Published online 29 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D963–D970***
doi: 10.1093/nar/gky963

DrugCentral 2018: an update

Oleg Ursu[1], Jayme Holmes[1], Cristian G. Bologa[1], Jeremy J. Yang[1], Stephen L. Mathias[1], Vasileios Stathias[2], Dac-Trung Nguyen[3], Stephan Schürer[2] and Tudor Oprea[1,*]

[1]Translational Informatics Division, Department of Internal Medicine, The University of New Mexico Health Science Center, Albuquerque, NM 87131, USA, [2]Center for Computational Science, Miller School of Medicine, University of Miami, Coral Gables, FL 33146, USA and [3]National Center for Advancing Translational Science, 9800 Medical Center Drive, Rockville, MD 20850, USA

Received September 12, 2018; Revised October 02, 2018; Editorial Decision October 04, 2018; Accepted October 26, 2018

ABSTRACT

DrugCentral is a drug information resource (http://drugcentral.org) open to the public since 2016 and previously described in the 2017 Nucleic Acids Research Database issue. Since the 2016 release, 103 new approved drugs were updated. The following new data sources have been included: Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS), FDA Orange Book information, L1000 gene perturbation profile distance/similarity matrices and estimated protonation constants. New and existing entries have been updated with the latest information from scientific literature, drug labels and external databases. The web interface has been updated to display and query new data. The full database dump and data files are available for download from the DrugCentral website.

INTRODUCTION

The DrugCentral database (http://drugcentral.org) was first made available to the public in 2016 and published in the 2017 NAR Database issue (1). Since its initial publication, the database has been cited 33 times according to Google Scholar and has been integrated into multiple open public resources, such as ChEMBL (2), ChEBI (3), PubChem (4), Probes and Drugs (5), Pharos (6) and repoDB (7). It is also available commercially via the Ontoforce DISQOVER platform (https://www.disqover.com). On average, there are 90–100 full database downloads/month. In the 2018 release, we have continued to update, curate and add additional data types to the DrugCentral database. We have added chemical structures, annotations and mappings to external resources for 103 new active pharmaceutical ingredients (APIs). All bioactivity profiles, pharmacological classifications, mechanism of action annotations, pharmaceutical formulations, indications etc. were updated using manual and automated pipelines to the most current versions of the data sources used by DrugCentral. Manual curation was performed to maintain the high accuracy of existing data. Sixteen entries having no evidence of approval for human use were deprecated compared to the 2016 release. Mechanism of action annotations are continuously updated to reflect current state of knowledge (8).

New data sources added to DrugCentral include FAERS (https://open.fda.gov/data/faers/), the FDA Orange Book (https://bit.ly/2o04xoA), distance and correlation matrices based on L1000 Connectivity Map perturbational profiles from LINCS (9), and protonation constants as estimated with the MOKA software (10). The criteria for inclusion of new data sources were: (i) provide new information on drugs and APIs compared to the existing version and (ii) information not already available in other open-access drug information resources. For example, while patent information from the FDA Orange Book is included in DrugBank (11), market exclusivity, pharmaceutical formulations and patent use codes are not available. Furthermore, we are not aware of open-access drug information repositories currently capturing FAERS data.

For each of the new sources of data, automated pipelines were scripted to enable automatic updates for future DrugCentral releases. When necessary, additional processing is performed to enhance information usefulness for the end user. FAERS data were subjected to extensive processing including aggregating adverse events by API starting from multiple pharmaceutical formulations, followed by computing statistical significance of the aggregated API-associated adverse events. Multiple user requests on how to query the database from both academic/private sectors indicate interest in our resource and prompted us to publish a list of example queries to be used with local instances of DrugCentral database. The updated DrugCentral database will continue to be released in open-access format similar to the initial release and can be accessed at http://drugcentral.org or downloaded from same URL. Along with regular releases, DrugCentral integration with PubChem and UniChem (12) will be continuously updated to ensure up-to-date mappings and data connectivity. Positive feedback received so far indi-

*To whom correspondence should be addressed. Tel: +1 505 925 4756; Fax: +1 505 925 7625; Email: toprea@salud.unm.edu

cates that our database is a valuable resource for researchers interested in drug discovery.

CURRENT DATA CONTENT

Active ingredients

The 2018 DrugCentral release captures 103 new APIs compared to 2016 (Table 1). Small organic molecules (49%) and biologics (43%) are the majority of the new additions. The number of new biologics captured by DrugCentral confirms approval trends observed in recent years i.e. an increased number of newly approved biologics compared to small molecules (see Figure 1). Since the 2016 release, regulatory approvals from five different regulatory agencies were indexed into DrugCentral: Most of the regulatory approval information was extracted from the US FDA (67 APIs) and the European Medicines Agency, EMA (41 APIs), with 31 APIs in common between these two agencies. Approval information from Japan Pharmaceuticals and Medical Devices Agency (7 APIs), China Food and Drug Administration (6 APIs) and South Korea Ministry of Food and Drug Safety (5 APIs) was also processed and integrated into DrugCentral. As of 1 October 2018, the ChEMBL database (2) indexes 88 of these 103 APIs, KEGG (13) indexes 89, DrugBank captures 85 and the IUPHAR/BPS Guide to Pharmacology (14) 53, respectively (see Supplementary Table S1).

Bioactivity data and mechanism of action

Since the 2016 release, 1656 new bioactivity data points were added to DrugCentral. As in the previous release, non-overlapping data were extracted from ChEMBL (41.91%), IUPHAR/BPS Guide to Pharmacology (28.74%) and PDSP (15) (19.34%) via automated pipelines; 10.02% of these data points were manually curated from scientific literature and drug labels. The 99 bioactivity values associated with new APIs were extracted, mostly via manual curation of scientific literature and drug labels (77.78%), followed by IUPHAR/BPS Guide to Pharmacology (18.18%) and ChEMBL (4.04%). Mechanism of action (MoA) annotations have been updated since the 2016 release, by adding new annotations to existing and new APIs and by removing inconsistencies. For example, MoA targets for antipsychotics are now mapped to the dopamine D2 and serotonin 2A receptor, respectively, unless there is evidence for additional MoA targets. Other examples of the MoA curations steps are described elsewhere (8). A total of 78 APIs were mapped to 79 MoA targets compared to the 2016 release including 60 new APIs; the 11 newly introduced MoA drug targets are shown in Table 2. The distribution of MoA target classes for new APIs confirms recent trends (8), with 50% of new APIs targeting enzymes and kinases, 15% targeting GPCRs and 6.7% targeting cytokine receptors.

We recently introduced a knowledge-based classification for proteins, derived from mapping knowledge from associated APIs, from molecular probes and from multiple types of experimental data (e.g. gene expression, pathways, function etc.), as identified in literature, patents, drug labels, NIH funded grants and other sources (16). This Target Development Level (TDL) classification scheme categorizes proteins into four groups—Tclin, Tchem, Tbio and Tdark—with respect to the depth of investigation from a clinical, chemical and biological standpoint. Briefly, Tclin are MoA-designated proteins; Tchem are proteins associated with small molecule probes; Tbio are proteins for which we have significant biological insight and Tdark refers to the understudied proteins ('dark genome'). Not surprisingly, the majority of API bioactivities in DrugCentral 2018 are associated with Tclin (Figure 2). However, DrugCentral also captures close to 4000 bioactivities that are not related to Tclin. Most of these pertain to Tchem, whereas some bioactivities are associated with Tbio and Tdark. These bioactivities reflect secondary pharmacology (17) annotations for APIs.

Pharmacological classification

The 2018 versions of the World Health Organization Anatomic, Therapeutic and Chemical classification system, WHO ATC (https://www.whocc.no/), the FDA Established Pharmacologic Class (EPC) (https://bit.ly/2OWiJdH), the Medical Subject Headings, MeSH (18) and ChEBI (3) pharmacological classifications are mapped to both new and existing APIs in DrugCentral, using adaptive mapping schemes. For example, WHO ATC codes are mapped using INN names (https://bit.ly/1Kx5Qmy), FDA EPC classes are mapped using FDA Substance Registration System (SRS) UNII codes (https://fdasis.nlm.nih.gov/srs/), and MeSH pharmacological actions and ChEBI roles are mapped using MeSH/ChEBI identifiers. A summary of new pharmacological classification additions is presented in Table 1.

External identifiers

DrugCentral mappings to external resources continues to enable integration with other drug information resources including: ChEMBL (through the UniChem service) (2), ChEBI ontology (3), PubChem (4), the Probes & Drugs portal (5), Pharos (6), e-Drug3D (19), MyChem.info (http://mychem.info/) and repoDB (7). The current release of DrugCentral maps 64 997 external identifiers from 21 different resources, representing an increase of 3624 (5.9%) compared to 2016 release. The largest increases in mapped identifiers are for ChEBI (1273), ChEMBL (533) and DrugBank (456). New APIs are mapped to 1185 external identifiers, with FDA SRS (98), KEGG (89), ChEMBL (88) and DrugBank (85) having the most coverage. These mapping are provided to enable end-user integration of different drug information resources across multiple domains.

Pharmaceutical formulations

FDA pharmaceutical formulations from DailyMed (https://dailymed.nlm.nih.gov/) were updated using the May 2018 DailyMed files download. A total of 10 420 new formulations were added to DrugCentral 2018 compared to 2016, with 56.55% of these representing over the counter (OTC) formulations. Compared to 2016, the percentage of OTC formulations shows little increase, from 55.77% to 55.87%.

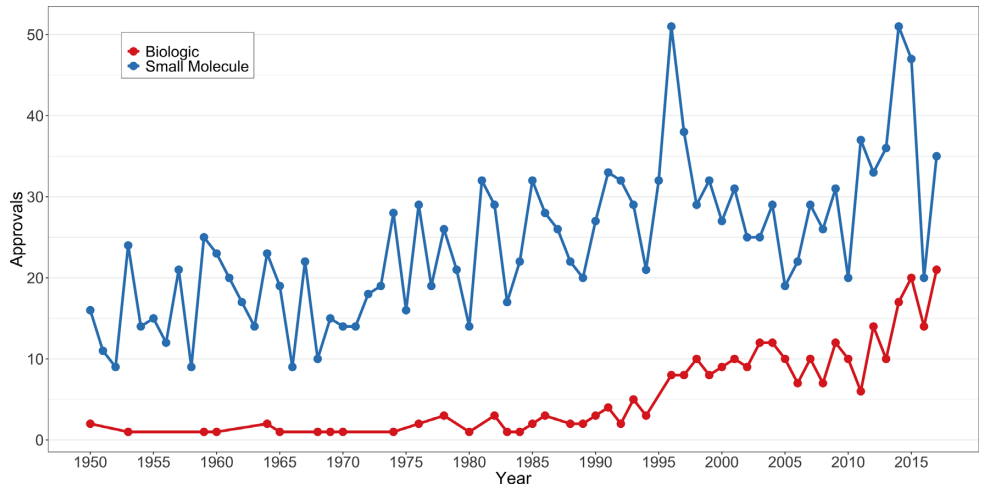

Figure 1. Small molecule and biologics approvals trends over the last six decades.

Table 1. Difference in data content between DrugCentral initial release (2016) and current release (2018)

	Entities (annotated APIs)	
	DrugCentral 2016	DrugCentral 2018
Active pharmaceutical ingredients	4444	4531
FDA drugs	2021	2094
Drugs approved outside US	2423	2437
Small molecules	3799	3825
Salts and inorganic molecules	112	115
Biologics and peptides	239	282
Other drugs	294	309
Parent molecules	199 (308)	211 (327)
Drug efficacy targets	837 (1689)	855 (1756)
Human protein targets	600 (1387)	613 (1447)
Infectious agents targets	194 (221)	197 (224)
Metabolites and biopolymers	43 (89)	45 (90)
Protein–drug crystal complex (PDB)	48 (82)	131 (236)
All protein–drug crystal complex (PDB)	1452 (283)	3991 (433)
Bioactivity data points	13 825 (1792)	15 481 (1911)
Human proteins	10 427 (1605)	11 241 (1692)
Other species	3398 (1002)	4240 (1175)
Pharmacological classification		
WHO ATC code	4195 (2941)	4889 (2978)
FDA Established Pharmacologic Class	428 (1165)	450 (1220)
MeSH pharmacological action	424 (2529)	457 (2615)
ChEBI ontology roles	285 (1487)	295 (1529)
Drug indications	2224 (2247)	2167 (2371)
Drug contra-indications	1458 (1376)	1407 (1379)
Drug off-label uses	847 (646)	817 (641)
Pharmaceutical products	67 064 (1660)	77 484 (1716)
Rx pharmaceutical products	29 665 (1561)	34 192 (1609)
OTC pharmaceutical products	37 399 (286)	43 292 (296)
External identifiers	61 349 (4444)	69 516 (4531)
CAS registry number	6072 (4444)	6200 (4531)
PubChem Compound Id	4175 (4175)	4289 (4308)
FDA Unique Ingredient Identifier (UNII)	4304 (4304)	4391 (4391)
ChEMBL-db id	5615 (4075)	6077 (4330)
WHO INN id	3519 (3519)	3589 (3589)
SNOMED-CT	4745 (2637)	4968 (2815)
KEGG DRUG	3501 (3501)	3576 (3576)
NDFRT	4171 (2406)	4256 (2479)
RxNorm RxCUI	2897 (2897)	2988 (2991)
IUPHAR/BPS ligand id	1345 (1345)	1391 (1395)
UMLS CUI	2839 (2839)	2835 (2835)
CHEBI	2557 (2557)	3824 (3830)
MeSH	4063 (3846)	4180 (3946)
DrugBank	2473 (2388)	2773 (2858)
Protein databank ligand id	646 (618)	713 (695)

Table 2. New APIs with novel mechanism of action approved since 2016 release of DrugCentral

Active ingredient	Indication	Approval	Target(s)
Brodalumab	Plaque psoriasis	VALEANT LUXEMBOURG (FDA, Feb, 2017)	Interleukin-17 receptor (IL17RA|IL17RC)
Telotristat ethyl	Diarrhea	LEXICON PHARMA INC (FDA, Feb, 2017)	Tryptophan 5-hydroxylase 1 (TPH1)
Dupilumab	Atopic dermatitis	REGENERON PHARMACEUTICALS (FDA, Mar, 2017)	Interleukin-4 receptor subunit α (IL4R)
Forodesine	Peripheral T-cell lymphoma	Mundi Pharma (PMDA, Mar, 2017)	Purine nucleoside phosphorylase (PNP)
Cenegermin	Neurotrophic keratitis	Dompe farmaceutici s.p.a. (EMA, Jul, 2017)	High affinity nerve growth factor receptor (NTRK1); Tumor necrosis factor receptor superfamily member 16 (NGFR)
Inotuzumab Ozogamicin	Pre B-cell acute lymphoblastic leukemia	Pfizer Limited (EMA, Jun, 2017)	B-cell receptor CD22 (CD22)
Enasidenib	Acute myeloid leukemia, disease	CELGENE CORP (FDA, Aug, 2017)	Isocitrate dehydrogenase [NADP], mitochondrial (IDH2)
Copanlisib	Follicular lymphoma	BAYER HEALTHCARE PHARMS (FDA, Sep, 2017)	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit α isoform (PIK3CA)
Benralizumab	Eosinophilic asthma	ASTRAZENECA AB (FDA, Nov, 2017)	Interleukin-5 receptor subunit α (IL5RA)
Elobixibat	Chronic constipation	EA Pharma (PMDA, Jan, 2018)	Ileal sodium/bile acid cotransporter (SLC10A2)

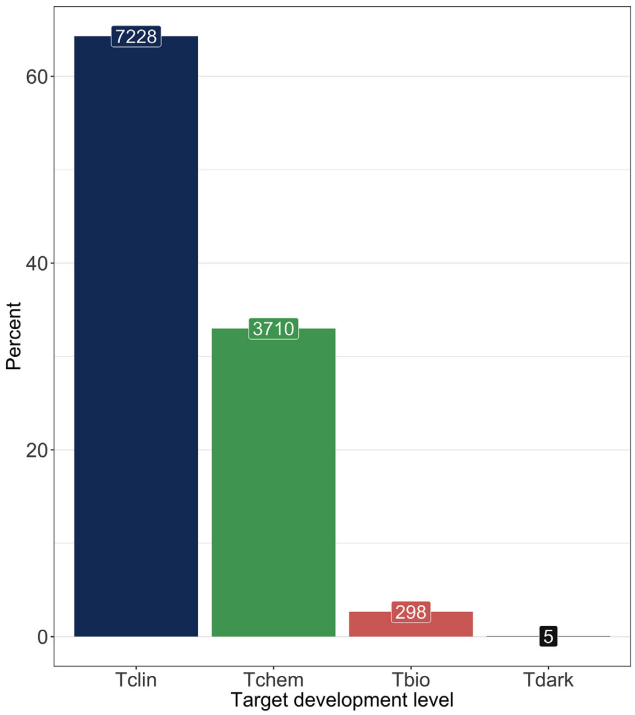

Figure 2. Distribution of APIs bioactivity data for human targets by target development levels.

Content from DailyMed drug labels is also indexed for full-text search functionality and is accessible via the DrugCentral website. While DailyMed files provide information on currently marketed pharmaceutical formulations in US, the FDA Orange Book provides information on both marketed and discontinued formulations. The Orange Book has been integrated in the 2018 release (*vide infra*).

NEW DATA AND FUNCTIONALITY

FDA Adverse events data

Adverse events data from FDA FAERS were downloaded, processed and integrated into the DrugCentral database. The data were processed as follows: (i) a total of 7437 pharmaceutical formulation and API names from FAERS were mapped onto 3307 DrugCentral identifiers; (ii) FAERS data not reported by healthcare professionals were discarded; (iii) only 'suspected drugs' were considered for each adverse event; (iv) only adverse events with three or more reports were considered for each API. After processing the data as described above, a total of 2023 unique APIs, associated with 10 779 unique MedDRA terms (Medical Dictionary for Regulatory Activities, https://www.meddra.org/), yielding 601 619 API—MedDRA term combinations were stored for further analyses. The Likelihood Ratio Test (LRT) signal detection procedure described by Huang *et al.* (20) was applied to identify API—MedDRA term combinations with disproportionally high reporting rates. Critical values based on P-value <0.05, under *null hypothesis* H_0 (reporting rate for an MedDRA term for an API equal to reporting rates for all other MedDRA terms for that same API), were computed and stored in DrugCentral in order to enable users to perform post-processing of adverse events using other criteria. According to LRT statistics, we identified 113 861 (18.93%) significant API—MedDRA term combinations for 1480 unique APIs and 7077 unique MedDRA terms. Figure 3 illustrates top 5 most reported APIs identified using LRT statistic grouped by the System Organ Class (SOC) MedDRA level; for clarity purposes only 9 SOC groups are shown. Acetylsalicylic acid (Aspirin) and Paracetamol (an API in Tylenol) are the top reported APIs in 7 out of 9 SOC groups. As noted by the authors of the LRT statistics (20), the signals detected are not always true signals; hence, a careful investigation is warranted before

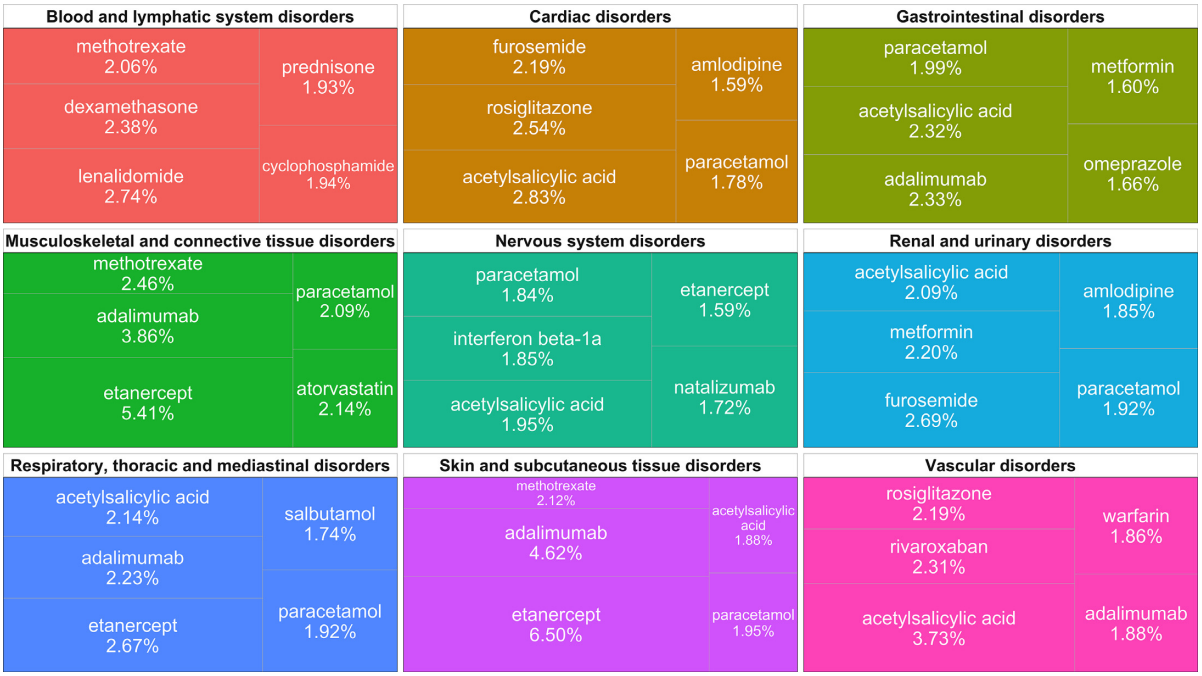

Figure 3. Top API and MedDRA system organ class terms associations computed using LRT statistic.

any critical decision is made. Some of the adverse events associated with Acetylsalicylic acid or Paracetamol might be false signals, and a case by case investigation may be required. The DrugCentral web application search functionality supports MedDRA terms querying and can be used to search for FAERS LRT-significant adverse events.

L1000 Connectivity Map perturbational profiles

Gene expression changes across multiple cell lines collected in the phase I (https://bit.ly/2PFHlZn) and phase II (https://bit.ly/2Luhj84) LINCS (Library of Integrated Cellular Signatures) data were mapped to 1613 unique DrugCentral APIs. The distance based on root mean square deviation (RMSD) and the Pearson correlation between perturbational profiles of APIs were computed across 81 cell lines and 8 757 622 active ingredients/cell lines combinations; these were stored in the DrugCentral database. Summary statistics of the distance metrics are presented in Table 3. A new search interface implemented in R-Shiny (https://shiny.rstudio.com/) was added to DrugCentral to enable browse and search of APIs having the most similar gene perturbation profiles (Figure 4); the web interface can be used to query and download correlation/distance profiles for any API. The complete correlation/distance matrix can be queried from a local instance of DrugCentral loaded from the provided database dump.

To provide useful reference points for end-users, we studied the distribution of these distance metrics across APIs sharing the same MoA. We extracted FDA established pharmacologic class annotations (http://bit.ly/2IqX5wi) from DrugCentral and identified API groups sharing MoA and API groups not sharing the same MoA, respectively. Both the Pearson correlation and RMSD (root-mean-square deviation) have overlapping distributions (see Supplementary Figure S1). The performance of both metrics was determined using the Receiver operating curve (ROC) analysis: Pearson correlation has an area under the curve (AUC) value of 0.76, sensitivity 0.69, specificity 0.7 and optimal cutoff value of 0.21, compared to normalized RMSD values of 0.72, 0.65, 0.69 and 1.08 respectively. Thus, the Pearson correlation has better performance across all ROC analysis parameters and was implemented as the default metric when ranking results in the DrugCentral web application.

Table 3. Summary statistic for drug gene perturbation profiles distance/correlation metrics

Metric	Mean (standard deviation)	Interquartile range
Pearson correlation	0.16 (0.13)	(0.07–0.24)
RMSD	2.58 (0.86)	(2.06–2.95)
Normalized RMSD	1.13 (0.20)	(1.00–1.23)

Protonation constants

Acid/base API properties have a profound effect on ingredient absorption, distribution, metabolism, excretion and toxicity (ADMET) (21). Active ingredient protonation constants (acid and base) were computed using the MOKA 3.0 software (10). Summary statistics of the estimated protonation constants are presented in Table 4. Most of the small molecule organic APIs (87.4%) are acidic or basic under the experimental pH range of 1 to 14, with 92% of acids and 97% of bases being monoprotic, diprotic or triprotic; another 407 APIs are neutral and lack ionizable groups (acidic or basic). The majority of acids (1265, 56%) are monoprotic and are relatively weak with a mean $pK_a = 7.87$ (IQR: 4.20–11.54). Two-thirds of the bases (1515, 65.64%) are weak

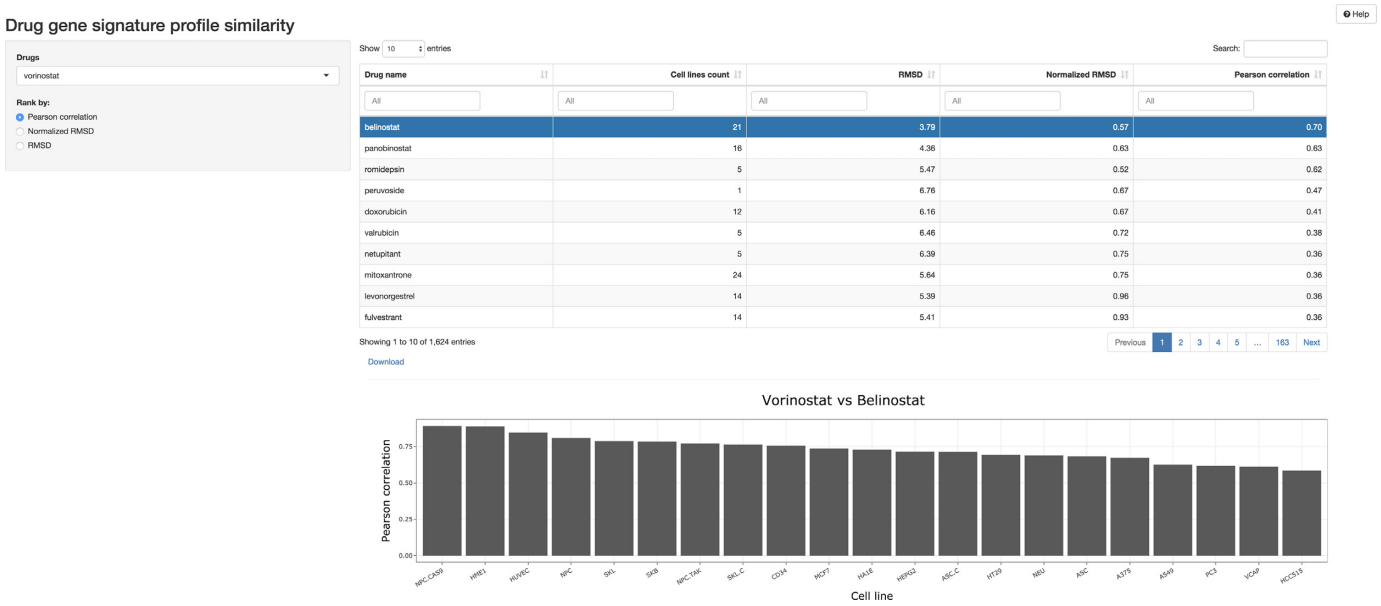

Figure 4. Web interface for query and browsing distance and correlation matrices based on L1000 Connectivity Map perturbational profiles from LINCS.

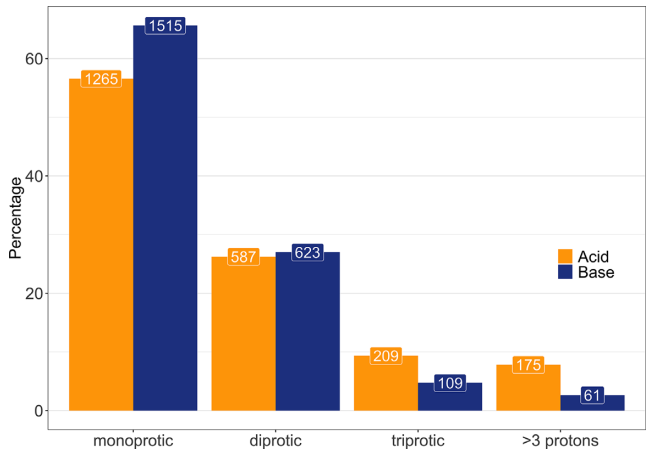

Figure. 5. Distribution of mono-, di-, tri- and multiprotic small molecule acids and bases API in DrugCentral.

monoprotic bases, having a mean pK_a = 7.65 (IQR: 6.84–9.14) (see Figure 5).

FDA Orange Book

To complement existing information on marketed pharmaceutical formulations, the FDA Orange Book was integrated in DrugCentral 2018. The Orange Book currently contains information on 34 850 currently marketed and 14 766 discontinued pharmaceutical products, as well as marketing exclusivity data, and patent information (Figure 6). Out of 20 084 unique marketed formulations in Orange Book, 96.56% are prescription (Rx) products while 3.44% are OTC. Information on market exclusivity data and patent information can be useful for, e.g., drug repurposing applications, and can provide an overview of the innovation landscape of FDA-approved drugs. For example, a query of APIs that have market exclusivity expirations in 2019 and have exclusivity code 'New Chemical Entity' shows that 32 APIs have exclusivity due to expire in 2019. Similar queries can be run to search for patent expiration data. A summary of these data is also available from the 'About' page on the DrugCentral website (http://drugcentral.org/about).

DATA ACCESS

Web interface

The DrugCentral web interface has been updated since the 2016 release to integrate novel data types and functionalities. All new data types, except for the L1000 signature distance/correlation matrices are integrated into drug card views. Furthermore, MedDRA adverse events terms were added to the search index and can be queried by typing full or partial search terms in the main query box. Due to the nature of the distance/correlation matrices based on the L1000 Connectivity Map perturbation profiles, we have designed a web interface tailored specifically for this data type (see Figure 4). A summary of the FDA Orange Book data is also available in the 'About' page on the DrugCentral website (http://drugcentral.org/about).

Download

For advanced users and for users who want to integrate DrugCentral data with other resources, download of the full database in PostgreSQL format has been available since the initial 2016 release. The full database dump can be loaded into a local PostgreSQL instance for advanced data query, export and integration. To facilitate the user interaction with the local instance, we have published multiple structured query language (SQL) examples available from the download section of the DrugCentral website. In addition to a full database dump, several user-requested downloads were also made available, including chemical structures of

Table 4. Summary statistic of estimated protonation constants for small molecule API

Protonation constant	Type	API count	Mean (standard deviation)	Interquartile range
pK_{a1}	acid	2236	7.87 (4.10)	(4.20–11.54)
	basic	1107	7.65 (2.30)	(6.84–9.14)
pK_{a2}	acid	971	10.19 (3.47)	(8.62–12.94)
	basic	1118	5.47 (2.90)	(3.04–8.09)
pK_{a3}	acid	384	11.42 (2.99)	(10.68–13.40)
	basic	600	4.46 (2.89)	(2.08–6.98)

Orange Book patent data (new drug applications)

Formulation strength	Trade name	Applicant	Application number	Approval date	Type	Dose form	Route	Patent number	Patent expiration date	Patent use
60MG	ERLEADA	JANSSEN BIOTECH	N210951	Feb. 14, 2018	RX	TABLET	ORAL	8802689	March 27, 2027	TREATMENT OF NON-METASTATIC, CASTRATION-RESISTANT PROSTATE CANCER (NM-CRPC)
60MG	ERLEADA	JANSSEN BIOTECH	N210951	Feb. 14, 2018	RX	TABLET	ORAL	8445507	Sept. 15, 2030	TREATMENT OF NON-METASTATIC, CASTRATION-RESISTANT PROSTATE CANCER (NM-CRPC)
60MG	ERLEADA	JANSSEN BIOTECH	N210951	Feb. 14, 2018	RX	TABLET	ORAL	9481663	June 4, 2033	TREATMENT OF NON-METASTATIC, CASTRATION-RESISTANT PROSTATE CANCER (NM-CRPC)
60MG	ERLEADA	JANSSEN BIOTECH	N210951	Feb. 14, 2018	RX	TABLET	ORAL	9884054	Sept. 23, 2033	TREATMENT OF NON-METASTATIC, CASTRATION-RESISTANT PROSTATE CANCER (NM-CRPC)

Orange Book exclusivity data (new drug applications)

Formulation strength	Trade name	Applicant	Application number	Approval date	Type	Dose form	Route	Exclusivity date	Description
60MG	ERLEADA	JANSSEN BIOTECH	N210951	Feb. 14, 2018	RX	TABLET	ORAL	Feb. 14, 2023	NEW CHEMICAL ENTITY

Figure 6. Patent information and market exclusivity data from the FDA Orange Book.

the APIs in structure-data file (SDF), SMILES and InChI formats, and APIs bioactivity profiles in tabular format.

SUMMARY

DrugCentral has been updated with the latest drug marketing approvals and annotations from external data sources. New data sources have been indexed and cross-referenced to APIs to enable new applications. The database core entities continue to remain pharmaceutical formulation–API–drug target–disease association, with numerous additional annotations around these core entities. DrugCentral has been integrated in 8 public and open resources in 2 years since its initial release and has been visited by nearly 40 000 unique visitors (see also http://drugcentral.org/about).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jeff Knockel, Ph.D., for contributions to the front-end website development of DrugCentral.

FUNDING

National Institutes of Health [U24 CA224370 to UNM and NCATS; U24 TR002278 to UNM and U. Miami; and OT3 HL142479 to UNM, respectively]. Funding for the open access charge: National Institutes of Health [U24 CA224370].

Conflict of interest statement. None declared.

REFERENCES

1. Ursu,O., Holmes,J., Knockel,J., Bologa,C.G., Yang,J.J., Mathias,S.L., Nelson,S.J. and Oprea,T.I. (2017) DrugCentral: online drug compendium. *Nucleic Acids Res.*, **45**, gkw993.
2. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrián-Uhalte,E. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
3. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
4. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem substance and compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.
5. Skuta,C., Popr,M., Muller,T., Jindrich,J., Kahle,M., Sedlak,D., Svozil,D. and Bartunek,P. (2017) Probes & Drugs portal: an interactive, open data resource for chemical biology. *Nat. Methods*, **14**, 759–760.
6. Nguyen,D.-T., Mathias,S., Bologa,C., Brunak,S., Fernandez,N., Gaulton,A., Hersey,A., Holmes,J., Jensen,L.J., Karlsson,A. *et al.* (2016) Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, 1–8.
7. Brown,A.S. and Patel,C.J. (2017) A standard database for drug repositioning. *Sci. Data*, **4**, 170029.
8. Santos,R., Ursu,O., Gaulton,A., Bento,A.P., Donadi,R.S., Bologa,C.G., Karlsson,A., Al-Lazikani,B., Hersey,A., Oprea,T.I. *et al.* (2016) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **16**, 19–34.
9. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A

next generation connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
10. Milletti,F., Storchi,L., Goracci,L., Bendels,S., Wagner,B., Kansy,M. and Cruciani,G. (2010) Extending pKa prediction accuracy: High-throughput pK a measurements to understand pKa modulation of new chemical series. *Eur. J. Med. Chem.*, **45**, 4270–4279.
11. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2017) DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
12. Chambers,J., Davies,M., Gaulton,A., Hersey,A., Velankar,S., Petryszak,R., Hastings,J., Bellis,L., McGlinchey,S. and Overington,J.P. (2013) UniChem: A unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.*, **5**, 3.
13. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
14. Harding,S.D., Sharman,J.L., Faccenda,E., Southan,C., Pawson,A.J., Ireland,S., Gray,A.J.G., Bruce,L., Alexander,S.P.H., Anderton,S. *et al.* (2018) The IUPHAR/BPS Guide to Pharmacology in 2018: Updates and expansion to encompass the new guide to immunopharmacology. *Nucleic Acids Res.*, **46**, D1091–D1106.
15. Roth,B.L., Lopez,E., Patel,S. and Kroeze,W.K. (2000) The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches? *Neurosci.*, **6**, 252–262.
16. Oprea,T.I., Bologa,C.G., Brunak,S., Campbell,A., Gan,G.N., Gaulton,A., Gomez,S.M., Guha,R., Hersey,A., Holmes,J. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.*, **17**, 317–332.
17. Valentin,J.-P. and Hammond,T. (2008) Safety and secondary pharmacology: Successes, threats, challenges and opportunities. *J. Pharmacol. Toxicol. Methods*, **58**, 77–87.
18. Nelson,S.J. (2009) Medical terminologies that work: the example of MeSH. In: *2009 10th Int. Symp.Pervasive Syst. Algorithms, Networks*. IEEE, Kaohsiung.
19. Pihan,E., Colliandre,L., Guichou,J.-F. and Douguet,D. (2012) e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design. *Bioinformatics*, **28**, 1540–1541.
20. Huang,L., Zalkikar,J. and Tiwari,R.C. (2011) A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *J. Am. Stat. Assoc.*, **106**, 1230–1241.
21. Manallack,D.T., Prankerd,R.J., Yuriev,E., Oprea,T.I. and Chalmers,D.K. (2013) The significance of acid/base properties in drug discovery. *Chem. Soc. Rev.*, **42**, 485–496.