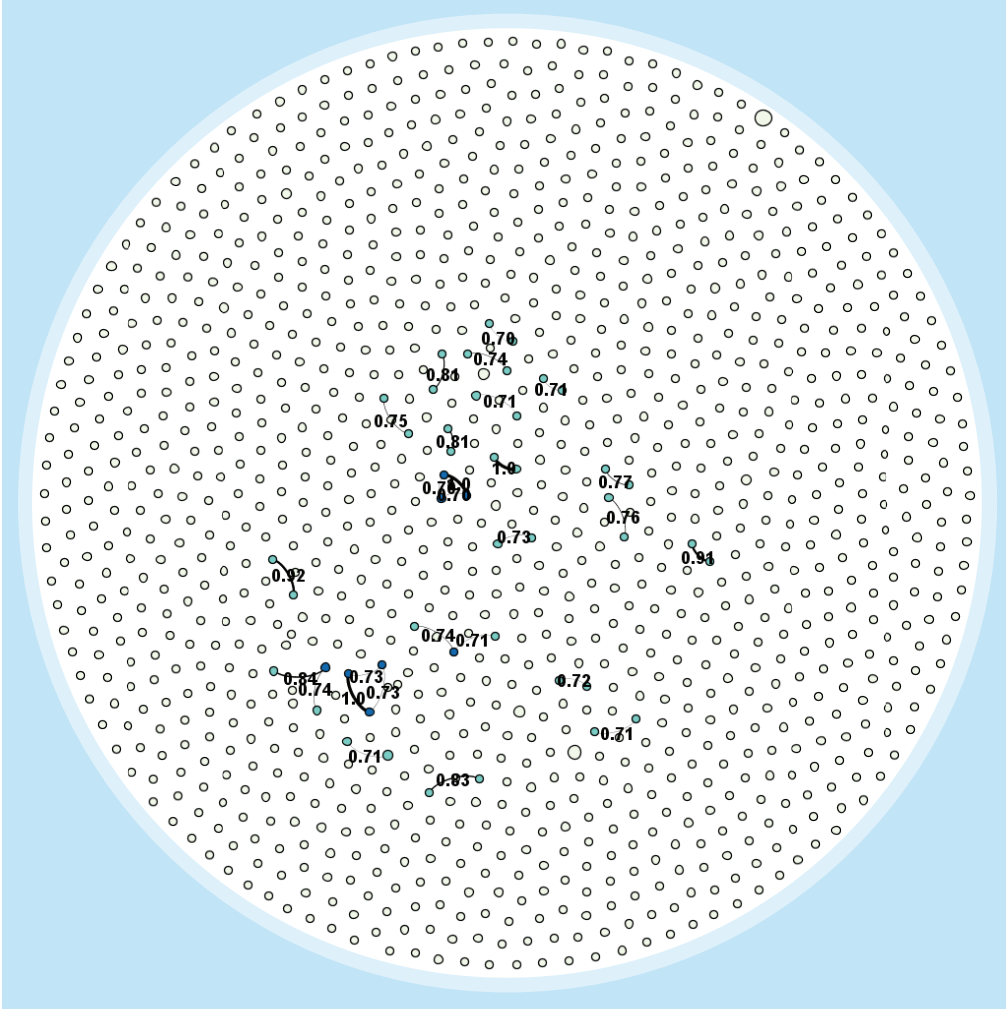



**Figure 3. Graphical representation of a PubChem extracted TRPV1 activation data set (AID623959).** The nodes represent Murcko scaffolds and their size encodes the amount of compounds sharing this scaffold. The edges represent similarity between scaffolds. The initial data set contained 1358 compounds, which resulted in 1184 different Murcko scaffolds.

might be efficient to use only the 'confirmatory' version of such assays as data sets for modeling.

In order to make open data more homogenous with respect to the assays, PubChem could be more strict in the data submission forms, by implementing the BAO vocabulary and by forcing submitters to provide standardized definitions for the respective target in the assay description. In other words, the bioactivity data would not be accepted for upload if these criteria are not met. The use of the same established vocabulary by all public compound-pharmacology-related data sources would not only improve the quality of open data, but also would enable quantitative comparison of results between assays and simplify cleaning steps when deciding which entries to keep in the collected data sets for computational studies.

With respect to chemical structures available in public databases, it is important to stress several problems that unfortunately have not been solved within the last 10 years [50]. First and foremost, publishers currently do not demand that authors upload chemical structures in a computer-readable format. Therefore, the quality of chemical structures in databases compiled from literature sources relies on correct structure depiction in the journals, and on correct transmission to the databases. As outlined in a recent paper by Tiikkainen *et al.* where they systematically studied error rates in public bioactivity databases, small-molecule structures seem to have the greatest estimated error rate (followed by target, activity value and activity type) [51].

Errors in representations could concern bonds, ring systems, chiral centers or the

number of hydrogen atoms. In addition, 2D depictions do not always allow the preservation of stereochemical features, or can adequately handle different sorts of isomerism (tautomerism, regioisomerism, optical isomerism and geometrical isomerism) [52]. Thus, chemical structure representation is not only very error-prone, it is also a long-standing matter of discussion regarding compound uniqueness and the type of canonical molecular identifiers that shall be used as a standard. Prominent examples for such chemical notation systems are the SMILES strings [53], the CACTVS hash codes [54] and the IUPAC InChIs, the latter being more and more accepted as a global standard. Errors in public data sources of course would be remarkably reduced if the authors were asked to provide the respective structures in a computer-readable format. Thus, we strongly recommend publishers to include this possibility into their electronic submission systems.

## Coverage of public data sources: how complete are the data?

In addition to those issues concerning the uniformity and the quality of small compounds and their respective bioactivity data in public databases, it is also the extent of data completeness that affects the usefulness of the extracted data set and subsequently the computational study performed with it. In open-data sources compiled from literature, data originate from only a limited number of peer-reviewed journals. As shown in **Figure 4** for ChEMBL version 17, most of the data are extracted from other journals. Thus, ChEMBL will certainly not cover the whole chemical space available for the respective biological target. To support this, we gathered data for a set of transporters of interest (BCRP, OATP1B1/B3, BSEP) in the light of the eTOX project [55]. Using only ChEMBL version 17 gave us approximately 200 BCRP inhibitors, while systematically checking articles in PubMed allowed us to retrieve a data set of 970 compounds [Montanari F, Unpublished Data]. An even more striking example is OATP1B1/B3, where an article reports the bioactivity data for 2000 compounds tested in the same assay [56]. However, this paper is not tracked in ChEMBL version 17 and thus only 285 and 258 distinct compounds are present in the database as OATP1B1 and OATP1B3 modulators, respectively [Kotsampasakou E, Unpublished Data]. For BSEP, 47 compounds have a reported activity end point in ChEMBL version 17 while in the work of Warner *et al.* [57] 624 compounds were measured for inhibition of BSEP in one single assay [Pinto M, Montanari F, Unpublished Data]. This discrepancy is most probably inherently linked to the aims and scope of the publications indicated, as they ask for new compounds rather than computational analysis of large-scale data sets. Therefore, in order to compile the most up-to-date compound collection we recommend to search for bioactivity data through all available sources, including literature search with tools such as SCAIVIEW developed by the Fraunhofer institute [58].

To improve the data completeness in the open domain, several approaches might be considered. The most straightforward approach would be to encourage EBI to expand the range of journals from which the pharmacological data are extracted. Moreover, authors could be provided with the possibility to directly upload their data to ChEMBL as an SD file (sdf) once their

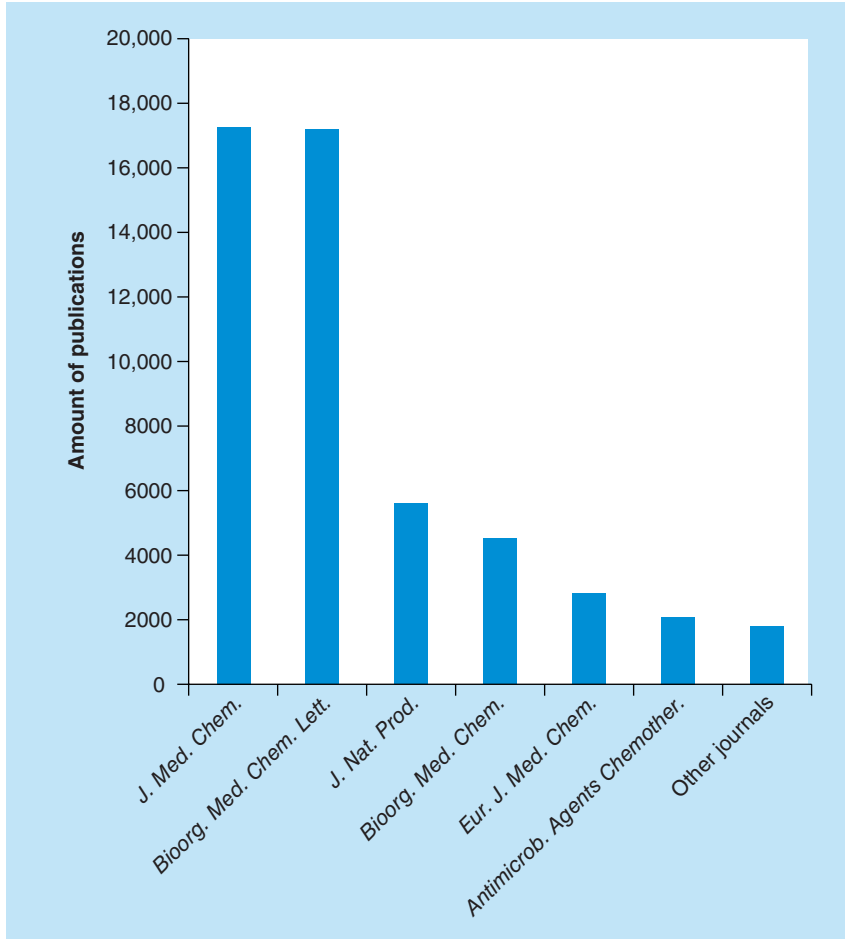



**Figure 4. The amount of publications from different journals included in ChEMBL DB version 17.** A list of other journals and corresponding amount of publications are given in the **Supplementary Data**.

manuscript is accepted. Even further, authors who used public money for conducting their research could be forced to deposit their data in a public depository.

Chemists and pharmacologists should get used to provide their research data the same way crystallographers are uploading protein structures to the PDB server. Although it would raise several questions on which structure format to use and on which vocabulary to agree on, it is feasible to create a domain of reusable open pharmacological data suitable for computational studies, and both ChEMBL and PubChem could become such databases.

## Expand the structural & biochemical space: does individual data upload serve this purpose?

We have broadly discussed several issues arising from inconsistency of public data, their quality and incompleteness. The availability to upload in-house data sets to open-access data sources could help to solve some of these problems, if accompanied by the agreement on and use of certain standards. Up to now, only a few open-source initiatives allow any organization or research group to contribute data. One of them is PubChem. The conditions for being able to deposit an assay are that the assay does not correspond to a virtual experiment and that the used methodology is properly described, either in a publication, or in the appropriate field of the submission form. Finally, the compounds tested should already be recorded in the PubChem substances database. Organizations may, therefore, register previously their small molecules using the sdf format to the PubChem substances before submitting their assay results.

Upon data submission, automated data validation is performed to ensure that the submitted assay data follows the data type description. PubChem staff also performs a final check on the data before attributing an assay identifier to the new bioassay and releasing it publicly [1]. It is possible to delay the publication of a new entry in the BioAssay database, namely if the results are linked to a paper which is not yet published. In this case, the content of the bioassay will be held from the users in a way that they are able to see the assay in their queries, but cannot access its content.

As for the assay results themselves, PubChem requires from the authors a summary result for each tested compound: a bioactivity outcome ('Active', 'Inactive', 'Unspecified', 'Inconclusive', and so on) and a bioactivity score. The bioactivity outcome is assigned subjectively by the depositor and may vary between identical assays performed by different laboratories.

To allow more flexibility to the depositors, the submission form corresponds to a loose data format, essentially when it comes to assay description [59]. As a result, similar assays are often described in different ways by different authors, making the compilation of data sets very complicated for the users if they want to consider data from different HTS runs.

Another public database, BindingDB, provides two possible ways for sharing data through it. The first is designed to connect already published bioactivity data to BindingDB: the online form for depositing data requires to insert the PubMed ID of a published article or to fill in corresponding fields manually (title, journal, issue date, institution credentials, author names and compound series) if the PubMed ID is missing. A second option allows the uploading of unpublished data using an Excel sheet provided on the website.

In both cases, compounds should be defined as SMILES strings and bioactivity data should be assigned to each of them together with the description of the experimental conditions. Furthermore, uploaded data will be reviewed and compounds, targets and assays already existing in public databases will be crosslinked via PubMed, ChEMBL and PubChem identifiers. New instances will receive a unique doi within BindingDB. In additon, while depositing data, the user can request to postpone open access to it until the respective publication is made available.

However, submission forms in BindingDB do not require rigorous description of assays and assay conditions. Furthermore, definition of compound structures using SMILES strings could cause errors in determining duplicated compounds. Although it might be tempting to perform a fast string comparison to find identical molecules in a database, actually different libraries or software could encode the same molecule in distinct (although valid and canonical) SMILES string.

In the case of ChEMBL, data upload by the authors themselves is not allowed yet, but larger data sets of special interest to the community might be donated. Pharmaceutical companies (e.g., GlaxoSmithKline [Brentford, UK], Novartis [Basel, Switzerland]) and institutions (e.g., Harvard [MA, USA]) donated

their malaria screening sets, and others donated kinase screening database [60]. However, in those cases, data curation and annotation is always performed by the ChEMBL curators to prevent errors and to maintain consistency. Still, more flexibility in this respect would help to increase the available amount of data points. Nevertheless, this comes with the risk of increased error rates [Senger S, Pers. Comm.].

Taken together, data upload to PubChem as well as to BindingDB is already possible, but would require more stringent rules in order to guarantee consistency among structure representations and bioassay descriptions.

## Future perspective

In our opinion there are three ways in which sharing of the published bioactivity data could be implemented by high-impact journals, open-access databases and European-funded projects. Each of these, however, has its advantages, drawbacks and potentials.

First, some publishers already allow the authors to provide their pharmacological and computational data sets along with their publications. In this case, authors have great freedom in the submission format of the chemical structure information and assay description. Most often compound structures are uploaded as an image, text or on rare occasions, in a machine-readable format in supplementary information. When journals require mol or mol2 file formats, structures are usually uploaded one by one and the biological activities are provided in a separate file. Since submitting a data set including a numerical field for activity value is not required, mistakes often arise in assigning of these values. Authors are neither obliged to use terms of the existing ontologies when describing the biological assay in their manuscripts. Consequently, this type of data sharing complicates immediate use of the published pharmacological data sets for computational studies. Ideally, publishers could require authors to provide any SAR table present in a paper in machine-readable format in supporting information, for example, as a list of standard InChI strings, canonical SMILES, archive of mol/mol2 files or a separate sdf, together with the numerical activity value; thus, facilitating the utilization of the published data. In addition, we want to highlight, that by providing the data sets as part of article's supporting information, authors could also share with the community their data on true inactives. As outlined before, for HTS data in PubChem such valuable data points are in stock. This is, however, not the case for literature-retrieved bioactivity data, as negative results are rarely reported. However, authors should be encouraged by publishers to provide inactive data because these undisclosed data are also useful for medicinal chemists to prevent retesting inactive compounds and, on the other hand, chemoinformaticians rely on the disposability of both, actives and inactives, in order to train their models and to be predictive in both directions.

Another possible solution would be that computational and biological studies identifying new chemical compounds and/or reporting bioactivities on certain targets are only accepted for publication in peer-reviewed journals if the utilized data sets are also uploaded to open-data sources at the same time. That could be implemented in a similar manner as for depositing crystallographic data in the PDB.

Since some open databases already allow authors to deposit their pharmacological data sets, we suggest that a clear, stringent protocol should be defined in each of the databases for the structure and bioactivity data submission, in order to improve the quality of published data. Structural information could be uploaded in the SMILES format, but for every compound the submitter should in addition define the InChI string and the UniChem ID [61]. Use of the InChIs would help the search algorithms, which are internally implemented in each database, to determine the correct structure of the compounds. Moreover, using these data format minimizes level of errors in the representation of chemical structures and allows fast extraction of the chemical information. If the UniChem ID is available, it will automatically crosslink identifiers of the existing compounds within databases, such as ChEMBL, PubChem, IUPHAR, ChEBI and ZINC, thus minimizing the amount of chemical duplicates in public sources, while new compounds would automatically receive a new identifier upon submission. It will be sufficient for the authors to deposit their data sets in one of the public databases, such as PubChem, ChEMBL or BindingDB, since the sources are interconnected.

A description of the biological assay should be given by the authors as precisely as possible, to allow fast and appropriate comparison of local data sets. In addition, the program interface for the assay submission to open databases could implement several fields corresponding to

**Key Term**

**Semantic integration:** Effort to interrelate diverse, often heterogeneous data sources by modeling the relationships of concepts within these sources through ontologies.

different assay details, such as biological target, target organism, assay organism, measured effect, reported end point value, and so on. Depending on the requirement of the database, standards and ontologies will be needed that capture the relations of concepts for describing a certain assay. It has to be pointed out that – according to the open biological and biomedical ontologies foundry principles [62] – new concepts should be adapted to existing ones rather than creating a multiplicity of new ontologies [63]. With respect to bioassays, the BAO is most suitable and is used by a number of projects such as PubChem, AstraZeneca's HTS annotation, and open PHACTS (full list is available online [64]). Therefore, while filling in the online assay submission form we recommend the author define all the entries using BAO terms as a vocabulary. In addition, we propose that authors use minimum information about a bioactive entity checklist as a standard for publication of bioactivity data [65]. We are convinced that those guidelines for description of synthesis and subsequent analysis of any potential bioactive entity in a publication could be used as gold standard in the field.

Third, one could imagine that data deposition in the public domain occurs independently from upload into distinct data sources, ideally provided in a schema-free and highly compatible format, such as RDF. Data provided in the RDF format would allow immediate **semantic integration** of the data into the linked open-data cloud. When providing data in RDF, the users first have to list the concepts (proteins, metabolites, genes, assays, and so forth) and the relations ('activates', 'inhibits', 'is toxic against', and so on) that link those concepts in their data. Next, the authors are prompted to use existing ontologies and vocabularies (using either ontology search engines or following RDF guidelines) that identify and describe concepts and relations between them. With this in hand, it is possible to convert these data into the RDF format using globally accepted identifiers for the used concepts, thus allowing fast comparison and integration of the data. The RDF versions of data could be deposited on the webpage of a scientific group and respective open-data sources could be notified about it. The open databases could then harvest the data directly from this website and integrate them into all other sources. These would create a highly interconnected, up-to-date, dynamically linked data cloud serving drug discovery. However, one disadvantage is the initial investment to establish an infrastructure to convert data to RDF. Another difficulty might be the search for identifiers for the concepts in common ontologies. In case some entries cannot be mapped to the existing ontologies, the user would have to make the new identifier openly available [66], which might not be trivial.

Finally, a major advantage of all these approaches (providing a data set in a machine-readable format in the **Supplementary Data**, uploading it to open databases or transferring it into RDF format) is that pharmacological data sets made publicly available will be immediately recognized and appreciated by the chemoinformatics community. These compound collections will appear in integrated systems such as Open PHACTS, or in interconnected databases enabling immediate access to data and allowing their free and comfortable use. The data will further be used for building computational models and the original article or the website will be cited. Thus, the data donators and the publishers will benefit through increasing awareness of these research contributions, which will also impact their respective bibliographic metrics.

## Supplementary data

*To view the supplementary data that accompany this paper please visit the journal website at:* **www.future-science.com/doi/full/10.4155/FMC.14.13**


## Acknowledgements

*The authors are grateful to M Pinto and E Kotsampasakou for providing data on BSEP, OATP1B1 and OATP1B3.*

## Financial & competing interests disclosure

*The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under Grant Agreements number 115191 and number 115002 (Open PHACTS and eTOX), resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007–2013) and EFPIA companies' in kind contribution. The authors also gratefully acknowledge the FWF doctoral program #W1232 (Molecular Drug Targets) and the FWF Special Research Program 35 (Transmembrane transporter in health anhd disease, project #F3502) for the financial support of this research, as well as the financial support provided by the University of Vienna, doctoral programme Biopromotion. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.*

*No writing assistance was utilized in the production of this manuscript.*

**Executive summary**

- When used properly (thoroughly), open data are a powerful means to enrich the chemical and biological space for chemoinformaticians.
- Current databases show limited coverage due to the way they are created and updated. Allowing individual data upload by scientists could improve the coverage, especially with respect to negative data.
- Biological data suffer from lack of standardization. Publishers are requested to enforce authors to describe their biological assays in a standardized way.
- Compound structures and bioactivity values that are uploaded to open databases also appear in integrated platforms, such as Open PHACTS and thus are more often used by the chemoinformatics community .

## References

Papers of special note have been highlighted as:
▪ of interest
▪▪ of considerable interest

1 Wang Y, Bolton E, Dracheva S *et al.* An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* 38(Suppl. 1), D255–D266 (2010).

2 Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40(D1), D1100–D1107 (2012).

3 Barnes Mr, Harland L, Foord Sm *et al.* Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discov.* 8(9), 701–708 (2009).

4 ChEMBL: European Bioinformatics Institute. www.ebi.ac.uk/chembldb

5 UniProt Consortium. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res.* 41(D1), D43–D47 (2013).

6 FTP directory. ftp://ftp.ebi.ac.uk/pub/databases/chembl/VM/myChEMBL/current

7 Seiler KP, George GA, Happ MP *et al.* ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36(Suppl. 1), D351–D359 (2008).

8 Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35(Suppl. 1), D198–D201 (2007).

9 Rosania Gr, Crippen G, Woolf P, States D, Shedden K. A cheminformatic toolkit for mining biomedical knowledge. *Pharm. Res.* 24(10), 1791–1802 (2007).

10 Sharman JL, Benson HE, Pawson AJ *et al.* IUPHAR-DB: updated database content and new features. *Nucleic Acids Res.* 41(D1), D1083–D1088 (2013).

11 IUPHAR-DB. www.iuphardb.org

12 Ozawa N, Shimizu T, Morita R *et al.* Transporter database, TP-Search: a web-accessible comprehensive database for research in pharmacokinetics of drugs. *Pharm. Res.* 21(11), 2133–2134 (2004).

13 TP-Search. http://125.206.112.67/tp-search

14 Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* 31(1), 294–297 (2003).

15 Seeman P. Drug dissociation constants for neuroreceptors and transporters. In: *Receptor Tables*. SZ Research, Toronto, Canada (1992).

16 Cutler D, Barbier A, Pestell K. In Brief. *Trends Pharm. Sci.* 23(6), 258–259 (2002).

17 BindingDB. www.bindingdb.org

18 Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res.* 28(1), 235–242 (2000).

19 Knox C, Law V, Jewison T *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* 39(Suppl. 1), D1035–D1041 (2011).

20 Drugbank. www.drugbank.ca

21 Chembank. http://chembank.broad.harvard.edu

22 Chemspider. www.chemspider.com

23 Azzaoui K, Jacoby E, Senger S *et al.* Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discov. Today* 18(17–18), 843–852 (2013).

24 Open PHACTS. www.openphacts.org

25 Williams AJ, Harland L, Groth P *et al.* Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* 17(21–22), 1188–1198 (2012).

▪▪ **Discusses semantic integration of heterogeneous chemical and biological data.**

26 Degtyarenko K, De Matos P, Ennis M *et al.* ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 36(Suppl. 1), D344–D350 (2008).

27 Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genetics* 25(1), 25–29 (2000).

28 Kelder T, Van Iersel MP, Hanspers K *et al.* WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* 40(D1), D1301–D1307 (2012).

29 Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res.* 28(1), 304–305 (2000).

30 ConceptWiki. http://ops.conceptwiki.org

31 Open PHACTS explorer. www.openphacts.org/explorer

32 McNaught A. The IUPAC international chemical identifier: InChl – a new standard for molecular informatics. *Chem. Int.* 28(6), 12–16 (2006).

33 Williams AJ, Ekins S, Tkachenko V. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discov. Today* 17(13–14), 685–701 (2012).

34 Hu Y, Bajorath J. Growth of ligand–target interaction data in ChEMBL is associated with increasing and activity measurement-dependent compound promiscuity. *J. Chem. Inf. Model.* 52(10), 2550–2558 (2012).

35 Kalliokoski T, Kramer C, Vulpetti A, Gedeck P. Comparability of mixed $IC_{50}$ data – a statistical analysis. *PLoS ONE* 8(4), e61007 (2013).

▪ **Comprehensively analyze the variability of pairs of $IC_{50}$ and $K_i$ values reported in ChEMBL.**

36 Cheng Y, Prusoff WH. Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition ($IC_{50}$) of an enzymatic reaction. *Biochem. Pharmacol.* 22(23), 3099–3108 (1973).

37 Tsareva DA, Ecker GF. How far could we go with open data – a case study for TRPV1 antagonists. *Mol. Inform.* 32(5–6), 555–562 (2013).

38 Bemis Gw, Murcko Ma: The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39(15), 2887–2893 (1996).

39 Zerhouni E. Medicine. The NIH Roadmap. *Science* 302(5642), 63–72 (2003).

40 Vempati UD, Przydzial MJ, Chung C *et al.* Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO). *PLoS ONE* 7(11), e49198 (2012).

41 BioAssay ontology. http://bioassayontology.org/wp

42 Malone J, Holloway E, Adamusiak T *et al.* Modeling sample variables with an experimental factor ontology. *Bioinformatics* 26(8), 1112–1118 (2010).

43 Zdrazil B, Pinto M, Vasanthanathan P *et al.* Annotating human P-glycoprotein bioassay data. *Mol. Inform.* 31(8), 599–609 (2012).

44 Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schurer SC. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinform.* 12, 257 (2011).

45 Rohrer SG, Baumann K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* 49(2), 169–184 (2009).

46 Feng BY, Shelat A, Doman TN, Guy RK, Shoichet BK. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* 1(3), 146–148 (2005).

47 Shoichet BK. Screening in a spirit haunted world. *Drug Discov. Today* 11(13–14), 607–615 (2006).

48 Bajorath J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1(11), 882–894 (2002).

49 Roche O, Schneider P, Zuegge J *et al.* Development of a virtual screening method for identification of 'frequent hitters' in compound libraries. *J. Med. Chem.* 45(1), 137–142 (2002).

50 Murray-Rust P, Mitchell JB, Rzepa HS. Communication and re-use of chemical information in bioscience. *BMC Bioinform.* 6, 180 (2005).

▪▪ **Discusses human-introduced errors in published chemical information (structures, spectras, synthetic details).**

51 Tiikkainen P, Bellis L, Light Y, Franke L. estimating error rates in bioactivity databases. *J. Chem. Inf. Model.* 53(10), 2499–2505 (2013).

▪ **Authors show that errors in the chemical structure representation are the most frequent ones in public databases.**

52 Gregori-Puigjane E, Garriga-Sust R, Mestres J. Indexing molecules with chemical graph identifiers. *J. Comp. Chem.* 32(12), 2638–2646 (2011).

53 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comp. Sci.* 28(1), 31–36 (1988).

54 Ihlenfeldt WD, Gasteiger J. Hash codes for the identification and classification of molecular structure elements. *J. Comp. Chem.* 15(8), 793–813 (1994).

55 Steger-Hartmann T, Pognan F, Sanz F, Diaz C. *In silico* prediction of *in vivo* toxicities (eTOX) – the Innovative Medicines Initiative approach. *Toxicol. Lett.* 189(Suppl.), S258 (2009).

56 De Bruyn T, Van Westen GJ, Ijzerman AP *et al.* Structure-based identification of OATP1B1/3 inhibitors. *Mol. Pharmacol.* 83(6), 1257–1267 (2013).

57 Warner DJ, Chen H, Cantin LD *et al.* Mitigating the inhibition of human bile salt export pump by drugs: opportunities provided by physicochemical property modulation, *in silico* modeling, and structural modification. *Drug Metab. Dispos.* 40(12), 2332–2341 (2012).

58 Friedrich CM, Dach H, Gattermayer T, Engelbrecht G, Benkner S, Hofmann-Apitius M. @neuLink: a service-oriented application for biomedical knowledge discovery. *Stud. Health Technol. Inform.* 138, 165–172 (2008).

59 Schurer SC, Vempati U, Smith R, Southern M, Lemmon V. BioAssay ontology annotations facilitate cross-analysis of diverse high-throughput screening data sets. *J. Biomol. Screen.* 16(4), 415–426 (2011).

60 Hersey A, Senger S, Overington JP. Open data for drug discovery: learning from the biological community. *Future Med. Chem.* 4(15), 1865–1867 (2012).

▪▪ **Supports the idea of direct submission of bioactivity data to the public domain and provides an example of data donation to ChEMBL.**

61 Chambers J, Davies M, Gaulton A *et al.* UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.* 5(1), 3 (2013).

62 Smith B, Ashburner M, Rosse C *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11), 1251–1255 (2007).

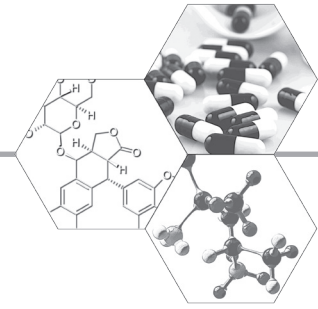
63 Zdrazil B, Chichester C, Zander Balderud L, Engkvist O, Gaulton A, Overington JP. Transporter assays and assay ontologies: useful tools for drug discovery. *Drug Discov. Today Technol.* (In Press) (2014).

64 The BioAssay Ontology. http://bioportal.bioontology.org/ontologies/BAO

65 Orchard S, Al-Lazikani B, Bryant S *et al.* Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10(9), 661–669 (2011).

▪▪ **Presents the guidelines for submission of bioactivity data using controlled vocabularies or ontologies.**

66 Persistent uniform resource locators. http://purl.oclc.org/docs/index.html





# Open source drug collaborations: a rational design approach

A decade ago, many biologists and legal scholars thought that open source (OS) methods would revolutionize drug discovery. These predictions were clearly disappointed. Not surprisingly, some scholars now take the opposite view that OS methods and drug discovery are inherently incompatible. This article argues for balance. Economists have observed that OS methods offer both advantages and disadvantages compared with conventional institutions. This article reviews what economists have learned regarding the strengths and weaknesses of OS methods and identifies specific drug-discovery tasks where OS methods are likely to work especially well. Successful OS collaborations should carefully focus on segments of the drug-discovery pipeline where the advantages of OS methods are particularly relevant and existing commercial methods are known to work badly. The article concludes by offering detailed suggestions for how existing OS software models can be modified and extended to conduct effective drug discovery.

**Stephen M Maurer**

Goldman School of Public Policy & Berkley Law School, University of California Berkley, CA, USA
Tel.: +1 510 725 5168
Fax: +1 510 643 9657
E-mail: smaurer@law.berkeley.edu

The rise of **open source** (OS) software, which delivers sophisticated commercial-grade products at little or no cost to users, struck most economists and legal scholars as a kind of miracle. The rub, with minor exceptions, is that the miracle is still almost entirely confined to software. The reason is not hard to see. Software, in economic jargon, is an 'information good': once a computer program is invented additional copies cost nothing. The genius of OS lies in realizing that software could be invented by volunteers working outside the patent system. Since copying is easy, producing and distributing the finished product to users was trivial at that point.

Bringing OS methods to the drug industry is clearly a harder problem. This is because drugs are only approximately an information good – indeed, manufacturing costs are approximately the same as R&D expense on a per-pill basis (non-physical costs, such as advertising and promotion, are artifacts of the patent system and would be much smaller for OS drugs). Even so, the 'information content' locked up in the average drug is still far larger than for conventional products such as automobiles or refrigerators. In the late 1990s, scholars almost always saw this as a hint that OS methods for producing software should work for drug discovery as well. The problem, a dozen years later, is that OS drug collaborations remain painfully few and far between. The case for OS drug discovery remains in the old Scottish phrase, 'not proven.' Scholars should accept much of the blame for this. Almost all of the early literature focused on asking how OS licenses that had originally been based on software could be rewritten to meet the very different demands of patent law. However necessary, this work almost always ignored the detailed physical differences between software and pharmaceutical R&D, most notably the fact that each step along the drug-discovery pipeline involved distinct scientific challenges and, more often than not, completely different researchers. In even less detail did they describe how the OS model should be customized for specific steps.

This article argues that these legal-formalist efforts were always doomed for failure. This is because OS is only secondarily a question of **intellectual property** (IP) and licensing law. Instead, the basic issues are economic. For this reason, it is better to start from an unsentimental look at the incentives that are known to drive conventional OS software production and then look for locations along the drug-discovery pipeline where similar mechanisms ought to work. As with drug discovery itself, our goal should be to move beyond intuition to rational, evidence-based design principles.

The work is underway. This article begins ('The need for rational design') by reviewing the rich theoretical literature that economists have developed over the past decade to identify the specific circumstances under which OS software collaborations are simultaneously feasible and likely to deliver better social outcomes than conventional IP incentives.

FUTURE SCIENCE part of fsg

**Key Terms**

**Open source:** R&D methods based on licenses that permit the royalty-free use of intellectual property-protected information. Many open source licenses require users who improve or extend the software to make these enhancements available to others, if at all, on identical open source terms.

**Intellectual property:** Government-enforced, statutory protection for authors and inventors against unlicensed copying. Includes patents for machines, chemicals and manufacturing processes; copyright for writings, images and software; and – in many jurisdictions outside the USA – special statutes protecting data. Biotechnology and pharmaceutical companies typically rely on patent rights far more than other forms of intellectual property.

It also discusses the secondary (but still significant) role that licenses play in stabilizing OS. The article then proceeds ('Negative reasons to use open source') with a survey of the drug-discovery pipeline and identifies multiple instances in which OS methods let economic actors avoid the well-known pathologies of IP. A third section ('Positive reasons to use open source') identifies additional locations along the pipeline where OS offers affirmative benefits that conventional IP incentives are hard-pressed to match. A fourth section ('Additional reasons to use open source') asks how OS drug-discovery collaborations should be organized and returns to the role of licenses. A fifth section ('Entrepreneurship') offers practical advice for researchers trying to organize working OS collaborations.

## The need for rational design

The fact that OS drug discovery has enjoyed only limited success to date has invited reaction. Belatedly, many observers now argue that OS drug discovery is impossible. First, they point out that OS collaborations must compete with existing IP rewards for volunteers and resources. Given the fabulous sums that can be earned by patenting drugs, why would anyone share discoveries *gratis*? While this argument has considerable merit, however, it is only true on average. In particular, there are still many instances in which OS methods are comparably appealing to patent rewards. This is particularly true where OS collaborations increase members' ability to make separate, patentable discoveries. Second, skeptics point out that drug discovery is much more expensive than writing software. This suggests that companies cannot support R&D unless they can recover their costs from consumers. Once again, the argument has a grain of truth. That said, however, it proves too much. After all, the software industry also spends billions of dollars creating its products. Despite this, OS continues to thrive.

The deep flaw in both arguments is that they assume that open and closed source developers are locked in some Manichean struggle which can only end when one or the other perishes. In the software realm, at least, this is clearly false. To the contrary, open and proprietary code developers routinely coexist and often need each other. Indeed, the most famous OS product of all – LINUX – supports a large for-profit sector that earns its living by customizing and trouble-shooting software for commercial users. From this standpoint, schemes that imagine monocultures comprised entirely of open or proprietary code are dangerously misleading. More usually, proprietary and open code coexist in ways that allow the former to fund the latter. The question then becomes which R&D areas should be open and which proprietary? This article argues that open and proprietary methods each have distinct areas of strength and weakness. The trick is to encourage OS in those circumstances where its strengths are likely to shine.

## Negative reasons to use OS: opting out of IP

These days, the case for IP is stated so insistently that it is easy to forget that economic actors often try to avoid patents and copyright altogether. For big corporations, such as AT&T or General Electric, this strategy often involves publishing new inventions in in-house journals. This 'defensive publishing' turns otherwise patentable discoveries into 'prior art' so that later independent inventors cannot patent the invention and charge for its use.

Within drug discovery, a similar logic was spectacularly responsible for the so-called 'single nucleotide polymorphisms (SNP) Consortium' of the late 1990s. In that era, big pharmaceutical companies observed that SNPs were easy to find, potentially patentable and provided little or no guidance on how to find new chemical entities. This created an obvious danger that big pharmaceutical companies could sink hundreds of millions of dollars into drug discovery before finding out that it had infringed an existing SNP patent. Trying to negotiate a fair royalty at this point would be almost impossible. The clear business solution was to prevent anyone (including the pharmaceutical companies themselves) from obtaining SNP patents. This was done by funding an academic consortium to find and defensively publish as many SNPs as it could. This deliberate effort to opt out of the patent system ultimately cost big pharmaceutical companies and various biotechnology companies more than £20 million [101]. Not surprisingly, the need for such models continues. For example, five big pharmaceutical companies have joined the Wellcome Trust and the Canadian government in funding the US$50 million Structural Genetics Consortium. This Consortium is annually responsible for discovering and depositing massive amounts of protein structure data into the public domain [102].

The SNP Consortium was, famously, dedicated to opting out of IP altogether. In principle, however, big pharmaceutical companies could have retained their IP and then offered their data to the world under OS licenses. Would this have yielded more or less benefit to society? This issue is returned to below. In the meantime, it is important to note that companies are only likely to adopt OS when IP's drawbacks exceed its benefits. The rest of this section discusses specific scenarios in which this is likely to happen.

### ■ When IP does not matter

Traditional IP business models depend on bringing goods to market. However, developing-world markets are often too poor to pay the high drug prices that would induce big pharmaceutical companies to invest in R&D for diseases such as Chagas disease or dengue fever. In these cases, IP should be more or less irrelevant. Worse, it may be malicious. Suppose that the Gates Foundation decides to create and fund its own nonprofit company to pursue promising ideas for curing dengue fever. If all of the good ideas have already been patented, the Gates Foundation will have to pay for development rights. This might be a good idea if there are so many competing patents that IP owners cannot charge the Foundation more than their ideas are worth. However, this condition is unlikely to hold for neglected diseases that are, by definition, understudied. If only a handful of ideas exist, IP owners could end up charging Gates far more than their ideas are worth.

Here, the obvious solution is to find new ideas and make them freely available on OS terms to would-be developers. Indeed, OS supplies a twin benefit: in addition to developing OS ideas directly, the Gates Foundation can threaten to use them when it negotiates with the owners of IP-protected ideas. In either case, increasing the supply of OS ideas cuts the Gates Foundation's development costs.

This is not to say that IP rights can never play a role in funding neglected disease research. For example, some economists have proposed schemes in which funding agencies offer subsidies to help developing countries pay higher prices for patented drugs [1]. However, IP is not really essential to these schemes since prizes can achieve identical results. More fundamentally, even the poorest communities can pay some money and these sums can be substantial for diseases such as malaria or tuberculosis. IP can potentially capture these rents so that the funder's limited funds go further. Even so, the case for IP in neglected disease research is often doubtful. In these cases, sponsors should seriously consider funding OS collaborations to find and validate new drug targets and lead compounds. This is particularly true where big pharmaceutical companies have no commercial incentive to perform follow-on steps such as compound optimization, animal testing or clinical trials. If nonprofit or government agencies are going to pay for the work in any case, IP rights add little and could easily make development more expensive.

### ■ Transaction costs

As Scotchmer has emphasized, the drawbacks of IP will often outweigh the benefits when multiple independent researchers contribute improvements to a shared product [2]. One might think that each researcher can fund their research by selling improvements on the open market. This is far from obvious, however, if each researcher also has to buy improvements so that their net income is zero. In this case, IP does nothing to fund innovation while forcing participants to incur a variety of 'transaction costs', such as billing customers, negotiating IP licenses and, inevitably, suing other IP owners. OS eliminates this drag by blocking IP so these expenses disappear.

As with computer software, the average drug candidate is frequently developed through a series of sequential innovations. This suggests that Scotchmer's scenario provides a strong argument for OS development. This is most likely to happen in the drug-discovery pipeline's early stages, where multiple researchers often trade small insights and improvements to a shared idea. Under the current system, IP concerns frequently prevent or delay sharing between academic researchers across universities. Yet this research is often so embryonic that the prospect of IP, let alone meaningful royalties, is remote. OS arrangements offer a natural way to avoid this impasse.

Using OS methods to develop drug candidates could still be counterproductive if IP rights were essential or even just useful in funding subsequent development steps. Formally, this problem can be solved by letting collaboration leaders award exclusive IP rights to companies that commit to invest in R&D. Still, such arrangements would be markedly less liberal than the openness found in OS software licenses and it is possible to imagine OS ideologues objecting. At the same time, a mature OS community should be willing to do whatever is necessary to turn its ideas into practical drugs. Furthermore, tough negotiators could demand agreements that hold exclusive

**Key Term**

**Information goods:** Goods whose total life-cycle cost is dominated by R&D expense. Software is an extreme example. Pharmaceuticals have very high R&D expenses and are often analyzed as information goods.

IP rights to a bare minimum. This could be done, for example, by limiting the companies' licenses to shorter terms than the patent statute provides; requiring companies to contribute cash and other resources to nonprofit research; or guaranteeing that a successful drug would be available at low or zero price to needy patients and/or entire countries.

### ■ Network effects

The traditional rationale for IP is that it lets owners share in the benefits that their discovery confers on consumers. This result only holds, however, if the price that consumers are willing to pay for a particular invention is proportional to its value or, more precisely, the improved performance that it delivers over earlier technologies. During the 1980s and 1990s, however, economists came to realize that this was not always the case. Instead, experience with the electronics and software industries showed that consumers often prefer to join 'networks' that use a common standard that allows them, say, to trade word processing documents. When this 'network externality' is large, consumers can and do choose popular standards over technically superior alternatives. This, of course, immediately decouples IP rewards from actual performance. In extreme cases, the IP's value is essentially identical to the value that consumers place on having a single standard. At this point, the IP reward becomes a kind of windfall that has little or nothing to do with owners' efforts to build a better invention.

But why should consumers pay this? Since the network effect exists regardless of IP, consumers who adopt 'open standards' can enjoy the network effect without paying any royalties at all. For large network effects, this is almost always a good trade. This does not, of course, mean that real markets always adopt open standards. After all, consumers are poorly organized and markets often 'tip' to a winning standard so quickly that open collaborations have no time to organize. Nevertheless, it makes sense for consumers to pre-empt this outcome where they can.

One might think that this problem was narrowly confined to the electronics industry. In fact, however, network effects are nearly as prevalent in drug discovery. Here, leading examples include model organisms, cell lines and standard biological parts [3]. Consider cell lines. The more scientists use a line, the more experience they have with it. This lets them share tricks and even special equipment for growing the line, trade information about known peculiarities, store cultures in community-wide repositories and compare new experiments with published results. These effects are even larger for commercial users who have invested large sums to optimize production around specific lines and/or have previously convinced regulators that particular lines are safe. The net result is that most researchers prefer to use the same handful of lines as everyone else – a network effect. This explains why most stem cell researchers, for example, only use a dozen or so lines from the more than 600 available. Even more strikingly, most of these lines date from the first dozen or so that were originally cultured in 1997. Scientifically, there is no reason to think that these lines are 'better' than others. Instead, their popularity is only explicable in terms of economics – in this case, network effects.

The existence of network effects raises deep strategic questions for companies trying to develop new stem cell therapies. Given that stem cell lines are *ex ante* indistinguishable, there is very little reason to choose one over another. At this point, even proprietary lines will earn few royalties. This changes, however, once a line becomes popular and acquires an experience base. At this point, choosing a popular line makes experiments cheaper and more successful. And the line's owner will set royalties accordingly. For reasons already stated, however, this value does not come from the owner. It comes from the experience base – that is, the researchers themselves. Standardizing around open lines assures researchers that no one can charge for this experience. Whether open lines are good for companies is a more difficult question. On the one hand, large companies could decide to hoard data. This could lead to a dynamic in which companies that invent more products have lower R&D costs so that they can invent still more products until their cost advantage becomes insurmountable. On the other hand, second- and third-place companies could decide that OS sharing is the only way to overcome the front-runners' information advantage. These sharing arrangements are explored further below.

## Positive reasons to use OS: altruism, reputation, own use & education

Inventors invariably confer benefits on consumers. IP allows inventors to recover or, in economics jargon, 'appropriate' part of this value. This makes sense for costly inventions that would otherwise not be invented. Many people find this

rationale so compelling that they ignore its loopholes. The magic of OS lies mostly in reminding us that these loopholes are substantial.

Start with the early days of OS software when collaborations were dominated by volunteers who received little or no commercial support. In the old economy, volunteers could never have covered the costs of manufacturing, say, an automobile for anyone who requested one. Needless to say, the economics are far more favorable for **information goods**. Here, only the first copy is expensive: having written an operating system once, additional copies cost nothing at all. This suggests that an average copy can be made for far less than consumers are willing to pay. In this context, letting IP owners grab approximately 50% of the benefit that consumers receive begins to look like overkill. Then too, the work of making even the first copy is often 'granular,' for example, can be split up into discrete tasks that an individual can perform in a few hours. Crucially, this investment is so small that cash-based incentives are no longer necessary. Softer incentives will do.

The rise of classical OS software collaborations is largely the story of how organizers identified and learned to harness these non-traditional motives. Some of these incentives are purely intrinsic, that is, provide rewards directly from the activity itself. Examples include the fun of programming and the joy of helping others (altruism). Other rewards are extrinsic. These include the quest for status and reputation, the desire to learn new skills, and the ability to demonstrate those skills to would-be employers. These relatively weak incentives would never have persuaded any single contributor to write an entire operating system. Because of granularity, however, this is unnecessary.

The trick for OS discovery is to find populations that are simultaneously susceptible to weak incentives and capable of doing useful scientific work. With respect to the first criterion, it is not hard to imagine that biologists would respond to the same weak incentives that computer scientists do, particularly for non-commercial projects where competing patent incentives are negligible. Even so, collaboration designers will need to think carefully about how to design incentives for each specific target group. We know from the software realm that incentives that drive some collaborations (and their volunteers) are almost irrelevant for others. Knowing which incentives to choose, in turn, says a great deal about how the consortium should be organized. For example, collaborations driven by reputation or signaling should feature honorifics and hierarchies. This will increase promotion prospects for members who make large contributions. Similarly, collaborations based on signaling should create opportunities for volunteers to work with corporate sponsors. More generally, collaboration organizers must be able to find out what prospective volunteers want and think up clever ways for them to get it. Unfortunately, this crucial skill is in short supply everywhere. This is especially true in grant-supported communities, where the experimenter's intelligence tends to be valued more than humbler virtues, such as 'pleasing the customer.'

The second criterion – finding a scientific project where volunteers can make useful contributions – is also important. In general, we expect weak incentives to be strongest where out-of-pocket costs are small. This is most likely to be true in the early stages of the drug-discovery pipeline. *In silico* research, which requires no reagents, is particularly promising in a world where computing power is cheap but human judgment is at a premium. However, many academic labs have facilities and even reagents that have already been paid for by grants. Students and faculty will often be able to scrounge such assets for OS research. Finally, drug libraries cost a great deal to establish but relatively little to share. In recent years, companies have frequently offered them to neglected disease researchers. It is also important to remember that the usefulness of research is always relative. Contributions that would be trivial for diseases typical of developed nations, such as rheumatism, can be extremely valuable for underfunded problems, such as malaria or dengue fever.

## ■ Scholarship

Claims that IP rights are essential to innovation are demonstrably false. For example, university researchers hardly ever patented their research prior to the 1970s. Instead, they were driven by other incentives, most notably reputation and (through reputation) promotion. Furthermore, researchers achieved these goals by publishing information far more often than hoarding it.

It is not surprising that such incentives produced institutions that foreshadow OS methods. Beginning in the 1940s, nuclear and particle physicists began building ultra-large databases that reported worldwide experimental values, recommended 'best values' where experiments conflicted, and exploited known physical laws

to calculate expected results for experiments that had not yet been performed. More recently, similar database projects have begun to enter drug research through so-called 'bio-wikis.' An estimated 2000 volunteer editors currently work on these projects [103].

Given that volunteers already work together to compile large databases, the transition to genuine OS production is probably short. The main challenge is purposefulness. Volunteer databases are notoriously based on whatever subjects members are enthused about. This means that a single physics volunteer's interest in, say, $^{8}$Li decay can distort the entire database. The difference for OS is that volunteers must subordinate personal interests to a single agreed end product. That said, the old physics databases have much to teach us. This is particularly true of their efforts to use existing data to predict future experiments. Strikingly, this goal is more or less identical to *in silico* biology's efforts to predict drug targets and lead compounds from existing data and simulations. This suggests that organizers of *in silico* drug-discovery collaborations should take academic incentives seriously. As with the physics database collaborations before them, this probably will include inventing institutions to document and publicize each volunteer's contributions. Detailed provisions for turning collaboration discoveries into academic papers – with explicit attribution for contributors – will be especially important. Collaborations that find a way to partner with leading science journals will be particularly good at convincing would-be volunteers that they will receive credit for discoveries.

### ■ Education, signaling & reputation

Many traditional OS software projects were driven by students who joined because they wanted to learn how to code and/or demonstrate skills to employers. Similar incentives should appeal to graduate and even undergraduate biology majors. In some cases – particularly for older, more established workers – non-monetary reputational benefits may also matter.

Student skills are particularly well suited to the wet chemistry and *in silico* discovery experiments that will be needed to find and validate drug targets and/or identify lead compounds in the early stages of drug discovery. These benefits could be still further amplified by attracting industry veterans to the collaboration. On the one hand, their presence would provide important guidance in practical drug discovery. University faculty are often weak in this area. On the other hand, industry volunteers would inevitably become talent scouts for employers. This would offer students who volunteered for OS collaborations a powerful incentive to demonstrate their wet chemistry and *in silico* skills.

### ■ Own use

Academic scientists often invent new tools for their own use. Once a tool has been invented, however, it is often cheap to share. At this point, even small benefits – for example, the reputation effects of publishing an article or the hope of receiving good ideas in return – often outweigh the costs. This explains, for example, why University of California (Berkeley, CA, USA) physicists went out of their way to spread particle accelerator technologies – 'cyclotron culture' – to other universities in the 1930s. Over 60 years later, the rise of OS methods spread this model to industry. Here, the most obvious example is the Apache community. Once corporate web designers have written a program for their own employer, sharing with the world costs very little. On the other hand, the benefits – reputation for the programmer, eliminating wasteful duplication for employers – are substantial. Far from being altruistic, sharing makes good business sense.

The tradition of sharing data and research tools is similarly strong in biology. It is difficult to be sure how this phenomenon would generalize to OS drug discovery. On the one hand, empirical studies demonstrate that sharing is often suppressed by donor fears that recipients will become stronger commercial or academic competitors [4]. These concerns will presumably increase to the extent that donors are asked to share their data and materials with entire collaborations. On the other hand, academics also report that sharing is deterred by 'transaction costs' such as obtaining permission from university authorities [4]. This deterrent effect should normally be smaller for OS, since donors who find transaction costs excessive for a single recipient are more likely to find them affordable when the promised sharing extends to multiple users.

### ■ Building practical collaborations

Many biologists respond favorably to projects that promise openness. Not surprisingly, these initiatives proliferated in the SNP Consortium's wake [5,6]. That said, most of the early collaborations were organized along very conventional lines with grant support from government and

industry. Despite sporadic and sometimes misleading press coverage, attempts to import OS methods from software production were few and far between. Furthermore, such initiatives seldom, if ever, proceeded beyond debating and posting OS licenses on the Web so that volunteers, to the extent that they joined at all, produced very little drug research [6]. One partial exception was Todd's schistosomiasis project, whose members worked to develop new production methods for the anti-schistosomiasis drug praziquantel. While Todd and his students did most of the work, approximately 30% of the project's online postings since 2006 seem to have been posted by outside volunteers [7].

In theory, at least, these hesitant starts are obsolete. In 1998, India's government committed to spend an astonishing $35 million to organize an open source drug development (OSDD) collaboration to find drugs for neglected diseases. According to organizers, OSDD will eventually conduct R&D at all stages along the drug-discovery pipeline from target discovery to lead optimization. To date, its accomplishments include mobilizing hundreds of students to annotate the tuberculosis genome and virtually screening 20,000 molecules to find 140 candidate compounds against tuberculosis and malaria [7]. That said, the scientific value of this work has been bitterly criticized and it is hard for outsiders to assess matters for themselves [104]. In the long run, OSDD's reputation will stand or fall with its success in finding and developing drugs.

## Positive reasons to use OS: shared commercial research

### ■ Complementary knowledge

Strikingly, recent OS collaborations frequently rely on conventional monetary incentives. Indeed, companies such as IBM often ask employees to 'volunteer' as part of their regular duties. How is this possible? The puzzle can be explained by noticing that IP is not the only and in many cases not even the best way to achieve appropriability. Instead, survey data routinely report that businesses in most industries rely on non-IP strategies such as being the first to market [8]. In these cases, the fact that information is produced under open methods hardly matters.

### ■ Sharing information

We have seen that big pharmaceutical companies are very unlikely to develop drugs for the so-called neglected diseases that afflict the developing world. Here, the good news is that big pharmaceutical companies also have very little to lose by helping others serve these communities. This is particularly true since some resources – notably data, research insights and access to existing chemical compound libraries – cost almost nothing to share. This explains why big pharmaceutical companies often donate these resources to neglected disease researchers.

The rub is that pharmaceutical companies could share even more data if they knew that the knowledge would not be diverted to competitors trying to develop drugs for rich nation diseases. Novartis has famously solved this problem by creating its own in-house neglected disease campus in Singapore. On the one hand, researchers have access to databases and human experts throughout the company. On the other, there is essentially zero chance that they will disclose the information to Novartis' rich nation competitors. The main drawback of this approach is that Novartis has a limited budget for neglected diseases research. One would ideally like to share data outside the company as well.

OS provides a natural way to do this. Suppose that an OS drug collaboration recruits one or more members from Novartis' research staff. Then these workers have dual loyalties and Novartis can count on them to respect its business need for confidentiality. This means that Novartis has no problem letting these workers sift through its worldwide internal databases. At the same time, small data disclosures seldom matter much. This suggests that when volunteers do find useful clues the company will let them divulge the information to the wider collaboration.

### ■ Sharing research costs

The modern world is awash in IP. This makes it hard to remember just how strange the idea actually is. We have already noted that information goods can be duplicated at essentially zero marginal cost. IP deliberately forfeits this advantage by giving owners the power to sell their discoveries at a high price. As with all monopoly pricing, this makes something that ought to be plentiful scarce. It also results in an inefficiency that economists call 'business stealing.' Suppose that a particular industry contains five firms. Then, an IP system requires each of them to produce its own software program from scratch. The good news is that the ensuing competition prevents any single firm from charging a monopoly price. The bad news is that society still has to pay for five duplicative R&D programs. One can, to be sure, imagine

circumstances under which these overlapping investments would be a good idea. For example, it may be impossible to predict how good a software program is in advance. In that case the existence of redundant programs would increase the odds that at least one good product emerges. Alternatively, each manufacturer could optimize its software for a different target audience. Nevertheless, the naive (and frequently correct) instinct is that business stealing is wasteful.

OS collaboration goes a long way toward solving this problem. On the one hand, it permits companies to share development costs. On the other, we have seen that companies often possess sufficient market power to recover at least some of the benefits that their research confers on consumers. For example, suppose that a company makes cell phones that run on software. Then the company knows that the value of its phones depends on, among other things, the quality of their operating system, and this is true whether or not the software is open. In these circumstances, the company will often find that investments in developing freely available OS software can be handsomely recouped in the form of increased cell phone sales.

Similar opportunities for shared research exist throughout drug discovery. For example, standard cell lines and DNA fragments ('standard biological parts') supply building blocks that can potentially be assembled into many different commercial products. In these circumstances, companies will often find it in their interest to develop the building blocks together and use them to develop different drugs. In many cases, each of the final drugs will serve distinct markets without competing. If anything, this scenario makes even more business sense than our hypothetical cell phone example.

### ■ Transparency & regulation

Transparency problems abound in the drug-discovery pipeline, where researchers almost always know far more about a particular compound's prospects than any outside observer. This poses a special problem for IP because researchers know that a positive test result can make them rich. This encourages them to overstate and even lie about results. OS incentives that reward researchers, regardless of whether or not compounds work, avoid this problem.

There are two places in which enhanced transparency would be particularly useful. The first involves the so-called 'valley of death' that occurs when biotechnologies are unable to obtain funding for promising early-stage compounds. The problem has to do with evidence. Investors know that they can trust insiders to reveal the good news about their compounds. But what about the bad? The result is that the good news must be very, very good before anyone will fund it (the problem is even more acute in neglected disease research, where funding is especially scarce). At least in principle, OS methods avoid this financial incentive to suppress bad results. This would go a long way toward suppressing valley of death problems.

The second place where OS could supply badly needed transparency involves bringing down the cost of human clinical trials. Over the years, IP incentives have produced repeated scandals in which commercial researchers have suppressed, falsified and even invented fictitious test results where none existed. This has understandably led regulators to insist on elaborate paperwork requirements that allow results to be audited. In principle, these costs could be much lower in an OS system. In this scheme, drug companies would supply test compounds so that doctors could treat patients and report results. Unlike the current system, however, doctors would receive no payment. Instead their reward would come from making and publishing discoveries. Indeed, Demonaco *et al.* point out that most of what we know regarding 'off-label' drugs already comes from papers published by unpaid physicians [9]. Still greater reliance on these non-financial incentives would dilute financial incentives to falsify results even further. The main problem with this scheme is that drug companies have no reason to support it unless it leads to lower costs. This, in turn, depends on regulators' willingness to relax existing paperwork requirements when OS methods are used.

Human trials account for three-quarters of all drug development costs. For this reason, applying OS methods to this phase offers the biggest potential payoff for any project discussed in this article. The problem, of course, is that the savings depend on regulators' ability to understand and appreciate that OS methods are inherently trustworthy. For now, this sounds like a political longshot.

## Entrepreneurship: making OS drug discovery happen

The analysis presented so far has focused almost entirely on finding specific instances along the drug-discovery pipeline where OS ought to thrive. Whether this actually happens will

depend on the dedication and shrewdness of entrepreneurs. Success will require as much art as science. Even so, it is possible to offer some reasonably specific guidance.

## ■ Creativity

For the most part, earlier sections of this article have focused on drug discovery as it exists today. But science is carried out by institutions. More than this, existing institutions tend to limit the scientific questions that we can ask and even think about. Modern academic research, for example, has been constructed around individual professors' laboratories. This favors questions that can be answered by small, well-funded teams. Successful OS collaborations, on the other hand, will most probably feature large, poorly-funded teams. OS software collaborations take advantage of this difference by focusing on problems – for example, bug hunting – that benefit from 'many eyeballs', that is, very large quantities of human judgment. The challenge for OS drug discovery is to find similar problems in biology. While some precedents exist, it is important for organizers to remember that most of today's conventional research and questions evolved in environments where 'eyeballs' were expensive and rare. For this reason, it is reasonable to think that the most productive OS collaborations will ask questions and use methods that are at least slightly different from what has come before. This puts a premium on scientific imagination, for example, the ability of organizers to ask new questions or develop new methods for existing ones.

Relatedly, OS organizers will often possess limited resources. Particularly in academic settings, organizers may have to use equipment and reagents that have already been purchased for other purposes. However, finding these resources will be difficult. Indeed, empirical studies demonstrate that universities typically know very little regarding what science is being done on campus [105]. Ironically, graduate and even undergraduate students may have a much better idea of what equipment and capabilities exist. In many cases, the answer can (and should) shape the science questions that OS collaborations set out to answer.

## ■ Listening

Traditional grant systems teach applicants to stress their own capabilities and ideas. This accentuates the natural human tendency to value our own judgments over competing viewpoints. OS collaborations are very different since a leader who fails to attract volunteers is (by definition) a failure. In the past, large scientific databases have often failed because leaders tried to force their own preferred nomenclatures and design choices onto contributors. The result, predictably, has been to drive away volunteers whose work could have supported the project. Conversely, successful computer science initiatives have frequently depended on leaders' willingness to accept solutions that they personally disagreed with [10]. Strangely, this willingness to listen is often associated with capitalism. Firms think constantly about what their customers and business partners would like, and might want, and how to please them. This backhanded altruism has almost nothing to do with affection. As in OS, however, success in the market means pleasing others.

Of course, OS leaders can also exercise influence. The reason has to do with what economists call 'coordination problems.' Collaboration members want to be useful, but this can only happen if everyone adopts the same plan. It follows that each individual will only adopt plans that have some chance of being implemented. Remarkably, this means that a leader's perceived power can be self-fulfilling. Indeed, even members who disagree with the leader's decision may go along on the theory that everyone else will. The result, as Tim Berners-Lee has jokingly remarked, is a kind of 'philosopher-king' system in which leaders' comments – although formally hortatory – frequently determine outcomes [10].

But when should leaders use this power? The obvious answer – when they have the best ideas – is not very helpful. A more precise insight starts with the proposition that studying issues and reaching decisions is costly. This suggests that followers can save time and effort by deferring to 'trusted intermediaries' who have already studied the problem. OS drug-discovery collaborations will continually encounter issues that the average member has no time to decide. These can include what experiments to do next, whether particular volunteers are producing good work, and even whether individual volunteers are trying to hijack or derail collaboration work for private ends. In all of these cases, followers will routinely turn to leaders for guidance not because they are smarter but because they have had more time to study the issues. At the same time, leaders must be careful not to outrun their influence. If followers reject their advice once,

**Key Terms**

**General Public License:** Early and extremely common open source license. It places strong restrictions on users' rights to reuse software, especially for commercial purposes.

**Embedded software:** Software used to operate electronic hardware ranging from cell phones to jet aircraft. Many embedded software developers license under open source licenses that suspend the duty to share their products for limited periods of time.

**Trade secrets:** Alternative, nonintellectual property strategy for protecting commercially valuable information based on enforcing actual or implied secrecy agreements. An important legal strategy in the biotechnology and pharmaceutical industries.

they are more likely to do so a second time. At this point leadership – and perhaps also the consortium's ability to hang together – will start to unravel.

## ■ Licenses

This article has stressed that rational collaboration design should start from a detailed substantive understanding of why volunteers might benefit from joining and supporting OS collaborations. The task of licenses and other paperwork is to reassure would-be volunteers that they will actually receive these benefits.

In the software world, licenses are mainly needed to stabilize the collaboration against free ridership. Absent licenses, each member would receive every other member's contributions at zero cost. In principle, these benefits are often more than enough repayment for joining the collaboration. In practice, however, members know that they can do even better by using IP to sell their contributions to other users. But if everyone does this the sharing unravels. Viral licenses block this outcome by requiring each member to contribute closely related improvements or extensions back to the collaboration. The question remains, however, just how broad this protection needs to be. In many if not most cases, selling small incremental improvements is impractical in any case. Here, viral licenses are superfluous. More generally, the breadth of viral licenses depends on the circumstances. For example, the well-known **General Public License** is almost certainly broader than stability requires [106]. In practice, most commercial OS collaborations use much narrower, Mozilla-type licenses and a few collaborations barely have any license at all [107]. Given that the whole point of OS is to provide knowledge to users, designers should avoid IP restrictions as much as possible [11].

OS drug-discovery collaborations may sometimes need licenses for a second reason. We have seen that commercial firms need significant appropriability before they invest in OS. This, however, may be inconsistent with the traditional OS software practice of requiring each collaboration member to share all improvements and extensions immediately. In this situation it may make more sense to let members keep their results secret for a short length of time and/or restrict shared results to collaboration members. This strategy is likely to be especially useful where sharing offers strong commercial benefits, most notably in industries where companies develop novel therapeutic organisms from shared building blocks. This would include, for instance, industries whose members try to make novel organisms from a common toolbox of DNA snippets ('standard biological parts') or a common cell line ('H1 stem cells'). In either case, allowing members to keep new results confidential for a limited period of time – for example, until a development project is abandoned, generates patents, or produces products that are sold to consumers – would offer significant appropriability. Once this initial period expired, however, members would still be required to disclose all of the project's test results to other collaboration members. Members would also be allowed to disclose their results earlier and/or share them with the public at large.

Allowing members to suppress results for limited periods would mark a clear departure from most software OS. That said, similar practices already exist in traditional OS communities, most notably in the collaborations that produce the '**embedded software**' that operates most electronic devices. Because of a loophole in General Public License, programmers in that industry typically have the legal right to suppress code for 18 months or so. Despite this, they choose to share approximately half of all code sooner than that. This is typically done to obtain error reports or extensions from other users. The embedded software example is important because it demonstrates that the OS model can be modified so that contributors receive more appropriability than they would under immediate sharing. This appropriability can, in turn, be used to fund investments that increase the volume and quality of software that is eventually shared. Similar strategies would be especially helpful in drug discovery, where the need to recover high R&D costs often rules out immediate sharing. At the same time, our embedded software example demonstrates that temporary embargos are less damaging than they appear since we can reliably expect many OS members to waive their rights early.

Remarkably, some drug-discovery initiatives are already experimenting with this kind of limited appropriability model. The proposed Arch-2POCM collaboration plans to spend upwards of $160 million to take novel drug candidates from discovery through Phase IIb clinical trials. None of the resulting data would be IP protected. Current negotiations envisage three to

five 'flagship' companies each paying $1.6 million per year to help fund the project. In return, flagship partners would receive any data in real time before it is made public [108,109].

### ■ IP rights

This article has argued that OS drug discovery primarily relates to incentives and transactions; from this viewpoint, legal issues are decidedly secondary. That said, we have seen that licenses are sometimes necessary. However, these are only enforceable if they are based on some existing IP right. In the software case, this right was almost always derived from copyright, which offers strong protection and is cheap to acquire. But drug discoveries cannot be copyrighted. This has led most commentators to consider alternative schemes based on patent protection. Yet this too presents a problem, as obtaining patents is expensive and likely to be unaffordable absent very generous grants.

Fortunately, patents are not the only option. Indeed, many conventional businesses protect and share data using a completely different principle – **trade secrets**. Legally, trade secrets have at least two drawbacks compared with copyright. First, they offer no protection against independent discovery: third parties who discover the secret can obtain patents and demand that the original discoverers pay royalties. Still, this objection cannot be fatal. After all, conventional businesses routinely manage trade secrets through a shrewd combination of secrecy agreements, defensive publishing and patenting. OS collaborations can do the same.

### ■ Limiting the community

The second drawback of trade secrets compared with copyright is that protected information cannot be shared too openly. Instead, recipients must explicitly agree to respect the secret before receiving it. This is unlikely to be a problem for companies, university laboratories, and other institutions that routinely sign and observe confidentiality agreements. However, individual students, and particularly undergraduates, are another matter. In some cases, at least, ideologically motivated volunteers could decide to break their agreements and disclose the trade secret. This would destroy confidentiality for all purposes.

The obvious answer is to limit membership to 'card carrying' faculty and institutional researchers. As a practical matter, this will probably do little to limit the talent pool available for OS. At the same time, open membership remains one of software OS's most attractive features. Collaborations may sometimes accept modest disclosure risks rather than give this up.

Finally, limited membership could also raise antitrust issues. So long as the collaboration remains open to any company willing to share data, these issues are almost certainly manageable.

## Future perspective

OS drug discovery is not a panacea and no one should expect it to displace conventional IP-driven research at each and every point along the drug-discovery pipeline. At the same time, this article has argued that OS methods often make good business sense. The only question is how quickly OS entrepreneurs can find and populate suitable locations along the pipeline. Early projects will probably focus on neglected disease research, where weak OS incentives do not have to compete with lucrative IP royalties. Given aggressive leadership, OS collaborations can easily be organized within a year or two to find early-stage drug targets and compounds, mine data that is currently locked inside large pharmaceutical houses, and bring badly needed transparency to projects that would otherwise languish in the valley of death. Successful neglected disease collaborations could, in turn, provide demonstration models for later and more ambitious collaborations. If this happens, we can reasonably expect the first commercial OS collaborations by the end of this decade. These will almost certainly be focused on fields where drug discovery is based on the use of shared building blocks such as standard organisms (e.g., stem cell therapies) or DNA fragments ('standard biological parts'). In the very long term OS methods could also offer an important alternative to the paperwork-intensive procedures that make late-stage drug discovery (and drugs generally) so expensive. Realizing these benefits will require understanding and cooperative regulators.

**Financial & competing interests disclosure**

*The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.*

*No writing assistance was utilized in the production of this manuscript.*

## Executive summary

***The need for rational design***

- Open source (OS) has become a powerful method for organizing software production. Scholars have long argued that similar methods can be used to accelerate drug discovery.
- Designers of successful OS drug-discovery initiatives must combine a careful understanding of when and how OS collaborations are possible with a sharp eye for potential opportunities along the drug-discovery pipeline.

***Negative reasons to use OS: opting out of intellectual property***

- Economists have documented various scenarios in which intellectual property incentives yield ineffective or socially perverse results. OS methods are often desirable in these circumstances.
- OS collaborations are promising vehicles for discovering and validating drug targets, identifying lead compounds and sharing standard research inputs, such as cell lines. The case for OS methods is particularly strong in neglected disease research.

***Positive reasons to use OS: altruism, reputation, own use & education***

- OS software volunteers frequently respond to non-monetary incentives based on altruism, reputation, education and own use. OS drug-discovery collaborations should be designed to make these incentives maximally attractive to potential volunteers.
- Attempts to launch working OS drug-discovery collaborations, including a US$35 million initiative funded by the Indian Government, are still in their infancy.

***Positive reasons to use OS: shared commercial research***

- Commercial software firms often use OS methods to pool R&D efforts. Similar methods make good business sense for many commercial drug-discovery tasks, especially in synthetic biology and stem cell research.
- Commercial researchers who find promising drug candidates often have difficulty convincing others. This leads to under-investment (the 'valley of death' problem) and high regulatory burdens. The inherent transparency of OS methods offers a potentially powerful tool for mitigating these problems.

***Entrepreneurship: making OS drug discovery happen***

- OS methods are best suited to 'many eyeballs' problems that bring massive amounts of human judgment to bear on a specific problem. Creative OS leaders should select and, if necessary, invent biology problems that play to this strength.
- Drug discovery will require new types of OS licenses. These will likely to be based on trade secret law and may permit members to delay information sharing for reasonable periods of time.

## References


1 Kremer M, Glennerster R. *Strong Medicine: Creating Incentives for Pharmaceutical Research on Neglected Diseases.* Princeton University Press, NJ, USA (2004).

2 Scotchmer S. Openness, open source, and the veil of ignorance. *Am. Econ. Rev.* 100, 165–171 (2010).

3 Henkel J, Maurer S. Network effects in biology. *Am. Econ. Rev.* 100, 159–164 (2010).

4 Campbell E, Clarridge B, Gokhale M *et al.* Data withholding in academic genetics: evidence from a national survey. *J. Am. Med. Assoc.* 287, 473 (2002).

5 Allarakhia M. Open source biopharmaceutical innovation – a mode of entry for firms in emerging markets. *J. Bus. Chem.* 6(1), 11–31 (2009).

6 Maurer S. Open source drug discovery: finding a niche (or maybe several). *UMKC Law Rev.* 76, 405 (2007).

7 Årdal C, Røttingen J-A. Open source drug discovery in practice: a case study. *PLoS Negl. Trop. Dis.* 6(9), e1827 (2012).

8 Graham SJH, Merges RP, Samuelson P, Sichelman T. High technology entrepreneurs and the patent system: results of the 2008 Berkeley patent survey. *Berkeley Technol. Law J.* 24, 255–327 (2009).

9 Demonaco HJ, Ali A, von Hippel E. The major role of clinicians in the discovery of off-label drug therapies. *Pharmacotherapy* 26(3), 323–332 (2006).

10 Berners-Lee T, Fischetti M. *Weaving the Web: The Design and Ultimate Destiny of the Worldwide Web by Its Inventor*. Harper-Collins, NY, USA (1999).

11 Maurer S. 'The Penguin and the Cartel,' *Utah Law Review* 2012(1), 269–318 (2012).

### ■ Websites

101 Wellcome Trust. 'SNP Consortium and International HapMap Project.' www.wellcome.ac.uk/Funding/Biomedical-science/Funded-projects/Major-initiatives/SNP-Consortium-and-International-HapMap/wtd003500.htm

102 Structural Genetics Consortium. 'What is The SGC?' www.thesgc.org/about/what_is_the_sgc

103 Callaway E. No Rest for the Bio-Wikis. Nature News. www.nature.com/news/2010/101115/full/468359a.html

104 Jayaraman KS. India's tuberculosis genome project under fire: sequence annotated by students should be peer reviewed, say scientists. Nature News. www.nature.com/news/2010/100609/full/news.285.html

105 Thursby J, Thursby M. University licensing under Bayh-Dole: what are the issues and evidence? www.gtrc.gatech.edu/otl.html

106 Free Software Foundation, 'General Public License'. https://gnu.org/licenses/gpl.html

107 The Mozilla Public License. www.mozilla.org/MPL

108 Arch2POCM Collaboration. 'Funding and Governance'. http://sagebase.org/WP/arch/faqs/148–142/

109 Arch2POCM Collaboration. 'Patent and Datasharing Policy'. http://sagebase.org/WP/arch/faqs/patent-and-information-policy