Journal of
CHEMICAL INFORMATION
AND MODELING

Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics

Ye Hu and Jürgen Bajorath*

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

PROMISCUITY

ABSTRACT: Compound activity data grow at unprecedented rates, and their complexity increases. This challenges compound data mining efforts and makes it difficult to draw reliable conclusions from data analysis. We have aimed to investigate the influence of individual parameters and data confidence levels on compound selection and property assessment. Therefore, alternative sets of bioactive compounds were systematically extracted from ChEMBL on the basis of iteratively expanding selection criteria with increasing stringency covering a variety of search parameters. The sequential application of criteria for the selection of high-confidence compound data was order-independent, as expected. Furthermore, the influence of separately applied selection criteria was analyzed. Criteria that largely influenced compound selection and compound promiscuity rates were identified. In the presence of stringent selection criteria and high data confidence, many compounds with likely assay artifacts or liabilities were eliminated from further consideration. Taken together, the findings of our analysis emphasize the need to carefully consider search parameters related to target organisms, confidence level of activity, and activity measurements and suggest reliable protocols for compound data mining.

INTRODUCTION

During the past decade, large amounts of compounds and activity data have become publicly available, and the data continue to grow at unprecedented rates.[1] A variety of public repositories for compound, structural, and activity data have been established with different focal points. Among others, these include the PubChem BioAssay database,[2] which mainly collects screening library compounds and high-throughput screening data; BindingDB[3] and ChEMBL,[4,5] in which most of the compound activity data from medicinal chemistry sources are assembled; DrugBank,[6] which collects approved and experimental drugs; and Open PHACTS,[7] which provides a meta search engine and pharmacology records for biological targets and/or small molecules. In addition, there are databases such as ZINC,[8] which archives compounds from vendor sources that might or might not be known to have biological activity. Taken together, these repositories currently already contain tens of millions of compounds that are biologically annotated at different levels, giving rise to the advent of the "big data" era in chemistry.[1]

Although it is generally appreciated that such large amounts of compound data are necessarily heterogeneous,[1] affected by data integrity issues, and thus often likely to have different confidence levels and error margins, only relatively few studies to date have addressed in detail the influence of data curation and selection criteria on the results of compound data mining or computational modeling. For example, in a recent study, compound data consistency and compatibility across the five databases mentioned above were assessed for selected examples of structurally related or functionally (therapeutically) related drugs.[1] It was shown that currently available activity data are highly complex and characterized by a high degree of heterogeneity when different data sources are compared. In a related analysis,[9] target annotations of drugs in the ChEMBL database were compared when different data selection criteria were applied. The number of target annotations was found to vary greatly when selection criteria reflecting different levels of stringency were evaluated. In another study, it was shown that the detectable growth of promiscuity (i.e., multitarget activity) of compounds across different protein families was largely dependent on the types of activity measurements that were considered.[10] Furthermore, the influence of different types of activity measurements and their confidence levels on the results of structure−activity relationship (SAR) analysis was investigated.[11] In part, significant differences in modeled activity landscapes of compound data sets and SAR patterns were detected when alternative activity annotations were utilized. In a conceptually different investigation,[12] the experimental uncertainties of K_i values (equilibrium constants) were analyzed for specific ligand−target interactions for which multiple measurements were available. Such measurements might originate from different laboratories or be obtained in assays of different designs and/or under different assay conditions. Experimental variances or uncertainties were estimated to result in a mean error of 0.44 pK_i units and a standard deviation of 0.54 pK_i units for individually published K_i values.[12] Data heterogeneity and different confidence levels (which might or might not be considered) generally affect compound and activity data selection from chemical repositories.[1,4,11−15]

Received: August 4, 2014
Published: October 20, 2014

© 2014 American Chemical Society

Table 1. Data Set Design[a]

set	description	specific selection criteria (ChEMBL)
1	all compounds with interactions against any target	none
2	all compounds with interactions against *human targets*	target organism: "*Homo sapiens*"
3	all compounds with *direct interactions* against human targets at the *highest confidence level*	assay relationship type: "D"; assay confidence score: "9"
4	all compounds with direct interactions against human *single protein* targets at the highest confidence level	target type: "Single Protein"
5	all compounds with direct interactions against human single protein targets at the highest confidence level; *potency measurement types K_i or IC_{50}*	standard activity type: "K_i" or "IC_{50}"
6	all compounds with direct interactions against human single protein targets at the highest confidence level; potency measurement types K_i or IC_{50}; *approximate potency values discarded*	standard activity relation: "="
7	all compounds with direct interactions against human single protein targets at the highest confidence level; potency measurement types K_i or IC_{50}; approximate potency values discarded; *activity unit "nM"*	standard activity unit: "nM"
8	all compounds with direct interactions against human single protein targets at the highest confidence level; potency measurement types K_i or IC_{50}; approximate potency values discarded; activity unit "nM"; *inactive compounds discarded*	activity comment: not equal to "Inactive", "Inconclusive", or "Not Active"

[a] For each data set, specific selection criteria are reported that were applied in a sequential manner. Additional criteria setting each set apart from its predecessor are given in *italics*. In each case, specific criteria are also reported using ChEMBL query terminology.

Taken together, the results of the studies discussed above indicate that conclusions drawn from data mining, SAR analysis, or other studies are, to a greater or lesser extent, influenced by compound data integrity and selection criteria, an issue that is often not explicitly taken into consideration when reporting computational studies, despite at least some general awareness.

High data confidence is a prerequisite for drawing meaningful conclusions from large-scale compound data mining studies, for example, in the context of promiscuity analysis. However, this does not mean that low-confidence data must per se be biologically irrelevant. They are, however, error-prone and as such are likely to bias knowledge extraction from data.[9] Given that large amounts of increasingly complex data find their way into public repositories, we have been interested in further exploring a number of questions that we consider to be of critical relevance for data mining. For example, how might compound coverage for given targets change under increasingly stringent selection criteria? Can we identify individual parameters or criteria that significantly impact compound selection? How might such criteria influence the assessment of compound promiscuity? In order to address these and related questions, we have systematically assessed how defined selection criteria and their sequential application change the composition of different data sets extracted from ChEMBL. Details of our analysis and the results are presented herein.

■ MATERIALS AND METHODS

ChEMBL. The ChEMBL database collects bioactivity information for large numbers of compounds from the medicinal chemistry literature and patent sources.[4,5] Most of the activity data are extracted from scientific publications in a variety of journals and also from the PubChem BioAssay database and manually curated. Currently, ChEMBL contains 1 411 786 distinct compounds with activity against 10 579 targets and a total of 12 843 338 activity measurements (interactions). As such, it is one of the largest public repositories of activity data from medicinal chemistry and an important knowledge base for drug discovery. In addition to access through an online search interface, the database content is also available in different formats, such as Oracle and MySQL that provide a relational database structure and enable data decomposition into individual tables.[4] For example, in ChEMBL release 18, a total of 51 different data tables are available. Hence, it is possible to systematically query compounds and corresponding activity information in a structured manner on the basis of various search parameters.

Compound Data Mining. From ChEMBL release 18, eight compound data sets were extracted by applying increasingly stringent selection criteria (Table 1):

Set 1: All compounds with any target annotations were selected; no additional criteria were applied, hence yielding the largest possible set of "bioactive" compounds.

Set 2: All compounds with reported interactions with human targets ("*Homo sapiens*") were extracted (in ChEMBL, targets originate from a total of 1639 different organisms).

Set 3: All compounds with "direct interactions" (i.e., ChEMBL assay relationship type "D") against human targets at the highest confidence level (i.e., assay confidence score of "9") were assembled. It is noted that ChEMBL utilizes two parameters to classify assay-to-target relationships, including the qualitative parameter "assay relationship type" and the quantitative parameter "assay confidence score". There are a total of six relationship types (Table 2) and 10 confidence scores (Table 3) that determine the level of confidence that the activity against a given target is evaluated in a relevant assay system. They range from "U" to "D" and "0" to "9" for increasingly defined relationships and increasing confidence levels, respectively.

Table 2. Assay Relationship Types[a]

relationship type	description
D	direct protein target assigned
H	homologous protein target assigned
M	molecular target other than protein assigned
N	nonmolecular target assigned
S	subcellular target assigned
U	default value—target has yet to be curated

[a] The six assay relationship types used in ChEMBL and their descriptions are provided.

Set 4: In addition to the criteria applied for set 3, this set required the target type "Single Protein", one of 26 different target types designated in ChEMBL (Table 4). Compounds active against any other target type were omitted.

Table 3. Assay Confidence Scores[a]

confidence score	description
0	default value—target unknown or has yet to be assigned
1	target assigned is nonmolecular
2	target assigned is a subcellular fraction
3	target assigned is a molecular nonprotein target
4	multiple homologous protein targets may be assigned
5	multiple direct protein targets may be assigned
6	homologous protein complex subunits are assigned
7	direct protein complex subunits are assigned
8	homologous single protein target is assigned
9	direct single protein target is assigned

[a] The 10 assay confidence scores used in ChEMBL and their descriptions are provided.

Table 4. Target Types[a]

target type	description: target is/has:
ADMET	not applicable for an ADMET assay
Cell-Line	specific cell line
Chimeric Protein	synthetic or naturally occurring fusion of two different proteins
Macromolecule	biological macromolecule
Metal	metal
Molecular	defined molecular entity
Non-Molecular	not defined at the molecular level; nonmolecular entity is affected
No Target	not applicable for a screening assay
Nucleic Acid	DNA, RNA, or PNA
Oligosaccharide	oligosaccharide
Organism	complete organism
Phenotype	biological phenotype or process
Protein	protein or group of proteins
Protein Complex	defined protein complex consisting of multiple subunits
Protein Complex Group	poorly defined protein complex with unknown subunit composition
Protein Family	group of closely related proteins
Protein−Nucleic Acid Complex	complex consisting of protein and nucleic acid components
Protein−Protein Interaction	protein−protein interaction
Selectivity Group	pair of proteins with known selectivity
Single Protein	single protein chain
Small Molecule	small molecule
Subcellular	subcellular fraction
Tissue	healthy or diseased tissue
Unchecked	not assigned
Undefined	not defined
Unknown	unknown molecular entity

[a] The 26 target types used in ChEMBL and their descriptions are provided.

Set 5: In addition, only two types of potency measurements were considered, i.e., (assay-independent) equilibrium constants (K_i values) and (assay-dependent) IC_{50} values. In ChEMBL, more than 5000 measurement types are considered including, for example, “%max”, “Activity”, “Efficacy”, “EC_{50}”, “K_d”, and “Residual Activity”. Compounds with measurements other than K_i or IC_{50} values were discarded.

Set 6: In addition, the “activity relation” was taken into consideration. In total, 22 relations are used to annotate activity measurements, including, for example, “>”, “=”, “<”, “~”, “≫”, and “=~~”. In this set, only compounds with explicitly defined K_i or IC_{50} values (i.e., standard activity relation “=”) were retained. Approximate measurements such as “>”, “<”, and “~” were eliminated.

Set 7: In addition to defining types of activity measurements and their relations, “activity units” were specified. In ChEMBL, there are over 1500 units available for activity measurements, including, for example, “%”, “degree”, “g”, “kCal”, “log10”, and “min”. Here, the standard activity unit was set to “nM”.

Set 8: For the most stringent selection, in addition to criteria applied for set 7, “activity comments” were further examined. Hence, all activity records, including the comment “Inactive”, “Inconclusive”, or “Not Active”, were discarded.

These eight compound data sets resulted from combinations of increasing numbers of query parameters. They were applied to generate well-defined data selection criteria of stepwise increasing stringency. Of course, selection criteria might be applied in different orders/sequences. To examine the influence of the order in which the selection criteria are applied, the selection protocol detailed above was also applied in reverse order, such that compounds with questionable activity measurements or inactivity annotations were initially removed. The inverse selection scheme resulted in selection sets 1′ to 8′, as summarized in Figure 1. Practically, database queries were carried out using SQL statements. Data set statistics are reported in Figure 3 and further discussed below. In the context of our analysis, it should be noted that there is a principal difference between “annotation” and “activity”. The term annotation is primarily qualitative in nature, i.e., it is used to determine whether a compound has any reported activity against a given target. On the other hand, activity is a quantitative attribute and refers to numerical (or pseudonumerical) measurements of affinity (at different confidence

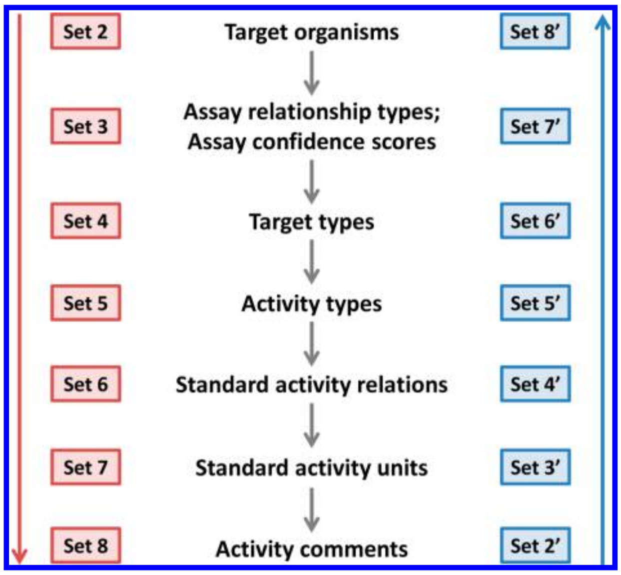


Figure 1. Data set organization. The organization of compound selection sets according to the original (left) and reverse (right) ordering of data selection criteria is summarized. The largest (initial) data sets (sets 1 and 1′) consist of the same set of compounds with reported activity against any targets and thus are not shown.

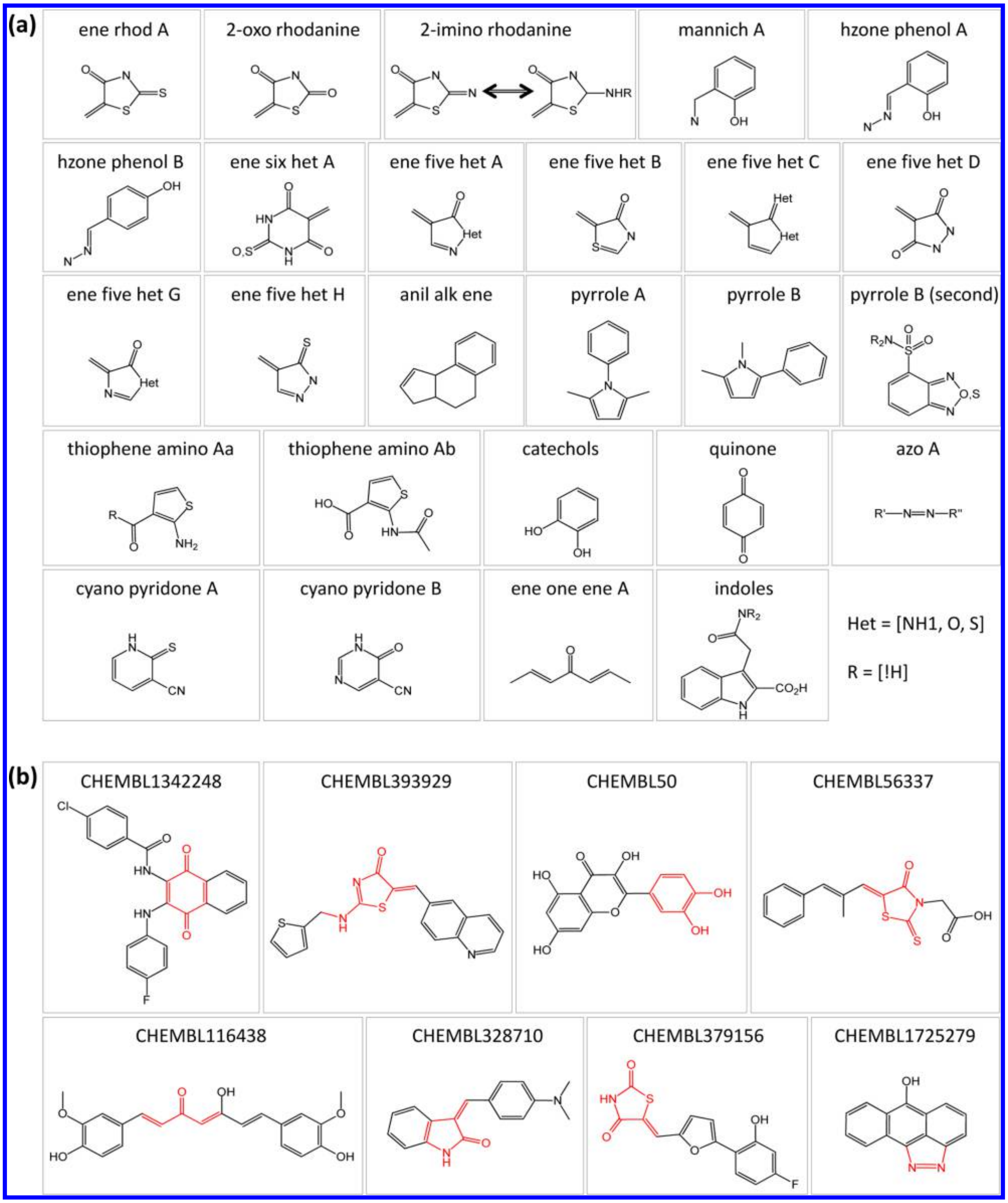


Figure 2. PAINS filter. (a) Set of 26 substructures used as a filter for possible "pan assay interference compounds" (PAINS). "Het" represents a secondary amine or an oxygen or sulfur atom, and "R" represents a non-hydrogen atom. (b) Eight ChEMBL compounds with potential PAINS liability. For each compound, its ChEMBL ID is provided and the detected PAINS substructure is highlighted in red.

levels). As such, activity is often synonymously used with potency (although this is formally not always correct).

Promiscuity Analysis. In the context of our analysis, compound promiscuity is rationalized as the ability of a small molecule to specifically interact with multiple targets.[16−18] So-defined promiscuity provides the molecular basis of polypharmacology, an emerging theme in drug discovery that refers to functional consequences of promiscuity, such as the engagement of a promiscuous compound in multiple signaling pathways.[16−20] Promiscuity rates were calculated here as a measure of compound data integrity (not to provide an update on most recent promiscuity assessments[16]). For the calculation of promiscuity rates, all ChEMBL target annotations of compounds were collected using increasingly stringent selection criteria as described above. Promiscuity rates calculated from ChEMBL data cannot be normalized by the number of instances of compounds that have been assayed because such statistics are not provided (and usually are not available) for active compounds retrieved from literature sources.[21] However, it has been observed previously that the majority of active compounds assembled from PubChem confirmatory assays have been tested in more than 50 different assays. These compounds were active against only ∼2.5 targets.[21]

Detection of Pan Assay Interference Compounds. So-called pan assay interference compounds (PAINS) display nonspecific apparent activity in a variety of assays and are typically false positives.[22] Often, such compounds are reactive or cause nonspecific aggregation effects. Baell and Holloway identified such problematic screening compounds and generated a set of 26 substructures that are indicative of PAINS liability.[22] These substructures are shown in Figure 2a. This set of substructures was assembled as a filter that was used to monitor the presence of compounds with potential PAINS liability in our data sets through substructure searching.

RESULTS AND DISCUSSION

Table 5. Data Set Composition[a]

set	no. of compounds	no. of targets	no. of interactions
(a) Original Order			
1	1291676	9223	8712130
2	936924	3934	6295086
3	605206	2144	2639767
4	605056	2123	2639440
5	179161	1527	376077
6	150379	1402	230493
7	150211	1398	230250
8	148373	1397	227311
(b) Reverse Order			
set	no. of compounds	no. of targets	no. of interactions
1′	1291676	9223	8712130
2′	1043487	9029	3453954
3′	828698	6968	2215503
4′	673089	6746	1486398
5′	482197	5722	992742
6′	330252	3799	579020
7′	201043	2891	314711
8′	148373	1397	227311

[a] For each data set, the numbers of compounds, targets, and compound−target interactions are reported for the (a) original and (b) reverse order of data selection criteria.

Data Sets. On the basis of the selection criteria detailed in Table 1, compound sets of very different composition were obtained, as reported in Table 5a. The comparison of Table 5a and Table 5b shows that the results of the final compound selection sets were not order-dependent; only the compositions of the intermediate data sets changed, as to be expected. Therefore, in the following, we discuss in detail only the results obtained from the original sequence of applied data selection criteria and present the statistics for both orderings in the following figures and tables. Set 1 contained nearly 1 300 000 compounds with reported activity against 9223 targets accounting for a total of 8 712 130 interactions. In set 2, in which only human targets were considered, the number of corresponding compounds and interactions were reduced by ∼28%, and the number of targets decreased by more than half, i.e., from 9223 to 3934 targets. Furthermore, in set 3, when the confidence level of assays was taken into consideration, an even more significant decrease was observed, as ∼58% of the interactions in set 2 were no longer detected. Accordingly, the majority of the interactions in ChEMBL had a low level of confidence and might often result from activities against orthologues. However, the compositions of sets 3 and 4 were comparable, indicating that most of the compounds in set 3 were active against single protein targets. Only a small number of compounds were annotated with targets belonging to other types. By contrast, in set 5, the limitation of activity measurements to K_i or IC_{50} values resulted in a dramatic further decrease. Compared with set 4, the number of compounds was reduced from more than 605 000 to less than 180 000, the number of targets from 2123 to 1527, and the number of interactions from ∼2 640 000 to 376 077. These changes indicated that the majority of the activity values represented approximate measurements. In sets 6 to 8, where activity relations, units, and comments were specified, respectively, changes in the data set composition were much smaller. However, the omission of approximate or implicit activity values had another significant effect. In set 8, obtained under the most stringent selection criteria, there were 148 373 bioactive compounds with activity against 1397 human targets, representing a total of 227 311 high-confidence interactions. Hence, under the final parameter settings, only a small subset of the entire database content was selected. As shown in Figure 3a, the most significant reductions in the number of compounds occurred when transitioning from set 1 to set 2, from set 2 to set 3, and from set 4 to set 5. The corresponding query parameters accounted for target organism, confidence level of assays, and types of activity measurements, respectively. As shown in Figure 3a, for targets, the largest reduction was observed in going from set 1 to set 2 when only human targets were considered. Furthermore, similar to the trends observed for compounds, the largest differences in the number of interactions were detected in comparisons of sets 1 and 2, sets 2 and 3, and sets 4 and 5, as shown in Figure 3a. Hence, target organism, confidence level of assays, and types of activity measurements represented overall the most influential parameters for compound selection.

Compound Promiscuity. For each data set, the number of compounds annotated with ≥2, >5, or >10 targets was determined, as reported in Table 6a. With increasingly rigorous selection criteria, the number of promiscuous compounds decreased to varying degrees. From ∼74% in set 1, the number of promiscuous compounds was reduced to ∼30% in sets 6 to 8. In set 1, 28% of all compounds were active against more than five targets, whereas the corresponding ratio decreased to only 1% in sets 6 to 8. In addition, for each data set, the average promiscuity rate (i.e., the average number of targets against which a compound was active) was determined, as reported in Table 6a and Figure 3a. The largest average rate was observed in sets 1 and 2. A compound in these two sets was on average active against 6.7 targets. Surprisingly, the average promiscuity rate was essentially the same for compounds in sets 1 and 2, although a significant decrease in the number of compounds, targets, and interactions was observed in set 2 relative to set 1, as shown in Figure 3a. Hence, the inclusion of activity against targets from other organisms did not increase the compound

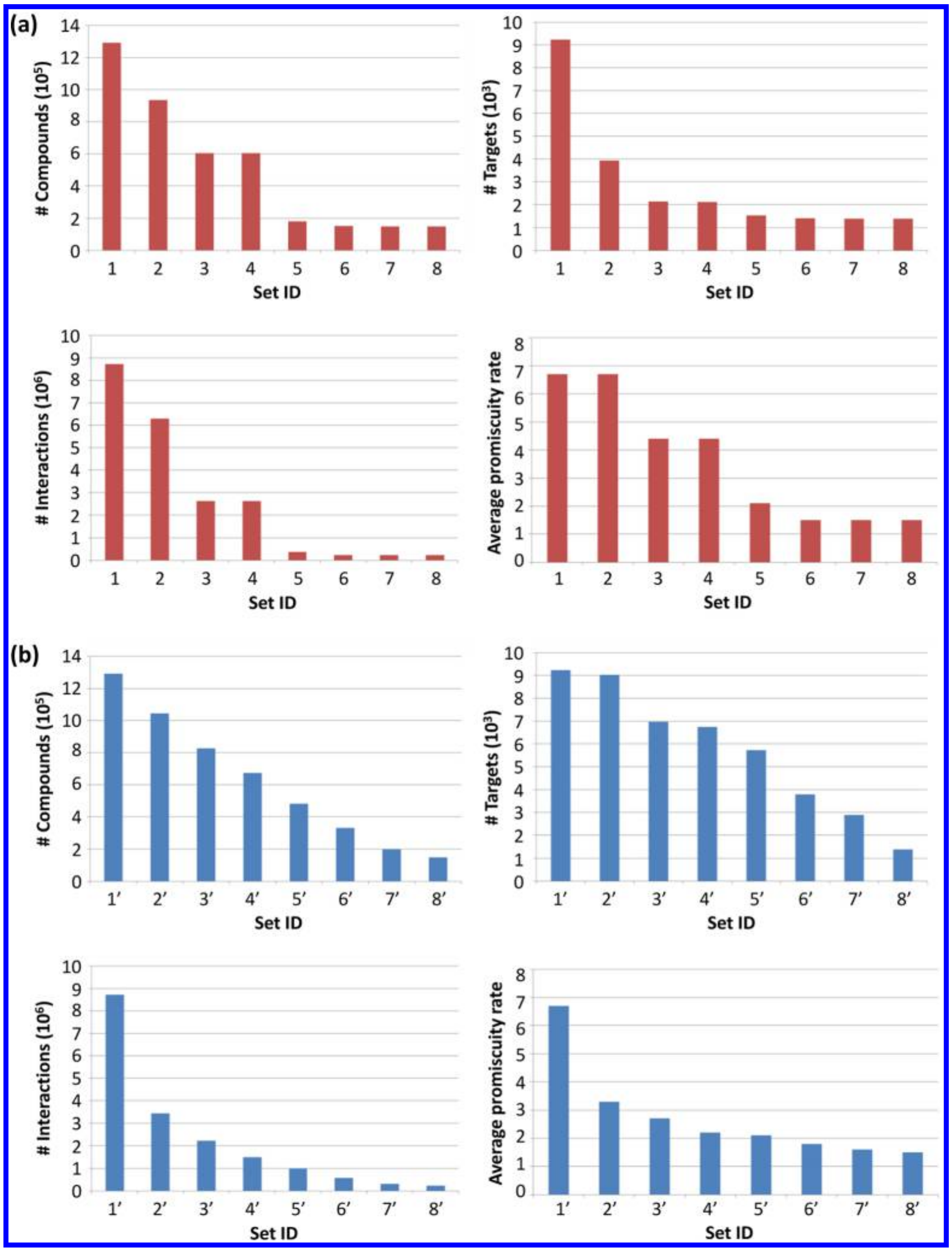


Figure 3. Data set statistics. (a) Distributions of compounds, targets, compound–target interactions, and average promiscuity rates over the eight data sets. (b) Corresponding results for the data sets obtained using the reverse order of data selection criteria.

promiscuity rate. In sets 3 and 4, the rate was reduced to 4.4 targets. Hence, taking the confidence level of assays into account resulted in a significant reduction in average compound promiscuity by more than two targets per compound. Moreover, when the type, relations, units, and comments parameters were applied to activity measurements in sets 5 to 8, the average promiscuity rate was further significantly reduced to 2.1 targets per compound (set 5) and 1.5 targets (sets 6 to 8). Hence, two parameters, i.e., the confidence level of assays and the types of activity measurements, substantially impacted the assessment of average promiscuity in a data set. When well-defined measurements at a high confidence level were considered, promiscuity rates were low, suggesting that compound promiscuity might often be overestimated because of the inclusion of low-confidence activity data.

PAINS Liability. The PAINS substructure filter was applied to all of the data sets. As reported in Table 7a, from sets 1 to 8, 55 447 to 5280 ChEMBL compounds were found to contain PAINS substructures. Exemplary compounds are shown in Figure 2b. Although the absolute numbers of compounds with possible PAINS liability significantly varied across sets 1 to 8, the proportions of PAINS with respect to the total number of compounds in a given set were comparable, ranging from 4.3% to 3.6%. Hence, stringency of target and activity parameters had only small effects on the global distribution of PAINS-positive compounds. However, when average promiscuity rates were calculated for compounds with PAINS liability in individual

Table 6. Compound Promiscuity[a]

	(a) Original Order			
	no. of promiscuous compounds (%)			
set	≥2 targets	>5 targets	>10 targets	average promiscuity rate
1	953261 (73.8%)	365033 (28.3%)	169311 (13.1%)	6.7
2	622581 (66.4%)	224999 (24.0%)	107470 (11.5%)	6.7
3	418295 (69.1%)	137256 (22.7%)	43803 (7.2%)	4.4
4	418265 (69.1%)	137241 (22.7%)	43802 (7.2%)	4.4
5	63007 (35.2%)	3632 (2.0%)	1351 (0.8%)	2.1
6	45431 (30.2%)	1468 (1.0%)	348 (0.2%)	1.5
7	45398 (30.2%)	1465 (1.0%)	347 (0.2%)	1.5
8	44731 (30.1%)	1441 (1.0%)	347 (0.2%)	1.5
	(b) Reverse Order			
	no. of promiscuous compounds (%)			
set	≥2 targets	>5 targets	>10 targets	average promiscuity rate
1′	953261 (73.8%)	365033 (28.3%)	169311 (13.1%)	6.7
2′	630097 (60.4%)	108668 (10.4%)	27262 (2.6%)	3.3
3′	430327 (51.9%)	58186 (7.0%)	15360 (1.9%)	2.7
4′	304589 (45.3%)	27660 (4.1%)	6153 (0.9%)	2.2
5′	200176 (41.5%)	13282 (2.8%)	3106 (0.6%)	2.1
6′	120847 (36.6%)	5624 (1.7%)	1772 (0.5%)	1.8
7′	64755 (32.2%)	2195 (1.1%)	497 (0.2%)	1.6
8′	44731 (30.1%)	1441 (1.0%)	347 (0.2%)	1.5

[a] For each data set, the numbers (and percentages) of compounds annotated with at least two, more than five, or more than 10 targets are provided for the (a) original and (b) reverse order. In addition, the average compound promiscuity rate is reported for each set.

Table 7. Compounds with Potential PAINS Liability[a]

	(a) Original Order	
set	no. of PAINS (%)	average promiscuity rate
1	55447 (4.3%)	9.8
2	38663 (4.1%)	10.8
3	20799 (3.4%)	7.4
4	20755 (3.4%)	7.4
5	6939 (3.9%)	2.4
6	5482 (3.6%)	1.6
7	5448 (3.6%)	1.6
8	5280 (3.6%)	1.6
	(b) Reverse Order	
set	no. of PAINS (%)	average promiscuity rate
1′	55447 (4.3%)	9.8
2′	49760 (4.8%)	4.4
3′	37937 (4.6%)	3.9
4′	31845 (4.7%)	2.9
5′	22845 (4.7%)	2.4
6′	12133 (3.7%)	1.8
7′	7914 (3.9%)	1.6
8′	5280 (3.6%)	1.6

[a] For each data set, the number (and percentage) of compounds containing PAINS substructures (according to Figure 2a) are reported for the (a) original and (b) reverse order. In addition, the average promiscuity rate of these compounds is given.

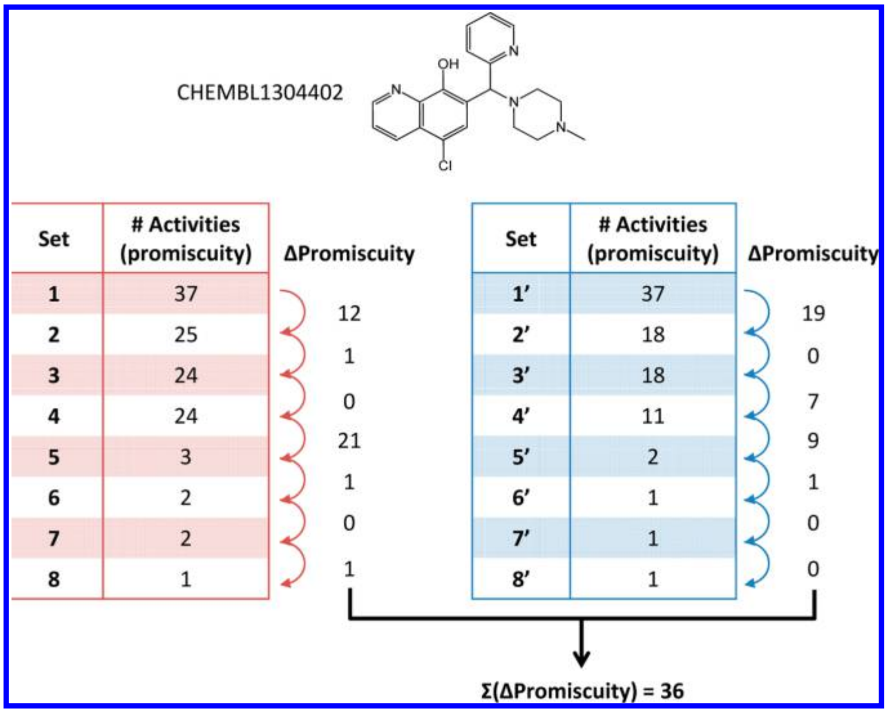
sets, as reported in Table 7a, sets 1 to 4 displayed an increase in promiscuity compared with the corresponding global (all-compound) rates in Table 6. For example, in set 1 the global rate of 6.7 increased to 9.8, in set 2 the rate further increased to 10.8, and in sets 3 and 4 the rate increased from 4.4 to 7.4. By contrast, in sets 5 to 8, which were characterized by higher activity data confidence, the promiscuity rates among PAINS-liable compounds and the global rates were comparable. Therefore, despite the presence of nearly 7000 PAINS-positive compounds in these sets, they did not tend to display more target interactions than the remaining compounds without characteristic PAINS substructures. By contrast, low-confidence activity data yielded many more interactions for PAINS-liable compounds, suggesting that PAINS-related effects might be controlled by applying rigorous data confidence criteria. This means that many PAINS-liable compounds will likely be eliminated from further consideration under these conditions because reliable measurements are not available.

In addition to ChEMBL compounds, we also determined the propensity of PAINS matches in drugs and screening hits. From DrugBank 3.0,[6] 1241 approved drugs with available target and structural information were assembled, 73 of which (nearly 6%) contained PAINS substructures, as reported in Table 8.

Table 8. Drugs and Screening Hits with PAINS Substructure Matches[a]

source	compound type		no. of compounds	no. of PAINS (%)
DrugBank 3.0	approved drugs		1241	73 (5.9%)
PubChem BioAssay	active	all	140112	7167 (5.1%)
		tested in ≥50 assays	108940	5482 (5.0%)
		tested in ≥50 assays; active in ≥30% of the assays	205	68 (33.2%)
	inactive	all	297176	4653 (1.6%)
		tested in ≥50 assays	179860	3005 (1.7%)

[a] For approved drugs assembled from DrugBank 3.0 and active or inactive compounds collected from PubChem confirmatory assays, the number (and percentage) of compounds containing PAINS substructures are reported.



Set	# Activities (promiscuity)	ΔPromiscuity
1	37	12
2	25	1
3	24	0
4	24	21
5	3	1
6	2	0
7	2	1
8	1	

Set	# Activities (promiscuity)	ΔPromiscuity
1′	37	19
2′	18	0
3′	18	7
4′	11	9
5′	2	1
6′	1	0
7′	1	0
8′	1	

Σ(ΔPromiscuity) = 36

Figure 4. Changes in promiscuity over different sets. For a selected compound (CHEMBL1304402), the degrees of promiscuity in different data sets are reported for both orderings, i.e., original order on the left and reverse order on the right. Differences in promiscuity (ΔPromiscuity) were calculated for pairs of adjacent sets. The sums of all differences (∑(ΔPromiscuity)) are also reported.

Furthermore, in 1085 confirmatory assays[21] collected from the PubChem BioAssay[2] database, 140 112 compounds were designated as "active" in one or more assays and 297 176 compounds were consistently designated as "inactive" in all assays in which they were tested (Table 8). Approximately 5.1% (active) and 1.6% (inactive) compounds were found to contain PAINS substructures. These percentages essentially remained constant for active or inactive compounds that were tested in 50 or more confirmatory assays. Thus, the proportion of PAINS-liable screening hits was generally higher than that of PAINS-liable bioactive compounds. However, ~33% of 205 screening hits that were tested in at least 50 assays and confirmed to be active in at least 30% of those assays were PAINS-positive. Hence, promiscuous screening hits from confirmatory assays contained PAINS substructures at a much higher rate than compounds from other sources, suggesting that particular care should be taken when considering promiscuous screening hits.

Table 9. Promiscuity Variation[a]

range of ∑(ΔPromiscuity)	no. of compounds (%)	no. of PAINS (%)
0	50316 (33.9%)	1597 (3.2%)
[1, 5]	77995 (52.6%)	1992 (2.6%)
[6, 10]	7973 (5.4%)	291 (3.6%)
[11, 20]	5755 (3.9%)	412 (7.2%)
[21, 50]	4526 (3.1%)	772 (17.1%)
[51, 100]	789 (0.5%)	127 (16.1%)
>100	1019 (0.7%)	89 (8.7%)

[a] For each ∑(ΔPromiscuity) value range, the corresponding number (and percentage) of compounds are reported. The percentages were calculated relative to all 148 373 compounds. For each value range, the number (and percentage) of compounds containing PAINS substructures are provided. Each percentage was calculated relative to the total number of compounds falling into the given value range.

Sequential Promiscuity Variation. In the next step, variation in promiscuity was analyzed for 148 373 compounds comprising high-confidence activity data set 8. For each of these compounds (occurring in all eight sets), its degree of promiscuity was monitored and compared over all sets, as illustrated in Figure 4. The exemplary compound displayed activity against 37 targets in set 1 and 25 targets in set 2. Thus, for this compound, the promiscuity difference (ΔPromiscuity) between these two sets was 12. The pairwise differences between adjacent sets (e.g., sets 1 and 2, sets 2 and 3, etc.) were calculated for each compound and summed to yield ∑(ΔPromiscuity). For example, for the compound shown in Figure 4, the ∑(ΔPromiscuity) value was 36. Then, all compounds in set 8 were ranked on the basis of ∑(ΔPromiscuity) in order to account for the variation of promiscuity across different sets. As reported in Table 9, over 98 000 compounds (~66%) yielded a nonzero ∑(ΔPromiscuity) value, indicating that these compounds showed varying degrees of promiscuity over one or more data sets. The majority of these compounds had only small ∑(ΔPromiscuity) values (ranging from 1 to 5), whereas 1019 compounds had values greater than 100. There were 10 compounds with ∑(ΔPromiscuity) values greater than 1000, as reported in Figure 5a. These compounds are shown in Figure 5b. The compounds in Figure 5 include marketed kinase inhibitors. Their high promiscuity rates can be rationalized by considering that these compounds have been extensively tested in many assays of which at least subsets must yield high-confidence data as a prerequisite for drug approval. Furthermore, since our analysis focused on human targets, interspecies promiscuity was not considered. Interestingly, among these top-10 compounds with apparently artificial target annotations were seven approved drugs (two of which also contained PAINS substructures, i.e., sunitinib and doxorubicin). Hence, in these cases, low-confidence target and activity data caused an

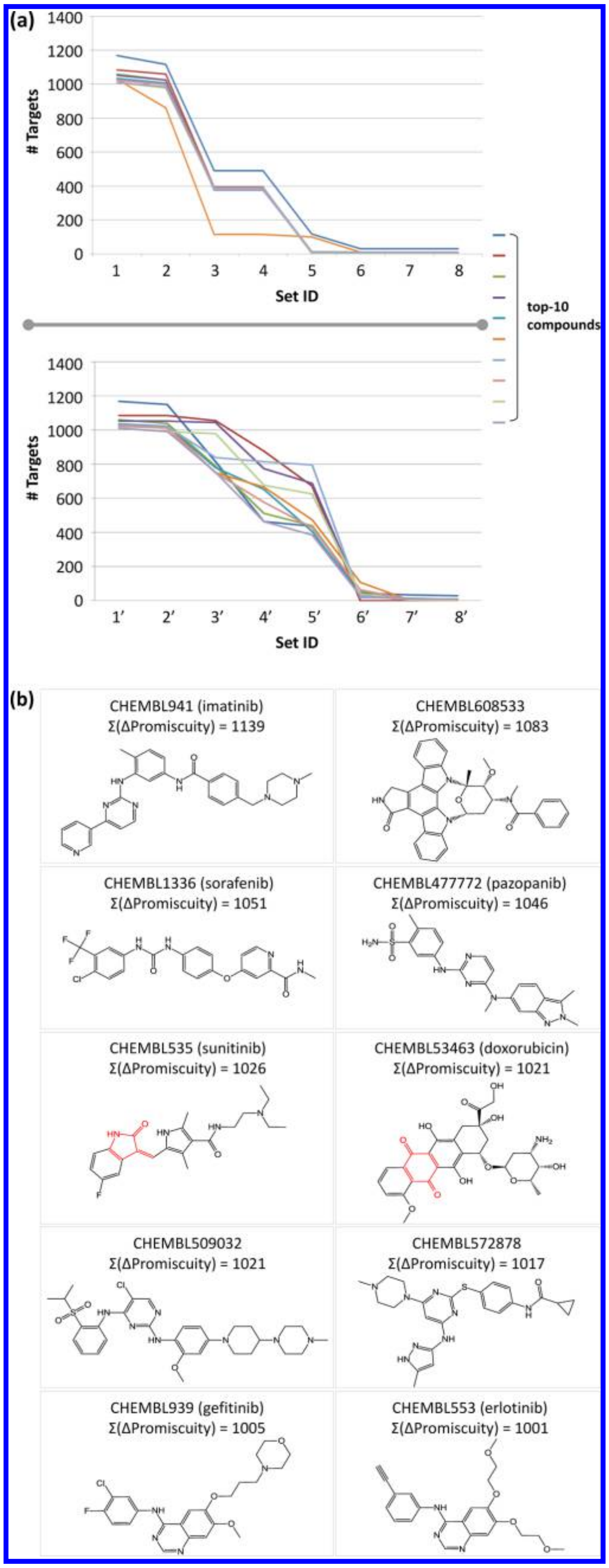


Figure 5. Top-10 compounds with largest changes in promiscuity. (a) Numbers of target annotations over all data sets for top-10 ChEMBL compounds with the largest changes in promiscuity for both orderings, i.e., original order at the top and reverse order at the bottom. (b) Structures of the top-10 compounds. For approved drugs, the name is provided in parentheses. $\sum(\Delta\text{Promiscuity})$ values are reported. In addition, for two compounds (approved drugs) with potential PAINS liability, detected structures (according to Figure 2a) are highlighted in red.

Figure 5. continued

inflation of apparent promiscuity rates. By contrast, over 50 000 compounds (~34%) displayed an essentially constant degree of promiscuity in all sets. These compounds were characterized by the exclusive (or nearly exclusive) presence of high-confidence activity data and the absence of additional activity information such as reported activity for orthologues. Figure 6 shows 12 exemplary compounds with constant promiscuity rates. In general, compounds with smaller variation in promiscuity were less likely to be PAINS-positive. As reported in Table 9, only ~3% of the compounds with $\sum(\Delta\text{Promiscuity})$ values ranging from 0 to 5 contained PAINS substructures. By contrast, from ~9% to 17% of the compounds having $\sum(\Delta\text{Promiscuity})$ values greater than 20 were PAINS-positive.

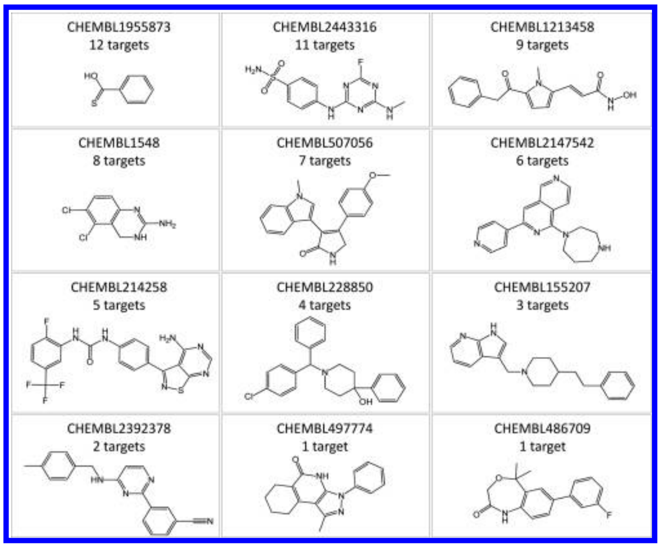


Figure 6. Compounds with constant promiscuity. Shown are 12 exemplary compounds that are active against varying numbers of targets. The degree of promiscuity of these compounds was conserved in all eight data sets.

Individual Parameter Contributions. In addition to assembling data sets under increasingly stringent selection criteria, additional data sets were generated by applying each individual criterion one at a time. It is re-emphasized that the composition of final selection sets through sequential application of filter criteria was order-independent, as demonstrated above. Moreover, data sets obtained by applying each criterion independently make it possible to assess the contribution of each individual parameter to the magnitude of the observed promiscuity. As reported in Table 10, the number of compounds in sets obtained after applying individual selection criteria ranged from 1 067 522 to 575 596. By applying the “target organism”, “target type”, “assay confidence level”, or “activity measurement type” criterion, the initial compound pool was reduced in size by more than 350 000 compounds. However, despite large differences in the numbers of compounds across the different sets, the proportion of

Table 10. Single-Parameter Data Sets[a]

parameters	no. of compounds	no. of PAINS (%)	average promiscuity rate		ΔPromiscuity
			all compounds	PAINS	
target organism	936924	38663 (4.1%)	6.7	10.8	4.1
target type	858742	30293 (3.5%)	4.5	7.2	2.7
assay confidence level	684061	24295 (3.6%)	4.7	7.8	3.1
activity measurement type	575596	28144 (4.9%)	2.5	2.9	0.4
activity relation type	1067522	47202 (4.4%)	3.2	4.2	1
activity unit	1064327	42879 (4.0%)	6.9	10.9	4
activity comment	1043487	49760 (4.8%)	3.3	4.4	1.1

[a] Seven data sets were generated by separately applying individual selection criteria. For each data set, the number of available compounds and the number (and ratio) of compounds containing PAINS substructures are reported. In addition, average promiscuity rates were determined for all compounds and PAINS-liable compounds, respectively. The difference in promiscuity rates ("Promiscuity) between all compounds and PAINS-liable compounds is reported.

PAINS-liable compounds varied only slightly, i.e., ~3.5% to ~4.9% of the compounds contained PAINS substructures.

Furthermore, we determined the average promiscuity rates for the single-parameter sets, which ranged from 2.5 to 6.9, as reported in Table 10. Compared with the initial compound pool, the degree of compound promiscuity differed from 0 to 4.2 targets across the single-parameter sets, as shown in Figure 7. Five of seven individual selection criteria (except "target organism" and "activity unit") were found to substantially affect the promiscuity rate. However, compared with the rate detected in the sets with highest confidence level (i.e., 1.5 targets for sets 8 and 8′), these data sets yielded much higher promiscuity rates, as to be expected given the large contributions of five individual selection criteria. Therefore, multiple parameters need to be applied to arrive at high-confidence promiscuity rates. In addition, Table 10 also reports the promiscuity rates for PAINS-liable compounds, which were consistently higher than the global rates (by, on average, 0.4 to 4.1 targets). However, in the highest-confidence sets (sets 8 and 8′), the global rates (i.e., 1.5) and promiscuity rates among PAINS-liable compounds (i.e., 1.6) were comparable. These findings further support the observation that low-confidence activity data yielded more interactions for PAINS-positive compounds. Therefore, the effect of PAINS liability on the assessment of compound promiscuity should be controlled by applying stringent compound selection criteria.

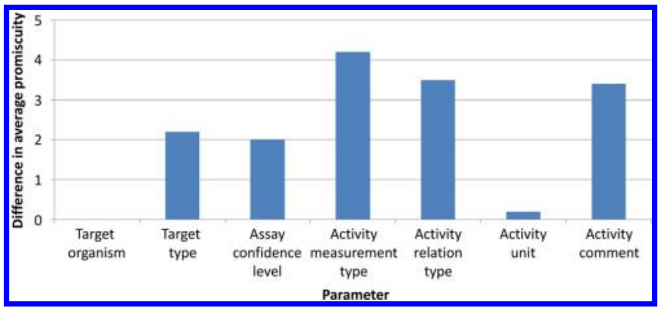


Figure 7. Influence of individual criteria on compound promiscuity. Shown are the differences in average promiscuity rates between the largest data set (i.e., sets 1 and 1′) and sets resulting from the application of individual selection criteria.

■ CONCLUDING REMARKS

From the ChEMBL database, eight sets of compounds were generated on the basis of defined selection criteria with increasing stringency covering a variety of data query parameters that are available to any user of the database. From set 1 to set 8, the numbers of compounds, targets, and interactions decreased significantly. This is important for database utility. Our analysis reveals that the combined application of these selection criteria profoundly influences data retrieval. Moreover, individual criteria that most strongly influenced the selection of compound activity data at different confidence levels were identified. Compound promiscuity was applied as an exemplary measure as to how data selection criteria at different stringency levels influence the results of data mining and conclusions drawn from such studies. Reversing the order in which compound selection criteria were applied did not change the final high-confidence data selection sets. This would be expected for sequential application of individual filter criteria in alternative orders and was confirmed in this analysis. In our experience, the types of activity measurements used are often not specified in compound data mining investigations, and assay confidence criteria are often not considered. The results of our analysis clearly show that these parameters must be carefully considered when selecting bioactive compounds. In addition, compounds with potential PAINS liability were detected and their promiscuity rates determined. When low-confidence activity data were applied, compounds containing PAINS substructures were found to have higher apparent promiscuity than other data set compounds. In contrast, when only high-confidence data were considered, such differences were not detected, which further emphasizes the critical role of data integrity for data mining. Furthermore, the majority of compounds entering the final, most stringently defined set 8 displayed varying degrees of promiscuity over different sets and a steady reduction in promiscuity when increasingly rigorous data confidence criteria were applied. We have analyzed ChEMBL as a leading public repository of compounds and activity data from pharmaceutical research, and there is every reason to anticipate that the findings reported herein will extend to other compound collections.

The critical role of high data confidence criteria for compound selection has been established by applying order-independent sequences of data selection criteria available in ChEMBL, which has been a central aspect of our investigation. These selection criteria can be applied by any user but require careful consideration. However, the problem of data integrity and conclusions drawn from data analysis are often not sufficiently addressed by considering data selection criteria available to users. This aspect has further implications that go beyond ChEMBL filtering criteria. For example, let us consider the types of activity measurements that are available. According to the selection criteria for sets 5 and 6 in Table 1, all compounds with precise K_i and/or IC_{50} potency measurements are selected, leading to the reduction of compounds in selection sets to those having well-defined potency measurements. Furthermore, in sets 7 and 8, criteria for accepting potency annotations are further refined, which represents the highest level of activity data confidence that can be achieved by rigorous applications of filtering criteria available in ChEMBL. However, for many applications this is not sufficient, and additional criteria must be considered. Importantly, however, it must be considered that assay-independent equilibrium constants (K_i values) and assay-dependent IC_{50} values, both of which are contained in the highest-confidence ChEMBL selection sets, are not directly comparable and that they must often be considered separately, depending on the application.[10] For example, if only the confirmed activity of a compound but not the magnitude of potency values is of relevance, such as for machine-learning-based compound classification (active vs inactive), K_i and IC_{50} measurements might be jointly considered. In contrast, for SAR analysis or any quantitative predictions of absolute potency values, they must be considered separately. Hence, to assure application-dependent data integrity, filtering according to criteria available in ChEMBL (or any other database) might not be sufficient, and additional data selection criteria (that depend on user knowledge) must often be applied.

There are more implications that go beyond what can be achieved by application of filtering criteria available in ChEMBL. For example, compounds with designated millimolar activities should best be omitted from any analysis or modeling efforts, even if defined measurements are available. This is the case because such activities are at or beyond the detection limits of many bioassays and prone to artificial readouts. Moreover, the availability of multiple well-defined activity measurements should be carefully investigated. If there are substantial differences between multiple reported K_i or IC_{50} measurements for given compounds that are active against the same targets (e.g., differences of 1 or 2 orders of magnitude, depending on the activity value distributions and assay statistics), the compounds should also be omitted from further consideration for all quantitative applications. This represents another generally important aspect of data integrity that is of high relevance for data mining and selection for computational modeling.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Hu, Y.; Bajorath, J. Learning from "Big Data": Compounds and Targets. *Drug Discovery Today* **2014**, *19*, 357−360.

(2) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.

(3) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein−Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(4) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.

(5) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.

(6) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A Comprehensive Resource for "Omics" Research on Drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.

(7) Williams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; Mons, B. Open PHACTS: Semantic Interoperability for Drug Discovery. *Drug Discovery Today* **2012**, *17*, 1188−1198.

(8) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool To Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(9) Hu, Y.; Bajorath, J. Many Structurally Related Drugs Bind Different Targets Whereas Distinct Drugs Display Significant Target Overlap. *RSC Adv.* **2012**, *2*, 3481−3489.

(10) Hu, Y.; Bajorath, J. Growth of Ligand−Target Interaction Data in ChEMBL Is Associated with Increasing and Measurement-Dependent Compound Promiscuity. *J. Chem. Inf. Model.* **2012**, *52*, 2550−2558.

(11) Stumpfe, D.; Bajorath, J. Assessing the Confidence Level of Public Domain Compound Activity Data and the Impact of Alternative Potency Measurements on SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 3131−3137.

(12) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K_i Data. *J. Med. Chem.* **2012**, *55*, 5165−5173.

(13) Olah, M. M.; Bologa, C. G.; Oprea, T. I. Strategies for Compound Selection. *Curr. Drug Discovery Technol.* **2004**, *1*, 211−220.

(14) Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. Mining the National Cancer Institute's Tumor-Screening Database: Identification of Compounds with Similar Cellular Activities. *J. Med. Chem.* **2002**, *45*, 818−840.

(15) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: Data Management and Interface Design. *Bioinformatics* **2002**, *18*, 130−139.

(16) Hu, Y.; Bajorath, J. Compound Promiscuity—What Can We Learn from Current Data. *Drug Discovery Today* **2013**, *18*, 644−650.

(17) Lu, J.-J.; Pan, W.; Hu, Y.-J.; Wang, Y.-T. Multi-Target Drugs: The Trend of Drug Research and Development. *PLoS One* **2012**, *7*, No. e40262.

(18) Jalencas, X.; Mestres, J. On the Origins of Drug Polypharmacology. *Med. Chem. Commun.* **2012**, *4*, 80−87.

(19) Hopkins, A. L. Network Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682−690.

(20) Peters, J.-U. Polypharmacology—Foe or Friend? *J. Med. Chem.* **2013**, *56*, 8955−8971.

(21) Hu, Y.; Bajorath, J. What is the Likelihood of an Active Compound To Be Promiscuous? Systematic Assessment of Compound Promiscuity on the Basis of PubChem Confirmatory Bioassay Data. *AAPS J.* **2013**, *15*, 808−815.

(22) Baell, J. B.; Holloway, G. A. 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*, 2719−2740.