Liang *et al. J Cheminform (2019) 11:47*
https://doi.org/10.1186/s13321-019-0370-7

Journal of Cheminformatics

SOFTWARE

Open Access

Bioactivity-explorer: a web application for interactive visualization and exploration of bioactivity data

Lu Liang[1], Chunfeng Ma[1,3], Tengfei Du[1], Yufei Zhao[1,3], Xiaoyong Zhao[1], Mengmeng Liu[1,3], Zhonghua Wang[2*] and Jianping Lin[1,2,3*]

Abstract

To better leverage the accumulated bioactivity data in the ChEMBL database, we have developed Bioactivity-explorer, a web application for interactive visualization and exploration of the large-scale bioactivity data in ChEMBL. Mining and integration of the Therapeutic Target Database disease-target mapping into the ChEMBL database has enabled Bioactivity-explorer to include 493,430 scaffolds, 31,400,000 matched molecular pairs, 1330,220 target–target interactions in terms of shared active compounds, 4526,718 target–target interactions in terms of shared active scaffolds, 97,041,700 molecule–molecule interactions and 14,974 disease-target mappings. This web tool is available at http://cadd.pharmacy.nankai.edu.cn/b17r. The source codes of the front end and back end, released under MIT license, can be found at GitHub.

Keywords: Network pharmacology, Web server, Matched molecular pair, Molecule scaffold, ChEMBL

Introduction

In the past decade, open bioactivity data accumulated rapidly in public data repositories such as ChEMBL [1], BindingDB [2] and PubChem BioAssay [3]. For example, the latest release of the ChEMBL database (v24) contains 12,091 targets, 1828,820 distinct compounds and 15,207,914 activities. With the plethora of publicly available bioactivity data, it becomes increasingly important to extract and explore information of interest in drug discovery, especially information derived from large-scale SAR data, for example, matched molecular pairs (MMPs), which describe how structural transformations affect compound properties, including biological activity [4]; the scaffold distribution of a specific target [5]; and network pharmacology [6], which relates targets in terms of shared active ligands, etc. Many studies have been performed to mine the information underlying experimental data repositories. Ye et al. explored all the scaffolds of compounds from the ChEMBL database [5, 7]. Paolini et al. [8] presented a global mapping of pharmacological space to explore the global relationships between chemical structure and biological targets. By employing the similarity ensemble approach, Keiser et al. related protein pharmacology to ligand chemistry, which was found to reveal unexpected protein target relationships that may be assayed using the ligands themselves. Wirth et al. [9] built an MMP database, named SwissBioisostere, by mining the bioactivity data in ChEMBL. Weber et al. also reported an MMP database with 3D information obtained by mining the data from ChEMBL and PDBbind.

However, to our knowledge, there is no tool or web server that provides comprehensive information mining from public bioactivity repositories, as mentioned above. Among the public bioactivity data repositories, the ChEMBL database hosted by the European Bioinformatics Institute is one of the most widely used data sources because of its well-curated bioactivity data and

*Correspondence: wang_zh@tib.cas.cn; jianpinglin@nankai.edu.cn
[1] State Key Laboratory of Medicinal Chemical Biology, College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Haihe Education Park, 38 Tongyan Road, Tianjin 300353, China
[2] Tianjin Institute of Industrial Biotechnology, Biodesign Center, Chinese Academy of Sciences, Tianjin, China
Full list of author information is available at the end of the article

user-friendly downloading format [1]. Accordingly, we have developed Bioactivity-explorer, a web application for the interactive visualization and exploration of large-scale bioactivity data in the ChEMBL database.

Implementation

Data preparation

The ChEMBL database is the central data repository for Bioactivity-explorer. To maintain consistency with the release pace of the ChEMBL database, we have developed an in-house workflow to help updating data which will be released on GitHub page. Current data were retrieved from the latest ChEMBL (v24) together with the target-diseases data (the ICD-10 diseases classification) from the TTD database [10], in which diseases are mapped to ChEMBL targets through Uniprot ID. For example, there are three targets, "Maltase-glucoamylase, intestinal", "Fibrin" and "Recombinant tetravalent attenuated live dengue", associated with Dengue fever(A90), and of them the target "Maltase-glucoamylase, intestinal" has Uniprot ID "O43451" which was then mapped to Maltase-glucoamylase (CHEMBL2074) in ChEMBL. To mine more information, essentially following our previous work [11], the bioactivity data from ChEMBL were first prepared by the following steps:

(1) Only activities with valid pChEMBL values (negative logarithms of molar IC50, EC50, Ki, etc.) were considered for future usage, such as scaffold activity, target–target network and MMPs generation, and for molecules with multiple potency measurements reported, the aggregate values (including mean, median, maximum and minimum value) were calculated.
(2) The Bemis-Murcko (BM) method [12], which is one of the most widely used scaffold-generation methods, was used to generate scaffolds for all structures. The algorithm implemented in RDKit was employed.
(3) As in the curation procedure for molecule activity, the aggregated BM scaffold [12] activity values were generated from molecules of the scaffold.
(4) The definition of "interaction" of target–target and molecule–molecule were derived from the network pharmacology work by Paolini et al. [8]. Similar ideas were successfully used in the visualization of the chemical space [13] and target prediction [14]. Hu et al. [15] employ a simple definition that two targets are related to each other if they share at least five active compounds. Inspired by the above studies, the target–target and molecule–molecule interaction relationships, which were defined by the number of shared active molecules/targets, were pre-calculated based on the processed molecule/scaffold activities mentioned above. For example, in the current database there are 57 and 143 ligands with valid pChEMBL against Sucrase-isomaltase (CHEMBL2748) and Maltase-glucoamylase (CHEMBL2074), respectively, and 51 of them are in common. This aggregated data was store in database for target–target network view. It should be mentioned that, in the molecule–molecule interaction generation process, for simplicity, only small molecules with more than 10 activities reported were considered, and only molecular pairs that shared more than 5 targets were stored in the database.

For MMP identification of compounds from the same assay, an algorithm implemented in RDKit [16] was employed, and of the generated transformations, the one with most heavy atoms (to maximize the transformation) was selected.

Construction

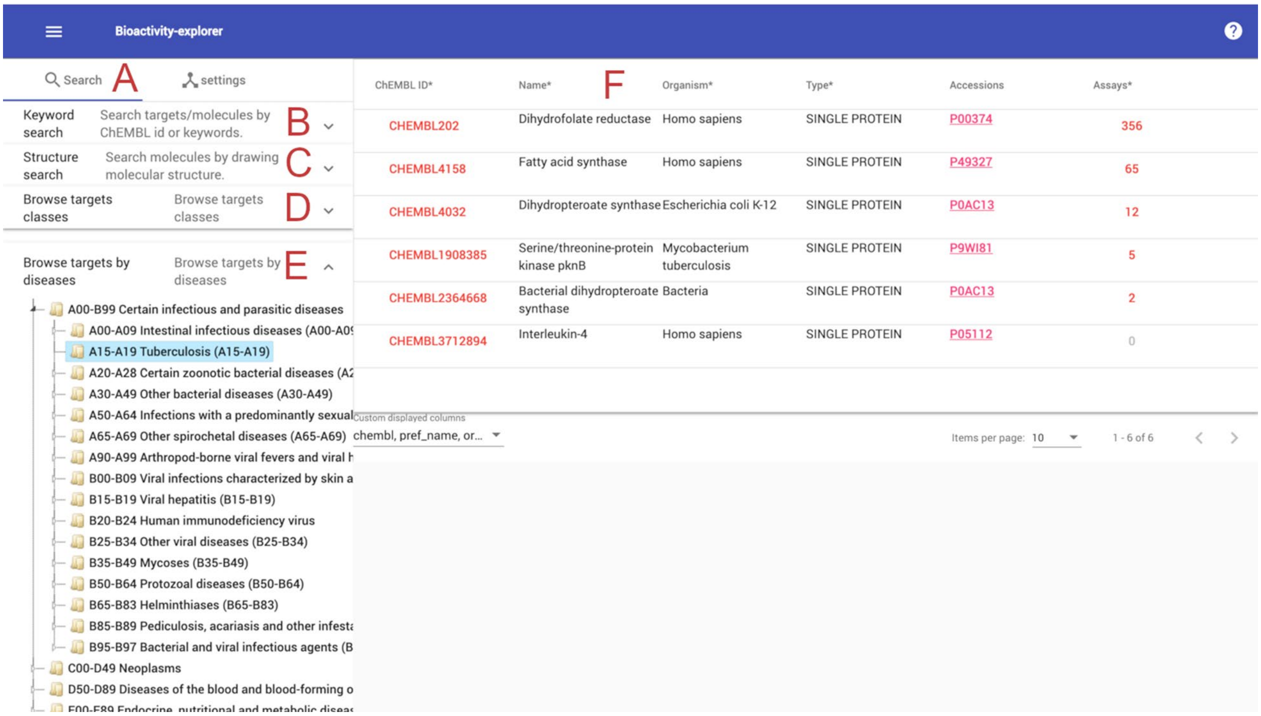
A user-friendly web interface was developed by combining the Angular framework (https://angular.io) and the Django framework (https://www.djangoproject.com/). All data used were stored in a PostgreSQL database. We also implemented a graphql interface (https://graphql.org/), which offers greater query flexibility for the user; a use case is provided in the Results section. The docker files for both the front end and the back end of the tool are provided to facilitate developing, deploying and running the application. Detailed setup instructions can be found at the GitHub repository.

Results

The front end of Bioactivity-explorer was built with modern web technology, Angular (www.angular.io) and Angular Material; therefore, to avoid compatibility problems, we strongly recommend the use of the latest versions of web browsers (Google Chrome, Firefox, etc.). Bioactivity-explorer consists of the following main modules: a target/molecule search panel, a target page, a molecule page, a scaffold page, an assay page and a document page.

Data search and results table

As shown in Fig. 1, Bioactivity-explorer provides a target/molecule keyword search, molecular structure search, and browser targets by classification. Bioactivity-explorer provides two kinds of target classification: (1) the protein classification taken from the official ChEMBL database and (2) the protein-disease classification derived from the latest Therapeutic Target Database (TTD) [10] by mapping targets to the UniProt accession id. The keyword search supports the ChEMBL ID, target and molecule name as input. For the structure search, the JSME



Keyword search — Search targets/molecules by ChEMBL id or keywords.

Structure search — Search molecules by drawing molecular structure.

Browse targets classes — Browse targets classes

Browse targets by diseases — Browse targets by diseases

- A00-B99 Certain infectious and parasitic diseases
  - A00-A09 Intestinal infectious diseases (A00-A0
  - A15-A19 Tuberculosis (A15-A19)
  - A20-A28 Certain zoonotic bacterial diseases (A2
  - A30-A49 Other bacterial diseases (A30-A49)
  - A50-A64 Infections with a predominantly sexual
  - A65-A69 Other spirochetal diseases (A65-A69)
  - A90-A99 Arthropod-borne viral fevers and viral h
  - B00-B09 Viral infections characterized by skin a
  - B15-B19 Viral hepatitis (B15-B19)
  - B20-B24 Human immunodeficiency virus
  - B25-B34 Other viral diseases (B25-B34)
  - B35-B49 Mycoses (B35-B49)
  - B50-B64 Protozoal diseases (B50-B64)
  - B65-B83 Helminthiases (B65-B83)
  - B85-B89 Pediculosis, acariasis and other infesta
  - B95-B97 Bacterial and viral infectious agents (B
- C00-D49 Neoplasms
- D50-D89 Diseases of the blood and blood-forming o
- E00-E89 Endocrine, nutritional and metabolic diseas

ChEMBL ID*	Name*	Organism*	Type*	Accessions	Assays*
CHEMBL202	Dihydrofolate reductase	Homo sapiens	SINGLE PROTEIN	P00374	356
CHEMBL4158	Fatty acid synthase	Homo sapiens	SINGLE PROTEIN	P49327	65
CHEMBL4032	Dihydropteroate synthase	Escherichia coli K-12	SINGLE PROTEIN	P0AC13	12
CHEMBL1908385	Serine/threonine-protein kinase pknB	Mycobacterium tuberculosis	SINGLE PROTEIN	P9WI81	5
CHEMBL2364668	Bacterial dihydropteroate synthase	Bacteria	SINGLE PROTEIN	P0AC13	2
CHEMBL3712894	Interleukin-4	Homo sapiens	SINGLE PROTEIN	P05112	0

Custom displayed columns: chembl, pref_name, or...

Items per page: 10   1 - 6 of 6

Fig. 1 Browsing the target activities of Tuberculosis (A15–A19). Users can perform a keyword search (B), perform a structure search (C), or browse targets by target classification (D) or ICD disease classification (E). Click the menu icon (A) to show or hide the search panel. The table in the right (F) is the result activities. Click ChEMBL ID, accession code and assay number to view corresponding page. Columns followed by "*" are sortable

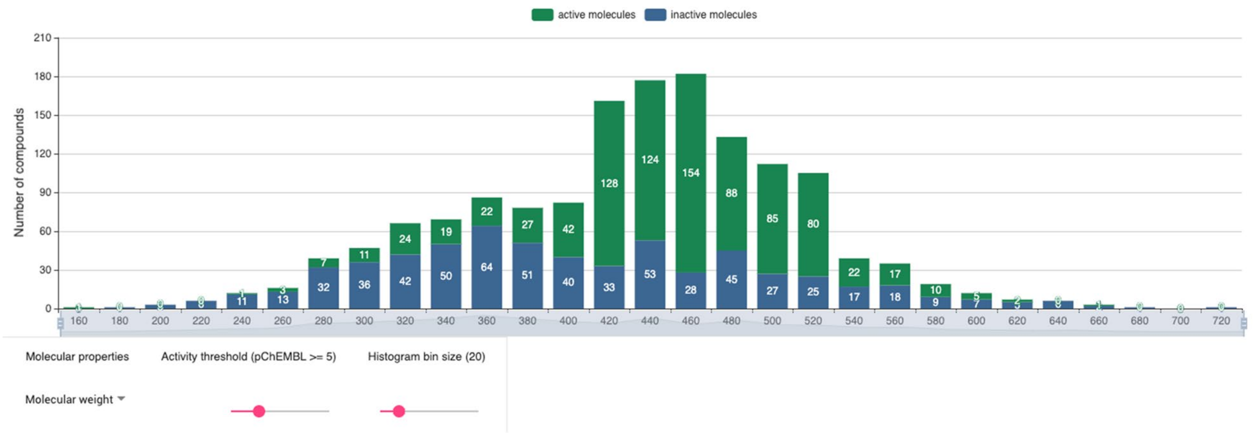
applet [17] was employed for drawing structures, and then structure and substructure searches of molecules or scaffolds can be performed. The Morgan fingerprint (implemented in RDKit) similarity threshold is adjustable from 0.7 to 1, and set to 1 enables an exact structure search.

Target information page

One of the central features of Bioactivity-explorer is the target information page. In addition to the basic information on a target, to improve usability, bioactivities, charts based on bioactivity data, target interaction network, MMPs and active scaffolds are also organized into this page. Compared to the charts of a target at the official ChEMBL site, Bioactivity-explorer provides charts with more functionalities. Taking the ligand property histogram as an example (Fig. 2), the chart in Bioactivity-explorer was further associated with bioactivities. Active and inactive molecules were shown in different colors; therefore, by selecting different activity thresholds and bin sizes, the user can view the compound properties interactively, including the molecular weight, polar surface area (PSA), AlogP, number of heavy atoms, number of aromatic rings, number of rotatable bonds (RTB), number of hydrogen bond donors (HBD) and number of hydrogen bond acceptors (HBA). There is an additional chart about the number of papers published each year on the target, which gives an overview of the research trend of this target.

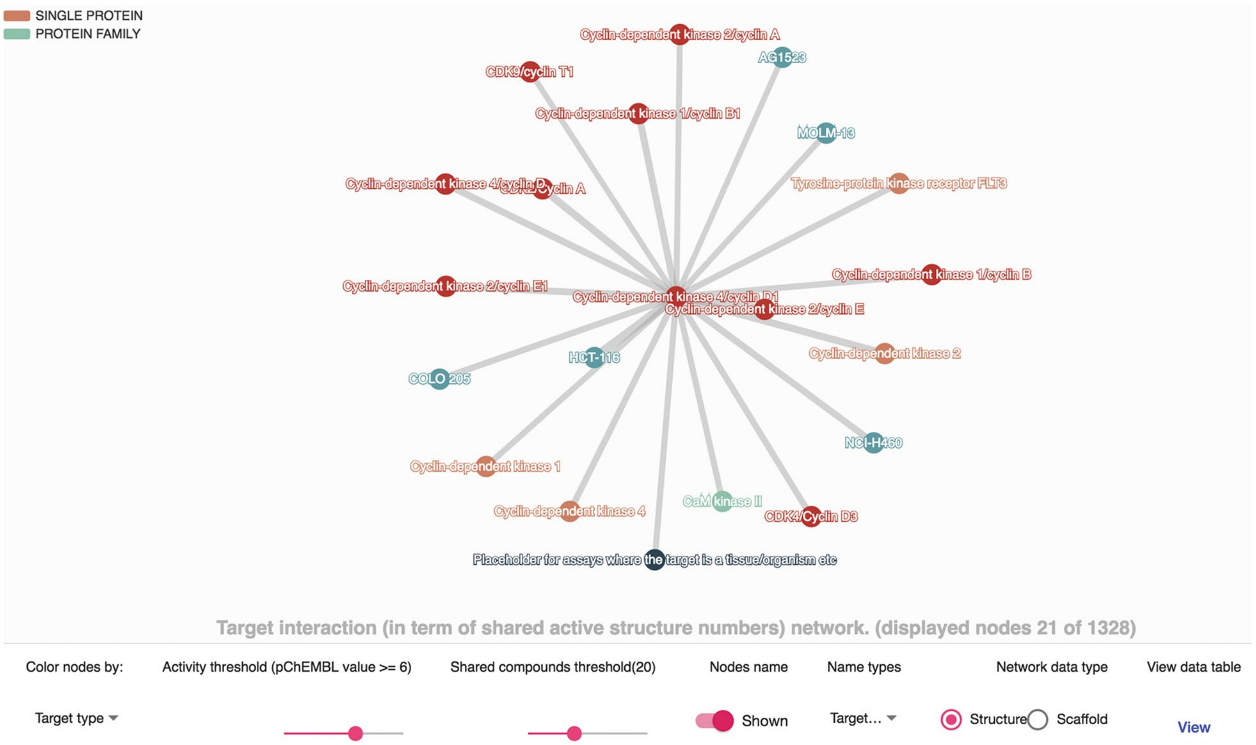
As stated above, the target interaction relationship here was defined by the number of shared active compounds or scaffolds. The "Network" tab on the target page contains a network, as shown in Fig. 3, illustrating these relationships. An edge that connects two nodes in the network denotes that the two targets have active compounds in common, and the width is proportional to the number of shared compounds. To construct an interactive network view, the node color (color by target type or organism), network type (shared compounds or scaffolds), activity threshold and number of shared compound thresholds are also customizable. The node and edge can be double-clicked to view the corresponding target and compound list, respectively. For some targets with a plethora of bioactivity data, it may take seconds to load the network.

The "MMP" tab on the target page contains a table of all MMPs derived from the assays of the target (Fig. 4). Each row of the MMP table is a matched molecular pair and consists of a right handed and left handed molecular structure, structural transformation, activity change, assays and physical property changes including molecular



Molecular properties
Activity threshold (pChEMBL >= 5)
Histogram bin size (20)
Molecular weight ▾

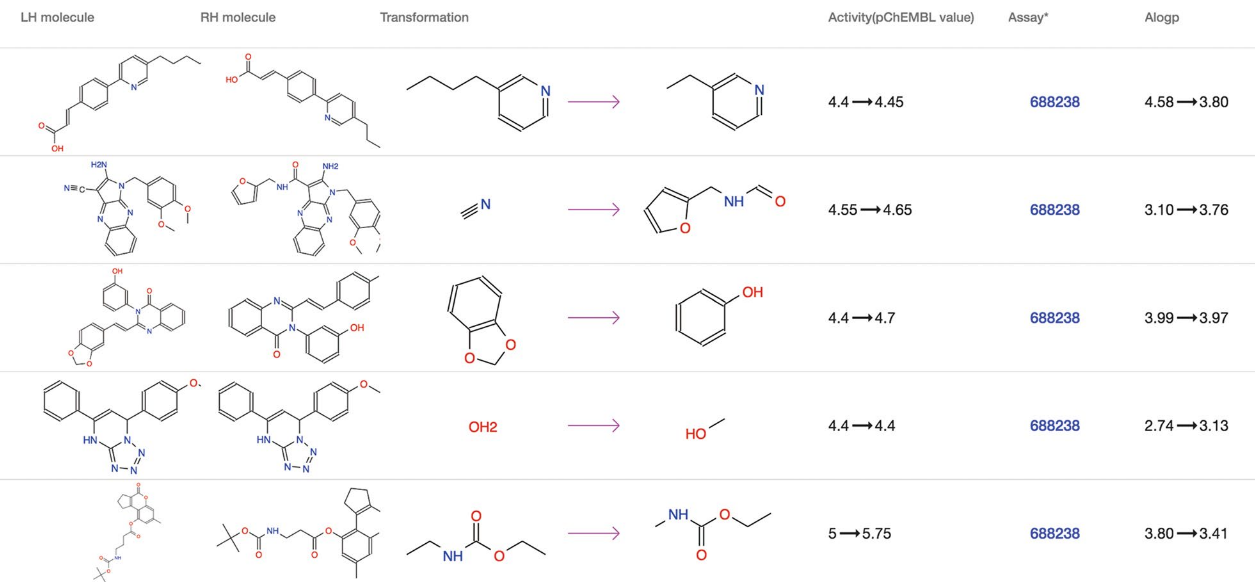
Fig. 2 Target-associated compound property histogram. Active and inactive molecules are shown in different colors. Below the histogram, the molecule property list, activity threshold and bin size threshold slider options could be used to interact with the chart


SINGLE PROTEIN
PROTEIN FAMILY
Cyclin-dependent kinase 2/cyclin A
AG1523
CDK9/cyclin T1
Cyclin-dependent kinase 1/cyclin B1
MOLM-13
Cyclin-dependent kinase 4/cyclin D
CDK2/cyclin A
Tyrosine-protein kinase receptor FLT3
Cyclin-dependent kinase 1/cyclin B
Cyclin-dependent kinase 2/cyclin E1
Cyclin-dependent kinase 4/cyclin D1
Cyclin-dependent kinase 2/cyclin E
HCT-116
Cyclin-dependent kinase 2
COLO 205
NCI-H460
Cyclin-dependent kinase 1
Cyclin-dependent kinase 4
CaM kinase II
CDK4/Cyclin D3
Placeholder for assays where the target is a tissue/organism etc
Target interaction (in term of shared active structure numbers) network. (displayed nodes 21 of 1328)

Color nodes by: Target type ▾
Activity threshold (pChEMBL value >= 6)
Shared compounds threshold(20)
Nodes name: Shown
Name types: Target... ▾
Network data type: ● Structure ○ Scaffold
View data table: View

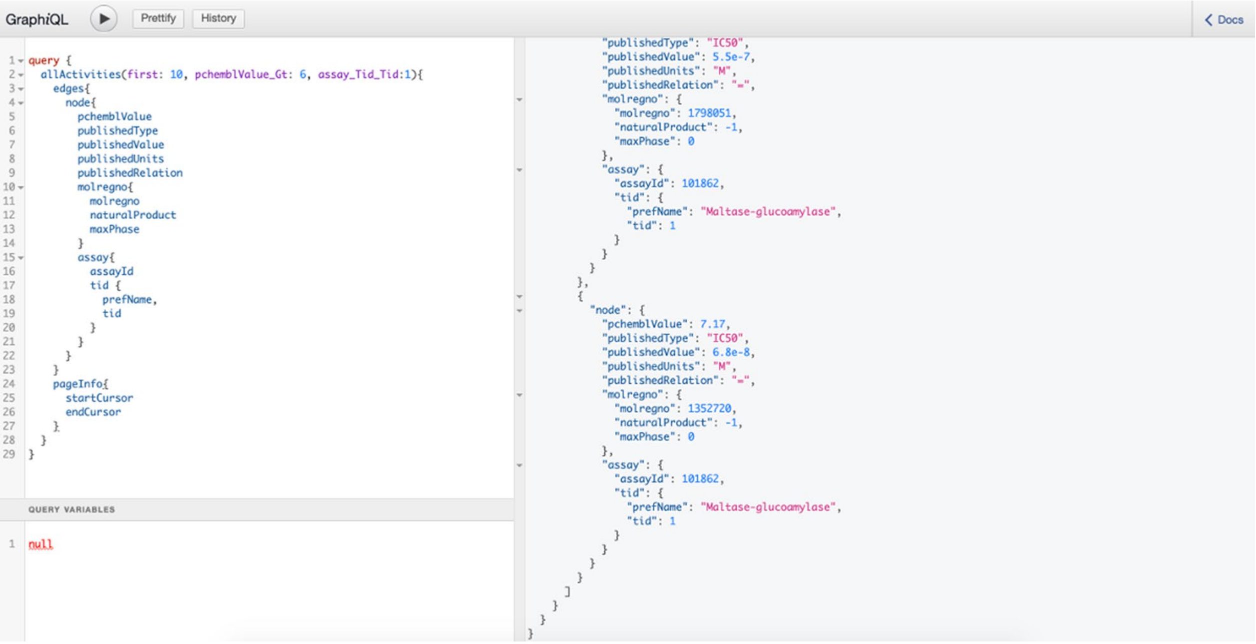
Fig. 3 Target interaction network of Cyclin-dependent kinase 4/cyclin D1. The node and edge denote the target and target–target interaction relationship, and double click on them to view target information and the shared compounds list, respectively. The options below were used to interact with the network

weight, PSA, RTB and Alogp. Finally, the “Scaffold” tab contains all BM scaffolds generated from active molecules of the target, together with the corresponding maximum, mean, median and minimum activity values of the scaffold and the number of underlying compounds.



LH molecule	RH molecule	Transformation	Activity(pChEMBL value)	Assay*	Alogp
			4.4 ⟶ 4.45	688238	4.58 ⟶ 3.80
			4.55 ⟶ 4.65	688238	3.10 ⟶ 3.76
			4.4 ⟶ 4.7	688238	3.99 ⟶ 3.97
			4.4 ⟶ 4.4	688238	2.74 ⟶ 3.13
			5 ⟶ 5.75	688238	3.80 ⟶ 3.41

Fig. 4 An example of the MMP table. Each row of the table represents an MMP, including the left handed (LH) and right handed (RH) molecule, structural transformation, activity and property changes



```
query {
  allActivities(first: 10, pchemblValue_Gt: 6, assay_Tid_Tid:1){
    edges{
      node{
        pchemblValue
        publishedType
        publishedValue
        publishedUnits
        publishedRelation
        molregno{
          molregno
          naturalProduct
          maxPhase
        }
        assay{
          assayId
          tid {
            prefName,
            tid
          }
        }
      }
    }
    pageInfo{
      startCursor
      endCursor
    }
  }
}
```

```
null
```

```
          "publishedType": "IC50",
          "publishedValue": 5.5e-7,
          "publishedUnits": "M",
          "publishedRelation": "=",
          "molregno": {
            "molregno": 1798051,
            "naturalProduct": -1,
            "maxPhase": 0
          },
          "assay": {
            "assayId": 101862,
            "tid": {
              "prefName": "Maltase-glucoamylase",
              "tid": 1
            }
          }
        }
      },
      {
        "node": {
          "pchemblValue": 7.17,
          "publishedType": "IC50",
          "publishedValue": 6.8e-8,
          "publishedUnits": "M",
          "publishedRelation": "=",
          "molregno": {
            "molregno": 1352720,
            "naturalProduct": -1,
            "maxPhase": 0
          },
          "assay": {
            "assayId": 101862,
            "tid": {
              "prefName": "Maltase-glucoamylase",
              "tid": 1
            }
          }
        }
      }
    ]
  }
}
}
```

Fig. 5 The Graphql query interface (graphiql)

A case study exploring data on the target Cyclin-dependent kinase 5 is provided in the Additional file 1.

Other pages were organized very similarly to the target page, except for specific data. For example, the "Related Molecules" tab on the molecule page contains molecules that have more than 5 active targets in common.

The graphql query interface

Graphql is a query language for APIs, and gives the user the power to ask for exactly what they need. The graphql query interface (named graphiq) and the documentation are located at http://cadd.pharmacy.nankai.edu.cn/b17r_api/graphql. Figure 5 demonstrates a graphql query (in the left panel) that retrieves the first 10 activities with a pChEMBL value greater than 6 of the target Maltase-glucoamylase (ChEMBL tid 1). The retrieved results are located in the right panel in JSON format.

Conclusion

We open sourced both the front end and back end of Bioactivity-explorer so that anyone can become directly involved in the development process on GitHub, and researchers who are interested in this tool could build their own local server to use it (which we strongly recommend). In addition to the ChEMBL database, there are still many open and valuable databases that are of great benefit to the drug discovery community; therefore, we expect Bioactivity-explorer to provide an integrated drug development environment for drug discovery and a starting point to integrate more databases.

In summary, based on the ChEMBL database, Bioactivity-explorer consists of 493,430 scaffolds, 31,400,000 MMPs, 1330,220 target–target interactions in terms of shared active compounds, 4526,718 target–target interactions in terms of shared active scaffolds, 97,041,700 molecule–molecule interactions and 14,974 disease-target mappings. For detailed instructions on using Bioactivity-explorer, please refer to the help page.

Additional file

Additional file 1. Case study: Exploring bioactivity data of Cyclin-dependent kinase 5.

Acknowledgements
We thank Dr. Pi Liu for valuable suggestions.

Authors' contributions
LL, ML and TD designed the algorithm for data curation. YZ, LL, XZ, CM and ZW implemented the front end and back end of the tool. ZW, LL and JL wrote the manuscript and discussed the results. All authors read and approved the final manuscript.

Funding
This work was supported by the National Key R&D Program of China (No. 2017YFC1104400).

Availability of data and materials
The front end and back end of the tool can be found at https://github.com/jianping-grp/ng-b17r and https://github.com/jianping-grp/b17r, respectively. The tool is also freely available at http://cadd.pharmacy.nankai.edu.cn/b17r.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] State Key Laboratory of Medicinal Chemical Biology, College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Haihe Education Park, 38 Tongyan Road, Tianjin 300353, China. [2] Tianjin Institute of Industrial Biotechnology, Biodesign Center, Chinese Academy of Sciences, Tianjin, China. [3] Platform of Pharmaceutical Intelligence, Tianjin International Joint Academy of Biomedicine, Tianjin 300000, China.

Received: 25 March 2019 Accepted: 2 July 2019
Published online: 10 July 2019

References

1. Gaulton A, Hersey A, Nowotka M et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954. https://doi.org/10.1093/nar/gkw1074
2. Gilson MK, Liu T, Baitaluk M et al (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053. https://doi.org/10.1093/nar/gkv1072
3. Kim S, Thiessen PA, Bolton EE et al (2016) PubChem substance and compound databases. Nucleic Acids Res 44:D1202–D1213. https://doi.org/10.1093/nar/gkv951
4. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool. J Med Chem 54:7739–7750. https://doi.org/10.1021/jm200452d
5. Hu Y, Stumpfe D, Bajorath J (2011) Lessons learned from molecular scaffold analysis. J Chem Inf Model 51:1742–1753. https://doi.org/10.1021/ci200179y
6. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690. https://doi.org/10.1038/nchembio.118
7. Hu Y, Stumpfe D, Bajorath J (2016) Computational exploration of molecular scaffolds in medicinal chemistry. J Med Chem 59:4062–4076
8. Paolini GV, Shapland RHB, Van Hoorn WP et al (2006) Global mapping of pharmacological space. Nat Biotechnol. https://doi.org/10.1038/nbt1228
9. Wirth M, Zoete V, Michielin O, Sauer WHB (2013) SwissBioisostere: a database of molecular replacements for ligand design. Nucleic Acids Res. https://doi.org/10.1093/nar/gks1059
10. Li YH, Yu CY, Li XX et al (2018) Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx1076
11. Wang Z, Li J, Dang R et al (2015) PhIN: a protein pharmacology interaction network database. CPT Pharmacometrics Syst Pharmacol 4:160–166. https://doi.org/10.1002/psp4.25
12. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928
13. Fechner N, Papadatos G, Evans D et al (2013) ChEMBLSpace-a graphical explorer of the chemogenomic space covered by the ChEMBL database. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts711
14. Keiser MJ, Roth BL, Armbruster BN et al (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol. https://doi.org/10.1038/nbt1284
15. Hu Y, Bajorath J (2010) Polypharmacology directed compound data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. J Chem Inf Model 50:2112–2118. https://doi.org/10.1021/ci1003637
16. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348. https://doi.org/10.1021/ci900450m
17. Bienfait B, Ertl P (2013) JSME: a free molecule editor in JavaScript. J Cheminform. https://doi.org/10.1186/1758-2946-5-24

Publisher's Note