Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D945–D954***
doi: 10.1093/nar/gkw1074

The ChEMBL database in 2017

Anna Gaulton[1], Anne Hersey[1], Michał Nowotka[1], A. Patrícia Bento[1,2], Jon Chambers[1], David Mendez[1], Prudence Mutowo[1], Francis Atkinson[1], Louisa J. Bellis[1], Elena Cibrián-Uhalte[1], Mark Davies[1], Nathan Dedman[1], Anneli Karlsson[1], María Paula Magariños[1,2], John P. Overington[1], George Papadatos[1], Ines Smit[1] and Andrew R. Leach[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK and [2]Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received September 23, 2016; Revised October 21, 2016; Editorial Decision October 24, 2016; Accepted October 30, 2016

ABSTRACT

ChEMBL is an open large-scale bioactivity database (https://www.ebi.ac.uk/chembl), previously described in the 2012 and 2014 Nucleic Acids Research Database Issues. Since then, alongside the continued extraction of data from the medicinal chemistry literature, new sources of bioactivity data have also been added to the database. These include: deposited data sets from neglected disease screening; crop protection data; drug metabolism and disposition data and bioactivity data from patents. A number of improvements and new features have also been incorporated. These include the annotation of assays and targets using ontologies, the inclusion of targets and indications for clinical candidates, addition of metabolic pathways for drugs and calculation of structural alerts. The ChEMBL data can be accessed via a web-interface, RDF distribution, data downloads and RESTful web-services.

INTRODUCTION

Since its inception a major component of ChEMBL's content has been bioactivity data regularly extracted from the medicinal chemistry literature (1,2). Among many other applications such data enables researchers to identify tool compounds for potential therapeutic targets, to probe the available SAR data for a target, investigate phenotypic data associated with similar compounds and to identify potential off-target effects of specific chemotypes. In order to provide a more complete perspective, based in part on user feedback, the scope and data types within ChEMBL have gradually expanded, with some major new areas included in recent releases: compounds in clinical development, data from patents, direct depositions for neglected diseases and agrochemical data.

Drug discovery remains a costly process with a high failure rate (3–6). To provide a more complete picture across the drug discovery and development process, and to help researchers better understand what makes a successful medicine, we have extended the ChEMBL data model to include, for the first time, data typically generated in the pre-clinical and clinical phases of drug discovery, specifically drug metabolism and disposition data. Another common approach to understanding pharmaceutical attrition is to learn from successful drugs and failed drug candidates (7–9). We have therefore extended our set of drug-target annotations to include those for clinical candidates and have also mapped these chemical entities to their therapeutic indications.

At the end of 2013, EMBL-EBI took over the operation, development and support of the SureChem patent system (now called SureChEMBL (10)) from Digital Science Ltd. Access to this resource has highlighted the potential value to scientists of bioactivity data not yet published in the scientific literature. However, the current SureChEMBL system only extracts compound structures from the patents and not associated bioactivity data. As a first step to address this opportunity we have worked with BindingDB to incorporate the BindingDB patent data into ChEMBL (11).

Neglected disease research continues to be a field of drug discovery conducted largely (though not exclusively) by not-for-profit organisations that aim to expedite research by encouraging sharing of experimental data with the commu-

[*]To whom correspondence should be addressed. Tel: +44 1223 494333; Fax: +44 1223 494468; Email: arl@ebi.ac.uk
Present addresses:
Louisa J. Bellis, Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK.
John P. Overington, Mark Davies and Anneli Karlsson, BenevolentAI, 40 Churchway, London NW1 1LW, UK.
Nathan Dedman, Local Measure, 87 Leonard St, London EC2A 4QS, UK.
George Papadatos, GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, Herts SG1 2NY, UK.

nity (12–15). Depositions of this type of data into ChEMBL have continued to increase since the original malaria depositions in 2010.

Whilst the pharmaceutical and drug discovery community continues to be the major user and consumer of ChEMBL data, other life sciences communities also work with similar types of data. The agrochemical industry is one such community where specific efforts have been made to widen the coverage of data relevant to the discovery and development of herbicides, pesticides and fungicides (16).

In the next sections, we describe the new data types now integrated into the ChEMBL database, the annotations we have undertaken to enable structured organisation and access to the data and how the new data and annotations can be viewed via the web interface.

DATA CONTENT

Current data content

ChEMBL's content continues to grow; release 22 of the database contains information extracted from more than 65 000 publications, together with 50 deposited data sets, and data drawn from other databases (Table 1). In total, there are >1.6 million distinct compound structures represented in the database, with 14 million activity values from >1.2 million assays. These assays are mapped to ∼11 000 targets, including 9052 proteins (of which 4255 are human).

Deposited data sets

ChEMBL continues to receive data sets from both not-for-profit and commercial organisations that wish to share data with the scientific community. These deposited data sets contain many novel chemical structures and associated bioactivity data. A good example is a library of small molecular weight (∼320 Da average) natural product-like compounds created at the University of Dundee with funding from the Gates Foundation. This library is being screened in a number of neglected disease assays. To date, information about the compound structures and their activity in cytotoxicity assays is available in ChEMBL; further assay data will be deposited as the screening is completed. Another example is the Malaria Box Compound Set (http://www.mmv.org/research-development/open-access-malaria-box), a set of 400 compounds with antimalarial activity that was made available by the Medicines for Malaria Venture (MMV) for research groups to request and screen (15). The latest ChEMBL releases include results from an additional 19 laboratories that have screened and deposited their results on these compounds. The Malaria Box was recently superseded by the Pathogen Box (http://www.pathogenbox.org), which is a new set of compounds with activity against a broader range of pathogens, also available from MMV for screening. These compounds have also been deposited in the ChEMBL database and experimental data will also be added as it becomes available. Four new sets of screening results from the Drugs for Neglected Diseases Initiative (DNDi, http://www.dndi.org) have also been incorporated into recent ChEMBL releases, providing data on compounds tested as potential drugs for neglected diseases such as Leishmaniasis and Chagas disease.

Beyond the area of neglected diseases, the University of Vienna and Roche have deposited supplementary data associated with publications already in ChEMBL (17,18). This is complementary to the similar sets already deposited by GlaxoSmithKline and we encourage similar depositions from other authors. AstraZeneca have taken a different approach to direct data deposition. They identified compounds already in ChEMBL and then provided data on these compounds from a variety of *in vitro* ADME and physicochemical screens including protein binding, microsome and hepatocyte clearance, solubility, pK_a and lipophilicity. It is important to note that for all such deposited data sets, ChEMBL provides a DOI so that the data can be cited in subsequent publications. For example, the DOI for the AstraZeneca deposited data is 10.6019/CHEMBL3301361 and this resolves to the ChEMBL Document Report Card for the complete data set.

Deposited data sets can be identified through the ChEMBL interface by selecting the relevant entry in the 'Activity Source Filter' (located to the right of the keyword search bar) and then performing a keyword search with a wildcard (*) against 'Documents' or 'Assays'. This will return all documents or assays associated with that source or depositor, from which point further information regarding compounds, targets and activity measurements can be navigated.

Crop protection data

In order to broaden the utility of ChEMBL in crop protection research, a data set of >40 000 compounds and 245 000 activity data points has been extracted from crop protection-related publications and added to ChEMBL (16). This data set significantly increased the content of pesticide, herbicide and insecticide assays in the database. The ChEMBL taxonomy tree browser has been extended to allow easier retrieval of this data (https://www.ebi.ac.uk/chembl/target/browser). In addition, known pesticides already in ChEMBL were assigned a mechanism of action classification, following the Fungicide Resistance Action Committee (FRAC: http://www.frac.info/publications), Herbicide Resistance Action Committee (HRAC: http://hrac.tsstaging.com/tools/classification-lookup) or Insecticide Resistance Action Committee (IRAC: http://www.irac-online.org/documents/moa-brochure/?ext=pdf) systems.

Patent data exchange

BindingDB is a database of affinity measurements for small molecules binding to protein targets (11). BindingDB curates data from journals complementary to ChEMBL and also incorporates relevant ChEMBL data. More recently, the BindingDB team have abstracted binding affinity data from granted US patents. We have worked with BindingDB to establish a data exchange mechanism for patent data and now include in ChEMBL the patent data extracted by the BindingDB team (ChEMBL source = 37). Identifiers for these patent documents are now included in the DOCS table as PATENT_ID. Since target information is al-

Table 1. Data sources included in the ChEMBL release 22

Short name	Source	No. compounds	No. assays	No. activities
LITERATURE	Scientific Literature	967 242	963 186	5 635 084
PUBCHEM_BIOASSAY	PubChem BioAssays	489 575	2937	7 559 601
GATES_LIBRARY	Gates Library compound collection	68 490	2	69 444
BINDINGDB	BindingDB Database	68 149	1317	99 061
GSK_TCMDC	GSK Malaria Screening	13 467	6	81 198
ST_JUDE_LEISH	St Jude Leishmania Screening	13 422	6	42 105
USP/USAN	USP Dictionary of USAN and International Drug Names	11 356	0	0
DNDI	Drugs for Neglected Diseases Initiative (DNDi)	7053	233	14 452
ASTRAZENECA	AstraZeneca Deposited Data	5799	15	11 687
NOVARTIS	Novartis Malaria Screening	5614	6	27 888
ORANGE_BOOK	Orange Book	2016	0	0
SUPPLEMENTARY	Deposited Supplementary Bioactivity Data	1786	13	4817
CANDIDATES	Clinical Candidates	1633	0	0
ST_JUDE	St Jude Malaria Screening	1524	16	5456
TP_TRANSPORTER	TP-search Transporter Database	1434	3592	6765
DRUGMATRIX	DrugMatrix	930	113 678	350 929
METABOLISM	Curated Drug Metabolism Pathways	828	0	0
GSK_TB	GSK Tuberculosis Screening	826	15	1814
WHO_TDR	WHO-TDR Malaria Screening	740	16	5853
GSK_TCAKS	GSK Kinetoplastid Screening	592	13	7235
MMV_MBOX	MMV Malaria Box	400	138	45 158
MMV_PBOX	MMV Pathogen Box	400	0	0
ATLAS	Gene Expression Atlas Compounds	398	0	0
DRUGS	Manually Added Drugs	378	0	0
GSK_PKIS	GSK Published Kinase Inhibitor Set	366	456	169 451
OSM	Open Source Malaria Screening	211	22	344
WITHDRAWN	Withdrawn Drugs	192	0	0
TG_GATES	Open TG-GATEs	160	158 199	158 199
SANGER	Sanger Institute Genomics of Drug Sensitivity in Cancer	137	714	73 169
FDA_APPROVAL	FDA Approval Packages	43	1386	1387
HARVARD	Harvard Malaria Screening	37	4	111

ready carefully manually curated in BindingDB, this information is retained in ChEMBL and simply mapped to the equivalent ChEMBL target. The data is taken from 1,015 granted US patents published between 2013 and 2015 and currently comprises 99 061 bioactivities on 68 149 distinct compounds binding to around 600 distinct targets. Of particular interest to drug discovery scientists is the fact that data often appears in patents earlier than in the traditional medicinal chemistry literature. This patent set contains data on 50 targets for which there was previously no data in ChEMBL and which may therefore represent novel targets of therapeutic interest.

NEW FUNCTIONALITY

Richer assay and target annotation

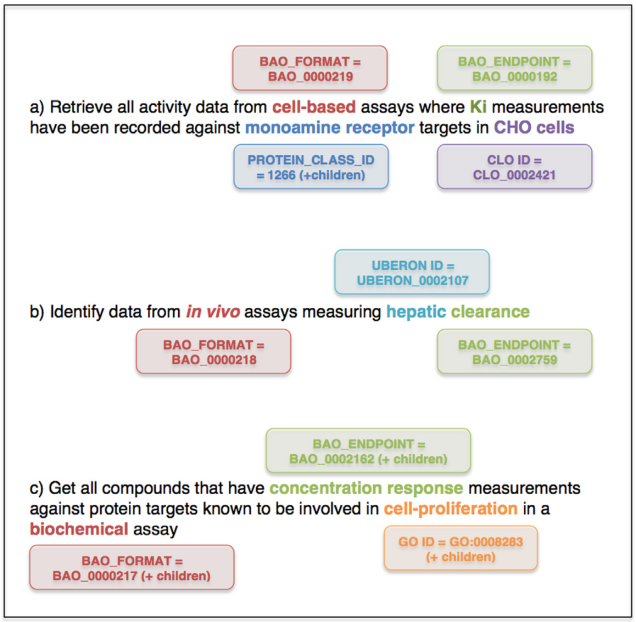
Typical entry points to ChEMBL have predominantly been compound-based or target-based searches. However, more than half of the activity data points in ChEMBL come from functional or phenotypic assays that cannot be assigned a molecular target. Since phenotypic screening is once more becoming commonplace in drug discovery (19), making this wealth of data more accessible is a priority. To this end, we have applied a number of ontologies to the ChEMBL assay and activity data, allowing them to be searched and filtered by cell-line, tissue or assay format, for example.

The BioAssay Ontology (BAO) (20,21) was chosen as a means of annotating ChEMBL assays for a number of reasons: this ontology has been developed specifically for small molecule screening data and so provides good coverage of the ChEMBL data; it has also been adopted by a number of other bioassay data providers and members of the drug discovery community, allowing for good data interoperability. ChEMBL standard activity types were manually mapped to corresponding BAO result terms (stored in the ACTIVITIES table as BAO_ENDPOINT). The resulting mappings cover 91% of the activity data points in ChEMBL. The remainder are mainly diverse phenotypic endpoints that are not covered by BAO (e.g. 'Tissue Severity Score', 'Anticonvulsant activity', 'Paw swelling', 'Relative uterus weight') or imprecise terms (e.g. 'Ratio', 'Selectivity', 'Response') that require further resolution. Similarly, ChEMBL standard activity units were mapped to Units Ontology (UO) terms (22) (which are also a component of BAO) and stored in the ACTIVITIES table as UO_UNITS. Where available, units were also mapped to terms from the Quantities, Units, Dimensions and Types ontology (http://www.qudt.org). These are stored in the ACTIVITIES table as QUDT_UNITS. The current mapping to Units Ontology covers 87% of ChEMBL activity data points (the remainder largely being complex units e.g. 'ng.h.ml^{-1}', 'ml.min^{-1}.g^{-1}' that are not covered by UO).

ChEMBL assays have also been annotated with BAO assay format terms, allowing users to distinguish biochemical, cell-based, tissue-based or organism-based assays. An automated, rule-based approach was used to classify the assay format for historical assays, based on information in assay descriptions and target assignments. In order to minimise false assignments, any assays where the format could not be determined unambiguously were left unclassified, pending further curation (11% of assays). The assay format is stored in the ASSAYS table as BAO_FORMAT. Complementary to this, we have also introduced annotation of the cells and tissues used in assays (stored as CELL_ID, ASSAY_CELL_TYPE, TISSUE_ID and ASSAY_TISSUE in the ASSAYS table). A ChEMBL CELL_DICTIONARY and TISSUE_DICTIONARY have been introduced and cell and tissue Report Card pages and search functionality created. It should be noted that the aim here is not to create new cell-line/tissue ontologies – entries in the ChEMBL dictionaries are imported from established vocabularies and ontologies wherever possible (e.g. Cell Line Ontology (23), Experimental Factor Ontology (EFO) (24), Cellosaurus (http://web.expasy.org/cellosaurus/) and LINCS cell dictionary (25) for cells; and Uberon (26), EFO, Brenda Tissue Ontology (27) and CALOHA for tissues (ftp://ftp.nextprot.org/pub/current_release/controlled_vocabularies/caloha.obo)) and mappings to these ontologies are provided.

Methods for browsing and retrieving protein target data in ChEMBL have also been enhanced. Improvements have been made to the existing protein family classification for ion channels and transporters to more closely align the ChEMBL system with other resources (the IUPHAR/BPS Guide To PHARMACOLOGY (28) and TCDB (29)) and new classes have been introduced for epigenetic regulators (following the ChromoHub database (30)). In order to allow browsing of ChEMBL targets by Gene Ontology terms, a new GO Slim has been created (31). This is a subset of GO terms that are enriched in ChEMBL targets and allows the implementation of a simplified Gene Ontology Tree for browsing (on the 'Browse Targets' tab).

These new features, particularly when combined, enable more sophisticated queries of the ChEMBL data to be performed, as illustrated in Figure 1.



a) Retrieve all activity data from cell-based assays where Ki measurements have been recorded against monoamine receptor targets in CHO cells

b) Identify data from *in vivo* assays measuring hepatic clearance

c) Get all compounds that have concentration response measurements against protein targets known to be involved in cell-proliferation in a biochemical assay

Figure 1. Examples of more complex queries that could be performed (e.g. using web services) by combining BioAssay Ontology, protein family and GO classifications.

Drug indications and clinical candidates

In previous releases of ChEMBL we annotated properties and efficacy targets for FDA approved drugs. We are now also incorporating drug indication information into ChEMBL to facilitate use-cases such as target validation (the existence of a drug that acts through a particular target and is prescribed for a given indication may be taken as evidence validating the relevance of that target for the indication). Drug indication data is collected from a variety of sources including DailyMed package inserts (https://dailymed.nlm.nih.gov), Anatomical Therapeutic Chemical (ATC) classification (http://www.whocc.no/atc_ddd_index/) (32) and ClinicalTrials.gov (https://clinicaltrials.gov) (33). Since these resources mainly provide free text information, a combination of text-mining (using BeCAS (34)), automated mapping and manual curation/validation is used to identify the indication terms and assign the corresponding disease terms in the Medical Subject Headings (MeSH) vocabulary (https://www.nlm.nih.gov/mesh/) and EFO (24). Two new tables have been added to the database schema to capture this information: DRUG_INDICATION and INDICATION_REFS. A new tab has been added to the ChEMBL interface to display the data—'Browse Drug Indications'. This view contains a list of all approved drugs and clinical candidates for which indication information is available, and displays the MeSH and EFO terms for the indications and the highest development phase at which each indication has been investigated. The view can be searched and filtered by typing keywords in the 'Search' box (e.g. 'diabetes' would filter the display to only those drugs that have an indication containing the word diabetes). More specific searches can be performed by entering a MeSH or EFO ID in the search box (e.g. 'EFO:0001360' would return drugs indicated for 'type II diabetes mellitus'). The whole or filtered data set can also be exported using the 'Downloads' option at the top right-hand side of the view. A summary of the indication data is also displayed on Compound Report Card pages for drugs.

Furthermore, we are now extending the annotation of efficacy targets and indications to cover drug candidates in clinical development. Mechanism of action and target information for 1,023 clinical candidates in phase I-III development (mainly targeting GPCR, kinase, nuclear hormone receptor and ion channel targets, which are the focus of the NIH-funded Illuminating the Druggable Genome project: https://commonfund.nih.gov/idg/) has already been curated and included in the database (ChEMBL source = 8), and targets for further candidates will be included in future releases. Indications for candidates have also been included using information in ClinicalTrials.gov. For each candidate-indication association, the highest development phase at which that indication has been investigated is also recorded. Currently 1,451 clinical candidates have associated indication information. Clinical candidates are now included in the 'Browse Drugs' tab on the ChEMBL interface and their mechanism of action and indication information (where annotated) can be viewed in the 'Browse Drug Targets' and 'Browse Drug Indications' tabs, respectively.

Finally, we have also added information regarding previously approved drugs that have been withdrawn for toxicity or efficacy reasons. Information regarding withdrawn drugs was collated from several sources: the FDA (http://www.fda.gov) and EMA (http://www.ema.europa.eu/ema/), the WITHDRAWN database (35), the US Electronic Code of Federal Regulations (http://www.ecfr.gov/cgi-bin/retrieveECFR?gp=2&SID=915cc9ab8176f1d1a2a355acf064ffe3&h=L&mc=true&n=sp21.4.216.b&r=SUBPART&ty=HTML#se21.4.216_124), Federal Register (https://www.gpo.gov/fdsys/pkg/FR-2014-07-02/pdf/2014-15371.pdf) and several review articles (36–38). Where available, the year of withdrawal, the applicable countries/areas and the reasons for the withdrawal are captured. This information has been added to the MOLECULE_DICTIONARY database table and is displayed on the 'Browse Drugs' tab and also on Compound Report Card pages (see Figure 2). The 'Molecule Features' icons have also been updated to include a new availability type icon. This icon now has four options: withdrawn, discontinued, prescription-only, over-the-counter. In cases where the drug has been withdrawn in one country but is still available in others, the icon will retain the prescription-only or over-the-counter status, rather than withdrawn.

Drug metabolism, toxicity and pharmacokinetic data

In vivo data can be extremely valuable but is also invariably complex, with experimental parameters typically being measured at varying doses, time points, routes of administration, in the presence of interacting compounds etc. In order to organise this information in a more structured way and to enable easier access of this information by ChEMBL users an ASSAY_PARAMETERS table has been included in the ChEMBL schema to record this information. This has been done retrospectively for all the ChEMBL assays and resulted in addition of 3.6 million parameter mappings for approximately one third of all the ChEMBL assays.

Additionally, new information has been added to ChEMBL on drug metabolism, pharmacokinetics and toxicology particularly for pre-clinical and clinical drug candidates. The data has been manually extracted from three new data sources and will be extended in future ChEMBL releases. Firstly ~260 articles covering the years 2011–2013 were extracted from the Journal of Drug Metabolism and Disposition. This resulted in data on ~2200 compounds tested in nearly 15 000 assays (included in ChEMBL source = 1). The second data set included are the FDA drug approval packages (39). These are summaries of the information submitted to the FDA when the drug was approved for use and are available for FDA drugs approved after 1997. To our knowledge this valuable data is not captured in a structured format elsewhere (ChEMBL source = 28).

Thirdly, given the need to understand the nature of drug metabolites and whether they themselves are therapeutically active or constitute reactive species that bind DNA or proteins resulting in toxicity (40,41), a project has been initiated to identify the metabolic pathways of approved drugs. This was achieved using a variety of data sources (ChEMBL sources 1, 28 and 31). The metabolic schemes are described in a data model that maintains the relationships between the various molecular entities (e.g. metabolite A may be formed directly from drug D, whereas metabolite B may result from the degradation of metabolite A). Where metabolising enzymes, species and tissues are available in the original publication this information is recorded in ChEMBL. In instances where the metabolite structure is known it is recorded as a chemical structure. If the exact structure is unknown the reaction is still recorded but with an undefined structure. The metabolism data is recorded in two new tables METABOLISM and METABOLISM_REFS. The metabolite pathway is shown on the ChEMBL interface as an interactive image with links to the data on the metabolites, the metabolising enzymes and the document sources for the information. An example for Simvastatin (https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL1064) is shown in Figure 3.

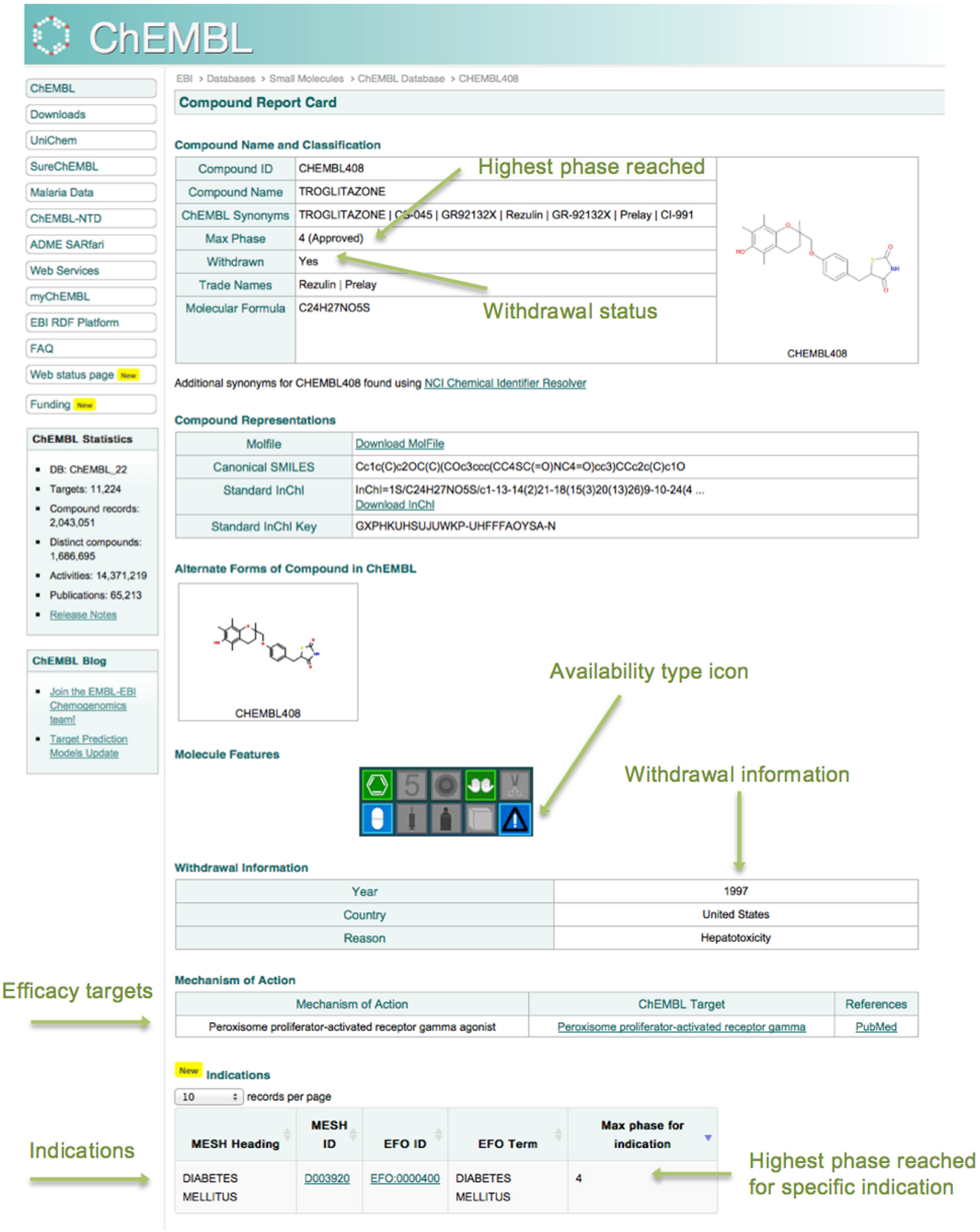
Compound structural alerts

In order to aid users in the selection of compounds, for example to create screening sets, we have compiled sets of publicly-available structural alerts as SMARTS (Pfizer LINT filters (42), Glaxo Wellcome Hard Filters (43), Bristol-Myers Squibb HTS Deck Filters (44), NIH MLSMR Excluded Functionality Filters (https://mlsmr.evotec.com/MLSMR_HomePage/pdf/MLSMR_Excluded_Functionality_Filters_200605121510.pdf), University of Dundee NTD Screening Library Filters (45) and Pan Assay Interference Compounds (PAINS) Filters (46), filters derived by Inpharmatica Ltd, set of alerts currently used in SureChEMBL https://www.surechembl.org/knowledgebase/169485).
All filters have been run against all ChEMBL compounds for which structures are available and the results included in the database in the COMPOUND_STRUCTURAL_ALERTS, STRUCTURAL_ALERTS and STRUCTURAL_ALERT_SETS tables. These filters can be used to identify compounds that may be problematic in a drug-discovery setting and so guide the selection of compounds and/or help in the interpretation of results. The filters typically represent substructures corresponding to chemically reactive functional groups, those that are associated with toxicity, interfere with certain assay formats or bind promiscuously to targets. Alerts are implemented as reported in the original source but should be interpreted with care, depending on the use-case, and not treated as a blanket filter (e.g. ~50% of approved drugs have one or more alerts from these sets). These alerts can be viewed on the Compound Report Card via the ChEMBL interface. The alerts are organised by the alert set that a specific functionality is found in and the functional group is highlighted on a depiction of the molecule.

DATA ACCESS

The ChEMBL interface

The ChEMBL database is accessible via a basic user interface at: https://www.ebi.ac.uk/chembl. This interface provides core search functionality for compounds (including



ChEMBL

- ChEMBL
- Downloads
- UniChem
- SureChEMBL
- Malaria Data
- ChEMBL-NTD
- ADME SARfari
- Web Services
- myChEMBL
- EBI RDF Platform
- FAQ
- Web status page New
- Funding New

ChEMBL Statistics

- DB: ChEMBL_22
- Targets: 11,224
- Compound records: 2,043,051
- Distinct compounds: 1,686,695
- Activities: 14,371,219
- Publications: 65,213
- Release Notes

ChEMBL Blog

- Join the EMBL-EBI Chemogenomics team!
- Target Prediction Models Update

EBI > Databases > Small Molecules > ChEMBL Database > CHEMBL408

Compound Report Card

Compound Name and Classification

Compound ID	CHEMBL408
Compound Name	TROGLITAZONE
ChEMBL Synonyms	TROGLITAZONE | CS-045 | GR92132X | Rezulin | GR-92132X | Prelay | CI-991
Max Phase	4 (Approved)
Withdrawn	Yes
Trade Names	Rezulin | Prelay
Molecular Formula	C24H27NO5S

CHEMBL408

Highest phase reached

Withdrawal status

Additional synonyms for CHEMBL408 found using NCI Chemical Identifier Resolver

Compound Representations

Molfile	Download MolFile
Canonical SMILES	Cc1c(C)c2OC(C)(COc3ccc(CC4SC(=O)NC4=O)cc3)CCc2c(C)c1O
Standard InChI	InChI=1S/C24H27NO5S/c1-13-14(2)21-18(15(3)20(13)26)9-10-24(4 ... Download InChI
Standard InChI Key	GXPHKUHSUJUWKP-UHFFFAOYSA-N

Alternate Forms of Compound in ChEMBL

CHEMBL408

Availability type icon

Molecule Features

Withdrawal information

Withdrawal Information

Year	1997
Country	United States
Reason	Hepatotoxicity

Efficacy targets

Mechanism of Action

Mechanism of Action	ChEMBL Target	References
Peroxisome proliferator-activated receptor gamma agonist	Peroxisome proliferator-activated receptor gamma	PubMed

New Indications

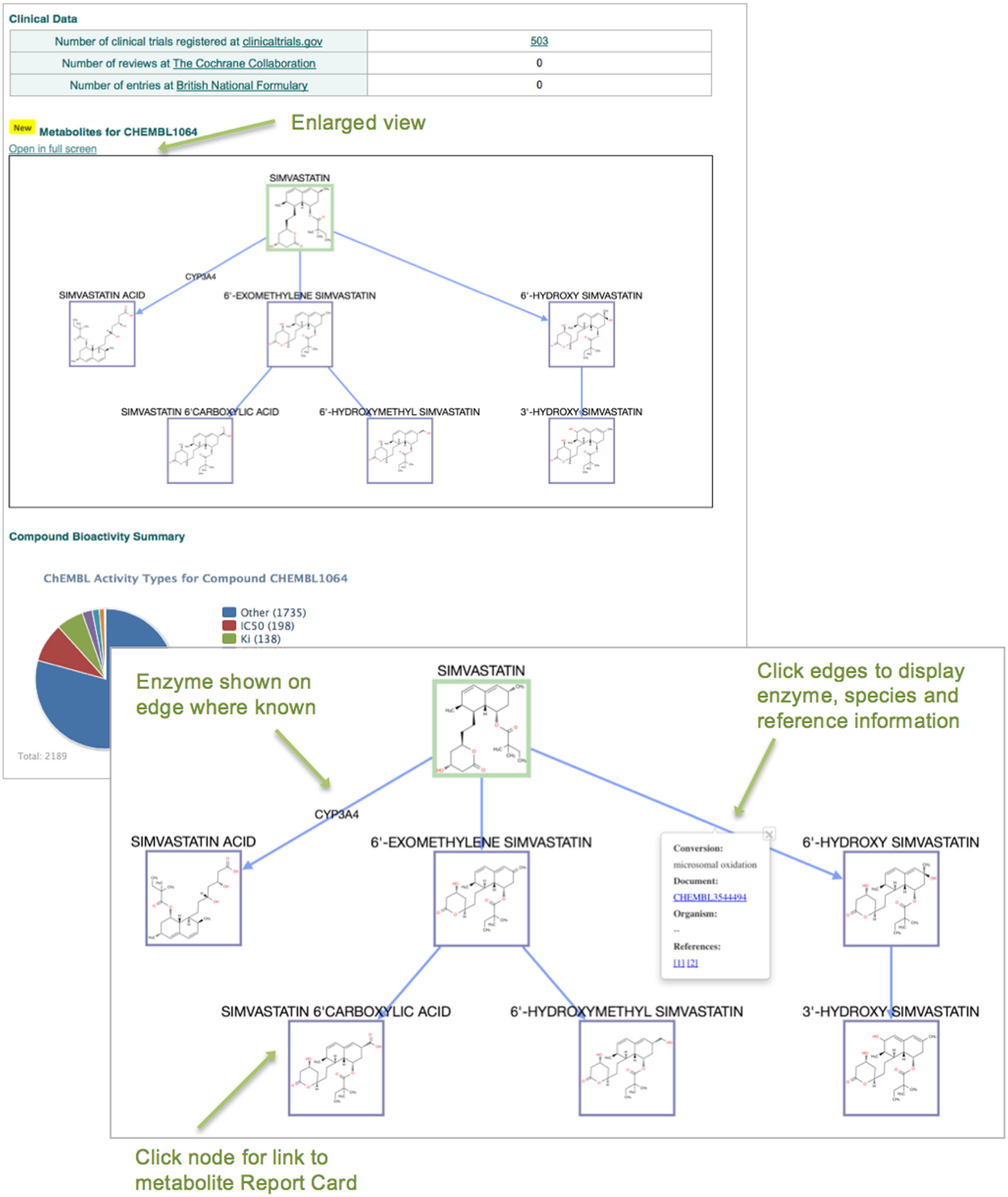
10 records per page

Indications

MESH Heading	MESH ID	EFO ID	EFO Term	Max phase for indication
DIABETES MELLITUS	D003920	EFO:0000400	DIABETES MELLITUS	4

Highest phase reached for specific indication

Figure 2. Compound Report Card for Troglitazone showing mechanism of action, indication and withdrawal information (https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL408).



Clinical Data

Number of clinical trials registered at clinicaltrials.gov	503
Number of reviews at The Cochrane Collaboration	0
Number of entries at British National Formulary	0

Conversion:
microsomal oxidation
Document:
CHEMBL3544494
Organism:
--
References:
[1] [2]

New
Metabolites for CHEMBL1064
Open in full screen
Enlarged view
SIMVASTATIN
CYP3A4
SIMVASTATIN ACID
6'-EXOMETHYLENE SIMVASTATIN
6'-HYDROXY SIMVASTATIN
SIMVASTATIN 6'CARBOXYLIC ACID
6'-HYDROXYMETHYL SIMVASTATIN
3'-HYDROXY SIMVASTATIN
Compound Bioactivity Summary
ChEMBL Activity Types for Compound CHEMBL1064
Other (1735)
IC50 (198)
Ki (138)
Total: 2189
Enzyme shown on edge where known
Click edges to display enzyme, species and reference information
Click node for link to metabolite Report Card

Figure 3. Metabolism scheme for Simvastatin (https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL1064).

substructure and similarity searching) and targets as well as keyword searching across assay, cell line and tissue information. Users can also retrieve and filter bioactivity information and browse drug and clinical candidate information (including targets and indications). More details of the user interface and its functionality can be found in previous publications (1,2).

Downloads and web-services

While the ChEMBL interface provides the functionality necessary for many simple use-cases, some users may prefer to download the database and query it locally (e.g. for use in data mining applications or to integrate with internal data). Each release of ChEMBL is available from our ftp site in a variety of formats including: Oracle, MySQL, PostgreSQL, SQLite, RDF, an SD file of compound structures and a FASTA file of the target sequences, under a Creative Commons Attribution-ShareAlike 3.0 Unported license (http://creativecommons.org/licenses/by-sa/3.0).

The myChEMBL software container is also available (47,48). It is distributed as a virtual machine disk image in many formats such as VMDK, QCOW2 and IMG that can be downloaded using FTP or Vagrant. Alternatively, myChEMBL can be used as a Docker container. The container is based on Ubuntu OS but there exists a CentOS version as well. The container consists of the ChEMBL PostgreSQL database, web services, open source cheminformatics tools (e.g. RDKit: http://www.rdkit.org, OSRA (49) and Open Babel (50)) and Jupyter Notebook tutorials (https://github.com/chembl/mychembl/tree/master/ipython_notebooks), providing a convenient environment for users to interrogate the data.

For users requiring programmatic access to ChEMBL, a comprehensive set of RESTful web services are provided (51), allowing retrieval of ChEMBL data in XML, JSON and YAML formats (see https://www.ebi.ac.uk/chembl/ws for more details and example queries). A Solr-based search functionality is also now available, permitting retrieval of compound, target and assay information via text queries. For example, the following query would retrieve all assays containing the term 'angiogenesis': https://www.ebi.ac.uk/chembl/api/data/assay/search?q=angiogenesis. An example Python web service client library is available: https://github.com/chembl/chembl_webresource_client, with usage examples covered in: https://github.com/chembl/mychembl/blob/master/ipython_notebooks/09_myChEMBL_web_services.ipynb.

Web services and myChEMBL are both open source projects and are available from the ChEMBL GitHub repository (https://github.com/chembl/), along with other useful resources to help integration of ChEMBL into software projects (https://arxiv.org/pdf/1607.00378.pdf). They are licensed under an Apache 2 license. Both projects converge to build a new version of the main ChEMBL web interface which will be built on top of web services and deployed and distributed using the myChEMBL container.

Inclusion of ChEMBL data in other resources

It is often advantageous to integrate ChEMBL data with data from other resources—either to maximise the amount of data that can be retrieved for a particular target/compound, or to provide relevant pharmacology and drug-target data in the context of other data types (e.g. pathway, expression or disease information). ChEMBL data is incorporated into a wide range of other resources including PubChem BioAssay (52), BindingDB (11), CanSAR (53), Open PHACTS (54), Open Targets (55) and the Target Central Resource Database/PHAROS (http://juniper.health.unm.edu/tcrd/, 56), so can also be accessed via these routes. However, since these other resources are different in scope, they do not all incorporate ChEMBL in full (e.g. BindingDB focuses only on binding measurements, while Open Targets incorporates data on drug–target and drug–indication linkage).

SUMMARY

ChEMBL continues to evolve and grow in order to address an ever-increasing range of use-cases, user communities and applications. The 'core' of the resource continues to be molecule–target interaction data curated from the published literature, but as outlined above and in previous reviews (1,2,51,57,58) this has been expanded both in terms of data content, annotation and infrastructure. Many of these developments have been in response to feedback and discussion from ChEMBL's growing user community; we welcome such engagement on the latest version of the resource.

ACKNOWLEDGEMENTS

Molecular Connections, Mike Gilson, Tiqing Liu, Katherine Doyle.

FUNDING

Strategic Awards from the Wellcome Trust [WT086151/Z/08/Z, WT104104/Z/14/Z]; Member States of the European Molecular Biology Laboratory (EMBL); Innovative Medicines Initiative Joint Undertaking [115191]; National Institutes of Health (NIH) Common Fund [U54CA189205]; European Union Seventh Framework Programme (FP7/2007–2013) [602156]; Syngenta; Open Targets. Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
2. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.
3. Arrowsmith,J. and Miller,P. (2013) Trial watch: phase II and phase III attrition rates 2011-2012. *Nat. Rev. Drug Discov.*, **12**, 569.
4. Bunnage,M.E. (2011) Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.*, **7**, 335–339.
5. Hay,M., Thomas,D.W., Craighead,J.L., Economides,C. and Rosenthal,J. (2014) Clinical development success rates for investigational drugs. *Nat. Biotechnol.*, **32**, 40–51.

6. Kola,I. and Landis,J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.*, **3**, 711–715.
7. Cook,D., Brown,D., Alexander,R., March,R., Morgan,P., Satterthwaite,G. and Pangalos,M.N. (2014) Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.*, **13**, 419–431.
8. Waring,M.J., Arrowsmith,J., Leach,A.R., Leeson,P.D., Mandrell,S., Owen,R.M., Pairaudeau,G., Pennie,W.D., Pickett,S.D., Wang,J. *et al.* (2015) An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.*, **14**, 475–486.
9. Morgan,P., Van Der Graaf,P.H., Arrowsmith,J., Feltner,D.E., Drummond,K.S., Wegner,C.D. and Street,S.D. (2012) Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. *Drug Discov. Today*, **17**, 419–424.
10. Papadatos,G., Davies,M., Dedman,N., Chambers,J., Gaulton,A., Siddle,J., Koks,R., Irvine,S.A., Pettersson,J., Goncharoff,N. *et al.* (2016) SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.*, **44**, D1220–D1228.
11. Gilson,M.K., Liu,T., Baitaluk,M., Nicola,G., Hwang,L. and Chong,J. (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.
12. Gamo,F.J., Sanz,L.M., Vidal,J., de Cozar,C., Alvarez,E., Lavandera,J.L., Vanderwall,D.E., Green,D.V., Kumar,V., Hasan,S. *et al.* (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature*, **465**, 305–310.
13. Ioset,J.R. and Chang,S. (2011) Drugs for Neglected Diseases initiative model of drug development for neglected diseases: current status and future challenges. *Future Med. Chem.*, **3**, 1361–1371.
14. Orti,L., Carbajo,R.J., Pieper,U., Eswar,N., Maurer,S.M., Rai,A.K., Taylor,G., Todd,M.H., Pineda-Lucena,A., Sali,A. *et al.* (2009) A kernel for open source drug discovery in tropical diseases. *PLoS Negl. Trop. Dis.*, **3**, e418.
15. Spangenberg,T., Burrows,J.N., Kowalczyk,P., McDonald,S., Wells,T.N. and Willis,P. (2013) The open access malaria box: a drug discovery catalyst for neglected diseases. *PLoS One*, **8**, e62906.
16. Gaulton,A., Kale,N., van Westen,G.J., Bellis,L.J., Bento,A.P., Davies,M., Hersey,A., Papadatos,G., Forster,M., Wege,P. *et al.* (2015) A large-scale crop protection bioassay data set. *Sci. Data*, **2**, 150032.
17. Ceccarelli,S.M., Chomienne,O., Gubler,M. and Arduini,A. (2011) Carnitine palmitoyltransferase (CPT) modulators: a medicinal chemistry perspective on 35 years of research. *J. Med. Chem.*, **54**, 3109–3152.
18. Kaiser,D., Terfloth,L., Kopp,S., Schulz,J., de Laet,R., Chiba,P., Ecker,G.F. and Gasteiger,J. (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. *J. Med. Chem.*, **50**, 1698–1702.
19. Lee,J.A., Uhlik,M.T., Moxham,C.M., Tomandl,D. and Sall,D.J. (2012) Modern phenotypic drug discovery is a viable, neoclassic pharma strategy. *J. Med. Chem.*, **55**, 4527–4538.
20. Abeyruwan,S., Vempati,U.D., Kucuk-McGinty,H., Visser,U., Koleti,A., Mir,A., Sakurai,K., Chung,C., Bittker,J.A., Clemons,P.A. *et al.* (2014) Evolving BioAssay Ontology (BAO): modularization, integration and applications. *J. Biomed. Semantics*, **5**, S5.
21. Visser,U., Abeyruwan,S., Vempati,U., Smith,R.P., Lemmon,V. and Schurer,S.C. (2011) BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, **12**, 257.
22. Gkoutos,G.V., Schofield,P.N. and Hoehndorf,R. (2012) The Units Ontology: a tool for integrating units of measurement in science. *Database (Oxford)*, **2012**, bas033.
23. Sarntivijai,S., Lin,Y., Xiang,Z., Meehan,T.F., Diehl,A.D., Vempati,U.D., Schurer,S.C., Pang,C., Malone,J., Parkinson,H. *et al.* (2014) CLO: The cell line ontology. *J. Biomed. Semantics*, **5**, 37.
24. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
25. Vempati,U.D., Chung,C., Mader,C., Koleti,A., Datar,N., Vidovic,D., Wrobel,D., Erickson,S., Muhlich,J.L., Berriz,G. *et al.* (2014) Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the library of integrated network-based cellular signatures (LINCS). *J. Biomol. Screen.*, **19**, 803–816.
26. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
27. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
28. Southan,C., Sharman,J.L., Benson,H.E., Faccenda,E., Pawson,A.J., Alexander,S.P., Buneman,O.P., Davenport,A.P., McGrath,J.C., Peters,J.A. *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–D1068.
29. Saier,M.H. Jr, Reddy,V.S., Tsu,B.V., Ahmed,M.S., Li,C. and Moreno-Hagelsieb,G. (2016) The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res.*, **44**, D372–D379.
30. Shah,M.A., Denton,E.L., Liu,L. and Schapira,M. (2014) ChromoHub V2: cancer genomics. *Bioinformatics*, **30**, 590–592.
31. Mutowo,P., Bento,A.P., Dedman,N., Gaulton,A., Hersey,A., Lomax,J. and Overington,J.P. (2016) A drug target slim: using Gene Ontology and Gene Ontology annotations to navigate protein-ligand target space in ChEMBL. *J. Biomed. Semantics*, **7**, 59.
32. WHO Collaborating Centre for Drug Statistics Methodology. (2015) *Guidelines for ATC classification and DDD assignment 2015*. WHO Collaborating Centre for Drug Statistics Methodology, Oslo.
33. Huston,M.W., Williams,R.J., Bergeris,A., Fun,J. and Tse,T. (2012) New style and new content for ClinicalTrials.gov. *NLM Tech. Bull.*, **387**, e5.
34. Nunes,T., Campos,D., Matos,S. and Oliveira,J.L. (2013) BeCAS: biomedical concept recognition services and visualization. *Bioinformatics*, **29**, 1915–1916.
35. Siramshetty,V.B., Nickel,J., Omieczynski,C., Gohlke,B.O., Drwal,M.N. and Preissner,R. (2016) WITHDRAWN–a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.*, **44**, D1080–D1086.
36. Bakke,O.M., Manocchia,M., de Abajo,F., Kaitin,K.I. and Lasagna,L. (1995) Drug safety discontinuations in the United Kingdom, the United States, and Spain from 1974 through 1993: a regulatory perspective. *Clin. Pharmacol. Ther.*, **58**, 108–117.
37. Qureshi,Z.P., Seoane-Vazquez,E., Rodriguez-Monguio,R., Stevenson,K.B. and Szeinbach,S.L. (2011) Market withdrawal of new molecular entities approved in the United States from 1980 to 2009. Pharmacoepidemiol. *Drug Saf.*, **20**, 772–777.
38. Fung,M., Thornton,A., Mybeck,K., Wu,J.H., Hornbuckle,K. and Muniz,E. (2001) Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets-1960 to 1999. *Drug Inf. J.*, **35**, 293–317.
39. Turner,E.H. (2013) How to access and process FDA drug approval packages for use in research. *BMJ*, **347**, f5992.
40. Foti,R.S. and Dalvie,D.K. (2016) Cytochrome P450 and Non-Cytochrome P450 oxidative metabolism: Contributions to the pharmacokinetics, safety, and efficacy of xenobiotics. *Drug Metab. Dispos.*, **44**, 1229–1245.
41. Kalgutkar,A.S., Gardner,I., Obach,R.S., Shaffer,C.L., Callegari,E., Henne,K.R., Mutlib,A.E., Dalvie,D.K., Lee,J.S., Nakai,Y. *et al.* (2005) A comprehensive listing of bioactivation pathways of organic functional groups. *Curr. Drug Metab.*, **6**, 161–225.
42. Blake,J.F. (2005) Identification and evaluation of molecular properties related to preclinical optimization and clinical fate. *Med. Chem.*, **1**, 649–655.
43. Hann,M., Hudson,B., Lewell,X., Lifely,R., Miller,L. and Ramsden,N. (1999) Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.*, **39**, 897–902.
44. Pearce,B.C., Sofia,M.J., Good,A.C., Drexler,D.M. and Stock,D.A. (2006) An empirical process for the design of high-throughput screening deck filters. *J. Chem. Inf. Model.*, **46**, 1060–1068.
45. Brenk,R., Schipani,A., James,D., Krasowski,A., Gilbert,I.H., Frearson,J. and Wyatt,P.G. (2008) Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem*, **3**, 435–444.
46. Baell,J.B. and Holloway,G.A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from

screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, **53**, 2719–2740.

47. Ochoa,R., Davies,M., Papadatos,G., Atkinson,F. and Overington,J.P. (2014) myChEMBL: a virtual machine implementation of open data and cheminformatics tools. *Bioinformatics*, **30**, 298–300.
48. Davies,M., Nowotka,M., Papadatos,G., Atkinson,F., van Westen,G.J., Dedman,N., Ochoa,R. and Overington,J.P. (2014) MyChEMBL: a virtual platform for distributing cheminformatics tools and open data. *Challenges*, **5**, 334–337.
49. Filippov,I.V. and Nicklaus,M.C. (2009) Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.*, **49**, 740–743.
50. O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.
51. Davies,M., Nowotka,M., Papadatos,G., Dedman,N., Gaulton,A., Atkinson,F., Bellis,L. and Overington,J.P. (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, **43**, W612–W620.
52. Wang,Y., Cheng,T., Wang,J., Gindulyte,A., Shoemaker,B.A., Thiessen,P.A., He,S., Zhang,J. and Bryant,S.H. (2016) PubChem BioAssay: 2017 update. *Nucleic Acids Res.*, doi:10.1093/nar/gkw1118.
53. Tym,J.E., Mitsopoulos,C., Coker,E.A., Razaz,P., Schierz,A.C., Antolin,A.A. and Al-Lazikani,B. (2016) canSAR: an updated cancer research and drug discovery knowledgebase. *Nucleic Acids Res.*, **44**, D938–D943.
54. Williams,A.J., Harland,L., Groth,P., Pettifer,S., Chichester,C., Willighagen,E.L., Evelo,C.T., Blomberg,N., Ecker,G., Goble,C. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, **17**, 1188–1198.
55. Koscielny,G., An,P., Carvalho-Silva,D., Cham,J.A., Munoz-Pomer Fuentes,A., Fumis,L., Gasparyan,R., Hasan,S., Karamanis,N., Maguire,M. *et al.* (2016) Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, doi:10.1093/nar/gkw1055.
56. Nguyen,D.T., Mathias,D., Bologa,C., Brunak,S., Fernandez,N., Gaulton,A., Hersey,A., Holmes,J., Jensen,L., Karlsson,A. *et al.* (2016) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, doi:10.1093/nar/gkw1072.
57. Chambers,J., Davies,M., Gaulton,A., Papadatos,G., Hersey,A. and Overington,J.P. (2014) UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. *J. Cheminform.*, **6**, 43.
58. Papadatos,G., Gaulton,A., Hersey,A. and Overington,J.P. (2015) Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided Mol. Des.*, **29**, 885–896.