RESEARCH ARTICLE

Towards a systematic assessment of assay interference: Identification of extensively tested compounds with high assay promiscuity [version 1; referees: 3 approved]

Erik Gilberg, Dagmar Stumpfe, Jürgen Bajorath

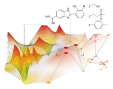
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Bonn, D-53113, Germany

v1 **First published:** 17 Aug 2017, **6**(Chem Inf Sci):1505 (doi: 10.12688/f1000research.12370.1)

Latest published: 17 Aug 2017, **6**(Chem Inf Sci):1505 (doi: 10.12688/f1000research.12370.1)

Abstract

A large-scale statistical analysis of hit rates of extensively assayed compounds is presented to provide a basis for a further assessment of assay interference potential and multi-target activities. A special feature of this investigation has been the inclusion of compound series information in activity analysis and the characterization of analog series using different parameters derived from assay statistics. No prior knowledge of compounds or targets was taken into consideration in the data-driven study of analog series. It was anticipated that taking large volumes of activity data, assay frequency, and assay overlap information into account would lead to statistically sound and chemically meaningful results. More than 6000 unique series of analogs with high hit rates were identified, more than 5000 of which did not contain known interference candidates, hence providing ample opportunities for follow-up analyses from a medicinal chemistry perspective.



This article is included in the Chemical Information Science gateway.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 17 Aug 2017	✓ report	✓ report	✓ report

1 **John A. Lowe III**, Jl3pharma LLC, USA

2 **José L Medina-Franco**, National Autonomous University of Mexico, Mexico **Fernanda I. Saldívar-González**, National Autonomous University of Mexico, Mexico

3 **Michael Walters**, University of Minnesota, USA

Discuss this article

Comments (2)

Corresponding author: Jürgen Bajorath (bajorath@bit.uni-bonn.de)

Author roles: Gilberg E: Data Curation, Formal Analysis, Methodology, Writing – Review & Editing; **Stumpfe D**: Data Curation, Formal Analysis, Methodology, Writing – Review & Editing; **Bajorath J**: Conceptualization, Formal Analysis, Methodology, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were declared.

How to cite this article: Gilberg E, Stumpfe D and Bajorath J. **Towards a systematic assessment of assay interference: Identification of extensively tested compounds with high assay promiscuity [version 1; referees: 3 approved]** *F1000Research* 2017, **6**(Chem Inf Sci):1505 (doi: 10.12688/f1000research.12370.1)

Grant information: DS is supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 17 Aug 2017, **6**(Chem Inf Sci):1505 (doi: 10.12688/f1000research.12370.1)

Introduction

Compounds with false-positive signals in biological assays cause substantial problems for biological screening and medicinal chemistry[1]. Assay artifacts often remain undetected or are unveiled only at later stages of compound development efforts, leading to substantial loss of time and resources. Moreover, once published, artificial activities spread through the scientific literature and potentially cause even more harm by inspiring follow-up investigations that are doomed to fail. Known assay interference compounds include colloidal aggregators[2–7] and many other compound classes that can react in different ways or are fluorescent under assay conditions[6–15]. Systematic efforts to identify interference compounds include the compilation of aggregators[2–4] and pan-assay interference compounds (PAINS)[8,9]. The latter comprise a set of 480 classes of compounds originally identified in AlphaScreen assays[8]. PAINS are typically contained as substructures in larger compounds. However, the assessment and prediction of assay interference is far from being a trivial exercise. For example, analysis of screening data from PubChem[16] has revealed that many compounds containing PAINS, including most reactive chemical entities, have very different hit rates or might be consistently inactive[17,18]. Moreover, analogs or different series of analogs containing the same PAINS substructure often have distinct activity profiles and are active against different targets[19]. Thus, interference characteristics of related compounds frequently differ and a substructure with interference potential does not necessarily give rise to false-positive assay signals. To further complicate matters, promiscuous compounds may also have true multi-target activities[20] that are relevant for polypharmacology[20–22]. Moreover, even highly promiscuous screening hits include molecules with no apparent liabilities, in addition to obvious interference compounds[12].

Without doubt, judging assay interference and candidate compounds requires profound chemical knowledge and experience. It is equally relevant, however, to strive for a data-driven assessment of promiscuity by exploring compound activity data on a large scale[20], aiming to identify compounds with interference potential for further analysis. Therefore, we have carried out a statistical analysis of hit rates of compounds that were extensively tested in screening assays. A special feature of this study has been its focus on pairs or larger series of analogs, rather than single compounds, which provides additional confidence criteria for activity assessment and further increases the information content of activity data analysis. Many series of analogs with much higher than typically observed hit rates and largely consistent activity profiles across many different assays were identified. This collection of series provides a basis for further investigating compounds with interference potential or true multi-target activities.

Methods

All calculations were carried out using in-house scripts and implementations.

Compounds

From the PubChem BioAssay database[16], 437,257 compounds were pre-selected that were tested in both primary and confirmatory assays, representing extensively assayed screening compounds[23]. Approximately 95% of these compounds were evaluated in more than 50 primary and/or confirmatory assays[23]. Primary PubChem assays report compound activity (e.g., percentage activity) for a single dose, while confirmatory assays are dose-response assays yielding titration curves and IC_{50} values. Our current analysis focused on primary assays, for which much larger data volumes were available than for confirmatory assay. Primary assays also included assays for which no target was specified (such as cell-based assays). For pre-selected compounds, hit rate statistics were determined.

Matched molecular pairs and series

A matched molecular pair (MMP) is a pair of compounds that are only distinguished by a chemical change at a single site[24], termed a chemical transformation[25]. As an extension of the MMP concept, a matched molecular series (MMS) was defined as the union of all MMP compounds that are only distinguished by chemical modifications at a given site[26]. Accordingly, an MMS represents a series of analogs sharing a single substitution site. To generate MMPs, exocyclic single bonds in screening compounds were systematically fragmented[25] following retrosynthetic fragmentation rules[27], yielding so-called RECAP-MMPs[28]. These MMPs were subject to transformation size restrictions in order to limit chemical changes to modifications typically observed in series of analogs[29]. An MMS was designated as redundant if it was a subset of a larger MMS or if there was another MMS representing the same series of analogs but having a larger MMP core. For screening compounds with high hit rates, non-redundant MMS were systematically determined.

MMS parameters

For each MMS, three parameters were calculated. First, the *MMS hit rate (HR)* was obtained from the union of all assays (i.e., the number of unique assays in which one or more analogs were tested in) and assays with activity signals (active assays, i.e., the number of unique assays in which one or more analogs were found to be active). Second, *assay overlap* was determined as the proportion of assays in which all MMS compounds were tested in (shared assays, i.e., the intersection of assays) relative to the union of assays. Third, from assay overlap, *assays with inconsistent activity* were calculated as the proportion of shared assays in which different MMS compounds were active or inactive.

Results and discussion

Study design

A statistical analysis of hit rates of extensively tested screening compounds is presented taking assay frequency into account. On the basis of the hit rate distribution, ranges of unusually high hit rates were determined. From compounds with high hit rates, analog series with single substitution sites (MMSs), i.e., “minimal” chemical modifications within series, were systematically extracted, which provided structural context information and hit rate controls for closely related compounds. For MMSs, different parameters were calculated, making it possible to compare and prioritize these series. The collection of MMSs with high hit rates provides a basis for investigating assay interference candidates, as well as chemical entities with potential multi-target activities.

Source compounds and assay distribution

Figure 1 (boxplot on the left) shows the global distribution of primary assays for 437,257 extensively tested PubChem compounds, with a median value of 347 assays per compound. From these, a subset of 327,532 compounds was selected that were tested in more than 257 primary assays, corresponding to the lower quartile boundary of the global distribution. For this subset, the assay distribution was separately monitored (Figure 1, boxplot on the right), yielding a median of 426 (and a maximum of 626) assays per compound. Hence, half of these compounds were tested in more than 426 primary assays.

Hit rate distribution

For 327,532 compounds tested in more than 257 assays, hit rates were determined. The distribution is reported in Figure 2 (boxplot on the left), resulting in a median hit rate of 0.4%. The

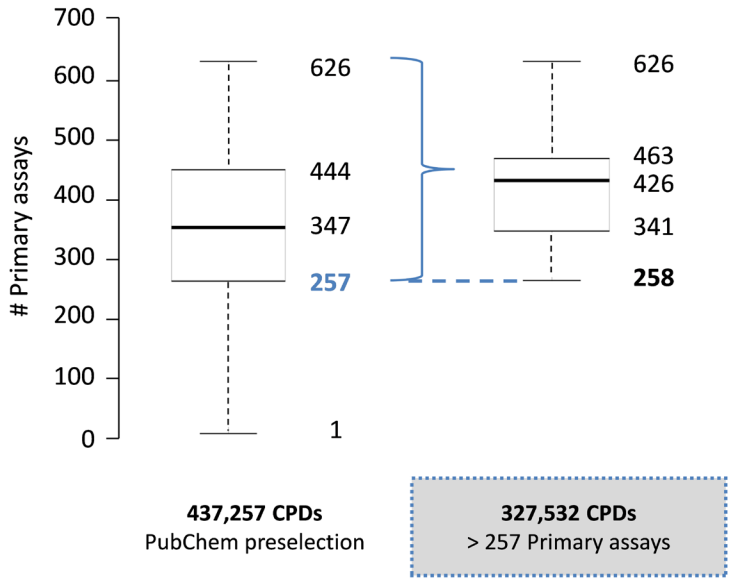


Figure 1. Assay frequency distribution. The frequency distribution of primary assays is shown in a boxplot format for 437,257 pre-selected PubChem compounds and a subset of 327,532 compounds. The plot gives the smallest number of primary assays (lower whisker), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), and largest number of assays (upper whisker). Outliers are not displayed. The dashed blue line indicates the selection criterion for the compound subset (i.e., tested in more than 257 primary assays).

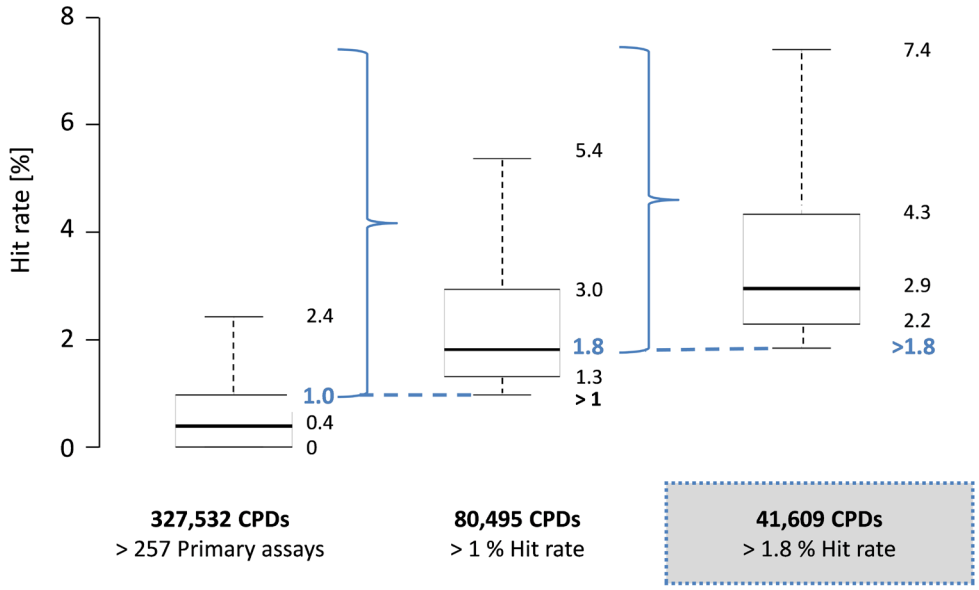


Figure 2. Hit rate distribution. For three different subsets of PubChem compounds, hit rate distributions are shown in boxplots according to Figure 1. The subsets are characterized by increasing hit rates (marked by dashed blue lines).

lower quartile boundary and lower whisker of the boxplot were identical and represented consistently inactive compounds, which were not of interest for our current analysis. On the basis of the distribution, the interval of "bulk hit rates" (b_hr) for these extensively assayed PubChem compounds was defined as 0% < b_hr ≤ 1.0%, covering the lower quartile, median, and upper quartile (and hence the "bulk" of the distribution). There were 80,495 compounds with hit rates ≥ 1.0%. The hit rate distribution of this compound subset is shown in Figure 2 (middle), yielding a median of 1.8%. This value was set as the hit rate threshold for most active screening compounds. The threshold was exceeded by 41,609 compounds, representing 12.7% of the initial compound pool. The hit rate distribution of these compounds is reported in Figure 2 (right), resulting in a median of 2.9%. We determined that 93.1% of the compounds with hit rates greater than 1.8% in primary assays were also active in confirmatory assays (yielding IC_{50} values). Hence, their activity was not confined to primary assays.

Compound series and parameters

From the 41,609 compounds with highest hit rates, MMSs were systematically extracted on the basis of RECAP-MMPs. After removal of redundant MMSs (see Methods), 6941 unique MMSs were obtained comprising 14,646 compounds, which represented our final hit rate- and series-based selection set. Table 1 reports the size distribution of the MMSs, ranging from two to 17 analogs per series. With 6111 instances, compound pairs and triplets dominated the distribution, but more than 800 larger MMSs were also obtained. As further discussed below, compound pairs and triplets already provide informative controls for activity analysis and enable a more confident assessment compared to the analysis of individual compounds. This was a major motivation for focusing the analysis on MMSs.

Table 1. Size distribution of matched molecular series (MMSs). The distribution of 6941 frequently active MMSs (#MMSs) over increasing numbers of compounds (#CPDs) is reported.

#CPDs	#MMSs
2	4965
3	1156
4	435
5	190
6	70
7	48
8	22
9	21
10	11
11	12
12	3
13	4
14	2
15	1
17	1

Figure 3a illustrates the derivation of three parameters for the characterization and comparison of MMSs (rationalized in the Methods section). The cumulative *MMS hit rate* is a direct measure for the activity of a series. In addition, *assay overlap* represents a confidence criterion for MMS assessment, i.e., large assay overlap of compounds comprising a series assigns high confidence to hit rate comparisons. By contrast, the proportion of *assays with inconsistent activity* should best be minimal to draw firm conclusions. Figure 3b reports the distribution of these three parameters for the 6941 MMSs. Assay overlap (upper left plot) and MMS hit rates (lower left) were generally high, with median values of 79.3% and 5.8%, respectively. By contrast, the proportion of inconsistent assays (upper right) was overall low, with a median of only 3.7%. Thus, the distributions of MMS parameters indicated that the set of MMSs was suitable for the analysis of series-based hit rates and hit rate comparison of compounds comprising individual MMSs. We note that MMSs can be ranked in the order of decreasing assay overlap and MMS hit rates and increasing inconsistent assays and prioritized, for example, on the basis of rank fusion calculations.

Target distribution in primary assays

Our analysis was intentionally focused on hit rates over assays (i.e., assay promiscuity) to take as many activity readouts as possible into account. Therefore, as a control, assay- and target-based hit rates were also compared. Compounds forming the 6941 MMS were evaluated in a total of 1213 assays. For 255 of these assays, no individual target was specified. The remaining 958 assays covered 426 different targets. Figure 3b reports the distributions of MMS hit rates over assays (lower left plot) and targets (lower right). The distributions were overall similar, with median values of assay- and target-based hit rates of 5.8% and 4.9%, respectively. Hence, despite the presence of multiple assays for a subset of targets, assay-based hit rates were only slightly higher than target-based rates, indicating that corresponding conclusions would be drawn from the analysis of these distributions.

Known interference candidates

The computational aggregation advisor[4] and compound strings taken from PAINS filters[30,31] (http://www.rdkit.org) were used to search the MMSs for known assay interference candidates. The 14,646 MMS compounds contained 783 aggregators (on the basis of 100% similarity) and 2381 compounds with PAINS substructures. There were 611 MMSs with one or more aggregators, 1139 MMSs with one or more PAINS, and 126 MMSs including aggregators and PAINS. However, 5065 MMSs with high hit rates did not contain known compounds with aggregation potential or PAINS substructures. Thus, the MMSs provide a large source of analogs for the exploration of other interference candidates, as well as compounds with true multi-assay/target activities.

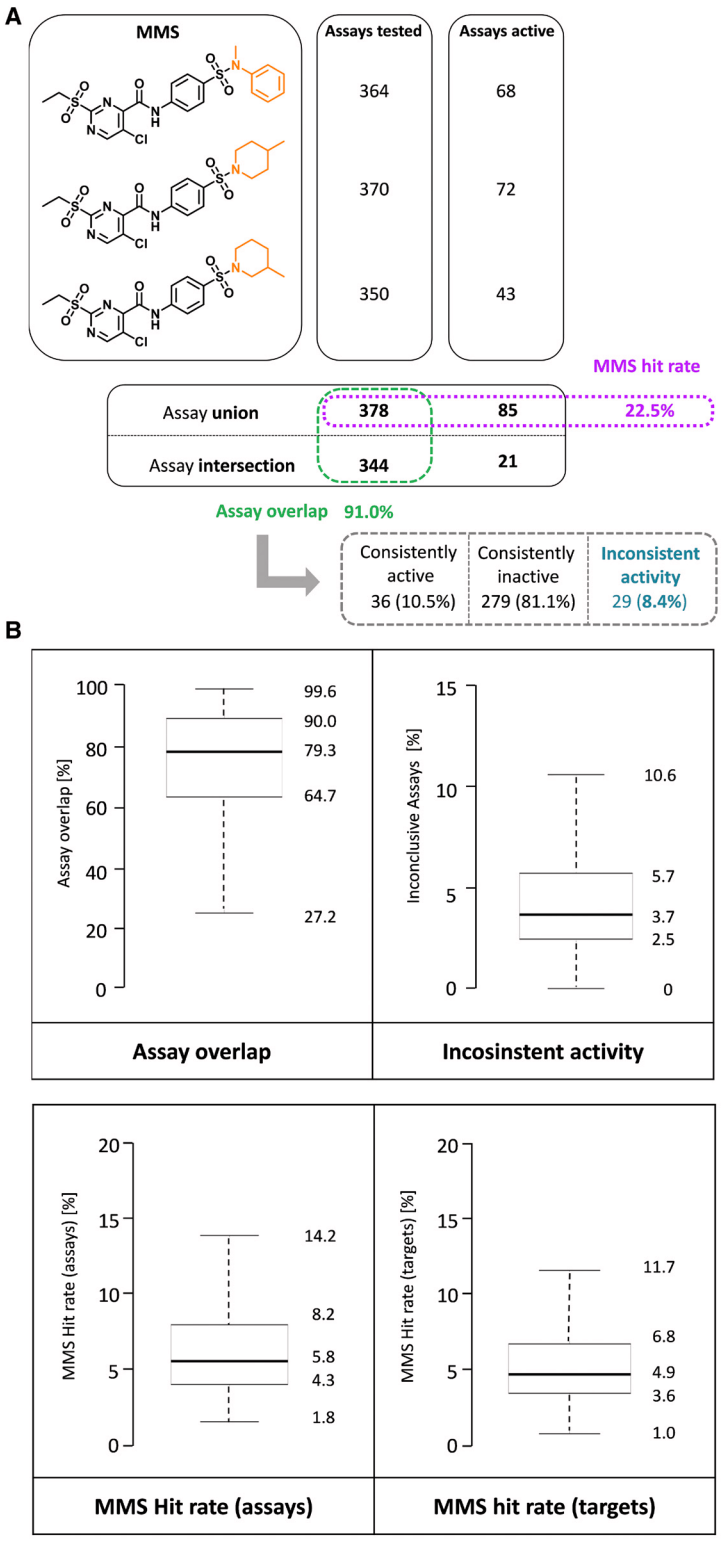
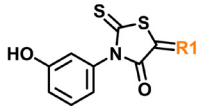
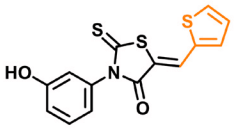
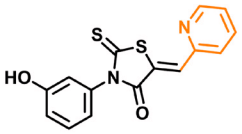


Figure 3. Characterization of matched molecular series (MMSs). (**a**) An exemplary MMS comprising three analogs is shown. The MMS core and varying substituents are colored in black and orange, respectively. For each compound, the number of assays it was tested and active in is reported, respectively. Furthermore, the assay union, intersection, and MMS hit rate (purple) are given. From these data, the assay overlap (green) of MMS analogs was determined as well as the proportion of assays with consistent activity, inactivity, and inconsistent activity (blue). (**b**) Boxplots are shown reporting the distribution of assay overlap, assays with inconsistent activity as well as assay- or target-based MMS hit rates for PubChem compounds with greater than 1.8% hit rate.

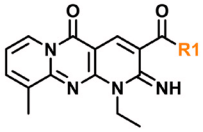
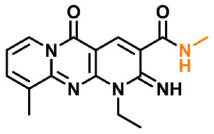
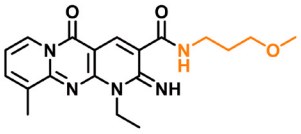
Exemplary series

Figure 4 shows exemplary compound pairs and triplets with high assay promiscuity. The two analogs in Figure 4a were tested in more than 380 assays with 93.5% assay overlap and only 1.6% inconsistent assays, yielding comparable hit rates of 2.8% and 3.2%, respectively, resulting in an MMS hit rate of 4.5%. These analogs contained a classical PAINS substructure (ene_rhodanine)[8,9]. Furthermore, compounds in Figure 4b were analogs of a molecule with aggregation potential[4]. They were tested in more than 300 and 400 assays, respectively, yielding a relatively low assay overlap of 59%, and had hit rates of 2.2% and 2.6%, respectively, resulting in a low MMS hit rate of 2.9%. Thus, these analogs were far from being consistently active, as one might assume for strong aggregators. In Figure 4c, a pair of thieno[2,3-d]pyrimidine-2-acetic acid ethyl ester analogs is shown that were tested in 442 assays with large overlap. These compounds had high hit rates of 5.9% and 7.9%, respectively, resulting in a high MMS hit rate of 9.8%. Moreover, Figure 4d shows a

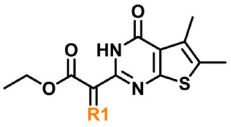
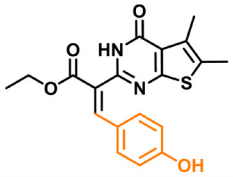
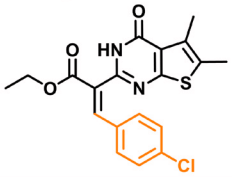
A

MMS core	MMS hit rate [%]	Assay overlap[%]	Inconsistent activitiy [%]
	**4.5**	**93.5**	**1.6**
CPDs			
#Primary assays	387	381	
#Active assays	11	12	
Hit rate [%]	2.8	3.2	

B

MMS core	MMS hit rate [%]	Assay overlap[%]	Incosistent activity [%]
	**2.9**	**59.0**	**0.7**
CPDs			
#Primary assays	318	416	
#Active assays	7	12	
Hit rate [%]	2.2	2.6	

C

MMS core	MMS hit rate [%]	Assay overlap[%]	Inconsistent activity [%]
	9.8	88.9	5.8
CPDs			
#Primary assays	442	442	
#Active assays	32	26	
Hit rate [%]	7.2	5.9	

D

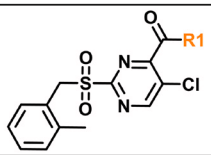
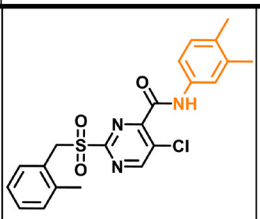
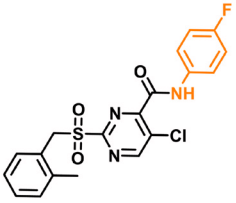
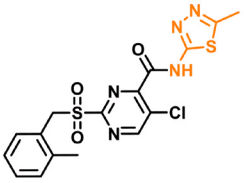
MMS core	MMS hit rate [%]	Assay overlap[%]	Inconsistent activity [%]
	17.8	89.7	8.0
CPDs			
#Primary assays	358	357	361
#Active assays	48	49	32
Hit rate [%]	13.4	13.7	8.7

Figure 4. Exemplary matched molecular series (MMSs). (**a**–**d**) Four exemplary MMSs (core, black; substituents, orange) are shown and the MMS hit rate, assay overlap, and proportion of assays with inconsistent activity are reported. In addition, for each individual analog, its assay frequency and hit rate are provided.

triplet of sulfonylpyrimidines that were tested in 357–361 assays with 89.7% overlap, having very high hit rates of 8.7% (one analog) and more than 13% (two analogs). Compounds forming each of the MMSs in Figure 4 displayed consistent hit rate characteristics, hence assigning confidence to their observed activity phenotype. The analogs in Figure 4c and Figure 4d have previously not been classified as interference candidates. However, they might well be reactive under assay conditions. Taken together, these examples of analog pairs and triplets (i.e., minimally sized MMSs) are indicative of the potential of well characterized MMSs for follow-up investigations focusing on assay interference and multi-target activities.

Conclusions

Herein, a detailed analysis of hit rates of nearly 440,000 extensively assayed screening compounds has been presented. On the basis of hit rate distributions, 12.7% of the compounds with highest hit rates were selected. From these compounds,

analog series with single substitution sites were systematically extracted to complement hit rate statistics with the assessment of structural relationships between active compounds. A total of 6941 unique MMSs were obtained comprising 14,646 compounds. These MMSs were characterized using different parameters prioritizing high-confidence series for activity analysis. A major goal of our study has been the data-driven generation of a pool of analog series for the evaluation of assay interference potential and multi-target activities. More than 5000 MMSs did not contain known interference candidates, providing an opportunity to evaluate compounds with interference potential on a large scale. In the next step, analog series will be evaluated from a medicinal chemistry perspective to complement and further extend statistical considerations. Annotated series and associated assay/target information will then be made freely available. The statistics and selection steps reported herein also make it possible to regenerate compound subsets at different hit rate levels and subject them to further analysis. In addition, large numbers of compounds with high hit rates that were not part of MMSs are also available. For reasons discussed, our preferred approach is taking compound series information into account when judging assay promiscuity.

Data availability

The data sets used in this study are freely available in PubChem and can be generated following the selection protocol reported in the Methods.

Competing interests

No competing interests were declared.

Grant information

DS is supported by *Sonderforschungsbereich 704* of the *Deutsche Forschungsgemeinschaft*.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Aldrich C, Bertozzi C, Georg GI, *et al.*: **The Ecstasy and Agony of Assay Interference Compounds.** *ACS Cent Sci.* 2017; **3**(3): 143–147.
PubMed Abstract | Publisher Full Text | Free Full Text
2. McGovern SL, Caselli E, Grigorieff N, *et al.*: **A Common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening.** *J Med Chem.* 1996; **45**(8): 1712–1722.
PubMed Abstract | Publisher Full Text
3. Shoichet BK: **Screening in a spirit haunted world.** *Drug Discov Today.* 2006; **11**(13–14): 607–615.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Irwin JJ, Duan D, Torosyan H, *et al.*: **An Aggregation Advisor for Ligand Discovery.** *J Med Chem.* 2015; **58**(17): 7076–7087.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Feng BY, Simeonov A, Jadhav A, *et al.*: **A high-throughput screen for aggregation-based inhibition in a large compound library.** *J Med Chem.* 2007; **50**(10): 2385–2390.
PubMed Abstract | Publisher Full Text
6. Jadhav A, Ferreira RS, Klumpp C, *et al.*: **Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease.** *J Med Chem.* 2010; **53**(1): 37–51.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Ferreira RS, Simeonov A, Jadhav A, *et al.*: **Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors.** *J Med Chem.* 2010; **53**(13): 4891–4905.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Baell JB, Holloway GA: **New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.** *J Med Chem.* 2010; **53**(7): 2719–2740.
PubMed Abstract | Publisher Full Text
9. Baell J, Walters MA: **Chemistry: Chemical con artists foil drug discovery.** *Nature.* 2014; **513**(7519): 481–483.
PubMed Abstract | Publisher Full Text
10. Dahlin JL, Nissink JW, Strasser JM, *et al.*: **PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS.** *J Med Chem.* 2015; **58**(5): 2091–2113.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Dahlin JL, Nissink JW, Francis S, *et al.*: **Post-HTS case report and structural alert: Promiscuous 4-aroyl-1,5-disubstituted-3-hydroxy-2*H*-pyrrol-2-one actives verified by ALARM NMR.** *Bioorg Med Chem Lett.* 2015; **25**(21): 4740–4752.
PubMed Abstract | Publisher Full Text
12. Gilberg E, Jasial S, Stumpfe D, *et al.*: **Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology.** *J Med Chem.* 2016; **59**(22): 10285–10290.
PubMed Abstract | Publisher Full Text
13. Baell JB: **Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS).** *J Nat Prod.* 2016; **79**(3): 616–628.
PubMed Abstract | Publisher Full Text
14. Bisson J, McAlpine JB, Friesen JB, *et al.*: **Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery?** *J Med Chem.* 2016; **59**(5): 1671–1690.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Nelson KM, Dahlin JL, Bisson J, *et al.*: **The Essential Medicinal Chemistry of Curcumin.** *J Med Chem.* 2017; **60**(5): 1620–1637.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Wang Y, Xiao J, Suzek TO, *et al.*: **PubChem's BioAssay Database.** *Nucleic Acids Res.* 2012; **40**(Database issue): D400–D412.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Capuzzi SJ, Muratov EN, Tropsha A: **Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS.** *J Chem Inf Model.* 2017; **57**(3): 417–427.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Jasial S, Hu Y, Bajorath J: **How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds.** *J Med Chem.* 2017; **60**(9): 3879–3886.
PubMed Abstract | Publisher Full Text
19. Gilberg E, Stumpfe D, Bajorath J: **Activity profiles of analog series containing pan assay interference compounds.** *RSC Adv.* 2017; **7**(57): 35638–35649.
Publisher Full Text
20. Hu Y, Bajorath J: **Compound promiscuity: what can we learn from current data?** *Drug Discov Today.* 2013; **18**(13–14): 644–650.
PubMed Abstract | Publisher Full Text
21. Paolini GV, Shapland RH, van Hoorn WP, *et al.*: **Global mapping of pharmacological space.** *Nat Biotechnol.* 2006; **24**(7): 805–815.
PubMed Abstract | Publisher Full Text
22. Boran AD, Iyengar R: **Systems approaches to polypharmacology and drug discovery.** *Curr Opin Drug Discov Devel.* 2010; **13**(3): 297–309.
PubMed Abstract | Free Full Text
23. Jasial S, Hu Y, Bajorath J: **Determining the Degree of Promiscuity of Extensively Assayed Compounds.** *PLoS One.* 2016; **11**(4): e0153873.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Griffen E, Leach AG, Robb GR, *et al.*: **Matched molecular pairs as a medicinal chemistry tool.** *J Med Chem.* 2011; **54**(22): 7739–7750.
PubMed Abstract | Publisher Full Text

25. Hussain J, Rea C: **Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets.** *J Chem Inf Model.* 2010; **50**(3): 339–348.
PubMed Abstract | Publisher Full Text

26. Wawer M, Bajorath J: **Local structural changes, global data views: graphical substructure-activity relationship trailing.** *J Med Chem.* 2011; **54**(8): 2944–2951.
PubMed Abstract | Publisher Full Text

27. Lewell XQ, Judd DB, Watson SP, *et al.*: **RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry.** *J Chem Inf Comput Sci.* 1998; **38**(3): 511–522.
PubMed Abstract | Publisher Full Text

28. de la Vega de León A, Bajorath J: **Matched molecular pairs derived by retrosynthetic fragmentation.** *Med Chem Commun.* 2014; **5**(1): 64–67.
Publisher Full Text

29. Hu X, Hu Y, Vogt M, *et al.*: **MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs.** *J Chem Inf Model.* 2012; **52**(5): 1138–1145.
PubMed Abstract | Publisher Full Text

30. Sterling T, Irwin JJ: **ZINC 15--Ligand Discovery for Everyone.** *J Chem Inf Model.* 2015; **55**(11): 2324–2337.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Gaulton A, Bellis LJ, Bento AP, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1100–D1107.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 1

Referee Report 02 October 2017

doi:10.5256/f1000research.13396.r25426

✔ **Michael Walters**
Department of Medicinal Chemistry, University of Minnesota, Minneapolis, MN, USA

This is an excellent manuscript that will certainly help researchers further understand the liabilities of "good-actors"; compounds that provide structure-interference relationships (SIR) that may not be pertinent to the biology being studied.

Notes:

1. Inconsistent is misspelled in Panel B of Figure 3.

2. Though this will need to be experimentally-verified, the sulfonylpyrimidines are almost certainly interfering by SnAr reactions. A literature search to gauge this potential reactivity of the core structure retrieved >100 articles and >1000 reactions. This group is certainly not in the PAINS definitions, but its assay interference potential can readily be ascertained by a quick reaction substructure search. Most importantly, this observation reinforces that lack of PAINS "flagging" is not sufficient to ensure that compounds don't have assay interference potential. The authors may wish to highlight this in their discussion.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: assay interference, medicinal chemistry

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 12 September 2017

doi:10.5256/f1000research.13396.r25674

José L Medina-Franco , **Fernanda I. Saldívar-González**
Department of Pharmacy, Faculty of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico

The manuscript addresses a timely topic and is part of an effort to understand the multitarget activities of small molecules and their potential utility or drawbacks for drug discovery.

Using the MMs formalism, the authors developed a protocol to identify analog series with high hit rates. The MMs approach responds to the need to generate SAR information of interference compounds. This approach allows for more reliable evaluation compared to the analysis of individual compounds.

The analogue series identified with data from PubChem should be useful for medicinal chemistry programs. Similarly, the protocol developed in this work using MMPs should be useful to mine other large screening data sources (either public or proprietary data sets).

Minor suggestions to further improve the quality of the manuscript:

- Comment on the manuscript the effect of the compound concentration that is used to define a “hit” compound in a given assay. In other words, since it is unlikely that the same compound concentration is used across all assays, how this variable influences the “hit rates” and conclusions of the study?
- The current analysis is made based primarily in primary assays in PubChem. In the Methods authors justify that there are larger data volumes for primary assays. We agree but would be nice to see in the manuscript a comment regarding the balance between volume vs. quality of the data.
- We found it very pertinent that the study focused on compounds with high rates in primary assays, since more than 90% of these compounds are also active in confirmatory trials. It is interesting to comment how the hit rate is related to the quality of the data.
- An earlier study [1] mentions that the assay frequency is not correlated with increased promiscuity. It is desirable to include a commentary in the manuscript when discussing the frequency of the test distribution and the generation of the first subset of 327,532 compounds.
- Three parameters for the characterization and comparison of MMSs were determined (hit rate (HR), assay overlap, assays with inconsistent activity). Does the MMS size affect these determinations?

Other suggestions (per section):

- Introduction, last paragraph: briefly summarize the major findings of previous works related to this study (e.g., refs. 18, 19, 23) and emphasize the novelty of this work.
- Shorten the title (and include the concept of MMPs).
- In the Introduction (last paragraph), include figure numbers for expressions such as “extensively”, “larger’, “many”, “much higher”. For instance, when the authors mention “extensively tested” in the title and through the manuscript, they mean “>10,000”, “>100,000”, “>400,000”, etc.
- Methods: elaborate a bit more on the in-house scripts, e.g., the program language.

References

1. Hu Y, Bajorath J: Entering the 'big data' era in medicinal chemistry: molecular promiscuity analysis revisited.*Future Sci OA*. 2017; **3** (2): FSO179 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 18 August 2017

doi:10.5256/f1000research.13396.r25140

John A. Lowe III
Jl3pharma LLC, Stonington, CT, USA

This article makes an important contribution to addressing a serious complication in screening for new biological activity, especially in the context of new drug discovery. Several researchers continue to point out the waste of time and money in following up on false positive hits from biological screening programs.

While paradigms to filter out false positives, such as aggregators or other “frequent hitters” such as PAINs, are becoming mainstream techniques, these may not be sufficient, either because they miss some false positives or because they mistakenly classify legitimate hits as false positives. By carrying out a statistical analysis of a large set of screening data for hit rates using matched molecular pairs and series, this article offers a valuable perspective on this issue.

There are several aspects of this article that are particularly valuable. For example, Figure 3b is an important control showing that using assay hit rate does not overstate promiscuity, in that target hit rate is similar, linking the results to a biological mechanism of action. In Figures 4c and 4d, the exemplified compounds appear to be Michael acceptors, and could react with Cys residues covalently, or sequester thiol reagents used in the assay, which could explain their promiscuity. It is worth pointing these out, as they would not be picked up in PAINs filters and only by a knowledgeable chemist. Overall, this work is as important contribution to the ongoing effort to reduce the occurrence of false positive hits serving as starting points for discovery programs.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Discuss this Article

Version 1

Author Response (*Member of the F1000 Faculty* and *F1000Research Advisory Board Member*) 02 Oct 2017

Jürgen Bajorath, LIMES Program Chem. Biol. & Med. Chem, University of Bonn, Germany

Thank you for your comment and interest. We are still investigating the systematic extraction of analog series from data sets of any composition and size and intend finalizing this methodological work before addressing code release issues. For the time being, the interested reader is referred to a variety of partly related implementations we have made freely available on the Zenodo open access platform.
On a more general note, although we are among the computational groups promoting open science by public release of many in-house generated data sets and software tools, our experiences with free data and tool sharing have not been entirely positive; another point of consideration going forward. Perhaps it might make sense to (re-)consider other collaborative models and release procedures.

Competing Interests: None

Reader Comment 06 Sep 2017

Greg Landrum, T5 Informatics GmbH, Switzerland

The approach for identifying chemical series in an automated and computationally efficient manner is an interesting one and I could imagine it being useful to other researchers. It's a challenging problem that comes up pretty frequently and for which no great solutions are available. The MMS-identification algorithm could be an interesting contribution on its own.

Is there any chance that the authors would be willing to make the code for doing the MMS identification available?

Competing Interests: No competing interests.