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

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

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

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

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

The big picture: error detection in databases

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

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

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

Antony J. Williams graduated with a Ph.D. in chemistry as an NMR spectroscopist. Dr Williams is currently VP, Strategic development for ChemSpider at the Royal Society of Chemistry. Dr Williams has written chapters for many books and authored or ≥120 peer reviewed papers and book chapters on NMR, predictive ADME methods, internet-based tools, crowdsourcing and database curation. He is an active blogger and participant in the internet chemistry network.

Sean Ekins graduated from the University of Aberdeen; receiving his M.Sc., Ph.D. and D.Sc. He is Principal Consultant for Collaborations in Chemistry and Collaborations Director at Collaborative Drug Discovery Inc. He has written more than 180 papers and book chapters on topics, including drug–drug interaction screening, computational ADME/Tox, collaborative computational technologies and neglected disease research. He has edited or co-edited 4 books.

Valery Tkachenko graduated from the Lomonosov Moscow State University, receiving his M.Sc. in Chemistry and B.Sc. in Computer Sciences. He is currently Chief Technology Officer of ChemSpider at the Royal Society of Chemistry. Over the course of the past 15 years he has participated in the development of several successful enterprise projects for large pharmaceutical companies and the public domain, including PubChem. He is the author of more than 20 peer reviewed papers and book chapters.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2012.02.013

claims [6]. Even manual curation of biological activity databases, such as the Accelrys (http://accelrys.com/) drug data report, MDL Drug Data Report (MDDR), has been proposed to have errors which have been compensated for by calculating activity–activity similarities [7]. It has also been suggested that automatic classification of molecules based on Simplified Molecular Input Line Entry Specification (SMILES) strings might be useful for error detection and aiding biochemical pathway database construction [8].

Error detection is important in clinical practice to avoid mortality owing to missed injury [9]. Errors in clinical research databases of thousands of patients have been shown in one study to vary from 2.3% to 26.9% [10] resulting in data analysis errors that have the potential to impact the standards of care for many thousands of patients. Medical records often contain multiple identifiers and are error prone when it comes to linkage to samples. Therefore, methods have been developed for identifier error detection (these types of approaches might be applicable to large chemical databases too) [11]. One group has also suggested that competition and conflicts of interest distort medical findings [12] and the same group has also suggested the need to improve validation practices in 'omics' data [13] and require that external validation of molecular classifiers is performed [14].

A recent multicenter analysis of a common biological sample by mass spectrometry-based proteomics identified generic problems in databases as the major hurdle for characterizing proteins in the test sample correctly. Primarily, search engines could not distinguish among different identifiers and many of these algorithms calculated molecular weight incorrectly [15]. Computational genomic annotation errors such as those in the *Brucella abortus* genome (which had seven annotation errors) have been corrected by proteomics-based confirmation of protein expression [16]. Methods have been developed for labeling error detection to improve analysis of microarray data and to discover the correct relationship between diseases and genes [17] (which suggests that many of the microarray databases might have mislabeled data). Simple rule-based methods for validating ontology-based instance data can detect curation errors during the development of biological pathways [18]. It has been suggested that functional annotation from well-studied proteins to new sequences hit a plateau as annotation transfer led to error propagation across databases, which resulted in follow up experiments that failed. Many proteins have completely wrong function assignments and one database had between 2.1% and 13.6% of annotated Pfam hits unjustified [19]. The authors also indicated difficulty in assessing incorrect annotations in public sequence databases because many of the sequences have not been studied experimentally. While most scientists think a ligand–protein X-ray structure is definitive others have highlighted how these can also have errors with far reaching consequences [20].

A recent review of data governance in predictive toxicology analyzed several public databases. The authors mentioned a lack of systematic and standard measures of data quality but did not address error rates or address molecule structure quality [21]. In our combined literature analysis independent groups have identified significant errors across all types of databases which if not checked and corrected will be an impediment to future science. We do not believe such studies identifying errors in databases across drug discovery and development have received the attention they deserve.

The quality of chemistry databases

Scientific knowledge is fragile, it demands incorruptible storage media and in today's electronic age the sheer amount of data underpinning this knowledge requires careful curation and verification, chemistry databases are no exception. There are now many such databases that are freely available on the internet [e.g. PubChem (PubChem Database: http://pubchem.ncbi.nlm.nih.gov/), ChemSpider (ChemSpider: http://www.chemspider.com), DrugBank (DrugBank: http://www.drugbank.ca) [22], among others] and we rely on them to be correct, often granting them 'trust' and declaring them as high quality without validation of these beliefs. We have also previously discussed the importance of chemical data curation [23].

For many types of scientific data minimal data standards have been created which provide confidence in data deposited in databases [24]. Unfortunately, for chemistry databases there are as yet no agreed upon standards and there is no freely available gold standard structure database which we can yet rely on. Despite the decades of experience that underpin the assembly of commercial molecule databases (e.g. Scifinder, MDDR, among others) primarily depending on skilled staff for curation and data checking, the delivery of online databases commonly appears to focus more on the development of the underlying cheminformatics architecture and platform rather than the delivery of a high quality resource of data. Also there are as yet no definitive guides for how to assemble and integrate disparate data sources and each of the individual groups appear to follow custom unproven and undocumented approaches in assembling data. Some of these databases are simply repositories whereby the data deposited to the system remain unedited even when the database hosts are well aware of errors in the data. In our chemistry domain there is a dire need for the suppliers of chemistry-based community resources to work together and develop best practices to reduce the amount of repetition and lessen the impact of poor data assembly. A highly curated online database of validated chemical name–structure relationships would probably provide for the underpinnings of a semantic web for chemistry, for chemistry text-mining, for integration to online resources and to enable the efficient disambiguation of chemical names. Unfortunately we currently do not have such a reliable resource which we would term a 'gold standard'.

A recently published paper describes an effort to assemble clinically approved drugs from the USA, EU, UK, Canada and Japan and the creation of a new database of molecular structures [25]. This group described it as a 'comprehensive and curated resource', they published the data in a cheminformatics browser and declared that it will be used along with the National Institutes of Health (NIH) Chemical Genomics Center (NCGC) screening resources as a component of the NIH therapeutics for rare and neglected diseases (TRND) program (TRND Program: http://www.nih.gov/news/health/may2009/nhgri-20.htm). The paper described in detail the considerable effort that went into dealing with semantic errors and sourcing 'correct structures'. The NPC browser was released to the public on 27th April 2011 (NPC browser press release: http://www.ncgc.nih.gov/docs/PressRelease-4_27_11.pdf). In keeping

with our ongoing efforts to aggregate and curate data for our own work in the fields of quantitative structure–activity relationship and drug repositioning we were interested in examining the quality of the data in the NIH Chemical Genomics Center Pharmaceutical Collection browser (NPC browser), and we have used it as an example in this article and elsewhere [23] for discussing the challenges of assembling high quality data and, specifically, as an example of how more foresight, consideration and care is required when releasing more chemistry related data into the public domain.

Trust and chemistry databases

While the quality of academic or commercial databases is rarely questioned the media have made much of the implicit trust granted to the online encyclopedia, Wikipedia, questioning whether a crowdsourced database can be as high-quality and as trusted as a highly curated and expert assembled resource, such as the Encyclopedia Britannica. Wikipedia has a great diversity of coverage, and offers unsurpassed immediacy although the question as to which is more 'trustworthy' is still an issue. Wikipedia chemical compound pages which have been developed by crowdsourcing are of high quality because there is debate between the editors (e.g. discussions regarding the structure of Tacrolimus on Wikipedia: http://en.wikipedia.org/wiki/Talk:Tacrolimus#IUPAC_Name_and_structure). Efforts have included dedicated time to validate chemical structures (ChemConnector, dedicating Christmas time to curating Wikipedia: http://www.chemconnector.com/2008/01/09/dedicating-christmas-time-to-the-cause-of-curating-wikipedia/), collaborations with CAS to ensure the validation of CAS Registry Numbers (CAS and Wikipedia: http://www.cas.org/newsevents/caswikipedia.html) and a dedicated Wikipedia project to validate the data in the ChemBoxes (ChemBox Validation: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Chemicals/Chembox_validation). A quantitative data validation exercise to review the structures for over 150 drug molecules has been performed comparing the data with several online databases, including ChemSpider, PubChem, DrugBank, among others, and shows that relative to the test set in question Wikipedia is the highest quality data available in online publicly accessible databases with one caveat [A.J. Williams, *et al.*, unpublished]. The data are 'NOT' available as a downloadable structure set that can be consumed by structure databases and the validation was performed by comparing the depicted chemical structures on Wikipedia with those available as electronic structure files.

Data quality can be compared with other resources, such as the Merck Index [5] and the US Pharmacopeia (US Pharmacopeia on Wikipedia: http://en.wikipedia.org/wiki/United_States_Pharmacopeia) which themselves are less frequently updated than Wikipedia. Errors detected in any of the published (book-bound) resources are not updated for many months after detection of errors and even then purchasers of a volume will not receive those corrections. Meanwhile, the obvious advantages of immediacy of editing and availability of an online data source, is clear. The devil's advocate and critic of the wiki approach would of course note that such edits could also degrade the quality of what is on Wikipedia. It is our anecdotal observation that Wikipedia information on chemical structures is considered trustworthy. We think that many of the standard reference resources will ultimately be replaced by Wikipedia or similar crowdsourced sites. The shift to mobile technologies and the expectation of improved searchability over thumbing physical paper pages is going to help to drive this shift.

It seems paradoxical if people question the validity of data and information captured on Wikipedia, while many are willing to grant trust to public databases and in particular, chemical compound databases at a surprisingly high level. One of the authors (Antony J. Williams) prepared an online survey (Fig. 1) requesting community feedback regarding the trust granted to online databases containing chemistry related information. The trust granted to online resources captured in the survey, considering the actual quality as discussed below, is consistent with our own experience of interacting with many scientists. We suggest that no database should always be trusted but that there is a different level of data-driven trustworthiness that can be granted to online chemistry databases, albeit at a specific point in time, as content continues to change over time and quality can certainly improve or degrade. We have recently introduced a wiki environment listing online scientific databases and contributions and commentary from the community might aid us in providing a ranking of quality (*vide infra*).

The initial data shown in Fig. 1 were gathered over four days and from 46 survey responders in 2010 before the release of the NPC browser (Views of the NPC browser data: http://www.chemconnector.com/2010/12/11/community-views-and-trust-in-public-domain-chemistry-resources/). It should be noted that the term 'trust' was purposely not explicitly defined and was left up to the interpretation of the surveyed population. The table (http://www.chemconnector.com/wp-content/uploads/2010/12/trust-2.png) represents several key points at the time of reporting:

(i) All responders were familiar with Wikipedia and the majority commonly trusted the resource.
(ii) Only one database had the majority of users 'always' trusting the resource, PubChem. It should be noted however that the distribution of trust appeared to be largest for this database.
(iii) ChemSpider is the database that is trusted by the largest population of users (see caveats below).

Approximately 50–70% of the responders had no experience with DrugBank (DrugBank: http://www.drugbank.ca/), ChemIDPlus (ChemIDPlus: http://chem.sis.nlm.nih.gov/chemidplus/), Psychoactive Drug Screening Program (PDSP; PDSP Ki Database: http://pdsp.med.unc.edu/kidb.php) and DailyMed (DailyMed: http://dailymed.nlm.nih.gov) resources. Some notable caveats regarding the data in the table are as follows. Because the author of the questionnaire is the host of the ChemSpider database a bias as to the trustworthiness of the ChemSpider resource is to be expected because the link to the original survey was posted on his blog and many blog posts regarding data quality in public domain databases and the efforts to curate ChemSpider had been posted there historically. The reverse is true in terms of the 'Always Trust' bias for PubChem. Several reports regarding quality issues about PubChem have also been made on blogs (Williams, A.J. All that glisters is not gold: Quality of Public Domain Chemistry Databases: http://blogs.scientificamerican.com/guest-blog/2011/08/02/all-that-glisters-is-not-gold-quality-of-public-domain-chemistry-databases/; Williams, A.J. The Messy World of Even Curated Chemistry on the Internet: http://www.chemconnector.

	Never trust	Rarely trust	Sometimes trust (Neutral)	Commonly trust	Always trust	No experience with this resource
DrugBank	1.6% (1)	7.8% (5)	6.3% (4)	20.3% (13)	9.4% (6)	**54.7% (35)**
DailyMed	3.2% (2)	4.8% (3)	4.8% (3)	7.9% (5)	7.9% (5)	**71.4% (45)**
Wikipedia	3.0% (2)	7.6% (5)	33.3% (22)	**48.5% (32)**	7.6% (5)	0.0% (0)
PubChem	3.1% (2)	13.8% (9)	18.5% (12)	26.2% (17)	**29.2% (19)**	9.2% (6)
ChemIDPlus	1.6% (1)	1.6% (1)	11.5% (7)	9.8% (6)	16.4% (10)	**59.0% (36)**
ChEBI/ChEMBL	1.6% (1)	1.6% (1)	6.3% (4)	34.9% (22)	14.3% (9)	**41.3% (26)**
CAS' Common chemistry	1.6% (1)	0.0% (0)	8.1% (5)	29.0% (18)	24.2% (15)	**37.1% (23)**
ChemSpider	1.5% (1)	1.5% (1)	7.5% (5)	**44.8% (30)**	31.3% (21)	13.4% (9)
PDSP	4.7% (3)	1.6% (1)	6.3% (4)	6.3% (4)	9.4% (6)	**71.9% (46)**

Drug Discovery Today

FIGURE 1

An online survey (performed by Antony J. Williams) requesting community feedback regarding trust in online chemistry databases.

com/2010/08/15/the-messy-world-of-even-curated-chemistry-on-the-internet/). It is assumed that the majority of responders were unaware of these commentaries in addition to others reported in mainstream journals [26]. The results of the questionnaire are probably representative of both the responders' beliefs and experience of the listed resources. In the case of this questionnaire, trust is probably rather an invested emotional response and for each individual might comprise, for example, belief in a resource (e.g. based on marketing, word-of-mouth or peer pressure) and quantifiable data-driven experience of a resource (the user might have downloaded the data and checked data quality and content thoroughly). These are really the two extremes and clearly there are various other biases inherent in the granting of trust. It should be noted that the survey remains online after it was originally posted and the bias of the questionnaire has changed as the number of responders has doubled. The data available as of October 2011 are presented in Fig. 2.

What is in a name: structure–identifier relationships in chemical databases

The relationship between chemical compounds and all possible identifiers leads to significant confusion in chemical databases. Some simple examples commonly observed in public domain databases are listed below:

(i) A chemical name can refer to a structure of a particular isomeric form but the stereochemistry might be confused. As an example, Taxol has a specific stereochemistry but in Pubchem a search returns five structures with different stereochemistry (Structures of Taxol on PubChem: http://www.ncbi.nlm.nih.gov/pccompound?term=Taxol%5Bcompletesynonym%5D).

(ii) A CAS registry number associated with a particular salt form can be incorrectly associated with the neutral compound. This experience comes from processing many tens of data sets supplied by chemical vendors for deposition to the ChemSpider database as vendors commonly enter only the neutral compound into their data sets and associate the chemical name and CAS number for the salt.

(iii) Systematic names containing all explicit stereochemical detail might be associated with a molecular skeleton with all stereochemistry absent. As an example, cholesterol has specific stereochemistry whereas one of the multiple forms of cholesterol on PubChem has no defined stereochemistry (cholesterol without stereochemistry: http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=304&loc=ec_rcs).

One example of the extreme nature and diversity of name–structure relationship errors in public domain databases is for the simplest organic molecule, methane. A review of all names and identifiers associated with methane provides a long list of obvious errors as listed below (list of chemical names for methane: http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?q=nama&cid=297).

1. Consider searching each of these chemical databases by chemical name (systematic name, trade name or synonym). Please mark each online resource according to how much you generally trust the results. Create chart Download

	Never trust	Rarely trust	Sometimes trust (Neutral)	Commonly trust	Always trust	No experience with this resource	Response count
DrugBank	1.1% (1)	5.3% (5)	9.5% (9)	24.2% (23)	9.5% (9)	**50.5% (48)**	95
DailyMed	2.2% (2)	7.5% (7)	5.4% (5)	8.6% (8)	6.5% (6)	**69.9% (65)**	93
Wikipedia	3.1% (3)	8.2% (8)	34.0% (33)	**46.4% (45)**	7.2% (7)	1.0% (1)	97
PubChem	3.2% (3)	12.6% (12)	14.7% (14)	**33.7% (32)**	25.3% (24)	10.5% (10)	95
ChemIDPlus	1.1% (1)	2.2% (2)	10.0% (9)	14.4% (13)	15.6% (14)	**56.7% (51)**	90
ChEBI/ChEMBL	1.1% (1)	1.1% (1)	7.5% (7)	34.4% (32)	15.1% (14)	**40.9% (38)**	93
CAS' Common chemistry	1.1% (1)	0.0% (0)	5.5% (5)	28.6% (26)	25.3% (23)	**39.6% (36)**	91
ChemSpider	1.0% (1)	2.0% (2)	6.1% (6)	**48.0% (47)**	28.6% (28)	14.3% (14)	98
PDSP	3.2% (3)	1.1% (1)	5.3% (5)	9.5% (9)	8.4% (8)	**72.6% (69)**	95
				Please recommend other resources you use Show responses			23

Drug Discovery Today

FIGURE 2

Updated results for online survey (performed by Antony J. Williams) requesting community feedback regarding trust in online chemistry databases.

These include: Furnace black; Graphitic acid; Mineral carbon; Royal spectra; Silver graphite; Special schwarz; GRAPHITE, NATURAL; Activated charcoal, iodinated; Carbon nanotube, single-walled; Fullerene soot; Carbon Activated; Diamond; CHARCOAL; (2*R*,3*R*)-Butanediol dimesylate; 1,3-DICHLORO-PROPAN-2-ONE; (2*R*,3*R*)-Butanediol bis(methanesulfonate); Ethyl-1-propenyl ether, mixture of cis and trans; Carbon, activated [UN1362] [Spontaneously combustible]; PSS-[2-[(Chloromethyl)phenyl]ethyl]-Heptaisobutyl substituted.

These can be categorized below:

(i) Representations of Carbon: graphite, diamond, soot, fullerene, coal;
(ii) Organic compounds;
(iii) Trade names.

This example is rather extreme in its nature whereas others given below are more general in nature and represent issues detected in many other public databases. It should be noted that the situation of confusion for multiple forms of carbon associated with methane is because that in many cases data depositions might be provided in the form of a SMILES format. The SMILES string C, rather than the equivalent alternative [CH4], corresponds to methane (SMILES string conversion: http://www.daylight.com/daycgi/depict?43). However, carbon, graphite, diamond or fullerene represented as carbon, as 'C' would then be associated with methane at deposition through conversion of the SMILES string. These issues can only be resolved through post-deposition curation or by predeposition filtering of existing name–structure relationships. This does not, however, account for the misassociations of the many other organic molecules.

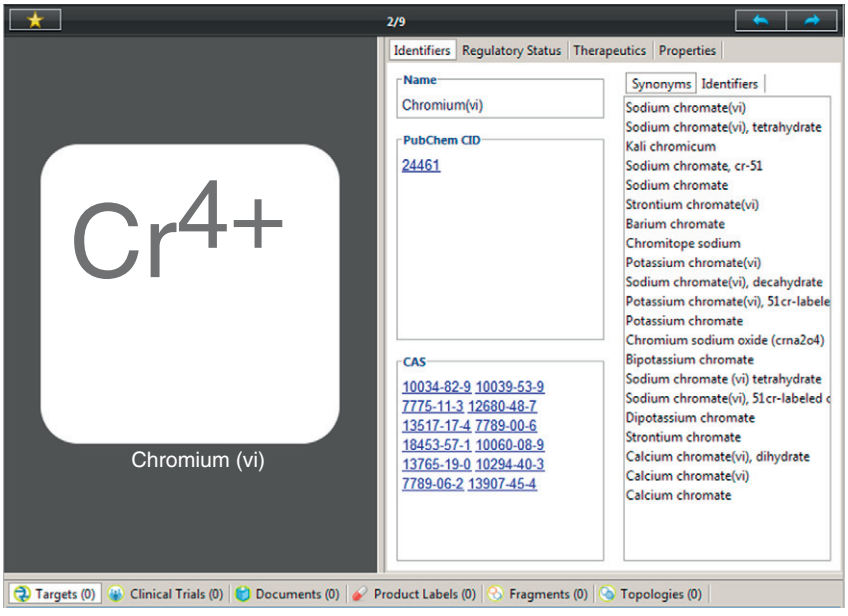
Data errors in the NPC browser: misassociations

Several data errors were detected with the 'originally' downloaded NPC browser [25]. Since the original release some of these errors have been resolved in the presently available dataset, some as a response to a series of public blog posts (Confusing Search Results in the NPC browser: http://www.chemconnector.com/2011/06/16/confusing-search-results-in-the-npc-browser/; Rabbits, Potatoes and other Vegetables in the NCGC Database: http://www.chemconnector.com/2011/07/19/rabbits-potatoes-and-other-vegetables-in-the-ncgc-database/; Duplicate compounds in the NPC browser and NCGC Dataset: http://www.chemconnector.com/2011/07/26/duplicate-compounds-in-the-npc-browser-and-ncgc-dataset/). For example,

some obvious misassociations are displayed in Figs 3–5. Figure 3 shows one of the nine chemicals returned following a search for 'chromium' in the NPC browser. The list of associated synonyms, do not coincide with the displayed chemical moiety of the bare chromium(IV) ion. The list is diverse and refers to several different chemicals and the species of interest is almost certainly meant to be a chromate ion based on the long list of identifiers, with the variation in the chemical only being the associated counterion. It should be noted that there is a long list of associated CAS numbers also, each probably associated with one of the many forms of chromate.

While this might also be deemed to be an extreme case, there are multiple other examples, including search results for manganese, titanium and chromium. It should be noted that the NPC browser is meant to represent the active drug moiety so there is an error in this respect. In the case of arsenic (Fig. 4) the species of interest is meant to be an arsenic oxide. The display of two equivalent trivalent arsenic cations is, in itself, clearly an error as there is no value in retaining two degenerate ions. Unfortunately the registration of chemical records containing degenerate compounds as doubles, triples, quads, among others, in online databases is not limited to the NPC browser, and these have been detected in other databases (discussed below).

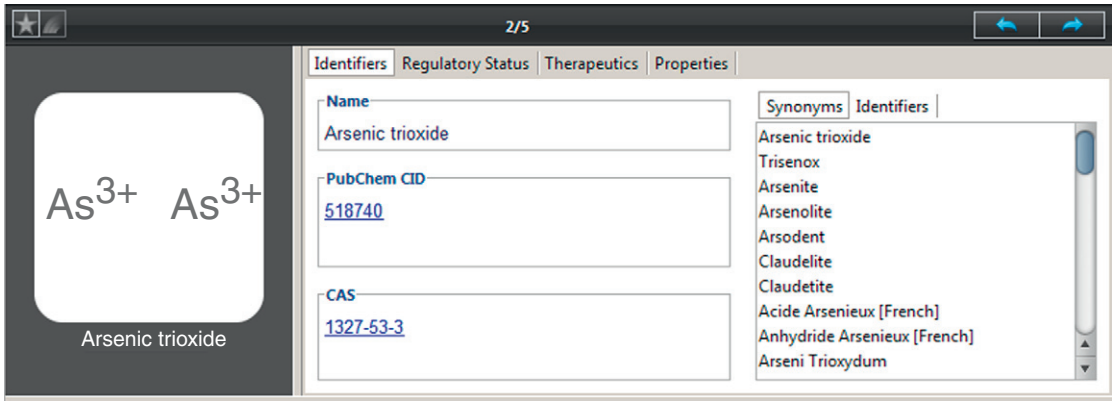
The misassociation of a chemical with one or more chemical names is not limited to inorganic ions. An example is shown in Fig. 5 for a well-known antibiotic 'neomycin'. Despite the fact that



Drug Discovery Today

FIGURE 3

Image from the original downloaded NPC browser showing the result of searching for chromium. Abbreviation: NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.



Drug Discovery Today

FIGURE 4

Image from the original downloaded NPC browser showing the result of searching for arsenic. Abbreviation: NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.

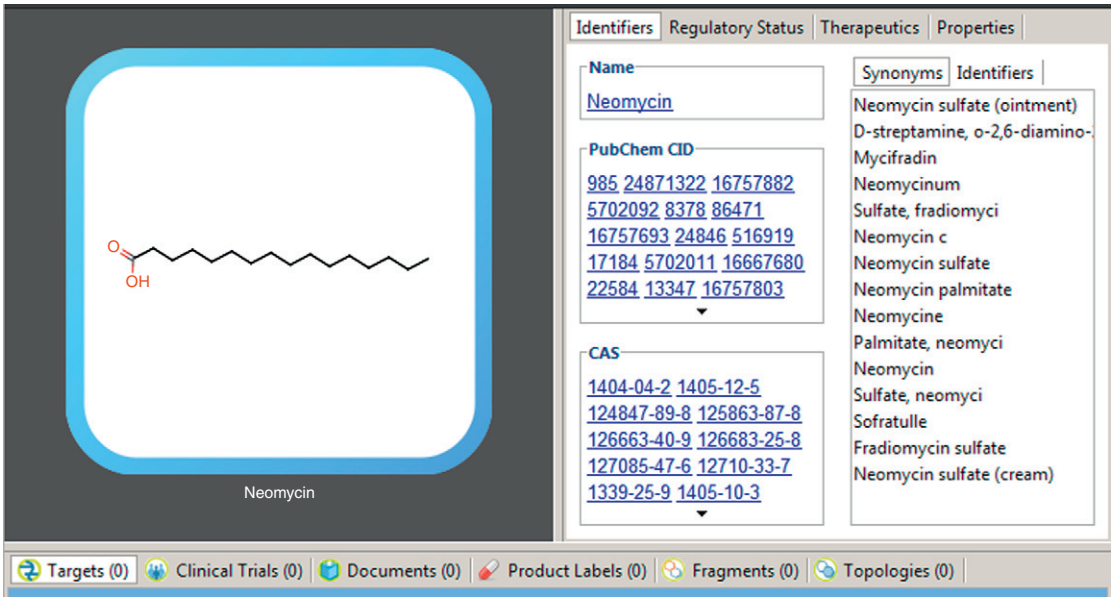


FIGURE 5

Image from the original downloaded NPC browser showing the result of searching for neomycin, an incorrect structure. Abbreviation: NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.

15 synonyms identify the displayed structure of hexadecanoic acid as an antibiotic, neomycin is in fact an aminoglycoside (the chemical structure of Neomycin: http://en.wikipedia.org/wiki/Neomycin) and this represents another clear example of misassociation. Clearly the effects of such egregious errors in databases, if the data is used elsewhere, can result in erroneous structure–property relationships when used for modeling, among others.

In our judgment the behavior of the search in the NPC browser itself is confusing as users would probably expect the chemical record associated with a search on a drug name to retrieve information about that particular drug only. For example, a search on neomycin would return a single record for that drug. However, the present text-based search is a search across all information in the entire database and therefore retrieves a total of 26 records; most of them are retrieved because some mention of the text string neomycin has been made in an associated document referring to that drug. That said, even that is not the complete explanation as there are three distinct drugs retrieved labeled as neomycin as shown in Fig. 6.

Such a 'full text' based search is similar to that available in PubChem. A naive user searching PubChem using a drug name as the input might expect to retrieve just the record associated with that drug. However, by default the search is a complete text search throughout the database and the user must be aware that only a constrained search of the form Name[CompleteSynonym] will retrieve the appropriate record(s). As an example, a search for the drug name aspirin retrieves 69 records whereas a search on Aspirin[CompleteSynonym] retrieves a single record in PubChem. It should also be noted however that such a [CompleteSynonym] search does not always retrieve a single record as retrieval depends on the quality of the data in the database. A search for the drug name Taxol in PubChem retrieves 59 records whereas a search on Taxol[CompleteSynonym] retrieves five records. This is confusing as it is not clear which is the actual structure. Closer examination shows that all are consistent in terms of connectivity but differ in stereochemistry and the majority of associated bioassay data are associated with the incorrect structure of Taxol. It should be noted that the correct structure is listed as one of the five.

Data errors in the NPC browser: analysis of steroids

To examine potential patterns in the quality of data contained within the original downloaded NPC browser 'HTS (high-throughput screening)' data set a series of three specific steroidal substructures were searched against the compounds contained within the data set (Table 1). These substructures were the gonane, gon-4-ene and gona-1,4-diene substructures as shown in Fig. 7. Each of these substructures were used as a separate search and the individual subsets of molecules examined. During the initial examination specific patterns emerged. For example, because the majority of steroids contain specific stereochemistry centers, the examination involved validating the stereospecific details of the structure against the structure expected based on the CAS number and chemical name combination as described earlier. As an example we will consider the structure of bufogenin. In the NPC browser the structure recovered is as shown in Fig. 8.

Figure 8 shows a single CAS number and a series of chemical names, all consistent with bufogenin. Using the series of databases listed earlier in Fig. 1 as validation sources the structure of bufogenin was identified by a cross-validation exercise. Comparing the determined structure with that in the NPC browser shows the differences illustrated in Fig. 9. Specifically, two stereocenters are missing and three stereocenters are inverted.

A similar analysis was performed for each of the retrieved chemicals in the three classes of steroids defined by the various

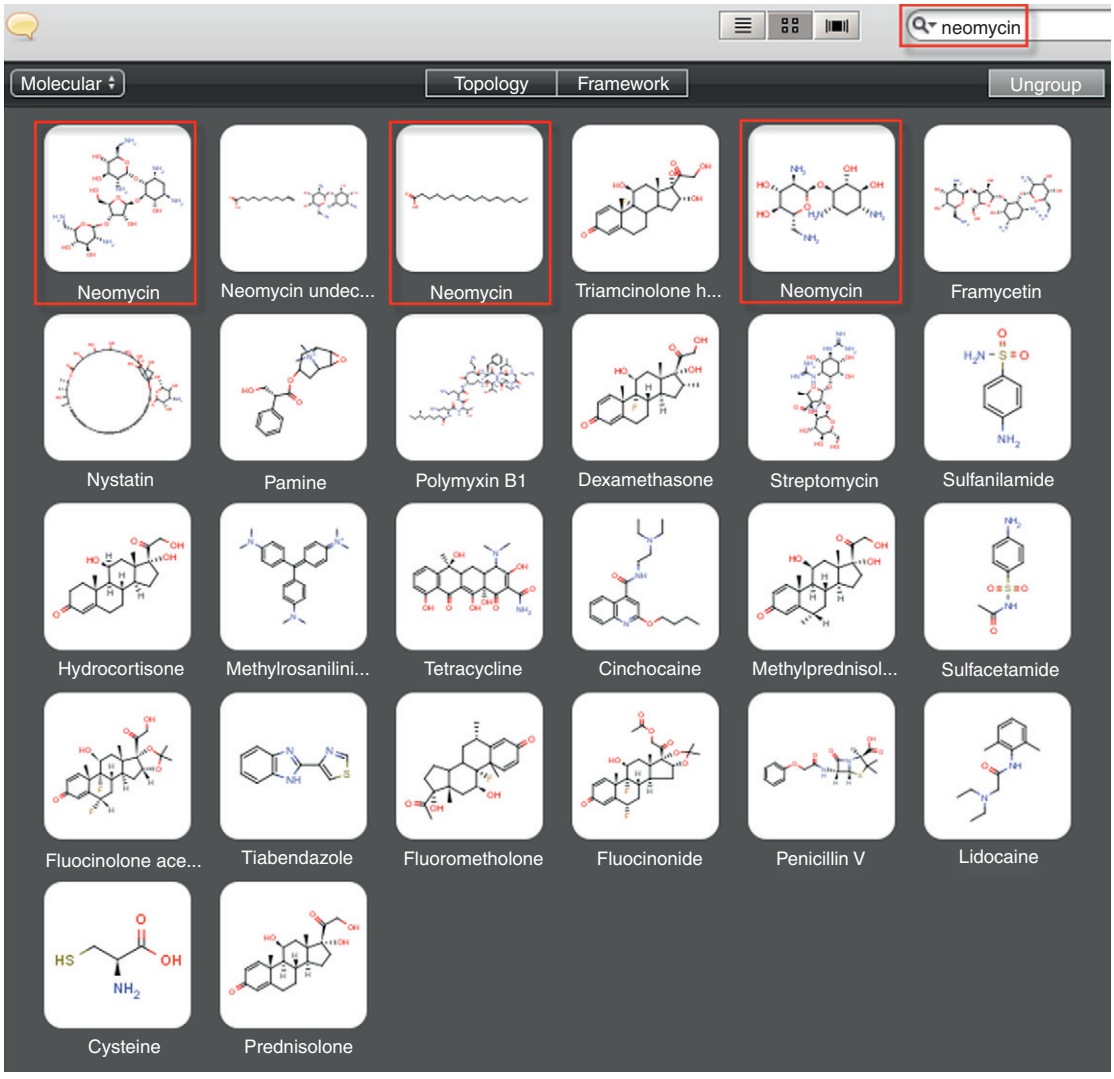


FIGURE 6

Image from the originally downloaded NPC browser showing the result of searching for neomycin, resulting in 26 structures. Abbreviation: NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.

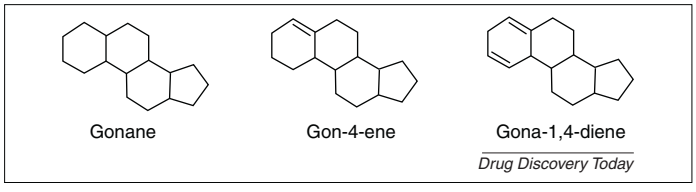


FIGURE 7

Steroidal substructures used for searching the NPC browser 'HTS screening' dataset. Abbreviations: HTS: high-throughput screening; NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.

substructures and the results were tabulated. A total of 149 unique compounds were retrieved and each was annotated as shown in Table 2. In each case only a fraction of the structures matched the correctly validated structure identified using the processes listed earlier. For each of the classes the majority of compounds had incomplete stereochemistry and for two of the classes, the gon-4-ene and gon-1,4-diene substructures, over 12% of the structures had complete but incorrect stereochemistry. This pattern was repeated in other cases throughout the database analysis.

TABLE 1

Summary of substructure search results for NPC browser searching with three steroid substructures.

Substructure	Number of hits	Number of correct hits	Number of stereochemistry	Incomplete stereochemistry	Complete but incorrect stereochemistry
Gonane	34	5	8	21	0
Gon-4-ene	55	12	3	33	7
Gon-1,4-diene	60	17	10	23	10

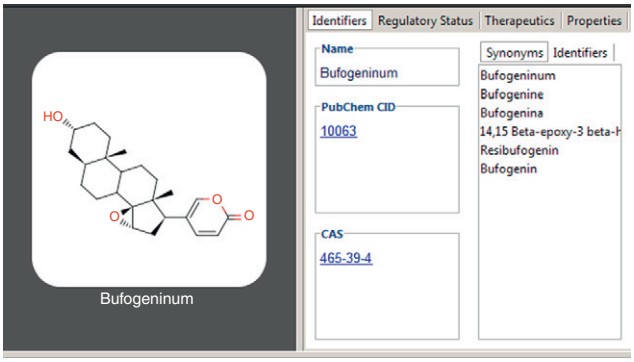


FIGURE 8

The result of searching for bufogenin in the original downloaded NPC browser data collection. Abbreviation: NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.

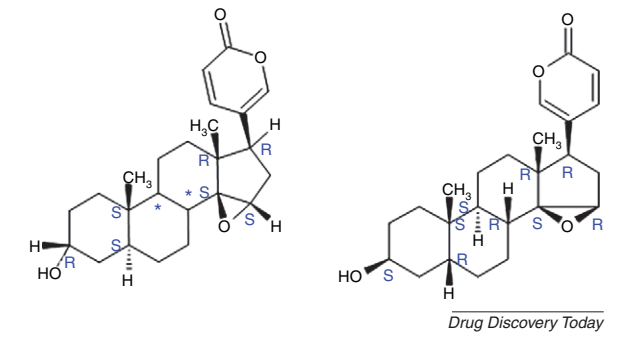


FIGURE 9

The structure of bufogenin: **(a)** the structural representation given in the NPC browser while **(b)** is determined from several validated sources. The *S* and *R* labels indicate specific stereocenters while the asterisks indicate undefined stereocenters. Abbreviation: NPC browser: NIH Chemical Genomics Center Pharmaceutical Collection browser.

Clicking on the hyperlinked CAS number in the NPC browser interface for bufogenin performs a search against PubChem for that CAS number. This search produces six compounds as shown in Fig. 10 (bufogenin structures in the PubChem database: http://

TABLE 2

A review of chemical structures retrieved based on name-based searches of the NPC browser (version 1.0.22) using a random selection of 50 of the top-selling US drugs.[a] The numbers in the error column refer to the list of errors given below the table.

Generic name	Correct structure	Number of hits	Error
Rosuvastatin		1	2
Zocor	×	1	
Thalidomide	×	1	
Taxol		1	2
Basen		1	1
Vytorin		1	4
Depakote		1	4
Symbicort*		1	3,4

TABLE 2 (*Continued*)

Generic name	Correct structure	Number of hits	Error
Spiriva		1	6
Prograf		1	2
Ezetimibe		2	5,8
Budesonide		1	3
Formoterol		3	3,8
Pioglitazone	×	2	2,8
Rabeprazole	×	1	
Anastrozole	×	1	
Nifedipine	×	1	
Goserelin		1	1
Sildenafil	×	1	
Cefdinir	×	1	
Cyclosporin		3	7,8
Clarithromycin		2	2,8
Tegaserod		1	6
Famotidine	×	1	
Drospirenone	×	1	
Tenofovir		4	3,8
Emtricitabine		4	3,8
Atorvastatin	×	1	
Clopidogrel	×	1	
Esomeprazole		3	1,8
Amlodipine	×	5	8
Olanzapine	×	3	8
Valsartan	×	3	8
Risperidone	×	4	8
Montelukast	×	1	
Quetiapine	×	1	
Lansoprazole	×	3	8
Losartan	×	1	
Alendronate	×	4	8
Escitalopram		1	1
Imatinib	×	3	8
Zolpidem	×	1	
Donepezil	×	2	8
Cetirizine	×	2	8
Irbesartan	×	2	8
Docetaxel	×	1	2
Sertraline	×	2	8
Oseltamivir	×	1	
Celecoxib	×	1	

[a] Source: Wikipedia (Wikipedia List of Bestselling Drugs: http://en.wikipedia.org/wiki/List_of_bestselling_drugs).
* Multicomponent combination drug.
1: Correct skeleton, no stereochemistry.
2: Correct skeleton, missing stereochemistry.
3: Correct skeleton, incorrect stereochemistry.
4: Single component of multicomponent structure.
5: Multiple components for single component structure.
6: No structure retrieved.
7: Incorrect skeleton.
8: Multiple structures based on name search.

PubChem SID	PubChem CID	StereoCenters	Original Source
96023536	10063	2 missing	KEGG
7848441	10063	2 missing	KEGG
153297	10063	2 missing	ChemIDPlus
99381037	6917974	Correct structure	Tractus
103121540	906501	Unrelated structure	AAA Chemistry
397470	259992	No stereochemistry	DTP/NCI

Drug Discovery Today

FIGURE 10

Six identifiers linked with bufogenin in PubChem.

www.ncbi.nlm.nih.gov/sites/entrez?db=pcsubstance&term=465-39-4). Each has the associated name bufogenin in addition to the associated CAS number of 465-39-4. There are four unique chemical structures contained within the PubChem database [listed as four different compound IDs (CIDs; 10063, 6917974, 906501, 259992) in Fig. 10]. Of the four-listed one is the correct structure, one has the correct connectivity but with no stereochemistry, one has the correct structure with two undefined stereocenters and one is a totally unrelated structure. This is typically representative of the confusions in name–structure relationships in the PubChem database.

Solutions to prevent data errors in databases: curation systems and validating data

Thus far we have provided some examples of quality in public domain chemistry databases. We clearly understand that no release of data, as highly curated or validated as it might (or might not) be, will be perfect. In the current age of nascent crowdsourcing, which we now find ourselves in, there is the ability to gather feedback from users of the database so that their comments can be reviewed and appropriate actions can be taken. This should be implicit in the design of any database. While the majority of chemistry databases online provide an email address to contact the database host this is rather an imperfect solution as any comments about a particular record are not associated with the record itself and are therefore unavailable to other users of the database to view. As a result obvious errors, even though reported to the database administrators, remain hidden from the community until appropriate action is taken. This is the case with some of the most 'trustworthy' databases including PubChem, ChemIDPlus and others. Sometimes responses are simply canned replies with no actions taken for months and, based on the experiences of the authors, in some cases years, and the feedback is not actioned. Other database hosts are responsive and pay attention to the feedback with a short cycle time. The authors have positive experiences with the hosts of the Drugbank, Chemical Entities of Biological Interest (ChEBI; http://www.ebi.ac.uk/chebi/) and ChEMBL (https://www.ebi.ac.uk/chembl/) databases specifically.

Other databases offer per record annotation and feedback as has been implemented on ChemSpider, DrugBank and on the NPC browser [25]. The Wolfram Alpha database (Wolfram Alpha database: http://www.wolframalpha.com/) provides the ability to submit feedback on individual records in the database. Based on our experiences all of our errors reported on the Wolfram database remain unresolved. The NPC browser includes a rather elegant approach to curation including the ability to edit the structure and remove and add synonyms, and secondary curators check the data and flag it further. As yet there does not appear to be a way to view a list of all curations made by a single contributor and this would make for an appropriate enhancement in our opinion.

The Nature Publishing Group (http://www.nature.com/npg), publisher of Nature Chemistry and Nature Chemical Biology, have contributed sincere efforts to ensuring that the chemistry data included within their articles is manually curated before publication (Jason Wilde, Adding structure to publishing chemistry: http://abstracts.acs.org/chem/239nm/program/view.php?obj_id=6897&terms). The data are also published into the PubChem database and as a result are made available to the community as a highly curated data set. The community would benefit from similar efforts by other chemistry publishers, especially if the chemistry content was provided with links to the published articles.

ChemSpider has implemented both curation and annotation capability and already offers the ability to registered users to provide direct feedback on a record-per-record basis regarding data quality. The 'dictionary' aspect of a chemical compound database, that is, the ability to retrieve high quality chemical compounds based on a name search, provides value to scientists [27], the greater value of course probably coming from the additional data and links associated with these records. Crowdsourced community participation has enabled for many tens of thousands of incorrect synonyms to be removed or validated on the database, thousands of incorrect chemical representations to be deprecated and new chemical compounds and spectra to be added. These contributions follow a similar power law trend to those demonstrated in internal projects, such as those at Pfizer (http://www.pfizer.co.uk/default.aspx) [28]. The number of curators is rather limited with less than 150 ever having contributed as of this writing. There are a very small number of dedicated contributors who have contributed thousands of curation actions to the database and these have proven to be of very high quality. In comparison, to date only 11 people ever have contributed to the curation of the data in the NPC browser, with five of them associated directly with the development of the software tool. While few people to date have actively contributed in these crowdsourcing efforts, this makes the task of cleaning up the databases immense for those who are actively involved, thus we require more automated mechanisms to ensure structures and data are correct.

Structure validation filters

The construction of a database of chemical compounds should attempt to deliver the highest quality of data possible to its users. For small collections of data, of a few thousand compounds for example, this is possible by manual curation of the data. As we have shown earlier with the analysis of data from the NPC browser even small databases are commonly not curated and can have many errors associated with them. For data aggregators of millions of chemicals, the curation of even the basic chemical compounds and associated identifiers is an enormous challenge when taking into account the scale of the data involved. However, it is possible to introduce some very basic structure validation filters into every system that can be used as immediate checks on the quality of data and provide automated flags for review as necessary. We are not

aware of any previous reports of such filters being compiled for use in such a manner, but recommend the following filters based on our experience in developing the ChemSpider database. These filters can be implemented in a manner that immediately reject the data at deposition or can be post-filtered following human review.

The suggested structure validation filters outlined might be largely applicable to new databases but can, of course, be applied to existing databases by reprocessing of the original data sets using the filters. Although it is a significant undertaking to rebuild these databases based on the proposed guidelines, and it might require additional funding, we do believe that such efforts will bring significant benefits.

Incorrect valence

Hypervalency is a rather common situation in many chemistry databases and the rejection of chemical compounds containing pentavalent carbon atoms, for example, should be a general rule. The implementation of such a rule would have removed such errors from the NPC browser (pentavalent carbons in the NCGC Collection in the NPC browser: http://www.chemconnector.com/2011/04/28/reviewing-data-quality-in-the-ncgc-pharmaceutical-collection-browser/). A related issue in some databases is the preference of display for nitro groups where some databases enable the functional group to be displayed with a pentavalent nitrogen while others prefer charge separated groups. It should be noted that this is a structure standardization choice and will be discussed below.

Atom labels

It is quite common for chemical compounds to be represented using atom labels of the type –Bz (for benzyl) and –Tos (for tosylate) and FMOC (for 9-fluorenylmethoxycarbonyl). In this case processing systems for the various databases hosting the data must correctly process and convert the data into an expanded form that accurately captures the intention of the label and provides an accurate structure representation that can be used for indexing within a database.

Aromatic bonds

While aromaticity is a well-known property of many organic chemicals, there are no agreed upon standards in representing aromatic molecules either graphically or encoded into chemistry format files. For example, for the SMILES format benzene might be represented either as c1ccccc1 or as C1=CC=CC=C1. There are various ways to depict aromatic rings including the Kekule form or explicitly designating aromatic bonds as solid-dashed parallel lines representing a bond order of 1.5. An alternative, often used to represent benzene, is a circle inside a hexagon. None of these approaches is 'chemically correct' *per se* and are simply representations of delocalization. Aromaticity handling by a particular cheminformatics software package requires rigor (Aromaticity detection, ChemAxon: http://www.chemaxon.com/marvin/help/sci/aromatization-doc.html) and the exchange of information between various software packages can be challenging. It is the logic in the various software packages that determines what kind of structure is encoded and how to standardize its representation. In this regard the interpretation depends on the software package and it is common to encounter multiple instances of the same chemical represented in different ways and listed in databases as different structures.

Non-zero total charge

It is quite common for chemicals to be represented in databases as active moieties associated with their salt counterions. For example, in Fig. 11 the compound is a disodium salt as evidenced by the associated chemical name and the 'SaltData' field. Unfortunately, when data is delivered in this format the structure deposited into the database will have obvious charge imbalance, the association with the chemical name will be incorrect and any experimental parameters in the file associated with a particular salt will be mismatched. Although there are appropriate cases where a chemical record should be associated with a charged species experience has proven that checking for a net zero charge is definitely of value in catching many errors at deposition. Many of the errors observed in the NPC browser in terms of compound-identifier mismatches probably result from aggregation of the data around the active drug component while ignoring the originally associated salt form.

Absent stereochemistry

As discussed earlier in the analysis of steroids in the NPC browser, it is rather common for stereochemistry to be excluded from chemical compound representations, whether intentionally or accidental. Steroids commonly show this issue as stereochemistry is generally 'assumed' based on standard steroid skeletons. Certain databases have also excluded stereochemistry from their collections when aggregating their data [e.g. the PDSP database (PDSP Ki Database: http://pdsp.med.unc.edu/kidb.php) sourced non-isomeric SMILES strings from PubChem for the structure representations]. While missed stereo bonds are of course acceptable for structure representations (e.g. for representation of racemates or

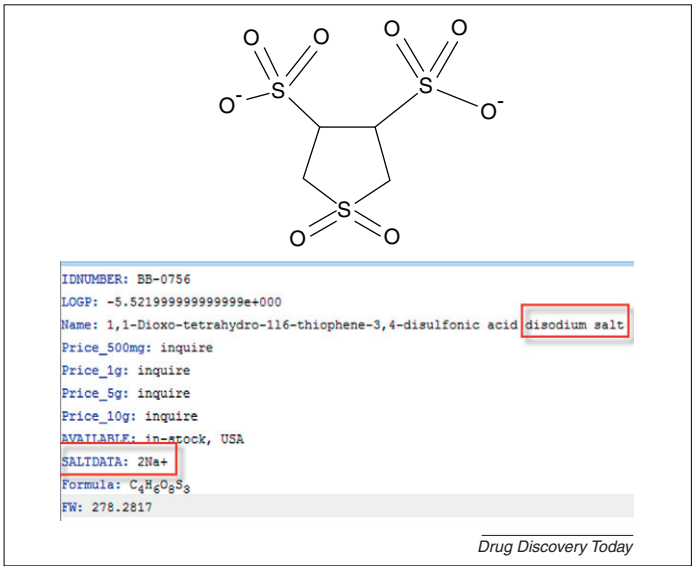


FIGURE 11

An example of a compound with charge imbalance and misassociation of name and structure. This commonly also leads to the misassociation of CAS numbers in chemical vendor files as vendors submit the chemical structure of a neutral compound but the CAS number for an associated salt. Abbreviation: CAS: Chemical Abstracts Service.

for unknown stereochemistry) it is the authors experience that in the majority of cases, as evidenced by the associated chemical name and/or identifier, that incomplete or absent stereochemistry is in fact an error. Flagging such compounds for manual review is an appropriate choice.

Salts with covalent bonds

While the majority of chemists would agree that a sodium carboxylate salt should be represented using a positively charged sodium ion and a negatively charged oxygen in the carboxylate anion, an alternative representation is a sodium atom covalently bonded to the oxygen as shown in Fig. 12. Using InChI [The IUPAC International Chemical Identifier (InChI): http://www.iupac.org/inchi/] the species are actually equivalent, as shown in the figure, whereas the SMILES string and standard molfile would clearly distinguish between them. For databases constructed using InChI as the basis of deduplication (e.g. ChemSpider) this equivalence has proven to be an issue in structure representation (ChemSpider: http://www.slideshare.net/AntonyWilliams/chemspider-an-online-database-and-registration-system-linking-the-web). In the case of the ChemSpider database several molfiles containing carboxylate groups with covalently bonded metals were deposited to the database and all future representations of the compound, whether ionic or not, were de-duplicated based on the InChI. Such misrepresentations on ChemSpider will be addressed in the future as part of a structure standardization project (*vide infra*). While clearly it is appropriate to have covalently bonded metals in many cases the identification of certain types of covalently bonded metals as a pre-filter is an appropriate step in validating data before deposition.

0D structure layout

The vast majority of molfiles contain 2D coordinates representing flat depictions of chemicals and, assuming appropriate layouts, are the most appropriate manner in which to depict chemicals. The submission of 3D molfiles for deposition can be quite common and, in general, is not an issue for simple molecules as the removal of the z-coordinate to flatten the molecule often produces an acceptable input. However, for complex molecules, and certainly for the majority of complex natural products, the resulting depiction can be hardly interpretable. For 3D conformations stereochemistry is encoded in the coordinates and stereo bonds (hash or wedge bonds) have no relevance. Chirality may (or may not) be encoded in the molfile but other types of stereochemistry (e.g. alkene) needs to be extracted from the 3D geometry before any flattening process takes place. Unfortunately, '0D molfiles', where all XYZ coordinates are set to zero, are also possible. The connectivities and bond orders between atoms are still contained within the file but all layout information is removed when all coordinates for all atoms are set to (0,0,0). In these cases the data should be pretreated using algorithmic 2D layout so as to ensure that visually interpretive data are available to the user. Without such 2D layouts the displayed compound will be confusing. An example of such an issue is where a 'hydroxyl group' shown in the NPC browser actually corresponds to the structure of silidianin, albeit without accurately encoded stereochemistry (Williams, A.J. Data Quality in the NCGC Pharmaceutical Collection Browser Part 4: http://www.chemconnector.com/2011/05/08/data-quality-in-the-ncgc-pharmaceutical-collection-browser-part-4/). It should be noted, however, that 0D structures can contain the encoded information for all supported types of stereochemistry if it was included.

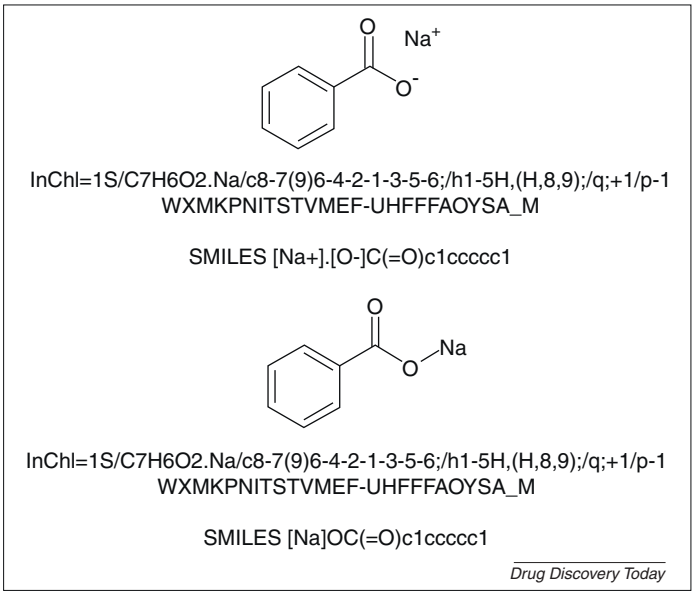


FIGURE 12

Alternative representations of a sodium atom with a carboxylic acid and the equivalency of InChIs versus SMILES strings. Abbreviations: InChI: International Chemical Indentifier; SMILES: Simplified Molecular Input Line Entry Specification.

Duplicated structures

Based on our experience of digesting data from various depositors into ChemSpider we have recognized that the submission of data containing multiple copies of the same compound is a rather common situation. The observation of a chemical record called 'terminal dimethyl', a record containing two methane molecules in the Drugbank database, enabled us to then trace the compound into the Wolfram Alpha database, PubChem, ChemSpider and others (Williams, A.J. 'Terminal dimethyl' means death by methane, twice: http://www.chemconnector.com/2011/10/31/terminal-dimethyl-means-death-by-methane-twice/). As a result of this observation we sought out similar records in ChemSpider ultimately removing almost 3000 of these twinned records. They probably arise as a result of attempting to represent racemates but we are not sure of the reason for their existence. Nevertheless, we see no value in storing such records in our database. The search and deprecation was also extended to larger multiples, such as triples, quads, among others, of the same compounds in a single record.

Data proliferation between databases

The distribution of online chemistry content is, in the opinion of the authors, dramatically overlapped in the majority of cases with only a small number of key resources adding data of value to the internet. PubChem serving the national library screening initiative is a valuable resource in terms of the hosting of bioassay data associated with hundreds of thousands of chemicals. The platform however offers many additional resources and, because the data are available for download and reuse (*vide infra*) many other

resources have used slices of the content as the basis of their own offerings, and generally link back to PubChem using CID links to drive traffic to their site. As a result data, both good and bad, proliferates among the databases and provenance is commonly confused or lost altogether. The situation is further confused when data providers who deposit to PubChem originally sourced the data from the same platform and in this manner they are simply looped depositions. Cheminformatics tools or standardization approaches used by the various hosts can differ and it is probably that data originally sourced from PubChem, when processed for hosting in a third party resource will be mapped back to a different, though related chemical. The cyclic processing of data through molfiles to SMILES to molfiles can introduce changes in stereochemistry and therefore add to the workload for those cleaning up the data.

There are numerous examples of data from PubChem being sourced and redeposited back into the database. For example, ChemSpider sourced their original seed set of 10.5 million chemicals from PubChem and redeposited the set back to PubChem later when the ChemSpider database was expanded to almost 20 million chemicals (Williams, A.J. PubChem Deposition of ChemSpider Data is Well Underway. My Favorite Color is Green: http://www.chemspider.com/blog/?p=279). NextBio sourced the majority of its content from PubChem and use it as part of their integrated content system on platforms, such as Elsevier (http://www.elsevier.com/wps/find/homepage.cws_home; Press Release: Elsevier and NextBio Sign Partnership to Enrich ScienceDirect Content: http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01249). NextBio also deposited their content back to PubChem (Structures from NextBio in PubChem: http://www.ncbi.nlm.nih.gov/sites/entrez?term=NextBio%5Bsourcename%5D&cmd=search&db=pcsubstance). Wolfram Alpha declares PubChem as one of their sources of information (Wolfram chemical data source information: http://www.wolframalpha.com/input/sources.jsp?sources=ChemicalData&sources=ElementData) but do not deposit, as yet, back to the platform. The Chemical Translation Service (CTS: Chemical Translation Service: http://uranus.fiehnlab.ucdavis.edu:8080/cts/homePage) hosted by the University of California [29] has clearly sourced the majority of their data from PubChem as the majority of erroneous name–structure synonyms have been migrated to the system. For example, the chemical record associated with methane is labeled as activated charcoal, includes most of the incorrect names from PubChem for methane (including diamond, soot, among others) but the issue is further compounded with errors by displaying incorrect structure representations as shown in Fig. 13.

Because data continues to proliferate between various online resources there is a continuous and growing problem as new databases appear. We are unaware of any reports detailing the processing and preparation of data other than the recent work describing the NPC browser but we can report on our experiences with developing the ChemSpider database.

ChemSpider has used both algorithmic and human curation to remove many thousands of inherited errors from the database. This has led it to being recognized as a valuable source of data which several online databases and publishers have shown an interest in linking to. ChemSpider data are now linked to other resources in two ways. A database host provides their set of chemical structures in either SDF format, as InChIs or in SMILES

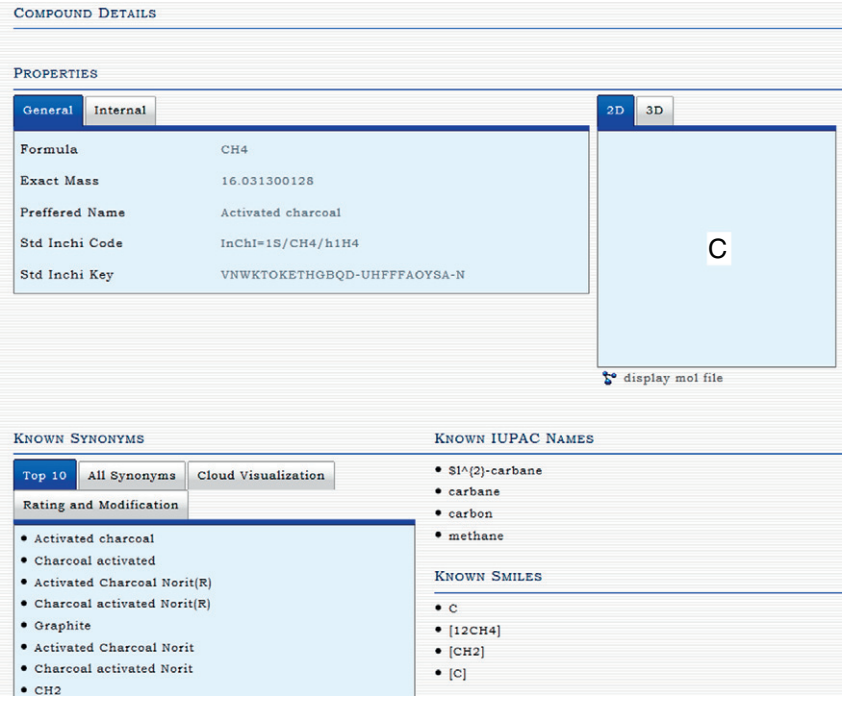


FIGURE 13

The chemical record for methane from the 'Chemical Translation Service' based on PubChem data. The molecular formula, InChI and mass agree with that of methane but the structure shown is of bare carbon. All displayed synonyms associated with the compounds are incorrect: charcoal, graphite, CH2, among others. Abbreviation: InChI: International Chemical Indentifier.

format. This file is then used to obtain associated ChemSpider IDs and the file is returned to the database host for them to insert the appropriate links to ChemSpider in their interface. This has already been done for the ChEBI database. An alternative manner in which to produce the links to ChemSpider is to use the ChemSpider web services (ChemSpider web services: http://www.chemspider.com/AboutServices.aspx) to search the database based on chemical structure (in one of several query formats). As a result of the web service query they will return the ChemSpider ID and insert it into their own database for linking. Several examples of this approach exist including that used by Nature Publishing Group to populate the chemical records associated with their articles in both Nature Chemistry and Nature Chemical Biology. Collaborative Drug Discovery also links to compounds that are registered in either public or private vaults [30]. It should be noted that the linking between the external sites and ChemSpider is based on retrieval of the ChemSpider ID associated with their query structure. This approach does not guarantee the validity of the association of the compound in their database with any chemical name, property or assertion. It is simply a link based on a look-up of the compound.

Structure standardization

An agreed upon set of standardization rules that can be agreed upon and implemented for all chemistry databases would greatly enable the interlinking between online resources and perhaps also help decrease errors. If both common standards and a common toolset were available then databases would be standardized in the same way and contain common identifiers for linking, for example, InChIs and SMILES, each generated using the same code base. The Open PHACTS project (Open PHACTS Project: http://www.openphacts.org/) has agreed on the need for a set of structure standardization rules that will be used to process all incoming chemical compound content that will be processed and hosted in the ChemSpider database serving the chemical services to the project. As the project is to serve the pharmaceutical industry it has been decided that the structure standardization guidelines provided by the FDA (Substance Registration System – Unique Ingredient Identifier (UNII): http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm) will form the foundation of the rule base, modified as necessary with the agreement of the EFPIA (EFPIA, European Federation of Pharmaceutical Industries and Associations: http://www.efpia.org/) members of the Open PHACTS consortium. Some basic rules extracted from the document include the handling of the nitro groups and salts as discussed earlier. It is to be expected that under the standardization rules that will be applied that we might see a significant reduction in the number of records in the aggregators database if the standardization process collapses tautomers.

Provenance in databases

The majority of online databases do not provide details regarding the provenance of all of their content. We think is an extremely challenging issue. As an example, for articles regarding chemical compounds on Wikipedia much of the content is aggregated by several co-authors, with only some of it sufficiently referenced, with the ChemBoxes (ChemBox Template in Wikipedia: http://en.wikipedia.org/wiki/Template:Chembox) or DrugBoxes (DrugBox template in Wikipedia: http://en.wikipedia.org/wiki/Template:Drugbox) containing various types of experimental data, identifiers and links to external resources. Until recently much of this data was not validated in any way but efforts are presently underway to validate the data and mark it as such (Wikipedia talk: WikiProject Chemistry/CAS validation: http://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Chemistry/CAS_validation). A ChemBox shows a chemical structure diagram, a systematic name, a list of identifiers, some links to online databases and a series of physicochemical properties. Ideally there needs to be some attribution as to where data or molecules came from. One of the reasons that provenance might not be provided is that the source of the data, and the associated license, might preclude such data sharing. Data licensing of online data is both a confusing and contentious issue but ultimately underlies the development of new systems, both commercial and public.

Crowdsourced review of public domain databases

Based on the discussions in this publication and others referenced herein, public domain databases contain data of variable quality. The value and utility of the databases depends not only on the quality and quantity of the content but also the mappings and associated metadata. Although we have focused on the quality of data based primarily on the mappings between chemical names and the correctness of the associated chemical structures, the overall value of the database is best defined by the users of the database resource and its content. We believe it is therefore appropriate to engage the community in providing their feedback regarding databases they use regularly. To facilitate this we are gathering input from the community through a Scientific Databases Wiki (Scientific databases wiki: http://www.scidbs.com/). The intention is to have both the hosts of scientific databases, as well as the users, contribute wiki pages. Because the wiki is an open environment anybody can register and contribute content. At present there are 15 chemistry related databases described on the database. For databases containing chemical compounds it is intended that some form of quantitative quality factor can be created that ranks the database. Williams [Structure representations in public chemistry databases: the challenges of validating the chemical structures for 200 top-selling drugs: http://www.slideshare.net/AntonyWilliams/structure-representations-in-public-chemistry-databases-the-challenges-of-validating-the-chemical-structures-for-200-topselling-drugs] has previously reported on a drug disambiguation exercise to validate the accuracy of the structure representations of more than 200 of the world's bestselling drugs in a series of databases. The work demonstrated that structure validation is a time-consuming and painstaking process susceptible to the performance of cheminformatics software tools and dependent on the cross-validation of various data sources. The result in this case was a quantifiable ranking of data accuracy in a series of well-known public databases.

Williams has reported that efforts are underway to share curation of the data on ChemSpider with other databases [ChemSpider – An Online Database and Registration System Linking the Web (Slide 21/73): http://www.slideshare.net/AntonyWilliams/chemspider-an-online-database-and-registration-system-linking-the-web]. A daily curation feed from ChemSpider has been established that summarizes the validation and deletion of

name–structure relationships on ChemSpider. The feed includes an International Chemical Identifier Key (InChIKey; InChIKey on the InChI Wikipedia Page: http://en.wikipedia.org/wiki/International_Chemical_Identifier#InChIKey) together with a list of validated names and deleted names. The InChIKey is a hashed version of the International Chemical Identifier, a text-based representation of a chemical structure, and can be used to check for the presence of the associated structure on another database. If the compound is detected through an InChIkey match then validated identifiers can be added and deleted identifiers can be removed thereby expanding the curation efforts of ChemSpider to other databases. An alternative use of the feed would be to use the validated names to search the database for the associated chemical and then compare the InChIKey from the feed with that associated with the chemical name in the database. If they do not match then the record can be flagged for manual inspection. Although it would be possible to pass out InChI strings, SMILES strings or molfiles in the validation feed the decision was made to not do this for the time being but to engage other databases in generating such a feed in a standard format. To date only the Drugbank database is utilizing the feed for validating their data but there is no reciprocal return as yet. The feed can be extended to include compound deprecation flagging, property value validation, among others, if the community chooses to engage in mutual sharing.

Concluding remarks

We have described some of the errors we are finding which are common to molecule databases. As chemistry content is expanding on the internet these errors are proliferating. Many errors can be identified quickly. For example, our analysis of the NPC browser 'HTS amenable compounds' subset of data for $\geq$7600 compounds identified fundamental errors in stereochemistry, valency issues and charge imbalances in a few minutes work using a rudimentary software tool. Such analyses can be performed by database owners before release. Even compounds that are suggested as having undergone 'quality control' have errors which could range from structural integrity to misassignment of synonyms, incorrectly associated CAS numbers or target mappings, among others. Correction of these errors manually in databases will be a considerable task. This raises several important questions such as how do we ensure that structures are as close to 100% correct as is possible based on assertion-based approaches and manual correction. Who corrects the errors in this database and who should be responsible for ensuring the integrity of such databases? The multitude of government funded databases, such as Environmental Protection Agency's (EPA's) Distributed Structure Searchable Toxicity (DSSTox) database [31], Aggregated Computational Toxicology Resource (ACTOR) [32] and Toxcast [33], NIH's PubChem [34] and the FDA's multitude of systems, and many others are all generating compound databases of differing quality and, surely, at this point, one would expect that it would be easy to gather a qualified set of well-known drug compounds with little effort. It is also probably a true statement that the quality of data in Wikipedia for many of these drugs is already of higher quality than most databases.

Although previous research efforts have gone into high-throughput analytical characterization of compound libraries [35] to identify impure or incorrectly synthesized compounds and to prevent ambiguous HTS results, there has been little research into compound quality in databases. Uses of compound databases include being combined with target annotation information and used to infer correspondence of molecule name and structure. These efforts are useful for ranking targets and looking at the druggability of targets and scaffold distribution. These types of metrics will be meaningless if the underlying structures are incorrect. The work also discusses commercial compound databases versus public databases [36]. Private databases are of a different scale. They did not look at quality of curation in the databases. Southan also described pairwise comparison of public versus private databases which will also be impacted by the fidelity of structures in each database.

Schuffenhauer *et al.* [37] described the use of ontologies for pharmaceutical compounds for pharmaceutical ligands and virtual screening. However, such efforts will certainly be nullified if the structures used in such databases are incorrect; thereby leading to incorrect classifications or retrieval of compounds from similarity searching that might be false positives. Chen *et al.* [38] described automated biochemotype annotation methods using PASS. If the underlying compound structures have errors then the predictions will also be erroneous.

The dangers of scientists taking the molecule structures in databases at face value are that the errors will profoundly impact their work [3,4,39]. Any computational models generated will be incorrect. If virtual hits are found by 3D screening this dataset they may also be misleading owing to the stereochemistry errors. It is not just an issue with a single database but many of these resources on the web in addition to commercial databases have errors [39]. As chemistry databases have proliferated in size these errors have accumulated. While some are checking for errors and correcting as they are suggested, this is the exception rather than normal. We have called for a good-faith effort for checking the data content carefully before making the database public [23]. We suggest that there needs to be a considerable investment in structure integrity checking software and more manual curation efforts.

There does not appear to be any change on the horizon in terms of the number and nature of these chemistry databases that, based on our evidence-based examination of data reuse and proliferation, will continue to distribute data of unknown quality across the internet. This disturbing and continuing trend needs to be managed. A basic gold-standard drug look-up dictionary of correct structure files for drugs with their associated synonyms is not yet available online. It has been acknowledged that even the FDA does not have a repository of approved drugs [25] which is quite shocking in this day and age. One of the reviewers for this article commented that the FDA cannot access information on a drug unless the Investigational New Drug (IND): http://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/investigationalnewdrugindapplication/default.htm tracking number is known but the IND tracking number is considered proprietary information and is not in the public domain. We can envisage a change in this situation on all fronts through collaboration and some of the large-scale semantic web efforts, such as Open PHACTS. While we have focused the majority of our discussion on small molecule databases we have also highlighted that others have found errors in various databases that are used throughout the drug discovery and development pipeline, these will also require careful curation. Ultimately data validation is a human activity performed

by experts and a clear path forward to engage community participation is likely to require direct funding, some form of rewards and recognition to encourage engagement, or depend on the charitable nature of skilled scientists to contribute. Perhaps a combination of all of the above might be necessary if we are ultimately going to achieve something closer to a gold standard database of molecules.

Conflicts of interest

Antony J Williams is employed by The Royal Society of Chemistry which produces ChemSpider discussed in this article. Sean Ekins consults for Collaborative Drug Discovery, Inc.

Licensing of data content

Online databases mix and aggregate content on a regular basis, linking to each other, losing provenance in many other cases and, in the process of passing through cheminformatics tools, sometimes changing the nature of the chemical compounds. An even larger issue is the potential fragility of the online databases based on poorly understood licensing for each of the databases. The chemical blogosphere has been host to many discussions regarding the need for clear data licensing definitions on chemistry related data. In particular, Murray-Rust (Peter Murray-Rust webpage: http://www.ch.cam.ac.uk/person/pm286) espouses the value of 'Open Data' (Open Data on Wikipedia: http://en.wikipedia.org/wiki/Open_data) to the scientific discovery process and encourages clear licensing of all chemistry data according to Open Knowledge Foundation licensing (Open Knowledge Foundation: Open Data Licensing: http://wiki.okfn.org/Open_Data_Licensing) and the so-called Panton Principles (Panton Principles: Principles for Open Data in Science: http://pantonprinciples.org/). It is generally accepted that individual data points cannot be copyrighted but that data collections might be copyrighted. Therefore a single data point such as a melting point cannot be copyrighted. Neither can a connection table, InChI or Simplified Molecular Input Line Entry Specification (SMILES). However, a depiction of the chemical compound as a figure in a publication 'can' be copyrighted, even though the appearance of FigShare (Figshare: http://figshare.com/) now facilitates making figures open and available to the community. In theory, if authors uploaded their figures to FigShare (or other online storage such as Flickr (Flickr photo sharing system: http://www.flickr.com/)) before publication even following copyright transfer to publishers near identical images will be available to be sourced from the internet. It is difficult to define where data transitions to become a copyrightable collection. Is a file containing 100 chemical structures, associated chemical identifiers and experimental parameters such as melting points copyrightable? Based on the activities of commercial businesses in this domain the answer is likely to be a yes.

Numerous well-known databases supporting the life sciences are freely available for download. These include, as mentioned earlier, PubChem, DrugBank, ChEBI, ChEMBL and the PHYSPROP data collection (PHYSPROP database: http://srcinc.com/what-we-do/product.aspx?id=133). A review of the licensing details for each provides a variety of details regarding the terms and conditions of usage. In general, of the many databases available online, the licensing of the majority of the data is undefined. The majority of SDF files downloadable from chemical vendor websites have no defined licenses at all. Despite the assumptions that PubChem data are 'Open', because the data are downloadable, they are not provided with any specific licenses *per se* but rather depositors assign rights simply by depositing data, thereby indicating acceptance of the depositors agreement. It is unlikely that the majority of scientists who download the data are aware of any license limitations constraining the data usage and have not concerned themselves with whether it is appropriate to monetize the data or repackage and redistribute under new licenses. It is just as unlikely that all depositors have fully understood that their data can be downloaded, redistributed and, ultimately, licensed, commoditized and monetized.

The ChEMBL database hosted by the European Bioinformatics Institute was recently released under a Creative Commons data license (ChEMBL Creative Commons Case Study: http://wiki.creativecommons.org/Case_Studies/ChEMBL) and it is hoped that more databases will be released with such clarity in the future. That said, even such well-defined and community accepted licenses can be abused. The continuation of an original license through other aggregators is also difficult to police and the deposition of ChEMBL data to PubChem is made under the PubChem data transfer agreement (PubChem data transfer agreement: http://pubchem.ncbi.nlm.nih.gov/deposit/docs/PubChem_Data_Agreement.pdf) and will probably confuse the majority of the community to believing, once again, that all data are public domain, therefore requiring no attribution.

Acknowledgement

The authors kindly acknowledge the reviewers for their constructive comments.

References

1 Brzustowicz, L.M. *et al.* (1993) Molecular and statistical approaches to the detection and correction of errors in genotype databases. *Am. J. Hum. Genet.* 53, 1137–1145

2 Migliavacca, E. *et al.* (2001) MDB: a database system utilizing automatic construction of modules and STAR-derived universal language. *Bioinformatics* 17, 1047–1052

3 Fourches, D. *et al.* (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 50, 1189–1204

4 Oprea, T.I. , Olah, M. , Ostopovici, L. , Rad, R. and Mracec, M. (2003) On the propagation of errors in the QSAR literature. In *EuroQSAR 2002 – Designing drugs and crop protectants: Processes, problems and solutions* (Ford, M., Livingstone, D., Dearden, J., Van de Waterbeemd, H., eds), pp. 314–315, Blackwell Publishing, New York, NY

5 O'Neil, M.J. (2006) *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals* (14th edn), Merck & Co.

6 Prinz, F. *et al.* (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10, 712

7 Sheridan, R.P. and Shpungin, J. (2004) Calculating similarities between biological activities in the MDL drug data report database. *J. Chem. Inf. Comput. Sci.* 44, 727–740

8 Wittig, U. *et al.* (2004) Classification of chemical compounds to support complex queries in a pathway database. *Comp. Funct. Genomics* 5, 156–162

9 Clarke, D.L. *et al.* (2008) Applying modern error theory to the problem of missed injuries in trauma. *World J. Surg.* 32, 1176–1182

10 Goldberg, S.I. *et al.* (2008) Analysis of data errors in clinical research databases. *AMIA Annu. Symp. Proc.* 242–246

11 Finney, J.M. *et al.* (2011) An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med. Inform. Decis. Mak.* 11, 7

12 Ioannidis, J.P. (2011) An epidemic of false claims. Competition and conflicts of interest distort too many medical findings. *Sci. Am.* 304, 16

13 Ioannidis, J.P. and Khoury, M.J. (2011) Improving validation practices in 'omics' research. *Science* 334, 1230–1232

14 Castaldi, P.J. *et al.* (2011) An empirical assessment of validation practices for molecular classifiers. *Brief Bioinform.* 12, 189–202

15 Bell, A.W. *et al.* (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* 6, 423–430
16 Lamontagne, J. *et al.* (2010) Proteomics-based confirmation of protein expression and correction of annotation errors in the *Brucella abortus* genome. *BMC Genomic* 11, 300
17 Zhang, C. *et al.* (2009) Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. *Bioinformatics* 25, 2708–2714
18 Jeong, E. *et al.* (2011) Ontology-based instance data validation for high-quality curated biological pathways. *BMC Bioinform.* 12 (Suppl. 1), S8
19 Wong, W.C. *et al.* (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput. Biol.* 6, E1000867
20 Davis, A.M. *et al.* (2008) Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today* 13, 831–841
21 Fu, X. *et al.* (2011) Data governance in predictive toxicology: a review. *J. Cheminform.* 3, 24
22 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36 (Database issue), D901–D906
23 Williams, A.J. and Ekins, S. (2011) A quality alert and call for improved curation of public chemistry databases. *Drug Discov. Today* 16, 747–750
24 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371
25 Huang, R. *et al.* (2011) The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.* 3, 80ps16
26 Baker, M. (2006) Open-access chemistry databases evolving slowly but not surely. *Nat. Rev. Drug Discov.* 5, 707–708
27 Hettne, K.M. *et al.* (2010) Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining. *J. Cheminform.* 2, 4
28 Ekins, S. *et al.* (2011) *Collaborative Computational Technologies for Biomedical Research*. Wiley
29 Wohlgemuth, G. *et al.* (2010) The chemical translation service – a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* 26, 2647–2648
30 Ekins, S. *et al.* (2011) Pioneering use of the cloud for development of the collaborative drug discovery (cdd) database. In *Collaborative Computational Technologies for Biomedical Research*, (Vols. 335–361) (Ekins, S. *et al.* eds), Wiley and Sons
31 Richard, A.M. (2006) DSSTox web site launch: Improving public access to databases for building structure–toxicity prediction models. *Preclinica* 2, 103–108
32 Judson, R. *et al.* (2008) ACToR – aggregated computational toxicology resource. *Toxicol. Appl. Pharmacol.* 233, 7–13
33 Dix, D.J. *et al.* (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* 95, 5–12
34 Wang, Y. *et al.* (2009) PubChem: a public information system for analysing bioactivities of small molecules. *Nucleic Acids Res.* 37 (Web Server issue), W623–W633
35 Kenseth, J.R. and Coldiron, S.J. (2004) High-throughput characterization and quality control of small-molecule combinatorial libraries. *Curr. Opin. Chem. Biol.* 8, 418–423
36 Southan, C. *et al.* (2009) Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminform.* 1, 10
37 Schuffenhauer, A. *et al.* (2002) An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* 42, 947–955
38 Chen, X. *et al.* (2006) Toward automated biochemotype annotation for large compound libraries. *Mol. Divers* 10, 495–509
39 Olah, M. *et al.* (2005) WOMBAT: world of molecular bioactivity. In *Chemoinformatics in Drug Discovery* (Oprea, T.I., ed.), pp. 223–239, Wiley