Data in Brief 32 (2020) 106189

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Kinase inhibitor data set for systematic analysis of representative kinases across the human kinome

Oliver Laufkötter[a], Stefan Laufer[b], Jürgen Bajorath[a,*]

[a] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, Bonn D-53115, Germany
[b] Department of Pharmacy and Biochemistry, Pharmaceutical/Medicinal Chemistry, TüCADD (Tübingen Center for Academic Drug Discovery), Eberhard Karls Universität Tübingen, Auf der Morgenstelle 8, Tübingen D-72076, Germany

ARTICLE INFO

Article history:
Received 5 July 2020
Accepted 12 August 2020
Available online 15 August 2020

Keywords:
Human kinome
Multi-kinase inhibitors
Activity annotations
Compound-kinase interactions
Network representations

ABSTRACT

A large set of multi-kinase inhibitors with high-confidence activity data was assembled and used to generate network representations revealing kinase relationships based upon shared inhibitors [1]. Compounds and activity annotations were originally selected from public repositories and organized in an in-house database from which the data set was extracted and curated. The new data set comprises more than 36,000 inhibitors with multiple activity annotations for a total of 420 human kinases (providing 81% coverage of the human kinome), representing a total of ~127,000 kinase-inhibitor interactions. Use of the data is not limited to the network application reported in [1]. It can also be used, for example, for different types of compound promiscuity analysis or machine learning (such a multi-task modeling). In addition, the data set provides a large resource for complementing kinase drug discovery projects with external compound information.

DOI of original article: 10.1016/j.ejmech.2020.112641
* Corresponding author.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

https://doi.org/10.1016/j.dib.2020.106189

Specifications Table

Subject	Drug discovery
Specific subject area	Computational analysis of compounds and activity data to explore inhibitor-based kinase relationships and identify representative kinases.
Type of data	Table Figure
How data were acquired	Data were acquired from a pre-established in-house database [2] and curated for applications using inhibitors with multi-kinase activity.
Data format	Secondary data Table (consistently formated)
Parameters for data collection	The following compound selection criteria were applied: (1) Inhibitors of human kinases, (2) multi-kinase activity, (3) valid SMILES representation [3], (4) standard potency measurements, (5) minimum potency of 1 μM.
Description of data collection	The source database of kinase inhibitors [2], from which the multi-kinase inhibitor data set reported herein was curated, was originally assembled from seven major compound repositories including ChEMBL [4], PubChem [5], Probes and Drugs Portal [6], BindingDB [7], PDBbind [8], ProteomicsDB [9], and Drug Target Commons [10].
Data source location	Department of Life Science Informatics, B-IT, University of Bonn, Endenicher Allee 19c, D-53,115 Bonn, Germany.
Data accessibility	The data set is freely available for download from the public university cloud as a formated data file (csv format) via the following link: https://uni-bonn.sciebo.de/s/reyHRZXYW1D26sq
Related research article	O. Laufkoetter, S. Laufer, J. Bajorath, Identifying representative kinases for inhibitor evaluation via systematic analysis of compound-based target relationships, Eur. J. Med. Chem. 204 (2020) 112641. https://doi.org/10.1016/j.ejmech.2020.112641 [1]

Value of the Data

- The large curated set of multi-kinase inhibitors with 81% coverage of the human kinome provides an extensive resource for promiscuity analysis and the exploration of inhibitor-based kinase relationships. Systematic organization of these relationships enabled the identification of small sets of kinases whose inhibitor binding profiles are representative of many others [1].
- The data set is designed to complement kinase drug discovery projects in academia and the pharmaceutical industry by providing a wealth of inhibitor information for kinases of interest as well as template compounds for multi-kinase drug design.
- Representative kinases can be used as primary targets for experimental evaluation of new inhibitors to estimate their potential for promiscuity across the human kinome. This substantially reduces initial experimental screening efforts. In addition, for individual kinase targets, the data set makes it possible to prioritize other kinases having very similar binding characteristics. These kinases can then be used as likely secondary targets to assess the potential selectivity of newly discovered kinase inhibitors.

1. Data description

The data set contains a total of 127,009 entries including 123,005 unique kinase-inhibitor interactions, 36,628 unique multi-kinase inhibitors, and 420 unique kinase targets. For each

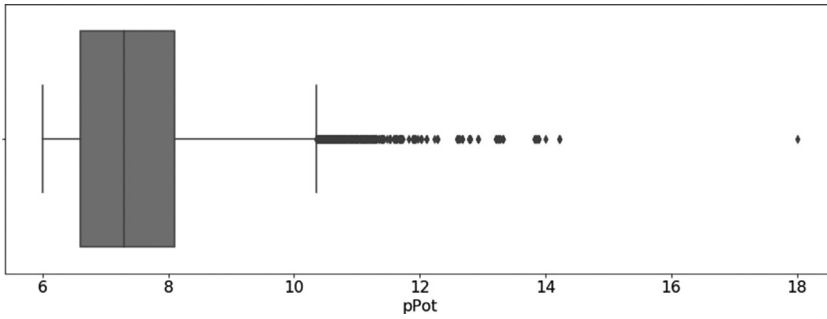

Fig. 1. Distribution of logarithmic potency values of data set compounds. The distribution is reported in a box plot. The vertical black line in the box indicates the median value.

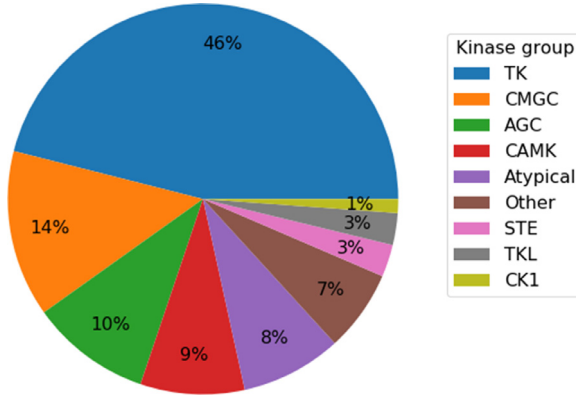

Fig. 2. Distribution of kinase-inhibitor interactions across kinase groups. Nine kinase groups representing the human kinome are listed and color-coded according to the pie chart.

interaction entry, the following information is provided: Compound identifier (CPD_ID), "Non-stereoAromatic" SMILES representation [3] of the inhibitor, standard potency measurement type, logarithmic potency value, full kinase name, standard abbreviation of the name, UniProt ID [11], kinase family of the target, kinase group [12], original source of the inhibitor, and a pan assay interference compound (PAINS) substructure alert, if it was detected using a PAINS filter [13]. The negative decadic logarithm of original potency measurements (IC_{50}, K_i, or K_d), was calculated to yield the "pPot" value. Only inhibitors with pPot $\geq$ 6.0 were retained.

Fig. 1 shows the compound potency distribution within the data set. With a median value > 7.0 more than half of the inhibitors are active in the nanomolar range.

Fig. 2 shows the distribution of kinase-inhibitor interactions over kinase groups. With 46% of the interactions, tyrosine kinases (group TK) dominate the distribution, followed by serine–threonine kinases (CMGC) with 14%. The remaining groups cover 1% - 10% of the interactions.

Fig. 3 shows the distribution of the promiscuity degree (number of kinase annotations) of the inhibitors. The data set contains 32 pan-kinome inhibitors with 100 to 346 kinase annotations. In addition, there are 453 inhibitors with 20 to 99 kinase annotations. However, most multi-kinase inhibitors have low to moderate promiscuity. Specifically, 98.7% (36,143) of the inhibitors comprising the data set are active against at most 10 kinases including 59.1% (21,650) with reported dual-kinase activity. The predominance of multi-kinase inhibitors with low promiscuity degrees mirrors the observed global distribution of promiscuous bioactive compounds [14].

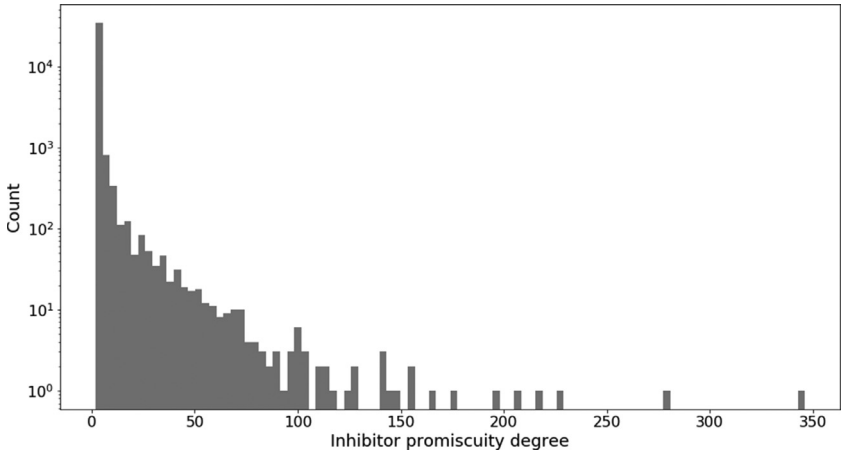

Fig. 3. Promiscuity degree of kinase inhibitors. The histogram reports the number of kinase inhibitors (y-axis, Count, on a logarithmic scale) with increasing promiscuity degree (x-axis, Inhibitor promiscuity degree). The promiscuity degree corresponds to the number of kinases an inhibitor is active against.

2. Experimental design, materials and methods

The source of multi-kinase inhibitor data set was an in-house compiled database of human kinase inhibitors assembled from public repositories comprising 112,624 unique inhibitors with activity against 426 kinases [1]. As potency measurements, IC_{50}, K_i, or K_d values were mostly used. For generating the data set reported herein, we only selected inhibitors with multi-kinase activity applying a general potency threshold of 1 μM (pPot = 6.0), resulting in 36,628 inhibitors with activity against 420 kinases, forming a total of 123,005 unique inhibitor-kinase interactions. The data set is made available in standard .csv format.

All data selection and preparation steps were carried out using a customized Python script covering the following six steps:

(1) Load all required libraries, (2) define functions, (3) load source data, (4) specify filtering variables, (5) Identify multi-target inhibitors and process activity data, (6) generate plots (Fig. 1-3).

The Python code is also made freely available together with the data set.

Ethics statement

This is a secondary data set and thus did not involve any human or animal testing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgment

The authors are grateful to Filip Miljković for assembling the in-house source database.

References

[1] O. Laufkoetter, S. Laufer, J. Bajorath, Identifying representative kinases for inhibitor evaluation via systematic analysis of compound-based target relationships, Eur. J. Med. Chem. 204 (2020) 112641.

[2] F. Miljković, J. Bajorath, Computational analysis of kinase inhibitors identifies promiscuity cliffs across the human kinome, ACS Omega 3 (2018) 17295–17308.

[3] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci 28 (1988) 31–36.

[4] A. Gaulton, A. Hersey, M. Nowotka, A.P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M.P. Magariños, J.P. Overington, G. Papadatos, I. Smit, A.R. Leach, The ChEMBLDatabase in 2017, Nucleic Acids Res. 45 (2017) D945–D954.

[5] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang, S.H. Bryant, PubChem substance and compound databases, Nucleic Acids Res. 44 (2016) D1202–D1213.

[6] C. Skuta, M. Popr, T. Muller, J. Jindrich, M. Kahle, D. Sedlak, D. Svozil, P. Bartunek, Probes & drugs portal: an interactive, opendata resource for chemical biology, Nat. Meth. 14 (2017) 759–760.

[7] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. 44 (2016) D1045–D1053.

[8] R. Wang, X. Fang, Y. Lu, C.Y. Yang, S. Wang, The PDBbind database: methodologies and updates, J. Med. Chem. 48 (2005) 4111–4119.

[9] T. Schmidt, P. Samaras, M. Frejno, S. Gessulat, M. Barnert, H. Kienegger, H. Krcmar, J. Schlegl, H.C. Ehrlich, S. Aiche, B. Kuster, M. Wilhelm, ProteomicsDB, Nucleic Acids Res. 46 (2018) D1271–D1281.

[10] J. Tang, Z.-U.-R. Tanoli, B. Ravikumar, Z. Alam, A. Rebane, M. Vähä-Koskela, G. Peddinti, A.J. van Adrichem, J. Wakkinen, A. Jaiswal, E. Karjalainen, P. Gautam, L. He, E. Parri, S. Khan, A. Gupta, M. Ali, L. Yetukuri, A.-.L. Gustavsson, B. Seashore-Ludlow, A. Hersey, A.R. Leach, J.P. Overington, G. Repasky, K. Wennerberg, T. Aittokallio, Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions, Cell. Chem. Biol. 25 (2018) 224–229.

[11] UniProt Consortium, UniProt, A hub for protein information, Nucleic Acids Res. 43 (2002) D204–D212.

[12] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S, The protein kinase complement of the human genome, Science 298 (2002) 1912–1934.

[13] J.B. Baell, G. A.Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, J. Med. Chem. 53 (2010) 2719–2740.

[14] Y. Hu, J. Bajorath, Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited, Future Sci. OA 3 (2017) FSO179.