D1100–D1107 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 23 September 2011*
doi:10.1093/nar/gkr777

ChEMBL: a large-scale bioactivity database for drug discovery

Anna Gaulton[1], Louisa J. Bellis[1], A. Patricia Bento[1], Jon Chambers[1], Mark Davies[1], Anne Hersey[1], Yvonne Light[1], Shaun McGlinchey[1], David Michalovich[2], Bissan Al-Lazikani[3] and John P. Overington[1,*]

[1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, [2]David Michalovich Scientific Consulting, London and [3]Cancer Research UK Cancer Therapeutics Unit, Institute of Cancer Research, 15 Cotswold Road, Belmont, Surrey, SM2 5NG, UK

Received August 15, 2011; Accepted September 5, 2011

ABSTRACT

ChEMBL is an Open Data database containing binding, functional and ADMET information for a large number of drug-like bioactive compounds. These data are manually abstracted from the primary published literature on a regular basis, then further curated and standardized to maximize their quality and utility across a wide range of chemical biology and drug-discovery research problems. Currently, the database contains 5.4 million bioactivity measurements for more than 1 million compounds and 5200 protein targets. Access is available through a web-based interface, data downloads and web services at: https://www.ebi.ac.uk/chembldb.

INTRODUCTION

A wealth of information on the activity of small molecules and biotherapeutics exists in the literature, and access to this information can enable many types of drug discovery analysis and decision making. For example: selection of tool compounds for probing targets or pathways of interest; identification of potential off-target activities of compounds which may pose safety concerns, explain existing side effects or suggest new applications for old compounds; analysis of structure–activity relationships (SAR) for a compound series of interest; assessment of *in vivo* absorption, distribution, metabolism, excretion and toxicity (ADMET) properties; or construction of predictive models for use in selection of compounds potentially active against a new target (1–5). Access to this information is especially important due to the continuing shift in fundamental research on disease mechanisms from the private to public sectors.

However, bioactivity data published in journal articles are usually found in a relatively unstructured format and are labour-intensive to search and extract. For example, compound structures are frequently depicted only as images and are not therefore searchable, protein targets may be referred to by a variety of synonyms or abbreviations with no reference to any database identifiers, and details of assays may be included only in Supplementary Data or by reference to previous publications. In addition, there is not currently any requirement by most journals for authors to deposit small-molecule assay results in public databases (as is the case for sequence, protein structure and gene expression data). Historically, therefore, the majority of the published small-molecule bioactivity data have only been readily available via commercial products.

In recent years, in response to the growing demand for open access to this kind of information, a variety of public-domain bioactivity resources have been developed. PubChem BioAssay (6) and ChemBank (7) are large archival databases providing access to millions of deposited screening results, typically from high-throughput screening (HTS) experiments. A number of other primary resources extract bioactivity data from literature, but tend to focus on particular thematic areas, and primarily on binding affinity information. For example, BindingDB contains quantitative binding constants manually extracted from publications, focusing chiefly on proteins that are considered to be potential drug targets (8). PDBBind (9), Binding MOAD (10) and AffinDB (11) contain binding affinity information for protein–ligand complexes found in the Protein Data Bank (PDB, 12). PDSP Ki database stores screening data from the National Institute of Mental Health's Psychoactive Drug Screening Program (13). BRENDA provides binding constants for enzymes (14), IUPHAR contains ligand information for receptors and ion channels (15), while GLIDA

*To whom correspondence should be addressed. Tel: + 44 (0) 1223 492 666; Fax: + 44 (0) 1223 494 468; Email: jpo@ebi.ac.uk

(16) and GPCRDB (17) provide information specifically for G-protein-coupled receptors. Other resources, such as DrugBank, provide detailed annotation around the properties and mechanism of action of approved drugs (18).

However, in order to make informed decisions in drug discovery or to design experiments to probe a biological system with chemical tools, it is important to consider not only the binding affinity of a compound for its target, but also its selectivity, efficacy in functional assays or disease models and the likely ADMET properties of the compound. Moreover, researchers need the ability to intelligently cluster relevant information across studies (based on target or compound similarities, for example) and to integrate data across therapeutic areas. ChEMBL aims to bridge this gap by providing broad coverage across a diverse set of targets, organisms and bioactivity measurements reported in the scientific literature, together with a range of user-friendly search capabilities (19).

DATA CONTENT

Data extraction and curation

The core activity data in the ChEMBL database are manually extracted from the full text of peer-reviewed scientific publications in a variety of journals, such as *Journal of Medicinal Chemistry*, *Bioorganic Medicinal Chemistry Letters* and *Journal of Natural Products*. The set of journals covered is by no means comprehensive, but is selected to capture the greatest quantity of high-quality data in a cost, and time-effective manner. From each publication, details of the compounds tested, the assays performed and any target information for these assays are abstracted.

Structures for small molecules are drawn in full, in machine-readable format, despite the structure often being provided as a scaffold and a list of R-group substituents, or referred to only by name in the original publication. Information about the particular salt form tested is also captured, where available, although this is often inconsistent in the literature. Before loading to the database, structures are checked for potential problems (e.g. unusual valence on atoms, incorrect structures for common compounds/drugs), then normalized according to a set of rules, to ensure consistency in representation (e.g. compounds are neutralized by protonating/deprotonating acids and bases to ensure a formal charge of zero where possible). Preferred representations are used for certain common groups (e.g. sugars, sulphoxides and nitroxides). Some chemical structures are typically only reported in an implicit format, and this is checked and assigned on registration—for example, the stereochemistry of the steroid framework is invariably not published, but is assumed to be that of the naturally occurring configuration, unless otherwise defined. Common salts are also stripped from the extracted compounds, and both the salt form and the parent compound are entered into the database. This allows users to view all data associated with the same parent compound, regardless of the salt form tested, while still retaining the salt information if required.

Details of all types of assays performed are extracted from each publication, including binding assays (measuring the interaction of the compound with the target directly), functional assays (often measuring indirect effects of the compound on a pathway, system or whole organism) and ADMET assays (measuring pharmacokinetic properties of the compound, interaction with key metabolic enzymes or toxic effects on cells/tissues). The activity endpoints measured in these assays are recorded with the values and units as given in the paper, but for the purposes of improved querying are also standardized, where possible, to convert them to a preferred unit of measurement for a given activity type (e.g. IC50 values are displayed in nM, rather than μM/mM/M, half-life is reported in hours rather than minutes/days/weeks). This enables the user to more easily compare data across different assays.

To maximize the utility of bioactivity data, the targets of assays need to be represented robustly and consistently, in a manner independent of the various adopted names and synonyms used across different sources. To this end, detailed manual annotation of targets is carried out within ChEMBL. Where the intended molecular target of an assay is reported in a publication, this information is extracted, together with associated details of the relevant organism in which the assay was performed (or the organism from which the protein/cell-line was derived for an *in vitro* assay). Target assignments are carefully checked by our curators, and corrected where necessary, then further annotated where any ambiguity exists. For example, for an *in vitro* binding assay, it is often possible to determine the precise protein target with which the compound is interacting and assign a single relevant protein to the assay. However, in other cases this may not be possible. For example, an assay may describe interaction of a compound with a target which is known to be a protein/biomolecular complex (e.g. ribosomes, GABA-A receptors or integrins). In this case, several protein subunits may be assigned to the assay, but a 'complex' field in the database is used to record the fact that these proteins are associated as a specific protein complex. In other cases, the assay performed may not allow elucidation of the precise protein subtypes with which a compound is interacting (e.g. cell/tissue-based assays where several closely related subtypes of the protein are likely to be expressed, or those reported prior to the discovery of particular receptor/enzyme subtypes). Again, the assay may therefore be mapped to each of the possible protein targets, but a 'multi' field in the database records the fact that it is not clear whether the compound is interacting non-specifically with all of these proteins, and consequently less confidence should be placed in these assignments.

In many cases, such as whole organism-based phenotypic assays, it is not possible to unambiguously determine the protein target that is responsible for the observed effect of the compound. In these cases, the assay will be mapped to a ChEMBL target representing the non-molecular system on which an effect is observed. For example, an assay measuring the cytotoxicity of a compound against the human breast carcinoma-derived

MCF-7 cells would be mapped to a ChEMBL cell-line target representing MCF-7. An *in vitro* assay measuring inhibition of growth of *Mycobacterium tuberculosis* would be mapped to a ChEMBL organism target representing *M. tuberculosis*. This allows users to easily retrieve information about other assays performed on the same systems, even though the underlying mechanism of action of the compounds might be different. Protein targets are further classified into a manually curated family hierarchy, according to nomenclature commonly used by drug discovery scientists (e.g. ligand-based classification of G-protein-coupled receptors, division of enzymes into proteases/kinases/phosphatases etc.), and organisms are classified according to a simplified subset of the NCBI taxonomic structure (20). This also allows data to be queried at a higher level (e.g. for all protein kinases or *Mycobacterium* species).

Approved drugs

In addition to literature-derived data, ChEMBL also contains structures and annotation for Food and Drug Administration (FDA)-approved drugs. For each drug entry, any information about approved products (from the FDA Orange Book, 21) including their trade names, administration routes, dosage information and approval dates is included in the database. Structures for novel drug ingredients are manually assigned, and for protein therapeutics, amino-acid sequences may be included, where available. Each drug is also annotated according to the drug type (synthetic small molecule, natural product-derived small molecule, antibody, protein, oligosaccharide, oligonucleotide, inorganic etc.), whether there are 'black box' safety warnings associated with a product containing that active ingredient, whether it is a known prodrug, the earliest approval date (where known), whether it is dosed as a defined single stereoisomer or racemic mixture, and whether it has a therapeutic application (as opposed to imaging/diagnostic agents, additives etc.). This information allows users of the bioactivity data to assess whether a compound of interest is an approved drug and is therefore likely to have an advantageous safety/pharmacokinetic profile or be orally bioavailable, for example.

Data model

The most important entity types within ChEMBL are documents (from which the data are extracted), compounds (substances that have been tested for their bioactivity), assays (individual experiments that have been carried out to assess bioactivity) and targets (the proteins or systems being monitored by an assay). Each extracted document has a list of associated compound records and assays, which are linked together by activities (i.e. the actual endpoints measured in the assay with their types, values and units).

Since the same compound may have been tested multiple times in different assays and publications, the compound records are collapsed, based on structure, to form a non-redundant molecule dictionary. Standard IUPAC Chemical Identifier (InChI) representation (22) is used to determine which compounds are identical and which should be registered with new identifiers. In general, the Standard InChI representation distinguishes stereoisomers of a compound, but not tautomers. Hence, stereoisomers will be given unique identifiers, but tautomers will not. We have taken the view that although a particular binding interaction may involve a specific ionization or tautomer state, in a biological assay, there will be interconversion and equilibration across these forms. A smaller number of protein therapeutics and substances with undefined structures are also included in the molecule dictionary. Additional information is then associated with the entries in this table, such as structure representations, calculated properties, synonyms, drug information and parent–salt relationships.

Similarly, a non-redundant target dictionary stores a list of the proteins, nucleic acids, subcellular fractions, cell-lines, tissues and organisms that are subject to investigation. Each assay is then mapped to one or more entries in this dictionary, as described above. Further information, such as protein family classification, is also linked to the target dictionary.

Each record in the documents, assays, molecule dictionary and target dictionary tables is assigned a unique ChEMBL identifier, which takes the form of a 'CHEMBL' prefix followed immediately by an integer (e.g. CHEMBL25 is the compound aspirin, CHEMBL210 is the human β-2 adrenergic receptor target). In addition, external identifiers are recorded for these entities where possible. For example, all small molecule compounds with defined structures are assigned ChEBI identifiers (23) and Standard InChIKeys. Where data are taken from other resources, the original identifiers are also retained (e.g. SIDs and AIDs for PubChem substances and assays, HET codes for PDBe ligands). PubMed identifiers or Digital Object Identifiers (DOIs) are stored for documents (20,24). Protein targets are represented by primary accessions within the UniProt protein database (25), and organism targets are assigned NCBI taxonomy IDs and names.

Data exchange

The PubChem BioAssay database accepts deposited results from many laboratories and screening centres and contains a large quantity of data, primarily from high-throughput screening experiments, measuring inhibition of a target by large numbers of compounds, often at a single compound concentration. As such, the number of data points within PubChem is huge, but a very small proportion of these represent compounds with dose–response measurements (e.g. IC50, Ki) of an affinity likely to specifically perturb a biological system. In contrast, due to extraction from published pharmacology and drug discovery literature, ChEMBL contains a much larger proportion of active compounds identified using dose–response assays. The number of distinct protein targets with dose–response measurements recorded in PubChem is also smaller (currently fewer than 700 proteins, compared with more than 4000 in ChEMBL). However, there are also novel protein targets in PubChem that are not currently included in ChEMBL. Therefore, the types of data reported in

PubChem and ChEMBL are distinct and complementary. To maximise the utility of the two data sets to users, we have worked with the PubChem group to develop a data exchange mechanism. All ChEMBL literature-derived assays are now included in PubChem BioAssay, and a subset of PubChem assays (confirmatory and panel assays with dose–response endpoints) have been loaded into ChEMBL. Assays from PubChem are clearly marked, both on the ChEMBL interface and in the database, allowing users to easily determine where data have originated, while benefiting from being able to retrieve more information through a single point of access.

Similarly, compounds and binding measurements from ChEMBL have been integrated into BindingDB, and the reciprocal incorporation of BindingDB data into ChEMBL is planned.

Current content

Release 11 of the ChEMBL database contains information extracted from more than 42 500 publications, together with several deposited datasets, and data drawn from other databases (Table 1). In total, there are more than 1 million distinct compound structures represented in the database, with 5.4 million activity values from more than 580 000 assays. These assays are mapped to 8200 targets, including 5200 proteins (of which 2388 are human).

DATA ACCESS

The ChEMBL interface

The ChEMBL database is accessible via a simple, user-friendly interface at: https://www.ebi.ac.uk/chembldb. This interface allows users to search for compounds, targets or assays of interest in a variety of ways.

For example, users wishing to retrieve potential tool compounds for a target of interest can perform a keyword search of the database using a protein name, synonym, UniProt accession or ChEMBL target identifier of interest. Alternatively, targets can be browsed according to protein family (e.g. to retrieve all chemokine receptors), or organism (e.g. to retrieve all *Plasmodium falciparum* targets). Since the database only includes protein targets for which bioactivity data are available, users can also perform a BLAST search of the ChEMBL target dictionary with a protein sequence of interest. This can be useful to identify closely related proteins with activity data, even if the sequence of interest is not represented in the database (e.g. activity data for a mouse orthologue of a human target).

Having retrieved a target, or multiple targets, of interest, a simple drop-down menu allows users to display all associated bioactivity data, or to filter the available data to select activity types of interest (for example to include only IC50 and Ki measurements below a given concentration threshold, or only certain ADMET endpoints, see Supplementary Figure 1). The resulting bioactivity table gives details of each compound that was tested (together with the particular salt form used in the assay), the measured activity type, value and units, a description of the assay, details of the target (including the organism) and, importantly, a link to the publication from which the data have been extracted. Data from this view can be exported as a text file or spread sheet for further analysis.

Alternatively, users may have a particular compound of interest and wish to retrieve potency, selectivity or ADMET information for this, or closely related compounds. Again, users can search for compounds using a keyword search with names/synonyms or ChEMBL identifiers. However, a more effective strategy will often be to search by compound structure. The interface provides a choice of several different drawing tools (26), allowing users to sketch in a structure or substructure of interest (Figure 1). A compound similarity or substructure search of the database (implemented using the Accelrys Direct Oracle Cartridge: http://accelrys.com/products/informatics/cheminformatics/accelrys-direct.html) can then be carried out to retrieve ChEMBL compounds similar to, or containing, the input structure.

Having retrieved a list of compounds of interest, a variety of calculated properties such as molecular weight, calculated lipophilicity (AlogP, 27) and polar surface area (28) can be viewed and filtered via a graphical display. This may be useful to restrict the set of compounds to those that are likely to have appropriate

Table 1. Sources of compound and bioactivity data in ChEMBL_11

Data Source	Number of compound structures	Number of assays	Number of activity results	Number of targets	Number of protein targets	Number of organisms
ChEMBL literature extraction	629 943	580 624	3 282 945	7 957	5104	1552
PubChem BioAssay[a]	364 203	1636	2 079 974	681	647	63
GSK TCAMS Malaria Data (32)	13 467	6	81 198	3	0	2
PDBe Ligands	12 337	0	0	0	0	0
Novartis-GNF Malaria Data (33)	5675	4	22 788	3	0	2
St Jude Children's Hospital Malaria Data[b] (34)	1524	16	5456	8	0	5
Guide to Receptors and Channels (35)	560	344	801	239	239	6
Sanger Institute Genomics of Drug Sensitivity in Cancer	17	352	5984	352	0	1

[a]PubChem BioAssay set includes only confirmatory/panel assays from PubChem that have dose–response end points.
[b]Only compounds with dose-response measurements from the St Jude malaria screening data set have been incorporated into ChEMBL, but the full high-throughput screening data can be downloaded from the ChEMBL-NTD website: https://www.ebi.ac.uk/chemblntd.

Parent	Ingredient	Bioactivity	Activity Comment	Operator	Value	Units	Assay ChEMBL ID	Assay Source	Assay Type	Description	ChEMBL Target ID	Target Name	Organism	Target Mapping	Curated By	Reference	Name in Reference
CHEMBL100150	CHEMBL100150	Ki		=	19	nM	CHEMBL760198	Scientific Literature	B	Tested in vitro for antagonistic activity to displace [3H]-PAF from rabbit platelet membrane PAF receptors	CHEMBL250	Platelet activating factor receptor	Homo sapiens	Homologous protein	Expert	Bioorg. Med. Chem. Lett., (1995) 5:23:2903	11b
CHEMBL100251	CHEMBL100251	Ki		=	6	nM	CHEMBL760198	Scientific Literature	B	Tested in vitro for antagonistic activity to displace [3H]-PAF from rabbit platelet membrane PAF receptors	CHEMBL250	Platelet activating factor receptor	Homo sapiens	Homologous protein	Expert	Bioorg. Med. Chem. Lett., (1995) 5:23:2903	11c
Chiral CHEMBL100257	Chiral CHEMBL100257	Ki		=	10	nM	CHEMBL760192	Scientific Literature	B	Binding affinity for platelet activating factor receptor using [3H]-PAF in rabbit platelet membranes	CHEMBL250	Platelet activating factor receptor	Homo sapiens	Homologous protein	Expert	Bioorg. Med. Chem. Lett., (1995) 5:23:2913	3i
CHEMBL10103	CHEMBL10103	IC50		=	1800	nM	CHEMBL640415	Scientific Literature	B	Binding affinity towards AMPA receptor using [3H]-AMPA as radioligand	CHEMBL3502	Glutamate receptor ionotropic, AMPA 1	Mus musculus	Multiple homologous proteins	Autocuration	Bioorg. Med. Chem. Lett., (1994) 4:2:371	7

Figure 1. Retrieving bioactivity data with a substructure search. A choice of sketchers allows the user to enter a structure of interest and search the database for compounds similar to, or containing that substructure (**a**). The resulting list of compounds can then be filtered graphically, according to their physicochemical properties (e.g. calculated lipophilicity AlogP and molecular weight) using the sliders and 'update chart' button (**b**). When a suitable compound set has been created, a drop-down menu allows the user to retrieve all relevant bioactivity results from the database, or filter the results further by activity type (**c**).

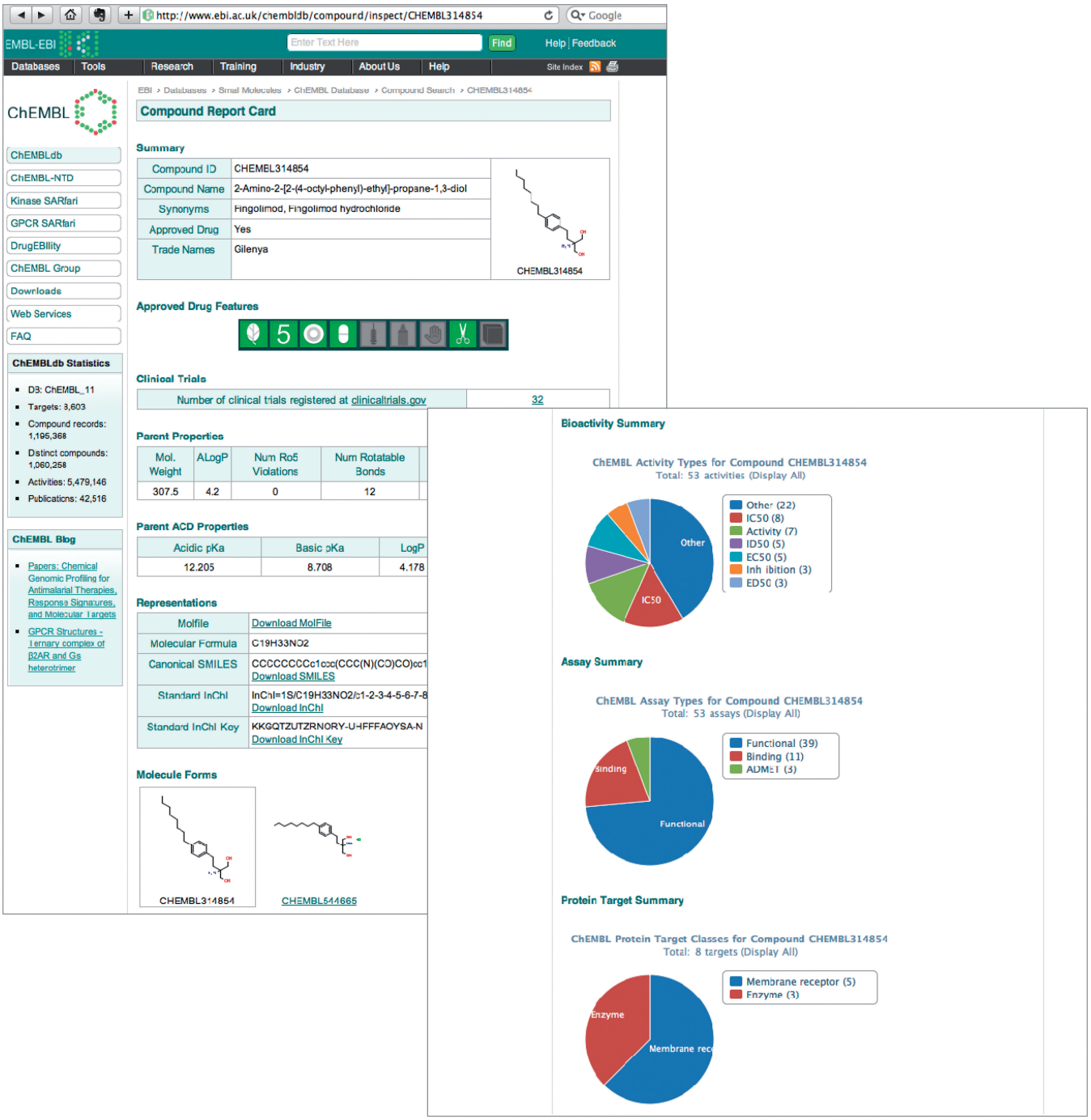

Figure 2. Compound report card for Fingolimod (CHEMBL314854) showing synonyms, approved drug features (see Supplementary Figure 2), a link to retrieve clinical trial data, calculated compound properties and structure representations, and different salt forms of the molecule (in this case, a hydrochloride salt). The lower portion of the page has a series of clickable widgets, showing breakdown of the activity data for this compound by activity type (e.g. IC50, EC50), assay type (e.g. binding/functional/ADMET) or target type (e.g. enzyme, receptor). Clicking on a portion of one of the pie charts takes the user directly to the relevant bioactivity results.

drug-like properties (29), before retrieving or filtering the associated bioactivity data.

For each of the main data types in ChEMBL (compounds, targets, assays and documents), report card pages are available. These provide further details about the entity of interest, such as names and synonyms (for targets and compounds), journal/abstract details (for documents), drug annotation, structures and calculated physicochemical properties (for compounds), together with cross-references to other resources (e.g. UniProt, PDBe, ChEBI, DrugBank and CiteXplore: http://www.ebi.ac.uk/citexplore). Each report card also contains a

series of clickable graphical 'widgets' summarizing and providing rapid access to all of the bioactivity data available for that entity (Figure 2).

A table view of approved drugs is also provided, with relevant annotation (e.g. drug type, administration route, 'black box' safety warnings) indicated by a series of sortable icons (see Supplementary Figure 2). Users can download the structures for these drugs or go to report cards to access further information, such as bioactivity data.

Downloads and web services

While the ChEMBL interface provides the functionality required for many common use-cases, some users may prefer to download the database and query it locally (for use in large-scale data mining, to integrate with their own proprietary data, or due to data security policies around the use of chemical structures at their institutions, for example). Each release of ChEMBL is freely available from our ftp site in a variety of formats, including Oracle, MySQL, 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).

In addition, a set of RESTful web services is provided (together with sample Java, Perl and Python clients), to allow programmatic retrieval of ChEMBL data in XML or JSON formats (see https://www.ebi.ac.uk/chembldb/ws for more details).

Finally, to allow greater interoperability of the ChEMBL data with molecular interaction and pathway data (e.g. for annotation of pathways with chemical tools), a subset of the database (compounds active in binding assays against protein targets) is available in PSI-MITAB 2.5 format (30) via PSICQUIC web services (31).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to former colleagues at Inpharmatica Ltd., our data extractors, part-time curators and interns for their contributions to the database. We thank Yanli Wang and Evan Bolton for their assistance with the PubChem data integration. We also greatly appreciate and acknowledge the feedback from users on data content and organization of the database.

FUNDING

A Strategic Award for Chemogenomics from the Wellcome Trust [086151/Z/08/Z]; and the European Molecular Biology Laboratory. Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Paolini,G.V., Shapland,R.H.B., van Hoorn,W.P., Mason,J.S. and Hopkins,A.L. (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.
2. Mestres,J., Gregori-Puigjané,E., Valverde,S. and Solé,R.V. (2009) The topology of drug–target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.*, **5**, 1051–1057.
3. Wassermann,A.M. and Bajorath,J. (2011) Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med. Chem.*, **3**, 425–436.
4. Papadatos,G., Alkarouri,M., Gillet,V.J., Willett,P., Kadirkamanathan,V., Luscombe,C.N., Bravi,G., Richmond,N.J., Pickett,S.D., Hussain,J. *et al.* (2010) Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of HERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.*, **50**, 1872–1886.
5. Keiser,M.J., Setola,V., Irwin,J.J., Laggner,C., Abbas,A.I., Hufeisen,S.J., Jensen,N.H., Kuijer,M.B., Matos,R.C., Tran,T.B. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.
6. Wang,Y., Bolton,E., Dracheva,S., Karapetyan,K., Shoemaker,B.A., Suzek,T.O., Wang,J., Xiao,J., Zhang,J. and Bryant,S.H. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–D266.
7. Seiler,K.P., George,G.A., Happ,M.P., Bodycombe,N.E., Carrinski,H.A., Norton,S., Brudz,S., Sullivan,J.P., Muhlich,J., Serrano,M. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.
8. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
9. Wang,R., Fang,X., Lu,Y., Yang,C. and Wang,W. (2005) The PDBBind database: Methodologies and updates. *J. Med. Chem.*, **48**, 4111–4119.
10. Benson,M.L., Smith,R.D., Khazanov,N.A., Dimcheff,B., Beaver,J., Dresslar,P., Nerothin,J. and Carlson,H.A. (2008) Binding MOAD, a high-quality protein–ligand database. *Nucleic Acids Res.*, **36**, D674–D678.
11. Block,P., Sotriffer,C.A., Dramburg,I. and Klebe,G. (2006) AffinDB: a freely accessible database of affinities for protein–ligand complexes from the PDB. *Nucleic Acids Res.*, **34**, D522–D536.
12. Velankar,S., Alhroub,Y., Alili,A., Best,C., Boutselakis,C.H., Caboche,S., Conroy,M.J., Dana,J.M., van Ginkel,G., Golovin,A. *et al.* (2011) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **39**, D402–D410.
13. Roth,B.L., Kroeze,W.K., Patel,S. and Lopez,E. (2000) The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist*, **6**, 252–262.
14. Scheer,M., Grote,A., Chang,A., Schomburg,I., Munaretto,C., Rother,M., Sohngen,C., Stelzer,M., Thiele,J. and Schomburg,D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, **39**, D670–D676.
15. Sharman,J.L., Mpamhanga,C.P., Spedding,M., Germain,G., Staels,B., Dacquet,C., Laudet,V. and Harmar,A.J. (2011). NC-IUPHAR. (2011) IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic Acids Res.*, **39**, D534–D538.
16. Okuno,Y., Yang,J., Taneishi,K., Yabuuchi,H. and Tsujimoto,G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.*, **34**, D673–D677.
17. Horn,F., Weare,J., Beukers,M., Horsch,S., Bairoch,A., Chen,W., Edvardsen,O., Campagne,F. and Vriend,G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.*, **26**, 275–279.
18. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

19. Warr,W.A. (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput. Aided Mol. Des.*, **23**, 195–198.
20. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
21. U.S. Department of Health and Human Services. (2011). *Approved Drug Products with Therapeutic Equivalence Evaluations*, 31st edn. U.S. Government Printing Office, Washington DC.
22. Stein,S.E., Heller,S.R. and Tchekhovskoi,D. (2003) An open standard for chemical structure representation: The IUPAC Chemical Identifier. *Proceedings of the 2003 International Chemical Information Conference (Nîmes)*. Infonortics, Tetbury, pp. 131–143.
23. De Matos,P., Alcantara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical entities of biological interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
24. Paskin,N. (2010) Digital Object Identifier (DOI®) System. In: Bates,M.J. and Maack,M.N. (ed). *Encyclopedia of Library and Information Sciences*, 3rd edn. Taylor & Francis, London pp. 1586–1592.
25. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
26. Ertl,P. (2010) Molecular structure input on the web. *J. Chemoinform.*, **2**, 1.
27. Ghose,A.K. and Crippen,G.M. (1987) Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.*, **27**, 21–35.
28. Ertl,P., Rohde,B. and Selzer,P. (2000) Fast calculation of molecular polar surface area as a sum of fragment based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, **43**, 3714–3717.
29. Lipinski,C.A., Lombardo,F., Dominy,B.W. and Feeney,P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **23**, 3–25.
30. Kerrien,S., Orchard,S., Montecchi-Palazzi,L., Aranda,B., Quinn,A.F., Vinod,N., Bader,G.D., Xenarios,I., Wojcik,J., Sherman,D. *et al.* (2007) Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.
31. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E. *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods*, **8**, 528–529.
32. Gamo,F.-J., Sanz,L.M., Vidal,J., de Cozar,C., Alvarez,E., Lavandera,J.-L., Vanderwall,D.E., Green,D.V.S., Kumar,V., Hasan,S. *et al.* (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature*, **465**, 305–310.
33. Plouffe,D., Brinker,A., McNamara,C., Henson,K., Kato,N., Kuhen,K., Nagle,A., Adrian,F., Matzen,J.T., Anderson,P. *et al.* (2008) In silico acitivity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl Acad. Sci. USA*, **105**, 9059–9064.
34. Guiguemde,W.A., Shelat,A.A., Bouck,D., Duffy,S., Crowther,G.J., Davis,P.H., Smithson,D.C., Connelly,M., Clark,J., Zhu,F. *et al.* (2010) Chemical genetics of *Plasmodium falciparum*. *Nature*, **465**, 311–315.
35. Alexander,S.P.H., Mathie,A. and Peters,J.A. (2009) Guide to Receptors and Channels (GRAC), 4th edn. *Br. J. Pharmacol.*, **158**, S1–S254.