

Analysis of Multitarget Activities and Assay Interference Characteristics of Pharmaceutically Relevant Compounds

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

SWARIT JASIAL

aus Hoshiarpur, Indien

Bonn

January, 2019

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Thomas Schultz
Tag der Promotion: 26th April, 2019
Erscheinungsjahr: 2019

Abstract

The availability of large amounts of data in public repositories provide a useful source of knowledge in the field of drug discovery. Given the increasing sizes of compound databases and volumes of activity data, computational data mining can be used to study different characteristics and properties of compounds on a large scale. One of the major source of identification of new compounds in early phase of drug discovery is high-throughput screening where millions of compounds are tested against many targets. The screening data provides opportunities to assess activity profiles of compounds.

This thesis aims at systematically mining activity data from publicly available sources in order to study the nature of growth of bioactive compounds, analyze multitarget activities and assay interference characteristics of pharmaceutically relevant compounds in context of polypharmacology. In the first study, growth of bioactive compounds against five major target families is monitored over time and compound-scaffold-CSK (cyclic skeleton) hierarchy is applied to investigate structural diversity of active compounds and topological diversity of their scaffolds. The next part of the thesis is based on the analysis of screening data. Initially, extensively assayed compounds are mined from the PubChem database and promiscuity of these compounds is assessed by taking assay frequencies into account. Next, DCM (dark chemical matter) or consistently inactive compounds that have been extensively tested are systematically extracted and their analog relationships with bioactive compounds are determined in order to derive target hypotheses for DCM. Further, PAINS (pan-assay interference compounds) are identified in the extensively tested set of compounds using substructure filters and their assay interference characteristics are studied. Finally, the limitations of PAINS filters are addressed using machine learning models that can distinguish between promiscuous and DCM PAINS. Structural context dependence of PAINS activities is studied by assessing predictions through feature weighting and mapping.

Acknowledgments

I would like to express my gratitude to Prof. Dr. Jürgen Bajorath for giving me opportunity to pursue my PhD in his group. I am highly grateful for his invaluable guidance during my doctoral studies. Thank you for your scientific inspiration and continuous support. I would like to thank Prof. Dr. Thomas Schultz for being the co-referent of my thesis.

I would also like to offer my sincere thanks to Dr. Ye Hu and Dr. Martin Vogt for their guidance and useful suggestions in all our collaborative projects. I am highly thankful to all the members of LSI group for providing a nice and interactive working environment. Special thanks to my dear colleagues Dr. Andrew Anighoro, Filip Miljković, Dr. Tomoyuki Miyao, Dr. Ryo Kunitomo, Dimitar Yonchev, Huabin Hu, and J Jesús Naveja for all the good times we shared. Many thanks to Erik Gilberg and Thomas Blaschke for productive and pleasant collaborations. I would particularly like to pay gratitude to Dr. Shilva Kayastha who always helped me in everything since the first day I arrived in Bonn. She will always be remembered as an amazing friend.

Being a firm believer in God, I would like to thank Almighty for giving me courage and strength in every sphere of life. I am deeply grateful to all my friends in Bonn for always being there on my side and uplifting my spirits. Special thanks go to Pranika Singh Rana, Meemansa Sood, Arka Mallick, Gurnoor Singh and Aaqib Parvez. I would also like to express my gratitude to all my friends in India specially Abhinav Anand and Prateek Mahajan for always keeping faith and believing in me.

I thank Prof. Dr. Andreas Weber and Prof. Dr. Michael Gütschow for agreeing to be members in my doctoral committee. Last but not the least, I want to thank my parents and family who were always there to support and encourage me during my studies.

Contents

1	Introduction	1
1.1	Publicly Available Compound Databases	1
1.1.1	ChEMBL	2
1.1.2	PubChem	2
1.1.3	ZINC	3
1.1.4	DrugBank	3
1.2	Promiscuity	3
1.3	Assay Interference	4
1.4	Dark Chemical Matter	6
1.5	Data Representations	7
1.5.1	Molecular Descriptors	8
1.5.2	Molecular Fingerprints	8
1.5.3	Scaffolds	10
1.6	Similarity Assessment	11
1.7	Machine Learning	14
1.7.1	Support Vector Machines	14
1.7.2	Random Forests	17
1.7.3	Neural Networks	20
1.8	Thesis Outline	22
2	Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping	25
	Introduction	25
	Publication	27
	Summary	35
3	Determining the Degree of Promiscuity of Extensively Assayed Compounds	37
	Introduction	37
	Publication	39
	Summary	55

4	Dark Chemical Matter in Public Screening Assays and Derivation of Target Hypotheses	57
	Introduction	57
	Publication	59
	Summary	65
5	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	67
	Introduction	67
	Publication	69
	Summary	77
6	Machine Learning Distinguishes with High Accuracy between Pan-Assay Interference Compounds That Are Promiscuous or Represent Dark Chemical Matter	79
	Introduction	79
	Publication	81
	Summary	91
	Conclusion	93
	Bibliography	97

Chapter 1

Introduction

In pharmaceutical research and drug discovery, increasing numbers of compounds and activity data are becoming available with the advent of methods such as high-throughput screening (HTS)¹ where increasingly large libraries of compounds are screened against drug targets in a short period of time. The increase in data volumes has also been accompanied by increasing data complexity and heterogeneity indicating that the “big data” phenomena originating from biology and bioinformatics have also entered medicinal chemistry, though still at a lesser magnitude.²⁻⁵ These large volumes of complex activity data provide considerable challenges to drug discovery scientists.⁶ Although it is difficult to analyze rapidly growing numbers of compounds and publicly available activity data, it represents a valuable source of knowledge and provides opportunities to learn in the field of drug discovery.⁷ Several efforts have been made to build publicly available databases in order to store and maintain information concerning increasing numbers of compound structures and their biological activity records against different targets.

1.1 Publicly Available Compound Databases

Four major public compound repositories are discussed in detail in the following:

1.1.1 ChEMBL

ChEMBL is an annotated public database containing activity data for small drug-like bioactive compounds.^{8,9} It is maintained by European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL). It provides information regarding binding, functional, and ADMET (absorption, distribution, metabolism, excretion, toxicity) properties of large number of compounds.^{8,9} The core activity data is manually extracted from the published medicinal chemistry literature on a regular basis and then curated and standardized to enhance the quality and utility across different drug discovery problems.⁸ ChEMBL version 24 contains nearly 1.8 million compounds from medicinal chemistry sources that are active against 12,091 targets, forming a total of more than 15 million ligand-target interactions. ChEMBL also incorporates activity data from PubChem database.

1.1.2 PubChem

PubChem is a public repository for chemical structures and their activities against biological assays.¹⁰ It is administered by the US National Institutes of Health (NIH). The information in PubChem is organized into three related databases: Substance, Compound and BioAssay. The Substance database contains contributed sample descriptions provided by depositors whereas the Compound database contains unique chemical structures derived from the substance depositions.¹¹ The PubChem BioAssay database contains compound screening data.¹² The screening data is structured into three types of records: Summary, Primary and Confirmatory. A Summary record gives the overview of an experiment. A Primary record contains results from the initial screen in which the activity assessment is based on percentage inhibition from a single dose. A Confirmatory record reports the effective concentrations of compounds passing the primary screen based on multi-concentration dose-response behavior.¹¹ The PubChem BioAssay database comprises more than 1 million assays with nearly 3.4 million tested compounds, yielding a total of more than 237 million activity annotations covering more than 11,000 biological targets.

1.1.3 ZINC

ZINC is a publicly accessible database that collects compounds relevant for medicinal chemistry from vendor sources as well as other databases.¹³ It is maintained by the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). The current release of ZINC (ZINC 15) reports approximately 230 million commercially available compounds.¹⁴ ZINC compounds are widely used for virtual screening applications.^{13,14}

1.1.4 DrugBank

DrugBank is a freely accessible web resource containing information about drugs and drug candidates.¹⁵ It contains detailed drug records such as chemical, pharmaceutical and pharmacological data associated with comprehensive drug target information such as sequences, structures and pathways. The latest release of DrugBank contains 11,885 drug entries; including 2528 FDA approved small molecule drugs and 5132 unique protein sequences.¹⁶

1.2 Promiscuity

Large volumes of compounds and activity data present in the databases can be used to analyze structure-activity relationships on a large scale which can help in chemical optimization. It is also possible to study the binding characteristics of different targets¹⁷ and systematically explore multitarget activities of small molecules on the basis of available data.⁷ These multitarget activities provide the foundations of polypharmacology. Polypharmacology is an emerging concept in drug discovery according to which many pharmaceutically relevant compounds elicit their therapeutic effects by acting on multiple biological targets.¹⁸⁻²² The molecular basis of polypharmacology is provided by compound promiscuity, which is defined as the ability of compounds to specifically interact with multiple targets.²³⁻²⁵ Promiscuity can be estimated computationally by mining rapidly increasing amounts of compound activity data and systematically assembling target annotations for compounds.^{21,24,25} However, data integrity and confidence levels have to be carefully taken into consideration

in order to arrive at reliable promiscuity estimates.²⁶ In recent years, several studies involving rigorous data-driven analysis have provided different promiscuity measures of bioactive compounds and drugs. It has been shown by data analysis that approved drugs directed against different target families bind to an average of two to seven targets.²¹ Furthermore, analysis of high-confidence data from ChEMBL for bioactive compounds indicated that they interact with an average of one to two targets and the most promiscuous compounds with on average two to five targets belonging to same target family.^{25,27} Drug and compound promiscuity were also monitored over time for a period of 14 years (2000 to 2014). Average degree of promiscuity of drugs extracted from Drug-Bank increased from 1.5 in 2000 to 3.2 in 2014 whereas it remained constant at 1.5 targets for bioactive compounds extracted from high-confidence data in ChEMBL, despite the massive growth of compound activity data during that time.^{28,29} Thus, computational data mining studies lead to the conclusions that active compounds and drugs have overall low degrees of promiscuity and drugs on average have a higher degree of promiscuity compared to bioactive compounds. Promiscuity estimates obtained from computational studies are often questioned because of data sparseness³⁰ as all active compounds are not tested against all the targets. However, given the large size of data samples and consistency in results, these observations might not be largely determined by data incompleteness and should be statistically meaningful.

1.3 Assay Interference

High-throughput screening is a key technology used by pharmaceutical industry to find potential drug candidates.¹ However, HTS data is always prone to false positives or false hits due to undesirable mechanisms of action.³¹⁻³³ Promiscuity should be clearly distinguished from these non-specific interactions or assay artifacts.³⁴ Not all the interactions between compounds and multiple targets make positive contributions to polypharmacology. It is important to identify compound classes that are frequently responsible for false positive assay readouts or “bad” promiscuity. These interaction artifacts are generally caused by compounds prone to colloidal aggregation³⁵⁻³⁷ or interference compounds that are highly reactive under assay conditions.^{38,39} Compounds responsible for as-

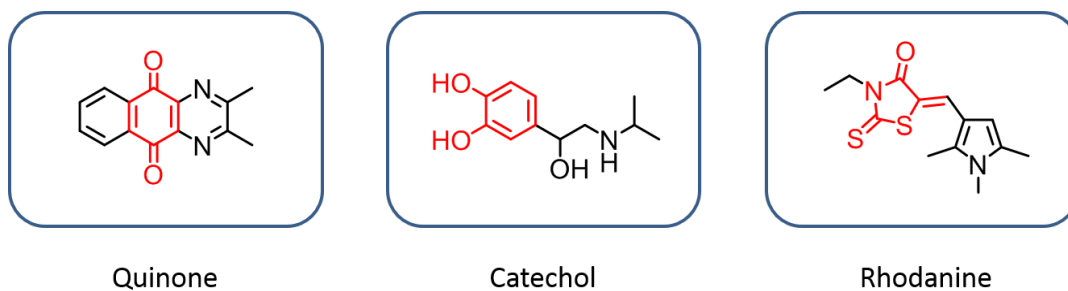


Figure 1.1: Exemplary PAINS. Compounds belonging to three different PAINS classes are shown. PAINS substructures are highlighted in red.

say interference have been found to originate from both synthetic and natural sources.^{40,41} Various mechanisms of compound-based assay interference include autofluorescence and quenching,⁴² covalent modification of proteins and assay reagents,⁴³ redox reactivity⁴⁴ or metal chelation.⁴⁵

Assay artifacts are often difficult to detect and are recognized in the later stages of drug discovery programs thereby leading to substantial loss of time and resources. False-positive activities as a result of assay interference spread throughout the scientific literature and cause problems in further investigations.⁴⁶ Therefore, it is important to filter compounds with liable mechanism of action before proceeding to biological or chemical optimization. Systematic efforts have been made to identify, select and filter compounds that can cause assay artifacts. In a landmark study, Baell and Holloway carried out an analysis of compounds that demonstrated activity in multiple AlphaScreen assays and put forward 480 chemical classes as candidates for assays interference.³⁸ Compounds containing these reactive chemical entities were referred to as pan-assay interference compounds.^{38,39,47} PAINS-defining moieties typically represent a compound's substructure and include various classes such as anilines, rhodanines, curcuminoids, Michael acceptors or Mannich bases. **Figure 1.1** shows exemplary compounds from three different PAINS classes namely quinones, catechols and rhodanines.

PAINS classes do not cover the entire spectrum of assay interference mechanisms, still they provide a basis for exploring compounds with potential in-

interference characteristics. The extrapolative power of PAINS filters has been called into question as they were derived from limited experimental data.^{48,49} Furthermore, evidence of PAINS has been seen in the marketed drugs^{50,51} indicating that overestimating the magnitude of assay interference can lead to exclusion of useful compounds with desired activities. Therefore, one has to be careful while filtering any potentially reactive compound as predicting assay interference requires thorough chemical knowledge and experience. Data-driven studies related to promiscuity and PAINS can help to better understand the interference potential of liable compounds.

1.4 Dark Chemical Matter

In HTS campaigns, considerable efforts have been made to design screening libraries focusing on chemical diversity and good quality of hits.^{52–54} Millions of compounds are subjected to screening and evaluated on the basis of activity profiles against diverse targets. However, it has been seen that large numbers of compounds in screening decks do not show any significant biological activity even after being tested in hundreds of biochemical or cellular assays.⁵⁵ These consistently inactive compounds either have specific properties that make them biologically inert or they have not been exposed to the appropriate target. In this context, a broad analysis of the bioactivities of small molecules in the Novartis and NIH Molecular Libraries screening collections⁵⁶ was carried out and a large fraction of compounds were found to be consistently inactive. The compounds which were lacking a biological profile despite having been screened in hundreds of high-throughput assays were termed as “dark chemical matter”.⁵⁷ DCM compounds were notably smaller, more soluble and less hydrophobic than other compounds in the screening library. Furthermore, when DCM compounds were tested in additional high-throughput screens, they were found to be active. Follow-up dose-response experiments confirmed that DCM compounds were potent hits.⁵⁷ This illustrates that although DCM compounds are less active than other screening compounds under normal HTS conditions, they are far from being biologically inert. In fact, DCM compounds might prove to be a potentially valuable resource for finding hits that are less likely to be false positives caused by assay interference. Thus, DCM yields less promiscuous

hits that might have unique selectivity towards a particular target and might become small molecule probes or lead candidates.

Some of the major concepts in chemoinformatics that have been used in this thesis for the analysis of data growth, promiscuity, assay interference and DCM compounds are discussed in next sections.

1.5 Data Representations

Molecules can be easily represented as graphs which are two-dimensional (2D) representations of chemical structures where nodes correspond to atoms and edges to bonds. 2D molecular graphs provide information about the connectivity between atoms and the topology of the molecules. They are simplified versions of molecular structures and are easily interpretable by medicinal chemists. They can be converted into machine interpretable form using connection tables consisting of atom coordinates, bond orders and hybridization states. However, storing large data sets of complex structures as molecular graphs is not computationally efficient. Therefore, chemical languages such as the simplified molecular-input line entry system (SMILES) were introduced to facilitate storage, retrieval, and modeling of chemical structures and chemical information.⁵⁸⁻⁶⁰ SMILES transform molecular graphs into strings of ASCII characters based on predefined rules for representing molecular structures. SMILES strings are compact compared to other methods of representing molecular structures and reduces the size of large databases considerably. Atoms are represented by their atomic symbols and branching is denoted by parentheses in SMILES notation. There are special symbols to denote chirality, isotopes, aromaticity and stereochemistry of a molecule. SMILES arbitrary target specification or in short SMARTS is an extension of SMILES which introduces atom and bond labels containing logical operators.⁶¹ It is generally used for searching patterns or substructure queries in databases. PAINS which are usually present as substructures in a molecule are represented in the form of SMARTS patterns. 2D molecular representations cannot account for spatial arrangement of atoms or conformations in a small molecule. For this purpose, three-dimensional (3D) representations such as pharmacophore models and molecular surfaces are used.

1.5.1 Molecular Descriptors

For many chemoinformatics applications, compounds are represented using molecular descriptors, which numerically describe molecular structure and properties. Different types of molecular descriptors are available to account for physicochemical, topological, surface and other properties of small molecules.⁶²⁻⁶⁴ Molecular descriptors can be classified as 1D, 2D or 3D based on the molecular representation from where they are derived.⁶⁵ 1D descriptors are calculated from the molecular formula and are very simple such as atom counts and molecular weights. 2D descriptors are calculated from molecular graphs, for example, models for estimating the water solubility of a compound are based on 2D representations. 3D descriptors are determined from molecular conformations.

1.5.2 Molecular Fingerprints

One of the most popular types of molecular descriptors are molecular fingerprints which are generally defined as bit string representations of molecular structure or properties. Each bit position usually accounts for the presence or absence of a given feature in a binary fingerprint although different types of fingerprints vary substantially in their design and length. If the feature is present in a molecule, the bit is set to '1' and if the feature is not present, it is set to '0'. There are also nonbinary versions of fingerprints where in addition to presence or absence of a feature, frequency of occurrence of a feature in a molecule is also monitored.^{66,67} Such fingerprints are referred to as count fingerprints. A variety of molecular fingerprints have been developed over the years that differ in composition, complexity and length.⁶⁸ Substructure-based fingerprints are one of the most common and widely used 2D fingerprints that represent dictionaries of predefined structural fragments. The length of these type of fingerprints is fixed. A classical example of substructure-based fingerprints is Molecular ACCess System (MACCS) structural keys,⁶⁹ the publicly available version of which consists of 166 bits, each of which accounts for the presence or absence of a structural pattern. **Figure 1.2** depicts an example of substructure-based fingerprint of fixed length 11.

Fingerprints representing molecular topology are another major class of struc-

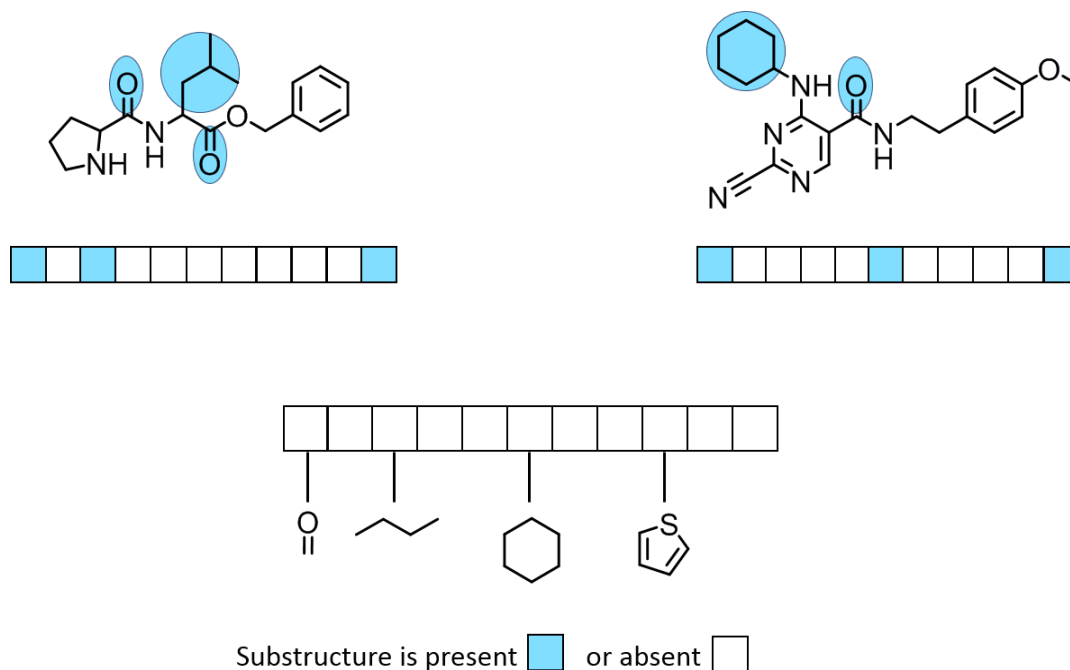


Figure 1.2: Substructure-based fingerprint. A substructural-type fingerprint representing 11 features is shown for two different compounds. If a feature is present in a compound, the corresponding bit is set on as indicated by sky blue color, otherwise it is set off. Substructural features that account for a particular bit to be set on are also highlighted in the compound structures. The figure has been adapted from reference [68].

tural 2D fingerprints. One prominent example of topological fingerprints is the extended connectivity fingerprint (ECFP).⁷⁰ It generates different layers of circular atom environments at a specified distance (depending on the specified bond diameter) centered around a non-hydrogen atom. **Figure 1.3** gives an illustration of how atom environments are calculated for topological fingerprints. In case of ECFP4, the maximum bond distance between atoms considered in the neighborhood of the central atom is four. Each unique environment is mapped to an integer using a hashing function and the collection of integers forms the fingerprint. ECFP4 can theoretically represent about 4 billion features. However, the features differ from molecule to molecule; hence they are variable in length. Both MACCS and ECFP4 fingerprints are frequently used to represent compounds in many machine learning applications.

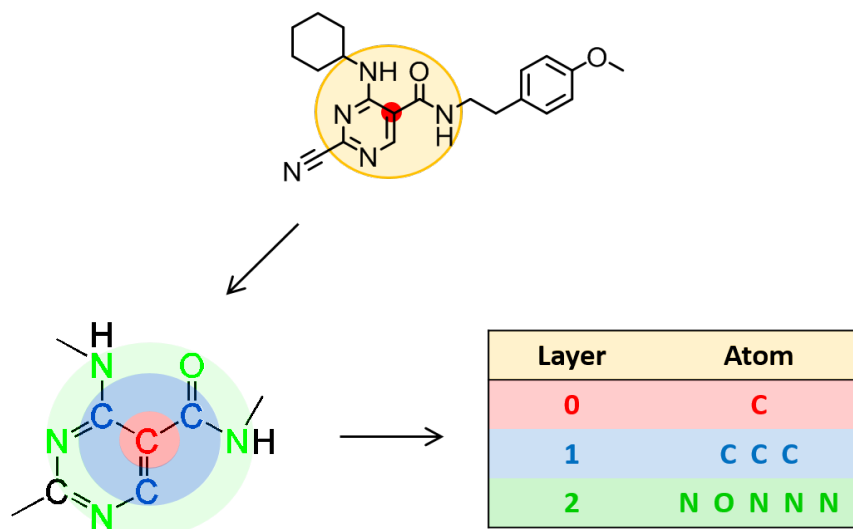


Figure 1.3: Topological fingerprint. Calculation of two atom environment layers with a carbon atom (red) at the center are shown. The first layer is colored in blue and the second layer in green. The resulting circular environments are hashed. The figure has been adapted from reference [68].

1.5.3 Scaffolds

Another important concept that is widely used in medicinal chemistry to describe core structures of bioactive compounds is the molecular scaffold. This concept is important for hierarchically classifying compounds present in large compound databases or screening libraries.⁷¹ In general terms, a “scaffold” refers to a molecular core to which functional groups are attached. The scaffold concept has several definitions in chemoinformatics and is applied in a subjective manner.^{72,73} The most extensively applied scaffold definition is the Bemis-Murcko scaffold (BM scaffold) according to which scaffolds are obtained from compounds by the removal of all non-ring R-groups while retaining all ring structures and linker fragments connecting the ring structures.⁷⁴ One can further abstract from chemical structures by generating cyclic skeletons from BM scaffolds by converting all heteroatoms to carbon and setting all bond orders to one. According to the molecular hierarchy, a BM scaffold can represent many compounds and a CSK can represent a set of topologically equivalent BM scaffolds.^{75,76} A compound-scaffold-CSK hierarchy is depicted in **Figure 1.4** with the help of two examples.

Different applications of the scaffold concept in medicinal chemistry include the search for privileged substructures or core structures that preferentially interact with a specific target family.⁷⁷ Furthermore, identification of structurally distinct compounds having similar activity, a task known as “scaffold hopping” is another major application of scaffold concept. Scaffold hopping refers to the ability of computational methods to recognize active compounds with different core structures.⁷⁸ Ligand-based virtual screening approaches start from known active compounds for a given target as reference molecules or search templates and try to identify novel compounds active against the same target. It is one of the major aims of virtual screening calculations to detect scaffold hops.⁷⁹

1.6 Similarity Assessment

The assessment of structural similarity of compounds is regarded as a key concept in chemoinformatics and drug discovery. Similarity assessment is generally required in order to relate structure and biological activity of compounds to each other. Also, several learning algorithms use similarity measures to quantify the similarity between two compounds. Computational similarity assessment mainly depends on the molecular representation and similarity metric used.^{80,81} The similarity between two compounds is typically calculated by a fingerprint comparison or in other words as overlap between fingerprint strings. For this purpose, many different similarity coefficients have been introduced such as Tanimoto coefficient (Tc), Dice coefficient (Dc), Tversky coefficient (Tv) and Cosine coefficient.^{81,82} The most popular measure of fingerprint similarity in chemoinformatics is the Tanimoto or Jaccard coefficient.^{82,83} For two binary molecular fingerprints A and B, the Tc is defined as

$$\text{Tc}(A, B) = \frac{c}{a + b - c}$$

where a and b are the number of bits set on in fingerprints A and B, respectively and c corresponds to the number of bits set on in both the fingerprints. Tc values range from 0 to 1 with 0 corresponding to no fingerprint overlap and 1 to

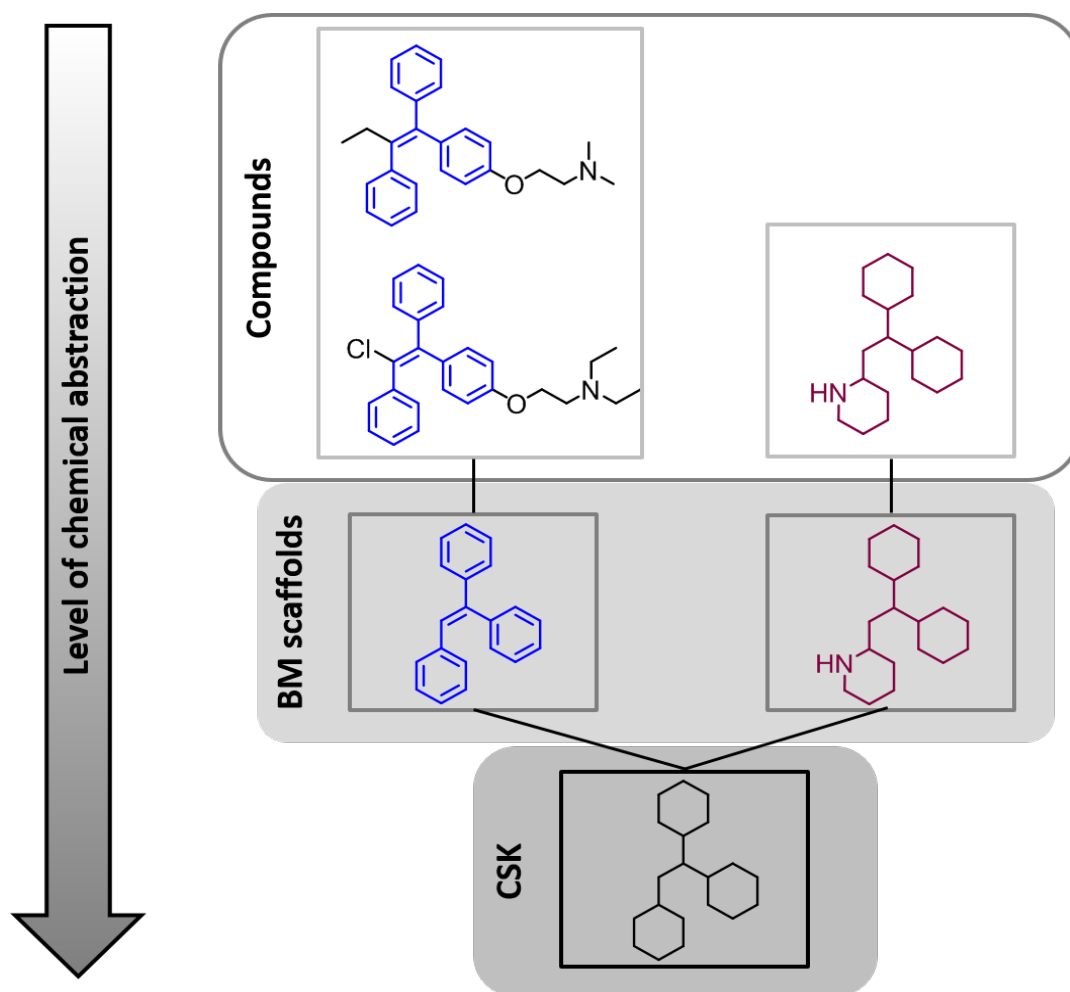


Figure 1.4: Compound-scaffold-CSK hierarchy. For three exemplary compounds, corresponding scaffolds and CSKs are shown. Two compounds on the left have the same scaffold highlighted in blue and the scaffold of the third compound is highlighted in maroon. These structurally distinct scaffolds are represented by the same CSK and therefore are topologically equivalent. The figure has been adapted from reference [76].

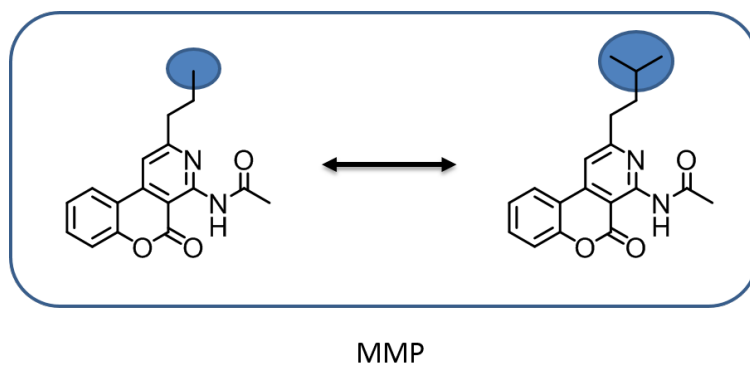


Figure 1.5: Matched molecular pair. Two compounds forming an MMP relationship are shown. The exchanged substructures representing a chemical transformation are highlighted in blue. The remaining part of the compounds represent the common core.

identical fingerprints but not necessarily identical molecules. Tc is an example of a symmetrical similarity coefficient, which means the value of comparing A to B is the same as for comparing B to A. The similarity value distributions of Tc values for MACCS and ECFP4 fingerprints are generally centered on a mean value of 0.4 to 0.6 and 0.25, respectively.⁸⁴

Another popular similarity concept for chemical structures in medicinal chemistry is the matched molecular pair (MMP), which is defined as pair of compounds that only differ by a structural change at a single site.^{85,86} This single modification to convert one compound to another is termed “chemical transformation”. **Figure 1.5** shows an example of two compounds forming an MMP.

MMPs are mostly generated by fragmentation or maximum common substructure-based algorithms. MMPs are chemically intuitive compared to numerical similarity measures as chemical transformations such as R-group replacements or core structure modifications can directly be associated with activity or other properties of a molecule. MMPs can further be extended to generate matched molecular series (MMSs) which are series of structurally analogous compounds that are only distinguished by chemical modifications at a single site.⁸⁷ MMPs are usually size-restricted and only small structural changes are allowed in chemical transformations. MMPs can also be generated on the basis of retrosynthetic fragmentation by applying reaction rules following the retrosynthetic combinatorial analysis procedure (RECAP).⁸⁸ These MMPs which are based on chemi-

cal reactions are termed as RECAP-MMPs.⁸⁹ Furthermore, global network representations can be generated in which nodes represent compounds and edges represent pairwise RECAP-MMP relationships. Each separate cluster in this network represents a unique analog series (AS).⁹⁰

1.7 Machine Learning

Machine learning methods are used to develop computational models that have the ability to learn from patterns in the data and make predictions. They have become increasingly popular in the field of drug discovery over the past years for classification of compounds and property predictions. Machine learning methods are extensively used in ligand-based virtual screening in order to rank database compounds.⁷⁹ Some of the most prominent machine learning methods are naïve Bayesian classification (NB),⁹¹ random forests (RF),⁹² neural networks (NN)⁹³ and support vector machines (SVM).⁹⁴

1.7.1 Support Vector Machines

Support vector machines are among the most widely used machine learning algorithms in chemoinformatics⁹⁵ mainly for compound activity predictions. SVM is a supervised learning method originally used for binary object classification⁹⁶ but has also been adapted for multitarget predictions⁹⁷ and compound ranking.⁹⁸ SVMs have gained popularity due to their ability to reach higher performance levels than other prediction methods in many applications.

During learning, SVM uses a set of n training instances $\{\mathbf{x}_i, y_i\}$ ($i = 1, \dots, n$) where $\mathbf{x}_i \in R^d$ is the feature vector and $y_i \in \{-1, 1\}$ is the class label (positive or negative) of a training compound i . Positive and negative training objects are projected into a feature space. A hyperplane H is derived that best separates positive and negative objects:

$$H = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$$

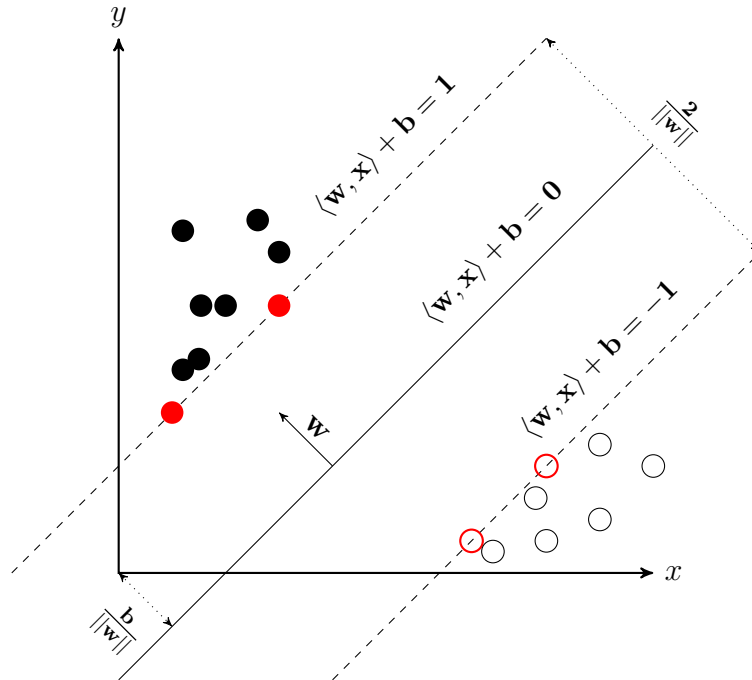


Figure 1.6: Support vector machine algorithm. A maximum margin hyperplane for classifying linearly separable data is derived. Positive and negative data are represented by filled and empty circles, respectively. The optimal hyperplane is shown by the solid black line. Data points determining the hyperplane or the support vectors are depicted by red circles.

where \mathbf{w} is the normal vector, b is the bias, and $\langle \cdot, \cdot \rangle$ is a scalar product.

As there can be many hyperplanes that can correctly classify linearly separable data, the SVM algorithm chooses an optimal hyperplane that maximizes the distance between the nearest training objects and the hyperplane. This distance is called as margin. **Figure 1.6** shows a representation of an SVM for linearly separable data. In this figure, positive data is represented by filled circles towards the y axis and negative data is represented by empty circles towards the x axis. $b/\|\mathbf{w}\|$ is the offset of the hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ from the origin. The two dotted hyperplanes $\langle \mathbf{w}, \mathbf{x} \rangle + b = 1$ and $\langle \mathbf{w}, \mathbf{x} \rangle + b = -1$ represent the two borders of the margin with $2/\|\mathbf{w}\|$ as the distance between them. In order to maximize this margin, $\|\mathbf{w}\|$ should be minimized. Positive and negative data on the borders of the margin (depicted by red circles) are called support vectors. Support vectors determine the position of the hyperplane.

For data that is not linearly separable, the parameters \mathbf{w} and b of the hyperplane are derived by solving the following optimization problem:

$$\text{minimize: } \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i$$

$$\text{subject to: } y_i (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i \quad \text{with } \xi_i \geq 0 \text{ and } i \in \{1, \dots, n\}$$

where ξ_i represent the slack variables⁹⁹ that are added to permit errors for training instances falling within the margin or on the incorrect side of the hyperplane and C is the cost or regularization hyperparameter introduced to balance training errors and margin size.

Instead of directly solving the primal optimization problem, it is also possible to formulate an equivalent dual problem using Lagrangian multipliers:¹⁰⁰

$$\text{maximize: } \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

$$\text{subject to: } \sum_{i=1}^n \lambda_i y_i = 0 \quad \text{with } 0 \leq \lambda_i \leq C \text{ and } i \in \{1, \dots, n\}$$

where λ_i are the Lagrangian multipliers. Dual expression makes it possible to compute the normal vector of the hyperplane as:

$$\mathbf{w} = \sum_{i=1}^n \lambda_i y_i \mathbf{x}_i$$

Lagrangian multipliers can be non-zero only for training examples that fall onto the margin of the hyperplane or are misclassified. This subset of training examples with non-zero coefficients falling onto the margin represents support vectors. Hence, the majority of training examples other than support vectors can be discarded following the training phase, which makes SVM modeling suitable for large data sets.

Test data is then projected into the feature space and classified depending

on the side of the plane onto which they fall, as determined by the following decision function:¹⁰¹

$$f(\mathbf{x}) = \text{sgn} \left(\sum_i \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right)$$

Test points with $f(\mathbf{x}) = 1$ are classified as positive and $f(\mathbf{x}) = -1$ are classified as negative. Furthermore, the decision function can be transformed to a ranking function by removing the signum function from the above equation and thus, generating real values for test examples:

$$g(\mathbf{x}) = \sum_i \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b$$

The dual formulation also enables the use of the “kernel trick”,¹⁰² which is of critical relevance for SVM modeling. If linear separation of training classes in a given feature space is not possible, the scalar product $\langle \cdot, \cdot \rangle$ is replaced by a kernel function $K(\cdot, \cdot)$, which corresponds to evaluating the scalar product in higher dimensional space (without an explicit feature representation in that space). A variety of kernel functions have been developed, including the Gaussian or radial basis function kernel, the Tanimoto kernel, and more complex graph kernels.^{103,104} The Tanimoto kernel, defined in accordance with the Tanimoto coefficient, is one of the most popular kernel function in chemoinformatics. It is defined as follows for two compound fingerprints \mathbf{u} and \mathbf{v} :

$$K(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle}$$

SVMs utilizing kernels usually have much higher prediction capacity compared to linear models. However, the use of kernel functions comes at the price of lacking model interpretability due to black box character of the resulting models.

1.7.2 Random Forests

Random Forest is a machine learning classification method based on an ensemble of decision trees.⁹² Each tree is built from a bootstrapped sample of

training data using a recursive partitioning method. A random subset of features is considered during node splitting for the construction of trees, which avoids the presence of correlated trees because of feature dominance. The final outcome depends on the prediction by the majority of trees i.e., a consensus prediction. **Figure 1.7** shows an example of a small random forest of five decision trees. RF has proven to be a very successful method in chemoinformatics¹⁰⁵ in different contexts such as QSAR (Quantitative Structure-Activity Relationship) modeling¹⁰⁶ and predicting protein-ligand binding affinity.¹⁰⁷

Suppose we aim to build an ensemble of B trees $\{T_1(X), \dots, T_B(X)\}$, where $X = \{x_1, \dots, x_p\}$ is a p -dimensional vector of molecular descriptors. For training, let there be a set of n molecules, $D = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where X_i , $i = 1, \dots, n$, is a vector of descriptors and Y_i is the corresponding class label i.e., active or inactive or any property of interest. The training algorithm proceeds as follows:

1. From the training data, a bootstrap sample is drawn i.e., a random sample with replacement from n molecules.
2. For each bootstrap sample, a tree is grown where the best split is chosen from a randomly selected subset of descriptors at each node. The tree is grown to the maximum size and not pruned back.
3. The above steps are repeated until the required number of trees (B) are grown.

The training forest is then used to predict test data. The ensemble produces B outputs $\{\hat{Y}_1 = T_1(X), \dots, \hat{Y}_B = T_B(X)\}$ where \hat{Y}_b , $b = 1, \dots, B$, is the prediction for a molecule by the b th tree. Outputs of all the individual trees are combined to produce one final prediction. For classification tasks, \hat{Y} is the class label predicted by the majority of the trees whereas for regression, it is the average of individual tree predictions.^{92,106}

RF calculations have relatively low computational costs and a large number of trees can easily be generated. RF is computationally efficient even for very large numbers of descriptors.

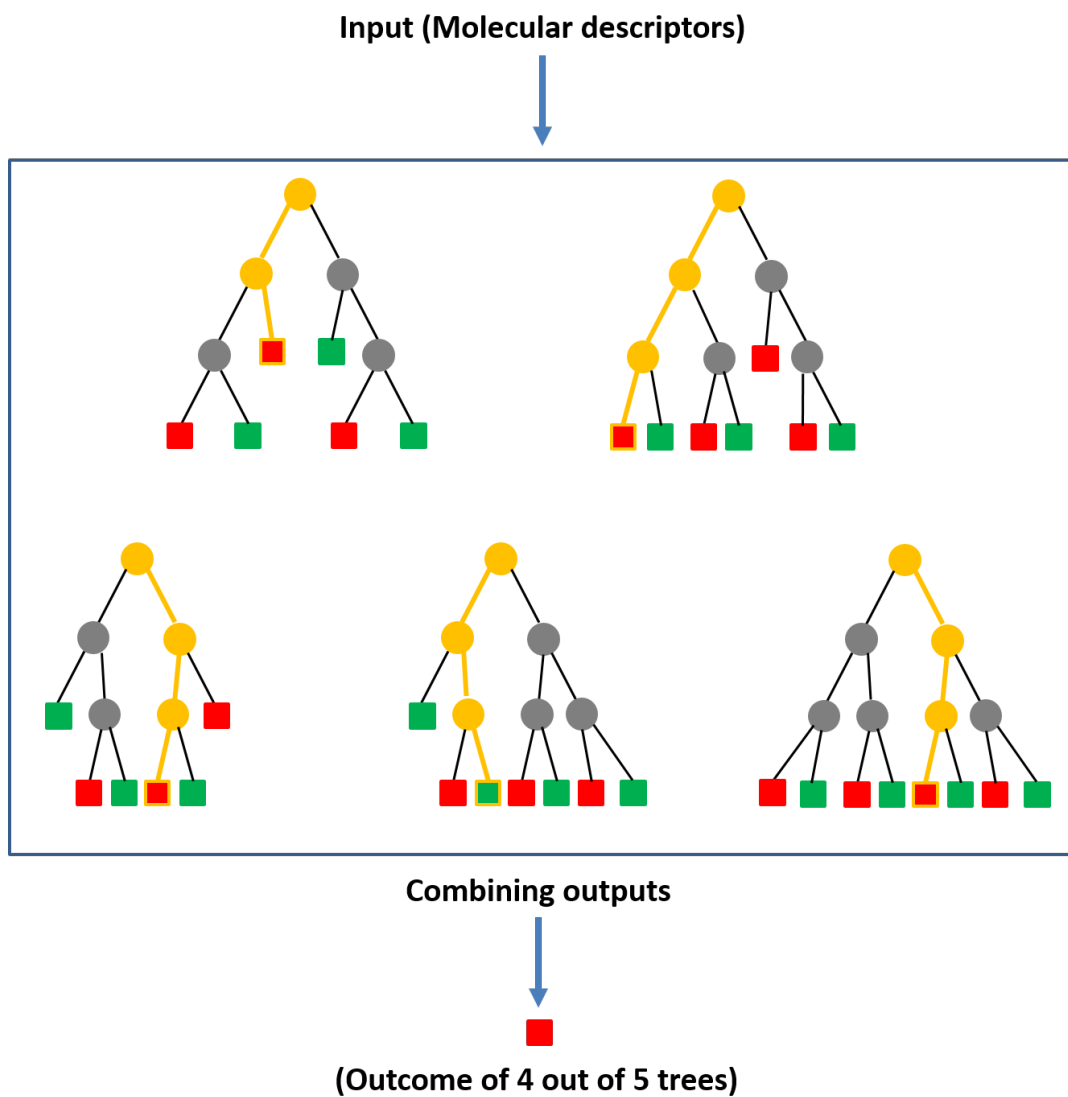


Figure 1.7: Random forest. Five exemplary decision trees are shown, which form a random forest for classification. The leaf nodes can either be red or green depending upon the class. The path taken by each test instance is highlighted in yellow. Four out of five decision trees predict that the test instance belongs to the red class, whereas one decision tree predicts that it belongs to the green class. Therefore, the final prediction is “red” by the random forest as it has the majority vote amongst the built trees. The figure has been adapted from reference [105].

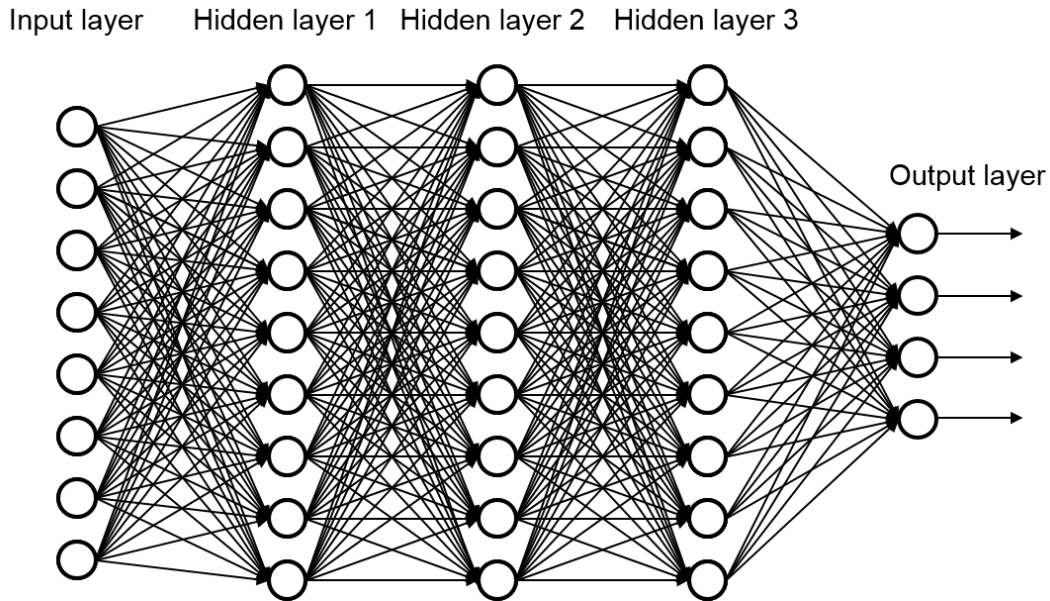


Figure 1.8: Deep Neural Network. A fully connected feed-forward deep neural network is shown with three hidden layers. Figure has been adapted from reference [108].

1.7.3 Neural Networks

An artificial neural network (ANN) is an interconnected group of nodes, similar to the vast network of neurons in brain. ANN consists of three basic layers: input, hidden and output layer. Input variables are taken by input nodes and the variables are transformed through hidden nodes, and in the end output values are calculated at output nodes. The output values of a hidden unit are calculated from input values via an activation function which is generally a nonlinear function to transform linear combination of input signal from input nodes to an output value. The output value Y_i of the node i is calculated as shown below:

$$Y_i = g \left(\sum_j W_{ij} * a_j \right)$$

where a_j are input variables, W_{ij} is the weight of input node j on node i and g is activation function. The training of ANN is done by modifying the weight values iteratively in the network.

Deep Neural Networks (DNNs)¹⁰⁸ contain larger number of hidden layers compared to traditional ANNs, which accounted for one or two hidden layers

due to limitation of computational power. Availability of more powerful CPUs and GPU hardware has allowed NN to use many more nodes in each layer. **Figure 1.8** shows an example of a fully connected DNN with three hidden layers. A fully connected deep feed-forward NN comprises of hundreds of non-linear process units in multiple hidden layers. DNNs can take large number of input features and the nodes can automatically extract features at different hierarchical levels.¹⁰⁹

In summary, a variety of machine learning concepts are applied in chemoinformatics and computational medicinal chemistry.

1.8 Thesis Outline

This thesis comprises five studies organized into individual chapters. The main focus of the studies has been to analyze the growth of bioactive compounds over time as well as to assess the activity profiles of compounds present in screening data. Promiscuity and assay interference of screening compounds are explored in detail. Methods are presented to deduce target hypotheses for inactive compounds and to refine and extend PAINS filters.

- *Chapter 2* presents a study focusing on the target-dependent growth of bioactive compounds and scaffolds over time. Structural diversity of compounds and topological diversity of scaffolds were explored on a time scale by applying compound-scaffold-CSK hierarchy. Implications for small molecule drug discovery were discussed.
- In *Chapter 3*, the promiscuity of screening compounds was analyzed with respect to assay frequency. Most extensively assayed compounds tested against hundreds of assays were extracted from PubChem and their promiscuity was systematically determined.
- In *Chapter 4*, consistently inactive or DCM compounds were systematically identified from screening data and other bioactive compounds were searched for analogs. Analog series were generated consisting of DCM and bioactive compounds and target hypotheses for DCM compounds were derived.
- *Chapter 5* presents a large-scale analysis of PAINS in biological screening assays. Activity profiles and hit rates of extensively assayed screening compounds detected by PAINS filters were studied in detail.
- In *Chapter 6*, machine learning models were introduced in order to distinguish between PAINS with high and low frequency of activity. SVM, RF and DNN models were trained using promiscuous and DCM PAINS data sets and ways to investigate structural context of PAINS were discussed.

Chapter 7 summarizes the major findings of all the studies in this thesis and contains concluding remarks.

Chapter 2

Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping

Introduction

In medicinal chemistry, the growth of bioactive compounds over time has recently been nearly exponential. This increasing volume of data provides a rich source of knowledge for exploring structure-activity relationships. Furthermore, activity data can be used to study interaction of ligands with therapeutic targets or promiscuity of bioactive compounds.

Activity data can also be used to determine the structural diversity of compounds globally as well as for individual targets. In this context, scaffolds are often used in order to represent the core structures of active compounds. Scaffolds are systematically extracted from active compounds by removing all substituents while retaining all ring structures and linker fragments between ring structures. As a further abstraction from scaffolds, CSKs are generated to focus on the molecular topology. The hierarchical organization from compounds to scaffolds to CSKs facilitates comparison of structures at different levels of abstraction.

In this work, we intended to explore the nature of compound data growth and investigate if the newly available active compounds for targets were structurally similar or diverse. The growth of bioactive compounds and scaffolds was monitored over a span of 15 years for five major target families. Scaffolds and CSKs were systematically extracted for compounds with high-confidence activity data from ChEMBL database and compound-scaffold-CSK hierarchy was employed to analyze structural diversity of compounds and topological diversity of scaffolds over time for all major target families.

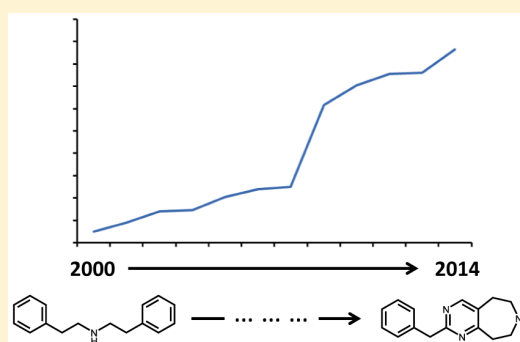
Reprinted with permission from “Jasial, S.; Hu, Y.; Bajorath J. Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping. *Journal of Chemical Information and Modeling* **2016**, *56*, 300-307”. Copyright 2016 American Chemical Society

Assessing the Growth of Bioactive Compounds and Scaffolds over Time: Implications for Lead Discovery and Scaffold Hopping

Swarit Jasial, 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

ABSTRACT: The increase in compounds with activity against five major therapeutic target families has been quantified on a time scale and investigated employing a compound–scaffold–cyclic skeleton (CSK) hierarchy. The analysis was designed to better understand possible reasons for target-dependent growth of bioactive compounds. There was strong correlation between compound and scaffold growth across all target families. Active compounds becoming available over time were mostly represented by new scaffolds. On the basis of scaffold-to-compound ratios, new active compounds were structurally diverse and, on the basis of CSK-to-scaffold ratios, often had previously unobserved topologies. In addition, novel targets emerged that complemented major families. The analysis revealed that compound growth is associated with increasing chemical diversity and that current pharmaceutical targets are capable of recognizing many structurally different compounds, which provides a rationale for the rapid increase in the number of bioactive compounds over the past decade. In light of these findings, it is likely that new chemical entities will be discovered for many small molecule targets including relatively unexplored ones as well as for popular and well-studied therapeutic targets. Moreover, given the wealth of new “active scaffolds” that have been increasingly identified for many targets over time, computational scaffold-hopping exercises should generally have a high likelihood of success.



INTRODUCTION

In pharmaceutical research, increasing volumes of compounds and activity data are becoming available. Not only data volumes but also complexity and heterogeneity are increasing, giving rise to the advent of big data phenomena in medicinal chemistry,^{1,2} similar to developments in biology and bioinformatics over the past decade,³ albeit still at lesser magnitude. Although large volumes of complex activity data are difficult to analyze, these data represent a valuable knowledge base for the large-scale exploration of structure–activity relationships and compound design.⁴ Analysis of activity data also helps to better understand ligand binding characteristics of therapeutic targets⁵ or promiscuity among bioactive compounds,^{6,7} which is defined as the ability of small molecules to specifically interact with multiple targets, a prerequisite for polypharmacological effects.^{8–10}

Activity data can also be related to structural classification schemes. For example, the scaffold concept has been applied over the last two decades to define core structures of compounds in a consistent manner.¹¹ Scaffolds are typically extracted from compounds by systematic removal of substituents.¹² Accordingly, a series of analogs yields the same scaffold. The scaffold concept has provided a basis for the generation of data structures such as the scaffold tree¹³ to systematically organize compound collections and annotate them with activity information. Scaffold-based compound organization can be extended through the generation of carbon skeletons, also termed cyclic skeletons (CSKs),¹⁴ which

represent a further abstraction from chemical structures focusing on molecular topology and enable the implementation of compound–scaffold–CSK hierarchies for structural organization and data analysis.¹⁵ A CSK represents a set of scaffolds that share the same topology and are only differentiated by heteroatom replacements and/or bond order variations. Scaffold and CSK analysis is often carried out to assess the structural diversity of compound collection, which is from a chemical perspective more intuitive than the calculation of descriptor-based similarity values.^{16–18}

The compound–scaffold–CSK hierarchy was previously employed by us to systematically explore structural relationships between scaffolds across bioactive compounds and study the potency range distribution of compounds sharing the same activity that were represented by different scaffolds.¹⁵ A major finding of this analysis was that many pairs of structurally distinct scaffolds represented highly potent compounds.¹⁵

The scaffold concept has also been applied to introduce “scaffold hopping”,^{19,20} which refers to computer-aided identification of compounds that share the same activity but differ in their core structures. Scaffold hopping through virtual compound screening is often regarded as one of the central tasks in computational medicinal chemistry.

We have been interested in exploring the nature of compound data growth in relation to scaffold growth and

Received: December 1, 2015

Published: February 2, 2016

diversity. How fast are volumes of bioactive compounds increasing and why are they increasing? Might the increase largely be due to extension of known compound series (perhaps reflecting a form of chemical “me-too-ism”)? Or is diversity generated among novel active compounds? Alternatively, might the increase be due to the emergence of novel targets for which new active compounds are identified? We have set out to explore these previously unaddressed questions.

Therefore, the increase in bioactive compounds over time was quantified for five major target families, and the compound–scaffold–CSK hierarchy was employed to characterize increasing volumes of bioactive compounds and analyze compound-to-scaffold ratios. For the first time, the growth of bioactive compounds and scaffolds extracted from them was followed on a time course over 15 years. This made it possible to monitor compound-to-scaffold ratios during periods of largest compound and activity data growth and compare the progression to earlier years when compound and data volumes were limited. A major and rather unexpected finding of our analysis has been that target-based growth of active compounds was consistently paralleled by increases in scaffold diversity across all major target families, independent of compound and data volumes. This has several implications for small molecule discovery as also discussed herein.

MATERIALS AND METHODS

Data Selection and Curation. Compounds and activity data were extracted from ChEMBL (release 20).²¹ Only compounds active against targets belonging to five major families were considered, including class A G protein-coupled receptors (GPCRs), ion channels, protein kinases, nuclear receptors, and proteases. These target families were organized following the UniProt²² and ChEMBL target classification schemes.

To ensure high data confidence, several preselection criteria were applied as implemented in ChEMBL. Compounds were extracted for which direct interactions (i.e., assay relationship type “D”) with human single-protein targets at the highest confidence level (assay confidence score 9) were reported. The two parameters, “assay relationship type” and “assay confidence score”, qualify and quantify the level of confidence that a compound is tested against a given target in a relevant assay system, respectively. Relationship type “D” and confidence score 9 indicate the highest level of confidence for activity data from ChEMBL. Furthermore, two types of potency measurements were considered including (assay-independent) equilibrium constants (K_i) and (assay-dependent) IC_{50} values. Only explicitly specified K_i and IC_{50} values were taken into account, and all approximate measurements such as “>”, “<”, or “~” were discarded. In addition, activity records with comments “inactive”, “inconclusive”, or “not active” were removed. Furthermore, compounds with activity records that did not contain publication dates were disregarded.

For a given target, activity data were examined to detect compound potency values reported in the same or different years. The following selection criteria were applied. Compounds with multiple potency measurements for the same year that differed by more than 1 order of magnitude were removed. In addition, compounds having multiple measurements in multiple years that differed by more than 1 order of magnitude were also discarded. However, if multiple measurements were reported in different years that fell into the same order of magnitude, the compound and the first reported potency value

were retained. For example, if a potency value of 4 nM was reported for a given compound in 2010 and 3 nM in 2011 for the same target, the compound was selected and 2010 potency value (4 nM) was assigned to the target. Furthermore, compounds with a single qualifying potency measurement were also retained.

Selected compounds and their activity data were assigned to individual years from 2000 to 2014 (all data reported prior to 2000 were assigned to year 2000).

The final curation step yielded all target-based compound sets for the five major families. Table 1 reports the total number

Table 1. Target Family-Based Compound Sets^a

target family	number of			
	all targets	all compounds	qualifying targets	qualifying compounds
GPCRs	165	46,905	153	46,885
kinases	276	21,756	176	21,699
ion channels	80	10,748	50	10,723
nuclear receptors	27	5032	24	5026
proteases	143	17,534	104	17,492

^aFor each family, the total number of targets and compounds available in ChEMBL and the number of targets and compounds qualifying for our analysis are reported.

of targets and compounds available for individual families. A target belonging to any one of these families was only considered if at least 10 active compounds were available. Accordingly, new targets or targets that had for other reasons very low compound coverage were omitted from further analysis. The number of these largely unexplored targets varied from three (nuclear receptors) to 100 (kinases). The number of compounds that were exclusively active against these targets and were also excluded from the analysis only ranged from six (nuclear receptors) to 57 (kinases). Hence, major target families contained novel and unexplored targets. For these targets, only small numbers of compounds were available, and their exclusion could not possibly bias the analysis given the large number of more than 100,000 qualifying compounds reported in Table 1. Moreover, when unexplored targets were omitted, a total of 507 qualifying targets with different degrees of chemical exploration remained that were associated with 101,825 active compounds including unique 99,216 molecules (and 2609 “promiscuous” compounds belonging to more than one target family). Given the large number of qualifying targets and compounds, the probability that the analysis might be biased by individual targets was extremely low.

Compound–Scaffold–Skeleton Hierarchy. For compounds of all target sets, a molecular hierarchy was generated. First, scaffolds were systematically extracted from active compounds by removing all substituents and retaining ring systems and linkers between them.¹² For each target, the scaffold-to-compound ratio was calculated by dividing the number of unique scaffolds available in a given year by the total number of compounds these scaffolds represented. Hence, a ratio of 1 meant that each compound contained a unique scaffold (reflecting highest possible scaffold-based diversity). In addition, cyclic skeletons (CSKs) were derived from scaffolds by converting all heteroatoms to carbon and all bond orders to single bonds.¹⁴ A given CSK might represent multiple scaffolds with conserved topology. For each target, the CSK-to-scaffold

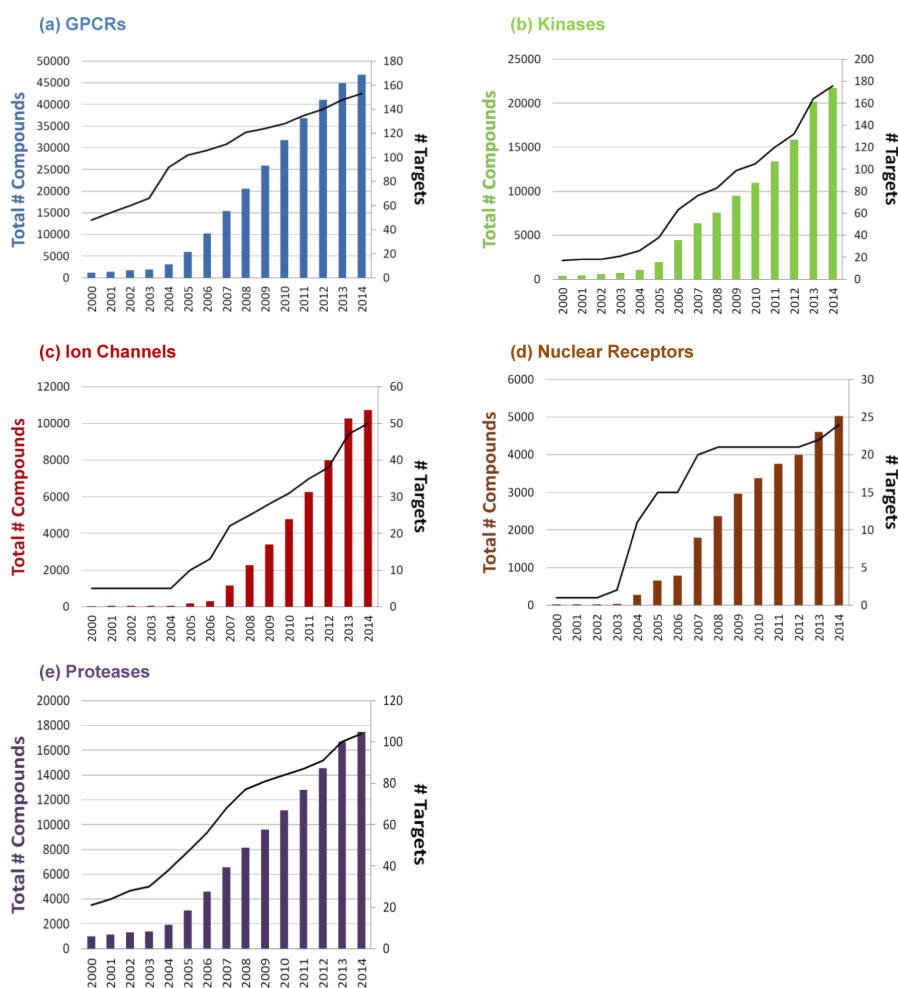


Figure 1. Growth of compounds and targets. In panels (a–e), the growth of compounds and targets is reported for the five target families. For each year, the cumulative number of compounds is shown using bar graphs (scale on the left vertical axis). In addition, the cumulative number of targets is traced using a black line (scale on the right axis).

ratio was also calculated by dividing the number of unique CSKs in a given year by the total number of scaffolds they represented. A ratio of 1 reflected the highest level of topological diversity within a scaffold set (i.e., all scaffolds were topologically distinct). All scaffolds and CSKs were calculated using in-house implementations that utilize the OpenEye toolkit.²³

RESULTS AND DISCUSSION

Analyzing the Increase in Active Compounds and Scaffolds over Time. The compound selection strategy applied here made it possible to follow target- and family-based growth of active compounds over time. For each year, the number of newly reported compounds and the cumulative number of active compounds were determined on a per-target basis and monitored for each family. Corresponding scaffolds and CSKs were also determined. The numbers of compounds, scaffolds, and CSKs were then related to each other, hence permitting a target family-based assessment of scaffold growth and topological diversity accompanying the increase in compound volumes over time.

Growth of Compounds and Targets. Figure 1 reports the family-based growth of active compounds and targets over

different years. Beginning in about 2006, a significant general increase in the number of compounds and targets was observed. For example, for class A GPCRs and the kinase family, the number of active compounds increased from 1227 to 46,885 (Figure 1a) and 380 to 21,699 (Figure 1b), respectively. The number of corresponding targets increased at different rates. The growth rate of targets from the class A GPCR family was slower than the growth of the corresponding compounds, as reported in Figure 1a. Steady increase in the number of targets was detected for the kinase, ion channel, and protease family, which paralleled the compound growth, as shown in Figure 1b, c, and e, respectively. A different pattern was observed for nuclear receptors where the number of targets significantly increased in 2004 and then essentially remained constant between 2008 and 2012 (Figure 1d).

Increase in Scaffolds. Figure 2 reports the increase in the number of scaffolds in target sets over time. Here, different observations were made. Both on the basis of the maximal and average number of scaffolds, striking scaffold growth was observed for all target families from 2000 to 2014. As a representative example, the number of available scaffolds for histamine H3 receptor increased from only one in 2000 to 1270 in 2014. Furthermore, compounds with HERG antitarget

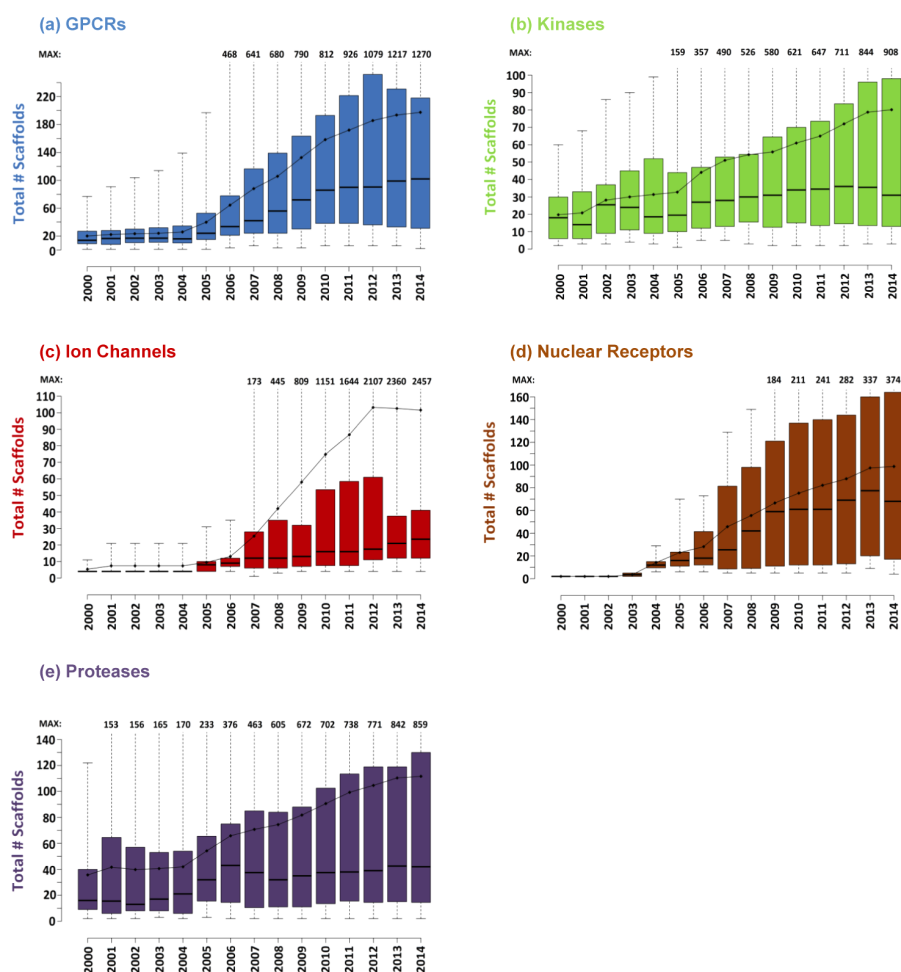


Figure 2. Growth of scaffolds. In panels (a–e), the growth of scaffolds is reported for the five target families using box plots for each year. A box plot gives the smallest number of scaffolds per year (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), and largest number of scaffolds per year (top line). The largest number of scaffolds per year is explicitly reported for years when this number exceeds the maximal value given on the vertical axis. The average number of scaffolds per year is traced using a black line.

activity were first detected in 2004. By 2014, HERG ligands were represented by 2457 scaffolds.

Interestingly, significant differences were evolving between the median and average number of scaffolds across all families, except nuclear receptors. The differences indicated that the scaffold distribution was increasingly dominated by a subset of the target sets with rapidly growing numbers of scaffolds, as indicated by the maximal numbers reported in Figure 2. On the basis of median numbers of scaffolds reported in Table 2, there was substantial overall growth in scaffold numbers since 2000.

Table 2. Scaffold Medians^a

target family	median number of scaffolds	
	2000	2014
GPCRs	14	102
kinases	18	31
ion channels	4	23.5
nuclear receptors	2	68
proteases	16	42

^aFor each family, the median number of unique scaffolds over all target sets is compared for 2000 and 2014.

However, different target families exhibited scaffold growth at varying magnitude. For example, the median number of scaffolds for class A GPCR targets increased from 14 in 2000 to 102 in 2014. Furthermore, in 2000, only 24 compounds represented by two scaffolds were detected having activity against a single nuclear receptor. However, by 2014, the median number of scaffolds for this family was 68 (Table 2). The overall smallest increase in the median number of scaffolds, from 18 to 31, was detected for kinases, although large increases were observed for a limited number of kinase sets, as reflected by the differences between median and average numbers.

The relationship between compound and scaffold growth was analyzed for all families and was found to be highly correlated, with correlation coefficients between 0.97 and 0.99. This strong correlation indicated that data increase was largely due to the addition of new compounds represented by new scaffolds. As a control calculation, the correlation between the number of compounds and scaffolds was also analyzed for all available targets, regardless of their family relationships. Nearly perfect correlation was also observed in this case.

Taken together, the findings revealed a steady and significant increase in the amount of scaffolds for all major target families,

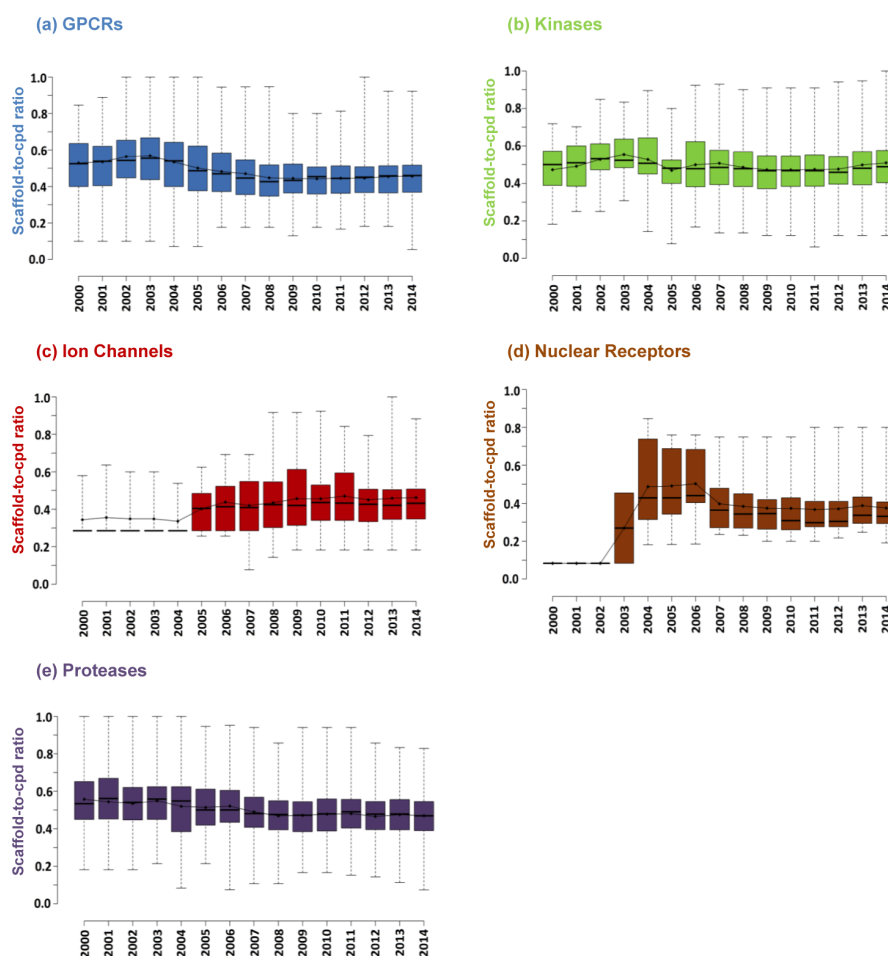


Figure 3. Scaffold-to-compound ratio. In panels (a–e), the scaffold-to-compound ratios are reported for the five target families in a box plot format according to Figure 2. The average ratio per year is traced using a black line.

which strongly correlated with the growth of compounds. These families included target sets for which very large numbers of scaffolds became available over time. New compounds becoming available each year were predominantly represented by new scaffolds.

Structural Diversity of Compounds. The scaffold-to-compound ratio was calculated to quantitatively assess the structural diversity of active compounds. Figure 3 reports the distribution of scaffold-to-compound ratios for targets from all five families. Different trends were observed. As shown in Figure 3a, b, and e, targets from the class A GPCR, kinase, and protease family displayed a wide range of scaffold-to-compound ratios. In these families, targets interacting with structurally diverse compounds (i.e., ratio close to 1) or structurally homogeneous compounds (ratio close to 0) frequently occurred. However, the median and average scaffold-to-compound ratios remained nearly constant over time, i.e., close to 0.5. Hence, on average, an individual scaffold represented two compounds across these target families.

By contrast, notable fluctuations in the distribution of the scaffold-to-compound ratios over time were observed for ion channels and nuclear receptors, as reported in Figure 3c and d, respectively. These fluctuations likely resulted from the presence of relatively small numbers of targets and active compounds for these two families during early years (giving rise

to statistical imbalances). In addition, the mean and average ratios were lower for these than the other three larger families. Hence, scaffold-based structural diversity of compounds active against different ion channels and nuclear receptors was generally limited.

Topological Diversity of Scaffolds. A CSK represents a set of topologically equivalent scaffolds. Thus, the CSK-to-scaffold ratio for a target set is an indicator of the degree of topological diversity among scaffolds. Distributions of the CSK-to-scaffold ratios for the five target families are reported in Figure 4. Average and median CSK-to-scaffold ratios were above 0.7 for most of the years. Fluctuations in these ratios over time and differences within and between target families were limited. Therefore, a high degree of topological diversity of scaffolds was observed for the majority of target sets.

Scaffold-to-Compound vs CSK-to-Scaffold Ratios. The CSK-to-scaffold ratios (Figure 4) were generally higher than the scaffold-to-compound ratios (Figure 3). On average, less than two scaffolds were represented by a given CSK. There was no detectable correlation between these two ratios over target sets and families (data not shown), indicating that high scaffold diversity (high scaffold-to-compound ratios) did not necessarily lead to topological diversity (high CSK-to-scaffold ratios) and vice versa. Figure 5 shows scaffolds from two exemplary target sets containing compounds with varying degrees of structural

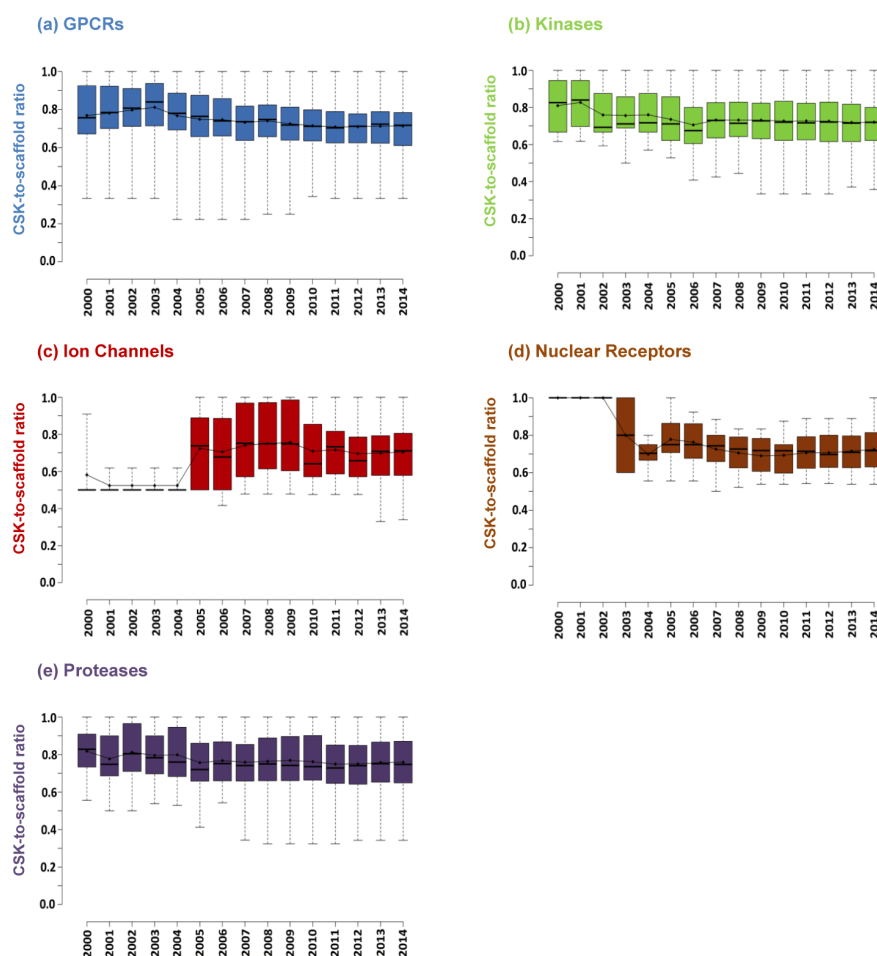


Figure 4. CSK-to-scaffold ratio. In panels (a–e), the CSK-to-scaffold ratios are reported for the five target families in a box plot format according to Figure 2. The average ratio per year is traced using a black line.

and topological diversity. In Figure 5a, scaffolds from beta-2 adrenergic receptor ligands are shown. Over time, a total of 355 compounds were found to be active against this receptor that were represented by 173 unique scaffolds and 137 unique CSKs. Accordingly, the scaffold-to-compound and CSK-to-scaffold ratios in 2014 were 0.49 and 0.79, respectively. Hence, in this case, a scaffold represented on average two compounds, and most of the scaffolds were topologically distinct. By contrast, Figure 5b shows scaffolds from inhibitors of MAP kinase-activated protein kinase 5 that yielded a much higher scaffold-to-compound ratio and much lower CSK-to-scaffold ratio, i.e., 0.95 and 0.50, respectively. Thus, in this case, scaffolds had equivalent topology more frequently, but most of the scaffolds represented only a single inhibitor. In particular, recurrent scaffold topologies were detected in 2012 and 2013. Scaffolds sharing these topologies only differed by one or more heteroatoms, as illustrated in Figure 5b. A variety of combinations between scaffold-to-compound and CSK-to-scaffold ratios were observed.

Implications and Conclusions. Herein, we have presented a systematic analysis of growth of compounds active against major target families and the scaffolds and CSKs these compounds contain. The analysis was inspired by our interest to better understand what might be major cause(s) for the increase in the number of compounds active against major

target families and how the increase might relate to chemical diversity and the emergence of new targets. The compound–scaffold–CSK hierarchy was employed as an indicator of structural and topological diversity, taking into consideration that boundaries between existing compound series and new structural classes are often fluid. Compound series representing a spectrum of structural relationships might often yield distinct scaffolds due to heteroatom substitutions in core structures or ring additions. However, the compound–scaffold–CSK hierarchy is a robust and consistently applicable analysis scheme to organize compound populations and assess structural diversity. As quantified in our analysis, there has been rapid growth of compounds active against major target families over the past decade. In addition, new targets have emerged over time complementing these families. As reported herein, compound growth is accompanied by a significant increase in the amount of scaffolds for all major target families. Importantly, new active compounds mostly contain new scaffolds. Hence, on the basis of scaffold-to-compound ratios, new active compounds are structurally diverse and, on the basis of CSK-to-scaffold ratios, frequently display new topologies. Therefore, the picture is emerging that major targets interact with many chemically diverse compounds, giving rise to substantial growth of bioactive compounds. Although an earlier study had shown that many compound activity classes were characterized by high

(a) Beta-2 adrenergic receptor

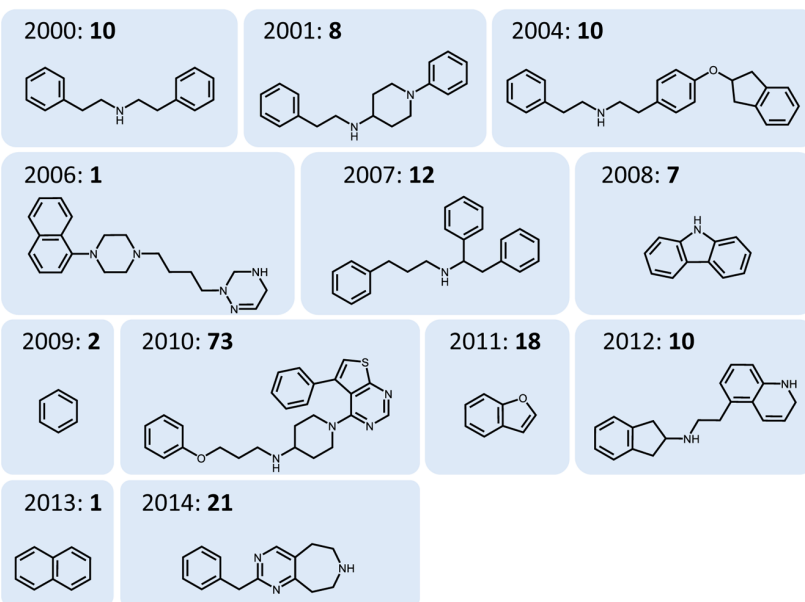
Compounds: 355

Scaffolds: 173

CSKs: 137

Scaffold-to-compound ratio: 0.49

CSK-to-scaffold ratio: 0.79

**(b) MAP kinase-activated protein kinase 5**

Compounds: 19

Scaffolds: 18

CSKs: 9

Scaffold-to-compound ratio: 0.95

CSK-to-scaffold ratio: 0.50

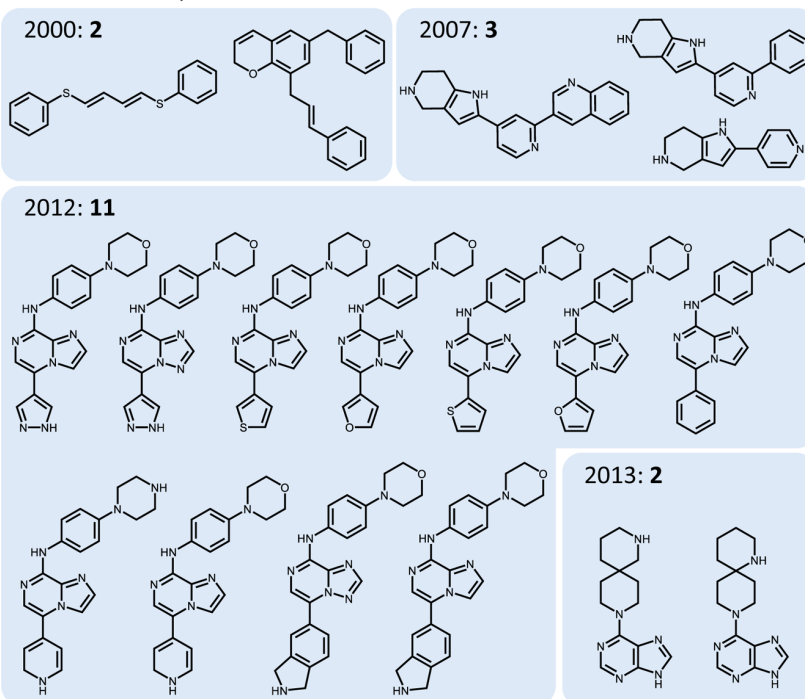


Figure 5. Exemplary scaffolds. Shown are scaffolds extracted from compounds active against (a) beta-2 adrenergic receptor and (b) MAP kinase-activated protein kinase 5. For each target, the total number of compounds, scaffolds, and CSKs available in 2014 is reported together with the scaffold-to-compound and CSK-to-scaffold ratio. For each year, the number of novel scaffolds is given in bold. For example, “2001: 8” means that eight new scaffolds became available in 2001. Panel (a) shows a representative scaffold per year. Panel (b) shows all new scaffolds for different years.

scaffold diversity,²⁴ the creation of chemical diversity as a major cause of compound growth was not anticipated. This also implies that current pharmaceutical targets are capable of

recognizing many structurally distinct compounds. This ability is essentially at the root of the rapid growth of bioactive compounds and exploited using increasing numbers of

bioassays. On the basis of these findings, it is likely that new active compounds with distantly related or novel structures will continue to be identified for major target families. By definition, most of newly identified compounds in recent years represent “scaffold hops”. This has implications for computational studies aiming at scaffold hopping. Since compound growth correlates with scaffold growth, predicting new active compounds with new scaffolds on the basis of known reference molecules should generally be a promising exercise, even for well-established targets, rather than a difficult task. The ultimate question will be which compound features might present “true” structural/chemical novelty; a question, however, which is essentially subjective in nature and difficult to address from first principles.

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.

ACKNOWLEDGMENTS

We thank OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits.

ABBREVIATIONS

CSK, cyclic skeleton; GPCR, G protein-coupled receptor

REFERENCES

- (1) Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-driven Medicinal Chemistry in the Era of Big Data. *Drug Discovery Today* **2014**, *19*, 859–868.
- (2) Hu, Y.; Bajorath, J. Learning from ‘Big Data’: Compounds and Targets. *Drug Discovery Today* **2014**, *19*, 357–360.
- (3) Schadt, E. E.; Linderman, M. D.; Sorenson, J.; Lee, L.; Nolan, G. P. Computational Solutions to Large-Scale Data Management and Analysis. *Nat. Rev. Genet.* **2010**, *11*, 647–657.
- (4) Hu, Y.; Bajorath, J. Compound Promiscuity: What Can We Learn from Current Data? *Drug Discovery Today* **2013**, *18*, 644–650.
- (5) Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130–137.
- (6) Hu, Y.; Bajorath, J. How Promiscuous are Pharmaceutically Relevant Compounds? A Data-Driven Assessment. *AAPS J.* **2013**, *15*, 104–111.
- (7) Hu, Y.; Bajorath, J. Monitoring Drug Promiscuity over Time. *F1000Research* **2014**, *3*, 218.
- (8) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
- (9) Boran, A. D.; Iyengar, R. Systems Approaches to Polypharmacology and Drug Discovery. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 297–309.
- (10) Jalencas, X.; Mestres, J. On the Origins of Drug Polypharmacology. *MedChemComm* **2013**, *4*, 80–87.
- (11) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742–1753.
- (12) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (13) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree—Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
- (14) Xu, Y.-J.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Model.* **2001**, *41*, 181–185.
- (15) Kayastha, S.; Dimova, D.; Stumpfe, D.; Bajorath, J. Structural Diversity and Potency Range Distribution of Scaffolds from Compounds Active against Current Pharmaceutical Targets. *Future Med. Chem.* **2015**, *7*, 111–122.
- (16) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17272–17277.
- (17) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **2011**, *51*, 2174–2185.
- (18) Taylor, R. D.; MacCoss, M.; Lawson, A. D. Rings in Drugs. *J. Med. Chem.* **2014**, *57*, 5845–5859.
- (19) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping” by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
- (20) Schuffenhauer, A. Computational Methods for Scaffold Hopping. *Wires Comput. Mol. Sci.* **2012**, *2*, 842–867.
- (21) 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.* **2012**, *40*, D1100–D1107.
- (22) UniProtConsortium. Reorganizing the Protein Space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71–D75.
- (23) OEChem, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012. <http://www.eyesopen.com> (accessed November 2015).
- (24) Hu, Y.; Bajorath, J. Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Targets. *MedChemComm* **2010**, *1*, 339–344.

Summary

A systematic analysis of compound data growth was carried out in relation to chemical diversity and emergence of new targets for five major target families. There has been a rapid increase in the number of bioactive compounds and targets for these target families in the past decade. Furthermore, a significant increase was found in the number of scaffolds over time and a strong correlation between compound growth and scaffold growth indicated that new active compounds were represented mostly by new scaffolds. Scaffold-to-compound ratios and CSK-to-scaffold ratios were calculated for all the targets in target families and their average and median values were close to 1. This implied that the new compounds available for target families were structurally diverse and new scaffolds displayed different topologies.

Hence, current pharmaceutical targets belonging to major target families are capable of interacting with structurally diverse compounds which might be a strong reason for the rapid increase in number of bioactive compounds over years. New compounds with distinct scaffolds will likely continue to be identified for these targets in future and will provide interesting insights for studies aiming at scaffold hopping.

The current analysis was based on the activity data from ChEMBL database which incorporates data mostly from medicinal chemistry literature. It provides activity annotations for a compound against a particular target but does not provide information about how many times a compound has been tested and against which targets. For this purpose, screening data available in PubChem BioAssay database can be utilized.

In the next chapter, we focus on the extraction of screening data from PubChem and determining the degree of promiscuity of the most extensively assayed public domain compounds.

Chapter 3

Determining the Degree of Promiscuity of Extensively Assayed Compounds

Introduction

Promiscuity is defined as the ability of small molecules to specifically interact with multiple targets. It can be rationalized as the molecular basis of polypharmacology, which aims at finding drugs with multitarget activities. As discussed in the previous chapter, large volumes of activity data are available in compound databases. Therefore, it is possible to estimate compound promiscuity through computational data mining. Promiscuity of drugs and bioactive compounds has so far been analyzed on the basis of activity annotations only, mainly from public domain databases such as DrugBank and ChEMBL where activity data is collected from literature. However, these databases do not provide information about assay frequency and inactivity records.

In this study, we extended the promiscuity analysis by taking assay frequency into account from screening data available in PubChem database so as to address the issue of data sparseness related to promiscuity estimates. It is not possible to directly extract assay frequency information on a per compound basis from PubChem database. Therefore, in the first step, data was curated to determine assay and activity profiles of screening compounds in primary

and confirmatory assays. In the next step, the most extensively assayed public domain compounds were identified and their promiscuity was systematically analyzed.

Reprinted with permission from “Jasial, S.; Hu, Y.; Bajorath J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PloS One* **2016**, *11*, e0153873”. Copyright 2016 PLOS

RESEARCH ARTICLE

Determining the Degree of Promiscuity of Extensively Assayed Compounds

Swarit Jasial, Ye Hu, Jürgen Bajorath*

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

* bajorath@bit.uni-bonn.de



CrossMark
click for updates

OPEN ACCESS

Citation: Jasial S, Hu Y, Bajorath J (2016) Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS ONE* 11(4): e0153873. doi:10.1371/journal.pone.0153873

Editor: Yoshihiro Yamanishi, Kyushu University, JAPAN

Received: January 14, 2016

Accepted: April 5, 2016

Published: April 15, 2016

Copyright: © 2016 Jasial et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data used for our analysis have been made freely available as a ZENODO deposition: Jasial S, Hu Y, Bajorath J. PubChem compounds tested in primary and confirmatory assays. ZENODO 2016; DOI: [10.5281/zenodo.44593](https://doi.org/10.5281/zenodo.44593).

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

In the context of polypharmacology, an emerging concept in drug discovery, promiscuity is rationalized as the ability of compounds to specifically interact with multiple targets. Promiscuity of drugs and bioactive compounds has thus far been analyzed computationally on the basis of activity annotations, without taking assay frequencies or inactivity records into account. Most recent estimates have indicated that bioactive compounds interact on average with only one to two targets, whereas drugs interact with six or more. In this study, we have further extended promiscuity analysis by identifying the most extensively assayed public domain compounds and systematically determining their promiscuity. These compounds were tested in hundreds of assays against hundreds of targets. In our analysis, assay promiscuity was distinguished from target promiscuity and separately analyzed for primary and confirmatory assays. Differences between the degree of assay and target promiscuity were surprisingly small and average and median degrees of target promiscuity of 2.6 to 3.4 and 2.0 were determined, respectively. Thus, target promiscuity remained at a low level even for most extensively tested active compounds. These findings provide further evidence that bioactive compounds are less promiscuous than drugs and have implications for pharmaceutical research. In addition to a possible explanation that drugs are more extensively tested for additional targets, the results would also support a “promiscuity enrichment model” according to which promiscuous compounds might be preferentially selected for therapeutic efficacy during clinical evaluation to ultimately become drugs.

Introduction

Polypharmacology is an emerging theme in pharmaceutical research [1–3]. It refers to increasing evidence that the therapeutic efficacy of many drugs depends on multi-target engagement. For example, this is by now well established for protein kinase inhibitors used in cancer therapy [4]. In the context of polypharmacology, compound promiscuity has been defined as the ability of small molecules to specifically interact with multiple targets [5,6], as opposed to engaging in non-specific or apparent interactions. Accordingly, so-defined promiscuity should not be confused with undesired pan-assay interference (PAINS) [7] or aggregator characteristic of compounds, giving rise to many false-positive assay readouts and doomed compound optimization

efforts. PAINS are typically reactive under assay conditions and the different types of undesired reactions associated with major classes of PAINS have been detailed [8]. Rather, promiscuity can be rationalized as the molecular basis of polypharmacology, which might also result in unwanted side effects due to specific target engagement.

Given the increasing sizes of compound databases and volumes of activity data, promiscuity of drugs and bioactive compounds can be estimated through computational data mining. Several studies have attempted to determine the numbers of targets drugs or bioactive compounds are known to be active against, focusing on premier public domain databases such as DrugBank [9], a major source of drug-target annotations, ChEMBL [10,11], the major public repository of compound activity data from medicinal chemistry, or the PubChem BioAssay collection [12], the major public repository of screening data, as well as various commercial compound databases. For example, surveys of drug targets have indicated that drugs interact on average with two to seven targets, depending on their primary target families and therapeutic areas, and that more than 50% of current drugs might interact with more than five targets [3]. On the basis of most recent estimates focusing on high-confidence activity data (i.e., well-defined single-target assays and precise activity measurements), approved drugs bind on average to 5.9 targets, whereas bioactive compounds from medicinal chemistry sources bind to 1.5 targets [13]. Interestingly, the average degree of compound promiscuity (i.e., average number of targets a compound is active against) was not notably higher for compounds active against major therapeutic targets such as G protein coupled receptors (GPCRs) or protein kinases [13]. Furthermore, mean degrees of promiscuity were not significantly higher for active compounds from confirmatory assays with, on average, 2.5 targets per compound [13,14]. Moreover, the degree of promiscuity of bioactive compounds covering the current spectrum of therapeutic targets did not significantly increase over time when high-confidence activity data were analyzed, despite the rapid growth in assay and activity data during recent years. For example, between 2004 and 2014, when most significant data growth occurred, detectable compound promiscuity remained essentially constant, with on average 1.5 targets per bioactive compound [15]. When promiscuity of drugs was followed over time, moderate increases in the degree of promiscuity were detected, albeit larger than for bioactive compounds, with the average degree increasing from 1.5 in 2000 to 3.2 in 2014 [16]. It was also observed that average degrees of promiscuity of drugs were frequently influenced by small numbers of highly promiscuous drug molecules [13]. Taken together, these studies have indicated that drugs are on average much more promiscuous than bioactive compounds, which are overall characterized by relatively low degrees of detectable promiscuity [13,15,16], especially on the basis of high-confidence activity data.

Considering the very large amounts of compound activity data that are already available [17,18], data mining should be expected to yield statistically meaningful promiscuity estimates [18]. On the other hand, there is the frequently discussed issue of data incompleteness [19], referring to the fact that not all available compounds have been tested against all targets. The generation of a complete compound-target activity matrix has been put forward as the ultimate goal of chemogenomics [20], which will most likely remain elusive. Regardless, due to data sparseness, the detectable degree of compound promiscuity might often be lower than true promiscuity, although it is unclear how large discrepancies might be.

In this context, it must also be taken into consideration that major compound repositories such as ChEMBL and DrugBank, upon which promiscuity estimates are based, collect activity annotations of compounds reported in the literature, but do not contain assay frequency or inactivity information, which is typically not reported. No major public compound database contains information of how many times a compound might have been tested so far against how many targets. Therefore, it is not possible, for example, to relate promiscuity degrees to assay frequency across different targets.

One possibility to extend promiscuity analysis through inclusion of assay frequency information is provided by screening data available in the public domain, with PubChem being the major repository. While it is not possible to directly access assay frequency information on a per compound basis, the data are available and it can be determined how many times a compound was tested in different screening assays and how often -and against which targets- it was found to be active. Recently, a web-based search tool has been introduced to retrieve such information from PubChem for individual query compounds [21]. However, for global and large-scale promiscuity analysis, assay and activity profiles must be determined systematically for all source compounds and analyzed in context.

In light of the above, we have reasoned that computational compound promiscuity analysis might be brought up to the next level by examining activity profiles of compounds that have been extensively assayed, thus addressing data sparseness issues in a previously unconsidered manner. To these ends, we have undertaken a large-magnitude analysis on the basis of currently available PubChem assay data. In a first data curation step, it was determined for each screening compound how often it was assayed and found to be active in primary screens as well as confirmatory assays. In the second step, promiscuity analysis was carried out for a large number of extensively tested compounds. In the following, our analysis and the results are presented in detail.

Material and Methods

Assay Categories

Assay data were taken from the PubChem BioAssay collection (accessed on 7th September 2015) [12], which contains different categories of assays including primary and confirmatory assays. Primary assays represent original screening data in which the activity assessment is based on percentage inhibition from a single dose. In this case, a compound is classified as active if it reduces target activity below an assay-specific threshold of residual activity. The threshold is often determined on the basis of the activity value distributions resulting from the screen. Accordingly, primary screens produce activity annotations of test compounds (i.e., active vs. inactive) but often not activity values. By contrast, confirmatory assays monitor activity measurements at varying compound concentrations and typically yield IC_{50} values derived from titration curves. In biological screening, it is common practice to re-evaluate initial screening hits in confirmatory assays. However, not all primary assays in PubChem have confirmatory counterparts and vice versa, for at least two reasons. First, primary or confirmatory assays are often independently deposited; second, increasing numbers of initial screens also use varying concentrations of test compounds for activity measurements and are thus confirmatory in nature. In general, activity annotations from primary screens have lower confidence than activity values from confirmatory assays, suggesting to best analyze them separately.

Data Collection

Primary and confirmatory assays were selected, as described below. From all available primary assays, only RNA interference (RNAi) screens were removed. Accordingly, all chemical screens were retained including primary cell-based assays for which no individual target was specified. For confirmatory assays, a series of selection criteria was applied using the PubChem BioAssay search interface [22]. First, “*On Hold BioAssays*” was set to “no hold”. Second, the type of bioassays was specified by setting “*Substance type*” to “chemical”; “*Screening stage*” to “confirmatory, dose-response”; and “*Target*” to “single”. Third, the “*Target type*” was set to “protein target”. Accordingly, all confirmatory assays in which chemical compounds were tested against single

target proteins with dose-response measurements were selected. Fourth, “Activity (IC_{50} , etc)” was set to “specified” and “Activity outcome” to “active”.

From each qualifying primary or confirmatory assay, only compounds classified as active or inactive were taken, whereas compounds with designations such as unspecified or inconclusive were discarded. For promiscuity analysis, compounds were prioritized that were tested in both primary and confirmatory assays, as rationalized below. For each compound, its identifier in PubChem (i.e., PubChem cid), the number of primary and confirmatory assays it was tested in, the number of primary and confirmatory assays in which it was active, and the number of unique targets from primary and confirmatory assays with activity were recorded.

The complete set of 437,257 compounds with assay and activity information has been made freely available as a ZENODO deposition [23].

Assay vs. Target Promiscuity

In our analysis, two types of promiscuity were distinguished. The degree of assay promiscuity was defined as the number of assays in which a compound was active. Assay promiscuity was determined by collecting all activity annotations from primary and confirmatory assays, respectively. Hence, different assays for the same target were counted individually. In addition, the degree of target promiscuity was defined as the number of unique targets a compound was active against across all assays. As a hypothetical example, a compound C was tested in assays 1–5 for a target T_1 and in assays 6–10 for another target T_2 and found to be active in assays 1, 2, 3, 8, and 10. Then, the corresponding assay and target promiscuity for C was five and two, respectively, indicating that the compound was active in a total of five assays against two targets. If another compound would be tested in 50 assays and found to be active in, for example, 14 against the same two targets, its assay promiscuity would be 14 and its target promiscuity would still be two. Hence, this would further differentiate between compounds having the same degree of target promiscuity. Therefore, these two promiscuity measures are complementary in nature. If no large and/or systematic discrepancies between assay and target promiscuity would be observed, there would be no indication of potential assay bias or false negatives that might affect target promiscuity analysis. Hence, considering assay and target promiscuity in context provides additional information. We also note that the degree of assay promiscuity of a compound may exceed its degree of target promiscuity, whereas target promiscuity cannot exceed assay promiscuity. Assay and target promiscuity were separately determined for compounds from primary and confirmatory PubChem assays.

Results

Assay and Compound Selection Strategy

A total of 1358 qualifying primary and 1823 confirmatory assays were obtained. Primary assays included 297 cell-based assays from which only assay promiscuity but not target promiscuity was determined. From primary and confirmatory assays, 836,585 and 457,842 unique compounds were selected, respectively, as reported in Table 1. These assays were directed against 476 (primary assays) and 632 (confirmatory) targets. Taken together, these assays covered a total of 824 unique targets. Furthermore, from all assays, a total of 146,270,306 and 37,808,671 assay-compound records were assembled, each of which reported the activity or inactivity of a given compound in an individual assay (Table 1).

From the PubChem BioAssay collection, the number of qualifying primary and confirmatory assays and corresponding targets is reported. In addition, the number of unique compounds tested in these assays is given. Furthermore, the total number of assay-compound records including active and inactive compounds is provided.

Table 1. Assay, target, and compound statistics.

Number of		Primary	Confirmatory
Assays		1358	1823
Targets		476	632
Compounds		836,585	457,842
	All	146,270,306	37,808,671
Assay-compound records	Activity	1,313,226	611,968
	Inactivity	144,957,080	37,196,703

doi:10.1371/journal.pone.0153873.t001

Next, the two large sets of compounds from primary or confirmatory assays were further compared. A subset of 437,257 compounds was tested in both primary and confirmatory assays. The remaining 399,328 and 20,585 compounds were evaluated only in primary or confirmatory assays, respectively. Of nearly 400,000 compounds tested exclusively in primary assays, ~73% were only evaluated in one to 10 primary assays. By contrast, only 1.5% of these compounds were tested in more than 50 assays. Furthermore, nearly 91% of these compounds were found to be consistently inactive in all primary assays they were tested in. These findings indicated that compounds tested exclusively in primary assays had low assay frequency and were predominantly inactive and thus not suitable for our promiscuity analysis. Similarly, ~75% of the 20,585 compounds exclusively tested in confirmatory assays were only evaluated in one to 10 and only ~4% of these compounds were tested in more than 50 assays. Hence, these infrequently assayed compounds were also not considered suitable for promiscuity analysis.

By contrast, the 437,257 compounds that were tested in both primary and confirmatory assays exhibited distinctly different assay frequencies. In this case, ~95% of the compounds were tested in more than 50 primary and/or confirmatory assays. Moreover, ~85% of these compounds were evaluated in a total of more than 100 assays. Hence, this subset of 437,257 compounds was extensively tested in both assay categories and strongly preferred for our analysis.

Assay Frequency Distribution

[Fig 1](#) reports assay frequencies for the 437,257 compounds in detail. In [Fig 1A and 1B](#), the distribution of compounds over primary and confirmatory assays is shown, respectively. The majority of these compounds were tested in hundreds of primary assays, with a mean of 325 assays per compound and a median of 347 assays. In addition, many compounds were also evaluated in more than 100 confirmatory assays (with a mean of 86 and median of 93 assays per compound). [Fig 1C](#) shows the distribution for combined primary and confirmatory assays, which confirms that most compounds were extensively evaluated, with a mean of 411 assays per compound and a median of 437 assays. More than 287,000 compounds were tested in a total of 400–848 assays. Hence, the selected compounds provided an unprecedented source for promiscuity analysis.

Consistently Inactive Compounds

Although the compounds were tested in hundreds of assays against hundreds of targets, large numbers of consistently inactive compounds were detected, as reported in [Fig 2](#). In primary ([Fig 2A](#)) and confirmatory assays ([Fig 2B](#)), a total of 169,839 and 240,650 compounds were consistently inactive, respectively. Furthermore, 119,256 compounds were found to be

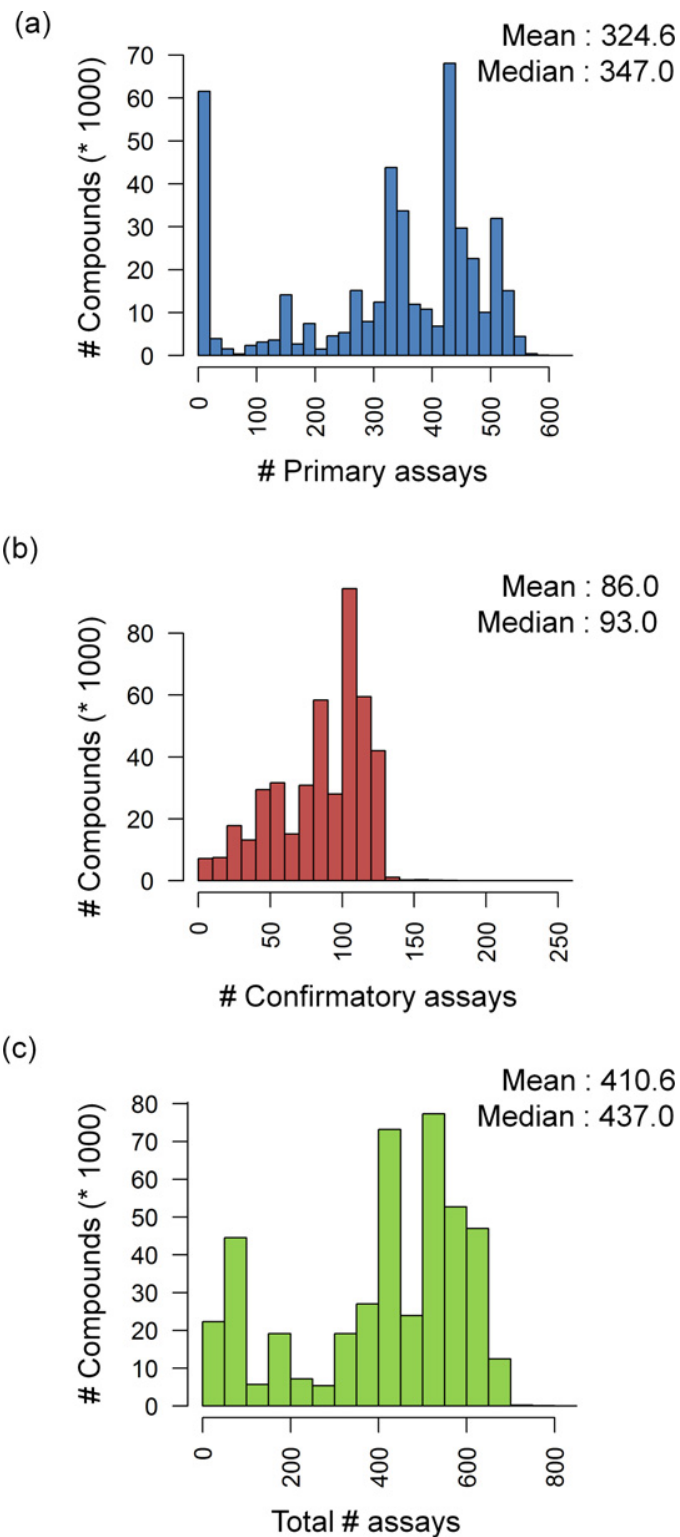


Fig 1. Assay frequency. Reported is the distribution of compounds tested in increasing numbers of (a) primary and (b) confirmatory assays. In (c), both assay categories are combined. In each case, the mean and median number of assays in which a compound was tested is provided.

doi:10.1371/journal.pone.0153873.g001

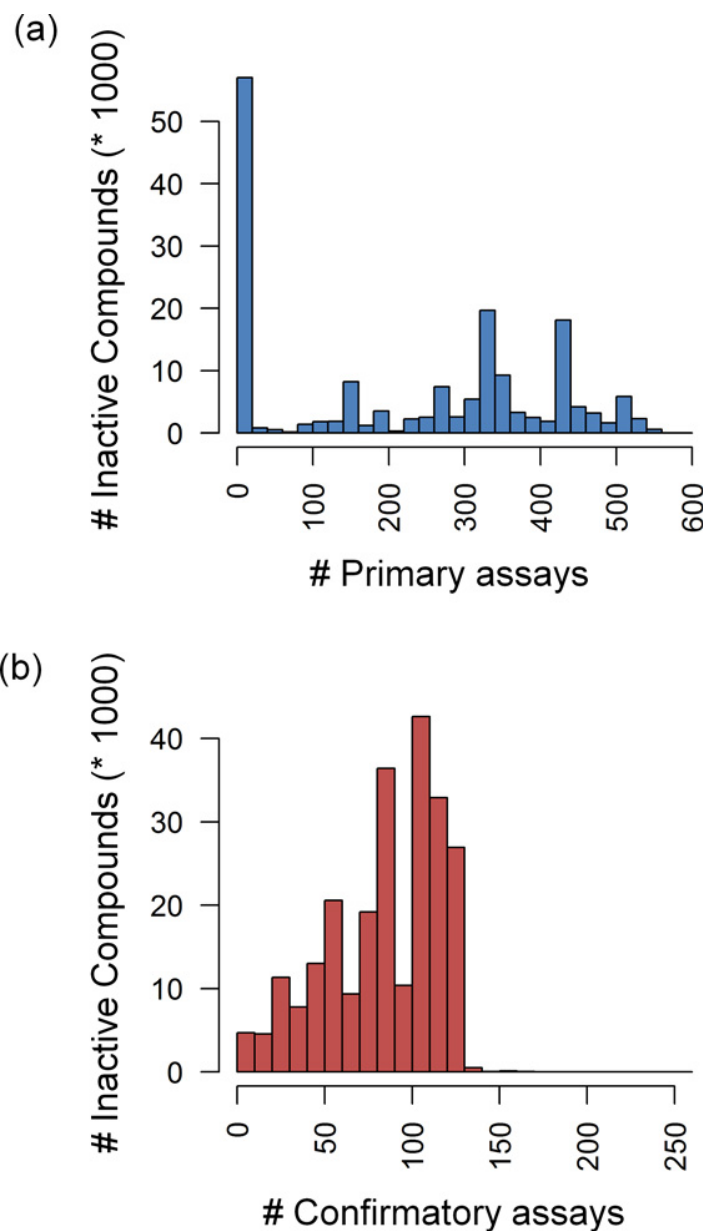


Fig 2. Inactive compounds. Reported is the distribution of compounds that were consistently inactive in increasing numbers of (a) primary and (b) confirmatory assays.

doi:10.1371/journal.pone.0153873.g002

consistently inactive in both primary and confirmatory assays. Fig 3 shows examples of structurally diverse compounds that were extensively tested, often in nearly or more than 700 assays, yet consistently inactive. The observation that 27.3% of the subset of extensively tested compounds was not active in any assay also indicated that there was no general tendency to produce false-positive assay signals, despite very large number of assays that were considered. Furthermore, these findings might also be viewed in light of recently described “dark chemical matter”, i.e., compounds that have been identified as consistently inactive in high-throughput screening assays of drug discovery projects but that might nonetheless have interesting activities and functional effects in other assay formats [24].

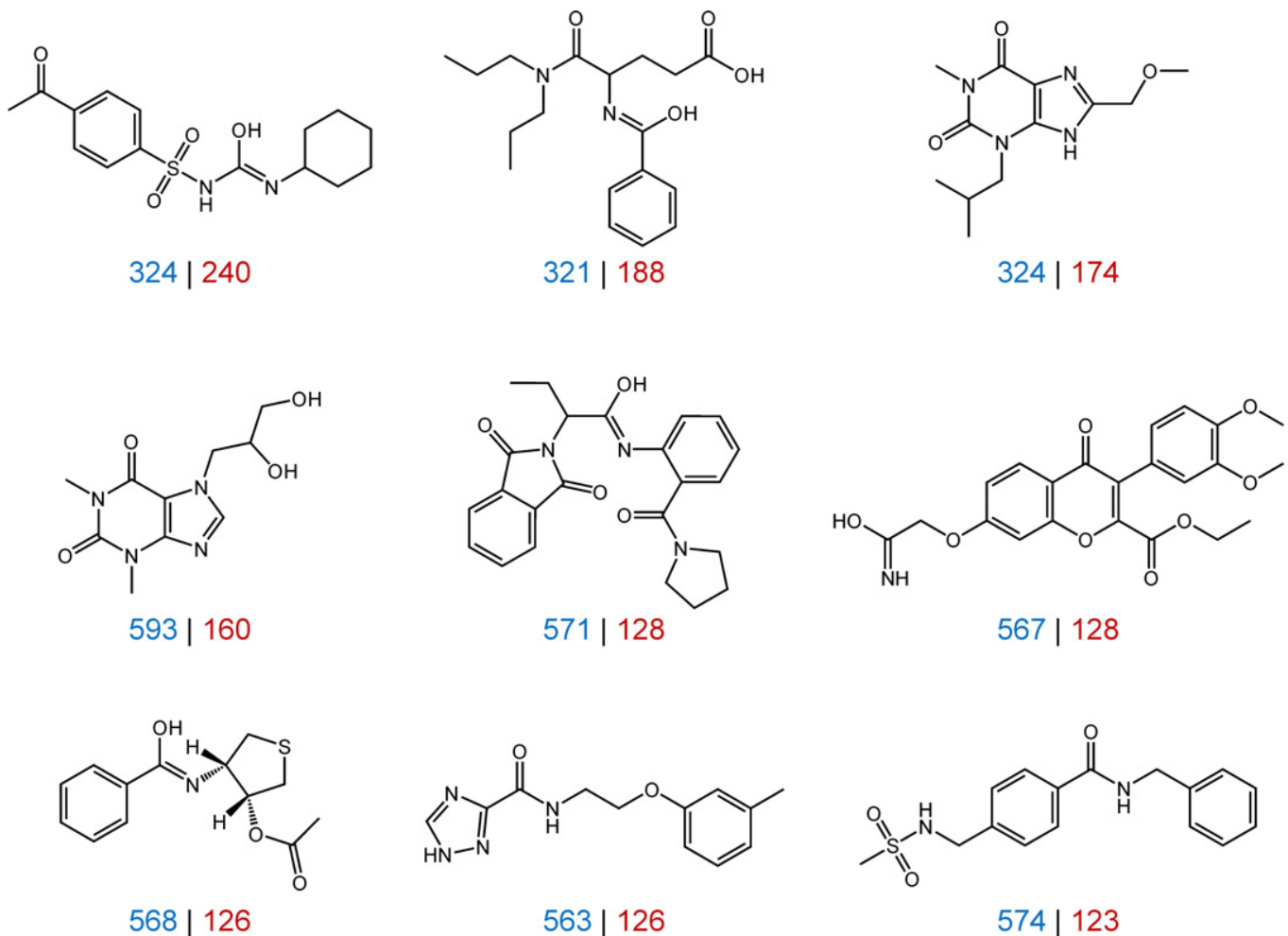


Fig 3. Exemplary inactive compounds. Shown are nine compounds that were consistently inactive in all assays. For each compound, the number of primary and confirmatory assays it was tested in is reported in blue and red, respectively.

doi:10.1371/journal.pone.0153873.g003

Compound Promiscuity

As the primary focal point of our analysis, we then systematically determined assay and target promiscuity for all active test compounds including 267,418 and 196,607 compounds from primary and confirmatory assays, respectively. Fig 4 shows the distribution of compounds that were active in increasing numbers of primary or confirmatory assays. In Fig 4A, assay promiscuity is monitored. On average, a compound was active in 4.7 primary and 3.0 confirmatory assays, with median values of 3.0 and 2.0, respectively. These values were lower than we anticipated. As shown in Fig 4B, and as expected, target promiscuity was lower than assay promiscuity. The average degree of target promiscuity in primary and confirmatory assays was 3.4 and 2.6, respectively, with a median degree of 2.0 in both cases. The observation that mean values were generally slightly or moderately higher than medians was attributed to the presence of a small proportion of highly promiscuous compounds, as further discussed below. Fig 5 reports changes in the degree of assay promiscuity for compounds tested in increasing numbers of primary (Fig 5A) and confirmatory assays (Fig 5B). In primary assays, median assay promiscuity essentially remained constant over increasing numbers of assays, except for a statistically small

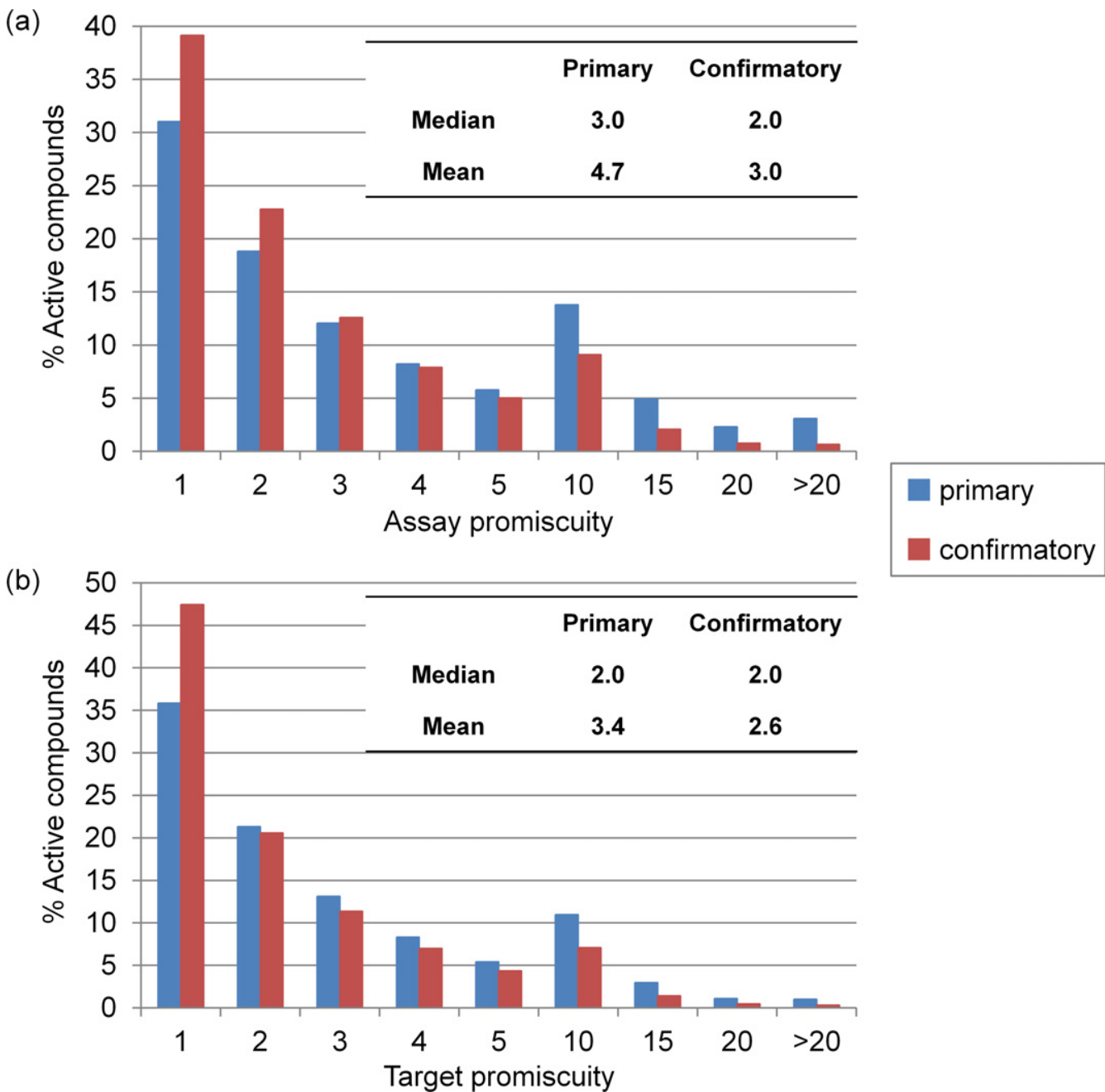


Fig 4. Assay and target promiscuity. Reported are the percentages of compounds with increasing degrees of (a) assay and (b) target promiscuity. In addition, average and median degrees of assay and target promiscuity are reported.

doi:10.1371/journal.pone.0153873.g004

sample of compounds tested in 600 to 700 assays where an increase was noted. Similar observations were made for confirmatory assays, with the exception of a moderate increase in the spread of promiscuity degrees for compounds tested in 150–250 assays. Fig 6 monitors changes in the degree of target promiscuity for compounds evaluated in increasing numbers of primary (Fig 6A) and confirmatory assays (Fig 6B). The distributions and median degrees of target promiscuity closely corresponded to those of assay promiscuity.

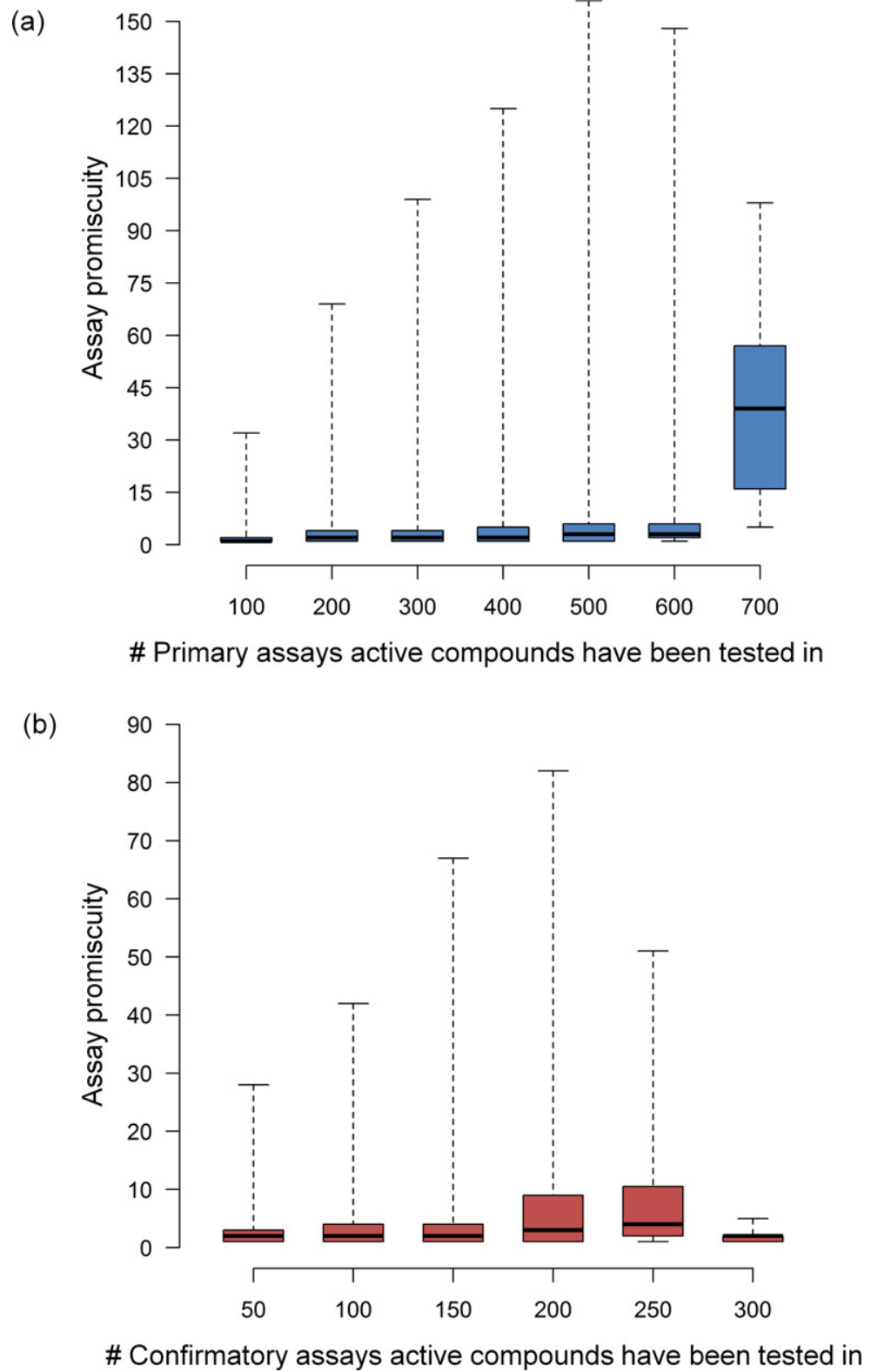


Fig 5. Assay frequency vs. assay promiscuity. For increasing numbers of (a) primary and (b) confirmatory assays, the distribution of assay promiscuity is reported in a box plot format. The plot gives the smallest

degree of assay promiscuity (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), and largest degree of assay promiscuity (top line).

doi:10.1371/journal.pone.0153873.g005

Fig 7 shows examples of highly promiscuous compounds that were active in more than 100 or 200 assays and largely responsible for increases in the average over median degree of promiscuity. Most of these compounds contained PAINS substructures [7,8] and were thus prone to assay artifacts. The filter for PAINS substructures in compounds was implemented using pattern checker [25] available in ZINC 15 in which a list of 480 SMARTS patterns was provided [26]. It should be noted that different implementations of PAINS might result in different mappings due to the conversion of original structural representations into SMARTS or the generation of different SMARTS variants [27]. In addition, different sets of fragments might be used or substructure search routines.

Taken together, the results revealed that assay promiscuity was higher than target promiscuity, as we would anticipate. However, the differences were small, as the average degree of assay promiscuity only increased by 1.3 and 0.4 in primary and confirmatory assays, respectively. The differences were even smaller for median promiscuity degrees. In addition, the mean and median degrees of assay or target promiscuity also only differed by less than 1 or 2.

Discussion

Target promiscuity of drugs and other bioactive compounds has thus far been studied on the basis of available activity annotations. Most recent surveys exclusively considering high-confidence activity data have resulted in average degrees of target promiscuity of 5.9 for approved drugs and 1.5 for bioactive compounds from medicinal chemistry sources [13]. Furthermore, the average degree of target promiscuity of compounds taken from confirmatory bioassays was 2.5 and thus also small [14]. Promiscuity estimates were generally higher for drugs than bioactive compounds. The higher degree of promiscuity among drugs might result from more extensive testing, but this remains uncertain. It is also possible that drug candidates that are successful in clinical trials might be more promiscuous than others.

Promiscuity analyses reported so far were based on known activity annotations, without taking assay frequencies or inactivity records into account, which are not available in major compound databases. This has generally been a point of concern, although very large volumes of activity data are already accessible, from which statistically meaningful trends can likely be derived. In light of data incompleteness or sparseness, it is frequently assumed that mining of compound activity annotations inevitably underestimates true compound promiscuity. This is likely the case although it remains unclear how large deviations from current promiscuity estimates might be.

We have set out to address these issues and further refine promiscuity analysis. Since it will hardly be possible to obtain a complete, or nearly complete, compound-target activity matrix any time soon, if at all, promiscuity analysis can at present only be further extended through incorporation of screening data. In addition, to address data sparseness concerns, compounds must be identified that have been extensively tested against many different targets.

Therefore, we have carried out a large-scale promiscuity analysis focusing on extensively assayed compounds. To our knowledge, this type of analysis is unprecedented. As a basis of our study, assay data were taken from PubChem and assay frequencies determined for all available compounds, which required substantial data curation efforts. For the first time, we also used primary screening data in promiscuity analysis to identify most extensively tested compounds. Because activity annotations from primary screening assays were only approximate in nature, multiple assays were frequently available for the same target, and a limited amount of

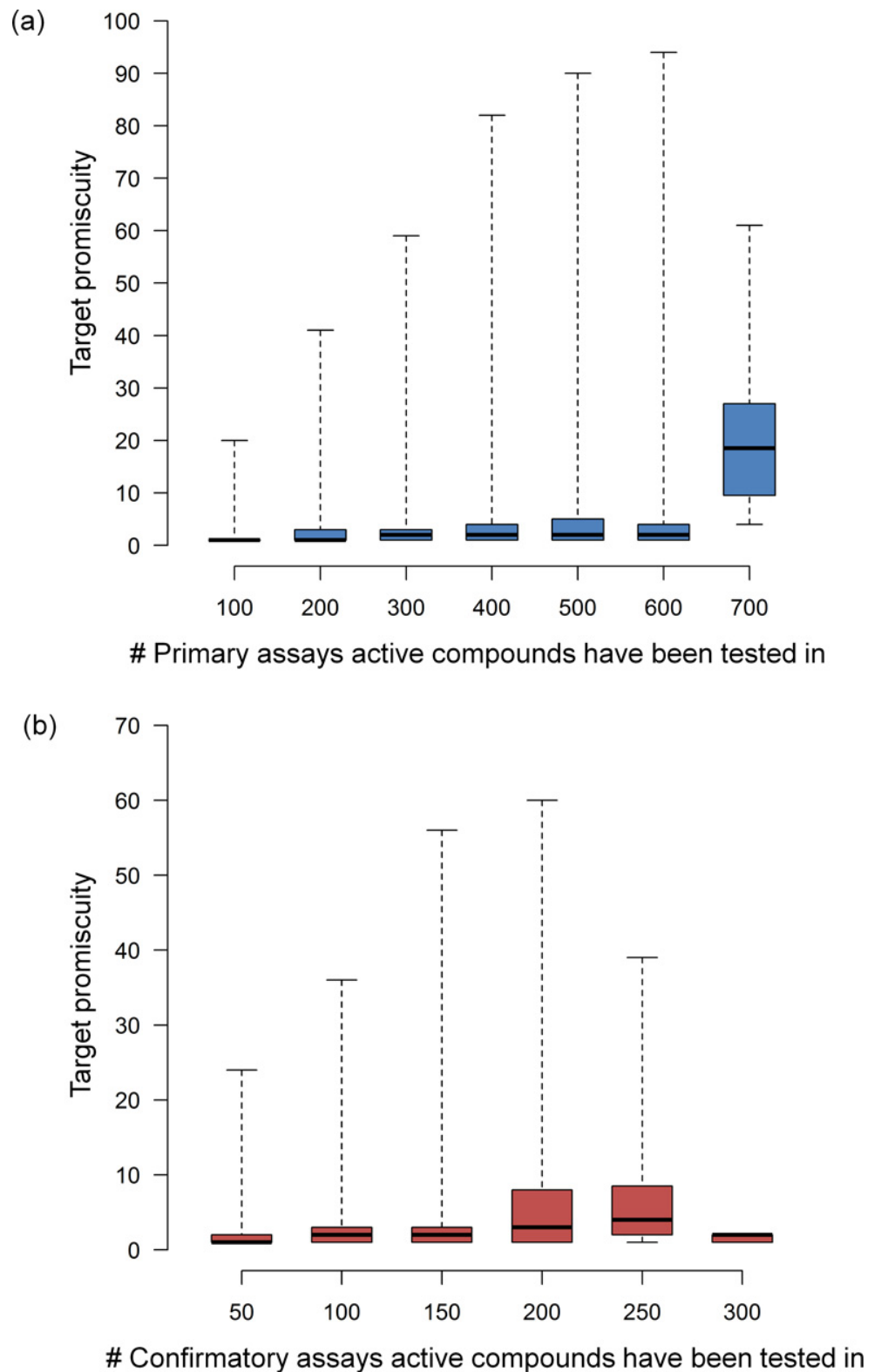


Fig 6. Assay frequency vs. target promiscuity. For increasing numbers of (a) primary and (b) confirmatory assays, the distribution of target promiscuity is reported in box plots according to Fig 5.

doi:10.1371/journal.pone.0153873.g006

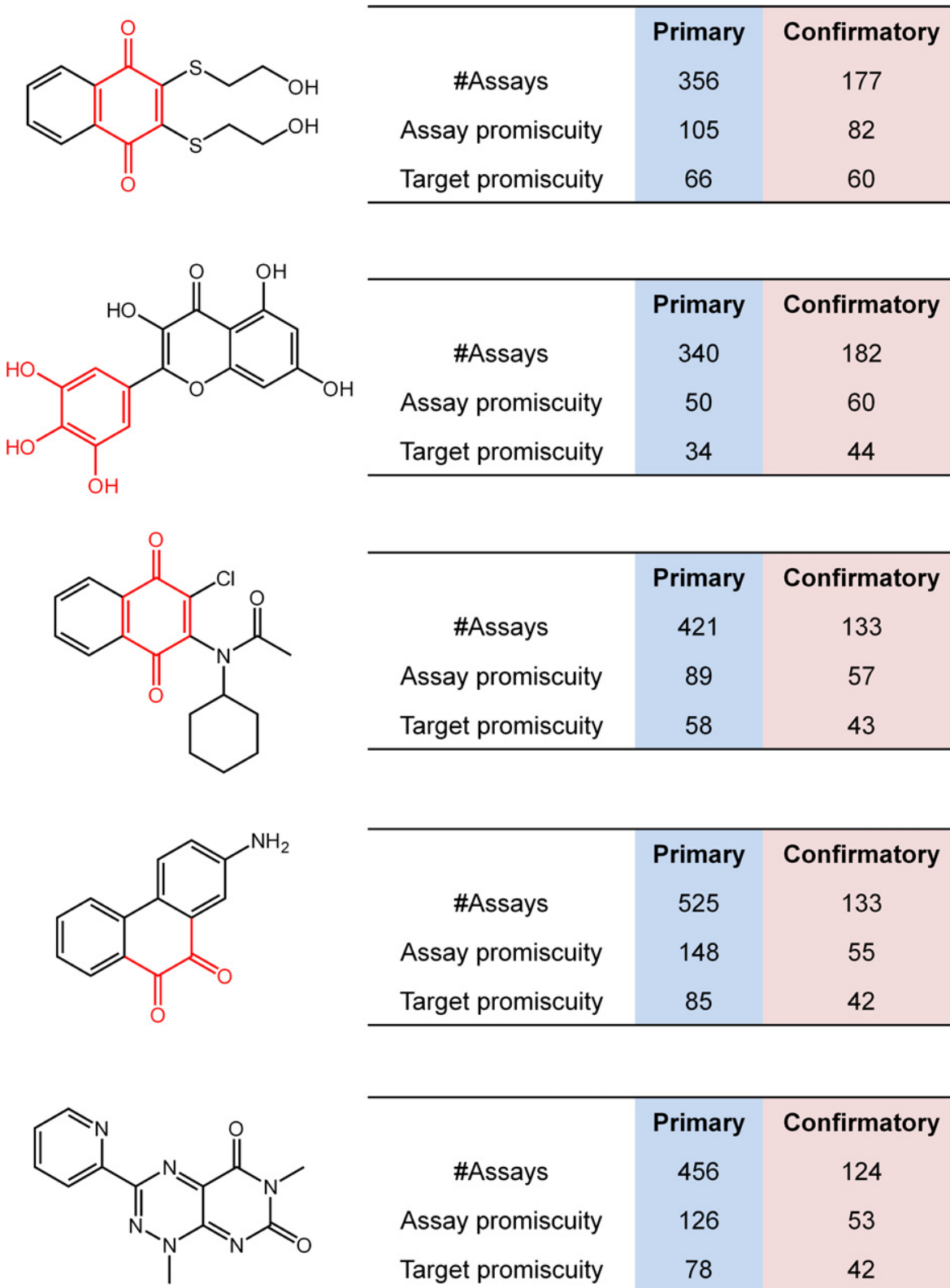


Fig 7. Highly promiscuous compounds. Shown are five exemplary highly promiscuous compounds. For each compound, the number of assays it was tested in and its assay and target promiscuity are reported. Four of these five compounds contain PAINS substructures (red).

doi:10.1371/journal.pone.0153873.g007

cell-based assays was also considered, assay promiscuity was distinguished from target promiscuity and separately analyzed.

A subset of ~437,000 compounds was identified that were extensively tested in hundreds of assays against hundreds of targets. These compounds were subjected to promiscuity analysis in which primary and confirmatory assay data were separately considered. As expected, we found that assay promiscuity was generally higher than target promiscuity. However, the differences were surprisingly small, only on the order of 1, as reported above.

Given that primary screening data and extensively assayed compounds were used in our analysis, it was anticipated to observe higher degrees of target promiscuity for active compounds than previously reported. Average degrees of target promiscuity of 3.4 and 2.6 were determined for primary and confirmatory assays, respectively. These promiscuity degrees were only moderately higher, even for primary screening assays, than previously determined for ChEMBL compounds with available high-confidence activity data. We also detected small subsets of highly promiscuous screening hits, which led to an increase in average target promiscuity over median promiscuity. Highly promiscuous compounds often contained PAINS substructures and were thus likely to cause assay artifacts. Accordingly, median values might better estimate promiscuity degrees, at least for compounds from screening sources. The median degree of target promiscuity was 2.0 for both primary and confirmatory assays and thus only slightly higher than the corresponding value of 1.5 for ChEMBL compounds.

In conclusion, as revealed by our analysis, target promiscuity remained at a low level for bioactive compounds, even when studying the most extensively assayed compounds that are currently available. These findings lend further support to previously drawn conclusions that bioactive compounds are in general only moderately promiscuous and less promiscuous than drugs. One possible explanation would be that drugs are much more intensively investigated and tested for additional targets than bioactive compounds, for example, in many drug repurposing projects. Alternatively, given that drugs originate from the pool of bioactive compounds, these results also support the idea of a “promiscuity enrichment model”. The underlying hypothesis is that promiscuous compounds are preferentially selected for therapeutic efficacy during clinical evaluation and ultimately become drugs. This requires, however, that desired therapeutic effects due to substantial promiscuity outweigh unwanted side effects that are also possible.

Author Contributions

Conceived and designed the experiments: JB YH. Performed the experiments: SJ YH. Analyzed the data: SJ YH JB. Contributed reagents/materials/analysis tools: YH. Wrote the paper: JB YH SJ.

References

1. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat. Biotechnol.* 2006; 24: 805–815. PMID: [16841068](#)
2. Boran AD, Iyengar R. Systems approaches to polypharmacology and drug discovery. *Curr. Opin. Drug Discov. Devel.* 2010; 13: 297–309. PMID: [20443163](#)
3. Jalencas X, Mestres J. On the origins of drug polypharmacology. *Med. Chem. Comm.* 2013; 4: 80–87.
4. Knight ZA, Lin H, Shokat KM. Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* 2010; 10: 130–137. doi: [10.1038/nrc2787](#) PMID: [20094047](#)
5. Hu Y, Bajorath J. Compound promiscuity—what can we learn from current data. *Drug Discov. Today* 2013; 18: 644–650.
6. Lu JJ, Pan W, Hu YJ, Wang YT. Multi-target drugs: the trend of drug research and development. *PLoS ONE* 2012; 7: e40262. doi: [10.1371/journal.pone.0040262](#) PMID: [22768266](#)

7. 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: 2719–2740. doi: [10.1021/jm901137j](https://doi.org/10.1021/jm901137j) PMID: [20131845](https://pubmed.ncbi.nlm.nih.gov/20131845/)
8. Baell JB, Walters MA. Chemical con artists foil drug discovery. *Nature* 2014; 513: 481–483. doi: [10.1038/513481a](https://doi.org/10.1038/513481a) PMID: [25254460](https://pubmed.ncbi.nlm.nih.gov/25254460/)
9. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 2014; 42: D1091–1097. doi: [10.1093/nar/gkt1068](https://doi.org/10.1093/nar/gkt1068) PMID: [24203711](https://pubmed.ncbi.nlm.nih.gov/24203711/)
10. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2011; 40: D1100–D1107. doi: [10.1093/nar/gkr777](https://doi.org/10.1093/nar/gkr777) PMID: [21948594](https://pubmed.ncbi.nlm.nih.gov/21948594/)
11. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 2014; 42: D1083–D1090. doi: [10.1093/nar/gkt1031](https://doi.org/10.1093/nar/gkt1031) PMID: [24214965](https://pubmed.ncbi.nlm.nih.gov/24214965/)
12. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, et al. PubChem's BioAssay database. *Nucleic Acids Res.* 2012; 40: D400–D412. doi: [10.1093/nar/gkr1132](https://doi.org/10.1093/nar/gkr1132) PMID: [22140110](https://pubmed.ncbi.nlm.nih.gov/22140110/)
13. Hu Y, Bajorath J. High-resolution view of compound promiscuity [v2; ref status: indexed, <http://f1000r.es/1ig>]. *F1000Res.* 2013; 2: 144.
14. 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. doi: [10.1208/s12248-013-9488-0](https://doi.org/10.1208/s12248-013-9488-0) PMID: [23605807](https://pubmed.ncbi.nlm.nih.gov/23605807/)
15. Hu Y, Jasial S, Bajorath J. Promiscuity progression of bioactive compounds over time. [v2; ref status: indexed, <http://f1000r.es/5h4>] *F1000Res.* 2015; 4: 118.
16. Hu Y, Bajorath J. Monitoring drug promiscuity over time [v2; ref status: indexed, <http://f1000r.es/4oa>] *F1000Res.* 2014; 3: 218.
17. Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, de Vlieg J. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today* 2014; 19: 859–868. doi: [10.1016/j.drudis.2013.12.004](https://doi.org/10.1016/j.drudis.2013.12.004) PMID: [24361338](https://pubmed.ncbi.nlm.nih.gov/24361338/)
18. Hu Y, Bajorath J. Learning from 'big data': compounds and targets. *Drug Discovery Today* 2014; 19: 357–360. doi: [10.1016/j.drudis.2014.02.004](https://doi.org/10.1016/j.drudis.2014.02.004) PMID: [24561327](https://pubmed.ncbi.nlm.nih.gov/24561327/)
19. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV. Data completeness—the achilles heel of drug-target networks. *Nat. Biotechnol.* 2008; 26: 983–984. doi: [10.1038/nbt0908-983](https://doi.org/10.1038/nbt0908-983) PMID: [18779805](https://pubmed.ncbi.nlm.nih.gov/18779805/)
20. Jacoby E. Chemogenomics: drug discovery's panacea? *Mol. BioSyst.* 2006; 2: 218–220. PMID: [16880939](https://pubmed.ncbi.nlm.nih.gov/16880939/)
21. Canny SA, Cruz Y, Southern MR, Griffin PR. PubChem promiscuity: a web resource for gathering compound promiscuity data from PubChem. *Bioinformatics* 2012; 28: 140–141. doi: [10.1093/bioinformatics/btr622](https://doi.org/10.1093/bioinformatics/btr622) PMID: [22084255](https://pubmed.ncbi.nlm.nih.gov/22084255/)
22. Available: <https://www.ncbi.nlm.nih.gov/pcassay/limits>.
23. Jasial S, Hu Y, Bajorath J. PubChem compounds tested in primary and confirmatory assays. *ZENODO* 2016; doi: [10.5281/zenodo.44593](https://doi.org/10.5281/zenodo.44593)
24. Wassermann AM, Lounkine E, Hoepfner D, Le Goff G, King FJ, Studer C, et al. Dark chemical matter as a promising starting point for drug lead discovery. *Nat. Chem. Biol.* 2015; 11: 958–966. doi: [10.1038/nchembio.1936](https://doi.org/10.1038/nchembio.1936) PMID: [26479441](https://pubmed.ncbi.nlm.nih.gov/26479441/)
25. Available: <http://zinc15.docking.org/patterns/home/>.
26. Sterling T, Irwin JJ. ZINC 15—ligand discovery for everyone. *J. Chem. Inf. Model.* 2015; 55: 2324–2337. doi: [10.1021/acs.jcim.5b00559](https://doi.org/10.1021/acs.jcim.5b00559) PMID: [26479676](https://pubmed.ncbi.nlm.nih.gov/26479676/)
27. Baell JB. Screening-based translation of public research encounters painful problems. *ACS Med. Chem. Lett.* 2015; 6: 229–234.

Summary

A large-scale promiscuity analysis was carried out on extensively assayed compounds extracted from PubChem BioAssay database. A subset of 437,257 compounds which were tested in both primary and confirmatory assays was used as the extensively assayed set. 95% of these compounds were tested in more than 50 assays. Assay promiscuity was calculated separately from target promiscuity for active compounds in primary and confirmatory assays to check for potential assay bias or false negatives. Assay promiscuity was found to be higher than the target promiscuity but the differences were surprisingly small. The median degree of target promiscuity was 2.0 for both primary and confirmatory assays, slightly higher than the corresponding value of 1.5 for ChEMBL compounds. Thus, target promiscuity remained at a low level even for extensively tested compounds which further gives evidence that bioactive compounds are less promiscuous than drugs. Small subsets of highly promiscuous compounds were also detected that were responsible for the increase in mean target promiscuity over median promiscuity. These subsets of compounds often contained PAINS substructures and therefore were prone to assay artifacts. While analyzing activity profiles of the extensively tested compound set, large numbers of consistently inactive compounds were detected. Some of these compounds, which were tested in hundreds of assays, qualify as DCM.

The analysis of DCM compounds extracted from screening data and derivation of their target hypotheses are presented in the next chapter.

Chapter 4

Dark Chemical Matter in Public Screening Assays and Derivation of Target Hypotheses

Introduction

A large number of compounds are tested against hundreds of targets during high-throughput screening campaigns in search of high-quality hits and new chemical entities. However, many screening compounds are generally found to be consistently inactive in all the assays they are tested in. These compounds are termed dark chemical matter.

It has been found that DCM compounds are not completely inert biologically. When tested in assays for novel targets, DCM compounds may yield selective hits. Thus, DCM can provide interesting starting points for finding lead candidates with high target selectivity.

In this work, DCM compounds were systematically extracted from extensively tested screening compounds and structural relationships between DCM and bioactive compounds were explored in order to derive target hypotheses for DCM. The key point was to check whether DCM compounds occur in analog

series with bioactive compounds that have target annotations in high-confidence activity data.

Reproduced from “Jasial, S.; Bajorath J. Dark Chemical Matter in Public Screening Assays and Derivation of Target Hypotheses. *Medicinal Chemistry Communications* **2017**, *8*, 2100-2104” with permission from The Royal Society of Chemistry.



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

Received 18th August 2017,
Accepted 20th October 2017

DOI: 10.1039/c7md00426e

rsc.li/medchemcomm

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

Swarit Jasial and Jürgen Bajorath *

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

Introduction

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

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

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

Methods and materials

Extensively assayed PubChem compounds

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

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

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

ChEMBL compounds with high-confidence activity data

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

Identification of analog series

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

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

Results and discussion

Dark chemical matter

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

Assay frequency

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

Overlap between PubChem and ChEMBL

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

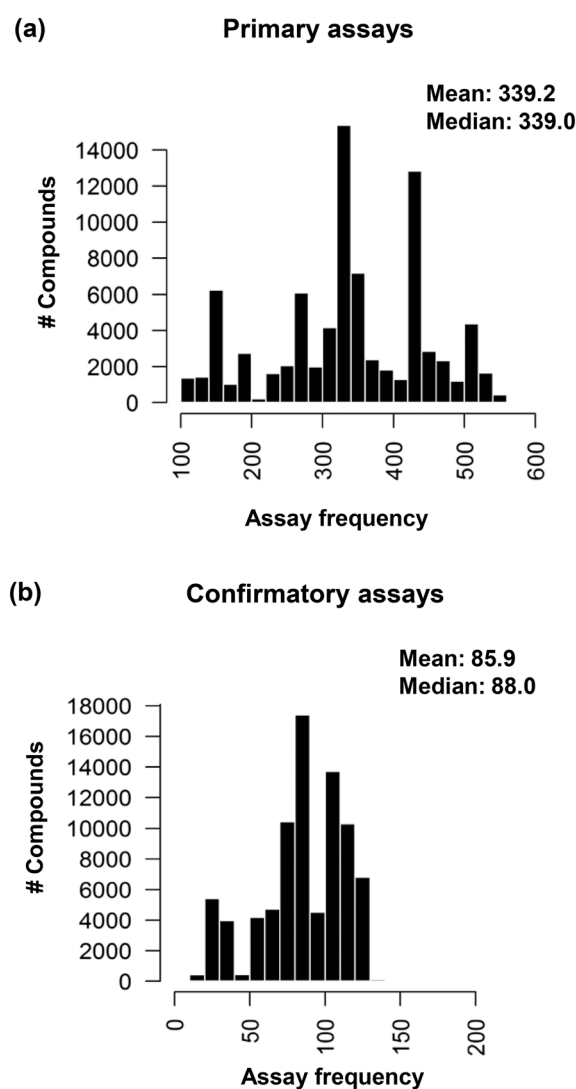


Fig. 1 Assay frequency distribution for DCM. Histograms show the distribution of (a) primary and (b) confirmatory assays in which DCM compounds from PubChem were tested.

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

Searching for analog series

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

Exemplary series

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

Table 1 Analog series containing DCM and ChEMBL compounds

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

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

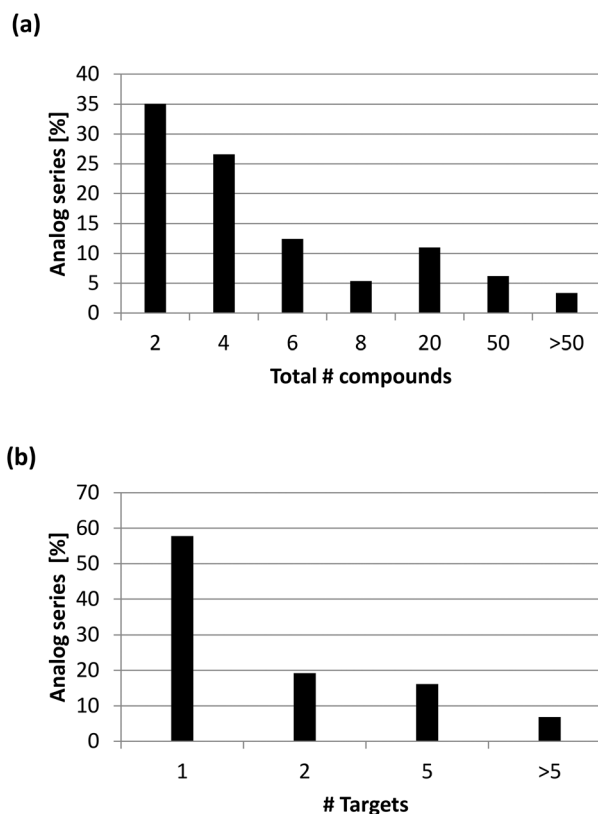


Fig. 2 Size and target distribution of analog series. For analog series including DCM and ChEMBL compounds, the (a) size and (b) target distribution is reported. For each series, the total number of unique targets of ChEMBL analogs was determined.

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

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

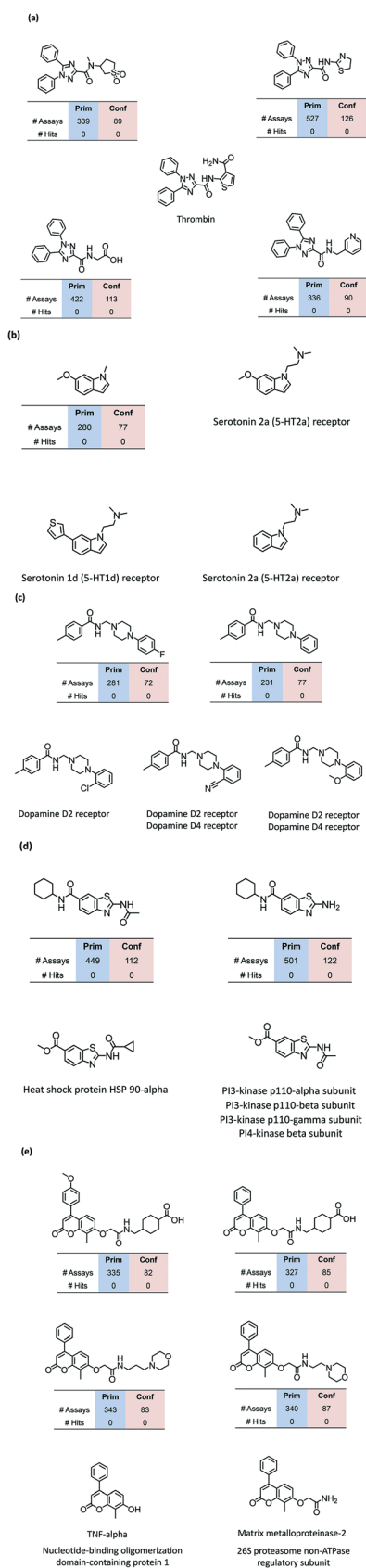


Fig. 3 Exemplary analog series. In (a)–(e), different examples of series containing DCM and ChEMBL analogs are presented. For DCM and ChEMBL compounds, assay statistics and target annotations are provided, respectively.

Conclusions

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

Conflicts of interest

The authors declare no competing interest.

Acknowledgements

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

References

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

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

Summary

A total of 81,597 unique compounds extracted from extensively assayed compound set were found to be consistently inactive in all primary and confirmatory assays they were tested in. A systematic search for bioactive analogs using high-confidence data from ChEMBL was carried out for these DCM compounds. For 10.5% of DCM (~ 8500 compounds), varying number of ChEMBL analogs were identified. Furthermore, 1400 analog series with different compositions of DCM and ChEMBL compounds was obtained. The ChEMBL compounds present in the DCM/ChEMBL analog series had activity annotations against well-known pharmaceutical targets which can provide target hypotheses for their DCM analogs. Thus, the analog series containing DCM and ChEMBL compounds can be a good starting point for further exploring activities of DCM compounds against targets of their bioactive analogs.

During the analysis of activity profiles of extensively assayed compounds from PubChem (*in Chapter 3*), we also found counterparts of DCM compounds that were highly promiscuous. These compounds often contained PAINS substructures and were likely to cause assay interference. However, activity profiles of PAINS have not been explored on a large scale.

In the next chapter, we present a systematic analysis of PAINS extracted from screening data.

Chapter 5

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

Introduction

PAINS are small molecules that might be reactive under assay conditions and might produce false-positive assay signals, which cause substantial problems for biological screening and medicinal chemistry. So far, 480 compound classes have been designated as PAINS, which are typically contained as substructures in larger compounds. Computational filters encoding PAINS can be used to detect compounds with potential chemical liabilities that require follow-up analysis. Such filters are controversially viewed in the field but provide first-path alerts for potential liabilities.

Herein, interference characteristics of PAINS have been computationally investigated by systematically analyzing publicly available screening data and determining activity profiles of screening compounds with PAINS substructures. The major goal of the analysis was to examine whether assumed interference characteristics of PAINS were supported by activities in biological screening assays.

Reprinted with permission from “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. *Journal of Medicinal Chemistry* **2017**, *60*, 3879-3886”. Copyright 2017 American Chemical Society

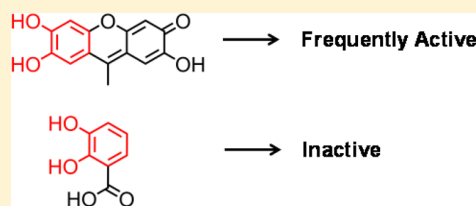


How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds

Swarit Jasial, Ye Hu, and Jürgen Bajorath*^{1b}

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

ABSTRACT: Undetected pan-assay interference compounds (PAINS) with false-positive activities in assays often propagate through medicinal chemistry programs and compromise their outcomes. Although a large number of PAINS have been classified, often on the basis of individual studies or chemical experience, little has been done so far to systematically assess their activity profiles. Herein we report a large-scale analysis of the behavior of PAINS in biological screening assays. More than 23 000 extensively tested compounds containing PAINS substructures were detected, and their hit rates were determined. Many consistently inactive compounds were identified. The hit frequency was low overall, with median values of two to five hits for PAINS tested in hundreds of assays. Only confined subsets of PAINS produced abundant hits. The same PAINS substructure was often found in consistently inactive and frequently active compounds, indicating that the structural context in which PAINS occur modulates their effects.



INTRODUCTION

Pan-assay interference compounds (PAINS) cause false-positive assay signals due to reactivity under assay conditions, including covalent modifications or redox effects, chelation, autofluorescence, or degradation.^{1–3} More than 450 compound classes have been designated as PAINS to date, including, for example, rhodanines, isothiazolones, enones, and quinones^{1–3} as well as pharmaceutically intensely explored compounds such as curcuminoids.^{3,4} Classified PAINS are typically small reactive or otherwise liable molecules that are contained as substructures in larger compounds. In addition to PAINS, other compounds might also be reactive under assay conditions and elicit artifacts,⁵ thus widening the spectrum of possible interference compounds. The most promiscuous compounds identified across PubChem assays included a variety of putative interference compounds not classified as PAINS. Although PAINS and other interference molecules were prevalent among highly promiscuous compounds, they also contained chemical entities with no apparent liabilities.⁵ Molecules having a tendency to cause assay artifacts are not limited to synthetic compounds but are also found among natural products.^{6,7} A small set of natural products exhibiting a plethora of artificial biological activities has recently been termed invalid metabolic panaceas (IMPs).⁷

Efforts to systematically identify and classify PAINS^{1–3} are complemented by case studies that carefully evaluate unrecognized interference compounds and uncover chemical liabilities.^{8,9} Occasionally, other investigations are reported that carefully analyze an observed activity associated with PAINS and ultimately confirm this activity using orthogonal assay

systems or X-ray crystallography.¹⁰ However, such investigations are exceptions. Publications reporting PAINS activities including unrecognized artifacts in good faith frequently appear, and individual examples propagate through the literature.⁴

Not only are potential PAINS widely distributed, but false-positive assay readouts caused by PAINS are often difficult to identify.^{2,3} Consequently, undetected PAINS may proceed far during doomed optimization efforts until a roadblock is hit and their artificial nature becomes evident. This presents a substantial problem for medicinal chemistry.

Although there are controversial views about interference compounds, increasing awareness of PAINS liabilities is evident in the field. In fact, an emerging trend can be observed to exclude a priori any potentially reactive compound from further consideration, akin to "PAINS paranoia". This is another point of concern. Disregarding all potentially reactive or problematic compounds would be detrimental for medicinal chemistry, just as much as ignoring PAINS would be. Ultimately, orthogonal assays and mechanistic studies are required to firmly establish the validity of active compounds, irrespective of whether they are designated as PAINS or not, as also emphasized in a recent editorial by editors of several journals of the American Chemical Society.¹¹

Many PAINS have been identified in individual medicinal chemistry and assay campaigns¹ from which general assay interference was extrapolated and supported by literature data and/or experience values. A key question is whether

Received: January 30, 2017

Published: April 19, 2017

interference characteristics are also supported by considering large numbers of different assays.

To arrive at a balanced and data-centric assessment of PAINS and the magnitude of their effects, we have set out to systematically explore the activity profiles of PAINS in biological screening assays. Only little prior knowledge exists. For example, an earlier study reported the results of a dose–response screen of nearly 200 000 compounds against the cysteine protease cruzain using a fluorescence assay system.¹² Of the initially detected active compounds, 23.7% were detergent-sensitive indicating aggregation, 3.1% were auto-fluorescent, and 2.6% contained reactive or undesirable groups. Hence, there was a substantial fraction of interference compounds among cruzain screening hits. A similar study reported a screen on β -lactamase.¹³ In this case, 95% of the hits were detergent-sensitive. Follow-up investigations analyzed the mechanisms of artifactual inhibitors identified in these two screens.^{14,15} Taken together, these studies predominantly focused on colloidal aggregation.

Herein we report a systematic analysis of PAINS in publicly available screening compounds. Importantly, while our paper was under review, a closely related investigation was published.¹⁶ In their comprehensive study,¹⁶ Capuzzi et al. analyzed PAINS in AlphaScreen assays (the technology utilized in the original PAINS report¹) available in PubChem¹⁶ as well as other assay formats, PAINS in dark chemical matter and drugs, and random PAINS in PubChem.¹⁵ In the context of their analysis, Capuzzi et al. reevaluated the original investigation by Baell and Holloway¹ and concentrated on the question of whether the use of PAINS filters is scientifically justified. The study of Capuzzi et al. and our analysis share a global investigation of PubChem compounds. In both cases, similar and consistent results were obtained, yielding equivalent conclusions, as further discussed below. In the following, we present our analysis of PAINS in extensively assayed compounds originating from PubChem. Taken together, these at least in part unexpected results provide some fresh insights into the breadth of interference effects as observed by analyzing a large body of experimental data.

METHODS AND MATERIALS

Compound Data Collection. From the PubChem BioAssay database,¹⁷ compounds were selected that were tested in both primary assays (resulting in percentage of activity from a single dose) and confirmatory assays (dose–response assays yielding IC_{50} values). This selection criterion was applied to focus the analysis on extensively assayed compounds. From primary assays, RNA interference screens were excluded. Compounds designated as “active” or “inactive” were selected, whereas compounds with attributes such as “unspecified” or “inconclusive” were disregarded. Confirmatory assays were only considered if single protein targets were specified. No threshold was applied to IC_{50} values so that weak activities of PAINS were also taken into consideration (leading to an upper-limit assessment of assay signals from PAINS).

A total of 437 257 compounds were obtained that were active against 824 targets. More than 95% of these compounds were tested in more than 50 primary and/or confirmatory assays, with a mean and median of 411 and 437 assays per compound, respectively. This large set of extensively assayed compounds and their associated activity/inactivity records provided the basis for our analysis.

Identification of PAINS. The 437 257 compounds were screened in silico for PAINS extracted from three public filters, including RDKit,¹⁸ ZINC,¹⁹ and ChEMBL.²⁰ Three filters were used to account for possible implementation discrepancies. SMARTS strings of PAINS were exported from RDKit (480 strings), ZINC (480), and ChEMBL

(481) and used as substructure queries to search the collection of extensively assayed compounds for PAINS.²¹ PubChem compounds were represented as canonical SMILES generated from hydrogen-suppressed graphs.

RESULTS AND DISCUSSION

PAINS and Assay Frequency. The set of 437 257 compounds tested in both primary and confirmatory assays contained 27 520 compounds detected as PAINS by at least one of the filters. These screening compounds included a total of 270 PAINS as substructures. Figure 1 reports the assay

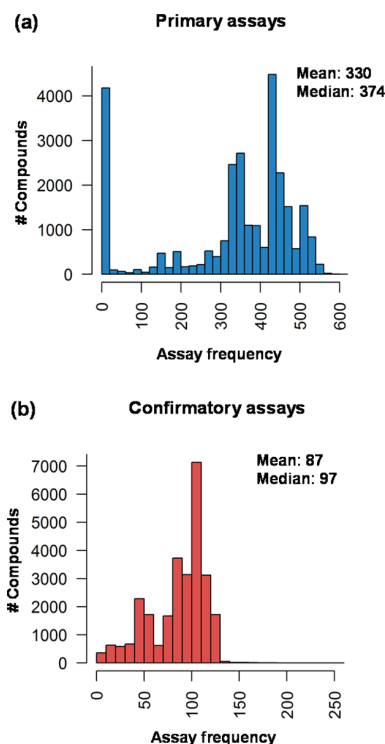


Figure 1. Assay frequency of PAINS. Histograms report the assay frequency of PAINS from (a) primary and (b) confirmatory assays.

frequency of the detected PAINS-containing compounds, confirming that many of these compounds were extensively tested. For PAINS from primary assays, the median frequency was 374 assays per compound, and for PAINS from confirmatory assays, the median was 97.

Active and Inactive PAINS. We determined that 23 036 PAINS-containing compounds (with 265 distinct PAINS substructures) were tested in at least 100 primary assays and 23 377 compounds (with 266 distinct PAINS substructures) in at least 50 confirmatory assays. The following analysis was based upon these most extensively tested compounds. Figure 2 reports their assay results. In primary assays, 18 248 compounds (with 262 PAINS substructures) were active at least once, whereas 4788 compounds (121 PAINS substructures) were consistently inactive in 100 to 594 assays. In confirmatory assays, 15 659 compounds (258 PAINS substructures) were active in at least one assay and 7718 (151 PAINS substructures) were consistently inactive. Primary and confirmatory assays shared 13 394 active and 2792 inactive compounds.

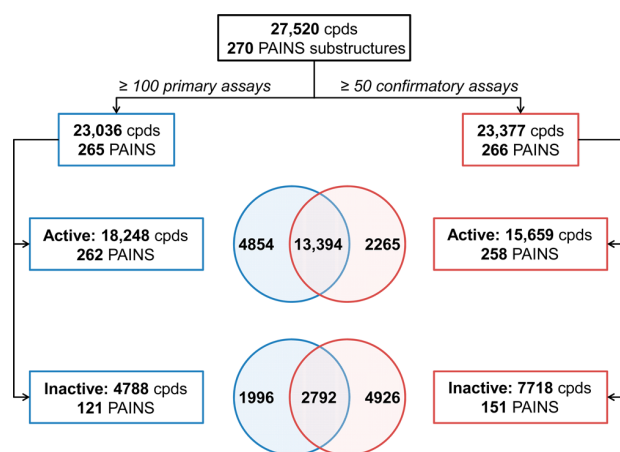


Figure 2. Active vs inactive compounds. The flowchart reports the numbers of active and consistently inactive PAINS-containing compounds from primary and confirmatory assays. The Venn diagrams in the center reveal the overlaps between active and inactive compounds from primary and confirmatory assays. The numbers of PAINS contained as substructures in active and inactive compounds are also given.

When primary and confirmatory assays were combined and PAINS substructures that were represented by at least five compounds were selected, 20 660 active and 3800 consistently inactive compounds were obtained that contained 176 and 98 unique PAINS substructures, respectively. All of these 98 PAINS substructures were found in both active and consistently inactive compounds.

Thus, the majority of PAINS-containing compounds was active in one or more assays. We note that compounds with activity in only a single assay cannot be regarded as PAINS, irrespective of whether they contain PAINS substructures.^{1,16} However, there also were large numbers of extensively tested compounds (i.e., ~21%, primary assays; ~33%, confirmatory) that were consistently inactive in all assays, which was an unexpected finding. Consistently inactive compounds contained many different PAINS. Similarly, the top 20% of compounds with the largest numbers of hits in primary or confirmatory assays were also widely distributed over different PAINS. In both primary and confirmatory assays, quinones were the class of PAINS that produced most hits, with 388 and 348 active compounds, respectively.

Hit Frequency. After identifying many consistently inactive PAINS-containing compounds, we determined how frequently active PAINS produced hits in different assays. Figure 3 reports the distribution of hits for PAINS and equally sized random samples of other screening compounds tested in increasing numbers of primary and confirmatory assays. For both PAINS and non-PAINS, the results were comparable for both assay categories. Overall, PAINS displayed slightly higher hit rates than non-PAINS. While there were small numbers of PAINS that produced large numbers of hits, with up to 150 and 82 hits in primary and confirmatory assays, respectively, the hit frequency of compounds with PAINS substructures was overall surprisingly low. In primary assays (Figure 3a), median values were consistently between three and five hits per compound over the entire range of 100 to more than 500 assays. For comparison, for non-PAINS, the corresponding medians were two to three hits per compound. For 7888 PAINS tested in 401–500 assays, the median value was five hits per compound,

and for 2185 PAINS tested in more than 500 assays, the median was four hits. As shown in Figure 3a, at least 75% of all PAINS produced fewer than 10 hits, with a median of four hits per compound. For non-PAINS, at least 75% of the compounds produced fewer than seven hits, with medians of two to three hits per compound. Similar observations were made for confirmatory assays (Figure 3b), where medians for PAINS over increasing numbers of assays were two to three hits per compound. For non-PAINS compounds, the corresponding median values were also two to three hits per compound. For PAINS, there was one exception. For compounds tested in more than 150 assays, the median was 15 hits per compound. However, this sample only included 59 compounds and was thus too small for reaching statistically sound conclusions. For 8136 compounds tested in 101–150 dose–response assays, the median value was three hits per compound, and the same median was obtained for all PAINS tested in confirmatory assays. As shown in Figure 3b, at least 75% of the PAINS compounds produced fewer than seven hits. For comparison, at least 75% of the non-PAINS compounds produced fewer than five hits. Thus, hit rates were comparable for the majority of PAINS and non-PAINS and only slightly higher for PAINS.

For the top 20% of PAINS with the largest numbers of hits, the median values were 20 (primary) and 12 (confirmatory) hits per compound; for the top 10%, the medians were 30 (primary) and 17 (confirmatory) hits.

Since no activity threshold was applied, weak activities were taken into account, which resulted in an upper-limit assessment of hit rates. Furthermore, given that more than 23 000 compounds were tested in hundreds of assays, a proportion of the observed activities was expected to represent true positives. Accordingly, potential false-positive rates of PAINS were overall lower than anticipated for compounds that are thought to produce assay artifacts. For such compounds, one would intuitively expect observing multiassay activities in magnitude at least comparable to the top 10 or 20% as a rule rather than an exception. Instead, the majority of PAINS-containing compounds only produced hits in a few of many single-dose and dose–response assays, might they be artifacts or not.

Table 1 reports a small subset of PAINS substructures with large numbers of compounds having hit rates of at least 10%, including popular candidates.³ However, even these PAINS substructures represented in part large numbers of consistently inactive compounds. For comparison, Table 2 reports individual compounds with very high hit rates not classified as PAINS. These molecules have potential chemical liabilities⁵ and are most likely interference compounds.

We also determined the ratio of active PAINS-containing compounds relative to all active compounds. For extensively assayed compounds, the ratio was 0.08 and 0.09 for primary and confirmatory assays, respectively. For those compounds having hit rates of at least 10%, the ratio was 0.3 in both cases. Hence, there was a moderate increase in the proportion of PAINS among compounds with overall highest hit rates, due to a small subset of PAINS according to Table 1.

Exemplary Compounds. Figure 4 shows further examples of PAINS-containing compounds and their activity profiles, which illustrate the entire spectrum of observations made. In the left column, examples of aminothiophenes and phenyl-sulfonamides are shown that were tested in about 650 assays and never produced a hit. Furthermore, thiazolone- and pyrrole-containing compounds that were tested in nearly or

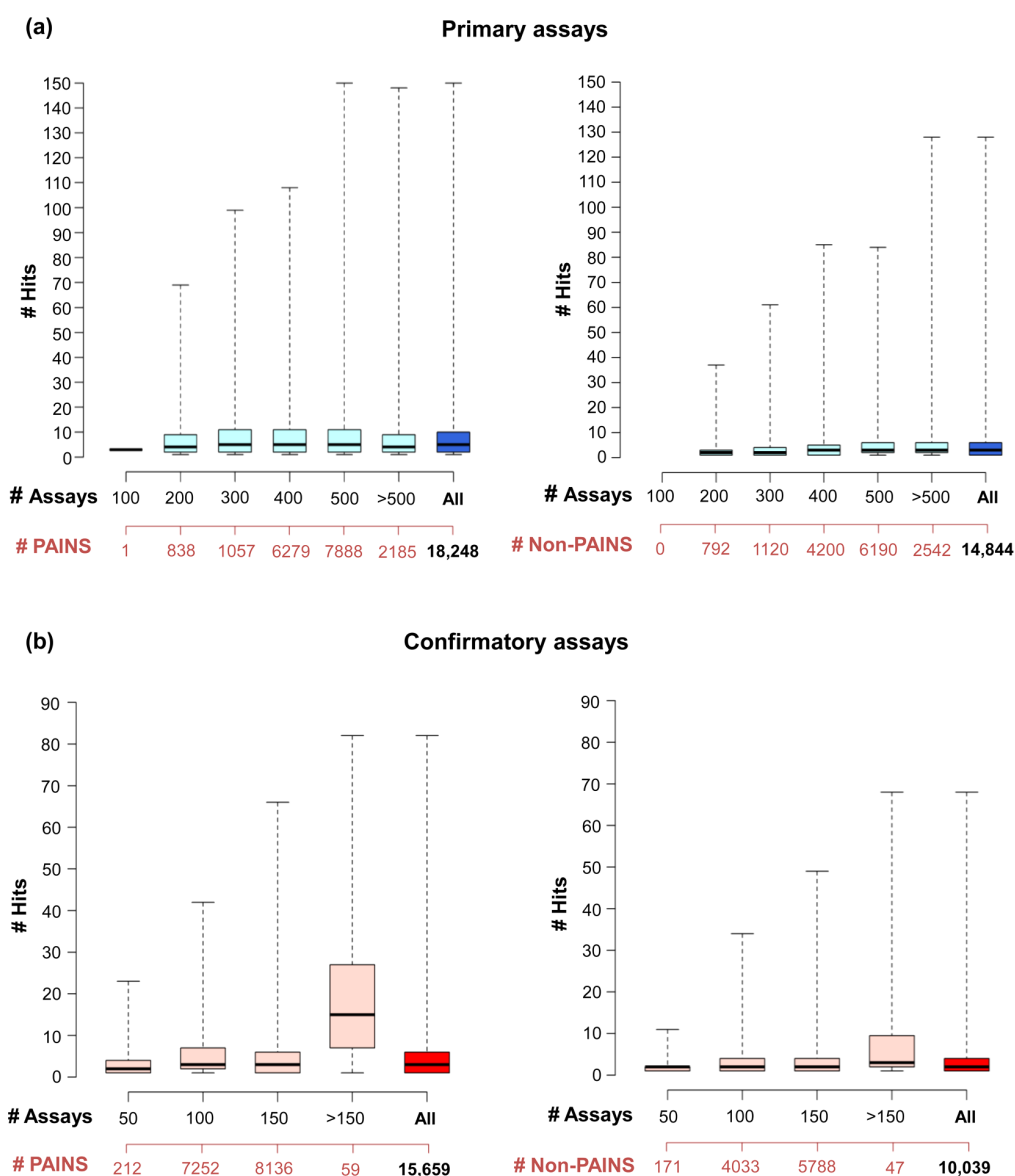


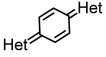









Figure 3. Activity profiles. For PAINS (left) and equally sized random samples of non-PAINS (right) from (a) primary and (b) confirmatory assays, box plots report the distributions of hits for subsets of compounds tested in increasing numbers of assays. “All” (bold) shows the distribution for all active PAINS and non-PAINS in primary or confirmatory assays. Box plots report the smallest value (bottom line), the first quartile (lower boundary of the box), the median value (thick line), the third quartile (upper boundary of the box), and the largest value (top line).

more than 600 assays yielded a total of only four or five hits. Furthermore, in the middle, compounds are shown that produced increasing numbers of hits, ranging from 17 for a thialindolizine and 45 for an aminoacridine to 69 and 107 for an aminomethylphenol and dimethylaminostyrene, respectively. The latter compounds provide examples of small subsets of compounds with PAINS substructures with large numbers of hits that were most likely artifacts. Moreover, in the right column, exemplary pairs of compounds containing classical PAINS substructures are shown that were either consistently inactive or active with high frequency. For example, the two *p*-benzoquinones were tested in comparable numbers of 252 and 261 assays, respectively. However, the first compound was consistently inactive, whereas the second produced a total of 91 hits. In addition, the small catechol was tested in 369 assays and was consistently inactive, whereas the larger catechol produced

135 hits in 592 assays. In the latter case, it was implausible that so many hits could be genuine.

Taken together, these examples illustrate the variety of PAINS assay phenotypes that were identified for more than 23 000 most extensively tested screening compounds containing PAINS substructures. The majority of PAINS were active in only a small number of assays. Variation of the assay conditions is expected to modulate the magnitude of compound reactivity and other potential causes of artifacts. However, comparison of these exemplary compounds and many others also suggested that the structural context in which PAINS are presented might often play an important role for their potential to produce artifacts. Accordingly, exploring PAINS substructure embedding in a systematic fashion might be an attractive goal for future research, further extending structure–interference analysis, for which the first examples have been reported.^{8,9}

Table 1. PAINS with High Hit Rates⁴²

PAINS designation	Substructure	Primary assays		Confirmatory assays	
		#cpds with hit rate >= 10%	#inactives	#cpds with hit rate >= 10%	#inactives
quinone_A(370)		150	18	305	34
catechol_A(92)		41	3	217	6
ene_six_het_A(483)		16	391	217	573
ene_rhod_A(235)		34	111	172	232
mannich_A(296)		39	178	144	400
anil_di_alk_A(478)		12	815	158	1074
imine_one_A(321)		43	49	103	79
anil_di_alk_B(251)		51	24	94	33
quinone_D(2)		42	3	85	9
ene_one_hal(17)		33	14	78	40

⁴²Listed are PAINS with the largest numbers of compounds having hit rates of at least 10% in primary and confirmatory assays.

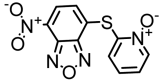

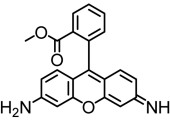

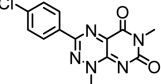
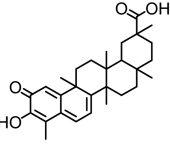
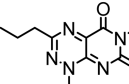
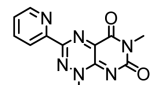
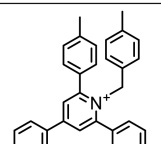
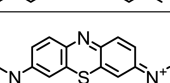
Concluding Remarks. In this work, we have systematically studied the behavior of compounds containing PAINS substructures in biological screening assays. There is no doubt that PAINS present a substantial problem for biological screening and medicinal chemistry. Waste of time and resources is inevitable if undetected PAINS with artificial activity enter compound optimization efforts. Moreover, false-positive activities reported in the literature are misleading at best and may catalyze further research that is doomed to fail. Clearly, indications of potential assay interference, for the detection of which current PAINS and aggregation filters—albeit imperfect—are helpful, must give rise to orthogonal assays and further experimental follow-up.¹¹

While many compound classes have been designated as PAINS, a critically important question is to what extent the notion of PAINS and artificial frequent hitter characteristics are supported by experimental data. Therefore, we have carried out a large-scale analysis of PAINS in primary and confirmatory assays and determined their activity profiles. Extensively assayed compounds were found to contain a subset of 270 PAINS, and nearly all of them were detected in both active compounds and thousands of others that were consistently inactive in single-dose and/or dose–response assays. For active compounds with PAINS substructures tested in increasing numbers of assays, hit rates were generally low, with median

values of two to five hits per compound. Only small subsets of compounds produced an abundance of hits. Moreover, the same PAINS substructure was often found in consistently inactive and frequently active compounds, suggesting that the structural context in which PAINS occur plays a role in causing undesired effects. Taken together, the results of our analysis reveal that PAINS are generally far from being excessively active. Rather, the hit frequencies vary greatly, and many PAINS are consistently inactive in different assays.

We return to the study of Capuzzi et al.¹⁶ Although analysis details differ, as expected for independent studies, the results reported herein are very similar to the findings of Capuzzi et al. for random PubChem compounds. Corresponding findings include assay frequencies of PAINS, the identification of the most reactive PAINS, and detection of many consistently inactive compounds. In fact, on the basis of both analyses, essentially equivalent conclusions can be drawn, including, among others, the likely structural context or “molecular environment”¹⁶ dependence of PAINS. At times when reproducibility of scientific studies—and the lack thereof—is a major issue, it is without doubt reassuring that these two independently performed analyses of PAINS in screening compounds are consistent. The interested reader may want to consider them side by side. Moreover, both the work of Capuzzi et al. and our analysis clearly indicate that the PAINS

Table 2. Other Compounds with High Hit Rates^a

ID	Structure	Primary assays			Confirmatory assays		
		#assays	#hits	hit rate	#assays	#hits	hit rate
313619		330	115	0.35	123	67	0.54
66541		380	125	0.33	138	53	0.38
24207752		312	100	0.32	82	24	0.29
460747		490	156	0.32	149	52	0.35
460749		372	117	0.31	92	32	0.35
4274774		216	61	0.28	62	29	0.47
3164059		355	99	0.28	129	43	0.33
906542		456	126	0.28	124	53	0.43
12647563		319	84	0.26	86	19	0.22
6099		463	84	0.18	131	49	0.37

^aExtensively tested compounds not classified as PAINS are shown with their PubChem IDs that produced highest hit rates in primary and confirmatory assays.

concept and practical assessment of PAINS are subject to further evaluation and refinement, just as much as the concept of colloidal aggregation, the result of another pioneering effort to control compound activity artifacts, has been continuously refined and evolved over time.^{12–15} The results of Capuzzi et al. and our analysis raise awareness of PAINS-relevant issues that require further exploration. However, neither Capuzzi et al. nor we offer practical solutions to address these issues. Thus, future investigations will be highly encouraged to translate the findings of rigorous large-scale data analysis into practical guidelines with utility for medicinal chemistry—and ultimately new computational tools.

As a step in this direction, the results of our analysis are made freely available. In our study, 270 of 480 PAINS listed in public filters were found in extensively assayed PubChem compounds. For each of these 270 PAINS, we make the total number of assayed compounds, the number of active compounds, their mean hit rates, and the number of consistently inactive compounds available in an open access deposition.²² These data should be helpful to investigators interested in reviewing or revising current PAINS filters.

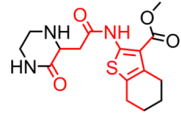
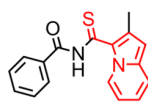
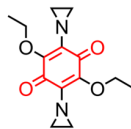
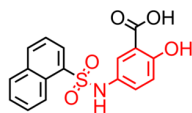
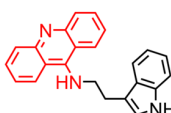
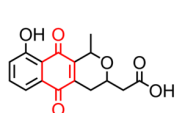
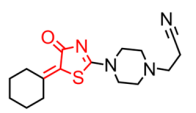
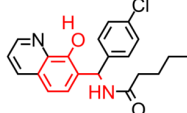
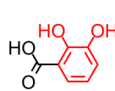
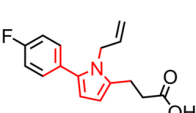
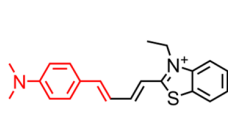
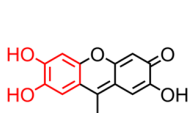
<p>aminothiophene</p> 	<p>thialindolizine</p> 	<p>p-benzoquinone</p> 																											
<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>531</td> <td>125</td> </tr> <tr> <td># Hits</td> <td>0</td> <td>0</td> </tr> </tbody> </table>		Prim	Conf	# Assays	531	125	# Hits	0	0	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>448</td> <td>99</td> </tr> <tr> <td># Hits</td> <td>12</td> <td>5</td> </tr> </tbody> </table>		Prim	Conf	# Assays	448	99	# Hits	12	5	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>197</td> <td>55</td> </tr> <tr> <td># Hits</td> <td>0</td> <td>0</td> </tr> </tbody> </table>		Prim	Conf	# Assays	197	55	# Hits	0	0
	Prim	Conf																											
# Assays	531	125																											
# Hits	0	0																											
	Prim	Conf																											
# Assays	448	99																											
# Hits	12	5																											
	Prim	Conf																											
# Assays	197	55																											
# Hits	0	0																											
<p>phenylsulfonamide</p> 	<p>aminoacridine</p> 	<p>p-benzoquinone</p> 																											
<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>521</td> <td>125</td> </tr> <tr> <td># Hits</td> <td>0</td> <td>0</td> </tr> </tbody> </table>		Prim	Conf	# Assays	521	125	# Hits	0	0	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>351</td> <td>77</td> </tr> <tr> <td># Hits</td> <td>34</td> <td>11</td> </tr> </tbody> </table>		Prim	Conf	# Assays	351	77	# Hits	34	11	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>207</td> <td>54</td> </tr> <tr> <td># Hits</td> <td>70</td> <td>21</td> </tr> </tbody> </table>		Prim	Conf	# Assays	207	54	# Hits	70	21
	Prim	Conf																											
# Assays	521	125																											
# Hits	0	0																											
	Prim	Conf																											
# Assays	351	77																											
# Hits	34	11																											
	Prim	Conf																											
# Assays	207	54																											
# Hits	70	21																											
<p>thiazolone</p> 	<p>aminomethylphenol</p> 	<p>catechol</p> 																											
<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>523</td> <td>124</td> </tr> <tr> <td># Hits</td> <td>2</td> <td>2</td> </tr> </tbody> </table>		Prim	Conf	# Assays	523	124	# Hits	2	2	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>355</td> <td>59</td> </tr> <tr> <td># Hits</td> <td>52</td> <td>17</td> </tr> </tbody> </table>		Prim	Conf	# Assays	355	59	# Hits	52	17	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>317</td> <td>52</td> </tr> <tr> <td># Hits</td> <td>0</td> <td>0</td> </tr> </tbody> </table>		Prim	Conf	# Assays	317	52	# Hits	0	0
	Prim	Conf																											
# Assays	523	124																											
# Hits	2	2																											
	Prim	Conf																											
# Assays	355	59																											
# Hits	52	17																											
	Prim	Conf																											
# Assays	317	52																											
# Hits	0	0																											
<p>pyrrole</p> 	<p>dimethylaminostyrene</p> 	<p>catechol</p> 																											
<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>484</td> <td>111</td> </tr> <tr> <td># Hits</td> <td>3</td> <td>2</td> </tr> </tbody> </table>		Prim	Conf	# Assays	484	111	# Hits	3	2	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>410</td> <td>97</td> </tr> <tr> <td># Hits</td> <td>77</td> <td>30</td> </tr> </tbody> </table>		Prim	Conf	# Assays	410	97	# Hits	77	30	<table border="1"> <thead> <tr> <th></th> <th>Prim</th> <th>Conf</th> </tr> </thead> <tbody> <tr> <td># Assays</td> <td>486</td> <td>106</td> </tr> <tr> <td># Hits</td> <td>106</td> <td>29</td> </tr> </tbody> </table>		Prim	Conf	# Assays	486	106	# Hits	106	29
	Prim	Conf																											
# Assays	484	111																											
# Hits	3	2																											
	Prim	Conf																											
# Assays	410	97																											
# Hits	77	30																											
	Prim	Conf																											
# Assays	486	106																											
# Hits	106	29																											

Figure 4. Exemplary compounds. Extensively tested compounds with PAINS substructures are shown that were consistently inactive or rarely active in all assays (left column) or produced hits with increasing frequencies (middle column). In addition, two pairs of compounds containing the same PAINS substructure are shown in which one compound was consistently inactive and the other was active with high frequency (right column). PAINS substructures are highlighted in red.

AUTHOR INFORMATION

Corresponding Author

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

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Alexander Tropsha for many constructive comments on our revised manuscript. We also thank Erik Gilberg for helpful discussions concerning PAINS substructures. The use of OpenEye's toolkits was made possible by their free academic licensing program.

ABBREVIATIONS USED

IMP, invalid metabolic panacea; PAINS, pan-assay interference compounds

REFERENCES

- (1) 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.
- (2) Baell, J. B.; Ferrins, L.; Falk, H.; Nikolakopoulos, G. PAINS: Relevance to Tool Compound Discovery and Fragment-Based Screening. *Aust. J. Chem.* **2013**, *66*, 1483–1494.
- (3) Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.
- (4) Nelson, K. M.; Dahlin, J. L.; Bisson, J.; Graham, J.; Pauli, G. F.; Walters, M. A. The Essential Medicinal Chemistry of Curcumin. *J. Med. Chem.* **2017**, *60*, 1620–1637.

(5) Gilberg, E.; Jasial, S.; Stumpfe, D.; Dimova, D.; Bajorath, J. Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology. *J. Med. Chem.* **2016**, *59*, 10285–10290.

(6) Baell, J. B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2016**, *79*, 616–628.

(7) Bisson, J.; McAlpine, J. B.; Friesen, J. B.; Chen, S. N.; Graham, J.; Pauli, G. F. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? *J. Med. Chem.* **2016**, *59*, 1671–1690.

(8) Dahlin, J. L.; Nissink, J. W.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med. Chem.* **2015**, *58*, 2091–2113.

(9) Dahlin, J. L.; Nissink, J. W.; Francis, S.; Strasser, J. M.; John, K.; Zhang, Z.; Walters, M. A. Post-HTS Case Report and Structural Alert: Promiscuous 4-Aroyl-1,5-Disubstituted-3-Hydroxy-2H-Pyrrol-2-One Actives Verified by ALARM NMR. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 4740–4752.

(10) Kilchmann, F.; Marcaida, M. J.; Kotak, S.; Schick, T.; Boss, S. D.; Awale, M.; Gönczy, P.; Reymond, J. L. Discovery of a Selective Aurora A Kinase Inhibitor by Virtual Screening. *J. Med. Chem.* **2016**, *59*, 7188–7211.

(11) Aldrich, C.; Bertozzi, C.; Georg, G. I.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz, K. M., Jr.; Schepartz, A.; Wang, S. The Ecstasy and Agony of Assay Interference Compounds. *ACS Cent. Sci.* **2017**, *3*, 143–147.

(12) Jadhav, A.; Ferreira, R. S.; Klumpp, C.; Mott, B. T.; Austin, C. P.; Inglese, J.; Thomas, C. J.; Maloney, D. J.; Shoichet, B. K.; Simeonov, A. Quantitative Analyses of Aggregation, Autofluorescence, and Reactivity Artifacts in a Screen for Inhibitors of a Thiol Protease. *J. Med. Chem.* **2010**, *53*, 37–51.

(13) Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *J. Med. Chem.* **2007**, *50*, 2385–2390.

(14) Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens Against Beta-Lactamase. *J. Med. Chem.* **2008**, *51*, 2502–2511.

(15) Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Eidam, O.; Mott, B. T.; Keiser, M. J.; McKerrow, J. H.; Maloney, D. J.; Irwin, J. J.; Shoichet, B. K. Complementarity Between a Docking and a High-Throughput Screen in Discovering New Cruzain Inhibitors. *J. Med. Chem.* **2010**, *53*, 4891–4905.

(16) Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 417–427.

(17) 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.

(18) RDKit: Cheminformatics and Machine Learning Software (2013). <http://www.rdkit.org>.

(19) Sterling, T.; Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.

(20) 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.* **2012**, *40*, D1100–D1107.

(21) Saubern, S.; Guha, R.; Baell, J. B. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol. Inf.* **2011**, *30*, 847–850.

(22) <http://www.zenodo.org>.

Summary

In our study, $\sim 23,000$ extensively tested compounds containing 270 PAINS substructures were identified and their activity profiles were analyzed. Surprisingly, many compounds containing PAINS substructures ($\sim 21\%$ in primary assays; $\sim 33\%$ in confirmatory assays) were found to be consistently inactive. PAINS with largest number of hits in primary and confirmatory assays as well as consistently inactive compounds were widely distributed over different PAINS substructures. Hit rates of PAINS varied but were often low, with median values of two to five hits for compounds tested in increasing numbers of assays. Only confined subsets of compounds produced an abundance of hits.

A variety of activity profiles were identified for extensively tested compounds containing PAINS substructures. Furthermore, the same PAINS substructure was often found in consistently inactive and frequently active compounds, suggesting that the structural context in which PAINS are presented plays an important role for interference potential. Thus, a majority of PAINS were not frequently active in screening assays. In fact, many PAINS were found to have DCM character. Therefore, it is important to distinguish between PAINS, which display high and low frequency of activity.

In the next chapter, we discuss how machine learning models can be used to characterize highly promiscuous and DCM PAINS.

Chapter 6

Machine Learning Distinguishes with High Accuracy between Pan-Assay Interference Compounds That Are Promiscuous or Represent Dark Chemical Matter

Introduction

Systematic analysis of compound activity data associated with PAINS has shown that PAINS have diverse activity profiles and PAINS substructures are found in compounds that are specifically active or inactive (*Chapter 5*). The structural environment of PAINS substructures is likely to play an important role for assay interference as well as for specific activity/inactivity.

PAINS filters have been viewed critically in the field as they are used to filter potentially liable compounds by only considering the presence of PAINS substructure, without taking structural context information into account. As PAINS filters are not fully reliable, exploring structural context of PAINS is of

high relevance for medicinal chemistry.

Given the large number of PAINS and the high variability of structural environments, it is difficult to formulate structural rules for differentiating between PAINS effects. Therefore, in this study, we used machine learning to distinguish between PAINS that are highly promiscuous from those having DCM character. Machine learning models were built using three methods including SVM, RF and DNN. Feature weighting and mapping were carried out to extract features responsible for correct predictions thus providing interesting insights into the structural environment of PAINS.

Reprinted with permission from “Jasial, S.; Gilberg, E.; Blaschke, T.; Bajorath J. Machine Learning Distinguishes with High Accuracy between Pan-Assay Interference Compounds That Are Promiscuous or Represent Dark Chemical Matter. *Journal of Medicinal Chemistry* **2018**, *61*, 10255-10264”. Copyright 2018 American Chemical Society

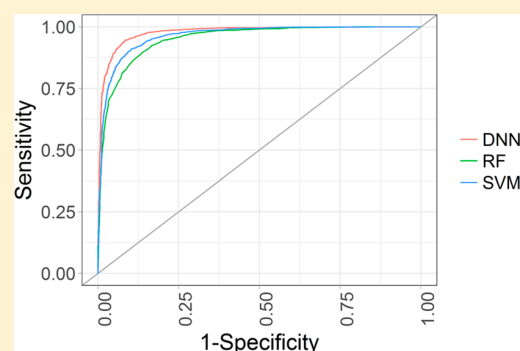
Machine Learning Distinguishes with High Accuracy between Pan-Assay Interference Compounds That Are Promiscuous or Represent Dark Chemical Matter

Swarit Jasial, Erik Gilberg, Thomas Blaschke, and Jürgen Bajorath*¹

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Endenicher Allee 19c, Rheinische Friedrich-Wilhelms-Universität, D-53115 Bonn, Germany

Supporting Information

ABSTRACT: Assay interference compounds give rise to false-positives and cause substantial problems in medicinal chemistry. Nearly 500 compound classes have been designated as pan-assay interference compounds (PAINS), which typically occur as substructures in other molecules. The structural environment of PAINS substructures is likely to play an important role for their potential reactivity. Given the large number of PAINS and their highly variable structural contexts, it is difficult to study context dependence on the basis of expert knowledge. Hence, we applied machine learning to predict PAINS that are promiscuous and distinguish them from others that are mostly inactive. Surprisingly accurate models can be derived using different methods such as support vector machines, random forests, or deep neural networks. Moreover, structural features that favor correct predictions have been identified, mapped, and categorized, shedding light on the structural context dependence of PAINS effects. The machine learning models presented herein further extend the capacity of PAINS filters.



INTRODUCTION

Compound screening is the major source of new chemical entities for drug discovery. However, biological screening is prone to artifacts due to assay interference for which a variety of potential mechanisms exist.^{1–10} False-positive hits that remain unrecognized and enter chemical optimization programs ultimately cause substantial waste of time and resources. Similarly, false-positives that propagate through the scientific literature may trigger further research activities that are doomed to fail. Hence, much awareness has been raised in recent years of compound classes that might cause assay interference by different mechanisms, which is often difficult to recognize.^{3,5,7,9} Clearly, confirmation of a true mechanism of action responsible for a specific biological activity of a compound is a crucial step before subjecting a candidate to chemical optimization and further development.^{1–5} Compound liabilities leading to assay interference include, among others, aggregation effects, autofluorescence and quenching, covalent modification of target proteins and assay reagents, redox effects, or metal chelation.^{6–9}

Although interference phenomena strongly depend on experimental conditions, systematic attempts have been made to identify problematic compound classes. Such efforts have uncovered colloidal aggregators^{2,3} and pan-assay interference compounds (PAINS),^{6,7,9} which include many different compound classes such as anilines, rhodanines, curcuminoids, Michael acceptors, or Mannich bases.^{6,7,9–11} Typically, PAINS occur as substructures in larger compounds. Originally, 480

compound classes have been designated as PAINS.⁶ Some of these classes and underlying interference mechanisms have been further explored and detailed, providing multifaceted views of interference phenomena and recommendations how to best address them.^{8,12–14}

However, systematic analyses of compound activity data associated with proposed PAINS have also shown that PAINS often have very different activity profiles and that PAINS substructures are also found in compounds that are specifically active or inactive.^{15–17} For example, X-ray structures of protein–PAINS complexes confirmed that prominent PAINS such as catechols or alkylindoles⁷ can engage in specific target–ligand interactions.^{18,19} Moreover, in some instances, proposed interference mechanisms and complex formation revealed by X-ray crystallography follow similar routes.¹⁸

Different lines of evidence suggest that the structural context in which PAINS substructures are presented plays a decisive role for assay interference and false-positive signals on one hand and specific activity or inactivity on the other.^{15,16,18,20,21} For example, chemical modifications of structural analogs containing PAINS substructures have been shown to strongly influence their hit rates in biological screening assays.²¹ Furthermore, significant numbers of compounds with PAINS substructures have been identified that were consistently inactive in screening assays^{15–17} including compounds tested

Received: September 10, 2018

Published: November 13, 2018

in hundreds of primary or confirmatory assays.¹⁷ Hence, the latter screening compounds qualify as “dark chemical matter” (DCM),²² i.e., small molecules for which any potential biological activity is yet to be confirmed.

Given that PAINS cover a wide spectrum of activity profiles, ranging from inactive over specifically active to highly promiscuous compounds, the PAINS concept has also been viewed critically, especially with respect to PAINS filters.¹⁶ Such filters compile PAINS for substructure searching to flag compounds with potential liabilities.⁶ For example, a point of critique has been that original PAINS filters were based on limited assay data, calling generalization into question.¹⁶ Moreover, PAINS filters do not contain structural context information. However, albeit imperfect, such filters provide initial alerts to carefully consider the apparent activity of compounds with PAINS substructures, for example, by testing them in orthogonal assays.¹⁴

Given the intrinsically limited reliability of PAINS filters to detect artificial activities, exploring the structural context dependence of PAINS effects is of high relevance for medicinal chemistry. However, while this structural context dependence has been analyzed on a case-by-case basis,^{18,21} large-scale analysis on the basis of expert knowledge is essentially prohibitive. Furthermore, it is difficult to formulate structural rules for differentiating between PAINS effects, given the large number of potentially liable substructures and the high variability of structural contexts.

In light of this situation, we have reasoned that machine learning might be investigated to distinguish between PAINS with high and low frequency of activity and build models to predict if a compound containing a PAINS substructure is likely to display assay promiscuity. Therefore, we have generated data sets containing highly promiscuous PAINS (PROM_PAINS) and PAINS having DCM character (DCM_PAINS) and investigated different machine learning algorithms for their ability to distinguish between these PAINS phenotypes. The resulting models have been unexpectedly successful in predicting promiscuous and inactive compounds with PAINS substructures. Feature weighting and mapping were carried out to rationalize predictions and better understand the structural context dependence of PAINS effects. The results of our analysis are presented in the following.

RESULTS

PAINS Characteristics and Distribution. The analysis was deliberately focused on extensively tested compounds with PAINS substructures that displayed unusually high hit rates in screening assays and others that were consistently inactive in all assays they were tested. Thus, these compound sets marked opposite ends of the PAINS activity spectrum. Although PROM_PAINS might have true biological activities, their unusually high hit frequency is indicative of artifacts. On the other hand, some DCM_PAINS might be compromised or insoluble under assay conditions, which could explain why they do not display any activity. However, the availability of more than 3000 qualifying DCM_PAINS that were extensively tested makes it highly unlikely that many of these compounds might be compromised in one way or another.

Our training and test sets covered a total of 212 different PAINS, 74 of which were shared by PROM_PAINS and DCM_PAINS. Hence, 118 and 20 PAINS substructures and

compounds containing them were unique to PROM_PAINS and DCM_PAINS, respectively.

Machine Learning and Resulting Models. Given the large number of different machine learning algorithms that are available, the choice of support vector machine (SVM) and random forest (RF) algorithms was motivated by their typically high performance in compound classification and ranking. In addition, the deep neural network (DNN) method was selected in light of the increasing popularity of deep learning in chemistry. Despite algorithmic differences, machine learning methods used for classification have in common that they associate molecular features with given class labels (e.g., PROM_PAINS versus DCM_PAINS) during model building. Hence, training data representing different class labels must be available. It is important to note that activity or reactivity is not used as a parameter during learning. Rather, classification is solely guided by associating feature distributions with given class labels. A known conundrum in machine learning is model derivation on the basis of unbalanced data sets, due to the availability of many more negative than positive training instances and vice versa. Data imbalance often reduces model quality or biases classification calculations, depending on the composition of test data sets. Therefore, it is important to take data imbalance into account during machine learning and derive alternative models on the basis of training sets with varying composition. This has also been taken into consideration when deriving classification models for PAINS, as discussed in the following.

Global Models. First, global models were built using SVM, RF, and DNN to distinguish between PROM_PAINS and DCM_PAINS taking all 212 PAINS substructures into account. Alternative performance measures were applied including the area under the ROC curve (AUC ROC), Matthew's correlation coefficient (MCC), and balanced accuracy (BA). Table 1 reports prediction results for global models using ECFP4 and MACCS as molecular representations. In addition, Figure 1 shows ROC curves for individual trials. Prediction accuracy of different models was generally high and comparable for SVM, RF, and DNN. In fact, differences

Table 1. Performance of Global Models^a

		ECFP4		
		SVM	RF	DNN
AUC ROC	Mean	0.960	0.946	0.952
	SD	0.001	0.002	0.003
MCC	Mean	0.797	0.750	0.776
	SD	0.007	0.010	0.011
BA	Mean	0.902	0.886	0.887
	SD	0.004	0.006	0.004
		MACCS		
		SVM	RF	DNN
AUC ROC	Mean	0.935	0.933	0.912
	SD	0.001	0.001	0.002
MCC	Mean	0.732	0.720	0.680
	SD	0.008	0.006	0.006
BA	Mean	0.863	0.870	0.835
	SD	0.004	0.003	0.004

^aReported are the mean and standard deviation (SD) of AUC ROC, MCC, and BA values for 10 independent trials using SVM, RF, and DNN global models with ECFP4 and MACCS fingerprints as descriptors.

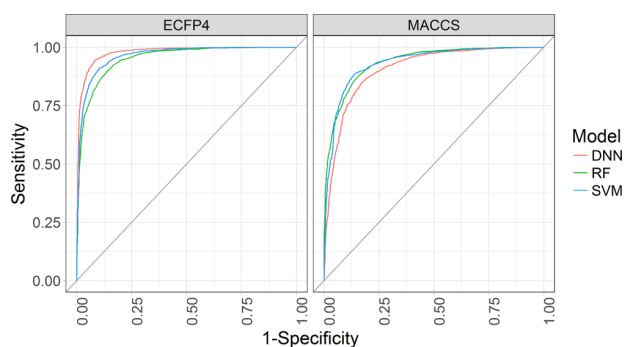


Figure 1. Receiver operating characteristic curves for global models. ROC curves are shown for an individual trial distinguishing PROM_PAINS and DCM_PAINS using SVM (blue), RF (green), and DNN (red) global models on the basis of ECFP4 and MACCS.

between these methods were only small. Furthermore, standard deviations of independent trials were also small in all cases (Table 1), indicating the presence of stable predictions. For ECFP4, high AUC ROC values of ~ 0.950 were obtained for SVM, RF, and DNN as well as high MCC and BA values, ranging from 0.750 (RF) to 0.797 (SVM) and from 0.886 (RF) to 0.902 (SVM), respectively.

When using the simpler MACCS fingerprint instead of ECFP4 as a molecular representation, only slightly lower values were obtained (Table 1). ROC curves further illustrate the comparable performance using ECFP4 and MACCS (Figure 1). There was no notable advantage of DNN over SVM and RF. All methods reached high performance levels, more so than we anticipated, leaving only little room for further improvements.

From Global to Balanced Models. We considered that the high accuracy of global models might be attributable to 118 PAINS substructures that were only present in PROM_PAINS and 20 others only present in DCM_PAINS. Compounds containing class-specific substructures might be straightforward to classify, which would at least in part explain the observed accuracy. In addition, the PAINS substructures shared by PROM_PAINS and DCM_PAINS were represented by larger numbers of PROM_PAINS, as shown in Figure 2, which might also favor their identification.

To investigate potential substructure bias and data imbalance, we generated training and test sets that exclusively contained shared PAINS substructures and the same number of PROM_PAINS and DCM_PAINS per substructure. These sets, which were used to generate and evaluate balanced models, comprised 54 shared substructures, and each contained a total of ~ 1900 compounds (see Experimental Section). They provided challenging conditions for predictions because accurate results could only be anticipated here if models were able to distinguish between different structural contexts in which shared PAINS substructures were embedded. We expected that the variety of structural contexts in which PAINS substructures were found in active and inactive compounds^{17,18,21} would make these predictions very difficult.

Balanced Models. Table 2 reports prediction results for balanced SVM, RF, and DNN models, and Figure 3a shows ROC curves for individual trials. Balanced models displayed the same characteristics as global models. The relative performance of balanced SVM, RF, and DNN models was

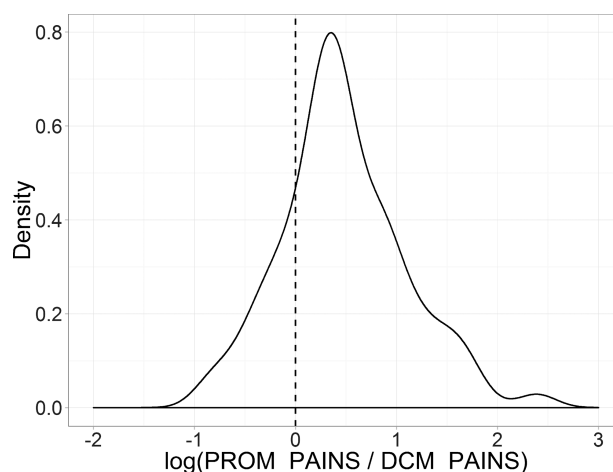


Figure 2. Data imbalance. A density plot is shown for the logarithmic ratio of PROM_PAINS and DCM_PAINS for shared PAINS substructures. The shift of the distribution toward positive log values indicates the presence of more PROM_PAINS than DCM_PAINS.

Table 2. Performance of Balanced Models^a

		ECFP4		
		SVM	RF	DNN
AUC ROC	Mean	0.920	0.910	0.901
	SD	0.006	0.005	0.008
MCC	Mean	0.681	0.658	0.651
	SD	0.018	0.012	0.014
BA	Mean	0.841	0.830	0.825
	SD	0.009	0.006	0.007
		MACCS		
		SVM	RF	DNN
AUC ROC	Mean	0.878	0.882	0.837
	SD	0.005	0.005	0.007
MCC	Mean	0.603	0.597	0.529
	SD	0.012	0.012	0.014
BA	Mean	0.802	0.799	0.764
	SD	0.006	0.006	0.007

^aReported are the mean and standard deviation (SD) of AUC ROC, MCC, and BA values for 10 independent trials using SVM, RF, and DNN balanced models with ECFP4 and MACCS fingerprints as descriptors.

very similar and, again, only slightly better using ECFP4 than MACCS. Balanced models also produced stable predictions with small standard deviations across different trials. They yielded high AUC ROC values greater than 0.900 as well as high MCC and BA values, ranging from 0.651 (DNN) to 0.681 (SVM) and from 0.825 (DNN) to 0.841 (SVM), respectively (Table 2, ECFP4). Compared to global models, the performance of balanced models was only lower by a small margin, on the order of 5%, when assessed on the basis of AUC ROC and BA, and 10% on the basis of MCC values. Thus, although balanced models were trained to distinguish between PROM_PAINS and DCM_PAINS containing the same PAINS substructure, they yielded accurate predictions similar to global models, another surprising finding. It indicated that alternative models were capable of distinguishing between PAINS substructures embedded in compounds representing different structural environments, which were characteristic of

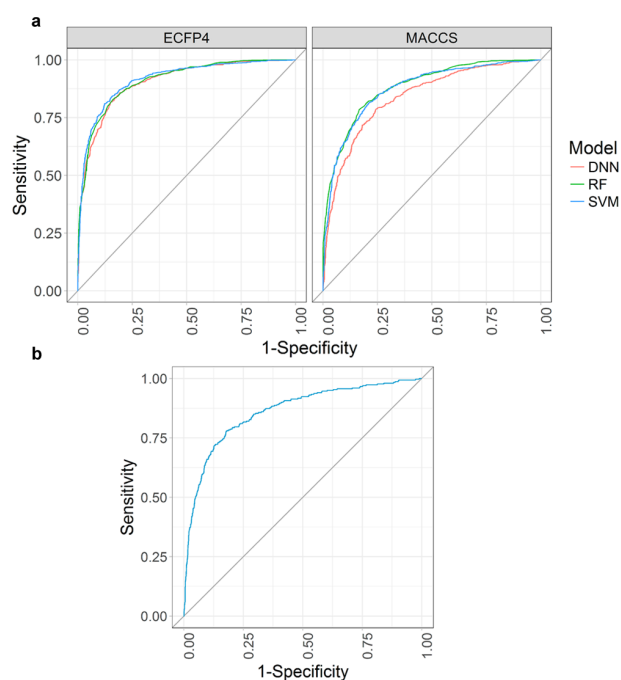


Figure 3. Receiver operating characteristic curves for balanced models. (a) ROC curves are shown for an individual trial distinguishing PROM_PAINS and DCM_PAINS using SVM (blue), RF (green), and DNN (red) balanced models on the basis of ECFP4 and MACCS. (b) A ROC curve is shown for an individual trial distinguishing PROM_PAINS, DCM_PAINS, and randomly selected ZINC compounds using the balanced SVM on the basis of ECFP4.

PROM_PAINS on the one hand and DCM_PAINS on the other.

To assess the performance of the balanced SVM model as a PROM_PAINS filter for compound libraries, we carried out five virtual screening-type calculations using ECFP4 as a molecular representation. For each trial, a screening library consisting of 300 randomly selected PROM_PAINS from the test set, 300 random DCM_PAINS from the test set, and 3000 randomly chosen ZINC compounds was assembled. DCM_PAINS and ZINC compounds were regarded as false positives. ZINC compounds were required to generate ECFP4 features present in PROM_PAINS. The SVM model yielded an average ROC AUC of 0.87 and achieved a TP rate of 62% within the top 100 ranked compounds. Figure 3b shows a ROC curve for an individual trial. Although the model was trained to distinguish between PROM_PAINS and DCM_PAINS, it was also capable of distinguishing PROM_PAINS from random ZINC compounds.

Taken together, the results revealed a level of accuracy of global and balanced models in differentiating PROM_PAINS, DCM_PAINS and randomly selected compounds that was higher than anticipated. Moreover, in single SVM trials, only one of 54 PAINS classes (code “thiaz_ene_B”) was entirely incorrectly predicted, as shown in Figure 4. In this case, however, only two training and test compounds were available. Hence, effective learning was essentially impossible due to data sparseness, providing an explanation for the incorrect prediction.

Model Diagnostics. In light of our findings, we investigated how to further assess and better understand

PAINS class	PAINS substructure	
Thiaz_ene_B		
Structure	Prediction	Experimental
	DCM_PAINS	PROM_PAINS
	PROM_PAINS	DCM_PAINS

Figure 4. Incorrectly predicted PAINS class. Shown are test compounds containing a PAINS substructure (code: thiaz_ene_B) that were incorrectly predicted using the balanced SVM model. In this case, only two training and test compounds were available. The PAINS substructure is shown in red.

successful predictions. Due to the black box character of machine learning models, especially DNN, there was no direct access to determinants of predictions. However, we adapted a feature weighting approach developed for SVM (see Experimental Section) to further explore the predictions. Given the consistency of predictions using different models, understanding which topological (ECFP4) features were prevalent in PROM_PAINS vs DCM_PAINS and vice versa (and hence contributed differentially to predictions) provided an opportunity to interpret classification results.

Feature Weights. For all ECFP4 features of test compounds in balanced data sets, cumulative positive and negative feature weights (see Experimental Section) were determined. Then, features were ranked for PROM_PAINS and DCM_PAINS on the basis of positive and negative weight sums, respectively. For example, for a given balanced test set, a total of 19 668 ECFP4 features were detected, 2573 and 2527 of which were found to have positive and negative feature sums, respectively. Top ranked features were identified for PROM_PAINS (termed positive features) and DCM_PAINS (negative features) and further analyzed.

Feature Analysis. Highly ranked positive and negative ECFP4 features typically differed. Negative features preferentially included aliphatic carbon atoms with varying levels of hydrogens, cyclic aliphatic ethers, and sp^2 -hybridized oxygens of carbonyl and sulfonyl groups. In contrast, highly ranked positive features included patterns from conjugated ring systems, chlorine and sulfur atoms, and β -unsaturated carbons bisecting a ring system. Hence, positive features were preferentially associated with reactive moieties, whereas negative features were chemically more inert, which we considered an interesting finding. For example, Figure 5a shows an N,N-disubstituted aniline pattern that is a part of a PAINS substructure and preferentially found in PROM_PAINS. In fact, aromatic tertiary amines are present in 23 PAINS substructures^{6–10} and thought to compromise fluorometric assays due to quenching.²³ Figure 5b shows a feature comprising doubled bonded sp^2 -hybridized carbons representing an unsaturated bond connecting two ring systems, which was also prevalent in PROM_PAINS. This atom arrangement is a part of a recurrent Michael acceptor motif

a

Contribution	SMARTS representation	Rank
Positive	<chem>C=C=C=C</chem>	10
Examples:		

b

Contribution	SMARTS representation	Rank
Positive	<chem>C=C=C</chem>	14
Examples:		

Figure 5. Positive ECFP4 features. Shown are features (in SMARTS representation) from SVM predictions that were highly ranked for PROM_PAINS. In exemplary PROM_PAINS, the mapped structural feature is traced on a gray background and the PAINS substructure is colored red. (a) shows an *N,N*-disubstituted aniline ECFP4 feature detected in different types of aniline PROM_PAINS (PAINS codes `anil_di_alk_A` and `anil_di_alk_D`). (b) shows an ECFP4 feature consisting of three doubled-bonded sp²-hybridized carbon atoms found in PROM_PAINS (codes `het_pyridinium_A` and `ene_ene`).

that is reactive and found in a variety of PAINS substructures. By contrast, Figure 6a and Figure 6b show examples of highly ranked negative features representing amide bond and morpholino patterns that mostly occur outside PAINS substructures and are stable under physiological conditions. In Figure 7, exemplary features with positive weight in PROM_PAINS are shown and their likely chemical reactivity is given. The frequency of occurrence of these features in PROM_PAINS was higher than in DCM_PAINS, as also reported.

The analysis revealed that calculated ECFP4 features distinguishing PROM_PAINS and DCM_PAINS often accounted for structural patterns associated with potential reactivity of PROM_PAINS, which provided a rationale for successful global classification using different methods. However, these observations did not explain why compounds containing the same PAINS substructure were correctly classified as promiscuous or inactive.

Structural Context Analysis. Differences in the activity of PAINS containing the same substructure and their correct prediction can only be rationalized by studying the structural environment. To address this issue at the level of model-based classification, top ranked ECFP4 features were mapped onto PROM_PAINS and DCM_PAINS and categorized with respect to their structural context. Accordingly, *subset* features

a

Contribution	SMARTS representation	Rank
Negative	<chem>O=C(N)</chem>	10
Examples:		

b

Contribution	SMARTS representation	Rank
Negative	<chem>C-O-C-N</chem>	20
Examples:		

Figure 6. Negative ECFP4 features. Shown are features (in SMARTS representation) from SVM predictions that were highly ranked for DCM_PAINS. In exemplary DCM_PAINS, the mapped structural feature is traced on a gray background and the PAINS substructure is colored red. (a) shows an amide-bond feature found in DCM_PAINS (codes `dyes5A` and `ene_rhod_A`) and (b) an aliphatic heterocycle (codes `ene_six_het_A` and `azo_A`).

Feature	Chemical reactivity	Ratio (PROM_PAINS/DCM_PAINS)	PROM_PAINS example
<chem>C=C=C</chem>	electrophile	3.21	
Cl	electron withdrawing	1.53	
<chem>C=C=C=C</chem>	quenching	2.48	

Figure 7. Prominent features with positive weights. For exemplary positive features, the SMARTS representation and their likely chemical reactivity are provided. In addition, the ratio of their frequency of occurrence in PROM_PAINS versus DCM_PAINS is reported. For each feature an exemplary compound is given. The structural feature is traced on a gray background, and the PAINS substructure is colored red.

were entirely contained in PAINS substructures, *intersection* features were part of a PAINS substructure and part of its structural environment, and *distinct* features completely mapped outside PAINS substructures.

Figure 8a shows exemplary positive ECFP4 features belonging to different categories for unsaturated rhodanines

a			b		
PAINS class	PAINS substructure		PAINS class	PAINS substructure	
Ene_rhod_A			Sulfonamide_B		
SMARTS representation	Context category	Example	SMARTS representation	Context category	Example
1	C	subset	$C=C=C=C$	subset	
2	S	subset	$C=C=C=C$	intersection	
3	$C=C=C$	intersection	$C=C=C$	intersection	
4	$C=C=C=C$	distinct	$C=C=C=C$	intersection	
5	Cl	distinct			

Figure 8. Structural context analysis. For two prominent PAINS classes⁶ including (a) ene_rhod_A and (b) sulfonamide_B, positive ECFP4 features are mapped onto exemplary compounds. Mapped features are traced on a gray background, and the PAINS substructure is colored red. For each feature, the structural context category is given, as defined in the text.

(PAINS code “ene_rhod_A”), a prominent PAINS class. Reported interference potential of these five-membered heterocycles with an exocyclic double bond includes protein reactivity, covalent modification, or metal chelation,^{24–26} despite their popularity as scaffolds in drug discovery.²⁷ The potential Michael acceptor activity of the exocyclic double bond provides a prime example for structural context dependency since many compounds containing this substructure have distinct activity profiles, including PROM_PAINS and DCM_PAINS. The identification of positive and negative features not only provided a rationale for successful classification, as discussed above, but their mapping and categorization also helped to explain the influence of structural contexts on predictions. In Figure 8a, positive features, which were absent or underrepresented in DCM_PAINS, are mapped onto correctly predicted PROM_PAINS. Both subset feature 1 and intersection feature 3 cover a sp^2 hybridized exocyclic carbon at which the attack of a reactive nucleophile such as a thiol compound occurs. Moreover, the intersection feature 3 highlights the critically important role of an additional ring system conjugated with the Michael acceptor for its activity. Consequently, the presence of an aromatic feature 4 in the vicinity of the reactive double bond also favors Michael acceptor reactivity, which is further supported by increasing the electrophilicity of the double bond, for example, by a chlorine substitution (feature 5).

Figure 8b shows positive features characterizing *p*-hydroxyarylsulfonamides (code “sulfonamide_B”), another prominent PAINS class prone to redox and thiol reactivity.^{8,28,29} The *p*-hydroxyarylsulfonamides typically display assay interference if they contain a naphthalene core,⁸ which is covered by the intersection features.

Feature mapping did not always provide consistent and easily interpretable results. An example is shown in Figure 9. Positively weighted features detected for alkylindoles (code “indol_3yl_alk”) could not be directly associated with the

PAINS class	PAINS substructure	
Indol_3yl_alk		
SMARTS	Context category	Example
1	$C-C=N$	intersection
2	$C=C=C$	distinct
3	$C-O-CH_3$	distinct

Figure 9. Noninterpretable feature mapping. For the PAINS class indol_3yl_alk, positive ECFP4 features are mapped onto exemplary compounds. For each feature, the structural context category is given. Mapped structural features are traced on a gray background, and the PAINS substructure is colored red.

likely interference mechanism of this class, which is proposed to rely on the Michael acceptor reactivity of the indoline 3-position of the corresponding tautomer.⁶ Here, positive features in test compounds mapped to conjugated ring systems distant from the reactive moiety, which could not be interpreted in chemical terms.

However, we also found that the absence or presence of a positive or negative feature in structurally similar compounds could be directly related to correct predictions by the SVM model. Figure 10 shows two exemplary PAINS classes. In the

PAINS class	PROM_PAINS	DCM_PAINS
Ene_five_het_A		
Ene_cyano_A		

Figure 10. Correctly predicted PROM_PAINS and DCM_PAINS. For two PAINS classes (codess ene_five_het_A and ene_cyano_A), exemplary PROM_PAINS and DCM_PAINS are shown in the absence or presence of mapped ECFP4 features. Mapped structural features are traced on a gray background, and the PAINS substructure is colored red.

first example, the conjugated ring system in the immediate neighborhood of the exocyclic double bond of the PAINS substructure (code “ene_five_het_A”) rationalized the correct prediction of PROM_PAINS. By contrast, when this feature was absent and replaced by a secondary amine, another compound was correctly predicted as a DCM_PAINS. In the second example, reactivity of the PAINS motif (code “ene_cyano_A”) might be regulated by the electron density at the exocyclic unsaturated carbon. Introducing an electron donating methoxy substituent is expected to decrease the likelihood of nucleophilic attacks at this site. As shown in Figure 10, a methoxy group represented a feature negatively weighted by the SVM model and its presence supported the correct prediction of DCM_PAINS.

DISCUSSION AND CONCLUSIONS

In this work, we introduce machine learning models to distinguish promiscuous PAINS from those that were inactive across many different screening assays. The large number of available PAINS, the complex nature of PAINS effects, and the critical role of the structural context in which PAINS are presented have thus far precluded the derivation of expert rules for predicting PAINS with different activity. In light of these challenges, we have investigated machine learning to facilitate such predictions. Accurate models were obtained, as revealed using different performance measures. Importantly, model performance did not significantly depend on the methods that were used, and deep learning was not required to achieve accurate predictions. To exclude the influence of data imbalance and statistical bias on predictions, we also built models for balanced data sets that exclusively consisted of

substructures shared by PROM_PAINS and DCM_PAINS. Although these data sets provided challenging conditions for predictions, the high performance level of all models nearly remained constant. Taken together, these findings indicated that structural/topological patterns distinguished promiscuous from inactive PAINS that could be readily explored and exploited by machine learning. However, given the black box character of machine learning models, a general and widely appreciated shortcoming, the predictions were difficult to interpret. The lack of interpretability is a major issue for machine learning applications in medicinal chemistry, much more so than for data mining. Therefore, we went a step further and indirectly assessed predictions through feature weighting and mapping. First, we identified highly weighted positive and negative features and found that many positive features of PROM_PAINS were associated with PAINS substructures and reactive moieties, which provided an explanation for their relevance. Moreover, in exemplary cases, mapping and categorization of features shed light on important differences in structural contexts, lending credence to predictions. However, not all successful and unsuccessful predictions can be rationalized on the basis of feature mapping, which reflects the general weakness of machine learning approaches referred to above. Another weakness of machine learning is the known compound class dependence of activity predictions, which is difficult to rationalize and also likely to influence predictions of PROM_PAINS under varying calculation conditions. By contrast, a particular strength of machine learning models for PAINS-based activity prediction is that they take structural context information for PAINS substructures into account, as shown herein, which sets such models apart from simple PAINS substructure filters. This is an important aspect because expert analysis will not be capable of exploring structural features giving rise to context dependence on a large scale, due to the very large number of structural patterns that need to be explored. Large-scale analysis of PAINS substructure environments inevitably depends on machine learning.

Taken together, the findings reported herein were encouraging, and we therefore intend to further expand mapping of topological features to a variety of PAINS classes and reconcile the results with observed differences in activity. This might provide an improved basis for deriving general rules to predict context-dependent PAINS reactivity.

As a part of our study, we have made our PROM_PAINS and DCM_PAINS data sets as well as the global and balanced SVM models available as open access deposition on the ZENODO platform.³⁰ Since the models were capable of accurately predicting PROM_PAINS, they further extend PAINS substructure filters providing initial alerts for follow-up investigations.

EXPERIMENTAL SECTION

Promiscuous and Consistently Inactive Compounds with PAINS Substructures. A large set of 437 257 extensively tested compounds³¹ was obtained from PubChem BioAssays.³² All of these compounds were tested in both primary assays (percentage of inhibition from a single dose) and confirmatory assays (dose-response assays yielding IC₅₀ values). From three public PAINS filters available in RDKit,³³ ZINC,³⁴ and ChEMBL,³⁵ SMARTS³⁶ representations of PAINS were obtained and used as substructure queries to search the set of extensively assayed PubChem compounds for PAINS. SMARTS strings from three PAINS filters were used to account for possible implementation discrepancies.

A subset of 27 520 screening compounds were found to contain substructures representing 270 different PAINS.¹⁵ The assay activity profiles of all compounds containing PAINS substructures were determined. On the basis of the activity profiles, two data sets were generated. The first set consisted of compounds with PAINS substructures that were consistently inactive in at least 100 primary and varying numbers of confirmatory assays they were tested in and hence had DCM character (termed DCM_PAINS). The DCM_PAINS set contained 3059 compounds. The second set consisted of promiscuous compounds with PAINS substructures (termed PROM_PAINS). Compounds were classified as PROM_PAINS if they were tested in at least 100 primary and varying numbers of confirmatory assays and were active in 10 or more assays (4944 compounds). In addition, compounds were selected that were tested in at least 50 confirmatory assays and varying numbers of primary assays and were active in 10 or more assays (279 compounds). Accordingly, the PROM_PAINS set contained a total of 5223 compounds. DCM_PAINS and PROM_PAINS represented 94 and 192 PAINS substructures, respectively, 74 of which were common to both sets. Table S1 of the Supporting Information reports the 74 shared PAINS substructures and the number of DCM_PAINS and PROM_PAINS compounds representing them. It cannot be ruled out that some apparent DCM_PAINS might result from highly reactive PAINS or might not be stable under experimental conditions. For example, if PAINS already react when solubilized for an experiment, they might be recorded as DCM_PAINS although the original reactive molecule is not tested.

Training and Test Sets for Classification Models. *Global Models.* For the generation of “global models”, the PROM_PAINS and DCM_PAINS sets were randomly divided 10 times into equally sized training (50%) and test sets (50%). Ten global models were independently generated, and their results were averaged. Global models were based on all PAINS substructures, regardless of the number of PROM_PAINS and DCM_PAINS that represented them.

Balanced Models. Classification models were also built for “balanced” training sets in which the number of PROM_PAINS and DCM_PAINS was adjusted for shared PAINS substructures. Therefore, 54 (of 74) shared PAINS substructures were selected that were represented by at least two PROM_PAINS and DCM_PAINS (Table S1). For each substructure, the smaller subset of PROM_PAINS or DCM_PAINS was selected and the same number of PROM_PAINS or DCM_PAINS was randomly sampled for the larger subset. Accordingly, each individual substructure was represented by the same number of PROM_PAINS and DCM_PAINS compounds, which were then randomly divided into training (50%) and test compounds (50%). Ten independent sampling trials were carried out yielding 10 balanced training and test sets, each consisting of ~1900 compounds (since some compounds contained more than one PAINS substructure, actual numbers slightly varied over different sampling trials). Classification models built and evaluated on the basis of balanced sets are referred to as “balanced models”.

Molecular Representations. Compounds were represented as extended connectivity fingerprints of bond diameter 4 (ECFP4)³⁷ and MACCS structural keys.³⁸ ECFP4 is a representative feature set fingerprint enumerating layered atom environments, which are encoded by integers using a hashing function. Feature-to-bit mapping was recorded to enable mapping of fingerprint features to test compounds. Each feature was also represented as a SMARTS pattern. MACCS is a binary-keyed fingerprint comprising 166 bits, each of which accounts for the presence or absence of a structural pattern. ECFP4 and MACCS fingerprints were generated using in-house Python scripts based upon the OEChem³⁹ and RDKit³³ toolkits, respectively.

Machine Learning Methods. Models were generated using SVM, RF, and DNN algorithms and applied to classify PROM_PAINS and DCM_PAINS. For building predictive models, training instances were defined by a feature vector $x \in \mathcal{X}$ and a class label $y \in \{-1, 1\}$.

Support Vector Machine. SVM is a supervised learning algorithm that derives a separating hyperplane H such that the distance between objects with different class labels, the so-called margin, is maximized.⁴⁰ This hyperplane is defined by a weight vector w and a bias b so that $H = \{x | \langle w, x \rangle + b = 0\}$. For better model generalization, slack variables are added to permit errors for training instances falling within the margin or on the incorrect side of the hyperplane. To balance training errors and margin size, the cost or regularization hyperparameter C is introduced, resulting in a primal optimization problem that can be expressed using Lagrangian multipliers λ_i . Dual expression makes it possible to compute the normal vector of the hyperplane as $w = \sum \lambda_i y_i x_i$. Lagrangian multipliers can be nonzero only for training examples that fall onto the margin of the hyperplane or are misclassified. This subset of training examples with nonzero coefficients falling onto the margin represents support vectors. Test data are projected into the feature space and classified depending on the side of the plane onto which they fall, i.e., $f(x) = \text{sgn}(\sum \lambda_i y_i \langle x_i, x \rangle + b)$ or ranked using the real value, i.e., $g(x) = \sum \lambda_i y_i \langle x_i, x \rangle + b$.⁴¹

The dual formulation also enables the use of the “kernel trick”, which is of critical relevance for SVM modeling. If linear separation of training classes in a given feature space is not possible, the scalar product $\langle \cdot, \cdot \rangle$ is replaced by a kernel function $K(\cdot, \cdot)$, which projects the data into a higher dimensional space where linear separation might be possible. One of the most popular kernel function for fingerprint representations is the Tanimoto kernel⁴² used herein:

$$K(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle}$$

The hyper-parameter C was optimized using 10-fold cross-validation on training data with candidate values of 0.01, 0.1, 1, 10, 100, and 1000. SVM calculations and data analysis protocols were implemented in Python using Scikit-learn.⁴³

Random Forest. RF is an ensemble recursive partitioning method where each decision tree is built from a bootstrapped⁴⁴ sample of training compounds. A random subset of features is considered during node splitting for the construction of trees.⁴⁵ The number of trees was set to 100 and class weights were applied. The minimum number of samples required to reach a leaf node (`min_samples_leaf`) was optimized via 10-fold cross-validation. Candidate values for `min_samples_leaf` included 1, 5, 10, 50, 100, 200, and 500. In addition, the number of features for identifying the best data split (`max_features`) was set to square root of the total number of features. RF calculations were carried out using Scikit-learn.⁴³

Deep Neural Network. A deep feed-forward neural network models a given function f . For a classification task, $y = f(x)$ maps an input x to a category y . A feed-forward NN defines a mapping $y = f(x; \theta)$ and learns the value of the parameters θ that result in the best function approximation.⁴⁶ DNNs consist of multiple layers of “neurons”.⁴⁷ Each neuron accepts an n -dimensional input x and generates a linearly transformed m -dimensional vector $y = W^T x + b$, where W and b are parameters of dimension (m, n) and m , respectively. A nonlinear activation function $h(y)$ is then applied to the output of a neuron to approximate nonlinear functions. Passing the output of one neuron to another neuron creates a new layer. We generated an input layer with 256 output dimensions, two hidden layers with 256 output dimensions each, and a final output layer yielding a single number. As an activation function, the rectified linear unit was used defined as $h(y) = \max(0, y)$ except for the final output layer for which the sigmoid activation function was used defined as $f(y) = \frac{1}{1 + e^{-y}}$. This DNN architecture was implemented using PyTorch 0.4.1.⁴⁸ Network parameters were trained applying the Adam optimizer and cross-entropy loss with a learning rate of 10^{-5} until convergence was reached, typically requiring 22 epochs.

Performance Measures. To assess model performance, three different measures were applied including the area under the receiver operating characteristic (ROC) curve (AUC),⁴⁹ Matthew’s correlation coefficient (MCC),⁵⁰ and balanced accuracy (BA).⁵¹ MCC and BA are defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

$$\text{BA} = \frac{0.5\text{TP}}{\text{TP} + \text{FN}} + \frac{0.5\text{TN}}{\text{TN} + \text{FP}}$$

where the abbreviations are as follows: TP, true positives; TN, true negatives; FP, false positives; FN, false negatives.

For AUC calculations, class label predictions were transformed into rankings of test compounds in the order of decreasing probability of promiscuity.

Feature Weighting. For SVM, a feature weighting method⁵² was applied to assign different weights to individual fingerprint features for predictions corresponding to coefficients of primal optimizations. For the nonlinear Tanimoto kernel, feature weights cannot be calculated directly because an explicit mapping to high-dimensional space is not available. However, the Tanimoto kernel can be expressed as the sum of feature contributions by using a normalization factor to obtain a constant denominator for each individual support vector.⁵² If $fc(\mathbf{x}, d)$ is the contribution of feature d to an individual SVM prediction, the normalization factor is determined by the following equation:

$$fc_{\text{Tanimoto}}(\mathbf{x}, d) = \sum_{\text{support vectors}} \frac{y_i \lambda_i x_{id} x_d}{\langle \mathbf{x}_p, \mathbf{x}_i \rangle + \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}_p, \mathbf{x} \rangle}$$

where \mathbf{x} is the test instance, \mathbf{x}_i is the support vector, and y_i and λ_i are support vector coefficients for the dual solution.

Feature Mapping. Features were ranked according to their preferential occurrence in either PROM_PAINS or DCM_PAINS, and the top 30 features from each ranking were transformed into SMARTS strings to search for PAINS substructures using the KNIME implementation⁵³ of the RDKit substructure filter. For each successful match, the list of atom indices was retained for mapping of features to test compounds.

Mapped features were assigned to three *structural context categories* depending on their structural embedding in test compounds. These categories included features (i) representing a subset of a PAINS substructure (*subset*), (ii) overlapping with a PAINS substructure and the remaining structure of a compound (*intersection*), and (iii) mapping to a region in a compound outside the PAINS substructure (*distinct*).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: [10.1021/acs.jmedchem.8b01404](https://doi.org/10.1021/acs.jmedchem.8b01404).

PAINS substructures shared by PROM_PAINS and DCM_PAINS and compounds representing these substructures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-73-69100. Fax: +49-228-73-69101. E-mail: bajorath@bit.uni-bonn.de.

ORCID

Jürgen Bajorath: [0000-0002-0557-5714](https://orcid.org/0000-0002-0557-5714)

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The use of OpenEye's toolkit was made possible by their free academic licensing program. We thank Michael Gütschow for helpful discussions.

■ ABBREVIATIONS USED

AUC, area under the curve; BA, balanced accuracy; DCM, dark chemical matter; MCC, Matthew's correlation coefficient; PAINS, pan-assay interference compounds; PROM, promiscuous; ROC, receiver operating characteristic

■ REFERENCES

- (1) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.
- (2) Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
- (3) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076–7087.
- (4) Ingles, J.; Johnson, R. L.; Simeonov, A.; Xia, M.; Zheng, W.; Austin, C. P.; Auld, D. S. High-Throughput Screening Assays for the Identification of Chemical Probes. *Nat. Chem. Biol.* **2007**, *3*, 466–479.
- (5) Thorne, N.; Auld, D. S.; Ingles, J. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Curr. Opin. Chem. Biol.* **2010**, *14*, 315–324.
- (6) 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.
- (7) Baell, J. B.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.
- (8) Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *J. Med. Chem.* **2015**, *58*, 2091–2113.
- (9) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 36–44.
- (10) Baell, J. B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2016**, *79*, 616–628.
- (11) Nelson, K. M.; Dahlin, J. L.; Bisson, J.; Graham, J.; Pauli, G. F.; Walters, M. A. The Essential Medicinal Chemistry of Curcumin. *J. Med. Chem.* **2017**, *60*, 1620–1637.
- (12) Gilbert, E.; Jasial, S.; Stumpfe, D.; Dimova, D.; Bajorath, J. Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but also Candidates for Polypharmacology. *J. Med. Chem.* **2016**, *59*, 10285–10290.
- (13) Mendgen, T.; Steuer, C.; Klein, C. D. Privileged Scaffolds or Promiscuous Binders: A Comparative Study on Rhodanines and Related Heterocycles in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 743–753.
- (14) Aldrich, C.; Bertozzi, C.; Georg, G. I.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz, K. M., Jr.; Schepartz, A.; Wang, S. The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 387–390.
- (15) 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*, 3879–3886.
- (16) Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 417–427.
- (17) Jasial, S.; Bajorath, J. Dark Chemical Matter in Public Screening Assays and Derivation of Target Hypotheses. *MedChemComm* **2017**, *8*, 2100–2104.

- (18) Gilberg, E.; Gütschow, M.; Bajorath, J. X-ray Structures of Target–Ligand Complexes Containing Compounds with Assay Interference Potential. *J. Med. Chem.* **2018**, *61*, 1276–1284.
- (19) Siramshetty, V. B.; Preissner, R.; Gohlke, B. Exploring Activity Profiles of PAINS and Their Structural Context in Target–Ligand Complexes. *J. Chem. Inf. Model.* **2018**, *58*, 1847–1857.
- (20) Gilberg, E.; Stumpfe, D.; Bajorath, J. Towards a Systematic Assessment of Assay Interference: Identification of Extensively Tested Compounds with High Assay Promiscuity. *FI1000Research* **2017**, *6*, e1505.
- (21) Gilberg, E.; Stumpfe, D.; Bajorath, J. Activity Profiles of Analog Series Containing Pan Assay Interference Compounds. *RSC Adv.* **2017**, *7*, 35638–35647.
- (22) Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol.* **2015**, *11*, 958–966.
- (23) Ferroud, C.; Rool, P.; Santamaria, J. Singlet Oxygen Mediated Alkaloid Tertiary Amines Oxidation by Single Electron Transfer. *Tetrahedron Lett.* **1998**, *39*, 9423–9426.
- (24) Carlson, E. E.; May, J. F.; Kiessling, L. L. Chemical Probes of UDP-Galactopyranose Mutase. *Chem. Biol.* **2006**, *13*, 825–837.
- (25) Metz, J. T.; Huth, J. R.; Hajduk, P. J. Enhancement of Chemical Rules for Predicting Compound Reactivity towards Protein Thiol Groups. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 139–144.
- (26) Voss, M. E.; Carter, P. H.; Tebben, A. J.; Scherle, P. A.; Brown, G. D.; Thompson, L. A.; Xu, M.; Lo, Y. C.; Yang, G.; Liu, R.-Q.; Strzemieniski, P.; Everlof, J. G.; Trzaskos, J. M.; Decicco, C. P. Both 5-Arylidene-2-Thioxodihydropyrimidine-4,6(1H,5H)-Diones and 3-Thioxo-2,3-Dihydro-1H-Imidazo[1,5-a]Indol-1-Ones Are Light-Dependent Tumor Necrosis Factor- α Antagonists. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 533–538.
- (27) Tomasić, T.; Masic, L. P. Rhodanine as a Privileged Scaffold in Drug Discovery. *Curr. Med. Chem.* **2009**, *16*, 1596–1629.
- (28) McCallum, M. M.; Nandhikonda, P.; Temmer, J. J.; Eyermann, C.; Simeonov, A.; Jadhav, A.; Yasgar, A.; Maloney, D.; Arnold, A. L. High-Throughput Identification of Promiscuous Inhibitors from Screening Libraries with the Use of a Thiol-Containing Fluorescent Probe. *J. Biomol. Screening* **2013**, *18*, 705–713.
- (29) Soares, K. M.; Blackmon, N.; Shun, T. Y.; Shinde, S. N.; Takyi, H. K.; Wipf, P.; Lazo, J. S.; Johnston, P. A. Profiling the NIH Small Molecule Repository for Compounds That Generate H₂O₂ by Redox Cycling in Reducing Environments. *Assay Drug Dev. Technol.* **2010**, *8*, 152–174.
- (30) Distinguishing between Pan Assay Interference Compounds (PAINS) That Are Promiscuous or Represent Dark Chemical Matter—Data Set and Prediction Models. <https://doi.org/10.5281/zenodo.1453913>.
- (31) Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS One* **2016**, *11*, e0153873.
- (32) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.
- (33) RDKit: Cheminformatics and Machine Learning Software, 2013. <http://www.rdkit.org> (accessed July 01, 2018).
- (34) Sterling, T.; Irwin, J. J. ZINC 15—Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- (35) 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.* **2012**, *40*, D1100–D1107.
- (36) James, C. A.; Weininger, D.; Delany, J. *SMARTS Theory. Daylight Theory Manual*; Daylight Chemical Information Systems: Laguna Niguel, CA, 2000.
- (37) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (38) *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.
- (39) *OEChem. TK*, version 2.0.0; OpenEye Scientific Software: Santa Fe, NM.
- (40) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
- (41) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.
- (42) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18*, 1093–1110.
- (43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
- (44) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
- (45) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (46) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning* (e-book); MIT Press: Cambridge, MA, 2016.
- (47) Nielsen, M. A. *Neural Networks and Deep Learning*; Determination Press, 2015.
- (48) Ketkar, N. Introduction to PyTorch. *Deep Learning with Python*; Apress: Berkeley, CA, 2017; pp 195–208.
- (49) Bradley, A. P. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recogn.* **1997**, *30*, 1145–1159.
- (50) Matthews, B. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct.* **1975**, *405*, 442–451.
- (51) Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. *2010 20th International Conference on Pattern Recognition*; IEEE, 2010; DOI: [10.1109/ICPR.2010.764](https://doi.org/10.1109/ICPR.2010.764).
- (52) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **2015**, *55*, 1136–1147.
- (53) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Preisach, C., Burkhart, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin, Germany, 2008; pp 319–326.

Summary

In this study, we generated machine learning models to distinguish PAINS that were highly promiscuous from others that were consistently inactive across hundreds of assays. Surprisingly high prediction accuracy was achieved using different models as shown by the performance measures. Also, model performance was found to be similar using different machine learning methods. There was no significant advantage in using deep learning models over SVM and RF for the prediction of PAINS. To rule out any statistical bias in the predictions due to data imbalance, balanced data sets were generated that exclusively consisted of shared PAINS substructures between promiscuous PAINS and DCM PAINS sets. These data sets provided challenging conditions for predictions. However, the machine learning models still performed well indicating that machine learning was able to identify structural patterns that distinguished these two PAINS phenotypes. Although it is difficult to interpret predictions due to black box character of machine learning models, feature weighting was applied to rank features responsible for the predictions of test compounds using SVM. Highly weighted positive features, which were obtained on the basis of weight sums over all predictions, could be associated with reactive moieties. In exemplary cases, mapping and categorization of features helped to explain the influence of structural contexts on predictions.

My major contributions to this study included the generation of promiscuous and DCM PAINS data sets, derivation of SVM and RF models for classification, and application of feature weighting methodology to test compounds. Taken together, machine learning models presented herein further extend the capacity of PAINS filters as they also take structural context information into account.

Conclusion

In this thesis, computational data mining studies are presented to better understand the nature of compound data growth and to explore target binding characteristics of different types of compounds present in screening data. Activity profiles of screening compounds have been analyzed in detail to study multitarget activities and assay interference in context of polypharmacology. Furthermore, a machine learning approach has been introduced for the classification of assay interference compounds. Major results are summarized in this chapter and conclusions are drawn.

The first representative study (*Chapter 2*) provided an analysis of the growth of bioactive compounds over time in relation to scaffold growth and diversity. The goal of the analysis was to understand possible reasons behind the increase in number of active compounds against five major target families. The new active compounds were found to be structurally diverse and new scaffolds were topologically diverse as shown by scaffold-to-compound ratios and CSK-to-scaffold ratios monitored on a per target basis for different target families. Thus, compound growth was associated with chemical diversity and targets were able to recognize structurally distinct compounds, which provided a rationale for the rapid increase of compounds. Since these structurally diverse active compounds for a particular target represent scaffold hops, scaffold hopping projects will be of high interest in future as it is likely that current targets will continue to interact with structurally diverse compounds.

Next studies focused on the analysis of screening data. The goal of the second study (*Chapter 3*) was to determine the promiscuity of extensively tested screening compounds. In light of data incompleteness or sparseness, it is often

assumed that data mining underestimates true compound promiscuity. Therefore, to address this concern, most extensively assayed compounds were identified and promiscuity analysis was performed through the inclusion of assay frequency information from PubChem database. Assay promiscuity was distinguished from target promiscuity to check for discrepancies in order to identify assay bias or false negatives. The promiscuity estimates from extensively tested compounds were found to be low with median value of 2.0 for both primary and confirmatory assays. The degree of promiscuity remained constant irrespective of the number of assays in which the compounds were tested. This provided further evidence that bioactive compounds had moderate or low degrees of promiscuity compared to drugs. These results gave rise to the idea of a “promiscuity enrichment model” in pharmaceutical research according to which promiscuous compounds might be preferentially selected during clinical evaluation as potential drug candidates, provided they have no unwanted side effects.

The next study was based on the analysis of DCM compounds present in extensively assayed compounds (*Chapter 4*). DCM compounds are known to be highly selective and can provide potential hits when tested further under right assay conditions and against appropriate target. Most of the compounds in screening decks remain consistently inactive. They can be utilized to find interesting leads that are less promiscuous and less prone to assay artifacts. A search was carried out in ChEMBL for bioactive analogs of DCM compounds on the basis of high-confidence activity data. More than 8000 DCM compounds were found to occur in a variety of analog series with bioactive compounds making it possible to derive target hypotheses for these compounds. Some of the DCM compounds were found to be analogs of ChEMBL compounds having target annotations against well-known pharmaceutical targets. Given the high degree of similarity between structural analogs, DCM compounds should be tested against these targets. Therefore, analog series generated in the study provide good starting points for lead discovery and have been made freely available publicly in an open access deposition.

In another analysis, activity profiles of PAINS present in extensively assayed screening compounds were determined and evaluated (*Chapter 5*). The key

question was whether assay interference characteristics of PAINS were prevalent in biological screening assays. A subset of 23,000 compounds were identified as PAINS consisting of 270 PAINS substructures. Large-scale analysis of PAINS revealed surprising observations because many PAINS, which are assumed to be frequent hitters, were found to be consistently inactive. Furthermore, the overall hit frequency of PAINS was low. Only few PAINS produced an abundance of hits. Thus, PAINS showed diverse activity profiles and were far from being excessively active. Most importantly, many consistently inactive and highly active compounds contained same PAINS substructure indicating that the occurrence of PAINS substructure is not solely responsible for frequent hitter characteristics. Instead, the structural environment in which a PAINS substructure is embedded in a molecule is important for determining its activity or inactivity. Therefore, PAINS filters do provide initial alerts for flagging compounds prone to assay interference but they should be tested further for their activity in orthogonal assays before removing any potential liable compound.

A major drawback of PAINS filters is that they do not take structural context information into account. Also, given the large number of PAINS substructures and variety of structural environments in which they are embedded, it is not possible to derive rules for classifying frequently active and inactive PAINS based on expert knowledge. Therefore, in the last study (*Chapter 6*), we attempted to address these issues by applying machine learning to distinguish promiscuous PAINS from inactive ones. SVM, RF and DNN models were built using promiscuous and DCM PAINS data sets represented as MACCS and ECFP4 fingerprints. Surprisingly accurate models were obtained with area under the curve (AUC) and Matthew’s correlation coefficient (MCC) values reaching around 0.95 and 0.80, respectively. The models were also built for balanced data sets that exclusively contained shared PAINS substructures and same number of promiscuous and DCM compounds per substructure in order to avoid any potential statistical bias. The performance of these models reached AUC and MCC of 0.90 and 0.68, respectively. Hence, the models performed well on the balanced data sets indicating that machine learning was able to identify and exploit structural patterns distinguishing between promiscuous and DCM PAINS. These models further extend the capacity of PAINS

filters by considering structural context information. However, due to the black box character of models, it was difficult to interpret the predictions. For that purpose, a feature weighting scheme for SVM model was applied in order to find positive and negative features for an individual prediction. The features were ranked according to their weights and highly ranked features were used to study structural context of PAINS in detail. However, not all successful and unsuccessful predictions could be rationalized as final predictions depended on the combination of features existing within a compound.

In conclusion, data mining and machine learning studies discussed above provide insights that are useful for drug discovery. It will be promising to develop new approaches in future that are able to identify structurally distinct compounds for a particular target. Promiscuity analysis has implications for polypharmacology. Analog series of DCM and bioactive compounds can be used to find potential targets for DCM and further tests should be performed for DCM against these targets. The analysis of PAINS in screening data puts emphasis on the careful use of PAINS filters and structural context dependence of PAINS activity. Lastly, machine learning provides a way to improve PAINS filters as the models consider structural environments. These results pave the way for further exploration of combinations of structural features and patterns in PAINS that determine their activity or inactivity.

Bibliography

- [1] Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nature Reviews Drug Discovery* **2011**, *10*, 188–195.
- [2] Lusher, S. J.; McGuire, R.; van Schaik, R. C.; Nicholson, C. D.; de Vlieg, J. Data-Driven Medicinal Chemistry in the Era of Big Data. *Drug Discovery Today* **2014**, *19*, 859–868.
- [3] Hu, Y.; Bajorath, J. Learning from 'Big Data': Compounds and Targets. *Drug Discovery Today* **2014**, *19*, 357–360.
- [4] Al-Lazikani, B.; Workman, P. Minimizing Bias in Target Selection by Exploiting Multidisciplinary Big Data and the Protein Interactome. *Future Medicinal Chemistry* **2016**, *8*, 1711–1716.
- [5] Hu, Y.; Bajorath, J. Entering the 'Big Data' Era in Medicinal Chemistry: Molecular Promiscuity Analysis Revisited. *Future Science OA* **2017**, *3*, FSO179.
- [6] Bajorath, J.; Overington, J.; Jenkins, J. L.; Walters, P. Drug Discovery and Development in the Era of Big Data. *Future Medicinal Chemistry* **2016**, *8*, 1807–1813.
- [7] Hu, Y.; Bajorath, J. How Promiscuous are Pharmaceutically Relevant Compounds? A Data-Driven Assessment. *The AAPS Journal* **2013**, *15*, 104–111.

- [8] 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 Research* **2012**, *40*, D1100–D1107.
- [9] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Research* **2017**, *45*, D945–D954.
- [10] Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: A Public Information System for Analyzing Bioactivities of Small Molecules. *Nucleic Acids Research* **2009**, *37*, W623–W633.
- [11] Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 6987–7002.
- [12] Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 update. *Nucleic Acids Research* **2017**, *45*, D955–D963.
- [13] Irwin, J. J.; Shoichet, B. K. ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 177–182.
- [14] Sterling, T.; Irwin, J. J. ZINC 15—Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.
- [15] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration. *Nucleic Acids Research* **2006**, *34*, D668–D672.
- [16] Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0:

- A Major Update to the DrugBank Database for 2018. *Nucleic Acids Research* **2018**, *46*, D1074–D1082.
- [17] Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome Through Polypharmacology. *Nature Reviews Cancer* **2010**, *10*, 130–137.
- [18] Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nature Biotechnology* **2006**, *24*, 805–815.
- [19] Hopkins, A. L. Network Pharmacology: The Next Paradigm in Drug Discovery. *Nature Chemical Biology* **2008**, *4*, 682–690.
- [20] Boran, A. D. W.; Iyengar, R. Systems Approaches to Polypharmacology and Drug Discovery. *Current Opinion in Drug Discovery and Development* **2010**, *13*, 297–309.
- [21] Jalencas, X.; Mestres, J. On the Origins of Drug Polypharmacology. *Medicinal Chemistry Communications* **2013**, *4*, 80–87.
- [22] Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *Journal of Medicinal Chemistry* **2014**, *57*, 7874–7887.
- [23] 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*, e40262.
- [24] Hu, Y.; Bajorath, J. Compound Promiscuity: What can we learn from Current Data? *Drug Discovery Today* **2013**, *18*, 644–650.
- [25] Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity. *F1000Research* **2013**, *2*.
- [26] Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *Journal of Chemical Information and Modeling* **2014**, *54*, 3056–3066.
- [27] Hu, Y.; Bajorath, J. Promiscuity Profiles of Bioactive Compounds: Potency Range and Difference Distributions and the Relation to Target

- Numbers and Families. *Medicinal Chemistry Communications* **2013**, *4*, 1196–1201.
- [28] Hu, Y.; Bajorath, J. Monitoring Drug Promiscuity over Time. *F1000Research* **2014**, *3*, 218.
- [29] Hu, Y.; Jasial, S.; Bajorath, J. Promiscuity Progression of Bioactive Compounds Over Time. *F1000Research* **2015**, *4*, 118.
- [30] Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. Data Completeness—The Achilles Heel of Drug-Target Networks. *Nature Biotechnology* **2008**, *26*, 983–984.
- [31] Diller, D. J.; Hobbs, D. W. Deriving Knowledge through Data Mining High-Throughput Screening Data. *Journal of Medicinal Chemistry* **2004**, *47*, 6373–6383.
- [32] Aldrich, C.; Bertozzi, C.; Georg, G. I.; Kiessling, L.; Lindsley, C.; Liotta, D.; Merz, K. M.; Schepartz, A.; Wang, S. The Ecstasy and Agony of Assay Interference Compounds. *Journal of Medicinal Chemistry* **2017**, *60*, 2165–2168.
- [33] Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chemical Biology* **2018**, *13*, 36–44.
- [34] McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *Journal of Medicinal Chemistry* **2002**, *45*, 1712–1722.
- [35] Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
- [36] Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *Journal of Medicinal Chemistry* **2015**, *58*, 7076–7087.

- [37] Feng, B. Y.; Simeonov, A.; Jadhav, A.; Babaoglu, K.; Inglese, J.; Shoichet, B. K.; Austin, C. P. A High-Throughput Screen for Aggregation-Based Inhibition in a Large Compound Library. *Journal of Medicinal Chemistry* **2007**, *50*, 2385–2390.
- [38] 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. *Journal of Medicinal Chemistry* **2010**, *53*, 2719–2740.
- [39] Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.
- [40] Bisson, J.; McAlpine, J. B.; Friesen, J. B.; Chen, S.-N.; Graham, J.; Pauli, G. F. Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery? *Journal of Medicinal Chemistry* **2015**, *59*, 1671–1690.
- [41] Baell, J. B. Feeling Nature’s PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *Journal of Natural Products* **2016**, *79*, 616–628.
- [42] Dahlin, J. L.; Nissink, J. W. M.; Strasser, J. M.; Francis, S.; Higgins, L.; Zhou, H.; Zhang, Z.; Walters, M. A. PAINS in the Assay: Chemical Mechanisms of Assay Interference and Promiscuous Enzymatic Inhibition Observed during a Sulfhydryl-Scavenging HTS. *Journal of Medicinal Chemistry* **2015**, *58*, 2091–2113.
- [43] Rishton, G. M. Nonleadlikeness and Leadlikeness in Biochemical Screening. *Drug Discovery Today* **2003**, *8*, 86–96.
- [44] Thorne, N.; Auld, D. S.; Inglese, J. Apparent Activity in High-Throughput Screening: Origins of Compound-Dependent Assay Interference. *Current Opinion in Chemical Biology* **2010**, *14*, 315–324.
- [45] Ingólfsson, H. I.; Thakur, P.; Herold, K. F.; Hobart, E. A.; Ramsey, N. B.; Periole, X.; de Jong, D. H.; Zwama, M.; Yilmaz, D.; Hall, K.; Marezky, T.; Hemmings, H. C.; Blobel, C.; Marrink, S. J.; Koçer, A.; Sack, J. T.; Andersen, O. S. Phytochemicals Perturb Membranes and

- Promiscuously Alter Protein Function. *ACS Chemical Biology* **2014**, *9*, 1788–1798.
- [46] Nelson, K. M.; Dahlin, J. L.; Bisson, J.; Graham, J.; Pauli, G. F.; Walters, M. A. The Essential Medicinal Chemistry of Curcumin. *Journal of Medicinal Chemistry* **2017**, *60*, 1620–1637.
- [47] Baell, J. B.; Ferrins, L.; Falk, H.; Nikolakopoulos, G. PAINS: Relevance to Tool Compound Discovery and Fragment-Based Screening. *Australian Journal of Chemistry* **2014**, *66*, 1483–1494.
- [48] Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *Journal of Chemical Information and Modeling* **2017**, *57*, 2640–2645.
- [49] Capuzzi, S. J.; Muratov, E. N.; Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS. *Journal of Chemical Information and Modeling* **2017**, *57*, 417–427.
- [50] Chai, C. L.; Mátyus, P. One Size Does Not Fit All: Challenging Some Dogmas and Taboos in Drug Discovery. *Future Medicinal Chemistry* **2016**, *8*, 29–38.
- [51] Senger, M. R.; Fraga, C. A. M.; Dantas, R. F.; Silva, F. P. Filtering Promiscuous Compounds in Early Drug Discovery: Is it a Good Idea? *Drug Discovery Today* **2016**, *21*, 868–872.
- [52] Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *Journal of Chemical Information and Modeling* **2006**, *46*, 512–524.
- [53] Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and Lead-likeness, Target Class, and Molecular Diversity Analysis of 7.9 Million Commercially Available Organic Compounds Provided by 29 Suppliers. *Journal of Chemical Information and Modeling* **2010**, *50*, 470–479.
- [54] Hajduk, P. J.; Galloway, W. R. J. D.; Spring, D. R. Drug Discovery: A Question of Library Design. *Nature* **2011**, *470*, 42–43.

- [55] Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J.; Selzer, P.; Glick, M. Biodiversity of Small Molecules—A New Perspective in Screening Set Selection. *Drug Discovery Today* **2013**, *18*, 674–680.
- [56] Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138–1139.
- [57] Wassermann, A. M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F. J.; Studer, C.; Peltier, J. M.; Grippo, M. L.; Prindle, V.; Tao, J.; Schuffenhauer, A.; Wallace, I. M.; Chen, S.; Krastel, P.; Cobos-Correa, A.; Parker, C. N.; Davies, J. W.; Glick, M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nature Chemical Biology* **2015**, *11*, 958–966.
- [58] Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.
- [59] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Modeling* **1989**, *29*, 97–101.
- [60] Weininger, D. SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures. *Journal of Chemical Information and Modeling* **1990**, *30*, 237–243.
- [61] James, C. A.; Weininger, D.; Delany, J. *SMARTS Theory. Daylight Theory Manual*; Daylight Chemical Information Systems: Laguna Niguel, CA, 2000.
- [62] Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of Descriptors from Molecular Structures. *Journal of Chemical Information and Modeling* **1976**, *16*, 105–110.
- [63] Glen, R. C.; Rose, V. S. Computer Program Suite for the Calculation, Storage and Manipulation of Molecular Property and Activity Descriptors. *Journal of Molecular Graphics* **1987**, *5*, 79–86.

- [64] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH, 2000.
- [65] Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nature Reviews Drug Discovery* **2002**, *1*, 882–894.
- [66] Ewing, T.; Baber, J. C.; Feher, M. Novel 2D Fingerprints for Ligand-Based Virtual Screening. *Journal of Chemical Information and Modeling* **2006**, *46*, 2423–2431.
- [67] Williams, C. Reverse Fingerprinting, Similarity Searching by Group Fusion and Fingerprint Bit Importance. *Molecular Diversity* **2006**, *10*, 311–332.
- [68] Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 260–282.
- [69] *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.
- [70] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.
- [71] Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Molecular Informatics* **2011**, *30*, 646–664.
- [72] Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *Journal of Chemical Information and Modeling* **2011**, *51*, 1742–1753.
- [73] Hu, Y.; Stumpfe, D.; Bajorath, J. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2016**, *59*, 4062–4076.
- [74] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.
- [75] Xu, Y.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 181–185.

- [76] Kayastha, S.; Dimova, D.; Stumpfe, D.; Bajorath, J. Structural Diversity and Potency Range Distribution of Scaffolds from Compounds Active Against Current Pharmaceutical Targets. *Future Medicinal Chemistry* **2015**, *7*, 111–122.
- [77] Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. L.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfield, J. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *Journal of Medicinal Chemistry* **1988**, *31*, 2235–2246.
- [78] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. “Scaffold-Hopping” by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angewandte Chemie International Edition* **1999**, *38*, 2894–2896.
- [79] Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *Journal of Chemical Information and Modeling* **2010**, *50*, 205–216.
- [80] Eckert, H.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening: Foundations, Limitations and Novel Approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
- [81] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2013**, *57*, 3186–3204.
- [82] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983–996.
- [83] Jaccard, P. Nouvelles Recherches Sur La Distribution Florale. *Bull Soc Vaudoise Des Sci Nat* **1908**, *44*, 223–270.

- [84] Vogt, M.; Bajorath, J. Introduction of the Conditional Correlated Bernoulli Model of Similarity Value Distributions and its Application to the Prospective Prediction of Fingerprint Search Performance. *Journal of Chemical Information and Modeling* **2011**, *51*, 2496–2506.
- [85] Kenny, P. W.; Sadowski, J. *Methods and Principles in Medicinal Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA, 2005; pp 271–285.
- [86] Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *Journal of Medicinal Chemistry* **2011**, *54*, 7739–7750.
- [87] Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *Journal of Medicinal Chemistry* **2011**, *54*, 2944–2951.
- [88] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *Journal of Chemical Information and Modeling* **1998**, *38*, 511–522.
- [89] de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Medicinal Chemistry Communications* **2014**, *5*, 64–67.
- [90] Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for the Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *Journal of Medicinal Chemistry* **2016**, *59*, 7667–7676.
- [91] Duda, R.; Hart, P.; Stork, D. *Pattern classification*; 2nd ed.; Wiley-Interscience: New York, 2000.
- [92] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- [93] Nielsen, M. A. *Neural Networks and Deep Learning*; Determination Press, 2015.

- [94] Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer New York, 1995.
- [95] Heikamp, K.; Bajorath, J. Support Vector Machines for Drug Discovery. *Expert Opinion on Drug Discovery* **2013**, *9*, 93–104.
- [96] Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *Journal of Chemical Information and Modeling* **2009**, *49*, 767–779.
- [97] Heikamp, K.; Bajorath, J. Prediction of Compounds with Closely Related Activity Profiles Using Weighted Support Vector Machine Linear Combinations. *Journal of Chemical Information and Modeling* **2013**, *53*, 791–801.
- [98] Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-Directed Similarity Searching Using Support Vector Machines. *Chemical Biology and Drug Design* **2010**, *77*, 30–38.
- [99] Cortes, C.; Vapnik, V. Support Vector Networks. *Machine Learning* **1995**, *20*, 273–297.
- [100] Burges, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* **1998**, *2*, 121–167.
- [101] Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *Journal of Chemical Information and Modeling* **2005**, *45*, 549–561.
- [102] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of The Fifth Annual Workshop on Computational Learning Theory* **1992**, 144–152.
- [103] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093–1110.
- [104] Gärtner, T.; Flach, P.; Wrobel, S. *Learning Theory and Kernel Machines*; Springer Berlin Heidelberg, 2003; pp 129–143.

- [105] Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 468–481.
- [106] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.
- [107] Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26*, 1169–1175.
- [108] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.
- [109] Lee, H.; Grosse, R.; Ranganath, R.; Ng, A. Y. Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. *Communications of the ACM* **2011**, *54*, 95–103.

Additional Publications

Hu, Y.; Jasial, S.; Bajorath J. Promiscuity Progression of Bioactive Compounds over Time. *F1000Research* **2015**, *4*, 118.

Jasial, S.; Hu, Y.; Vogt, M.; Bajorath J. Activity-Relevant Similarity Values for Fingerprints and Implications for Similarity Searching. *F1000Research* **2016**, *5*, 591.

Gilberg, E.; Jasial, S.; Stumpfe, D.; Dimova, D.; Bajorath J. Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology. *Journal of Medicinal Chemistry* **2016**, *59*, 10285-10290.

Hu, Y.; Jasial, S.; Gilberg, E.; Bajorath J. Structure-Promiscuity Relationship Puzzles-Extensively Assayed Analogs with Large Differences in Target Annotations. *The AAPS Journal* **2017**, *19*, 856-864.

Vogt, M.; Jasial, S.; Bajorath J. Extracting Compound Profiling Matrices from Screening Data. *ACS Omega* **2018**, *3*, 4706-4712.

Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; Vogt, M.; Bajorath J. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3*, 4713-4723.

Vogt, M.; Jasial, S.; Bajorath J. Computationally Derived Compound Profiling Matrices. *Future Science OA* **2018**, *4*, FSO327.