RESEARCH ARTICLE

[View Article Online](#)
[View Journal](#) | [View Issue](#)

Cite this: *Med. Chem. Commun.*,
2017, 8, 2100

Received 18th August 2017,
Accepted 20th October 2017

DOI: 10.1039/c7md00426e

rsc.li/medchemcomm

Dark chemical matter in public screening assays and derivation of target hypotheses

Swarit Jasial and Jürgen Bajorath *

Compounds that are consistently inactive in many screening assays, so-called dark chemical matter (DCM), have recently experienced increasing attention. One of the reasons is that many DCM compounds may not be fully inert biologically, but may provide interesting leads for obtaining compounds that are highly selective or active against unusual targets. In this study, we have systematically identified DCM among extensively assayed screening compounds and searched for analogs of these compounds that have known bioactivities. Analog series containing DCM and known bioactive compounds were generated on a large scale, making it possible to derive target hypotheses for more than 8000 extensively assayed DCM molecules.

Introduction

High-throughput screening (HTS) plays a critically important role in early-phase drug discovery as the primary source of new active compounds and starting points for medicinal chemistry.¹ Given current standards in the pharmaceutical industry, millions of compounds are often subjected to screening campaigns. Striving for chemical diversity and broad chemical space coverage and focusing on specific bioactivities continue to be primary design strategies for screening libraries.^{2–4} The major goal of library design is maximizing the number of high-quality hits. However, it has also been observed that significant numbers of compounds in screening decks were mostly or consistently inactive in assays they were tested in.^{5,6} In a milestone contribution analyzing in-house screening data of a major pharmaceutical company as well as screens carried out in the context of the NIH molecular libraries initiative,⁷ such consistently inactive compounds have been termed ‘dark chemical matter’ (DCM).⁶ In HTS, DCM provides a sharp contrast to molecules with true multi-target activities^{8,9} and assay interference compounds,^{10–14} which plague screening campaigns and medicinal chemistry programs. The DCM study showed that more than a third of the compounds tested in at least 100 NIH library program assays were consistently inactive.⁶ Furthermore, 14% of the compounds in a large pharmaceutical screening deck were inactive in at least 100 in-house assays.⁶ In the latter case, weak activities were also taken into consideration, providing an explanation for the observed discrepancy in the proportion of DCM between external and in-house screens. As

one would expect, DCM molecules were often smaller, less aromatic, and more soluble than other screening compounds. However, despite the lack of activity in large numbers of assays, at least some DCM molecules might also have a brighter side. Wassermann *et al.* confirmed that selected DCM compounds were active in additional assays. When evaluated in off-the-beaten-path assays, including novel targets, DCM compounds frequently yielded attractive hits. These findings led to the conclusion that DCM might not be entirely inert biologically, but may frequently have the potential to display specific activities.⁶ Thus, DCM compounds may or may not be consistently inactive. It follows that DCM should be of considerable interest in the search for chemical entities having high target selectivity or unusual activities. To these ends, structural relationships between DCM compounds and active molecules might be explored to derive target hypotheses for DCM.

Herein, we report a large-scale computational analysis with two primary goals. First, a systematic search for DCM in extensively tested screening compounds was carried out to identify all currently available DCM compounds. Publicly available assay data were collected and analyzed. Second, after identifying DCM molecules it was attempted to derive target hypotheses for them by systematically evaluating structural relationships to known bioactive compounds with available high-confidence activity data and generating analog series. The results of our analysis are reported in the following.

Methods and materials

Extensively assayed PubChem compounds

From the PubChem BioAssay database,¹⁵ compounds tested in both primary (percentage of inhibition from a single dose) and confirmatory assays (dose-response titration yielding IC₅₀ values) were selected.² A total of 437 257 screening

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany. E-mail: bajorath@bit.uni-bonn.de; Fax: +49 228 2699 341; Tel: +49 228 2699 306

compounds were obtained.⁹ For DCM analysis, PubChem compounds were selected that were tested in at least 100 primary assays and did not display activity in any primary or confirmatory assay.

ChEMBL compounds with high-confidence activity data

From ChEMBL¹⁶ release 22, compounds with available high-confidence activity data were selected. Qualifying compounds were required to form direct interactions (relationship type “D”) with human targets at the highest confidence level (confidence score 9). Furthermore, two types of potency measurements were considered including equilibrium constants (K_i) and IC_{50} values. Only compounds having numerically specified K_i or IC_{50} values were accepted and those with approximate measurements such as “>”, “<”, or “~” were discarded. Moreover, PubChem and ChEMBL compounds with PAINS substructures^{16–18} or aggregation potential¹⁹ were removed.

Identification of analog series

From DCM and ChEMBL compounds, analog series were extracted using a recently introduced method²⁰ based upon the matched molecular pair (MMP) concept.²¹ An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site, termed a transformation.²² For MMP generation, random fragmentation of exocyclic single bonds²² was replaced by fragmentation according to retrosynthetic rules,²³ generating so-called RECAP-MMPs.²⁴ Transformation size restrictions were applied to limit chemical changes to those typically observed in series of analogs.²⁵ On the basis of RECAP-MMPs, analog series were systematically generated and series containing DCM compounds from PubChem and bioactive analogs from ChEMBL were selected. Ligand-based target prediction has mostly been carried out on the basis of statistically supported Tanimoto similarity calculations.^{26,27} Compared to such whole-molecule similarity assessment, we give preference to the detection of analog relationships, which provide a more conservative assessment of structural relationships on the basis of which target hypotheses might be inferred.

All calculations reported herein were carried out using in-house scripts with the aid of a chemistry toolkit.²⁸

Results and discussion

Dark chemical matter

We identified 367 557 screening compounds from PubChem that were tested in at least 100 primary assays. For these compounds, all primary and confirmatory assay records were analyzed and 81 597 unique compounds were found to be consistently inactive in all primary and confirmatory assays they were tested in. These compounds represented an – at least to us – unexpectedly large DCM subset.

Assay frequency

For the 81 597 DCM compounds, assay frequency was determined, as reported in Fig. 1. On average, these compounds were tested in 339 primary and 86 confirmatory assays, with median values of 339 and 88 assays, respectively. Thus, DCM from PubChem was extensively tested in both primary and confirmatory assays, yet the compounds were consistently inactive.

Overlap between PubChem and ChEMBL

As a control, we mapped all DCM compounds from PubChem to ChEMBL. With 310 compounds, a minute proportion of 0.38% of DCM was detected in ChEMBL. These 310 compounds were annotated with one to 17 targets on the basis of high-confidence activity data, although they were consistently inactive in hundreds of PubChem screening assays. These

Fig. 1 Assay frequency distribution for DCM. Histograms show the distribution of (a) primary and (b) confirmatory assays in which DCM compounds from PubChem were tested.

findings provided a hint that it might be possible to derive target hypotheses for other DCM compounds by exploring narrowly confined chemical space around them.

Searching for analog series

Therefore, we systematically searched for analog series consisting of PubChem DCM and ChEMBL compounds with available high-confidence activity data. The underlying rationale was that the presence of analogs of DCM in ChEMBL might provide target hypotheses for these DCM compounds, taking into consideration that structurally very similar compounds often interact with the same target(s). As reported in Table 1, an unexpectedly large number of 1400 DCM/ChEMBL analog series was identified. These series contained a total of 14 796 analogs and included 8568 DCM compounds. Thus, for 10.5% of DCM, ChEMBL analogs with high-confidence target annotations were identified. These analogs were active against a total of 613 targets. Fig. 2 shows the compound and target distribution of these series. Statistics are reported in Table 1. The median size of the series was three compounds but series with up to 20 analogs were frequently detected. About half of the series were annotated with a single target but series with up to five targets were also frequently found. Hence, many series were available to compare DCM and ChEMBL analogs and deduce target hypotheses for DCM.

Exemplary series

Fig. 3 shows different examples of analog series containing DCM and ChEMBL compounds. In Fig. 3a, four DCM analogs are shown that were tested in more than 400 to 600 assays. This series contained a known thrombin inhibitor from ChEMBL. Given the high degree of structural similarity of these analogs, the DCM compounds should be tested for thrombin inhibition. If one or another analog would indeed be a thrombin inhibitor, it might be rather selective, given the inactivity of DCM analogs in very large numbers of assays. However, since only one bioactive analog was available in this case, attention must be paid to its activity records to exclude potential artifacts. This represents a prime reason for exclusively considering compounds with high-confidence activity data for analog series. In Fig. 3b, a series is shown

Fig. 2 Size and target distribution of analog series. For analog series including DCM and ChEMBL compounds, the (a) size and (b) target distribution is reported. For each series, the total number of unique targets of ChEMBL analogs was determined.

that consisted of a small DCM and larger ChEMBL analogs with activity against serotonin receptor isoforms. The small DCM analog lacked the tertiary amine, a hallmark for serotonin receptor activity.

Nonetheless, it is striking that this small DCM compound was inactive in all 357 assays it was tested in. In Fig. 3c, a series with two closely related DCM and three ChEMBL analogs is shown that were active against the dopamine D2/D4 receptor. In this case, chemical changes were confined to a terminal phenyl ring, revealing some puzzling observations. For example, the difference between a DCM compound and a D2 and D2/D4 receptor ligand was the change of a *para*-fluoro to an *ortho*-chloro and *ortho*-methoxy substituent, respectively. An unsubstituted phenyl ring was present in the other DCM compound. Hence, structure–activity relationships and DCM character should be further explored here. Fig. 3d shows two DCM analogs that were inactive in more than 500 and 600 assays, respectively, and two ChEMBL analogs with activity against HSP 90 and different PI3/4 kinase subunits, respectively. In addition, Fig. 3e depicts a subset of a series consisting of four DCM and two ChEMBL analogs with activity against pairs of distinct targets including novel target proteins. Taken together, these examples highlight other opportunities for deriving target hypotheses for compounds with DCM character.

Table 1 Analog series containing DCM and ChEMBL compounds

1400 analog series	
Total number of compounds	14 796
Number of unique targets	613
Number of ChEMBL compounds	6228
Compounds per series	2–754
Median of compounds per series	3
Targets per series	1–74
Median of targets per series	1

Compound and target statistics are provided for 1400 analog series consisting of DCM and ChEMBL compounds.

Fig. 3 Exemplary analog series. In (a)–(e), different examples of series containing DCM and ChEMBL analogs are presented. For DCM and ChEMBL compounds, assay statistics and target annotations are provided, respectively.

Conclusions

Herein we have reported a systematic analysis of DCM from public screening assays. From a large pool of extensively assayed compounds, more than 81 000 chemical entities were identified that were consistently inactive in all primary and confirmatory assays in which they were tested. There are multiple possible reasons for inactivity in assays, one of which is the lack of compound quality or stability. However, given the very large number of DCM compounds that were identified, consistent lack of activity could hardly be in general attributed to compound quality or concentration issues. Single instances likely exist, but DCM character prevails on a large scale. Identification of DCM was followed by a systematic search for bioactive analogs. For more than 8000 of these DCM compounds, varying numbers of ChEMBL compounds were identified, making it possible to evaluate potential targets for DCM. A variety of analog series with interesting composition were obtained also including series with multiple DCM and ChEMBL analogs having activity against well-studied pharmaceutical targets. Thus, DCM might not only fill niche positions in target space. The analog series we identified provide starting points for further exploring the assay behavior of DCM compounds, comparing them directly to known active analogs, and deriving new experimentally testable target hypotheses. Therefore, as a part of our study, the large number of series containing DCM and bioactive analogs is made freely available as an open access deposition.²⁹

Conflicts of interest

The authors declare no competing interest.

Acknowledgements

We thank the OpenEye Free Academic Licensing Program for providing an academic license for the chemistry toolkit.

References

- 1 R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer and G. S. Sittampalam, *Nat. Rev. Drug Discovery*, 2011, **10**, 188–195.
- 2 A. A. Shelat and R. K. Guy, *Nat. Chem. Biol.*, 2007, **3**, 442–446.
- 3 M. E. Welsch, S. A. Snyder and B. R. Stockwell, *Curr. Opin. Chem. Biol.*, 2010, **14**, 347–361.
- 4 P. J. Hajduk, J. Philip, W. R. J. D. Galloway and D. R. Spring, *Nature*, 2011, **470**, 42–43.
- 5 P. M. Petrone, A. M. Wassermann, E. Lounkine, P. Kutchukian, B. Simms, J. Jenkins, P. Selzer and M. Glick, *Drug Discovery Today*, 2013, **18**, 674–680.
- 6 A. M. Wassermann, E. Lounkine, D. Hoepfner, G. Le Goff, F. J. King, C. Studer, J. M. Peltier, M. L. Grippo, V. Prindle, J. Tao, A. Schuffenhauer, I. M. Wallace, S. Chen, P. Krastel, A. Cobos-Correa, C. N. Parker, J. W. Davies and M. Glick, *Nat. Chem. Biol.*, 2015, **11**, 958–966.

- 7 C. P. Austin, L. S. Brady, T. R. Insel and F. S. Collins, *Science*, 2004, **306**, 1138–1139.
- 8 Y. Hu and J. Bajorath, *Drug Discovery Today*, 2013, **18**, 644–650.
- 9 S. Jasial, Y. Hu and J. Bajorath, *PLoS One*, 2016, **11**, e0153873.
- 10 S. L. McGovern, E. Caselli, N. A. Grigorieff and B. K. Shoichet, *J. Med. Chem.*, 1996, **45**, 1712–1722.
- 11 B. K. Shoichet, *Drug Discovery Today*, 2006, **11**, 607–615.
- 12 J. B. Baell and G. A. Holloway, *J. Med. Chem.*, 2010, **53**, 2719–2740.
- 13 J. Baell and M. A. Walters, *Nature*, 2014, **513**, 481–483.
- 14 J. W. M. Nissink and S. Blackburn, *Future Med. Chem.*, 2014, **6**, 1113–1126.
- 15 Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant, *Nucleic Acids Res.*, 2012, **40**, D400–D412.
- 16 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
- 17 RDKit, 2013, <http://www.rdkit.org>.
- 18 T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.
- 19 J. J. Irwin, D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian and B. K. Shoichet, *J. Med. Chem.*, 2015, **58**, 7076–7087.
- 20 D. Stumpfe, D. Dimova and J. Bajorath, *J. Med. Chem.*, 2016, **59**, 7667–7676.
- 21 E. Griffen, A. G. Leach, G. R. Robb and D. J. Warner, *J. Med. Chem.*, 2011, **54**, 7739–7750.
- 22 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.
- 23 X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 511–522.
- 24 A. de la Vega de León and J. Bajorath, *Med. Chem. Commun.*, 2014, **5**, 64–67.
- 25 X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1138–1145.
- 26 M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin and B. K. Shoichet, *Nat. Biotechnol.*, 2007, **25**, 197–206.
- 27 M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijter, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet and B. L. Roth, *Nature*, 2009, **462**, 175–181.
- 28 OEChem TK, OpenEye Scientific Software, Inc., Santa Fe, NM, 2012.
- 29 <https://doi.org/10.5281/zenodo.890619>.