Published in final edited form as:
Drug Discov Today. 2019 April ; 24(4): 1010–1016. doi:10.1016/j.drudis.2019.01.010.

Assessing the public landscape of clinical-stage pharmaceuticals through freely available online databases

Rebekah H. Griesenauer, **Constantino Schillebeeckx**, **Michael S. Kinch**[*]
Center for Research Innovation in Biotechnology, Washington University in St Louis, MO 63130, USA

Abstract

Several public databases have emerged over the past decade to enable chemo- and bioinformatics research in the field of drug development. To a naive observer, as well as many seasoned professionals, the differences among many drug databases are unclear. We assessed the availability of all pharmaceuticals with evidence of clinical testing (i.e., been in at least a Phase I clinical trial) and highlight the major differences and similarities between public databases containing clinically tested pharmaceuticals. We review a selection of the most recent and prominent databases including: ChEMBL, CRIB NME, DrugBank, DrugCentral, PubChem, repoDB, SuperDrug2 and WITHDRAWN, and found that ˜11 700 unique active pharmaceutical ingredients are available in the public domain, with evidence of clinical testing.

Keywords

In silico drug discovery; chemoinformatics; bioinformatics; clinical trials

Introduction

The recent rise in the number of drug discovery databases has initiated interest in analyzing the usage and adoption of each database and the interconnections and cross-fertilization that exists between them. A variety of databases focusing on targets, proteins, metabolism and active pharmaceutical ingredients (APIs) has proliferated, particularly over the past 5 years. These databases are crucial resources for *in silico* drug discovery, for prioritizing repurposing opportunities and for identifying trends in the drug development enterprise. Drug repurposing has become increasingly attractive in recent years as a less expensive option with lower barriers to approval than traditional drug discovery and development. Therefore, it is of interest to assess the current public landscape of approved APIs and APIs that have been in clinical trials.

[*] *Corresponding author*: Kinch, M.S. (michael.kinch@wustl.edu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Teaser: This review highlights the major differences and similarities between public databases containing pharmaceuticals with evidence of clinical testing and assesses the current landscape of clinical-stage pharmaceuticals (experimental, failed and approved).

Assessing the public landscape of clinical-stage APIs involves comparison of various biopharmaceutical databases, which is a notoriously challenging task. Yonchev *et al.* [1] reported on the redundancies in PubChem and ChEMBL. The authors pointed out specific aspects that made comparing the two databases difficult. For example, in PubChem the same API can have multiple compound records. Furthermore, differences in terminology and ways in which the databases are structured impede straightforward comparison of database content. Furthermore, owing to the compound submission model of PubChem, a researcher might run the risk of extracting duplicate compounds. Southan *et al.* compared chemical structures in ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database in 2013 [2]. In 2018, Southan reviewed the largest chemical structure databases: PubChem, ChemSpider and UniChem, which contain 95, 63 and 154 million chemical structure records, respectively [3]. Here, Southan examined the databases' contributing sources and found that sources common among databases could have substantial differences in chemical structure count. Fourches *et al.* [4] provided guidance on reviewing and comparing chemogenomic datasets with suggestions for how to curate and clean chemical datasets and discuss the importance of properly cleaning chemical datasets, including removing duplicates. Ambiguity in and across databases can confound efforts to model and analyze data.

Although not immediately obvious, one fact emerging from our studies was that each database has a distinct emphasis and target audience. Chemistry-based databases, such as PubChem and ChEMBL, contain large-scale record counts of compounds with potential medicinal uses. Other databases, such as DrugBank, focus on unique APIs, most of which convey some evidence of clinical interest. Several open databases specialize in specific subsets of drugs, such as approved or withdrawn medicines. In this review, we have selected databases to meet the following criteria: (i) are public and freely available with downloadable data; (ii) are compound oriented and contain clinically tested compounds; (iii) have at least one peer-reviewed publication describing the content and construction of the database. The databases meeting these criteria are: ChEMBL [5], CRIB NME [6], DrugBank [7], DrugCentral [8], PubChem [9], repoDB [10], SuperDrug2 [11] and WITHDRAWN [12]. Whereas additional databases are undoubtedly available through commercial sources on a subscription basis or as the result of extensive competitive intelligence, these are not freely available and therefore not included in our present analysis. Instead, we focus on public and freely available databases.

For the work herein, a short summary of each selected database is provided. The usage and adoption of each database is discussed by analyzing peer-reviewed publications citing each database. The relationships between each database is examined according to the sources used for construction as well as the overlap in clinical-stage drug compounds in each database. Finally, we summarize the current number of pharmaceuticals in the public domain with evidence of inhuman experience. Raw data files and codes for this review can be found at: https://github.com/WUSTL-CRIB/clinical_databases_review.

Overview of selected databases

The past decade has witnessed an increase in publications citing public drug databases. In Figure 1, we review the largest and most cited databases: PubChem, Chembl and DrugBank. Whereas there were only two mentions of PubChem, ChEMBL or DrugBank in the literature in 2004, the field began to grow soon thereafter. By 2010, the rate of annual citations was nearing 500 and would more than double within 2 years. This trend continues as the rates of new citations have continued to climb after 2010. At present, the number of citations of these databases exceeds 2500 per year.

A summary of the different databases is provided in Table 1. The size of each database in terms of 'compound records' and a subset of 'clinical-stage compounds' is given. PubChem and ChEMBL dwarf the other databases in this analysis. PubChem, by far the largest, contains almost 100 million compound records. PubChem is an open database, which relies upon investigator submission of data. Beyond conveying the properties of the molecules themselves, PubChem aggregates supplementary data pertaining to the chemistry and bioactivity of compounds including chemical structures, physical properties, toxicity and safety information. The most impactful articles citing PubChem include the descriptions of the DrugBank database [13] (1318 citations) and the IUPHAR/BPS Guide to PHARMACOLOGY [14] (cited by 874), which both provide external links to PubChem. Other commonly cited articles referencing PubChem include a chemical visualization and analysis platform, Avogadro [15] (with 1318 citations), and an article focusing on quantitative HTS [16] (with 457 citations).

ChEMBL is a bioactivity database with >2 million compound records created by mining and extracting data from medicinal chemistry literature. Recently, ChEMBL expanded the focus of their offerings to aid the drug discovery and development process by including data generated during preclinical and clinical investigation (including metabolism, mechanisms of action and therapeutic indications) [5]. Some of the most impactful applications of the ChEMBL database include medicinal chemistry applications such as *in silico* target predictions [17], visualization tools including Reactome.org (714 citations) and biopharmaceutical databases created from subsets of ChEMBL data [18] or databases providing external links to ChEMBL [14].

DrugBank is the third-largest database and provides a rich source of quality data on drugs in the preclinical, clinical and approved stages of drug development. With >10 000 compounds, DrugBank pairs drug and target information in an encyclopedic manner, providing the chemo- and bio-informatics communities with a rich source of data. Articles describing the construction and content of the DrugBank database are themselves highly cited, with >4500 citations of their original publication and subsequent update publications in *Nucleic Acids Research*. Other impactful articles citing DrugBank include descriptions of drug discovery trends and the use of novel drug targets [19] (439 citations), a web server for identifying potential drug targets [20] (237 citations) and a protein database constructed from DrugBank data subsets [21] (193 citations).

The CRIB NME database is a collection of FDA-approved new molecular entities (NMEs). The unique features of the CRIB NME database are the inclusion of data on the organizations responsible for the R&D of each NME, linked patent information, approval dates for all NME data (including those not found within current FDA regulatory databases) and a focus on capturing historic NME data (with approval dates going back to and before the 1800s). Impactful articles reviewing pharmacogenomics in the clinic [22] (165 citations), natural product discovery [23] (125 citations) and reviews of oral drug candidates [24] (102) all cite the CRIB NME database paper [6].

The WITHDRAWN database contains 270 drugs withdrawn for toxicity and/or safety reasons and 308 drugs withdrawn for reasons other than safety (such as market obsolescence or ineffectiveness). WITHDRAWN provides information such as chemical structures, withdrawal reasons, toxicity information and protein targets. The focus of the WITHDRAWN database is toxicity and adverse events of drugs. A unique data field within the WITHDRAWN database is the curated ‘reason for withdrawal’ field, which has been manually extracted and labeled from scientific literature. The WITHDRAWN database was added to ChEMBL in 2017 [5]. Other impactful references to the WITHDRAWN database include looking at trends in molecular medicine [25] (40 citations) and studies looking at safety [26] and toxicity [27].

The motivation for SuperDrug2 was the burdensome nature of manually patrolling regulatory websites to investigate components for use with *in silico* drug discovery. The unique features of SuperDrug2 include the pharmacokinetic simulator and 3D drug conformer visualizations. Articles referencing SuperDrug2 include the 2018 Southan *et al.* review of major chemistry databases [3], and drug repositioning [28] and drug design studies [29].

DrugCentral also focuses on molecules approved from regulatory agencies across the world. DrugCentral contains unique regulatory information (such as FDA drug labels and approval dates), drug indications, contraindications and off-label indications. Another unique feature of DrugCentral is an emphasis upon frequent updates by continuously monitoring the activities of regulatory agencies worldwide. Impactful articles referencing DrugCentral include the IUPHAR/BPS Guide to PHARMACOLOGY [14], which again provides external database links to DrugCentral, an article providing a comprehensive map of drug targets [30] (192 citations) and a study using DrugCentral’s drug target bioactivity data to study the druggable genome [31] (19 citations).

As the name implies, repoDB has a drug repurposing focus. repoDB contains true positives (i.e., approved drug-indication pairs) and true negatives (i.e., failed drug-indication pairs) to train and validate predictive models for drug repurposing. The unique features of repoDB is that it contains drug-indication failure information to remove the bias in building predictive drug models on successful drug-indication pairs alone. As one might expect, articles citing repoDB are primarily concerned with drug repurposing [32–34].

In aggregate, these databases contain an overwhelming amount of information, which can unintentionally impede the overall impact of any given dataset. Immersed in data, many researchers have created or deployed visualization tools and applications that function as subsets of other databases with a focus on a particular aspect of drug development. For example, Reactome.org [35] created a website with tools for browsing pathway and interaction data by interacting with the ChEMBL database. Additionally, Mysinger *et al.* [18] used a subset of ChEMBL data to create a resource of ligands and decoys for improved benchmarking. PDTD is a web-accessible protein database created using a subset of DrugBank data [21] and WITHDRAWN was added as a data source to PubChem in 2017. Databases focused on approved drugs (SuperDrug2, CRIB NME and DrugCentral) were created to establish resources for analyzing clinically successful drugs. The WITHDRAWN database specializes in drug toxicities and, as the name suggests, archives information about medicines that have been approved and subsequently withdrawn from the market. The CRIB NME database focuses exclusively upon FDA-approved medicines, including withdrawn drugs. DrugBank is a curated database that contains preclinical, clinical-stage and approved drugs. An overview of the contents of each database in terms of drug development staging is shown in Table 2. ChEMBL and PubChem contain compounds at all stages of the drug development spectrum. Other databases focus on approved or withdrawn drugs only.

Usage and adoption

To quantify the adoption and usage of each database, peer-reviewed publications were extracted from the Elsevier Scopus literature database. For the top-three most-cited drug databases (ChEMBL, DrugBank and PubChem) the database name itself was used as the search term to obtain a set of publications that cite the database. For the remaining databases, articles citing the main manuscript containing the database description were included for analysis.

Table 3 summarizes the adoption of the selected databases from the citation perspective as well as geographic location of the authors. The most widely adopted database is PubChem with 6122 citation records, followed closely by DrugBank with 5675 citation records. Newer databases such as SuperDrug2 (2018), repoDB (2017) and DrugCentral (2017) have the least amount of citations records. Author affiliations gathered from the records citing each database provided a means to summarize the adoption of each database by country. Overall, the USA is the primary consumer of drug databases followed by the UK, France, China and India. However, we did identify apparent geographic preferences, such as a disproportionate citation of ChEMBL and WITHDRAWN databases by European (German and British) investigators.

Source analysis

We examined the relationships between sources used to create open drug databases along with external links to other databases to assess their interconnectivity. The sources of each database were compiled from publications describing the content and construction of the database. Online descriptions of data sources from the websites hosting each database were also included in this evaluation. A network graph of all sources and external links and their shared connections to each database is shown in Figure 2. Overall, the total number of unique data sources amounts to 688, with 87% of those being attributed to PubChem. PubChem and Drugbank share the most sources and external links, totaling 14.

This study highlighted the connectivity of the selected databases in terms of sources used for database construction and links to external databases. The network is color-coordinated to highlight which source is used for each database. When databases share a source or external link, two or more colored edges connect at a single node (or source). Data sources used most frequently are shown as numbered nodes in the network graph. These include the WHO ATC index [36], PubMed publications, FDA-approval packages, the KEGG database [37] and the Therapeutic Target Database (TTD) [38]. This network analysis reveals each database has a handful of unique sources but they share a significant portion of sources with other databases. The network also shows that many databases cite each other as sources or provide links to external databases for cross-referencing and user convenience. For example, ChEMBL, DrugBank, WITHDRAWN and DrugCentral use PubChem as a source or external link and PubChem uses DrugCentral, ChEMBL and DrugBank as sources or external links.

Content analysis

To assess the uniqueness and overlap of each drug database, compounds with evidence of clinical testing were extracted from each database. Evidence of clinical testing included links to any clinical trial information or records labeled as ‘approved’ or ‘withdrawn’ in any database. Data formats varied among the numerous data sources, therefore general coding ability in languages such as Python, Bash and Structured Query Language (SQL) was required to access and thoroughly evaluate all data included in this analysis. ChEMBL and DrugCentral make database dumps available for direct download. Database dumps contain records of the contents and structure of the database and are usually stored as SQL statements, allowing users to build local versions of a database. CRIB NME, repoDB and SuperDrug2 make all their data available in comma-separated value (CSV) formatted files. DrugBank and WITHDRAWN provide chemical-data SDF files, which were programmatically parsed into CSV-formatted files using custom python scripts. Finally, all clinicaltrials.gov sourced data were downloaded from PubChem using their pug_view API and then parsed using a python script. With all the data converted into a delimited format, they were then manually inserted into a single database for analysis.

A unique list of molecular entities was generated by merging together compounds by SMILES strings and synonyms provided by each database. First, we manually removed salts or solvents to achieve a list of APIs. When available, the SMILES string was downloaded from the source database. Then, the ChemDraw v16 plugin for Excel was used to convert any compound names into chemical structures using the Name>Struct functionality. The structures generated by the name-to-structure function were converted into SMILES strings using the CHEM_SMILES function. Afterwards, SMILE strings were checked during the merging process with the assumption that duplicate SMILES strings were the same API. After programmatic merging of compounds, the API list was manually reviewed to identify possible programmatic merging errors. This list was then used to probe each database and a link was created if the compound was present. Figure 3 shows the resulting overlap summaries. The number of overlapping clinical-stage compounds between any two databases is shown in Figure 3. With this visualization, one can deduce that 99% of repoDB compounds can be found in ChEMBL, DrugCentral and DrugBank. Furthermore, 98% of SuperDrug2 compounds can be found in DrugCentral but only 82% of DrugCentral compounds can be found in SuperDrug2. Ninety-nine percent of the WITHDRAWN database can be found in PubChem. Another major observation is that no database is a 100% subset of another.

Discussion

Barriers to entry can impede the usefulness for compound databases for certain analytical efforts. For example, selecting drug compounds that have been through clinical testing is not a straightforward task in large compound databases (such as PubChem and ChEMBL). A handful of databases contain clinical-stage drugs but the extent to which clinical trial data are linked is limited. It is not clear and very difficult to ascertain which compounds are unique to a particular source because many of the databases do not explicitly link source information to a compound. The explicit linking of source information at the record level to APIs would be a significant improvement to any biopharmaceutical database that does not already include the metadata.

There are several limitations to comparing the contents of pharmaceutical databases. For example, name-to-structure comparison and merging of entities based on synonyms can be error prone. There are several limitations to merging entities based on SMILES strings. One issue to note is that not all entries have SMILES strings, such as large macromolecules (i.e., biologics), and therefore names and synonyms alone were used to remove ambiguity. Of the 11 763 APIs included in this analysis, ~3840 were not small molecules. Furthermore, as pointed out by Fourches *et al.* [4], merging compounds using noncanonical SMILES strings can overlook duplicates. It is also possible that a small set of chemical structures was converted to SMILES strings incorrectly. Although we believe these errors to be small relative to the size of the whole dataset, improved chemical data curation methods, such as those outlined by Fourches *et al.* [4], could yield more duplicates than currently detected in our dataset. Overall, we do not recommend using this dataset in its current form for applications such as QSAR models because duplicates might still exist. Furthermore, most of these databases are constantly updated and new databases might have emerged while performing this study. Therefore, the contents of this review could already be slightly out of date.

In summary, PubChem and ChEMBL are large databases of compounds emphasizing medicinal chemistry. Some databases might have been originally built for one audience but later expanded their capabilities to reach a more diverse user pool. Given this complexity and specialty, the decision as to which database should be deployed to address problems outside their focus can present a considerable challenge. Based on our experience, PubChem and ChEMBL provide the greatest breadth of medicinal compounds but can be challenging to navigate. DrugBank is not as comprehensive as PubChem and ChEMBL but has reliable curated data that are relatively more approachable for non-experts to analyze. DrugCentral and SuperDrug2 are great resources for drugs that have successfully navigated the drug development pipeline. WITHDRAWN is specifically valuable for toxicology studies. The CRIB NME database contains useful information regarding NMEs and the organizations involved in their development but is again limited to FDA-approved therapeutics. Finally, repoDB, although relatively small, contains approved and investigational drug-indication pairs and is an excellent resource for drug repurposing studies. It is interesting to note, however, that there is not one database that represents a comprehensive source of data for compounds of clinical interest. During our content analysis, we combined all data sources into a comprehensive database and found that ˜11 700 unique APIs are currently available in public databases with some evidence of clinical experience. In future work, we will employ more-rigorous curation methods to study these clinical APIs to review overall trends in clinical-stage drug development.

Concluding remarks

In this study, we compiled >11 700 unique APIs with evidence of clinical testing. Online drug databases are crucial expeditors of *in silico* drug discovery and chemoinformatics, especially in academic, biotechnology startup and non-profit environments where high-cost subscription databases are undesirable. However, the cloudy existence of overlapping and unique compounds in common drug databases impedes comprehensive analysis.

Acknowledgments

Research reported in this publication was supported by the Washington University Institute of Clinical and Translational Sciences grant UL1TR002345, subaward TL1TR002344, from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.

References

[1]. Yonchev D et al. (2018) Redundancy in two major compound databases. Drug Discov. Today 23, 1183–1186 [PubMed: 29559364]

[2]. Southan C et al. (2013) Comparing the chemical structure and protein content of ChEMBL, DrugBank, human metabolome database and the therapeutic target database. Mol. Inform 32, 881–897 [PubMed: 24533037]

[3]. Southan C (2018) Caveat Usor: assessing differences between major chemistry databases ChemMedChem. 13, 470–481 [PubMed: 29451740]

[4]. 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 [PubMed: 20572635]

[5]. Gaulton A et al. (2016) The ChEMBL database in 2017. Nucleic Acids Res. 45, D945–954 [PubMed: 27899562]

[6]. Kinch MS et al. (2014) An overview of FDA-approved new molecular entities: 1827–2013. Drug Discov. Today 19, 1033–1039 [PubMed: 24680947]

[7]. Wishart DS (2017) et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46, D1074–1082

[8]. Ursu O et al. (2016) DrugCentral: online drug compendium. Nucleic Acids Res. 45, 932–939

[9]. Kim S et al. (2015) PubChem substance and compound databases. Nucleic Acids Res. 44, D1202–1213 [PubMed: 26400175]

[10]. Brown AS and Patel CJ (2017) A standard database for drug repositioning. Sci. Data 4, 170029 [PubMed: 28291243]

[11]. Siramshetty VB et al. (2017) SuperDRUG2: a one stop resource for approved/marketed drugs. Nucleic Acids Res. 46, D1137–1143

[12]. Gillespie LD et al. (2009) WITHDRAWN: interventions for preventing falls in elderly people. Cochrane Database Syst. Rev 2, CD000340

[13]. Wishart DS (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 34, D668–672 [PubMed: 16381955]

[14]. Southan C et al. (2016) The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res. 44, D1054–1068 [PubMed: 26464438]

[15]. Hanwell MD et al. (2012) Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J. Cheminform 4, 17 [PubMed: 22889332]

[16]. Inglese J et al. (2006) Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. Proc. Natl. Acad. Sci. U. S. A 103, 11473–11478 [PubMed: 16864780]

[17]. Koutsoukas A et al. (2013) *In silico* target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window. J. Chem. Inf. Model 53, 1957–1966 [PubMed: 23829430]

[18]. Mysinger MM et al. (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J. Med. Chem 55, 6582–6594 [PubMed: 22716043]

[19]. Rask-Andersen M et al. (2011) Trends in the exploitation of novel drug targets. Nat. Rev. Drug Discov 10, 579–590 [PubMed: 21804595]

[20]. Liu X et al. (2010) PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res. 38, W609–614 [PubMed: 20430828]

[21]. Gao Z et al. (2008) PDTD: a web-accessible protein database for drug target identification. BMC Bioinformatics 9, 104 [PubMed: 18282303]

[22]. Relling MV and Evans WE (2015) Pharmacogenomics in the clinic. Nature. 526, 343–350 [PubMed: 26469045]

[23]. Katz L and Baltz RH (2016) Natural product discovery: past, present, and future. J. Ind. Microbiol. Biotechnol 43, 155–176 [PubMed: 26739136]

[24]. Doak BC et al. (2014) Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. Chemistry and Biology. 21, 1115–1142 [PubMed: 25237858]

[25]. Dallmann R et al. (2016) Dosing-time makes the poison: circadian regulation and pharmacotherapy. Trends Mol. Med 22, 430–445 [PubMed: 27066876]

[26]. Sala L et al. (2017) Integrating cardiomyocytes from human pluripotent stem cells in safety pharmacology: has the time come? Br. J. Pharmacol doi: 10.1111/bph.13577

[27]. Ivanov DP et al. (2016) Separating chemotherapy-related developmental neurotoxicity from cytotoxicity in monolayer and neurosphere cultures of human fetal brain cells. Toxicol. Vitr 37, 88–96

[28]. Lagarde N et al. (2018) Online structure-based screening of purchasable approved drugs and natural compounds: retrospective examples of drug repositioning on cancer targets. Oncotarget 9, 32346–32361 [PubMed: 30190791]

[29]. Wang C et al. (2016) Current strategies and applications for precision drug design. Front. Pharmacol 9, 787

[30]. Santos R et al. (2016) A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov 16, 19–34 [PubMed: 27910877]

[31]. Nguyen D-T. et al. (2017) Pharos: collating protein information to shed light on the druggable genome. Nucleic Acids Res. 45, D995–1002 [PubMed: 27903890]

[32]. March-Vila E et al. (2017) On the integration of *in silico* drug design methods for drug repurposing. Front. Pharmacol 8, 298 [PubMed: 28588497]

[33]. Himmelstein DS et al. (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife 6, 2017

[34]. Bisgin H et al. (2014) A phenome-guided drug repositioning through a latent variable model. BMC Bioinformatics 15, 267 [PubMed: 25103881]

[35]. Croft D et al. (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 39, D691–697 [PubMed: 21067998]

[36]. World Health Organization (2015) ATC/DDD Index 2015. Available at: https://www.whocc.no/atc_ddd_index/

[37]. Kanehisa M et al. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–361 [PubMed: 27899662]

[38]. Yang H et al. (2016) Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. Nucleic Acids Res. 44, D1069–1074 [PubMed: 26578601]

Highlights

- The content of biopharmaceutical databases is dictated by the intended audience
- Databases share sources in their construction and often cite each other as sources
- No single database captures comprehensive information
- Each database has at least a small number of distinct clinical drug compounds

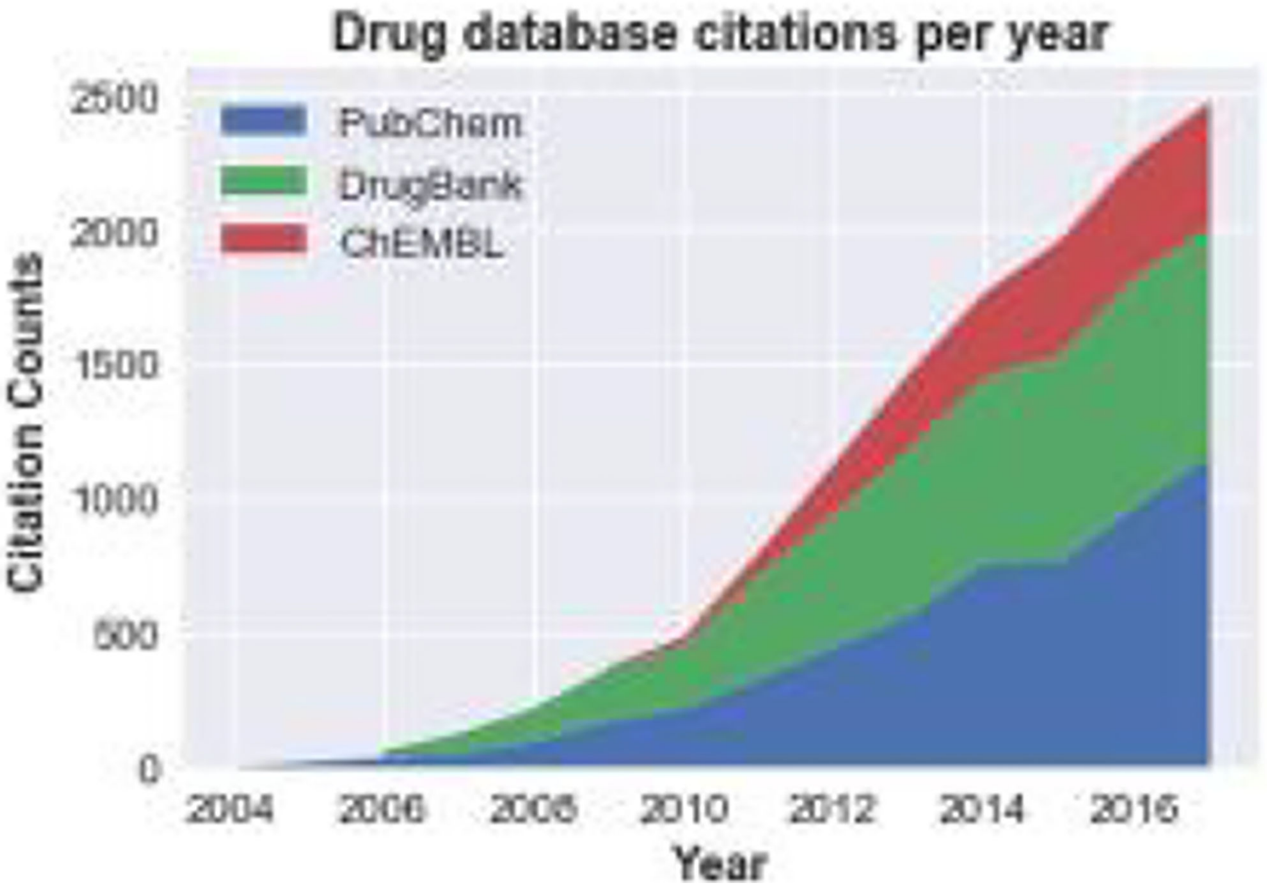
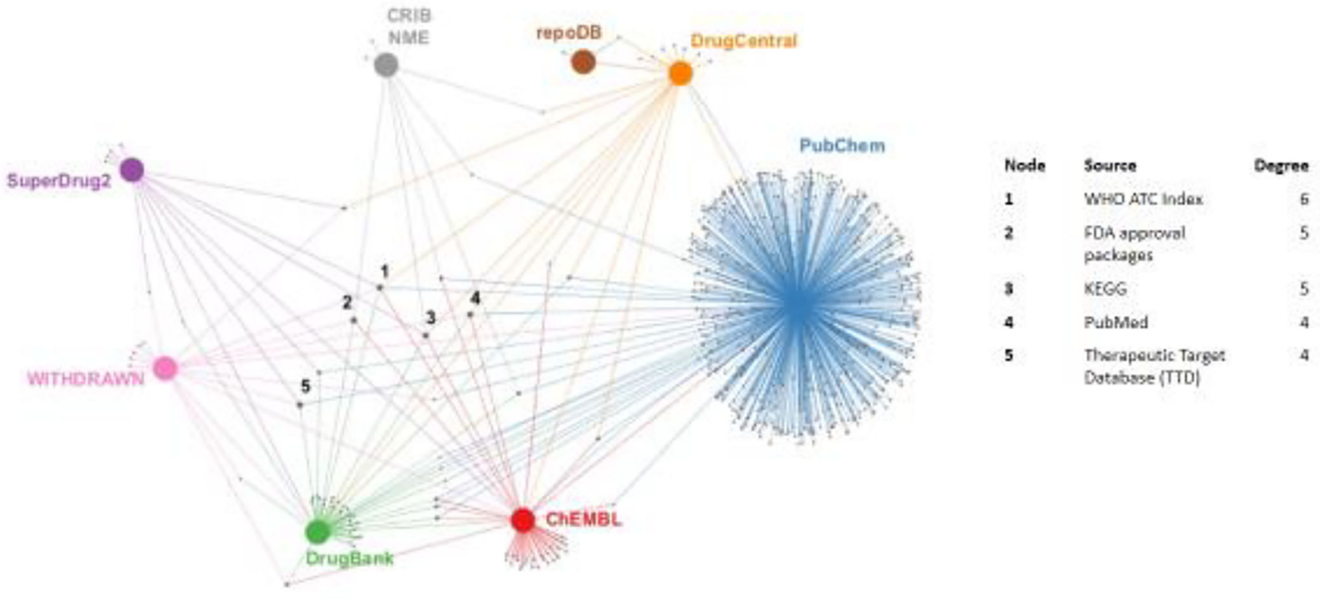


Figure 1.

Increasing use of public and freely available drug databases for research. The total number of record counts is plotted over time since 2004. The total record counts were compiled by searching the databases selected for this review as search terms in Scopus.



Node	Source	Degree
1	WHO ATC Index	6
2	FDA approval packages	5
3	KEGG	5
4	PubMed	4
5	Therapeutic Target Database (TTD)	4

Figure 2.

Network graph displaying the interconnectivity of sources used to construct each database. Each node represents a unique data source. The nodes with the highest degrees represent data sources shared between several databases and are labeled 1–5 in the graph.

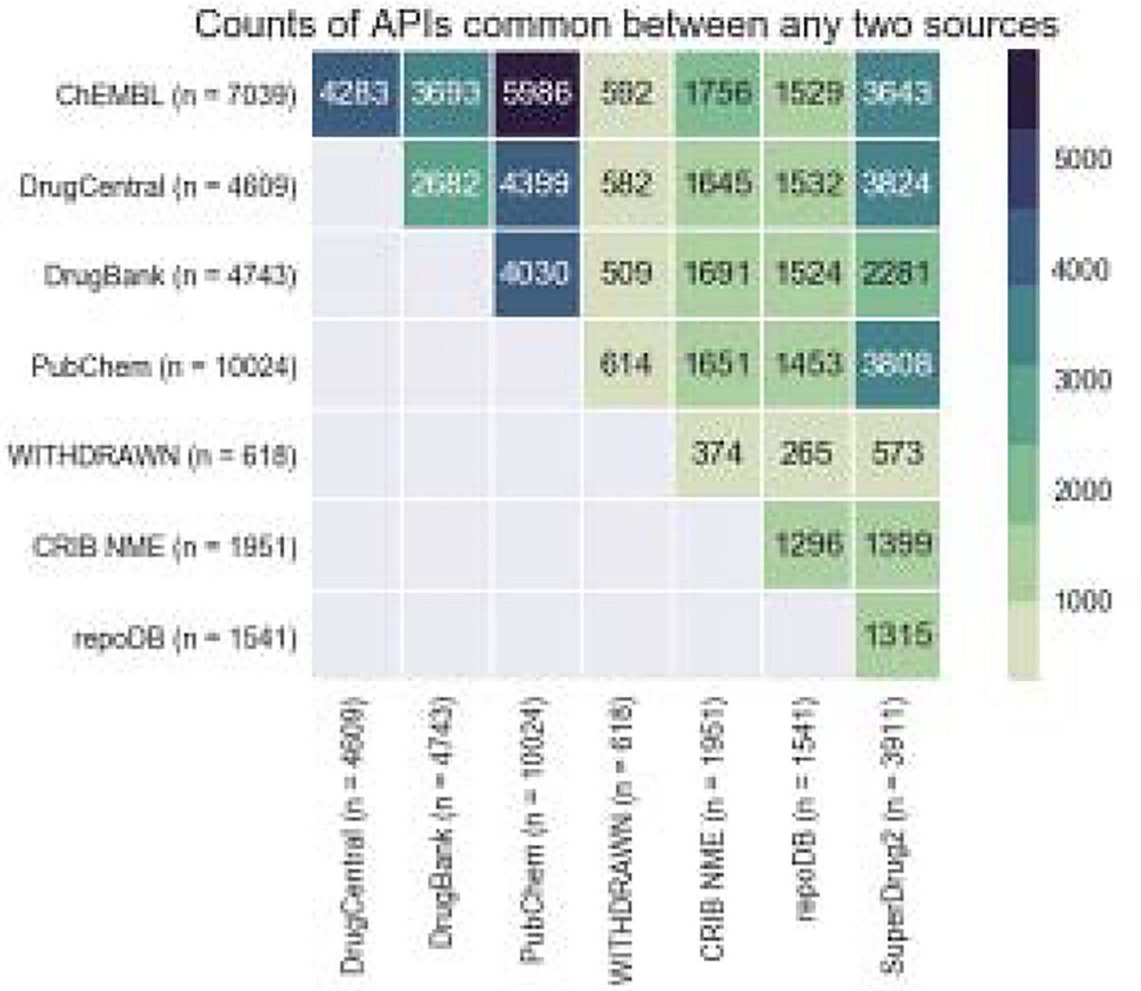


Figure 3.

Heatmap displaying the overlap in clinical-stage active pharmaceutical ingredients (APIs) between any two databases. The coloring and number displayed between any two databases is total number of shared clinical-stage drugs.

Table 1.

Brief description, size and intended audience of select drug databases

Database	Emphasis	Access	Compound Records	Clinical-Stage Compounds
PubChem	Chemical entities and their bioactivities	https://pubchem.ncbi.nlm.nih.gov	96398953	10028
ChEMBL	Bioactivity for drug discovery	https://www.ebi.ac.uk/chembl	2275906	7045
DrugBank	*in silico* d rug discovery and exploration	https://www.drugbank.ca	10562	4743
DrugCentral	Active pharmaceutical ingredients approved by FDA and other agencies	http://drugcentral.org	4608	4608
SuperDrug2	Marketed drugs	http://cheminfo.charite.de/superdrug2	3910	3910
CRIB NME	FDA approved molecular entities and biopharmaceutical organizations	http://cribdb.wustl.edu	1950	1950
repoDB	Drug repurposing	http://apps.chiragjpgroup.org/repoDB	1541	1541
WITHDRAWN	Withdrawn or discontinued drugs	http://cheminfo.charite.de/withdrawn	618	618

Table 2.

Overview of database content broken down by drug development stage

Database	Approved	Withdrawn	Clinical stage	Preclinical	Bioactive
PubChem	X	X	X	X	X
ChEMBL	X	X	X	X	X
DrugBank	X	X	X	X	
CRIB NME	X	X			
repoDB	X		X		
WITHDRAWN		X			
DrugCentral	X				
SuperDrug2	X				

Approved: approved by regulatory agencies in the USA, Canada, Europe, Japan, Korea or China. **Withdrawn**: previously approved drugs withdrawn from the market for any reason. **Clinical-stage investigational**: drugs with evidence of first-in-human testing (i.e., in clinical trials). **Preclinical**: drug candidates with preclinical testing. **Bioactivity**: compounds showing activity in biological assays.

Table 3.

Geographic adoption of drug databases and total number of associated citation records extracted from Scopus

Database	Origin	Geographic Adoption	Citations
PubChem	U.S.	United States (23.2%), India (9.4%), China (7.1%)	6122
DrugBank	Canada	United States (19.7%), China (10.8%), India (10.8 %)	5675
ChEMBL	U.K.	United States (18.7%), United Kingdom (11.7%), Germany (10.3%)	2424
CRIB NME	U.S.	United States (48.8%), France (5.8%), India (5.8%)	67
WITHDRAWN	Germany	United Kingdom (19.3%), United States (19.3%), China (9.6%)	22
DrugCentral	U.S.	United States (28.9%), France (13.15%), United Kingdom (10.5%)	20
repoDB	U.S.	Turkey (16.6%), United States (16.6%), Egypt (8.3%)	8
SuperDrug2	Germany	United Kingdom (100%)	1