

# Computational Analysis of Assay Interference and Compound Promiscuity in Medicinal Chemistry

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)  
der Mathematisch-Naturwissenschaftlichen Fakultät  
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Apotheker

ERIK GILBERG

aus Bad Honnef

Bonn, 18.02.2019



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

1. Gutachter: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
  2. Gutachter: Univ.-Prof. Dr. rer. nat. Michael Gütschow
- Tag der Promotion: 07.05.2019  
Erscheinungsjahr: 2019





# Abstract

Compound promiscuity offers both opportunities and perils for medicinal chemistry and drug discovery. On the one hand, it is well-established that compounds elicit their therapeutic potential by engaging with multiple proteins, giving rise to polypharmacology. On the other hand, promiscuous small molecules must be treated with caution, as their activity towards many targets is often associated with non-specific binding and assay interference. Thus, an essential step in drug discovery is to confirm true, beneficial, promiscuity by distinguishing it from artificial multitarget activity. This thesis aims to elicit molecular mechanisms of true multitarget activity, to separate them from those of assay interference, and to identify properties of molecules that can be exploited for polypharmacology.

Hit compounds that originate from biological screening assays display a major resource for the exploration of promiscuity. Taking potential chemical liabilities of these compounds into account, 480 substructure patterns of pan-assay interference compounds (PAINS) have been put forward and are frequently used as structural alerts in screening efforts. In this thesis, the utility of extensively assayed promiscuous compounds as a starting point for the study of pharmacology is explored. In addition, limitations of PAINS filters are elaborated on the basis of crystallographic target-PAINS complexes and extensively assayed compounds. Further, series of analogs containing PAINS are generated to draw structure-activity relationships between PAINS displaying different activity profiles. To elucidate the structural context dependence of PAINS activities, structural features that favor correct predictions of PAINS activities by machine learning models are investigated with respect to their chemical interpretability. Moreover, analog series of extensively tested compounds displaying high hit rates are provided, allowing a systematic analysis of assay interference by circumventing shortcomings of PAINS filters.

Uncertainties associated with screening data are avoided by validating the promiscuity of small molecules through their presence in experimentally determined target-ligand complexes. First, ligands are identified that are present in multiple complexes with distantly related or unrelated targets. These multifamily ligands are utilized for the generation of template structures that allow the design of polypharmacology candidates. Finally, chemical properties and binding modes of multifamily ligands are explored, revealing insight into the mechanisms that allow molecular recognition of ligands across distinct biological targets.

# Acknowledgements

First of all, I would like to thank my supervisors Prof. Dr. Jürgen Bajorath and Prof. Dr. Michael Gütschow for their scientific inspiration, continuous support of my scientific and personal development as well as for their invaluable guidance and encouragement throughout my PhD study.

I further thank all my colleagues at the b-it and the Pharmaceutical Institute for providing a helpful and friendly working environment. Special thanks are given to Dr. Norbert Furtmann and Dr. Janina Schmitz for introducing me to medicinal and computational chemistry and for all the good times we shared. Especially, I would like to thank Dr. Dagmar Stumpfe for her continuous support as a scientist and friend.

To Carina Lemke, Martin Mangold, Jim Küppers, Christian Breuer, and Lorenzo Cianni, thank you for the fruitful collaboration on your experimental projects and the friendship that grew out of it. To Dr. Dilyana Dimova, Dr. Andrew Anighoro, Swarit Jasial, Thomas Blaschke, and Christian Feldmann, thank you for being excellent team partners and for all you taught me. Moreover, thanks to Dimitar Yonchev, Thomas Blaschke, and Dr. Dagmar Stumpfe for making everyday at the b-it a joyful experience. I also want to thank Georg Rolshoven and Markus Ries for being great friends throughout my studies.

Finally, I owe my deepest gratitude to my parents, my brother, and Marie for being the greatest inspiration and for their everlasting support and encouragement.



# Contents

1	Introduction	1
1.1	Drug Discovery . . . . .	1
1.1.1	Assay Interference . . . . .	4
1.1.2	Promiscuity and Polypharmacology . . . . .	6
1.2	Structure-Activity Relationship . . . . .	8
1.2.1	Molecular Similarity . . . . .	10
1.2.2	Molecular Representations . . . . .	11
1.2.3	Matched Molecular Pairs . . . . .	14
1.2.4	Scaffolds . . . . .	16
1.3	Three-Dimensional Target Structures . . . . .	18
1.3.1	The Protein Data Bank . . . . .	19
1.3.2	Target-Ligand Interactions . . . . .	20
1.4	Outline of the Thesis . . . . .	21
2	Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology	23
3	X-ray Structures of Target-Ligand Complexes Containing Compounds with Assay Interference Potential	33
4	Activity Profiles of Analog Series Containing Pan-Assay Interference Compounds	47

5	Machine Learning Distinguishes with High Accuracy Between Pan-Assay Interference Compounds that are Promiscuous or Represent Dark Chemical Matter	61
6	Towards a Systematic Assessment of Assay Interference: Identification of Extensively Tested Compounds with High Assay Promiscuity	77
7	X-ray-Structure-Based Identification of Compounds with Activity Against Targets from Different Families and Generation of Templates for Multi-target Ligand Design	97
8	Promiscuous Ligands from Experimentally Determined Structures, Binding Conformations, and Protein Family Dependent Interaction Hotspots	107
	Bibliography	126

# Acronyms

**AS** analog series

**ASB** analog series-based

**BM** Bemis Murcko

**ECFP** extended connectivity fingerprints

**FP** fingerprint

**HTS** high-throughput screening

**MACCS** molecular ACCess system

**MCS** maximum common substructure

**MMP** matched molecular pair

**MMS** matched molecular series

**PAINS** pan-assay interference compounds

**PDB** Protein Data Bank

**PLIF** protein ligand interaction fingerprint

**QSAR** quantitative SAR

**RECAP** retrosynthetic combinatorial analysis procedure

**REOS** rapid elimination of swill

**RMMP** RECAP-MMP

**RMMS** RECAP-MMS

**SAR** structure-activity relationship

**SMARTS** SMILES arbitrary target specification

**SMILES** simplified molecular input line entry specification

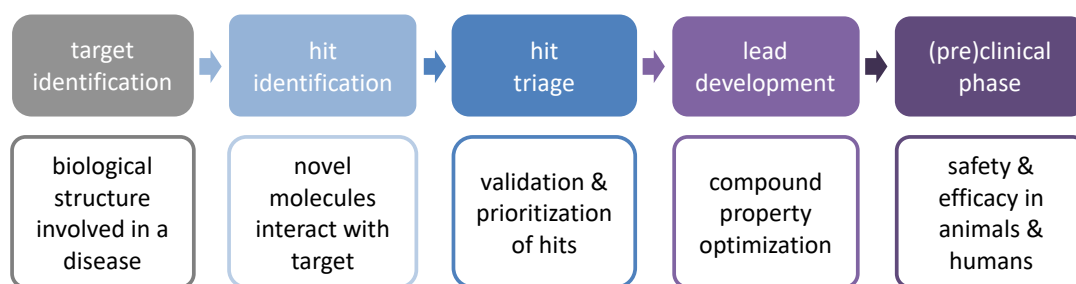


# 1 Introduction

## 1.1 Drug Discovery

Modern pharmaceutical research originated in the 19th century, when advances in organic chemistry led to the first synthetically manufactured drugs.<sup>1</sup> Based on by-products of the textile and dye industry, benzene derivatives were among the first drugs used in analgesia and antipyresis.<sup>1,2</sup> Until then, medicines were limited to substances that were individually isolated from natural products.<sup>3</sup> Originally driven by chemistry, but increasingly influenced by pharmacology and medicine, the first pharmaceutical companies emerged.<sup>1</sup> Today, the pharmaceutical industry has developed into a billion dollar business and the hunt for new therapeutics has become a highly complex and long-lasting challenge.<sup>4,5</sup> While first discoveries of bioactive molecules were made on the basis of serendipity, modern drug discovery focuses on rationalized and thoroughly organized processes covering a wide range of scientific fields (Figure 1).<sup>1,2</sup>

The unmet need for therapies for diseases initiates the development of a new drug.<sup>5</sup> First, physiological and pathophysiological pathways associated with a given clinical condition must be explored. In these pathways, potential molecular targets must be identified whose manipulation through pharmaceutical agents can induce a therapeutic effect.<sup>5,6</sup> These targets comprise biological structures that differ in composition and function, including enzymes, genes, receptors, transporters, ion channels, or RNA.<sup>7</sup> However, potential targets must undergo a validation process and their role in diseases must be clearly defined before drugs are sought to act against it.<sup>8</sup> Identification and validation steps require the interplay of different disciplines, such as chemical proteomics, genetic screening, and patient studies.<sup>9-11</sup>



**Figure 1: Drug discovery process.** A schematic representation of the drug discovery process is shown.

Once suitable experimental assays have been developed, confirmed targets will be interrogated into screening campaigns. Here, the search for molecules that modulate target functions and thus elicit a pharmacological effect begins. These 'hit' compounds interact with biological macromolecules in different ways, including the inhibition of enzymes, activation or deactivation of receptor signaling, and modification of ion channels.<sup>12</sup> For hit-identification, a variety of screening paradigms exist to investigate biochemical and cellular effects of molecules.<sup>5</sup> A widespread method involves high-throughput screening (HTS), in which large compound libraries are automatically evaluated for activities against a given target.<sup>13,14</sup> Apart from HTS, other screening methods have been established that, depending on approach and prior knowledge, utilize DNA encoded compound repositories, reduced sets of molecules with preferential structure classes, or fragment based libraries.<sup>5,14-17</sup> However, hits from initial screening campaigns represent a heterogeneous mixture of different chemotypes and are often associated with chemical liabilities that cause artificial assay readouts. Therefore, hit compounds need to be validated and characterized by a second stage of confirmatory experiments, often referenced as the triaging process.<sup>18</sup> Here, truly competitive behavior of hit candidates is verified by orthogonal testing, dose-response experiments, and X-ray crystallographic target-hit complexes.<sup>17-19</sup> Once the integrity of a hit compound is confirmed, chemical series are prioritized by their synthetic tractability, drug-likeness, and patentability, to be included in hit-to-lead optimization programs.<sup>17</sup>

Within this process, hit series are structurally modified to provide candidates with high potency and selectivity, low toxicity, and suitable bioavailability for eval-

uation in animal studies.<sup>5</sup> This multi-objective procedure usually requires many iterations and is closely linked to the preclinical phase in which pharmacokinetic and pharmacodynamic parameters as well as toxicology information and initial dosage profiles are generated in animal models.

The information obtained in these final steps of drug discovery serves for the design of clinical trials. Divided into three different phases, clinical studies investigate the therapeutic effect of drug candidates in humans versus placebo and/or established therapies.<sup>20</sup> In phases one and two, the safety, dosage, and route of administration is investigated in small to medium numbers of volunteers. Covering a large population of patients in phase three, efficacy and safety of candidates is confirmed by long-term clinical studies.<sup>20,21</sup> Ultimately, pharmaceuticals are granted market access if they meet regulatory requirements.

Drug discovery is a tedious and high-risk process. Research and development for most therapies available today has taken 12 to 24 years and the cost of this process rose sharply to an estimated 2.6 billion dollars per drug.<sup>22,23</sup> However, numerous expensive, long-running research projects fail to produce marketable drugs.<sup>24</sup> Often because hit candidates reveal undesired biological and pharmacological characteristics that were not properly identified in hit identification and confirmation steps.<sup>19,25</sup> Consequently, it is essential to validate and optimize early-stage compounds carefully and thus minimize the risk of failure. To achieve this, knowledge of the molecular basis on which drug candidates elicit desired or undesired actions in biological and physiological contexts is vital.

The systematic use of chemoinformatic approaches complements resource-intensive experimental research in this field.<sup>26-28</sup> It is a central objective of chemoinformatics to explore which structural features influence activity, properties, and the ability of molecules to interact with target structures.<sup>29,30</sup> Therefore, the confirmation of compound integrity and activity profiling of drug candidates naturally benefits from the application of chemoinformatic methods.

### 1.1.1 Assay Interference

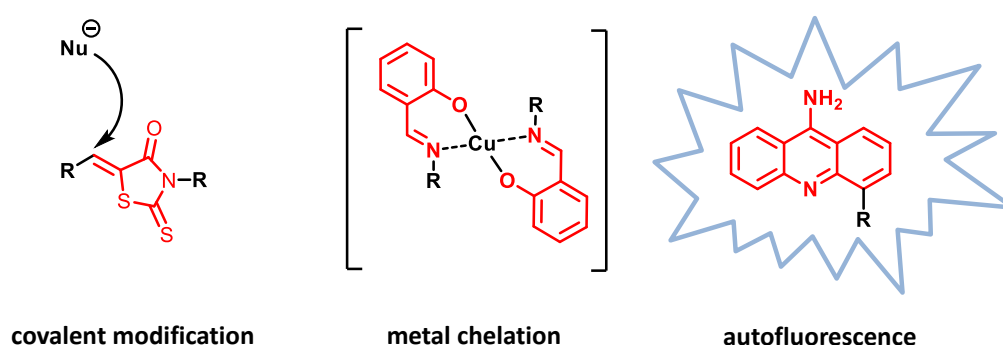
HTS is considered to be a primary source of new molecular structures in industrial and academic drug discovery.<sup>31,32</sup> Notably, screening data is often made publicly available and can be explored to study compound activity profiles on a large scale. For example, the open PubChem BioAssay database provides primary and confirmatory biological screening data.<sup>33</sup> However, the focus on biological screening hits also bears the risk of pursuing peculiar molecules, which turn out to be a nuisance in later stages and ultimately condemn research efforts to fail.<sup>19</sup> Primary HTS assays are typically single-point experiments that provide qualitative results. Therefore, lead development candidates are validated in a confirmatory phase, typically through the collaboration of experts from medicinal chemistry, biology, and chemoinformatics.<sup>17,18</sup>

Assay artifacts are multifactorial and hard to detect and rationalize. Often, however, liable compounds are active against a wide range of targets and display a promiscuous nature. Moreover, they display certain phenotypes in confirmatory experiments that raise suspicion. For instance, frequent hitters exhibit an all or nothing behavior in dose-response experiments and are highly sensitive to assay conditions. Thus, they are most likely interacting with a component of the assay system and not with the target protein.<sup>5,19</sup> Moreover, analogs derived from assay artifacts display flat structure-activity relationship (SAR). This indicates that biological activities are not associated with specific interactions between molecule residues and target structures, but related to the overall biophysical and chemical properties of frequent hitters and assay artifacts.<sup>19</sup>

A variety of computational approaches have been provided to improve the quality of experimental and virtual screening results. Compound libraries are initially refined by selecting structures that display favorable physicochemical properties,<sup>34</sup> e.g. as defined by Lipinski's rule of five for oral drug bioavailability.<sup>35</sup> Nevertheless, drug-likeness does not exclude artificial activities and toxicity of screening compounds.<sup>19</sup> For instance, at micro- to submicromolar concentrations, many drug-like organic molecules undergo colloidal aggregation in assay solutions and incorporate protein targets, thereby inhibiting their function.<sup>19,36</sup> Retro perspective analysis of aggregators led to the introduction of similarity-based computational methods, which assess a screening candidate's potential to aggregate.<sup>37</sup>

The 'Rapid Elimination Of Swill' (REOS) procedure was implemented as one of the first substructural filters that allowed the systematic identification of compound moieties that might elicit undesired reactivities or toxicity.<sup>13</sup> REOS represents a hybrid method that combines the assessment of physicochemical properties with filtering out reactive functional groups, e.g. peroxides and isocyanates.<sup>13</sup>

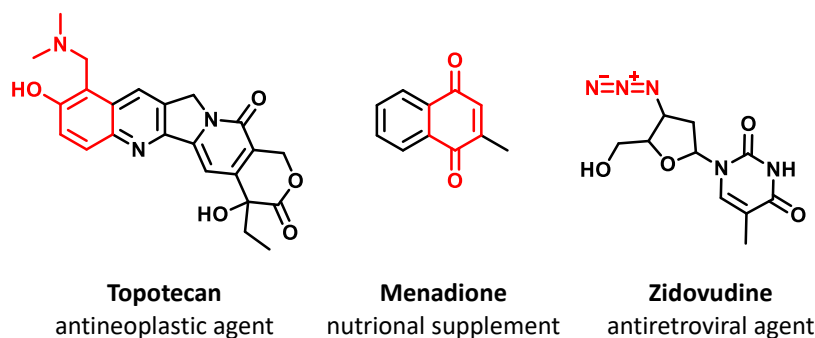
Expanding the concept of substructure filters, 'Pan-Assay INterference Com-pounds' (PAINS) represent 480 structure classes that are prone to show frequent-hitting behavior due to their assay interference potential.<sup>38</sup> PAINS were experimen-tally defined by monitoring activity profiles of approximately 100,000 compounds in six HTS campaigns, utilizing the AlphaScreen technology.<sup>38</sup> In this bead-based assay, a successful interaction between a small molecule and target protein results in the transfer of a singlet oxygen between donor and acceptor bead that causes a chemiluminescent signal.<sup>39</sup> PAINS filters have found wide acceptance in the academic and industrial world<sup>40</sup> and cover a wide range of structural classes with different interference mechanisms (Figure 2).



**Figure 2: Interference mechanisms.** Three exemplary PAINS substructures and their proposed mechanism of actions are shown schematically. The PAINS substructure is colored in red. The Michael-type reaction of a biological nucleophile (Nu) and the exocyclic double bond of a unsaturated rhodanine (left), copper chelation by hydroxyphenylhydrazones by hydroxyphenylhydrazones (middle), and the autofluorescence of aminoacridines (right) are depicted.

For example, unsaturated rhodanines, phenol-sulfonamides, and quinones are prone to non-specific covalent reactions with biological nucleophiles and protein residues.<sup>25,38,41</sup> These Michael-type reactions also occur for Mannich bases, hydroxyphenylhydrazones, and catechols, which can additionally inactivate proteins through complexation and precipitation of metal ions.<sup>38,42,43</sup> Individual mecha-

nisms of action have been identified including toxoflavins that create reactive peroxide species,<sup>44</sup> curcumin derivatives that interfere by membrane perturbation,<sup>45</sup> and aminoacridines, which cause photoinduced interference in fluorometric assays.<sup>46</sup> Many of these structure classes occur in a variety of natural products<sup>47</sup> or approved drugs,<sup>48</sup> and some of them have been considered as privileged scaffolds for drug discovery in the past (Figure 3).<sup>49-51</sup>



**Figure 3: PAINS substructures in natural products and approved drugs.** Exemplary natural products and approved drugs containing PAINS motifs are shown. PAINS substructures are colored in red.

Although structure filters such as PAINS draw attention to problematic compound classes and support the triaging process,<sup>18,48</sup> they are often viewed controversially. The statistical and experimental foundation of the PAINS concept is called into question and varying activity profiles of PAINS in different structural contexts raised awareness of an overestimation of PAINS filters.<sup>52-55</sup> Nevertheless, compound integrity often lies in the eyes of the beholder and successful hit conformation inevitable relies on systematic approaches,<sup>18,56-58</sup> especially, when there is need to distinguish between truly promiscuous compounds and frequent hitters, as discussed in the following chapter.

### 1.1.2 Promiscuity and Polypharmacology

From its beginning, drug discovery was guided by the idea of finding the ‘magic bullet’, a drug that elicits its desired therapeutic effect solely based on the interaction with a specific biological target.<sup>59,60</sup> Accordingly, the administration of a medicine with a broad spectrum of biological activities could lead to un-

predictable and eventually adverse reactions. However, this principle had to be carefully refined when it became apparent that drugs impacting many targets at the same time better control complex disease systems and are less likely to cause drug resistance.<sup>61,62</sup> The design of new pharmaceutical agents that achieve efficacy through their ability to influence multiple targets and disease pathways is referred to as 'polypharmacology'.<sup>61,63-66</sup>

Without neglecting the paradigm of single target exclusivity, polypharmacology provides additional possibilities for drug discovery.<sup>59,64</sup> For example, polypharmacology accounts for the balance between desired multitarget activity and harmful promiscuity, i.e. the extent to which promiscuous compounds interact with detrimental off-targets.<sup>64,67</sup> In turn, this knowledge can be exploited to understand and ultimately reduce side effects of drug candidates.<sup>68</sup> It has also been shown that drugs used to treat a particular disease can be repurposed for other therapeutic applications.<sup>69</sup> This implies that promiscuous compounds modulate additional targets relevant for new applications or that primary targets of drugs are found in multiple pathophysiological pathways.<sup>70</sup> Prime examples of repurposed drugs include thalidomide, which was initially used in the treatment of morning sickness and is now applied for leprosis and multiple myeloma,<sup>71</sup> and methotrexate, which in addition to cancer has found new applications in inflammatory diseases such as rheumatoid arthritis.<sup>72</sup> Most notably, in the treatment of multifactorial diseases the modulation of a well chosen array of targets by a promiscuous drug may be preferable to the application of a single target agent.<sup>59,61,63,64</sup> For example, promiscuous kinase inhibitors have proven their clinical utility in oncology by interfering with target networks and associated cellular signaling pathways.<sup>73,74</sup> In addition, multitarget strategies offer a promising approach for modulation of complex neurotransmitter systems involved in central nervous system disorders.<sup>64,75</sup>

The extent of multitarget activities among bioactive compounds and drugs is still being investigated and it is becoming apparent that expectations of a widespread drug promiscuity need to be balanced.<sup>63,76,77</sup> Although incompleteness and different confidence levels of available data certainly influence these estimates,<sup>78,79</sup> comprehensive analyses of activity profiles from biological screening assays and the medicinal chemical literature reveal that experimental candidates or drugs are often only active against a small number of targets, or even

consistently inactive.<sup>79–81</sup> Nevertheless, confined subsets of drugs and highly promiscuous compounds exist, thereby providing the basis for polypharmacology approaches.<sup>73,81</sup> Given the complex and challenging nature of polypharmacology, it is important to understand the molecular basis of multitarget activities. To this end, several computational approaches have been introduced to study promiscuity on a molecular level. For example, through systematic compound data mining the SAR of promiscuous compounds was explored and the rational design of multitarget ligands was addressed.<sup>81–85</sup> In addition, new targets of compounds were proposed on the basis of chemical similarities between multitarget ligands,<sup>70,86</sup> and crystallographic data was utilized to study binding sites and functional similarities of targets that bind promiscuous ligands.<sup>87,88</sup>

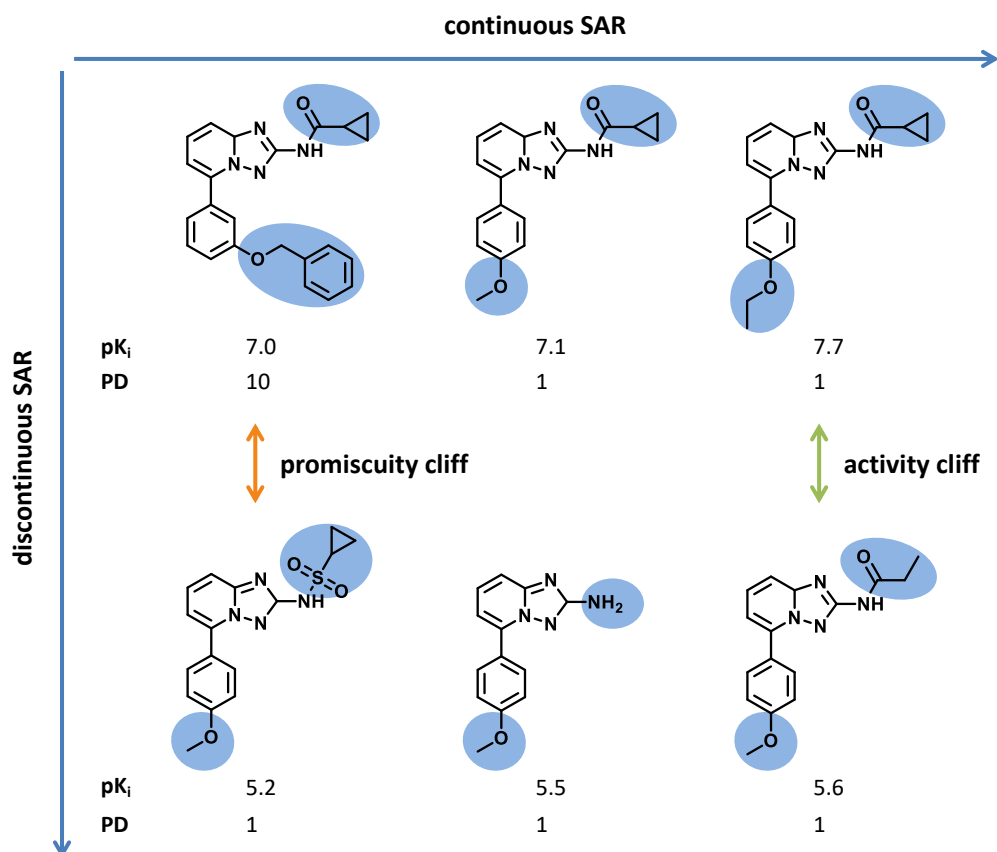
Importantly, compound promiscuity as the basis for polypharmacology does not include nonspecific binding events of frequent hitters due to assay interference or other compound liabilities.<sup>64</sup> Therefore, it is crucial to validate true multitarget activity before studying promiscuity. In the present work, chemoinformatic approaches are discussed that evaluate multitarget binding events on the basis of molecular similarity and molecular interactions of promiscuous compounds. This SAR analysis is based on biological screening data and crystallographic target-ligand complexes.

## 1.2 Structure-Activity Relationship

SAR analysis aims to establish the relationship between chemical structures and biological responses of active compounds. Traditionally these relationships are drawn for individual series of structurally homologous compounds that are tested against small sets of related targets and off-targets, respectively.<sup>89</sup> The emergence of larger data sets through efficient biological screening methods and publicly accessible experimental data enabled computational methods for a large-scale SAR exploration.

In general, these methods can be classified as *descriptive* and *predictive*. Descriptive methods identify SAR determinants by a retro perspective deconvolution and visualization of available SAR information. On the other hand, predictive methods, such as the quantitative SAR (QSAR), attempt to predict biological activity of novel, untested compounds by establishing a correlation between biological





**Figure 4: SAR characteristics.** A series of compounds that displays continuous SAR character is shown (left to right). On the vertical axis, compound pairs that display a discontinuous SAR are given. Below each molecule its biological activity (pK<sub>i</sub>) against tyrosine-protein kinase JAK1 and the promiscuity degree (PD) are given. One activity- and promiscuity cliff is highlighted (green and orange arrow, respectively). Substitution sites are traced in blue.

activity and structural or chemical properties.<sup>90</sup> Common to QSAR models is that they conceptually rely on the so-called ‘similarity property principle’.<sup>91</sup> It states that structurally related molecules should have similar properties. Accordingly, small structural changes should only result in minor potency variations and a continuous SAR (Figure 4). However, exceptions for this rule are frequently observed and represented by structurally homologous compounds that elicit heterogeneous biological activity, thereby displaying a discontinuous SAR (Figure 4). The most extreme form of SAR discontinuity are activity cliffs.<sup>92,93</sup> They are formed by

two structurally similar compounds having a large potency difference. Activity cliffs represent a particular rich source of SAR information by identifying small structural modifications having a major influence on compound potency (Figure 4).<sup>93</sup>

However, as previously discussed, biological activity is not the only determinant of a drug's therapeutic efficacy and utility for drug discovery efforts. To expand the scope of SAR analysis, additional compound properties are investigated. These can include physicochemical features or the promiscuity degree (PD), i.e. the number of targets a compound is active against. It has been shown that promiscuity of analogous compounds, similar to biological activity, is not invariably linear, and small structural modifications can lead to the formation of promiscuity cliffs between structurally related compounds (Figure 4).<sup>94</sup> In the light of apparent multitarget activities, descriptive SAR analysis provides a useful method to explore the molecular basis of assay interference and identify SAR determinants of undesired chemical reactivity.

### 1.2.1 Molecular Similarity

Molecular similarity refers to the concept of clustering compounds on the basis of structural features, biological effects or physicochemical properties.<sup>95,96</sup> This principle finds application in a variety of chemoinformatic methods that either aim to predict compound properties, search and extract large databases for similar molecules, or design new entities with desired descriptors.<sup>95</sup>

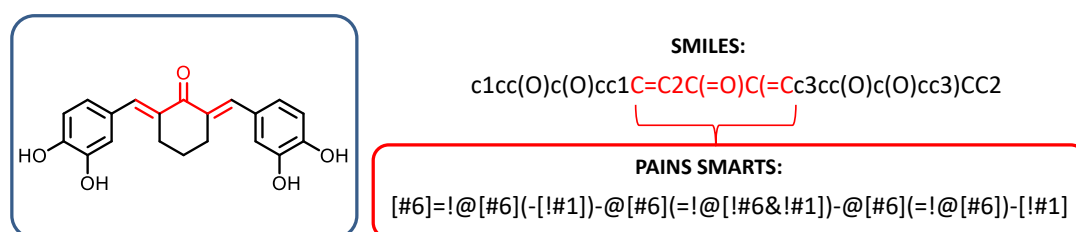
However, the medicinal chemist's perception of similarity is usually guided by intuition, knowledge, and individual experience.<sup>18,96</sup> There is often little agreement between experts compound classification based on desired or undesired structural features and results vary depending on the context in which structures are presented. In addition, similarity comparison is driven by the focus on local patterns that are responsible for a distinct biological response such as activity or interference.<sup>18,57,58,97</sup> On the other hand, chemoinformatic approaches towards a systematic similarity assessment offer the possibility to circumvent this subjective nature.<sup>96</sup> The computational evaluation of molecular similarities is based on two aspects. First, the representation of molecules covering relevant structural and functional features. Second, a method that extracts information encoded in such

representations and determines similarity between molecules. In general, similarity can be evaluated in a quantitative or qualitative manner and the results of similarity assessments depend on applied methodology and structural diversity of investigated compounds.<sup>96</sup> Therefore, different molecular representations and methods to determine similarity will be discussed in the following.

## 1.2.2 Molecular Representations

Molecular representations encode structural features and/or properties of a molecule. They vary in complexity and included chemical information depending on the desired application. Molecular representations can be subdivided into one-, two-, and three-dimensional ones (1D, 2D, and 3D, respectively).<sup>96,98</sup>

1D representations include the molecular formula, as well as the simplified molecular input line entry specification (SMILES).<sup>99</sup> Based on predefined rules, SMILES are linear notations that represent molecular structures by taking into account atom types, branching, stereochemistry, and aromaticity. Notably, PAINS substructures are annotated as SMILES arbitrary target specification (SMARTS), an extension of SMILES that introduces atom and bond labels containing logical operators (Figure 5). SMARTS are typically used for substructure searches in compound databases.<sup>38,100</sup>



**Figure 5: PAINS substructure.** A divinyl ketone that is classified as PAINS is shown. In the molecular graph representation of the compound, atoms and bonds that are part of the PAINS substructure are colored in red. Corresponding characters in the SMILES are also colored in red and the SMARTS string of the PAINS substructure is provided.

Molecular graphs are 2D representations of small molecules that provide an intuitive model of molecular topology and structure. In these graphs, nodes and edges correspond to atoms and bonds, respectively (Figure 6). To represent the steric and electronic properties of a molecule, 3D representations additionally take

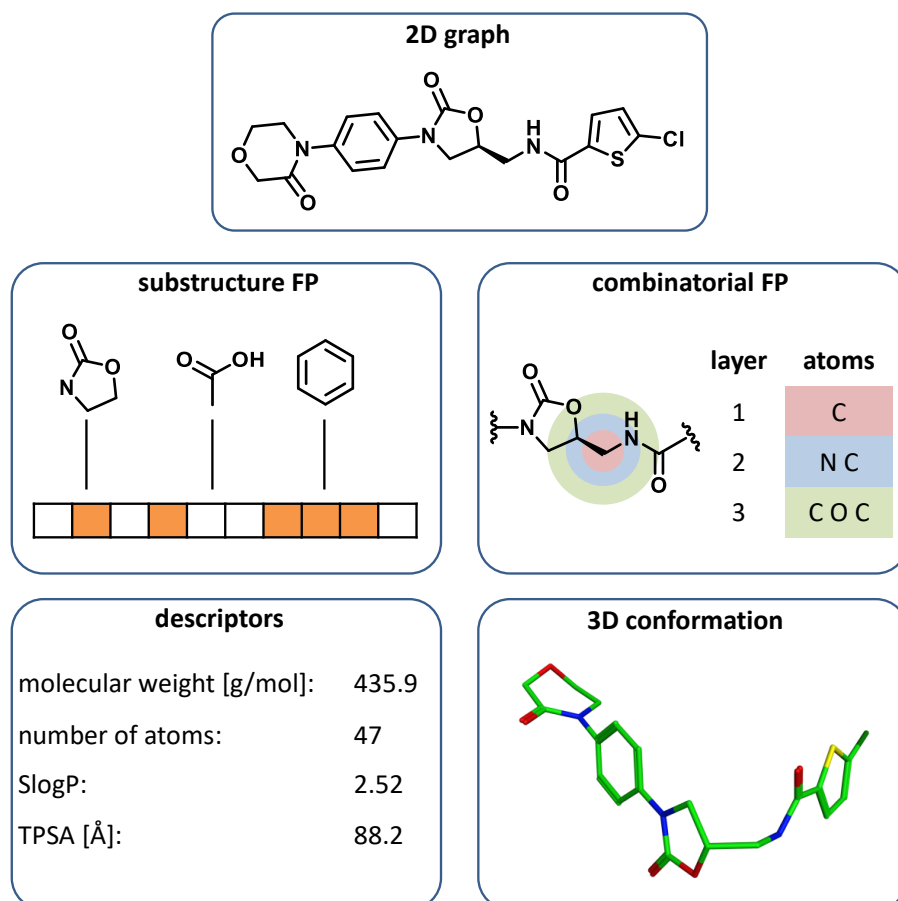
into account the spatial arrangement of atoms and bonds. These 3D representations are typically used to determine a molecule's conformation, surface, or volume.<sup>96</sup> To provide a computer-interpretable form of 2D or 3D representations, atom coordinates, bond orders, charges, and hybridization states can be encoded into connectivity tables, such as in MOL or PDB file types.<sup>101,102</sup>

On the basis of the aforementioned structure-based molecular representations, molecules can additionally be described in terms of numerical values. These molecular descriptors are mathematical models that capture a variety of chemical properties of compounds.<sup>98,103,104</sup> 1D descriptors are calculated on the basis of linear notations and do not consider atom connectivity information. These simple descriptors include for example atom counts and molecular weight. Descriptors that require the molecular topology and structure are called 2D descriptors. Such 2D descriptors are approximated physicochemical properties such as SlogP<sup>105</sup> or topological indices like the topological polar surface area (TPSA).<sup>106</sup> Finally, 3D descriptors, such as the solvent accessible surface area, are calculated on the basis of a compound's distinct 3D conformation.

Fingerprints (FPs) are a subset of molecular descriptors that are utilized in chemoinformatics to represent a molecule as a numerical vector (Figure 6).<sup>108</sup> In binary fingerprints, a position in the bit vector accounts for the absence or presence of a specific chemical feature. The presence of a feature will set the corresponding bit to 1, while its absence will set it to 0. In non-binary versions of fingerprints, each position numerically accounts for the frequency of occurrence of the underlying feature. Based on the molecular representation they originate from (i.e. 2D graph or 3D conformation), fingerprints are typically classified as 2D or 3D. In addition, fingerprints vary in the chemical features they use. Substructure fingerprints utilize dictionaries with predefined substructures and create fingerprints of fixed length. One prime example is the molecular ACCess system (MACCS) consisting of a set of 166 structural patterns.<sup>109</sup>

In contrast, combinatorial fingerprints are of variable length and molecule-specific as they do not use predefined substructures. For example, extended-connectivity fingerprints (ECFPs) encode all possible subgraphs of a given molecule up to a specific bond diameter.<sup>110</sup> For the comparison of molecules represented as fingerprints suitable similarity metrics, including a variety of coefficient and distance functions, are used.<sup>96</sup> However, these whole-molecule similarity assessments

are often difficult to reconcile with medicinal chemistry approaches. The outcome of a similarity assessment often changes with the used fingerprint. Moreover, no generally applicable thresholds for similarity coefficients that indicate structural or activity similarity exist. Hence, the use of a metric represents a quantitative assessment of molecular similarity in which compounds are ranked by a numeri-



**Figure 6: Molecular representations.** Rivaroxaban and schematic representations of fingerprints, descriptors, and a 3D conformation, are shown. For the substructure fingerprint (FP) the bit string is represented as a set of cells. Orange cells denote chemical patterns present in the molecule. White cells represent absent patterns. For the combinatorial FP, concentric layers with a maximum radius of 2 around an exemplary root atom are shown. For each layer, the type of additional atoms is reported. Finally, molecular descriptor values and the 3D binding conformation of rivaroxaban against factor Xa are given (PDB-ID: 2W26).<sup>107</sup>

cal value. In contrast, the application of a structure-based similarity assessment that is based on predefined rules provides a qualitative result, i.e. whether two compounds are similar to each other or not.<sup>89,96</sup>

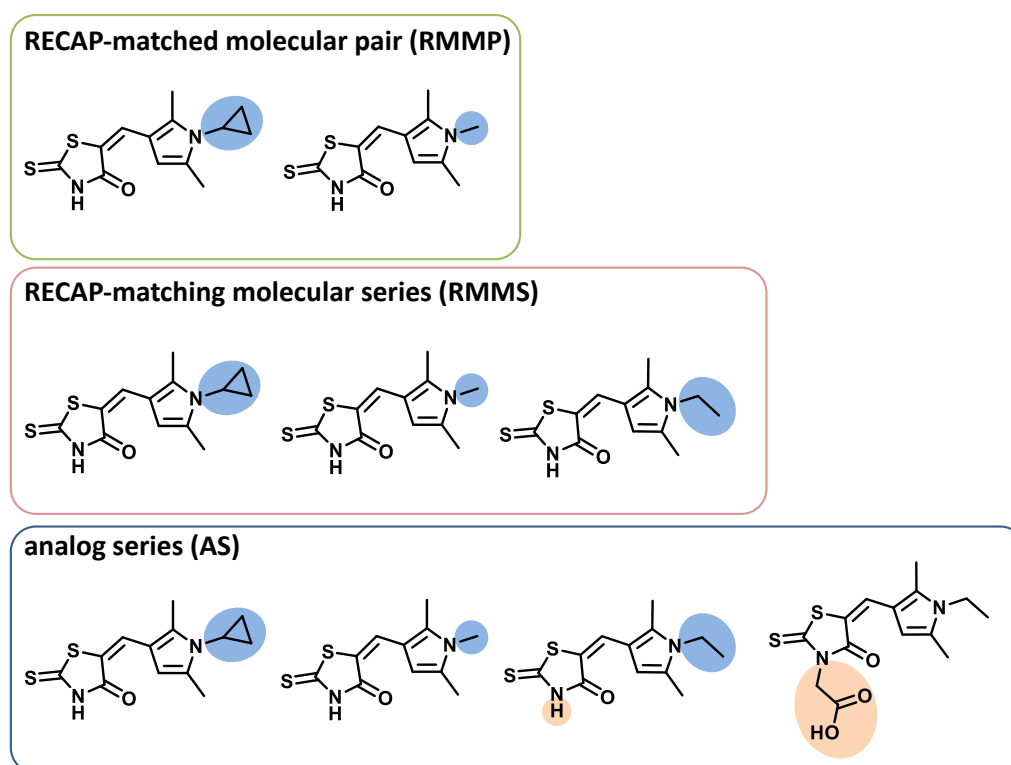
### 1.2.3 Matched Molecular Pairs

The concept of matched molecular pairs (MMPs) allows a structure-based comparison of molecular representations that circumvents the subjective nature of similarity thresholds and provides a chemically intuitive view on molecular similarity. MMPs are defined by a pair of molecules that only differs by a structural modification at a single site.<sup>111</sup> Consequently, compounds that display a MMP relationship can be interconverted into one another by the exchange of a substructure, often termed chemical transformation (Figure 7).<sup>112</sup> MMP-based analyses have found large acceptance and are used for a wide range of chemoinformatic applications, such as SAR matrices, and the exploration of activity cliffs.<sup>93,113,114</sup> One of the first methods to calculate MMPs utilized the maximum common substructure (MCS) of a compound pair.<sup>115</sup> In this method, the largest substructure shared between a pair of molecules is determined and identified as the shared core. The remaining part of the compounds represents the chemical transformation. However, MCS-based MMP generation is computationally not efficient as it requires the generation of MCS for all possible compound pairs in a data set.<sup>115,116</sup> Alternatively, shared substructures between compound pairs can be identified using fragmentation-based algorithms. In a simplified manner, these algorithms can be regarded as a two-step process. In the first step, a compound library is subjected to a rule-based fragmentation procedure. The second stage results in MMP identification by indexing and subsequent comparison of generated fragments. In comparison to MCS-based approaches, fragmentation-based algorithms can be applied to large data sets and are computationally more efficient since each molecule is processed only once.

A widely used fragmentation-based algorithm was introduced by Hussain and Rea.<sup>117</sup> In this fragmentation process, each exocyclic single bond between two non-hydrogen atoms is cleaved. These so-called 'single cuts' result in two fragments. The larger fragment of a compound is termed 'key' and the remaining smaller substructure as 'value'. Fragmentation can also occur at two or three

bonds at the same time, thereby generating 'double-' or 'triple-cut' fragments, respectively. Generated fragments are stored as key-value pairs in an index table. A key entry (i.e. the shared molecular core) that refers to two values (i.e. the different substituents) represents an MMP.

However, keys and values are often not chemically intuitive. Transformation size-restricted MMPs have been introduced to ensure that key fragments are larger than values. This confines MMPs to structurally analogous compounds that differ by interpretable substitutions, such as functional groups or single ring systems.<sup>114</sup> Additionally, fragmentation of molecules based on the well-known retrosynthetic combinatorial analysis procedure (RECAP)<sup>118</sup> allows for synthetically accessible transformations between MMPs, often referred to as RECAP-MMPs (RMMPs) (Figure 7).<sup>119</sup>



**Figure 7: Matched molecular pair, matching molecular series, and analog series (AS).** Shown are analogs of unsaturated rhodanines forming a RECAP-matched molecular pair (top), a RECAP-matching molecular series (middle), and an analog series (bottom). Exchanged substructures are highlighted in blue and orange, respectively.

Matching molecular series (MMS) represent an extension of the MMP concept.<sup>120</sup> MMSs are sets of three or more molecules that share a common substructure, i.e. a key fragment that has more than two value fragments. All compounds in the MMS display a pairwise MMP relationship and thus represent a series of analogs with structural modifications at a single site. MMSs that are derived from RECAP transformations are termed RECAP-MMSs (RMMSs) (Figure 7). Computational SAR exploration based on series, such as RMMSs, reflect the series-centric efforts that dominate medicinal chemistry programs.

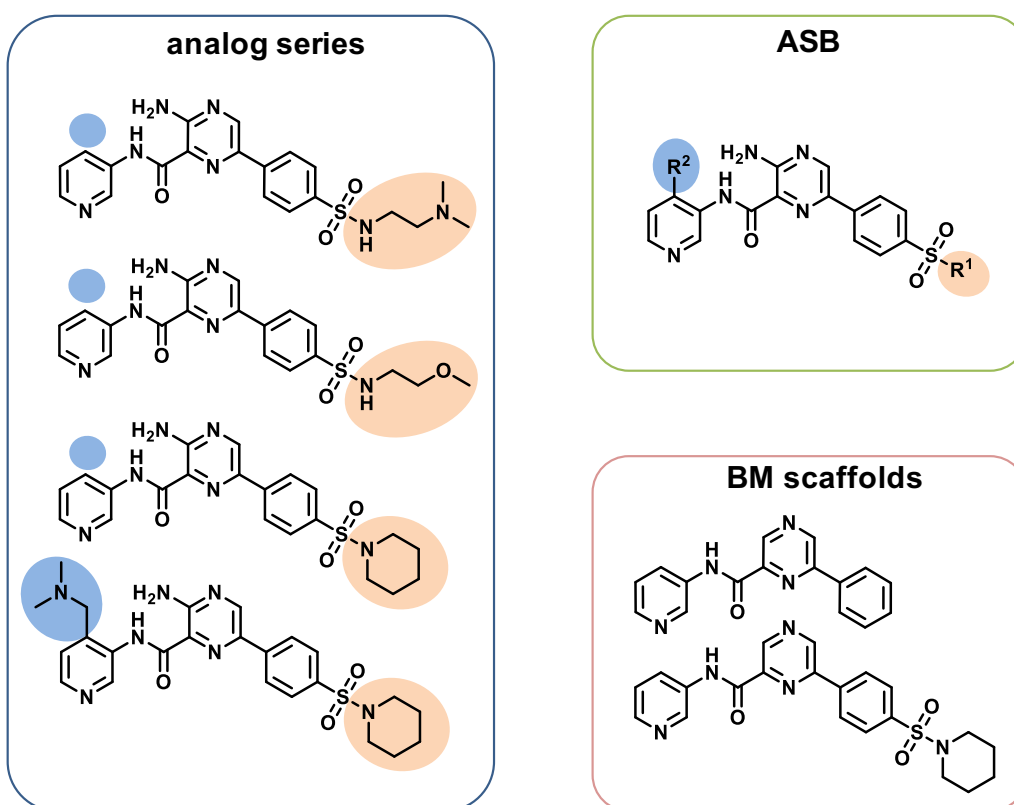
To comprehensively cover the analog space of a given compound data set, the concept of analog series (AS) has been introduced.<sup>121</sup> In this approach, RMMPs are systematically organized in global networks where nodes represent compounds and edges a pairwise RMMP-relationship. Each disjoint cluster represents a distinct AS. ASs may combine RMMSs with overlapping compounds. In contrast to RMMSs, where a compound can be part of multiple RMMSs, compounds are exclusively assigned to one specific AS. Thus, ASs cover all chemically explored substitution sites and available substitution patterns of analogous compounds (Figure 7).<sup>121</sup>

#### 1.2.4 Scaffolds

From a medicinal chemist's point of view, a scaffold represents the core structure of bioactive molecules to which structural modifications are synthetically introduced to exchange functional groups. Typically, this iterative process results in a series of analogous compounds, in which substitutions of the core structure are presented in R-group tables.<sup>114</sup> In this case, the definition of a scaffold relies on the structural context of the series and the subjective perception of a chemist. However, cheminformatic approaches aim to compare the SAR of different series, systematically replace core structures, or structurally organize large screening libraries. Hence, computational methods require a consistent definition of scaffolds.<sup>114</sup>

A widely applied molecular graph based definition was provided by Bemis and Murcko (BM).<sup>122</sup> This approach follows a molecular hierarchy by dividing a compound into ring systems, chemical linker fragments, and R-groups. A scaffold is obtained by removal of all R-groups while retaining ring systems and linker elements connecting two or more rings. From a chemical perspective, this method





**Figure 8: Scaffolds.** For an exemplary AS, corresponding BM and ASB scaffolds are shown. Exchanged substructures are colored according to the substitution site of the ASB scaffold (R1, orange and R2, blue).

contains intrinsic limitations. By definition, the addition of a ring to a BM scaffold yields a new scaffold, although experimental analog generation often introduces additional rings to core structures as R-groups. In addition, BM scaffolds are not generated considering chemical reaction information.<sup>122</sup>

As an alternative to the compound based definition of BM scaffolds, AS-based (ASB) scaffolds were introduced that are derived from ASs containing single as well as multiple substitution sites (Figure 8).<sup>123,124</sup> The generation of an ASB scaffold can be rationalized by considering RMMS (*vide supra*). By definition, a RMMS represents the most simple form of an AS that contains only analogs with a single substitution site. Consequently, ASs covering multiple substitution sites must consist of more than one overlapping RMMS. Thus, two overlapping RMMSs must share at least one analog with structural modifications at two distinct sites,

therefore participating in two RMMPs. The overlapping part of two RMMS-cores (i.e. the RECAP-MMP cores of two RMMSs) yields the ASB scaffold of the AS. In contrast to BM scaffolds, ASB scaffolds include retrosynthetic information and are not restricted by a predefined molecular hierarchy. In addition, the use of a BM scaffold determines the composition of a series and may possibly limit the chemical information it contains. ASB scaffolds, on the other hand, are derived from existing ASs. Thus, ASB scaffolds ensure that all conserved structural elements of ASs are covered (Figure 8) and, according to the properties of corresponding analogs, ASBs can be utilized as templates for the design of focused compound libraries.<sup>124</sup>

### 1.3 Three-Dimensional Target Structures

The completion of the human genome project led to the identification of disease-associated anomalies at the gene level and to complementary proteomic research characterizing (patho)physiological mechanisms of complex diseases. This has enabled the discovery of numerous potential drug targets whose structure and function is investigated at the molecular level of 3D target structures.<sup>125,126</sup> Typically, 3D structures of human and pathogenic proteins are determined by X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy.<sup>127-129</sup> Within the last years, an extensive knowledge base of structural data has been obtained that allows molecular details of a drug-target interaction to be investigated and exploited for the design of optimized compounds. This process is often referred to as structure-based drug design, and covers a variety of computer aided methods, including molecular docking and homology modeling.<sup>27,130</sup>

In the context of assay interference and multitarget activities, target-ligand complexes provide firm evidence for true binding events. Hence, they can be utilized to circumvent uncertainties associated with biological screening data to explore the molecular basis of promiscuity. Although pharmaceutical companies keep many experimental structures unpublished, a significant amount of structures was made publicly available.

### 1.3.1 The Protein Data Bank

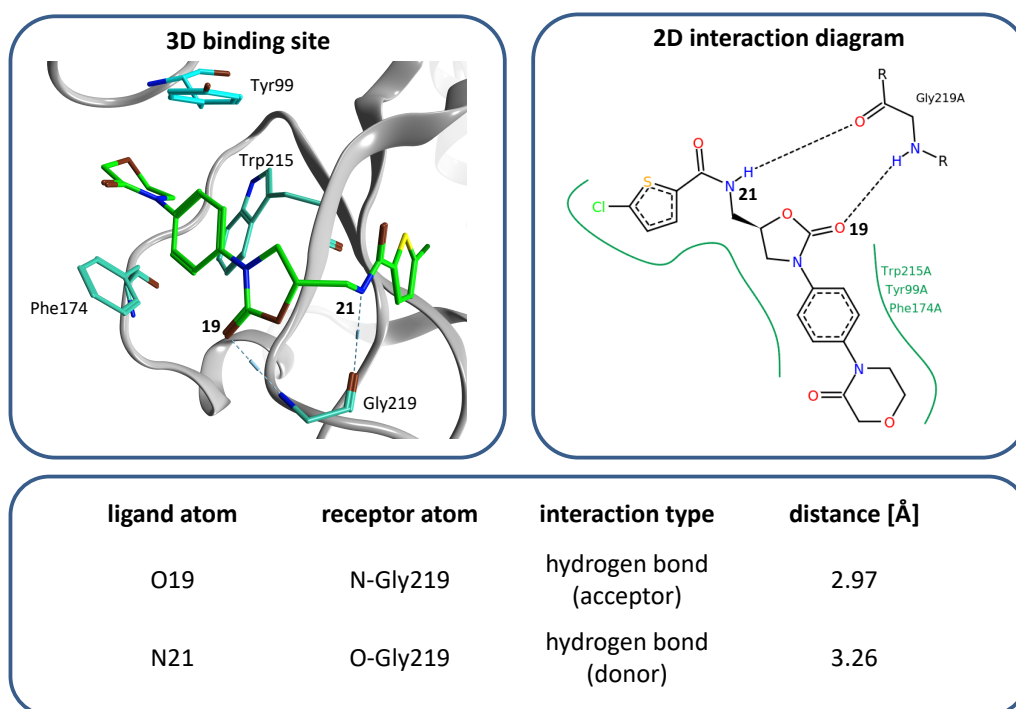
Since its launch in 1971, the Protein Data Bank (PDB) has grown to the largest source of publicly available experimental complex structures. Starting with only seven protein structures, the PDB archive has experienced a nearly exponential growth over the years and by now contains more than 147,633 entries (accessed on January, 16 2019).<sup>131</sup> Those include 132,376 X-ray, 12,494 NMR, and 2,763 electron microscopy structures. Proteins represent the major type of molecules in the PDB compared to nucleic acids, and protein nucleic acid complex structures with 2,2% and 5,0% of the total number of entries, respectively. The PDB presents a global view of structural biology that goes beyond the simple provision of structural data, i.e. atomic coordinates. For example, sequence and 3D structure similarity information is made available for deposited structures as well as corresponding metabolic pathways and gene information. In addition, co-crystal structures of bioactive ligands are mapped in DrugBank and PDBbind to provide potency annotations and further information about ligands.<sup>132,133</sup> This also includes interactive tools to visualize whole protein structures, target ligand interactions, and binding sites.<sup>134</sup> Although experimentally determined 3D structures are of great value for the binding mode analysis and optimization of drug candidates, they are associated with intrinsic limitations. First, X-ray complexes are not equally accessible for all target structures. For example, cell membrane-associated proteins, such as receptors or ion channels, can adopt different conformations and are hard to crystallize in their natural environment of the lipid bilayer.<sup>129,135</sup> This results in an over-representation of cytosolic and secreted proteins in the PDB.<sup>131</sup> Secondly, crystal structure coordinates provide a misleading static view of molecular interactions and protein-ligand binding events. In reality, it must be assumed that a macromolecular complex is an ensemble of several conformations that continuously transform into each other. Moreover, the underlying thermodynamic process includes desolvation, long-range interactions, and entropic effects involving protein, ligand, and solvent.<sup>136-138</sup> Therefore, a crystal structure provides a time-average view of the protein-ligand complex under experimental crystallization conditions. However, it is assumed that molecular recognition in biological systems relies on the existence of specific short-range interactions and a high degree of shape complementarity between protein and ligand.<sup>136,139</sup>

Thus, experimentally determined complexes do not only provide evidence of true binding events, but can also be used to explore molecular mechanisms that allow promiscuous compounds to bind distinct targets.

### 1.3.2 Target-Ligand Interactions

While molecular modeling techniques, such as quantum mechanics and molecular dynamics, investigate short-range interactions in a 3D manner, chemoinformatic approaches typically transform experimentally determined complexes into 2D or 1D representations.<sup>140</sup> Hence, the combination of both approaches allows for efficient analysis of a large array of 3D information. A prime example of this interplay represents the analysis of structural similar crystallographic ligands that display large differences in potency against a specific target. This study provided a structural rationale for activity cliffs.<sup>141,142</sup> Various chemoinformatic approaches have been introduced that can process large amounts of protein-ligand complexes. For example, schematic 2D diagrams of protein-ligand interactions are used to facilitate an initial assessment of structural binding site information and provide a summary of interaction distances and energies (Figure 9).<sup>143</sup>

Further simplifying 3D complexes, protein-ligand interaction fingerprints (PLIFS) encode the absence or presence of specific interactions between ligand atoms and protein residues, thus providing a 1D structural interaction profile of a ligand. PLIFs typically find application in the post-processing of docking results and the analysis of large data sets.<sup>140,144,145</sup> For example, they have been utilized for the assessment of ligand binding similarities across related target proteins.<sup>146</sup> Depending on the underlying method, interactions are identified by geometrical criteria, i.e. distance and angle between ligand and protein atoms (Figure 9).<sup>136,143</sup> Additionally, favorable interactions are often prioritized based on their interaction energies, which are typically derived from empirical- or knowledge-based scoring functions.<sup>147,148</sup> Recent advances in the implementation of molecular modeling applications<sup>149</sup> in modular pipelining tools<sup>150</sup> provide the possibility to rapidly convert large amounts of 3D structure data into easily interpretable 2D and 1D information. If an appropriate number of complexes of a given ligand with distantly related or unrelated targets is available, a structure-based investigation of promiscuity at the molecular level of detail is possible.



**Figure 9: Binding mode analysis.** The X-ray structure of human factor Xa in complex with rivaroxaban (green color) is shown (top, left). In the 3D depiction of the binding site, the protein backbone and protein residues are colored grey and blue, respectively. A ligand-target interaction diagram of the complex taken from the PDB is depicted (top, right; PDB-ID: 2W26).<sup>107</sup> Short-range interactions of rivaroxaban and human factor Xa are illustrated as dashed lines in the 3D and 2D depiction, respectively. For these interactions, involved atoms of ligand and receptor, interaction types, and distances are provided in a table format.

## 1.4 Outline of the Thesis

This thesis focuses first on the analysis of assay interference and promiscuity based on extensively tested compounds originating from biological screening assays. In *chapter 2*, a subset of highly promiscuous screening compounds is submitted to publicly available PAINS filters and the chemical integrity of filtered compounds is inspected. Shortcomings of structural alerts are elaborated on promiscuous compounds associated with chemical liabilities. Additionally, candidates for polypharmacology with no apparent liabilities are provided. In *chapter 3*, crystallographic complexes of PAINS and biological target structures

are explored. On the basis of structurally confirmed binding events different categories of PAINS-containing ligands are distinguished. Further, the structural context dependency of PAINS activities is rationalized on a structural level.

The following chapters focus on the SAR analysis of PAINS. In *chapter 4*, ASs containing PAINS are generated. Activity profiles of compounds with differently embedded PAINS substructures are studied and structural modifications that affect PAINS activities are elaborated. In *chapter 5*, ECFP4 structural features that favor the correct classification between promiscuous and consistently inactive PAINS by machine learning models are identified. Furthermore, these features are evaluated with respect to their role in the structural context dependence and chemical interpretability of PAINS activities. In *chapter 6*, a data driven statistical analysis of hit rates of extensively assayed compounds is carried out. Without prior filtering steps, MMSs containing compounds with high assay promiscuity are generated and provide a basis for the systematic assessment of assay interference, taking structural context information into account.

In the final chapters, the structure-based approach is revisited to explore compound promiscuity. In *chapter 7*, multitarget ligands and multifamily ligands are identified on the basis of structurally confirmed binding events. For these promiscuous ligands, additional analogs and target annotations are systematically derived from medicinal chemistry literature and ASB scaffolds are isolated that serve as templates for polypharmacological ligand design. In *chapter 8*, promiscuous ligands are characterized based on molecular properties. Moreover, binding conformations and interaction hotspots of multifamily ligands are comprehensively analyzed to rationalize their promiscuous binding behavior. The final chapter summarizes the main points of this work and serves as a conclusion of the thesis.

# 2 Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology

## Introduction

In the context of polypharmacology, the promiscuity of small molecules must be considered from different perspectives. On the one hand, it is well-known that multitarget-dependent pharmacological effects are mediated by promiscuous compounds interacting with multiple distantly or unrelated targets. On the other hand, the promiscuous behavior of small molecules is often associated with non-specific binding events or assay artifacts caused by undesired chemical and biological reactions.

In this chapter, the application of public structural filters is combined with the knowledge-based assessment of chemical liabilities of extensively tested screening compounds. Aggregators and PAINS found in highly promiscuous compounds are identified and potential shortcomings of currently available structural alerts are highlighted. Compound integrity is verified by visual inspection and consideration of high-confidence activity data from medicinal chemistry literature to provide candidate compounds for polypharmacology.

This study provides a starting point for the future improvement of detection methods and the exploration of molecular details of desired and undesired promiscuity.

My main contribution to this work was the visual inspection and knowledge-based identification of chemical liabilities in highly promiscuous screening compounds.

Reprinted with permission from 'E. Gilberg, S. Jasial, D. Stumpfe, D. Dimova, J. Bajorath. 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.' Copyright 2016 American Chemical Society

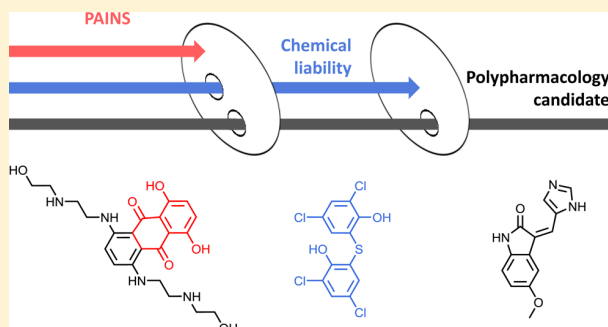


## Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology

Erik Gilberg, Swarit Jasial, Dagmar Stumpfe, Dilyana Dimova, and Jürgen Bajorath\*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

**ABSTRACT:** In PubChem screening assays, 466 highly promiscuous compounds were identified that were examined for known pan-assay interference compounds (PAINS) and aggregators using publicly available filters. These filters detected 210 PAINS and 67 aggregators. Compounds passing the filters included additional PAINS that were not detected, mostly due to tautomerism, and a variety of other potentially reactive compounds currently not encoded as PAINS. For a subset of compounds passing the filters, there was no evidence of potential artifacts. These compounds are considered candidates for further exploring multitarget activities and the molecular basis of polypharmacology.



## INTRODUCTION

Multitarget activities of small molecules can be considered from different viewpoints. For example, polypharmacology is an emerging theme in pharmaceutical research and results from compounds or drugs that act on multiple physiological targets and elicit multitarget-dependent pharmacological effects.<sup>1–3</sup> In the context of polypharmacology, compound promiscuity is defined as the ability of small molecules to specifically interact with multiple targets.<sup>4,5</sup> Such instances of “good” promiscuity must be distinguished from nonspecific interactions or other undesirable effects, leading to artificial activity readouts in biological assays (“bad” promiscuity).<sup>6–9</sup> These include aggregating compounds leading to nonspecific inhibition<sup>6,7</sup> and other compound classes known to produce assay artifacts referred to as pan-assay interference compounds (PAINS).<sup>8,9</sup> Exemplary PAINS are small molecules that are reactive under assay conditions and produce false-positive signals, which presents a major problem for medicinal chemistry.<sup>9</sup> So far, more than 400 compound classes are regarded as PAINS including, among others, rhodanines, curcumines, or quinones.<sup>8,9</sup> However, it is often complicated to unambiguously detect PAINS.<sup>10</sup> To these ends, implementations of PAINS filters have been made publicly available in software tools<sup>11,12</sup> or databases.<sup>13,14</sup> In addition, a public filter for aggregators has also been generated.<sup>15</sup>

In this study, we have determined the most promiscuous compounds from publicly available screening assays.<sup>16</sup> For the majority of these compounds, available activity data met high-confidence criteria. These promiscuous compounds were searched for PAINS and aggregators, which identified a large number of liable molecules. In some instances, PAINS motifs were not detected by standard filters and other compounds

passing the filters also included potentially reactive molecules or chelators not encoded as PAINS. For some highly promiscuous compounds, structural analogues with varying target annotations were identified. In addition, for a subset of promiscuous compounds passing the filters, no obvious liabilities were found. Thus, these compounds might be relevant for the study of polypharmacology.

## RESULTS

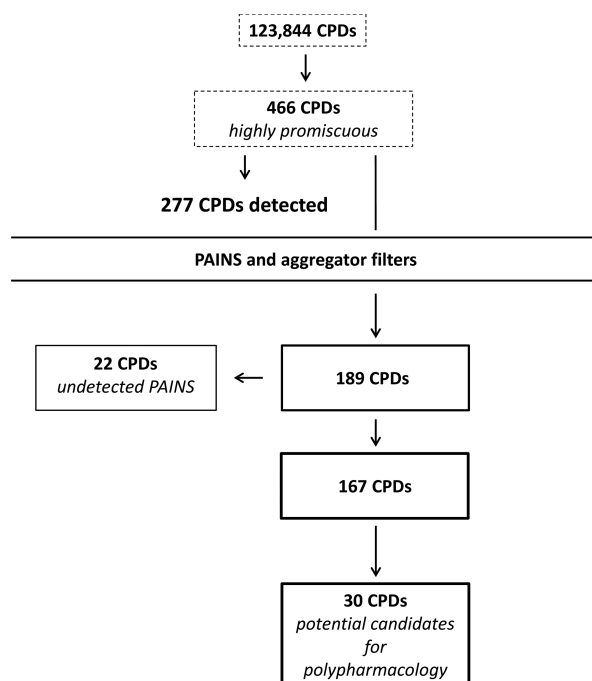
## Identification of Highly Promiscuous Compounds.

From PubChem,<sup>16</sup> compounds were extracted that were extensively tested in primary and confirmatory assays. Primary assays represent initial biological screening data and typically report compound activity as a percentage of inhibition using a single compound concentration. In confirmatory assays, activity measurements are performed at varying compound concentrations, yielding activity values derived from titration curves. We searched for PubChem compounds that were tested in more than 300 primary plus more than 50 confirmatory assays and identified 123,844 qualifying compounds. For each of these compounds, the promiscuity degree (PD; see *Methods and Materials*) was calculated separately for primary and confirmatory assays and compounds were ranked for each assay category in the order of decreasing PD values. Then the overlap among the top 1000 compounds in both rankings was determined, resulting in 466 highly promiscuous compounds (Figure 1).

These compounds were tested in a total of 373–775 assays and reported to be active against 36–128 unique targets on the

Received: September 1, 2016

Published: November 3, 2016



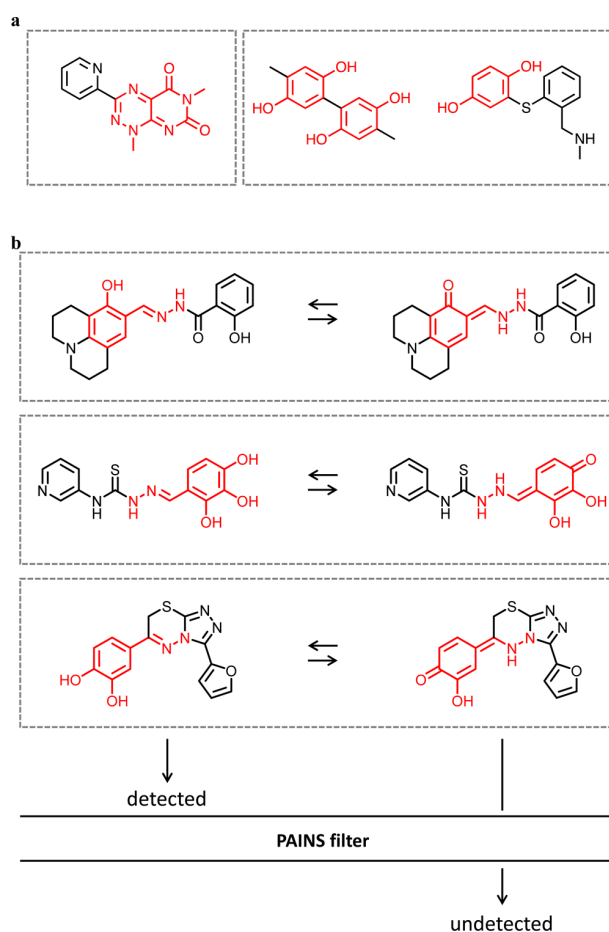
**Figure 1.** Highly promiscuous compounds and their analysis. The flowchart summarizes the steps involved in the identification and analysis of highly promiscuous compounds from PubChem.

basis of PubChem assay records. They represented the most promiscuous compounds we have been able to identify. One might anticipate that compounds having such high degrees of promiscuity are particularly prone to assay artifacts. Therefore, these compounds were initially screened for PAINS and aggregators.

**PAINS and Aggregator Screening.** Different public filters detected 210 of 466 highly promiscuous PubChem compounds as PAINS (45.1%). In addition, 67 compounds were found to be known aggregators (14.4%). The remaining 189 compounds (40.6%) (Figure 1) were further analyzed.

**Undetected PAINS.** Visual inspection of the 189 compounds passing all filters identified 22 compounds that were PAINS or tautomers of PAINS but not detected by filters. These compounds included a toxoflavin, two quinones, and 19 tautomers of hydroxyphenylhydrazones (Figure 2). These findings pointed at a problematic issue with PAINS implementations that do not take tautomerism into account.

**Compounds with Other Potential Liabilities.** Compounds passing the filters also included other reactive compounds, fluorescent molecules, or chelators that were not encoded as PAINS. For example, we identified 17 compounds for which data from the medicinal chemistry literature available in ChEMBL<sup>13</sup> identified additional targets. Figure 3 shows these compounds. They were active against a diverse array of targets from a variety of families, without a notable tendency of preferential activity against individual targets or families. Compounds in Figure 3 include examples of reactive compounds. For example, the photosensitive potential of biothionol **1** has been noted,<sup>19</sup> which led to its withdrawal from the market by the FDA.<sup>20</sup> Moreover, compounds **2** and **6** might undergo ring-opening reactions at sulfur or selenium atoms.<sup>21</sup> In addition, the sesquiterpenelactone helenalin<sup>22</sup> **11** and the maleimide **12** act as Michael acceptors<sup>23</sup> and

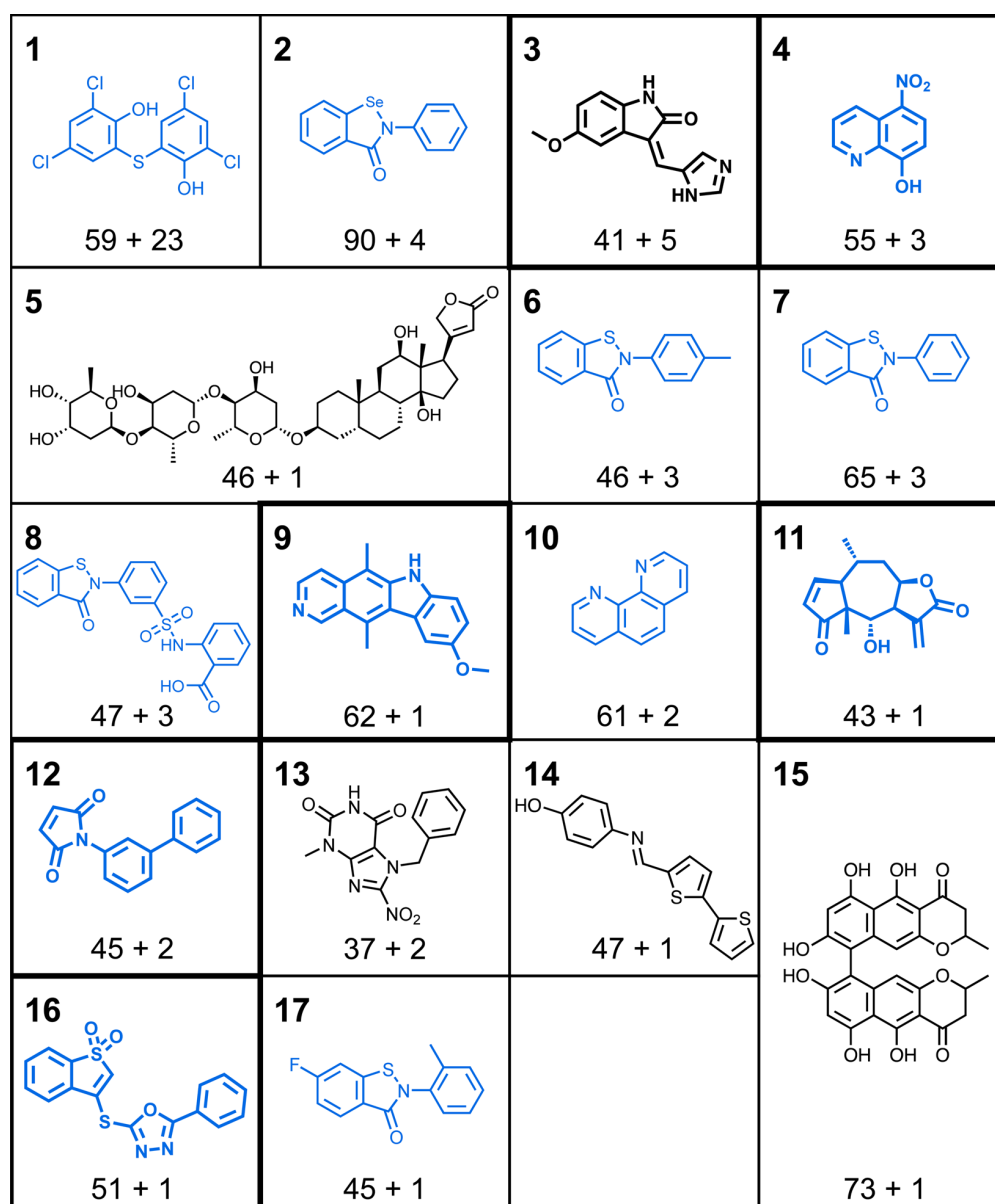


**Figure 2.** Undetected PAINS. Shown are (a) a toxoflavin and two quinones and (b) representative examples of tautomers of hydroxyphenylhydrazones that were not detected by public PAINS filters. Known PAINS substructures are colored red.

compound **16** was reported to react with thiols.<sup>24</sup> Furthermore, compound **4** contains a reactive nitro group and **10** is capable of chelating metal ions.<sup>25</sup> To what extent these properties give rise to systematic assay artifacts remains to be determined. Clearly, it is often difficult to draw a line between known PAINS and compounds containing other potentially reactive moieties. Figure 3 also contains other examples. Compounds **14** and **15**, for example, have no apparent liabilities, and compound **3** is a known promiscuous kinase inhibitor<sup>26</sup> for which 63 analogues with annotations for 16 kinases are available in ChEMBL. Compound **3** is a good example of a polypharmacological compound.

**Promiscuous Compounds with Analogues.** For 31 of the compounds passing the filters, 1–63 structural analogues with available high-confidence activity data were identified in ChEMBL. These analogues belonged to a total of 28 analogue series, 26 of which were annotated with multiple (2–47) targets reported in the medicinal chemistry literature.

**High-Confidence Activity Data.** For all 466 highly promiscuous compounds, a search for high-confidence activity data was carried out in ChEMBL. The availability of high-confidence activity data was thought to lend credence to multitarget activities observed in screening assays. ChEMBL contained assay data incorporated from PubChem that met at least in part high-confidence activity data criteria (see [Methods](#)



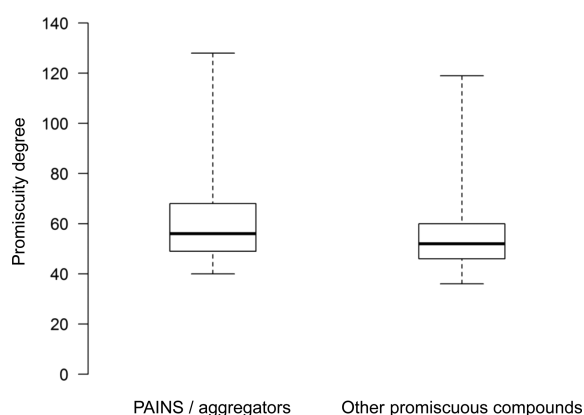
**Figure 3.** Highly promiscuous compounds with additional targets. Shown are 17 compounds from PubChem assays for which high-confidence activity data from the medicinal chemistry literature in ChEMBL identified additional targets. For each compound, target annotations are reported (bottom). “ $x + y$ ” means that “ $x$ ” unique targets originated from PubChem assay data and “ $y$ ” additional targets from separate activity records in ChEMBL. Compounds for which structural analogues with high-confidence activity data were found in ChEMBL are shown in bold. Compounds with potential liabilities not detected by filtering, as discussed in the text, are colored blue.

and Materials) for 206 of the 277 compounds detected as PAINS or aggregators and for 114 of the 167 compounds passing the filters. Thus, high-confidence activity data was available for most highly promiscuous compounds regardless of their structural characteristics.

Figure 4 shows the distribution of PD values for all PAINS and aggregators (with a median PD of 56) in comparison to promiscuous compounds passing the filters (median PD of 52). The distributions were rather similar and PD values were only slightly larger for known PAINS and aggregators than for other promiscuous compounds. Hence, there were also no significant differences in apparent promiscuity levels between these compound subsets. Clearly, from the magnitude of PD values

alone, no conclusions about artificial or true compound activities can be drawn.

**Candidates for Polypharmacology.** The 167 highly promiscuous compounds passing the filters were subjected to visual inspection in light of the PAINS results and other potential activities, as described above, and a subset of 30 compounds was identified that did not display obvious chemical liabilities. Figure 5 shows exemplary compounds. These compounds are thought to represent interesting candidates for the study of polypharmacology. It is noted that compound 21 might decompose over time into different products,<sup>27</sup> an example of a potential activity that may or may not affect biological activities or assay interference.



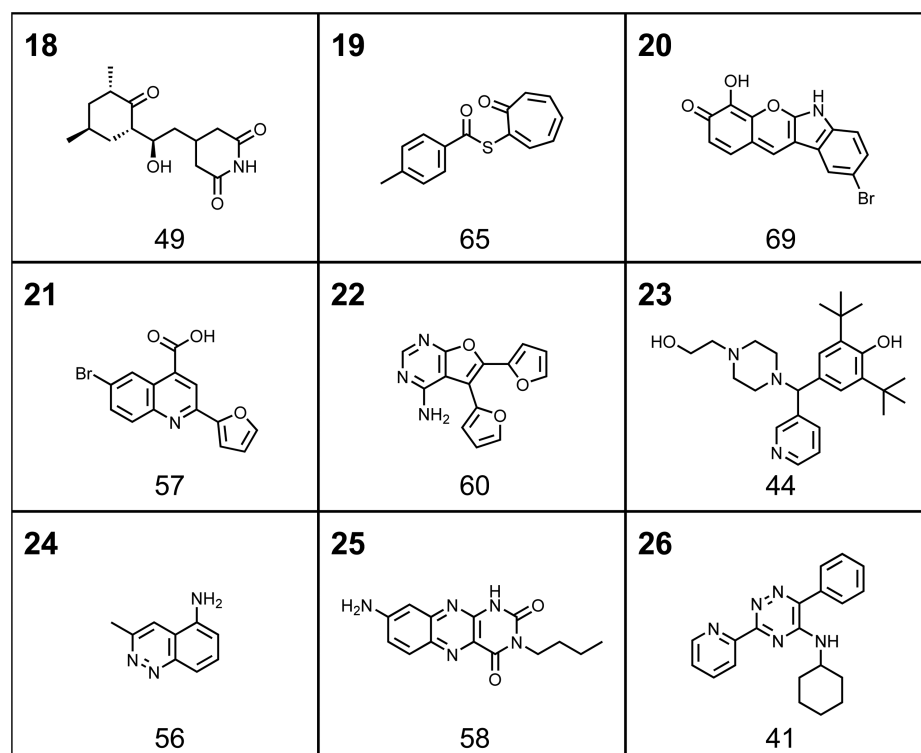
**Figure 4.** Class-specific promiscuity degrees. The PD value distribution of detected PAINS and aggregators is compared to the distribution of other promiscuous compounds passing the filters for which high-confidence data or analogues were available. Boxplots report the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), and largest value (top line).

We also emphasize that putative reactivities do not preclude the evaluation of candidate compounds for polypharmacology. In fact, putative liabilities often suggest experimentally testable hypotheses. For example, if a compound is thought to elicit “bad” promiscuity through chelator properties, examples are shown in blue in Figure 3, derivatives with substitutions of chelating atoms can be assayed, and it can be determined whether or not these derivatives become inactive. If not, the compound becomes even more interesting for the assessment

of multitarget activities. Hence, the subset of candidates for the study of polypharmacology might well be further extended beyond the 30 compounds we have prioritized.

## DISCUSSION AND CONCLUSIONS

In this work, compounds displaying high degrees of promiscuity in screening assays were identified and analyzed. We deliberately focused the analysis on the pinnacle of currently detectable promiscuity, taking into consideration that high degrees of promiscuity might often result from artifacts in screening assays. The set of 466 compounds upon which our analysis was based were the most promiscuous ones we have been able to identify. As anticipated, the majority of these compounds were detected as PAINS or aggregators and represented exemplary cases of “bad” promiscuity. In addition, false-negative PAINS were identified that remained undetected by public filters, mostly due to tautomerism, as well as other compounds potentially causing assay artifacts that were not encoded as PAINS. These problematic issues are expected to apply to many more classes of PAINS and other potentially reactive molecules than covered by our set of highly promiscuous compounds. Public PAINS filters and other detection methods are of critical importance for the medicinal chemistry community but they are currently not sufficiently developed to be fully reliable, as indicated by different types of false-negatives detected in our analysis. High-confidence activity data were available for many compounds detected as PAINS as well as others passing the filters. Thus, the availability of high-confidence activity data is considered a necessary but not sufficient condition for compound integrity.



**Figure 5.** Candidates for polypharmacology. Shown are exemplary promiscuous compounds that passed the filters and did not display obvious chemical liabilities. For each compound, the PD is reported on the basis of target annotations from PubChem (bottom).



Although the magnitude of the PAINS problem is clearly indicated by the results of our analysis, it should be noted that PAINS alerts do not necessarily disqualify compounds from further consideration or invalidate available data. Furthermore, no compound can a priori be disqualified as a bioactive chemical entity if it is potentially reactive or fluorescent. These issues further complicate the assessment of compound liabilities. There are numerous instances of compounds containing PAINS motifs with true activities including a number of marketed drugs.<sup>28</sup> Whether or not potentially reactive compounds give rise to artifacts often depends on specific experimental conditions and modes of action. Furthermore, although promiscuity degrees of bioactive compounds are in our experience often overestimated by taking low-confidence data into consideration, there is no doubt that compounds frequently display multitarget activities that result in side effects or desirable polypharmacology. Our analysis also identified a subset of highly promiscuous compounds for which no chemical liabilities were evident. Thus, although we cannot rule out the presence of other sources of artifacts that might be associated with at least some of these compounds, they are likely to represent interesting starting points for further exploring the molecular basis of polypharmacology. Similarly, reactive compounds not detected by PAINS filters provide opportunities for further studying origins of assay interference.

As a part of our study, all highly promiscuous compounds and the structural analogues we identified are made freely available as an open access deposition together with their target annotations and PAINS/aggregator detection status.<sup>29</sup> These compounds should be useful for further developing detection methods and exploring molecular detail of “good” or “bad” promiscuity.

## METHODS AND MATERIALS

**PubChem Data Collection.** Assay data were extracted from the PubChem BioAssay collection<sup>16</sup> containing primary and confirmatory assays. RNA interference (RNAi) screens were removed from primary assays, and only chemical screens were considered. Confirmatory assays were required to have specified compound activity values for individual protein targets resulting from dose–response curves (typically IC<sub>50</sub> values). From each qualifying assay, compounds classified as “active” or “inactive” were selected and molecules designated as “unspecified” or “inconclusive” were disregarded.<sup>17</sup>

**ChEMBL High-Confidence Activity Data.** In ChEMBL version 21,<sup>13</sup> all compounds were identified for which high-confidence activity data were available. We required the presence of direct interactions (i.e., assay relationship type “D”) with human single-protein targets at the highest confidence level (i.e., assay confidence score 9). Relationship type “D” and confidence score 9 represent the highest level of confidence for activity data from ChEMBL. In addition, only explicitly specified equilibrium constants ( $K_i$ ) and IC<sub>50</sub> values were considered as potency measurements. All approximate measurements such as “>”, “<”, or “~” and activity records including comments such as “inactive”, “inconclusive”, or “not active” were discarded. This protocol identified 174,839 qualifying compounds.

Because ChEMBL incorporates assay data from PubChem, these data were also filtered according to high-confidence criteria and distinguished from activity data originating from the medicinal chemistry literature.

**Promiscuity Analysis.** Target annotations were determined on the basis of assay activity records (PubChem) and high-confidence activity data (ChEMBL). The PD of a compound was defined as the number of unique targets it was reported to be active against. The PD was initially calculated on the basis of PubChem annotations and adjusted if additional targets were identified in ChEMBL.

**PAINS and Aggregators.** Promiscuous compounds were screened for PAINS using three public PAINS filters available in RDKit,<sup>12</sup> ChEMBL,<sup>13</sup> and ZINC.<sup>14</sup> Compounds were represented as canonical SMILES on the basis of hydrogen suppressed graphs and submitted to the respective filters. We note that ChEMBL does not provide an interactive filter but lists PAINS annotations in compound records instead, from which they can be retrieved. Following PAINS filtering, compounds were also screened for known aggregators using the ZINC aggregator filter.<sup>15</sup>

**Analogue Search.** Analogue searching in ChEMBL was carried out using a recently reported matched molecular pair-based methodology.<sup>18</sup>

## AUTHOR INFORMATION

### Corresponding Author

\*Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de).

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The use of OpenEye’s toolkits was made possible by their free academic licensing program. D.S. is supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft. We are grateful to a reviewer for pointing out specific examples of reactive compounds and relevant references.

## ABBREVIATIONS USED

PAINS, Pan-assay interference compounds; PD, promiscuity degree

## REFERENCES

- (1) 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.
- (2) Boran, A. D.; Iyengar, R. Systems Approaches to Polypharmacology and Drug Discovery. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 297–309.
- (3) Jalencas, X.; Mestres, J. On the Origins of Drug Polypharmacology. *MedChemComm* **2013**, *4*, 80–87.
- (4) Hu, Y.; Bajorath, J. Compound Promiscuity - What Can We Learn From Current Data. *Drug Discovery Today* **2013**, *18*, 644–650.
- (5) Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity. *F1000Research* **2013**, *2*, 144.
- (6) McGovern, S. L.; Caselli, E.; Grigorieff, N. A.; Shoichet, B. K. Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.
- (7) Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
- (8) 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.
- (9) Baell, J. B.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.
- (10) 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.
- (11) 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.
- (12) *RDKit: Cheminformatics and Machine Learning Software*; RDKit, 2013; <http://www.rdkit.org>.
- (13) 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.

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

(15) 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.

(16) 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.

(17) Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLoS One* **2016**, *11*, e0153873.

(18) 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. *J. Med. Chem.* **2016**, *59*, 7667–7676.

(19) Henry, B.; Foti, C.; Alsante, K. Can Light Absorption and Photostability Data be Used to Assess the Photosafety Risks in Patients for a New Drug Molecule? *J. Photochem. Photobiol., B* **2009**, *96*, 57–62.

(20) Pharmacy Compounding. *Code of Federal Regulations*, Section 216.24, Title 21, 2016.

(21) Sanchez, J. P. A Ring Opening Reaction of Benzisothiazolones. A New Route to Unsymmetrical Disulfides. *J. Heterocycl. Chem.* **1997**, *34*, 1463–1467.

(22) Lyss, G.; Knorre, A.; Schmidt, T. J.; Pahl, H. L.; Merfort, I. The Anti-Inflammatory Sqsuiterpene Lactone Helenalin Inhibits the Transcription Factor NF- $\kappa$ B by Directly Targeting p65. *J. Biol. Chem.* **1998**, *273*, 33508–33516.

(23) Arróniz, C.; Molina, J.; Abás, S.; Molins, E.; Campanera, J. M.; Luque, F. J.; Escolano, C. First Diastereoselective [3 + 2] Cycloaddition Reaction of Diethyl Isocyanomethylphosphonate and Maleimides. *Org. Biomol. Chem.* **2013**, *11*, 1640–1649.

(24) Dahlin, J. L.; Nissink, 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.

(25) Tabrizi, L.; Fooladivanda, M.; Chiniforoshan, H. Copper(II), Cobalt(II) and Nickel(II) Complexes of Juglone: Synthesis, Structure, DNA Interaction and Enhanced Cytotoxicity. *BioMetals* **2016**, DOI: 10.1007/s10534-016-9970-0.

(26) Anastasiadis, T.; Deacon, S. W.; Devarajan, K.; Ma, H.; Peterson, J. R. Comprehensive Assay of Kinase Catalytic Activity Reveals Features of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2011**, *29*, 1039–1045.

(27) Olson, M. E.; Abate-Pella, D.; Perkins, A. L.; Li, M.; Carpenter, M. A.; Rathore, A.; Harris, R. S.; Harki, D. A. Oxidative Reactivities of 2-Furylquinolines: Ubiquitous Scaffolds in Common High-Throughput Screening Libraries. *J. Med. Chem.* **2015**, *58*, 7419–7430.

(28) Senger, M. R.; Fraga, C. A. M.; Dantas, R. F.; Silva, F. P., Jr. Filtering Promiscuous Compounds in Early Drug Discovery: Is It a Good Idea? *Drug Discovery Today* **2016**, *21*, 868–872.

(29) <https://zenodo.org/record/164405>

## Conclusions

Herein, highly promiscuous compounds were explored with respect to their chemical integrity and potential usability in polypharmacology. A set of 466 compounds displaying the highest degrees of promiscuity in primary and confirmatory biological screening assays has been identified and was submitted to structural filters. A majority of those were detected as colloidal aggregators or PAINS, classifying them as exemplary cases of undesired promiscuity.

The remaining 189 compounds were visually inspected and false-negative PAINS as well as compounds with other potential liabilities were identified. These results revealed potential challenges with PAINS implementations that do not address tautomerism and do not fully cover chemical liabilities, including autofluorescence and chelation. Additional high-confidence activity annotations were available for both detected PAINS and compounds passing structural filters. This indicated that high-confidence activity data displayed a necessary but not sufficient prerequisite for compound integrity. Finally, 30 highly promiscuous compounds were identified for which no chemical liabilities were evident, providing a starting point for further research in the field of polypharmacology.

In this study, the knowledge-based analysis of extensively tested screening compounds showed that promiscuity degrees and structural filters alone are not sufficient indicators of the quality for a compound's promiscuity. Thus, upon the utility of promiscuous compounds for polypharmacology must often be decided on a case-by-case basis. In the next chapter, we explore crystallographic protein-ligand complexes for known PAINS and draw first structure-activity relationships between interference compounds.





# 3 X-ray Structures of Target-Ligand Complexes Containing Compounds with Assay Interference Potential

## Introduction

Assay artifacts compromise medicinal chemistry programs and false-positive activities make it difficult to identify genuinely promiscuous compounds. To address this, problematic PAINS classes were proposed as structural alerts for assay interference, which require caution when they occur as substructures in bioactive molecules. However, rationalizing and predicting interference compounds remains difficult, as PAINS have been demonstrated to exhibit heterogeneous activities and varying hit rates. These results indicated that PAINS activities are not exclusively attributed to undesired interference mechanisms.

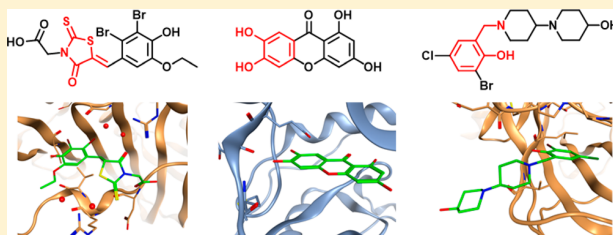
In this work, the analysis of PAINS is extended beyond biological assays and X-ray crystallographic protein-ligand complexes containing compounds with known interference potential are systematically studied. Further, interactions with target proteins are examined on the basis of structural data and different modes of action of PAINS are explored.

Reprinted with permission from 'E. Gilberg, M. Gütschow, J. Bajorath. X-ray Structures of Target-Ligand Complexes Containing Compounds with Assay Interference Potential. *Journal of Medicinal Chemistry* **2018**, 61, 1276–1284.' Copyright 2018 American Chemical Society

## X-ray Structures of Target–Ligand Complexes Containing Compounds with Assay Interference Potential

Erik Gilberg,<sup>†,‡</sup> Michael Gütschow,<sup>‡</sup> and Jürgen Bajorath<sup>\*,†</sup><sup>†</sup>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<sup>‡</sup>Pharmaceutical Institute, Rheinische Friedrich-Wilhelms-Universität, An der Immenburg 4, D-53121 Bonn, Germany

**ABSTRACT:** Pan assay interference compounds (PAINS) have become a paradigm for compound classes that might cause artifacts in biological assays. PAINS-defining substructures are typically contained in larger compounds. We have systematically examined X-ray structures of protein–ligand complexes for compounds containing PAINS motifs. In 2874 X-ray structures, 1107 unique ligands with PAINS substructures belonging to 70 different classes were identified. PAINS most frequently detected in crystallographic ligands included a number of prominent candidates such as quinones, catechols, or Mannich bases. However, on the basis of X-ray data, the presence of specific ligand–target interactions and reactivity under assay conditions were not mutually exclusive. In some instances, reactivity of ligands was likely responsible for complex formation. Different categories of PAINS-containing ligands were distinguished, which aided in the interpretation of specific interactions versus potential assay artifacts. Careful consideration of structural data adds another dimension to the analysis of interference compounds.



## INTRODUCTION

In biological screening and medicinal chemistry, newly identified active compounds and candidates for chemical optimization must be distinguished from others that cause assay artifacts. False-positive activity signals might be due to colloidal aggregation<sup>1–3</sup> or reactivity under assay conditions.<sup>4</sup> Various mechanisms of assay interference have been suggested including, among others, covalent modification of target proteins, autofluorescence, redox reactivity, or chelation.<sup>4,5</sup> Systematic attempts have been made to identify and collect potential aggregators<sup>3</sup> or compounds that cause assay artifacts due to reactivity or autofluorescence,<sup>4</sup> leading to the introduction of pan assay interference compounds (PAINS).<sup>4,5</sup> The latter include intensely studied chemical classes such as rhodanines, quinones, or curcuminoids.<sup>4–6</sup> PAINS-defining moieties are typically contained as substructures in other compounds.

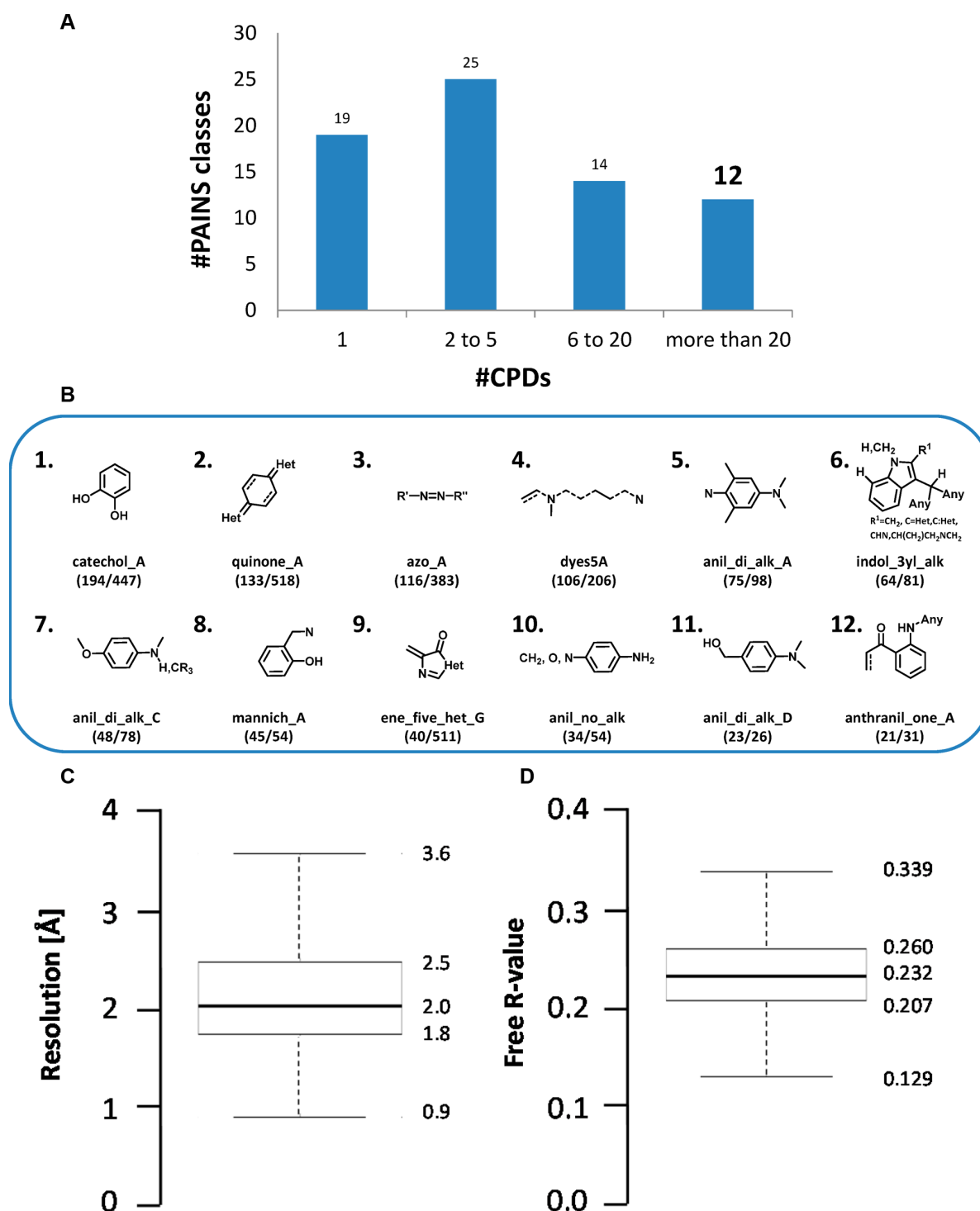
There is no doubt that assay interference represents a major problem for biological screening and medicinal chemistry.<sup>7</sup> However, rationalizing and predicting assay interference potential or molecular promiscuity are far from being simple tasks.<sup>8–12</sup> For example, it has been shown that compounds with PAINS substructures often had very different activity profiles in screening assays and also contained a variety of chemical entities that were consistently inactive.<sup>10,11</sup> It is very likely that the structural environment of PAINS substructures or other reactive moieties in compounds plays a critically important role for their potential to elicit artifacts.<sup>10–12</sup> Substantial evidence for this premise was provided by the study of analog series containing PAINS substructures, which made it possible to

evaluate PAINS in different structural contexts.<sup>12</sup> For example, a number of analog series containing Mannich bases, aniline derivatives, or alkylidene barbiturates were consistently inactive in screening assays, whereas others contained frequently active compounds.<sup>12</sup> Furthermore, structural modifications at a single site of notorious PAINS such as rhodanine or indol derivatives strongly influenced their assay hit rates.<sup>12</sup> In addition to PAINS, many other compounds have interference potential.<sup>8,9</sup> However, even for some highly promiscuous molecules, no evidence or indications of possible assay artifacts exists.<sup>8</sup> Clearly, relationships between inactivity, specific activity, and artificial activity of compounds with potential liabilities are often highly complex and difficult to unravel.

To further extend the analysis of compounds with interference potential beyond biological assays, we have searched X-ray structures of target–ligand complexes for compounds with PAINS substructures. The formation of complexes that can be crystallized is usually driven by specific interactions, which are revealed by X-ray structures. The presence of specific interactions between a ligand and a given target does not preclude desired or undesired interactions with other targets. Importantly, complex structures containing PAINS can be examined for modes of action, which can then be related to potential reactivity and artifacts, hence providing a new reference frame for judging interference potential.

Received: December 4, 2017

Published: January 12, 2018

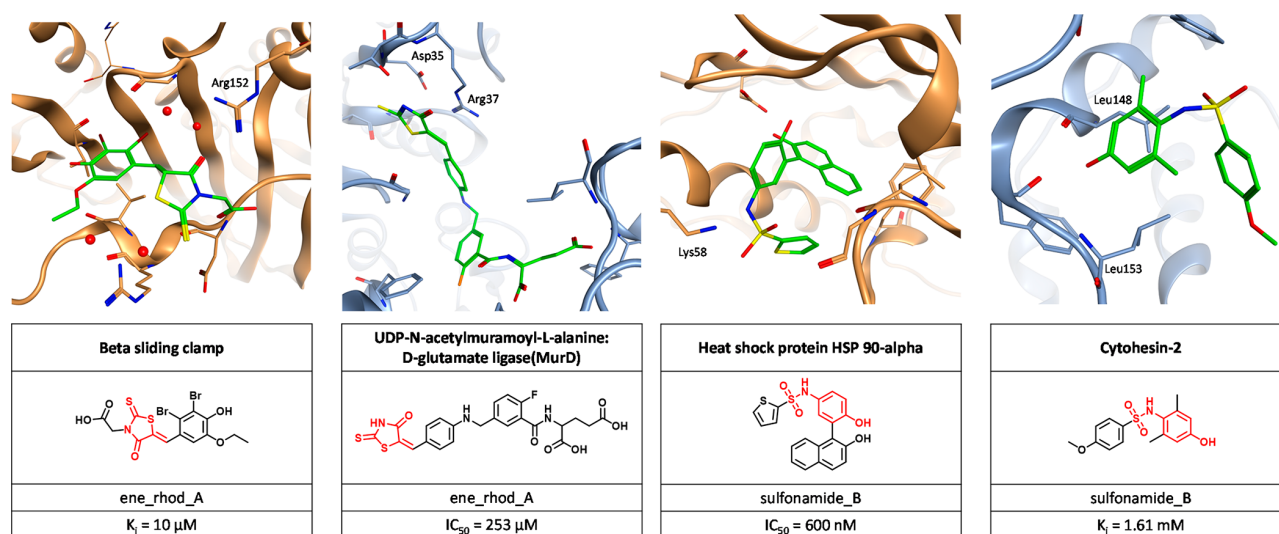


**Figure 1.** Classification of PAINS found in X-ray structures. (a) The distribution of PDB\_PAINS over 70 PAINS classes is shown in a histogram format. Bins report the number of PDB\_PAINS per class. In (b), 12 PAINS classes are shown that were each present in more than 20 PDB\_PAINS. Numbers in parentheses report the number of X-ray ligands belonging to a given PAINS class and the number of complexes in which they occurred. For example, (194/447) means that there were 194 PDB\_PAINS belonging to class catechol\_A that occurred in 447 complex structures. In (c) and (d), box plots report the distribution of crystallographic resolution (Å) and free R-values for all X-ray complexes with PAINS, respectively. Boxplots report the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), and largest value (top line).

## MATERIALS AND METHODS

**X-ray Structures.** Complex X-ray structures were selected from the Ligand Expo section<sup>13</sup> of RCSB Protein Data Bank (PDB),<sup>14</sup>

which provides standardized ligands.<sup>13</sup> X-ray structures were graphically analyzed using the Molecular Operating Environment.<sup>15</sup> Compounds extracted from X-ray structures (PDB ligands) were



**Figure 2.** Structures containing prominent interference compounds. Shown are exemplary X-ray structures with PDB\_PAIS (carbon atoms colored green) that contain prominent PAINS substructures known to frequently cause assay artifacts. Protein backbone ribbons and carbon atoms of protein residues are colored in gold or silver. Other ligand and protein atoms are shown using standard atom coloring. Residues discussed in the text are labeled. In molecular graphs of ligands, PAINS substructures are colored red. Names of target proteins, PAINS codes, and potency values (if available in the literature) are provided. Figures 3–6 are presented accordingly. (a) Two unsaturated rhodanine derivatives are bound to the  $\beta$  sliding clamp protein (PDB ID 3D1G, resolution: 1.64 Å, free  $R$ -value: 0.296) and bacterial MurD ligase (2Y68, 1.49 Å, 0.204), respectively. (b) Two  $p$ -hydroxyphenylsulfonamides in complex with heat shock protein HSP 90- $\alpha$  (2YE9, 2.2 Å, 0.289) and cytohesin-2 (4JMO, 1.8 Å, 0.210).

represented as aromatic nonstereo SMILES.<sup>16</sup> Proteins from X-ray structures were assigned to families on the basis of UniProt identifiers (IDs) and the UniProt classification scheme.<sup>17</sup>

**PAINS Analysis.** PDB ligands were screened *in silico* for compounds containing PAINS substructures (PDB\_PAIS) using SMARTS<sup>18</sup> strings obtained from three public filters available in ChEMBL release 23,<sup>19</sup> RDKit,<sup>20,21</sup> and ZINC.<sup>22</sup> Given possible implementation discrepancies of substructure strings, compounds were classified as PAINS if one or more filters yielded an alert. All calculations were carried out using Python scripts with the aid of the OpenEye chemistry toolkit,<sup>23</sup> RStudio,<sup>24</sup> and KNIME.<sup>25</sup>

PAINS filters are often viewed controversially<sup>10</sup> as indicators of assay interference.<sup>10–12</sup> However, they provide direct access to compound classes with potential liabilities classified as PAINS.<sup>7,11</sup> As such, their application was consistent with the goals of our analysis.

## RESULTS AND DISCUSSION

**PAINS in PDB Ligands.** From PDB entries, we extracted 22,467 ligands with unique structures. These PDB ligands contained 1107 PDB\_PAIS that were present in a total of 2874 complex X-ray structures. This was a much larger number of complex structures containing PAINS than we had expected. The 1107 PDB\_PAIS covered 70 of the 480 originally proposed PAINS classes.<sup>4</sup> Figure 1a shows the distribution of PDB\_PAIS over PAINS classes. The 12 classes of PAINS substructures most frequently detected in PDB ligands are shown in Figure 1b. Notably, these classes contained prominent PAINS such as quinones, catechols, or Mannich bases.<sup>5</sup> Furthermore, for 78 PDB\_PAIS, complex structures with different targets were available including 40 ligands with targets from different families. Hence, for these PDB\_PAIS there was evidence for specific multitarget activity, as further discussed below.

**X-ray Structures with PAINS and Their Analysis.** The 2874 X-ray structures containing PDB\_PAIS were analyzed in detail including one-by-one visual inspection. Following initial inspection, 119 X-ray complexes with low resolution

were omitted from further consideration, yielding a final set of 2755 structures for detailed analysis. The PDB IDs of these X-ray structures and 1094 corresponding PDB\_PAIS are made freely available in an open access deposition, as further specified below.

Figure 1c,d reports the crystallographic resolution and free  $R$ -values for the X-ray complexes, respectively, revealing that the majority of structures were well-refined at medium to high resolution, with a median value of 2.0 Å. We emphasize that care must be taken to avoid overinterpretation of X-ray data, taking resolution limits, potential local disorder, and refinement ambiguities into account, which are not necessarily reflected by global statistics. Structure factors and resulting electron densities are experimental observations, not the atomic models fitted into densities. At a resolution lower than 2 Å, it is becoming difficult to distinguish between covalent and noncovalent bonds, and case-by-case decisions must be made during refinement, which complicates the verification of covalent inhibition on the basis of X-ray structures. These caveats generally apply and must also be taken into consideration in analyzing complexes with PAINS. As discussed above, reactivity under assay conditions does not by default result in covalent inhibition of PAINS, which only applies to a subset of reactive compounds. In fact, on the basis of visual inspection of our collection of X-ray structures with PAINS, the majority of complexes formed were clearly noncovalent. On the other hand, the formation of a noncovalent complex does not preclude reactivity of a compound under different conditions. It proves, however, that a ligand is capable of forming specific interactions with a given target. Formally, it is also possible that cocrystallization might select nonreactive over reactive compounds. Hence, even at the level of well-refined crystallographic models of complexes, the analysis of desirable versus undesirable activities remains complex and overconclusions must be drawn with caution.



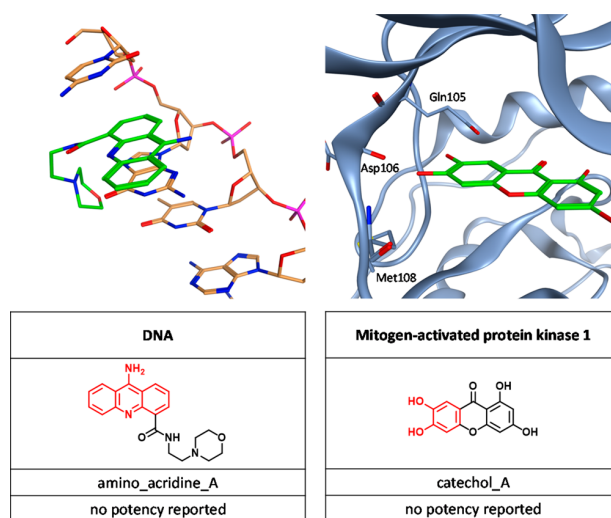
**PDB\_PAINS Categories.** On the basis of visual inspection of the large number of PAINS-containing complex structures, different categories of PDB\_PAINS were identified that were recurrent among X-ray structures. This made it possible to rationalize interactions of PAINS with different targets in light of assay interference potential, which represented a prime motivation for our study. In the following, exemplary structures are discussed that are representative of different categories of PDB\_PAINS we identified.

**Prominent Interference Motifs.** PAINS substructures that are known to frequently cause assay artifacts were found in a variety of complex structures. For instance, five-membered heterocycles such as rhodanines are among notorious PAINS,<sup>5</sup> for which reactions with protein residues have been observed including attacks by protein nucleophiles<sup>27</sup> or metal chelation.<sup>28</sup> The interference potential of rhodanines has been investigated in different structural environments, indicating a context dependence of reactivity.<sup>12,29</sup> On the other hand, rhodanines have also been put forward as “privileged scaffolds” for drug discovery.<sup>30</sup> Figure 2a shows two representative unsaturated rhodanines (PAINS code<sup>4</sup> ‘ene\_rhod\_A’) that contribute to specific molecular interactions in X-ray structures with different targets. For example, in subsite 1 of the  $\beta$  sliding clamp protein,<sup>31</sup> the carbonyl oxygen of the rhodanine heterocycle participates in a water-mediated hydrogen-bond network with the residue Arg152. Furthermore, the terminal arylidene moiety of rhodanine in complex with UDP-*N*-acetylmuramoyl-L-alanine:D-glutamate ligase<sup>32</sup> occupies the uracil-binding pocket and participates in a salt bridge interaction with residues Asp35 and Arg37.

Another prominent PAINS class are *p*-hydroxyarylsulfonamides that have redox activity<sup>33</sup> and the ability of covalent modification<sup>34</sup> and degradation.<sup>35</sup> The *p*-hydroxyarylsulfonamides typically only display assay interference if they contain a naphthalene core.<sup>35</sup> Figure 2b shows two examples of *p*-hydroxyarylsulfonamides from fragment-based drug design campaigns with and without a naphthalene substructure that have distinct binding modes. In a complex with heat shock protein 90,<sup>36</sup> one sulfonamide oxygen forms a hydrogen bond with Lys58. Alternatively, the phenolic hydroxyl group acts as an anchor in the Sec7 domain of cytohesin-2, a guanine nucleotide exchange factor,<sup>37</sup> where it forms hydrogen bonds with the carbonyl oxygen of Leu148 and the backbone amide of Leu153.

**Crucial Interactions Distinct from Interference Potential.** A number of cases were identified in which specific activity of PAINS substructures could be attributed to crystallographically observed interactions that were distinct from molecular origins of assay interference. For example, fluorescent dyes such as the 9-aminoacridine derivative shown in Figure 3a are potent DNA intercalating agents<sup>38</sup> that can act as topoisomerase inhibitors and mutagenic agents.<sup>39</sup> While the X-ray structure of its complex with DNA validates the contribution of the chromophoric acridine moiety to DNA intercalation, there is no doubt that photoinduction of 9-aminoacridines can cause false-positive signals in photometric and fluorometric assays.<sup>40</sup>

Catechols represent another prime PAINS motif that is not only present in synthetic compounds but also widely distributed among natural products.<sup>41</sup> The tendency of catechols to chelate metals, their redox activity, and reactivity against nucleophiles in oxidized form cause frequent assay artifacts.<sup>5,41</sup> However, catechols also have other potential activities. For example, norathyriol, the aglycon of the natural

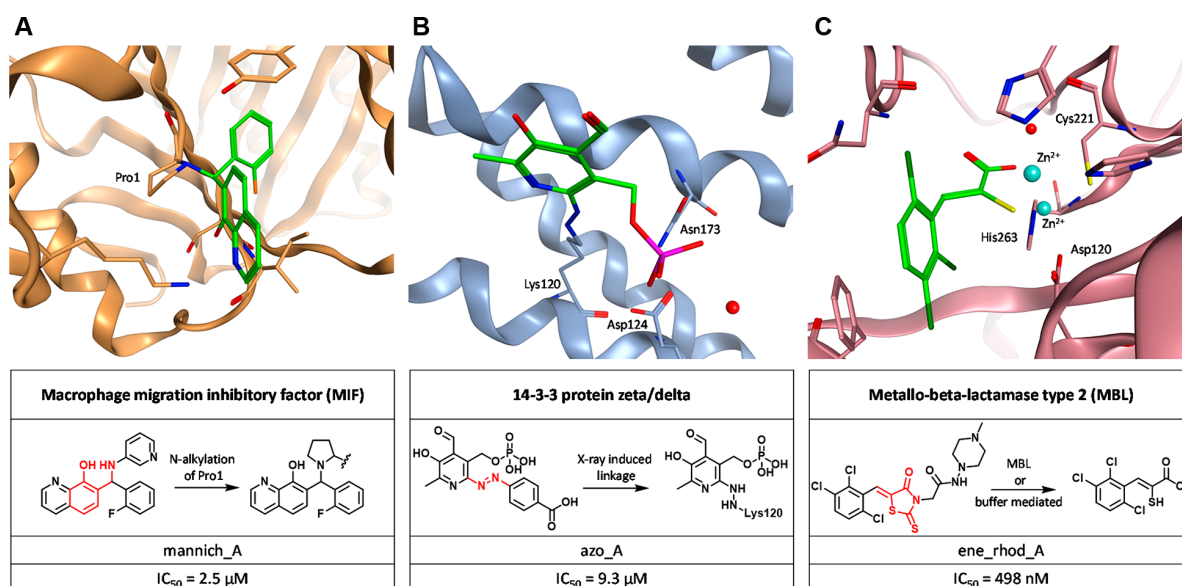


**Figure 3.** Interference motifs forming crucial interactions. (a) An aminoacridine derivative in complex with DNA (1KCI; 1.8 Å, 0.291, DNA carbon atoms colored gold). (b) Norathyriol bound to the ATP site of mitogen-activated protein kinase 1 (3SA0, 1.59 Å, 0.200).

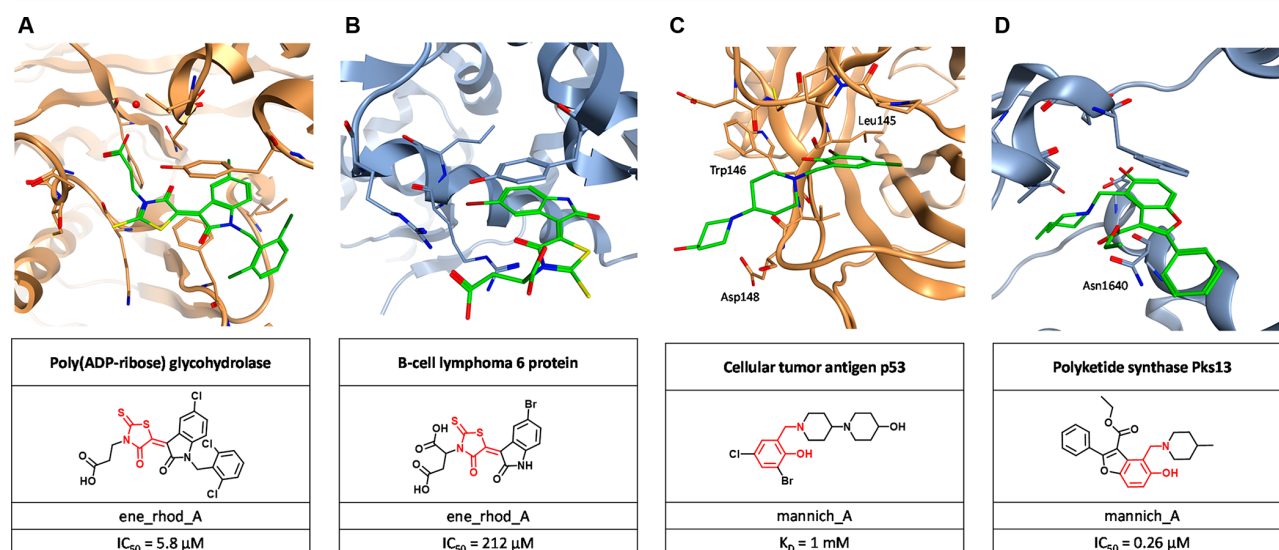
product mangiferin, is an ATP site-directed inhibitor of mitogen-activated protein kinase 1 (ERK2).<sup>42</sup> Figure 3b shows that the 1,2-dihydroxyphenyl group of the bound ligand forms three hydrogen bonds with the hinge region of ERK2 involving the side chain of Gln105 and the backbone carbonyl oxygens of Asp106 and Met108. Hence, specific interactions are responsible for an activity distinct from possible interference effects.

**Interference Reactions As a Mechanism Underlying Protein Binding.** Many interference mechanisms are reported to rely on complexation or covalent modification of target proteins.<sup>5,35,41</sup> Thus, it was particularly interesting to discover compounds that result from corresponding reactions and bind to different targets. For example, Figure 4a shows a covalently bound inhibitor of macrophage migration inhibitor factor (MIF).<sup>43,44</sup> This quinolol-type ligand contains a 2-hydroxybenzylamine substructure that has undesirable Mannich base reactivity leading to chelation of metal ions<sup>45</sup> or formation of reactive quinone methides, which are prone to unspecific nucleophilic attacks.<sup>46</sup> Furthermore, Cisneros et al. have shown that 2-hydroxybenzylamine containing compounds covalently inhibit MIF, giving rise to discrepancies in different assays.<sup>47</sup> In the MIF complex shown in Figure 4a, the formation of a Michael acceptor intermediate with subsequent nucleophilic attack of the Pro1 nitrogen generated the observed covalent protein–inhibitor complex. Hence, in this case, a Michael-type reaction led to modification of a PAINS motif and inhibition.

Depending on the assay setup, signals caused by compounds that might be colored should be carefully evaluated. Interestingly, for some colored compounds, additional interference mechanisms are reported. For instance, azo compounds are well-known dyes but may also undergo light- or X-ray-induced photolysis and react with biological nucleophiles, resulting in the formation of covalent bonds.<sup>48,49</sup> Such a bond was formed between an azo compound, a known protein–protein interaction inhibitor, and Lys120 of the 14-3-3 protein  $\zeta/\delta$ ,<sup>50</sup> shown in Figure 4b. Zhao et al. proposed that radiation exposure of this molecule caused diazene bond cleavage.<sup>50</sup> The resulting pyridoxal-



**Figure 4.** Interference mechanisms lead to binding. Shown are structures of complexes formed on the basis of PAINS reactivity. Reactions involving PAINS motifs that result in compounds forming specific interactions are summarized. (a) Covalent reaction of a Mannich base leads to alkylated macrophage migration inhibitor factor (3JSF, 1.93 Å, 0.218). (b) Cleavage of a diazene bond and subsequent formation of a covalent bond with 14-3-3 protein  $\zeta/\delta$  (3RDH, 2.39 Å, 0.287). (c) Generation of a zinc cation complexing metallo- $\beta$ -lactamase type 2 inhibitor following hydrolysis of a rhodanine (4PVO, 1.48 Å, 0.165).



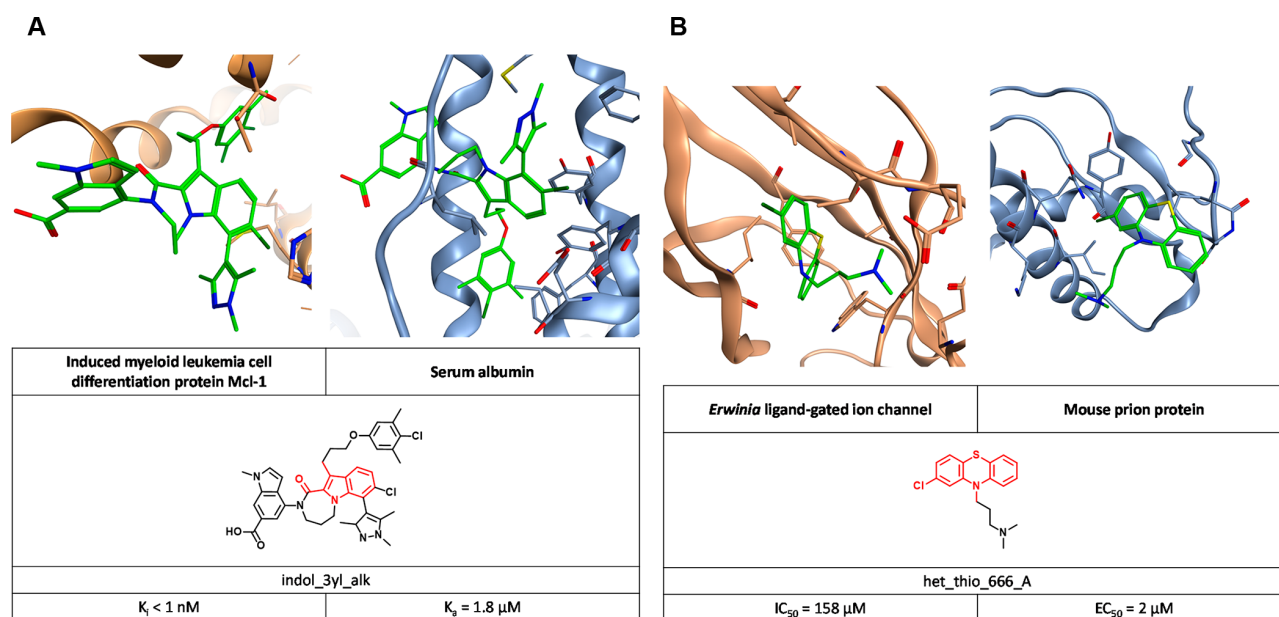
**Figure 5.** Structural context of PAINS motif. Examples of analogous PDB\_PAINS are shown in which the PAINS substructure is presented in a structural context that restricts reactivity, leading to the formation of specific interactions with different targets. (a) Two rhodanine derivatives containing an oxindole ring in complexes with poly(ADP-ribose) glycohydrolase (4EPQ) and B-cell lymphoma 6 protein (3LBZ, 2.3 Å, 0.269). (b) Mannich bases with a piperidinyl moiety in complexes with cellular tumor antigen p53 (5ABA, 1.62 Å, 0.203) and polyketide synthase Pks13 (5V3X, 1.94 Å, 0.247).

phosphate moiety was anchored by hydrogen-bond interactions with Asp124 and Asn173, correctly positioning the reactive nitrogen for a nucleophilic attack by residue Lys120.

Ring opening reactions and subsequent degradation also cause assay interference. For rationalizing such effects, the structure of a complex of metallo- $\beta$ -lactamase type 2 with an arylidene rhodanine inhibitor crystallized in a degraded state<sup>51</sup> was highly instructive, shown in Figure 4c. Following hydrolysis, the resulting thioenolate formed a network of specific interactions in the active site of  $\beta$ -lactamase. The thiol

function bridged the two zinc cations and the carboxylate moiety of Asp120, while the carboxylate oxygen of the ligand interacted with one of the zinc cations, the thiol group of Cys221, and the side chain nitrogen of His263. Hence, interference reactions might also lead to specific inhibition, which could be a more general way in which, for example, rhodanine-type ligands might interact with different targets.

*Exploring the Structural Context Hypothesis.* The analysis of analog series containing PAINS substructures has provided substantial support for the dependence of PAINS reactivity on



**Figure 6.** Activity of PDB\_PAINS across distinct target families. Examples of PDB\_PAINS are shown that form complexes with unrelated targets yielding different or similar binding modes. (a) An indole derivative bound to myeloid leukemia cell protein 1 (SIF4, 2.39 Å, 0.218) and serum albumin (SUJB, 2.7 Å, 0.215). The binding modes of the ligand differ. (b) Chlorpromazine, a phenothiazine derivative, in complexes with *Erwinia* ligand-gated ion channel (SLG3, 3.57 Å, 0.253) and a mouse prion protein (4MA8, 2.2 Å, 0.236), displaying similar binding modes.

the structural context in which they are presented.<sup>12</sup> This context hypothesis can also be investigated at the X-ray structural level. For example, Figure 5a shows two ene\_rhodanine-type ligands that were crystallized with poly(ADP-ribose) glycohydrolase (PARG)<sup>52,53</sup> and B cell lymphoma 6 protein (BCL6).<sup>54</sup> Both inhibitors were highly similar and shared a central unsaturated rhodanine heterocycle that was substituted with an aliphatic carboxyl moiety at the nitrogen and an oxindole ring at the unsaturated carbon. Given this structural context, the rhodanine heterocycle was not reactive and contributed to binding. In both complexes, the inhibitors were sandwiched via  $\pi$ - $\pi$  stacking interactions between the oxindole function and corresponding tyrosine and phenylalanine residues. The rhodanine ring was deeply bound in the ADP binding region of PARG and the lateral groove of BCL6, respectively.

Figure 5b shows complexes of cellular tumor antigen p53 and polyketide synthase PKs13 with phenolic Mannich bases possessing a tertiary amine as a part of an aliphatic piperidine ring system. In these structures, the positively charged tertiary amine contributes to charge assisted hydrogen bonds with aspartic acid and asparagine, respectively. In the p53 complex, the hydroxyphenyl moiety also contributes to backbone interactions with corresponding leucine and tryptophan moieties, while the hydroxypiperidin-1-yl ring remains solvent exposed. Interference mechanisms associated with phenolic Mannich bases such as chelation or the formation of reactive quinone methides depend on steric properties and the ability of the amine to act as a leaving group.<sup>45,46</sup> However, this ability was restricted in these cases. Therefore, different from other phenolic Mannich bases, these compounds were not reactive and interacted with both targets.

**Multitarget Activity Across Different Families.** Previously, compounds forming X-ray complexes with targets from different families were identified.<sup>55</sup> We have also searched for PDB\_PAINS bound to targets from distinct families. For 40

PDB\_PAINS, complexes with members of unrelated protein families were identified. Among these were 3-alkylindoles that may have Michael-type reactivity.<sup>4</sup> The indole derivative shown in Figure 6a was crystallized with myeloid cell leukemia protein 1 (Mcl-1).<sup>56</sup> In this compound, the PAINS motif is part of a tricyclic indole lactam that positions two ring systems for specific interactions in its bound conformation. Subsequently, binding of this compound to human serum albumin binding (HSA) was investigated.<sup>57</sup> Here, the rigid tricyclic indole lactam is tightly bound to drug site 3 of HSA. The binding modes of the ligand and its interaction patterns clearly differ for the unrelated targets. Clearly, the observed promiscuity of this indole derivative was not related to possible assay interference mechanisms, but was due to specific interactions in distinct binding sites resulting from the given structural embedding of the PAINS motif.

Figure 6b compares X-ray structures of two complexes involving the phenothiazine derivative chlorpromazine, which is typically used as a probe for nicotinic acetylcholine receptors. However, the structure of chlorpromazine in complex with *Erwinia* ligand-gated ion channel revealed binding of the ligand to an allosteric site.<sup>58</sup> Furthermore, a structure of chlorpromazine in complex with unrelated mouse prion protein was also available.<sup>59,60</sup> Interestingly, the binding modes of chlorpromazine were similar in these cases, although the protein environments and binding sites were distinct. In both cases, the phenothiazine moiety served as a hydrophobic anchor in lipophilic regions of the binding sites.

Examples such as the complex structures containing the 3-alkylindole derivative or chlorpromazine confirm multitarget activity of ligands with PAINS substructures and thus rationalize the promiscuity of these compounds.



## CONCLUDING DISCUSSION AND SUMMARY

In this study, we have systematically searched for and examined X-ray structures of target–ligand complexes with assay interference compounds. Our reasoning was that X-ray structures might provide most detailed information as to how compounds classified as PAINS might interact with targets, by desirable or undesirable mechanisms. Therefore, while very few individual examples of crystallized PAINS ligands have been noted,<sup>47,61</sup> we have carried out a first systematic analysis of PAINS in X-ray structures. A large number of crystallographic ligands containing PAINS substructures were identified. Careful analysis of these X-ray structures revealed different mechanisms of action for PAINS. In some cases, prime candidates for assay interference due to potential reactivity were found to form specific interactions with targets. In others, PAINS-containing ligands formed interactions that were not related to likely causes of assay artifacts. Moreover, in other instances, PAINS reactivity was responsible for the formation of specific target–ligand complexes. Finally, a number of PDB\_PAINS also bound to unrelated targets, displaying either different or similar binding modes. These findings provided a rationale for promiscuity of such compounds as a consequence of specific interactions with multiple targets.

As discussed, potential PAINS reactivity does not imply that covalent complexes must be formed. We also note that compound potency cannot be unambiguously correlated with interactions seen in X-ray structures because potency results from multiple effects including desolvation and entropic contributions. This also needs to be taken into consideration in interaction analysis.

On the basis of structural data, a differentiated picture of possible PAINS effects was obtained. Clearly, the formation of specific target–ligand interactions and the possibility of assay artifacts were not mutually exclusive. Small chemical modifications of PAINS substructures often restricted undesirable reactivity and enabled binding interactions, which provided direct support for structural context dependence of PAINS reactivity.

Together with the structural context dependency of PAINS activity,<sup>12</sup> the multifaceted picture of PAINS actions revealed at the level of complex X-ray structures reinforces the use of PAINS substructure filters, which are controversially viewed,<sup>10,62,63</sup> as an initial indicator of assay interference potential. PAINS filters cannot detect compounds that are certain to elicit artifacts.<sup>10,11</sup> Rather, they provide valuable structural alerts to raise awareness of potential caveats and prioritize candidate compounds for follow-up investigations,<sup>11</sup> which is well in accord with recommendations of leading journals in medicinal chemistry and related fields.<sup>7</sup>

In summary, more than 1000 PDB\_PAINS were detected in nearly 3000 X-ray structures of ligand–target complexes. Potentially reactive compounds were frequently found to engage specific interactions with variety of targets. On the basis of X-ray data, most complexes with PDB\_PAINS were noncovalent, but PAINS reactivity also resulted in the formation of complexes including covalent inhibition. A confined number of PDB\_PAINS was found to form complexes with unrelated targets adopting different binding modes. Furthermore, X-ray structures of noncovalent complexes with PDB\_PAINS provided further evidence for different embedding of PAINS-defining substructures in ligands restricting intrinsic reactivity but not precluding binding events.

Moreover, many PDB\_PAINS interacted with targets by mechanisms that were distinct from confirmed or assumed interference effects, demonstrating that specific binding and assay interference were not mutually exclusive.

As an outlook, the unexpected wealth of structural data available for ligands with PAINS substructures opens the door for further mechanism of action investigations of PAINS and other interference compounds. To these ends, PDB\_PAINS we identified are made freely available together with links to the X-ray structures from which they originated.<sup>64</sup>

## AUTHOR INFORMATION

### Corresponding Author

\*Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de).

### ORCID

Michael Gütschow: 0000-0002-9376-7897

Jürgen Bajorath: 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.

## ABBREVIATIONS USED

PAINS, pan-assay interference compounds; PDB, protein data bank

## REFERENCES

- (1) McGovern, S. L.; Caselli, E.; Griorieff, 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) 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.
- (5) Baell, J. B.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.
- (6) 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.
- (7) 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. Med. Chem.* **2017**, *60*, 2165–2168.
- (8) 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.
- (9) Gilberg, E.; Stumpfe, D.; Bajorath, J. Towards a Systematic Assessment of Assay Interference: Identification of Extensively Tested Compounds with High Assay Promiscuity. *F1000Research* **2017**, *6*, 1505.
- (10) 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.
- (11) 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.
- (12) Gilberg, E.; Stumpf, D.; Bajorath, J. Activity Profiles of Analog Series Containing Pan Assay Interference Compounds. *RSC Adv.* **2017**, *7*, 35638–35647.
- (13) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155.
- (14) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (15) *Molecular Operating Environment (MOE)*, version 2014.09; Chemical Computing Group ULC: Montreal, QC, Canada, 2014.
- (16) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.
- (17) UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
- (18) James, C. A.; Weininger, D.; Delany, J. *SMARTS Theory. Daylight Theory Manual*; Daylight Chemical Information Systems: Laguna Niguel, CA, 2000.
- (19) 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.
- (20) Saubern, S.; Guha, R.; Baell, J. B. A KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol. Inf.* **2011**, *30*, 847–850.
- (21) RDKit: *Chemoinformatics and Machine Learning Software*, Landrum, G., 2013.
- (22) Sterling, T.; Irwin, J. J. ZINC 15-Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
- (23) OEChem, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM.
- (24) RStudio: *Integrated Development Environment for R*, RStudio, Inc.: Boston, MA, 2015.
- (25) 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.
- (26) 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.
- (27) Carlson, E. E.; May, J. F.; Kiessling, L. L. Chemical Probes of UDP-Galactopyranose Mutase. *Chem. Biol.* **2006**, *13*, 825–837.
- (28) 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.; Strzemiński, 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-R Antagonists. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 533–538.
- (29) 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.
- (30) Tomasić, T.; Masic, L. P. Rhodanine as a Privileged Scaffold in Drug Discovery. *Curr. Med. Chem.* **2009**, *16*, 1596–1629.
- (31) Georgescu, R. E.; Yurieva, O.; Seung-Sup, K.; Kuriyan, J.; Kong, X.; O'Donnell, M. Structure of a Small-Molecule Inhibitor of a DNA Polymerase Sliding Clamp. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 11116–11121.
- (32) Tomasić, T.; Zidar, N.; Sink, R.; Kovac, A.; Blanot, D.; Contreras-Martel, C.; Dessen, A.; Muller-Premru, M.; Zega, A.; Gobec, S.; Kikelj, D.; PeterlinMasic, L. Structure-Based Design of a New Series of D-Glutamic Acid Based Inhibitors of Bacterial UDP-N-Acetylmuramoyl-L-alanine:D-glutamate Ligase (MurD). *J. Med. Chem.* **2011**, *54*, 4600–4610.
- (33) 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 H<sub>2</sub>O<sub>2</sub> by Redox Cycling in Reducing Environments. *Assay Drug Dev. Assay Drug Dev. Technol.* **2010**, *8*, 152–174.
- (34) McCallum, M. M.; Nandhikonda, P.; Temmer, J. J.; Eyermann, C.; Simeonov, A.; Jadhav, A.; Yasgar, A.; Maloney, D.; Arnold, L. A. 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.
- (35) 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.
- (36) Roughley, S. D.; Hubbard, R. E. How Well Can Fragments Explore Accessed Chemical Space? A Case Study from Heat Shock Protein 90. *J. Med. Chem.* **2011**, *54*, 3989–4005.
- (37) Rouhana, J.; Hoh, F.; Estaran, S.; Henriquet, C.; Boublik, Y.; Kerkour, A.; Trouillard, R.; Martinez, J.; Pugnieri, M.; Padilla, A.; Chavanieu, A. Fragment-Based Identification of a Locus in the Sec7 Domain of Arno for the Design of Protein-Protein Interaction Inhibitors. *J. Med. Chem.* **2013**, *56*, 8497–8511.
- (38) Adams, A.; Guss, M. J.; Denny, W. A.; Wakelin, L. P. G. Crystal Structure of 9-Amino-N-[2-(4-morpholinyl)ethyl]-4-acridinecarboxamide Bound to d(CGTAACG)<sub>2</sub>: Implications for Structure-Activity Relationships of Acridinecarboxamide Topoisomerase Poisons. *Nucleic Acids Res.* **2002**, *30*, 719–725.
- (39) Turhan, K.; Ozturkcan, S. A.; Turgut, Z.; Karadayi, M.; Gulluce, M. Inhibition of the Mutagenic Effects of N-Methyl-N'-nitro-N-nitrosoguanidine and 9-Aminoacridine by Indenopyridines in the *Salmonella typhimurium* Tester Strain 1537 and *E. coli*. *Drug Chem. Toxicol.* **2014**, *37*, 365–369.
- (40) Manivannan, C.; Sundaram, K. M.; Renganathan, R.; Sundararaman, M. Investigations on Photoinduced Interaction of 9-Aminoacridine with Certain Catechols and Rutin. *J. Fluoresc.* **2012**, *22*, 1113–1125.
- (41) 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.
- (42) Li, J.; Malakhova, M.; Mottamal, M.; Reddy, K.; Kurinov, I.; Carper, A.; Langfald, A.; Oi, N.; Kim, N. O.; Zhu, F.; Sosa, C. P.; Zhou, K.; Bode, A. M.; Dong, Z. Norathyriol Suppresses Solar UV-Induced Skin Cancer by Targeting ERKs. *Cancer Res.* **2012**, *72*, 260–270.
- (43) McLean, L. R.; Zhang, Y.; Li, H.; Lukasczyk, U.; Choi, Y. M.; Han, Z.; Prisco, J.; Fordham, J.; Tsay, J. T.; Reiling, S.; Vaz, R. J.; Li, Y. Discovery of Covalent Inhibitors for MIF Tautomerase via Cocrystal Structures with Phantom Hits from Virtual Screening. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 6717–6720.
- (44) Cournia, Z.; Leng, L.; Sunilkumar, G.; Du, X.; Bucala, R.; Jorgensen, W. L. Discovery of Human Macrophage Migration Inhibitory Factor (MIF)-CD74 Antagonists via Virtual Screening. *J. Med. Chem.* **2009**, *52*, 416–424.
- (45) Herzig, Y.; Lerman, L.; Goldenberg, W.; Lerner, D.; Gottlieb, H. E.; Nudelman, A. Hydroxy-1-aminoindans and Derivatives: Preparation, Stability, and Reactivity. *J. Org. Chem.* **2006**, *71*, 4130–4140.
- (46) Young, R. H.; Brewer, D.; Kayser, R.; Martin, R.; Feriozi, D.; Keller, R. A. On the Mechanism of Quenching by Amines: A New Method for Investigation of Interactions with Triplet States. *Can. J. Chem.* **1974**, *52*, 2889–2893.
- (47) Cisneros, J. A.; Robertson, M. J.; Valhondo, M.; Jorgensen, W. L. Irregularities in Enzyme Assays: The Case of Macrophage Migration Inhibitory Factor. *Bioorg. Med. Chem. Lett.* **2016**, *26*, 2764–2767.
- (48) Hoijsenberg, P. A.; Karlen, S. D.; Sanrame, C. N.; Aramedia, P. F.; Garcia-Garibay, M. A. Photolysis of an Asymmetrically Substituted

Diazene in Solution and in the Crystalline State. *Photochem. Photobiol.* **2009**, *8*, 961–969.

(49) Boulegue, C.; Loeweneck, M.; Renner, C.; Moroder, L. Redox Potential of Azobenzene as an Amino Acid Residue in Peptides. *ChemBioChem* **2007**, *8*, S91–S94.

(50) Zhao, J.; Du, Y.; Horton, J. R.; Upadhyay, A. K.; Lou, B.; Bai, Y.; Zhang, X.; Du, L.; Li, M.; Wang, B.; Zhang, L.; Barbieri, J. T.; Khuri, F. R.; Cheng, X.; Fu, H. Discovery and Structural Characterization of a Small Molecule 14–3-3 Protein-Protein Interaction Inhibitor. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16212–16216.

(51) Brem, J.; van Berkel, S. S.; Aik, W.; Rydzik, A. M.; Avison, M. B.; Pettinati, I.; Umland, K.; Kawamura, A.; Spencer, J.; Claridge, T. D. W.; McDonough, M. A.; Schofield, C. J. Rhodanine Hydrolysis Leads to Potent Thioenolate Mediated Metallo- $\beta$ -Lactamase Inhibition. *Nat. Chem.* **2014**, *6*, 1084–1090.

(52) Dunstan, M. S.; Barkauskaite, E.; Lafite, P.; Knezevic, C. E.; Brassington, A.; Marijan, A.; Hergenrother, P. J.; Leys, D.; Ahel, I. Structure and Mechanism of a Canonical Poly(ADP-ribose) Glycohydrolase. *Nat. Commun.* **2012**, *3*, 878–884.

(53) Finch, K. E.; Knezevic, C. E.; Nottbohm, A. C.; Partlow, K. C.; Hergenrother, P. J. Selective Small Molecule Inhibition of Poly(ADP-Ribose) Glycohydrolase (PARG). *ACS Chem. Biol.* **2012**, *7*, 563–570.

(54) Cerhietti, L. C.; Ghetu, A. F.; Zhu, X.; Da Silva, G. F.; Shijun, Z.; Matthews, M.; Bunting, K. L.; Polo, J. M.; Farés, C.; Arrowsmith, C. H.; Yang, S. N.; Garcia, M.; Coop, A.; MacKerell, D., Jr; Privé, G. G.; Melnick, A. A Small Molecule Inhibitor of BCL6 Kills DLBCL Cells in Vitro and in Vivo. *Cancer Cell* **2010**, *17*, 400–411.

(55) Gilberg, E.; Stumpfe, D.; Bajorath, J. X-Ray Structure Based Identification of Compounds with Activity against Targets from Different Families and Generation of Templates for Multi-Target Ligand Design. *ACS Omega* **2018**, *3*, 106–111.

(56) Lee, T.; Bian, Z.; Zhao, B.; Hogdal, L. J.; Sensintaffar, J. L.; Goodwin, C. M.; Belmar, J.; Shaw, S.; Tarr, J. C.; Veerasamy, N.; Matulis, S. M.; Koss, B.; Fischer, M. A.; Arnold, A. L.; Camper, D. V.; Browning, C. F.; Rossanese, O. W.; Budhraj, A.; Opferman, J.; Boise, L. H.; Savona, M. R.; Letai, A.; Olejniczak, E. T.; Fesik, S. W. Discovery and Biological Characterization of Potent Myeloid Cell Leukemia-1 Inhibitors. *FEBS Lett.* **2017**, *591*, 240–251.

(57) Zhao, B.; Sensintaffar, J.; Bian, Z.; Belmar, J.; Lee, T.; Olejniczak, E. T.; Fesik, S. W. Structure of a Myeloid Cell Leukemia-1 (Mcl-1) Inhibitor Bound to Drug Site 3 of Human Serum Albumin. *Bioorg. Med. Chem.* **2017**, *25*, 3087–3092.

(58) Nys, M.; Wijckmans, E.; Farinha, A.; Yoluk, Ö.; Andersson, M.; Brams, M.; Spurny, R.; Peigneur, S.; Tytgat, J.; Lindahl, E.; Ulens, C. Allosteric Binding Site in a Cys-Loop Receptor Ligand-Binding Domain Unveiled in the Crystal Structure of ELIC in Complex with Chlorpromazine. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E6696–E6703.

(59) Baral, P. K.; Swayampakula, M.; Rout, M. K.; Kav, N. N.; Spyrapoulos, L.; Aguzzi, A.; James, M. N. Structural Basis of Prion Inhibition by Phenothiazine Compounds. *Structure* **2014**, *22*, 291–303.

(60) Vogtherr, M.; Grimme, S.; Elshorst, B.; Jacobs, D. M.; Fiebig, K.; Griesinger, C.; Zahn, R. Antimalarial Drug Quinacrine Binds to C-Terminal Helix of Cellular Prion Protein. *J. Med. Chem.* **2003**, *46*, 3563–3564.

(61) 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.

(62) Lagorce, D.; Oliveira, N.; Miteva, M. A.; Villoutreix, B. O. Pan-Assay Interference Compounds (PAINS) That May Not Be Too Painful for Chemical Biology Projects. *Drug Discovery Today* **2017**, *22*, 1131–1133.

(63) Kenny, P. W. Comment on The Ecstasy and Agony of Assay Interference Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2640–2645.

(64) X-Ray Structures of Target–Ligand Complexes Containing Compounds with Assay Interference Potential; <https://www.zenodo.org/record/1144030>, accessed January 10, 2018.



## Conclusions

By taking X-ray structure data into account, a further dimension for the analysis of interference compounds was explored. In total, 70 different PAINS classes were represented by 1107 unique ligands found in 2874 crystallographic complexes, including prominent representatives such as rhodanines, catecholes, or Mannich bases. Analysis of this complexes allowed us to reveal different modes of action.

Crucial ligand interactions were often found to be distinct from interference potential and PAINS substructures contributed to specific target interactions. In some instances, PAINS reactivity was even responsible for the formation of complexes, for example, by covalent modification of target proteins. We also identified a series of X-ray structures, in which chemical modifications of reactive PAINS moieties restricted their ability to produce undesired chemical reactions. This finding suggested that the structural context in which PAINS are represented may play an important role for eliciting false-positive activities. The set of PAINS crystallographic ligands is made freely available.

The results of this work demonstrate the multifaceted picture of PAINS activities and provide a novel reference frame for the judgment of interference potential and multitarget activities. Following the structural context hypothesis, activity profiles of analog series containing PAINS are systematically identified and analyzed in the next chapter. This allows us to draw structure-activity relationships of interference mechanisms based on the structural embedding of PAINS substructures.



# 4 Activity Profiles of Analog Series Containing Pan-Assay Interference Compounds

## Introduction

Although there is no doubt that many PAINS generate false-positive activity data, the application of structural alerts should not result in automatically discarding potential candidates for medicinal chemistry. For example, systematic analyses of PAINS activity data have shown that interference compounds are often specifically active or consistently inactive. Moreover, X-ray structures of protein-PAINS complexes demonstrate that prominent PAINS engage in specific binding interactions and that interference mechanisms often provide the basis for target interactions. PAINS alerts intrinsically do not represent the entire structure of a compound. Therefore, the aforementioned findings provide evidence that the structural embedding of PAINS substructures plays a decisive role for assay interference.

Herein, the structural context of PAINS is explored by performing a systematic analysis of analog series containing assay interference compounds. Activity profiles of extensively tested PAINS and their structural analogs are compared.

My main contribution to this work was the generation of analogs series, identification of activity profiles, and the structural context analysis of PAINS activities.

Reprinted with permission from 'E. Gilberg, D. Stumpfe, J. Bajorath. Activity Profiles of Analog Series Containing Pan Assay Interference Compounds. *RSC Advances* **2017**, *7*, 35638–35647.' Copyright 2017 Royal Society of Chemistry





Cite this: *RSC Adv.*, 2017, 7, 35638

## Activity profiles of analog series containing pan assay interference compounds

Erik Gilberg, Dagmar Stumpfe and Jürgen Bajorath \*

Activity artifacts in assays present a major problem for biological screening and medicinal chemistry. Such artifacts are often caused by compounds that form aggregates or are reactive under assay conditions. Many pan assay interference compounds (PAINS) have been proposed to cause false-positive assay readouts. PAINS are typically contained as substructures in larger molecules. They are used as computational filters to detect compounds with potential chemical liabilities. Recent studies have shown that molecules containing the same PAINS substructure often have greatly varying hit rates in screening assays. Even the overall most frequently active PAINS substructures are found in compounds that are only rarely active or consistently inactive in many assays they are tested in. These observations suggest that the structural context in which PAINS are presented may play an important role for eliciting false-positive activities. However, this assumption remains to be investigated. Herein, we report the systematic identification of analog series of screening compounds that contain PAINS or exclusively consist of PAINS and the analysis of their activity profiles. Comparison of analogs or different series of analogs containing the same PAINS substructure provides structural context information. For many PAINS, extensively tested series with distinct activity profiles were detected. Furthermore, analogs within the same series often displayed significant differences in hit rates. The analog series reported herein organize PAINS in different structural contexts. Their activity profiles provide many opportunities for experimental follow-up investigations to better understand PAINS characteristics.

Received 16th June 2017  
Accepted 11th July 2017

DOI: 10.1039/c7ra06736d

rsc.li/rsc-advances

### Introduction

Activity artifacts in biological screening assays can be caused by compounds that are prone to colloidal aggregation<sup>1,2</sup> or that are chemically reactive under assay conditions.<sup>3,4</sup> A variety of mechanisms may lead to apparent inhibition and false-positive signals including, among others, fluorescence of small molecules, redox reactivity, or covalent modifications of target proteins.<sup>4-6</sup> Compounds with assay interference potential originate from both synthetic and natural sources<sup>7</sup> and include molecules that are intensely investigated in pharmaceutical research.<sup>8</sup>

There is no doubt that assay artifacts compromise medicinal chemistry programs and that false-positive activities cumulate in the scientific literature. This situation has triggered community efforts to raise awareness of assay interference.<sup>9</sup> Since it often remains unclear if a compound causes an artificial activity signal, careful experimental follow-up studies are required.<sup>2,9</sup> One way to proactively address this problem is the search for potential interference compounds

that require special attention if they are found to be active in assays.<sup>10</sup>

In a landmark study, 480 chemical classes have been put forward as candidates for assay interference.<sup>3</sup> To these ends, limited numbers of compounds were tested in AlphaScreen assays.<sup>3</sup> This set of so-called pan assay interference compounds (PAINS)<sup>3</sup> contains many small reactive chemical entities that often occur as substructures in larger molecules. While it cannot be expected that PAINS cover the entire spectrum of possible interference mechanisms, their identification has made it possible to implement substructure filters to flag potential interference compounds,<sup>3,10</sup> an important step toward the identification of questionable candidates.

However, the predictive value of PAINS filters has also been called into question, given that for many of the proposed structures only limited experimental support was available.<sup>11</sup> In general, although assay artifacts are a problematic issue, excluding any potentially reactive compound from further consideration would not be justifiable scientifically. Overestimating the magnitude of assay interference may lead to disregarding compounds that have desired activities and/or act by novel mechanisms.

Two recent studies, have systematically evaluated the activity of PAINS on the basis of publicly available screening data<sup>11,12</sup>

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



and other compound sources.<sup>11</sup> Both investigations revealed substantial heterogeneity in PAINS activities and greatly varying hit rates. Furthermore, many rarely active or consistently inactive molecules with PAINS substructures were detected.<sup>11,12</sup> While small subsets of PAINS including, for example, quinones, catechols, rhodanines, or Mannich bases often represented highly active compounds, most likely causing artifacts, other classes of PAINS did not display unusual hit rates. Moreover, even the most frequently active PAINS were also found in many consistently inactive compounds. Taken together, these observations indicated that the molecular environment<sup>11</sup> or structural context<sup>12</sup> in which PAINS are presented might play an important role for their ability to elicit desirable activities or artifacts. However, little has been done so far to address the question how structural embedding might modulate PAINS activity.

Therefore, we have carried out a systematic analysis of analog series containing PAINS, which provide structural context information. Analog series were systematically extracted from screening compounds. For series of extensively assayed PAINS, activity profiles were determined and studied in detail, yielding first insights into structural context-dependent modulation of PAINS actions. The results of our analysis are presented in the following.

## Methods and materials

### Compound activity data

A subset of 437 257 screening compounds from PubChem BioAssays<sup>13</sup> that were tested in primary assays (percentage of inhibition from a single dose) and confirmatory assays (dose-response assays yielding  $IC_{50}$  values)<sup>14</sup> provided our starting point. PubChem compounds for which data from both primary and confirmatory assays are available have usually been frequently tested. Hence, most of the pre-selected molecules were evaluated in more than 50 assays. For our analysis, only the most extensively assayed compounds were considered. Therefore, the global distribution of the number of assays in which the pre-selected compounds were tested was determined. Fig. 1a shows this distribution in a boxplot format. PubChem compounds that were tested in more than 257 primary assays, corresponding to the lower quartile boundary of the distribution, were selected for our analysis, yielding a total of 327 523 compounds.

### Identification of analog series

From these 327 523 compounds, analog series (ASs) were systematically extracted using a recently developed methodology,<sup>15</sup> which is based upon the matched molecular pair (MMP) formalism.<sup>16</sup> MMPs are pairs of compounds that are only distinguished by a chemical change at a single site,<sup>16</sup> often termed a chemical transformation.<sup>17</sup> To generate MMPs, exocyclic single bonds in screening compounds were systematically fragmented following retrosynthetic fragmentation rules,<sup>18</sup> yielding RECAP-MMPs.<sup>19</sup> Previously established transformation size restrictions were introduced to limit

transformations in MMPs to chemical modifications typically observed in analogs.<sup>20</sup> Once all possible RECAP-MMPs were generated, a global MMP network was constructed in which compounds were represented as nodes and edges accounted for pairwise MMP relationships. In this network representation, ASs form disjoint (isolated) clusters.<sup>15</sup> Each cluster contains all possible MMP relationships within a series, which cover all substitution sites and available R-groups. For 190 612 of the 327 523 extensively assayed compounds, analog relationships were detected, resulting in the formation of 34 300 individual clusters and ASs.

For each of the 34 300 ASs, assay and target information was compiled. For each AS, assay overlap was determined as the number of assays shared by all analogs. In addition, for pairwise comparison of ASs, the overlap was calculated as the number of assays common to both series.

### Hit rate intervals and activity profiles

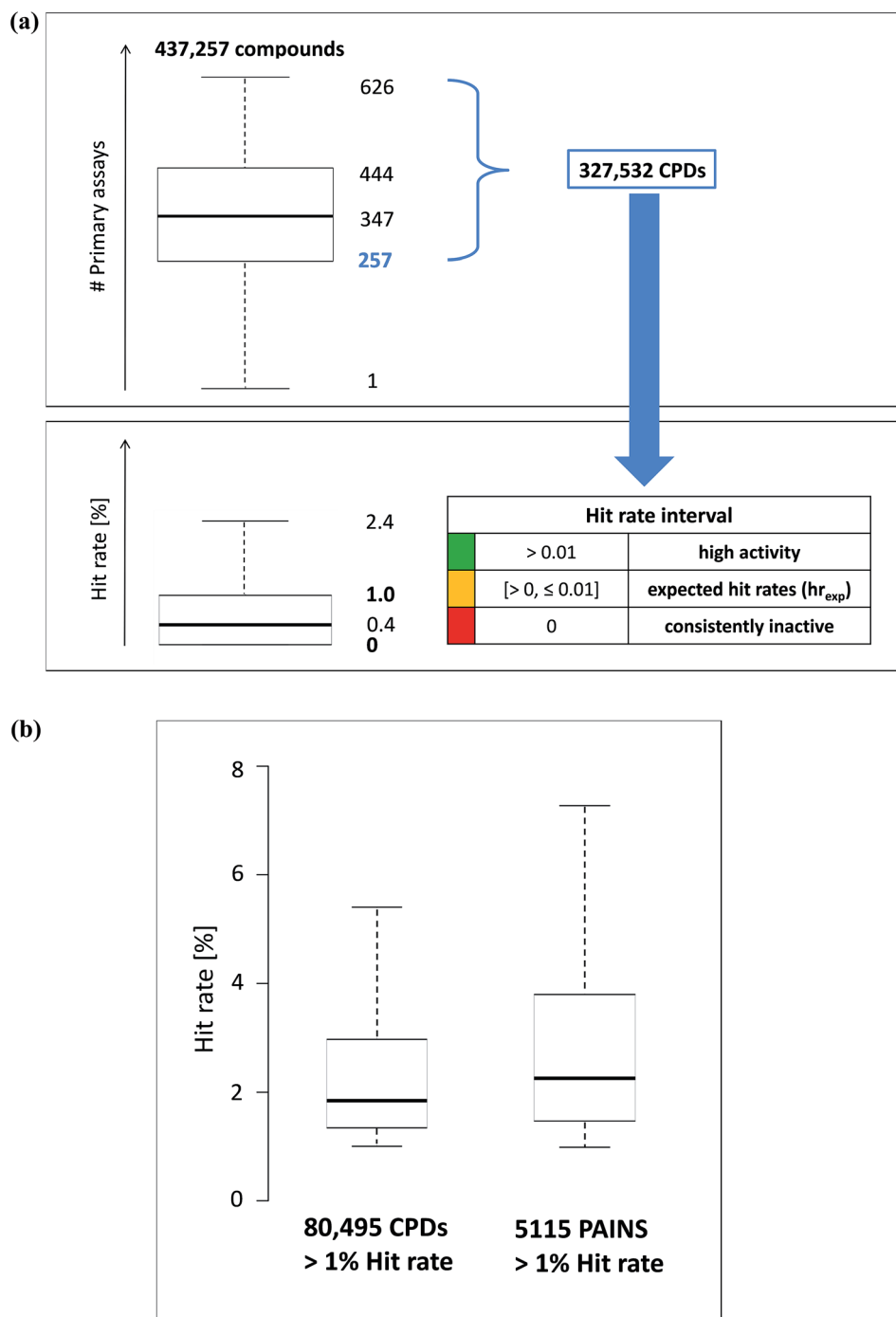
The hit rate of a compound was conventionally defined as the fraction of assays in which it was active. The distribution of hit rates over all compounds was captured in a boxplot yielding a median value of 0.4% (Fig. 1a). On the basis of this distribution, the interval of expected hit rates ( $hr_{exp}$ ) for active PubChem screening compounds was defined as  $0\% < hr_{exp} \leq 1.0\%$  covering the lower quartile, median, and upper quartile. Accordingly, hit rates exceeding 1.0% (upper whisker and outliers) were considered high. The lower whisker and lower quartile boundary of the boxplot were identical and represented consistently inactive compounds. Thus, activity profiles were defined on the basis of three hit rate intervals including consistent inactivity (0%), expected or average hit rates ( $0\% < hr_{exp} \leq 1.0\%$ ), and high hit rates ( $>1.0\%$ ) (Fig. 1a). Given that qualifying compounds were tested in at least 258 assays, high hit rates corresponded to activity in a minimum of three assays, while expected hit rates of active compounds corresponded to activity in one or two assays. Hence, as defined, the interval of high rates predominantly focused on promiscuous compounds. Apparent promiscuity might result from true multi-target activities or assay artifacts. The distribution of hit rates exceeding 1.0% was also monitored in boxplots for screening compounds that did not contain PAINS substructures (non-PAINS) and PAINS substructures (Fig. 1b).

The activity profile of an AS was then generated by combining hit rates of all participating analogs, as illustrated in Fig. 2.

### Detection of pan assay interference compounds

Analog series were screened *in silico* for PAINS using three public filters available in ChEMBL (481 substructures),<sup>21</sup> RDKit (480),<sup>22</sup> and ZINC (480).<sup>23</sup> For screening compounds, canonical SMILES representations<sup>24</sup> were generated. Compounds were classified as PAINS if a PAINS substructure was detected by at least one of the three filters (considering possible implementation discrepancies of substructure strings). Filtering identified 177 different PAINS substructures in 3473 ASs.





**Fig. 1** Assay frequency and hit rate distribution. (a) At the top, a boxplot shows the primary assay frequency distribution for 437 257 pre-selected PubChem compounds. Only compounds tested in more than 257 primary assays (lower quartile boundary) were considered for further analysis. At the bottom, the hit rate distribution for these 327 532 compounds is shown in another boxplot on the basis of which hit rate intervals were defined, as detailed in the text. In (b), the hit rate distribution of compounds with hit rates above 1% is shown in boxplots for screening compounds without PAINS substructures (non-PAINS, left) and PAINS (right).

All calculations were performed using in-house Java and R scripts with the aid of KNIME<sup>25</sup> protocols, the OpenEye<sup>26</sup> chemistry toolkit, and RStudio.<sup>27</sup>

### Control calculations

As a control, the analysis was repeated for ASs originating from compounds tested in 65-247 confirmatory assays. In

this case, 3459 ASs with PAINS substructures were identified, 1865 of which exclusively consisted of PAINS. The analysis of the activity profiles of this set of series yielded results that were readily comparable to those obtained for ASs originating from primary assays. In the following, we therefore concentrate on the results for ASs from primary assays.



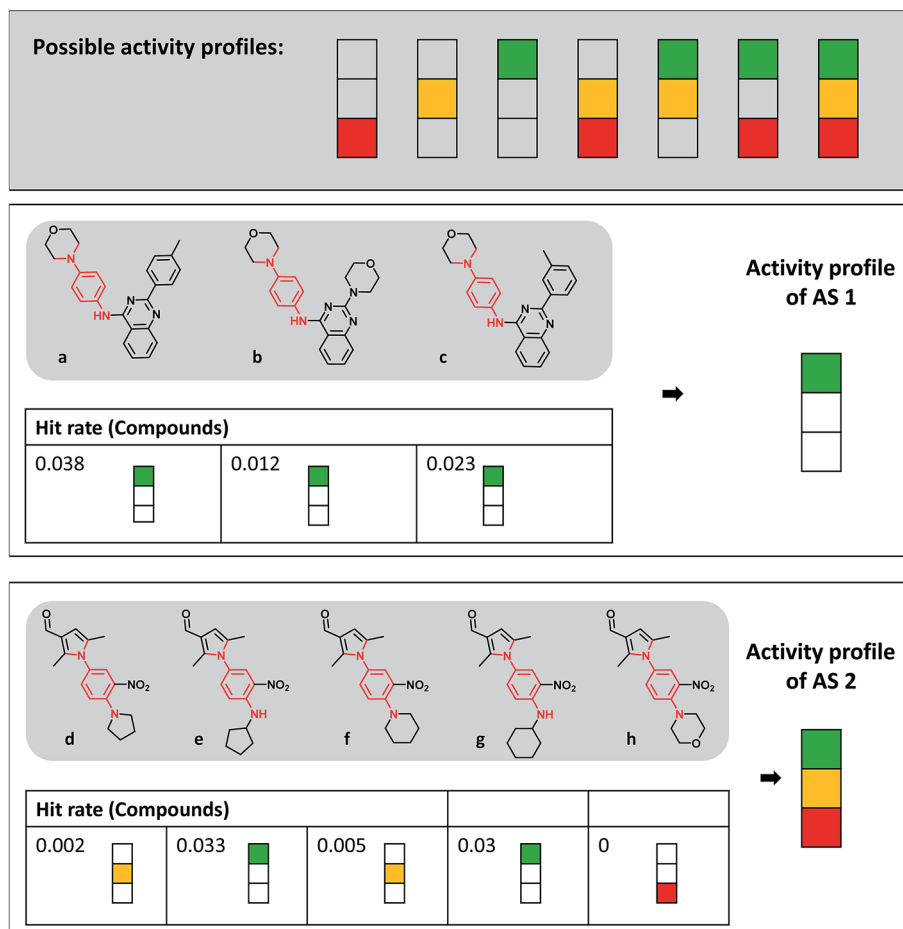


Fig. 2 Activity profiles and exemplary analog series. At the top, all possible activity profiles are displayed that represent different combinations of the three hit rate intervals according to Fig. 1a (consistently inactive, red; expected hit rates, yellow; high hit rates, green). Below the profiles, compounds forming two different ASs containing the same PAINS substructure (red) are shown. For each analog, the hit rate and corresponding interval are given and the resulting activity profile of the series is displayed.

## Results and discussions

### Analog series with PAINS

For 190 612 of 327 532 PubChem compounds tested in at least 257 primary assays, analog relationships were identified, yielding a total of 34 300 ASs. Compound and AS statistics are reported in Fig. 3 (bottom). PAINS were detected in 13 018 compounds from 3473 ASs. More than half of these ASs, *i.e.* 1876 series comprising 7969 compounds, exclusively consisted of PAINS. These ASs contained two to 190 analogs with on average four PAINS per series. In all ASs with PAINS, 177 of the 480 PAINS substructures were detected. ASs exclusively consisting of PAINS covered 140 different substructures. Furthermore, for 32 PAINS substructures, at least 10 ASs were identified. Thus, overall, a large number of PAINS-containing ASs was available, providing an extensive structural organization of PAINS and a sound basis for our analysis.

### Targets

For the ASs belonging to the three different categories according to Fig. 3 target statistics were determined. We found that 7.3%

of the ASs exclusively consisting of PAINS and 6.9% of ASs comprising PAINS and non-PAINS were only active against a single target (ST-ASs). For ASs only consisting of non-PAINS, the proportion of ST-ASs was 13.6%. Thus, most ASs in all three categories were multi-target ASs (MT-ASs). ST- and MT-ASs exclusively consisting of PAINS were active against a total of 385 unique targets, while ST- and MT-ASs with PAINS and non-PAINS covered 401 targets. In addition, non-PAINS ST- and MT-ASs were active against a total of 418 targets. Thus, target coverage of all three categories of ASs was extensive and comparable in magnitude. Notably, the 1873 ASs only consisting of PAINS were active against nearly as many targets (92.1%) as the  $\sim$ 16-fold larger number of non-PAINS ASs.

### Hit rates

Fig. 1a shows the distribution of hit rates of extensively assayed PubChem compounds on the basis of which hit rate intervals were determined, as detailed above. In addition, Fig. 1b compares the distribution of hit rates for those non-PAINS and PAINS having rates exceeding 1.0%. More than 75 000 non-PAINS had hit rates greater than 1.0% compared to 5115



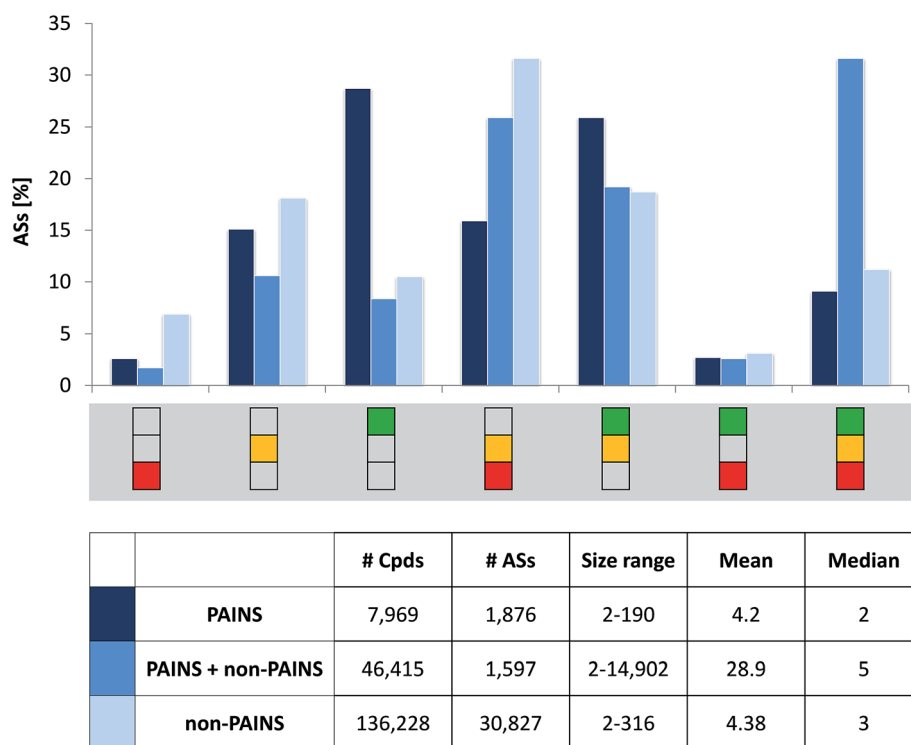


Fig. 3 Distribution of activity profiles for analog series of different composition. Global distributions of activity profiles for ASs exclusively consisting of PAINS (dark blue bars), combinations of PAINS and non-PAINS (blue), and only non-PAINS (light blue) are reported. At the bottom, compound and series statistics are provided.

compounds with PAINS substructures. Thus, 60.7% of all PAINS from ASs (7903 compounds) were consistently inactive or only active in one or two assays. As one would anticipate, within the hit rate interval exceeding 1.0%, PAINS had overall higher hit rates than non-PAINS but the differences were only small. As shown in Fig. 1b, the hit rate distributions were similar for PAINS and non-PAINS, with median values of slightly above and below 2.0%, respectively. Taken together, these observations made for PAINS with analog relationships corroborated earlier findings from global PubChem analysis.<sup>11,12</sup> For all ASs with PAINS, activity profiles were generated from their assay data.

### Activity profiles

Fig. 2 depicts the seven possible activity profiles for ASs that account for hit rate intervals and their combinations. Two exemplary ASs are shown. All analogs belonging to AS 1 were active and had high hit rates, resulting in the 'green-only' profile of the series. By contrast, four of five analogs of AS 2 were active and one consistently inactive. Two of the active analogs had high and two others expected hit rates. Thus, the activity profile of the series was the combination of all three intervals ('green-yellow-red').

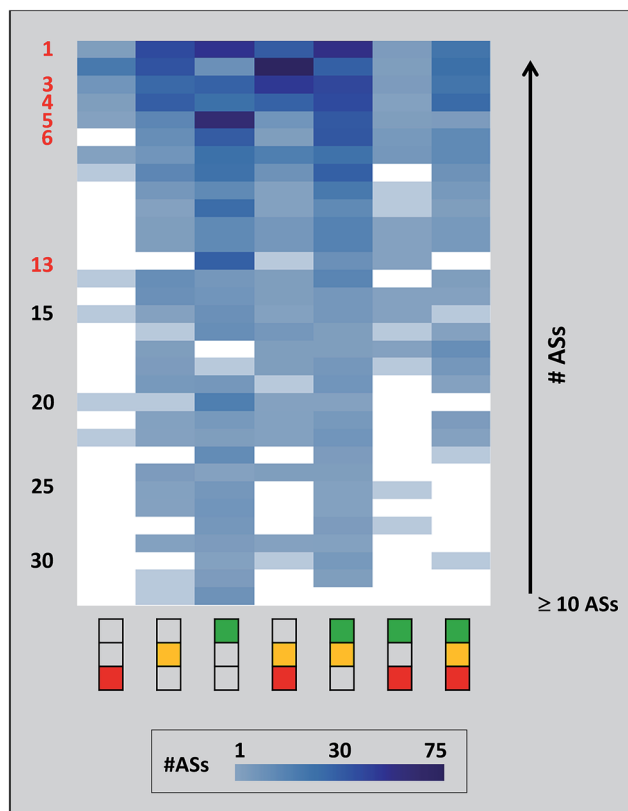
Activity profiles were systematically determined for all 34 300 ASs extracted from extensively assayed PubChem compounds. Therefore, the ASs were divided into three subsets: analogs having no PAINS substructures (30 827 series), analogs with and without PAINS substructures (1597), and analogs always containing PAINS (1876). Fig. 3 reports the distribution of these AS

subsets over different activity profiles in a histogram. Consistently inactive ASs ('red-only' profile) and ASs containing compounds having high hit rates and inactive analogs ('green-red') were rare. By contrast, nearly 30% of ASs exclusively consisting of PAINS displayed the 'green-only' (high hit rate) profile, which was a much larger proportion than obtained for the other two AS subsets (with close to 10%). Essentially inverse proportions were observed for ASs containing consistently inactive as well as active compounds with expected hit rates ('yellow-red'). Furthermore, more than 30% of ASs with non-PAINS and PAINS yielded the complete ('green-yellow-red') activity profile. Hence, these series contained analogs covering all hit rate intervals. Notably, about 55% of ASs exclusively consisting of PAINS yielded activity profiles covering multiple hit rate intervals, revealing that analogs with a given PAINS substructure often had different activities.

Fig. 4 shows the distribution of activity profiles for 32 PAINS for which 10 or more ASs were available that exclusively contained this PAINS substructure. Thus, this subset of PAINS was most frequently found in ASs. It included widely recognized PAINS such as anilines, rhodanines, or quinones.<sup>4</sup> The heatmap reveals the prevalence of the 'green-only' and 'green-yellow' profiles among the ASs of this subset of PAINS. However, the heatmap also shows that activity profiles were variably distributed across ASs with different PAINS. For example, the 'yellow-red' and complete activity profiles were also frequently observed. Hence, prevalent PAINS also displayed varying activities in ASs.







1	<b>ene_six_het_A</b> <i>Alkylidene barbiturates</i>	
3	<b>anil_di_alk_A</b> <i>Tertiary anilines</i>	
4	<b>indol_3yl_alk</b> <i>Indoles</i>	
5	<b>mannich_A</b> <i>Phenolic mannich bases</i>	
6	<b>ene_rhod_A</b> <i>Unsaturated rhodanines</i>	
13	<b>quinone_A</b> <i>Quinones</i>	

Fig. 4 Activity profile distribution for different PAINS. Activity profiles of ASs containing the same PAINS substructure are displayed in a heatmap. Each column corresponds to a given activity profile and each row represents an individual PAINS (sub)structure. Empty cells (white) indicate the absence of a profile. Occupied cells are color-coded according to increasing numbers of ASs displaying the same profile using a spectrum from light to dark blue. The heatmap only contains 32 PAINS with at least 10 different ASs. PAINS were ordered

### Context-dependent structure–activity relationships

The ASs containing PAINS substructures provided a series-based organization and reference frame for analyzing and comparing the activity of PAINS in different structural environments. A variety of interesting and in part puzzling relationships was observed.

Fig. 5a compares two rhodanine-based series with distinct hit rates and activity profiles. These ASs were tested in more than 300 assays with an assay overlap of 98%. Compounds forming the series on the left were at most active in a single assay, whereas compounds in the series on the right were active in six to eight assays. Both ASs shared a 5-phenylmethylene-3-rhodanine acetamide substructure that was modified at the nitrogen of the acetamide. Analogs in the series on the left had a tetrahydrothiophene-1,1-dioxide substituent in common, while the frequently active compounds in the ASs on the right shared a 2-(3-pyridinyl)-piperidine. Biologically relevant reactivities of rhodanines and related heterocycles have been intensely investigated and several plausible mechanisms of action have been proposed.<sup>28</sup> Often considered is a Michael addition *via* the exocyclic double bond.<sup>29</sup> In this case, observed differences in activity could not be attributed to a Michael-type reaction because the same rhodanine derivative occurred in both ASs. Instead, possible photochemical<sup>30</sup> or hydrolytic<sup>31</sup> reactivity might be modulated by different substituents at the acetamide.

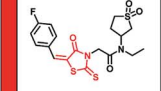
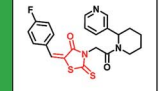
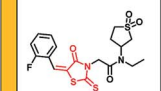
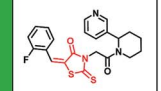
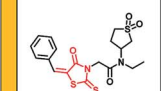
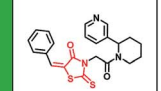
Fig. 5b depicts two ASs sharing a 3-methyl-indole core. Similar to the previous example, analogs forming the series on the left were only active in at most one assay, while analogs of the series on the right were active in five to nine assays. This was the case although the AS on the left was more extensively tested than the one on the right (in more than 500 vs. 300 assays). Different from the previous example, substitution patterns were more diverse here. Baell *et al.* discussed that 3-alkylindoles and indole-3-acetamide-2-carboxylic acids likely act as Michael acceptors and thereby cause artifacts.<sup>3</sup> However, in this case, the rarely active analogs in the ASs on the left contained a carboxylic acid function, which was replaced by a methyl group in the frequently active series on the right. Thus, the activity profiles of these ASs were opposite to expectations considering potential Michael acceptor reactivity.

Fig. 5c compares two series of 2-hydroxybenzylamine derivatives, one of which was consistently inactive in many assays (left), whereas analogs forming the other (right) were active in nine, 15, and 20 assays, respectively. Such high hit rates are likely to involve artifacts. The 2-hydroxybenzylamines may act as Mannich bases and elicit undesired activities by forming reactive quinone methides<sup>32</sup> or by chelating metal ions.<sup>33</sup> However, the striking difference in activity between these two ASs was not straightforward to rationalize. Notably, the 2-hydroxybenzylamine moiety in the series on the left was located at the terminus of the analogs, whereas it was fused with a pyridine

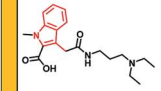
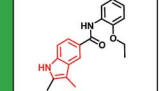
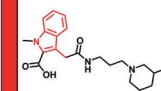
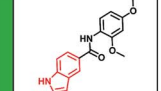
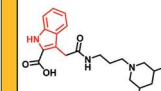
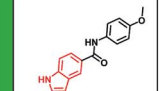
according to increasing numbers of ASs. Rows of PAINS for which specific examples are discussed in the text are numbered in red and these PAINS are specified below the heatmap.



(a)

PAINS class: ene_rhod A							
Assay overlap: 98.0%							
	# Assays [active]	# Assays	Hit rate		# Assays [active]	# Assays	Hit rate
	0	341	0		6	338	0.018
	1	337	0.003		8	341	0.023
	1	336	0.003		7	335	0.021

(b)

PAINS class: indol_3yl_alk							
Assay overlap: 59.4%							
	# Assays [active]	# Assays	Hit rate		# Assays [active]	# Assays	Hit rate
	1	526	0.002		5	342	0.015
	0	484	0		5	341	0.015
	1	510	0.002		9	313	0.029

(c)

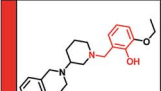
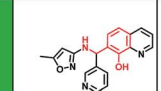
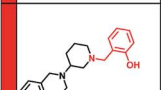
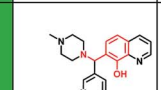
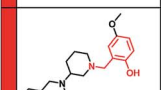
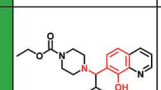
PAINS class: mannich_A							
Assay overlap: 61.2%							
	# Assays [active]	# Assays	Hit rate		# Assays [active]	# Assays	Hit rate
	0	339	0		20	520	0.038
	0	341	0		9	473	0.019
	0	307	0		15	462	0.032

Fig. 5 Analog series with PAINS having different activity profiles. In (a) to (c), pairs of ASs are shown that contain the same PAINS substructure (red) but have different activity profiles. For each compound, the number of assays it was tested in, the number of assays in which it was active, and the corresponding hit rate are reported. For each pair of ASs, assay overlap is quantified. (a) Rhodanines, (b) 3-alkylindoles, (c) Mannich bases.



ring in the compounds on the right and, in addition, bound to two other rings. Thus, the structural context in which the PAINS substructure was presented in these two ASs was distinct and one may hypothesize that a more or less constrained structural environment affects Mannich base reactivity.

In addition to comparing different ASs containing the same PAINS, it is also informative to analyze individual series with

(a)

PAINS class: ene_six_het_A			
Assay overlap: 58.0 %			
	# Assays [active]	# Assays	Hit rate
	33	405	0.081
	7	479	0.015
	1	451	0.02
	0	369	0

(b)

PAINS class: quinone_A			
Assay overlap: 56.3%			
	# Assays [active]	# Assays	Hit rate
	11	435	0.025
	10	347	0.03
	0	430	0
	7	383	0.018

Fig. 6 Analog series with PAINS having large variations in hit rates. Exemplary ASs are shown in which compounds display significantly different hit rates. The representation is according to Fig. 5. (a) Alkylidene thiobarbiturates, (b) quinone derivatives.

(a)

PAINS class: anil_di_alk_a			
Assay overlap: 68.6%			
	# Assays [active]	# Assays	Hit rate
	0	384	0
	0	284	0
	0	279	0
	0	391	0

(b)

PAINS class: imidazole_amino_A			
Assay overlap: 52.1%			
	# Assays [active]	# Assays	Hit rate
	7	459	0.015
	8	510	0.016
	9	458	0.02
	6	383	0.016

(c)

PAINS class: mannich_A			
Assay overlap: 48.2%			
	# Assays [active]	# Assays	Hit rate
	25	514	0.049
	19	444	0.043
	19	437	0.043
	0	268	0

Fig. 7 Analog series comprising PAINS and non-PAINS. Examples of 'mixed' ASs are shown that consist of compounds with and without PAINS substructures (red). The representation is according to Fig. 5. PAINS include (a) tertiary anilines, (b) amino imidazoles, and (c) phenolic Mannich bases.





different activity profiles. For example, Fig. 6a shows a series of alkylidene thiobarbiturates with varying hit rates. Here, replacement of a 1-methyl-pyrrol with a 3-pyridinmethanamine group greatly reduced hit rates or completely abolished activity. In addition, replacing the aromatic (4-fluorophenyl)-methyl substituent with increasingly aliphatic moieties might also contributed to a loss in activity. Hence, on the basis of these observations, several experimentally testable hypotheses can be formulated.

Fig. 6b depicts an AS of 9,10-dihydro-9,10-dioxo-2-anthracenesulfonamides containing a quinone substructure, a notorious PAINS<sup>4</sup> with one of the highest hit rates overall. However, in this AS having an unusual 'green-red' activity profile, one of the analogs was found to be consistently inactive in 430 assays. Compared to a closely related compound with activity in seven assays, the only modification was a *para-to-ortho* repositioning of methyl substituents at the phenyl moiety; a puzzling observation.

So far, only ASs exclusively consisting of PAINS were considered. However, series containing analogs with and without PAINS substructures also revealed interesting relationships. For example, Fig. 7a shows an ASs with a 'red-only' activity profile in which consistently inactive analogs contained a 1,4-diphenyl-2,6-piperidinedione core. Three of four analogs had different phenyl derivatives as substituents at the 2,6-piperidinedione nitrogen. Replacement of these groups with a *N,N*-dimethylaniline PAINS substructure also produced a completely inactive analog, although several likely interference mechanisms were proposed for tertiary anilines.<sup>34</sup> Thus, in this case, the 1,4-diphenyl-2,6-piperidinedione core restricted possible reactivity of different substituents.

Fig. 7b shows an ASs with a 'green-only' activity profile containing different amino imidazole derivatives, only one of which was a PAINS substructure. However, all analogs were active in seven to nine assays. Finally, Fig. 7c depicts an AS with three compounds containing a phenolic Mannich base that were active in 19 or 25 assays. In a fourth analog, methylation of the phenolic hydroxyl group of the Mannich base led to consistent inactivity. The only caveat in interpreting these results was that the inactive analog was tested in 268 assays, while the remaining active compounds were tested in more than 400 or 500 assays (mostly including the 268 assays). Thus, these differences in assay frequency might influence hit rates. Nonetheless, analysis of this series immediately provides the experimentally testable hypothesis that methylation of the reactive phenolic hydroxyl might 'disable' this PAINS structure. Many other ASs including PAINS are available to explore the dependence of assay interference on the structural context in which PAINS are presented.

## Conclusions

In this work, we have systematically extracted ASs with PAINS substructures from extensively assayed compounds, analyzed their activity profiles, and explored structure–activity relationships. These ASs provided an organization of PAINS according to varying structural contexts and a reference frame for studying

PAINS actions in different environments. A number of instructive examples have been identified, providing first insights into the structural context dependence of PAINS activities. As a part of our study, all ASs containing PAINS are made freely available (in an open access deposition referring to this work) to aid in theoretical and experimental follow-up investigations to further explore PAINS characteristics and the influence of structural embedding.<sup>35</sup>

## Conflict of interest

There are no conflicts of interest to declare.

## Acknowledgements

The use of OpenEye's toolkits was made possible by their free academic licensing program. D. S. is supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft.

## References

- 1 S. L. McGovern, E. Caselli, N. A. Grigorieff and B. K. Shoichet, *J. Med. Chem.*, 1996, **45**, 1712–1722.
- 2 B. K. Shoichet, *Drug Discovery Today*, 2006, **11**, 607–615.
- 3 J. B. Baell and G. A. Holloway, *J. Med. Chem.*, 2010, **53**, 2719–2740.
- 4 J. Baell and M. A. Walters, *Nature*, 2014, **513**, 481–483.
- 5 J. L. Dahlin, J. W. Nissink, J. M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang and M. A. Walters, *J. Med. Chem.*, 2015, **58**, 2091–2113.
- 6 E. Gilberg, S. Jasial, D. Stumpfe, D. Dimova and J. Bajorath, *J. Med. Chem.*, 2016, **59**, 10285–10290.
- 7 J. Bisson, J. B. McAlpine, J. B. Friesen, S. N. Chen, J. Graham and G. F. Pauli, *J. Med. Chem.*, 2016, **59**, 1671–1690.
- 8 K. M. Nelson, J. L. Dahlin, J. Bisson, J. Graham, G. F. Pauli and M. A. Walters, *J. Med. Chem.*, 2017, **60**, 1620–1637.
- 9 C. Aldrich, C. Bertozzi, G. I. Georg, L. Kiessling, C. Lindsley, D. Liotta, K. M. Merz Jr, A. Schepartz and S. Wang, *ACS Cent. Sci.*, 2017, **3**, 143–147.
- 10 S. Saubern, R. Guha and J. B. Baell, *Mol. Inf.*, 2011, **30**, 847–850.
- 11 S. J. Capuzzi, E. N. Muratov and A. Tropsha, *J. Chem. Inf. Model.*, 2017, **57**, 417–427.
- 12 S. Jasial, Y. Hu and J. Bajorath, *J. Med. Chem.*, 2017, **60**, 3879–3886.
- 13 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.
- 14 S. Jasial, Y. Hu and J. Bajorath, *PLoS One*, 2016, **11**, e0153873.
- 15 D. Stumpfe, D. Dimova and J. Bajorath, *J. Med. Chem.*, 2016, **59**, 7667–7676.
- 16 E. Griffen, A. G. Leach, G. R. Robb and D. J. Warner, *J. Med. Chem.*, 2011, **54**, 7739–7750.
- 17 X. Q. Lewell, D. B. Judd, S. P. Watson and M. M. Hann, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 511–522.



- 18 A. de la Vega de León and J. Bajorath, *MedChemComm*, 2014, **5**, 64–67.
- 19 J. Hussain and C. Rea, *J. Chem. Inf. Model.*, 2010, **50**, 339–348.
- 20 X. Hu, Y. Hu, M. Vogt, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2012, **52**, 1138–1145.
- 21 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.
- 22 . *RDKit: Cheminformatics and Machine Learning Software*, 2013, <http://www.rdkit.org>.
- 23 T. Sterling and J. J. Irwin, *J. Chem. Inf. Model.*, 2015, **55**, 2324–2337.
- 24 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
- 25 M. R. Berthold, N. Cebon, F. Dill, T. R. Gabriel, T. Kötter, T. Meinel, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, KNIME: The Konstanz Information Miner, in *Studies in Classification, Data Analysis, and Knowledge Organization*, ed. C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker, Springer, Berlin, Germany, 2008, pp. 319–326.
- 26 *OEChem TK*, OpenEye Scientific Software, Inc., Santa Fe, NM, U.S., 2012.
- 27 *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA, 2016.
- 28 T. Mendgen, C. Steuer and C. D. Klein, *J. Med. Chem.*, 2012, **55**, 743–753.
- 29 J. P. Powers, D. E. Piper, Y. Li, V. Mayorga, J. Anzola, J. M. Chen, J. C. Jaen, G. Lee, J. Liu, M. G. Peterson, G. R. Tonn, Q. Ye, N. P. C. Walker and Z. Wang, *J. Med. Chem.*, 2006, **49**, 1034–1046.
- 30 M. E. Voss, P. H. Carter, A. J. Tebben, P. A. Scherle, G. D. Brown, L. A. Thompson, M. Xu, Y. C. Lo, G. Yang, R. Liu, J. Strzemienski and G. Everlof, *Bioorg. Med. Chem. Lett.*, 2003, **13**, 533–538.
- 31 J. Brem, S. S. van Berkel, W. Aik, A. M. Rydzik, M. B. Avison, I. Pettinati, K. Umland, A. Kawamura, J. Spencer, T. D. W. Claridge, M. A. McDonough and C. J. Schofield, *Nat. Chem.*, 2014, **6**, 1084–1090.
- 32 Y. Herzig, L. Lerman, W. Goldenberg, D. Lerner, H. E. Gottlieb and H. E. Nudelman, *J. Org. Chem.*, 2006, **71**, 4130–4140.
- 33 M. J. Caulfield, D. J. McAllister, T. Russo and D. H. Solomon, *Aust. J. Chem.*, 2001, **54**, 383–389.
- 34 R. H. Young, D. Brewer, R. Kayser, R. Martin, D. Feriozi and R. A. Keller, *Can. J. Chem.*, 1974, **52**, 2889–2893.
- 35 <http://zenodo.org/>, data release upon publication.



## Conclusions

A first assessment of the structural context dependence of PAINS activities has been presented. A set of 177 individual PAINS classes were available in a large number of analog series of extensively tested screening compounds, thereby providing structural context information of PAINS. Activity profiles of series were systematically generated based on hit rates of analogs in screening assays. Depending on the PAINS substructure, analog series showed varying activity phenotypes. For example, quinones were predominantly found in series with high hit rates, while derivatives of tertiary anilines and indoles often displayed low activities or inactivity.

Moreover, analogs or different analogs series containing the same PAINS substructure were compared. Thus, structural modifications were identified that had significant influence on the hit rates of interference compounds. These included, among others, methylation of Mannich bases and different aromatic substitution of unsaturated rhodanines.

In this work, PAINS were organized according to varying structural contexts and activity profiles. Thereby a reference frame for the exploration of interference in different structural environments was provided. However, given the complexity of interference mechanisms and the variety of liable substructures, it remains difficult to formulate general rules of structural context dependency based on the study of instructive examples. Hence, in the next chapter we apply machine learning methods for a systematic classification of PAINS that are highly promiscuous from others that are consistently inactive. The analysis of structural features that favor correct predictions further supports the analysis of structural context dependence.



# 5 Machine Learning Distinguishes with High Accuracy Between Pan-Assay Interference Compounds that are Promiscuous or Represent Dark Chemical Matter

## Introduction

Structural alerts offer a good opportunity to detect potential false-positives and assay artifacts at an early stage of drug discovery programs. However, assay interference activities are strongly dependent on experimental conditions and the structural embedding of chemically liable substructures. Due to this complexity, it is difficult to predict undesired reactivities sufficiently, as decisions can often only be made on a case-by-case basis. To address this problem, machine learning methods were used to distinguish between promiscuous compounds and compounds that display consistent inactivity, often referred to as 'dark chemical matter' (DCM).<sup>151</sup> Based on their representation as ECFP4 and MACCS structural fingerprints, PAINS were classified using support vector machines, random forest, and deep neural networks.

In the following chapter, structural features that support correct predictions are identified and evaluated for their role in the structural dependence of PAINS activities.

My main contribution to this work included weighting, mapping, and analysis of EFCP4 features that favored correct classification by support vector machines.

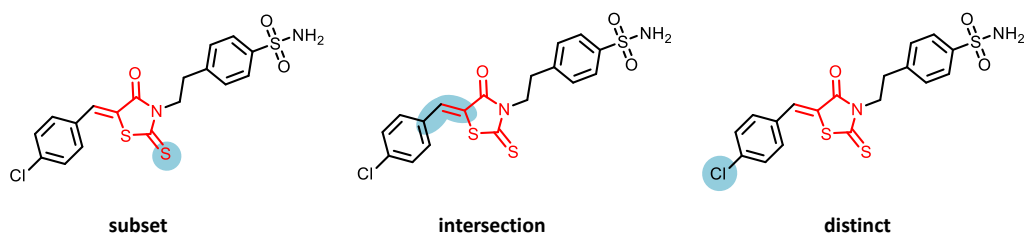
This study was published as: S. Jasial, E. Gilberg, T. Blaschke, J. Bajorath. Machine Learning Distinguishes with High Accuracy between Pan-Assay Interference Compounds that are Promiscuous or Represent Dark Chemical Matter. *Journal of Medicinal Chemistry* **2018**, 61, 10255–10264.

## Materials and Methods

A subset of 437,257 extensively assayed screening compounds<sup>152</sup> from PubChem BioAssays<sup>153</sup> that were tested in both primary assays (percentage of inhibition from a single dose) and confirmatory assays (dose-response assays yielding IC<sub>50</sub> values) provided the starting point of our analysis. The set of extensively assayed compounds was screened *in silico* for PAINS using three publicly available PAINS filters (RDKit,<sup>154</sup> ZINC,<sup>155</sup> and ChEMBL<sup>156</sup>). Filtering identified 270 different PAINS substructures present in 27,520 screening compounds. Based on assay activity profiles of all compounds containing PAINS, two data sets were generated. The first set consisted of 5233 promiscuous PAINS (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. Additionally, compounds qualified for PROM\_PAINS if they were tested in at least 50 confirmatory assays and varying numbers of primary assays and showed activity in 10 or more assays. The second set consisted of 3059 compounds with PAINS substructures that were consistently inactive in at least 100 primary and varying number of confirmatory assays, thus displaying dark chemical matter (DCM) character (termed DCM\_PAINS). DCM\_PAINS and PROM\_PAINS represented 94 and 192 PAINS substructures, respectively.

Weighting of structural features and their subsequent analysis was based on a classification model utilizing support vector machines (SVM). In this model, PROM\_PAINS and DCM\_PAINS were projected as vectors into a multidimensional feature space, based on their representation as an extended ECFP4 feature vector. SVMs are used to solve the classification problem by generating a hyperplane in feature space that best distinguishes training instances with different binary class labels (i.e. PROM vs. DCM\_PAINS).<sup>157</sup> Test compounds were then classified depending on which side of the separating hyperplane they fell.

The classification model was built for 'balanced' training sets in which the numbers of PROM\_PAINS and DCM\_PAINS were adjusted for shared PAINS substructures. Thus, 54 individual PAINS substructures were represented by at least two DCM\_PAINS and PROM\_PAINS distributed equally and then randomly divided into training (50%) and test compounds (50%). Ten independent sampling trials were carried out yielding ten balanced training and test sets. In each trial,



**Figure 1: Structural embedding of ECFP4 features.** The three structural context categories are shown by the example of a training set rhodanine derivative (PAINS code: *ene\_rhod\_A*). Mapped features are traced in blue and the PAINS substructure is colored red. The figure has been adapted from [159].

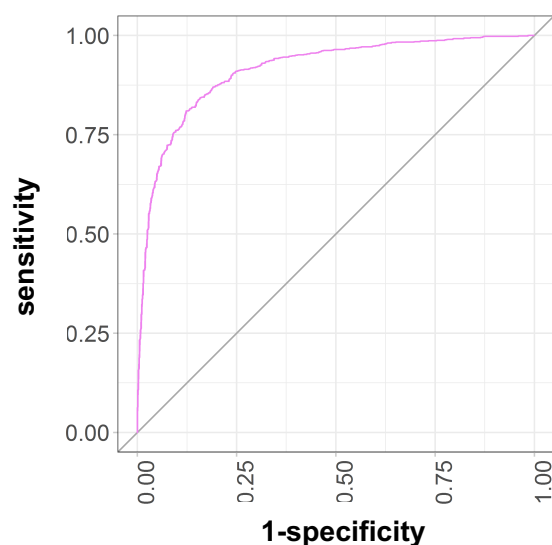
each training and test set contained approximately 1900 compounds.

A feature weighting method<sup>158</sup> was applied to assign different weights to individual ECFP4 features that favored correct predictions of the classification model. Features were ranked according to their preferred presence in either PROM\_PAINS or DCM\_PAINS. SMARTS string representations of the top 30 ECFP4 features from each ranking were used to search for substructures in PAINS utilizing the KNIME implementation<sup>150</sup> of the RDKit substructure filter.<sup>154</sup> For a successful match, the list of atom indices was used to map features to compounds in the test set. Depending on their structural embedding, mapped features were assigned to three structural context categories. As shown in Figure 1, these categories comprised features (i) representing a subset of a PAINS substructure (subset); (ii) being part of a PAINS substructure and part of its structural environment (intersection); and (iii) mapping to a region in a test compound outside of the PAINS substructure (distinct).

## Results and Discussion

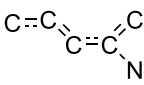
In the first part of this study, we examined three different machine learning algorithms with respect to their ability to distinguish between promiscuous PAINS and those that were consistently inactive through numerous biological screening assays. These classification methods included SVM, random forest, and deep neural networks and were applied using MACCS and ECFP4 structural fingerprints as molecular representations of PAINS. Independent of the used method, accurate

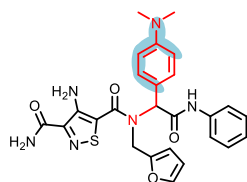




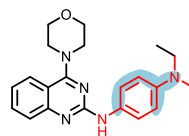
**Figure 2: Receiver operating characteristic (ROC) curve for a balanced SVM model.** A ROC curve is shown for an individual classification trial distinguishing PROM\_PAINS and DCM\_PAINS using SVM on the basis of ECFP4 structural fingerprints. The sensitivity indicates the proportion of correctly predicted PROM\_PAINS and is plotted on the y-axis (true positive rate). On the x-axis, the 1-Specificity value shows the ratio of DCM\_PAINS predicted as PROM\_PAINS (false positive rate). The figure has been adapted from [159].

classification models were obtained for a global data set and a balanced set. To account for potential substructure bias and data imbalance, the balanced set contained training and test compounds that exclusively originated from shared PAINS motifs of equally distributed PROM\_PAINS and DCM\_PAINS. Figure 2 shows a receiver operating characteristic (ROC) curve for an individual classification trial using a balanced SVM model on the basis of ECFP4 molecular representations. Notably, the classification was solely guided by association of structural feature distributions with the binary class label and did not consider additional parameters such as reactivity or activity annotations. Thus, the stable performance of the balanced model indicated that the applied machine learning methods were capable of distinguishing between structural contexts of actives and inactives in which shared PAINS substructure were embedded.


contribution	feature	rank
positive		<b>10</b>

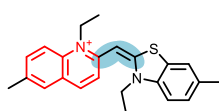


aniline (active)

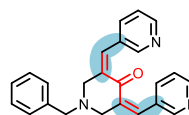


aniline (active)

contribution	feature	rank
positive		<b>14</b>



pyridinium (active)

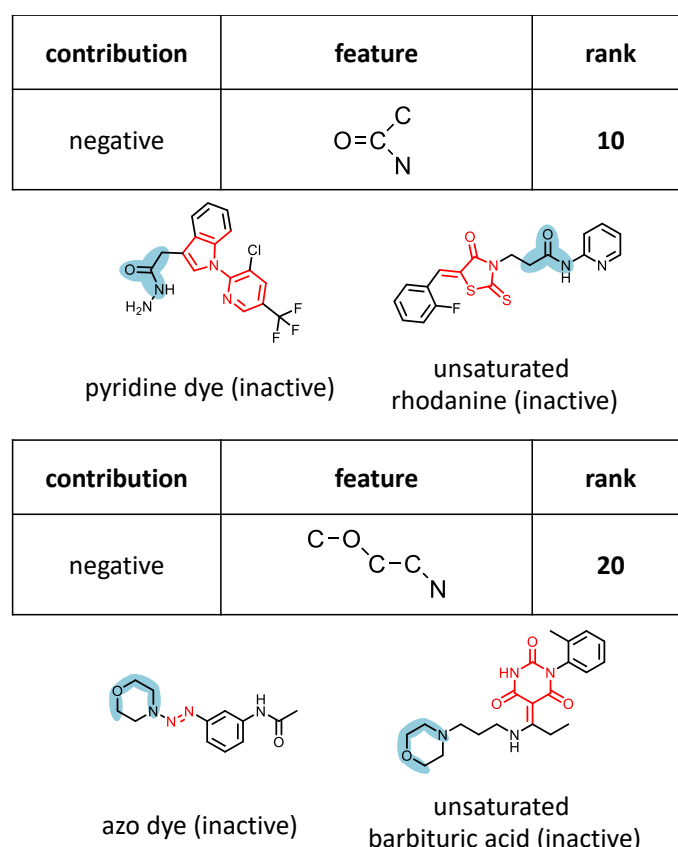


unsaturated  
ketone (active)

**Figure 3: Positive ECFP4 feature.** Shown are SMARTS representations of ECFP4 features from SVM predictions that were highly ranked for PROM\_PAINS. In exemplary PROM\_PAINS, mapped features are traced in blue and PAINS substructures are colored in red. At the top, a *N,N*-di-substituted aniline ECFP4 feature detected in aniline type PROM\_PAINS (PAINS codes: *anil\_di\_alk\_A* and *anil\_di\_alk\_D*) is shown. At the bottom, an ECFP4 feature covering an unsaturated alkene found in PROM\_PAINS is shown. (PAINS codes: *het\_pyridinium\_A* and *ene\_one\_ene*). The figure has been adapted from [159].

In the second part of this work, we intended to analyze the structural and topological patterns that were selected and utilized for correct classification by machine learning algorithms to expand our knowledge of the structural context underlying PAINS activities. Given the blackbox character of machine learning models, we adapted a feature weighting approach developed for SVM<sup>158</sup> to explore determinants of the predictions and calculated cumulative positive and negative feature weights for all ECFP4 features of test compounds in balanced data sets.

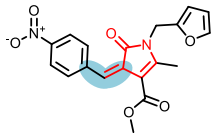
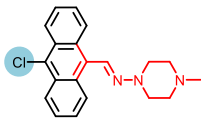
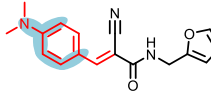
Then, features were ranked for PROM\_PAINS and DCM\_PAINS according to positive and negative weight sums, respectively. For example, a total of 19,668 ECFP4 features were detected for a given balanced test set, of which 2573 and 2527 features had positive and negative weight sums, respectively. On the basis of this result, 30 top ranked features were identified for each PROM\_PAINS (termed positive features) and DCM\_PAINS (negative features) and provided the basis of the feature analysis.



**Figure 4: Negative ECFP4 features.** Shown are features from SVM predictions that were highly ranked for DCM\_PAINS. Mapped structural features are marked in blue and PAINS motifs are colored red. On top, an amide pattern found in DCM\_PAINS (PAINS codes: *dyes5A* and *ene\_rhod\_A*) and on bottom the mapping of an aliphatic heterocycle (PAINS codes: *ene\_six\_het\_A* and *azo\_A*) is shown. The figure has been adapted from [159].

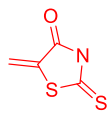
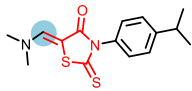
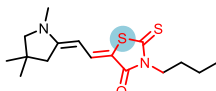
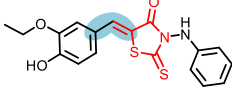
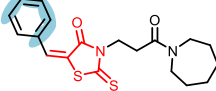
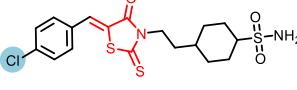
Frequent negative features included aliphatic carbon atoms with different degrees of hydrogen substitution, cyclic aliphatic ethers, and  $sp^2$ -hybridized oxygen atoms of carbonyl and sulfonyl groups. On the other hand, patterns of conjugated ring systems, chlorine and sulfur atoms, and unsaturated hydrocarbons adjacent to ring systems were among the preferred positive features. As a result, high-ranked positive and negative features were found to be composed differently.

Positive features were preferentially associated with reactive moieties, while negative features were chemically more inert. For example, Figure 3 (top) shows an *N,N*-di-substituted aniline motif that is a part of a PAINS substructure and preferentially emerges in PROM\_PAINS. Indeed, these aromatic tertiary amines are present in 23 PAINS substructures<sup>25,38,41,48</sup> and thought to interfere with fluorometric assays by quenching.<sup>160</sup> In addition, Figure 3 (bottom) shows a  $\beta$ -unsaturated alkene bisecting two ring systems, which was also prevalent in PROM\_PAINS. This atom arrangement is a part of a recurrent Michael acceptor motif that occurs in a variety of PAINS substructures. In contrast, Figure 4 shows examples of highly ranked negative features representing amide bond and morpholino patterns that preferably appear outside PAINS substructures and are not associated with known interference mechanisms.

feature	chemical reactivity	ratio	PROM_PAINS example
$C-C=C$	electrophile	3.21	
Cl	electron withdrawing	1.53	
$C=C=C=C$ N	quenching	2.48	

**Figure 5: Prominent features with positive weight.** The SMARTS representation of exemplary positive features and their possible chemical reactivity is provided. The ratio of their frequency of occurrence in PROM\_PAINS versus DCM\_PAINS is reported (ratio). For each feature an exemplary compound is shown. The structural feature is traced in blue and the PAINS substructure is colored in red. The figure has been adapted from [159].

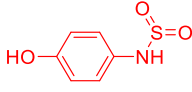
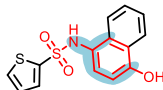
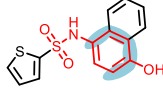
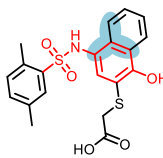
In Figure 5, exemplary features with positive weight in PROM\_PAINS are shown and their possible chemical reactivity is provided. The frequency of occurrence of these features in PROM\_PAINS was higher than in DCM\_PAINS.

PAINS code		PAINS substructure	
ene_rhod_A			
feature	context category	example	
1	C	subset	
2	S	subset	
3	C-C=C	intersection	
4	C=C=C=C=C	distinct	
5	Cl	distinct	

**Figure 6: Structural context analysis of unsaturated rhodanines.** For rhodanines (PAINS code: *ene\_rhod\_A*) positive features are mapped onto exemplary compounds. Mapped features are traced on a blue background and atoms of the PAINS substructure are colored red. For each feature, the structural context category is provided, as defined in the text. The figure has been adapted from [159].

The analysis showed that calculated ECFP4 features classifying promiscuous and inactive PAINS often constituted structural patterns related to the potential reactivity of PROM\_PAINS, providing a rationale for successful global classification. Yet, different activities of compounds belonging to the same PAINS class and their correct classification could only be rationalized by studying the structural environment in which the PAINS substructure was incorporated. To address this issue, top ranked ECFP4 features were mapped onto PROM\_PAINS and DCM\_PAINS and categorized with respect to their structural embedding. Accordingly, *subset* features were incorporated in PAINS substructures, *intersection* features were overlapping with the PAINS substructure and the remaining structure of a compound, and *distinct* features were entirely localized outside PAINS substructures.

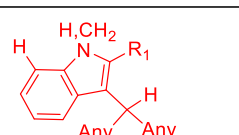

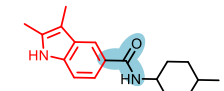

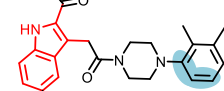

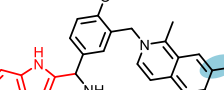
Figure 6 shows exemplary positive ECFP4 features belonging to different categories for unsaturated rhodanines. Five-membered heterocycles such as rhodanines represent a prominent PAINS class, but have also been considered as privileged scaffolds in drug discovery.<sup>49</sup> Unsaturated rhodanines are prone to ring opening reactions, metal chelation, and have a strong tendency to react with nucleophilic thiol groups through a Michael-type addition at the exocyclic double bond.<sup>161-163</sup> The potential reactivity of the exocyclic double bond towards biological nucleophiles is a prime example of structural context dependency, as many compounds containing this substructure elucidate varying activity profiles, including promiscuous and consistently inactive PAINS. As discussed above, identification of positive and negative features did not only put reason on successful classifications, but their mapping and categorization also provided an insight into the influence of structural context on predictions. Positive features that were absent or underrepresented in DCM\_PAINS characterizing accurately classified rhodanine PROM\_PAINS are shown in Figure 6. Subset feature 1 and intersection feature 3 cover a  $sp^2$ -hybridized exocyclic carbon, the position at which attacks of reactive nucleophiles, such as thiols, typically occur. Furthermore, intersection feature 3 underlines the importance of an additional ring system in vicinity of the Michael acceptor. Consequently, the presence of the aromatic feature 4 in conjugation with the unsaturated bond favors Michael acceptor activity, which is supported by further decreasing the electron density of the double bond, for example, by a chlorine substitution (feature 5).

PAINS code	PAINS substructure	
sulfonamide_B		
feature	context category	example
<b>1</b> $C=C=C=C$	subset	
<b>2</b> $C=C=C=C$	intersection	
<b>3</b> $C=C=C$	intersection	

**Figure 7: Structural context analysis of phenylsulfonamides.** For *p*-hydroxyarylsulfonamides (PAINS code: *sulfonamide\_B*) positive ECFP4 features are mapped onto exemplary compounds. Mapped features are traced on a blue background and atoms of the PAINS substructure are colored in red. For each feature, the structural context category is given, as defined in the text. The figure has been adapted from [159].

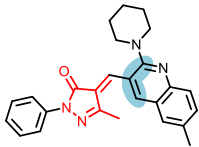
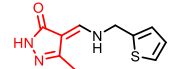
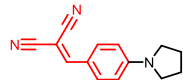
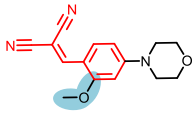
The structure class of *p*-hydroxyarylsulfonamides (PAINS code: *sulfonamide\_B*) typically displays undesired redox and thiol reactivity if they contain a naphthalene core.<sup>41,164,165</sup> As shown in Figure 7, the intersection features 2 and 3 mapped onto correctly predicted PROM\_PAINS cover this condition.



PAINS code	PAINS substructure	
indol_3yl_alk	 $R_1 = \text{CH}_2, \text{C=Het}, \text{C:Het}, \text{CHN}, \text{CH}(\text{CH}_2)\text{CH}_2\text{NCH}_2$	
feature	context category	example
1 	intersection	
2 	distinct	
3 	distinct	

**Figure 8: Non interpretable feature mapping.** For the PAINS class *indol\_3yl\_alk*, positive ECFP4 features are mapped onto exemplary compounds. The structural context category for each feature is provided. ECFP4 features are traced on a blue background, and PAINS substructures are colored red. The figure has been adapted from [159].

Despite consistent interpretation of feature mappings, positive features mapped onto test compounds could not always be easily interpreted in chemical terms. For example, positively weighted features detected for alkyloindoles (PAINS code: *indol\_3yl\_alk*) could not be directly associated with the likely interference mechanism of this class, as shown in Figure 8. Here, positive features in PROM\_PAINS displayed conjugated ring systems that were distant to the indoline 3-position whose tautomer is prone to unwanted Micheal-type reactivity.<sup>38</sup>

PAINS code	PROM_PAINS	DCM_PAINS
ene_five_het_A		
ene_cyano_A		

**Figure 9: Correctly predicted PROM\_PAINS and DCM\_PAINS.** Exemplary PROM\_PAINS and DCM\_PAINS are shown for two PAINS classes (PAINS codes: *ene\_five\_het\_A* and *ene\_cyano\_A*) in absence or presence of mapped ECFP4 features. PAINS substructures are colored in red and mapped features are traced on a blue background. The figure has been adapted from [159].

However, we also found that the correct classification by the SVM model could be directly related to the presence or absence of a positive or negative feature in structurally similar compounds. Figure 9 shows two exemplary PAINS classes.

In the first example, the conjugated quinolone bicyclic adjacent to the exocyclic double bond of the PAINS substructure (PAINS code: *ene\_five\_het\_A*) rationalized the correct prediction of PROM\_PAINS. On the contrary, when this feature was not present and replaced by a secondary amine, another compound was correctly predicted as a DCM\_PAINS. In the second example, reactivity of the PAINS motif (PAINS code: *ene\_cyano\_A*) might depend on the reactivity of the exocyclic unsaturated carbon, bisecting two carbonitrile moieties. Introducing an electron donating methoxy substituent is expected to increase electron density of the conjugated system and thus decrease the likelihood of nucleophilic attacks at this site. As shown in Figure 9, a methoxy group represented a feature negatively weighted by the SVM model and its presence supported the correct prediction of DCM\_PAINS.

## Conclusions

In this chapter it has been demonstrated that machine learning algorithms have the ability to distinguish highly promiscuous from consistently inactive PAINS. Classification models were applied to balanced data sets in which compounds of a given PAINS substructure were equally distributed among promiscuous and inactive compounds. Therefore, it was concluded that machine learning methods must recognize and utilize differences in topology and structure of differently active PAINS to ensure accurate classification.

To interpret these models, a feature weighting and mapping method was applied that indirectly assessed the predictions of support vector machines. A number of positively weighted ECFP4 features were identified which included PAINS substructures and reactive moieties that were predominantly found in promiscuous PAINS. Moreover, the localization of positive and negative features in correctly classified PAINS allowed further exploration of the structural environment of PAINS displaying different activity profiles. For example, electron withdrawing substituents in the vicinity of reactive Michael acceptors frequently favored the classification of PAINS as promiscuous.

Nevertheless, it must be taken into account that not all correct or incorrect predictions can be easily rationalized by analyzing structural features and that the performance of machine learning models generally depends on calculation conditions. However, as shown in the previous studies, key for an accurate assessment of assay interference is the large scale exploration of the complex interplay between structural features that cause undesired chemical reactivities. Therefore, the use of classification models that take available structural context information into account provides a valuable addition to the use of PAINS substructure filters and knowledge-based assessment of chemical liabilities.

In the next chapter, the SAR analysis of PAINS is complemented by a data-driven assessment of assay promiscuity. Therefore, a systematic analysis of hit rates and structural relationships of extensively assayed screening compounds was carried out.



# 6 Towards a Systematic Assessment of Assay Interference: Identification of Extensively Tested Compounds with High Assay Promiscuity

## Introduction

In the previous chapter it has been demonstrated that distinguishing potential liabilities from true multitarget activities of promiscuous screening compounds remains a difficult task. Especially the sole use of substructure filters does not do justice to the multifaceted picture of assay interference. For many PAINS classes, interference properties between derivatives may differ according to structural and experimental contexts. Moreover, as discussed in the first chapter, publicly available structural alerts do not comprehensively cover interference mechanisms.

This chapter addresses these challenges through a data-driven analysis of promiscuous compounds, without reducing available compounds through the prior use of structural filters. Taking structural context into account, MMSs of extensively tested screening compounds are generated and characterized according to their series hit rates and experimental assay overlap. Data sets are made freely available for follow-up investigations.

My main contribution to this work was the statistic analysis of extensively tested screening compounds and generation and classification of MMS.

Reprinted with permission from 'E. Gilberg, D. Stumpfe, J. Bajorath. Towards a Systematic Assessment of Assay Interference: Identification of Extensively Tested Compounds with High Assay Promiscuity. *F1000Research* **2017**, 6, e1505.' Copyright 2017 Science Navigation Group



## RESEARCH ARTICLE

**REVISED** Towards a systematic assessment of assay interference: Identification of extensively tested compounds with high assay promiscuity [version 2; referees: 3 approved]

Erik Gilberg, Dagmar Stumpfe, Jürgen Bajorath

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

**v2** First published: 17 Aug 2017, 6(CHEM INF SCI):1505 (doi: 10.12688/f1000research.12370.1)

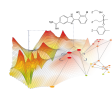
Latest published: 09 Oct 2017, 6(CHEM INF SCI):1505 (doi: 10.12688/f1000research.12370.2)

**Abstract**

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

**Keywords**

Biological screening data, statistical analysis, active compounds, assay frequency, hit rate distribution, assay interference

This article is included in the **Chemical Information Science gateway**.**Open Peer Review**

Referee Status:

	Invited Referees		
	1	2	3
<b>REVISED</b>			
<b>version 2</b>		report	
published 09 Oct 2017			
<b>version 1</b>			
published 17 Aug 2017	report	report	report

- 1 **John A. Lowe III** , JI3pharma LLC, USA
- 2 **José L Medina-Franco** , National Autonomous University of Mexico, Mexico  
**Fernanda I. Saldívar-González**, National Autonomous University of Mexico, Mexico
- 3 **Michael Walters** , University of Minnesota, USA

**Discuss this article**

Comments (2)

**Corresponding author:** Jürgen Bajorath ([bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de))

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

**Competing interests:** No competing interests were declared.

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

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

**Copyright:** © 2017 Gilberg E *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

**How to cite this article:** Gilberg E, Stumpfe D and Bajorath J. **Towards a systematic assessment of assay interference: Identification of extensively tested compounds with high assay promiscuity [version 2; referees: 3 approved]** *F1000Research* 2017, 6(Chem Inf Sci):1505 (doi: [10.12688/f1000research.12370.2](https://doi.org/10.12688/f1000research.12370.2))

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



**REVISED Amendments from Version 1**

We thank all reviewers for their positive and encouraging comments. In our revision, we have addressed all points raised by Dr. José L. Medina-Franco and Dr. Fernanda I. Saldívar-González as suggested in report 2, except one, i.e., we prefer retaining the current title as it is. In addition, we thank Dr. Michael Walters for noticing the typographical error in Figure 3, which has been corrected, and for emphasizing the interference potential of sulfonylpyrimidines, which we presented as an example for frequently active compounds. The interference potential of sulfonylpyrimidines has been further discussed, as suggested by Dr. Walters.

See referee reports

## Introduction

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

Without doubt, judging assay interference and candidate compounds requires profound chemical knowledge and experience. It is equally relevant, however, to strive for a data-driven assessment of promiscuity by exploring compound activity data on a large scale<sup>20</sup>, aiming to identify compounds with interference potential for further analysis. Previously, we have determined that increasing assay frequency of pairs of structural analogs did not correlate with differences in promiscuity<sup>23</sup>. The current analysis was focused on hit rates of individual compounds that were extensively assayed to identify the overall

most active chemical entities. Therefore, we have carried out a statistical analysis of hit rates of compounds that were extensively tested in screening assays. These compounds were evaluated in on average 411 assays (with a median value of 437 assays per compound). A special feature of this study has been its focus on pairs or larger series of analogs, rather than single compounds, which provides additional confidence criteria for activity assessment and further increases the information content of activity data analysis. Many series of analogs with much higher than typically observed hit rates and largely consistent activity profiles across many different assays were identified. This collection of series provides a basis for further investigating compounds with interference potential or true multi-target activities.

## Methods

### Compounds

From the PubChem BioAssay database<sup>16</sup>, 437,257 compounds were pre-selected that were tested in both primary and confirmatory assays, representing extensively assayed screening compounds<sup>23</sup>. Approximately 95% of these compounds were evaluated in more than 50 primary and/or confirmatory assays<sup>23</sup>. Primary PubChem assays report compound activity (e.g., percentage activity) for a single dose, while confirmatory assays are dose-response assays yielding titration curves and IC<sub>50</sub> values. Our current analysis focused on primary assays, for which much larger data volumes were available than for confirmatory assay. Primary assays also included assays for which no target was specified (such as cell-based assays). In addition to larger volumes, primary assays are more prone to false-positives than confirmatory assays, thus providing an upper-level estimate of compound promiscuity consistent with the goals of our analysis. Assignments of active compounds were taken from each individual assay as reported. Screening parameters such as compound concentration and activity criteria were assay-dependent. For pre-selected compounds, hit rate statistics were determined.

### Matched molecular pairs and series

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

### MMS parameters

For each MMS, three parameters were calculated. First, the *MMS hit rate (HR)* was obtained from the union of all assays (i.e., the

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

All calculations were carried out using in-house Java scripts and KNIME<sup>30</sup> protocols with the aid of the OpenEye<sup>31</sup> chemistry toolkit.

## Results and discussion

### Study design

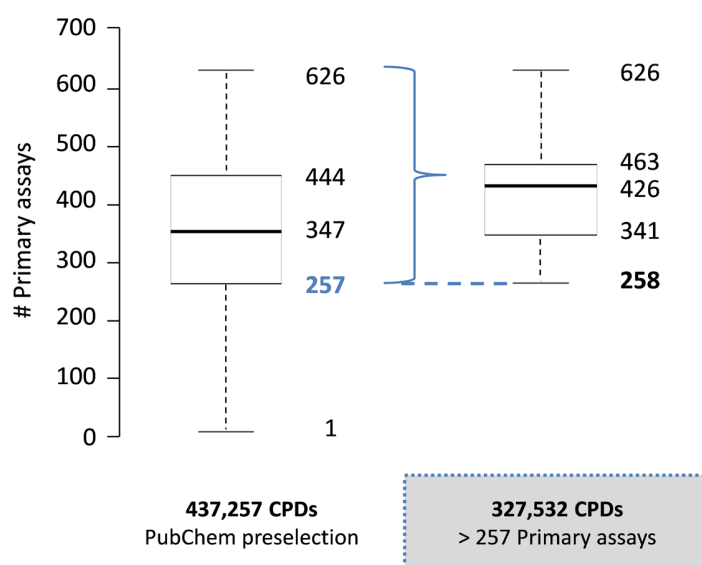
A statistical analysis of hit rates of extensively tested screening compounds is presented taking assay frequency into account. On the basis of the hit rate distribution, ranges of unusually high hit rates were determined. Since the majority of compounds that were active in primary assays were also active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

### Source compounds and assay distribution

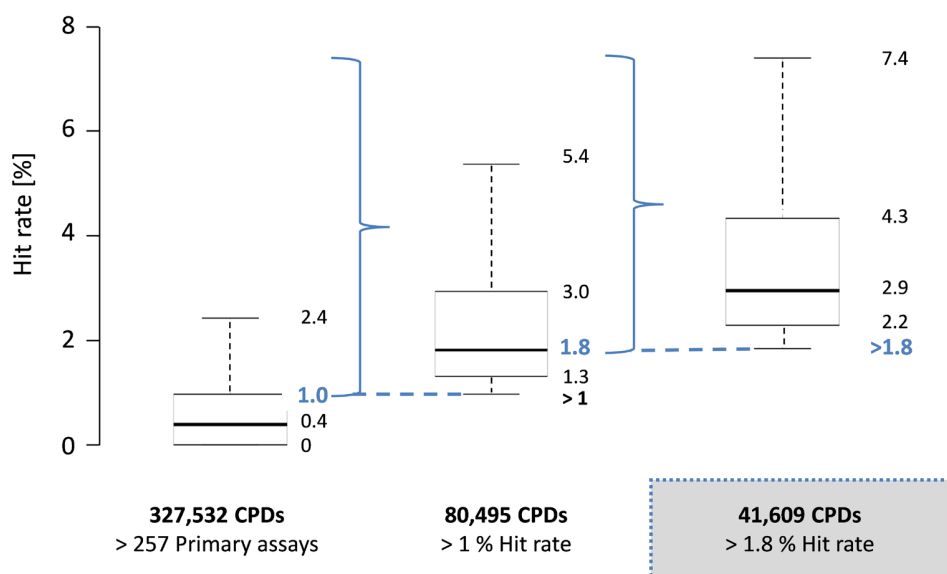
Figure 1 (boxplot on the left) shows the global distribution of primary assays for 437,257 extensively tested PubChem compounds, with a median value of 347 assays per compound. From these, a subset of 327,532 compounds was selected that were tested in more than 257 primary assays, corresponding to the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

### Hit rate distribution

For 327,532 compounds tested in more than 257 assays, hit rates were determined. The distribution is reported in Figure 2 (boxplot on the left), resulting in a median hit rate of 0.4%. The lower quartile boundary and lower whisker of the boxplot were identical and represented consistently inactive compounds, which were not of interest for our current analysis. On the basis of the distribution, the interval of “bulk hit rates” ( $b_{hr}$ ) for these extensively assayed PubChem compounds was defined as  $0\% < b_{hr} \leq 1.0\%$ , covering the lower quartile, median, and upper quartile (and hence the “bulk” of the distribution). There were 80,495 compounds with hit rates  $\geq 1.0\%$ . The hit rate distribution of this compound subset is shown in Figure 2 (middle), yielding a median of 1.8%. This value was set as the hit rate threshold for most active screening compounds. The threshold was exceeded by 41,609 compounds, representing 12.7% of the initial compound pool. The hit rate distribution of these compounds is reported in Figure 2 (right), resulting in a median of 2.9%. We determined that 93.1% of the compounds with hit rates greater than 1.8% in primary assays were also active in confirmatory assays (yielding  $IC_{50}$  values). Hence, their activity was not confined to primary assays.



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



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

### Compound series and parameters

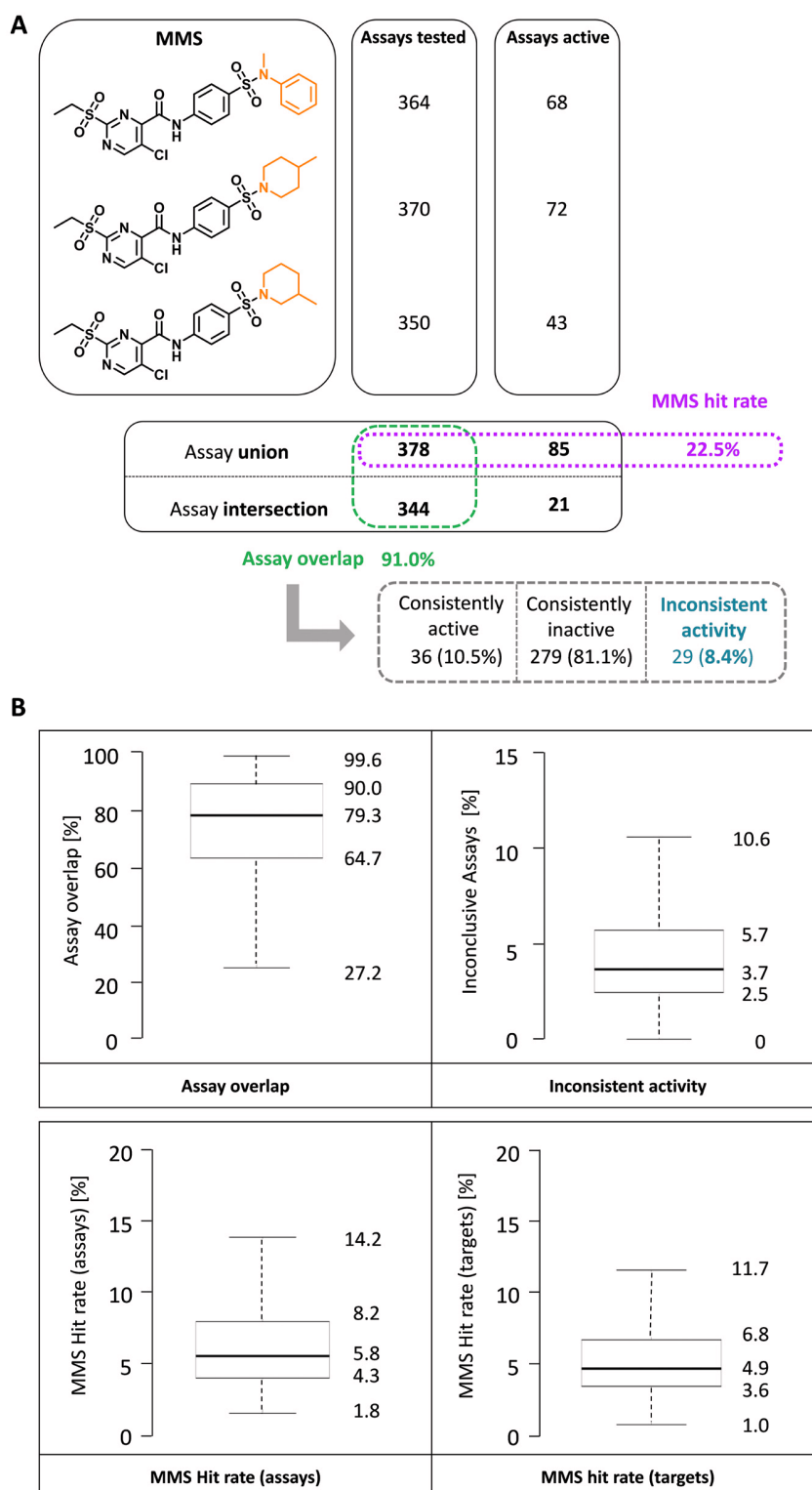
From the 41,609 compounds with highest hit rates, MMSs were systematically extracted on the basis of RECAP-MMPs. After removal of redundant MMSs (see Methods), 6941 unique MMSs were obtained comprising 14,646 compounds, which represented our final hit rate- and series-based selection set. Table 1 reports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

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

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

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

compounds comprising individual MMSs. We note that MMSs can be ranked in the order of decreasing assay overlap and MMS hit rates and increasing inconsistent assays and prioritized, for example, on the basis of rank fusion calculations.



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

### Target distribution in primary assays

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

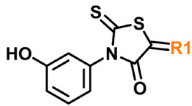
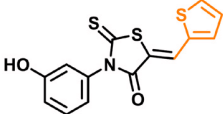
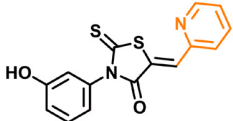
### Known interference candidates

The computational aggregation advisor<sup>4</sup> and compound strings taken from PAINS filters<sup>32,33</sup> (<http://www.rdkit.org>) were used to search the MMSs for known assay interference candidates. The 14,646 MMS compounds contained 783 aggregators (on the basis of 100% similarity) and 2381 compounds with PAINS substructures. There were 611 MMSs with one or more aggregators, 1139 MMSs with one or more PAINS, and 126 MMSs including aggregators and PAINS. However, 5065 MMSs with high hit rates did not contain known compounds with aggregation potential or PAINS substructures. Thus, the MMSs provide a large source of analogs for the exploration of other interference candidates, as well as compounds with true multi-assay/target activities.

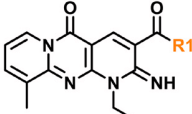
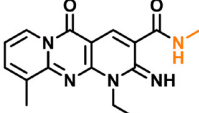
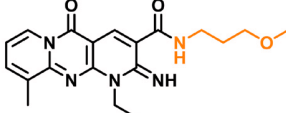
### Exemplary series

**Figure 4** shows exemplary compound pairs and triplets with high assay promiscuity. The two analogs in **Figure 4a** were tested in

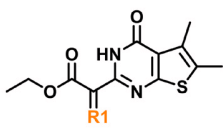
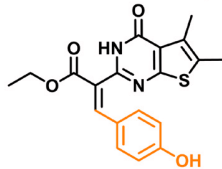
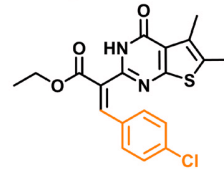
**A**

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

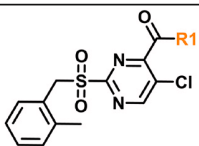
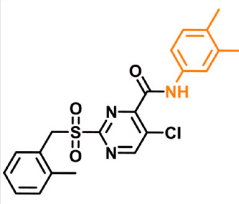
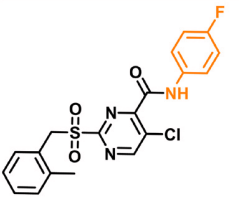
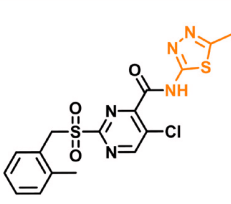
**B**

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

**C**

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

**D**

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

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

more than 380 assays with 93.5% assay overlap and only 1.6% inconsistent assays, yielding comparable hit rates of 2.8% and 3.2%, respectively, resulting in an MMS hit rate of 4.5%. These analogs contained a classical PAINS substructure (ene\_rhodanine)<sup>8,9</sup>. Furthermore, compounds in Figure 4b were analogs of a molecule with aggregation potential<sup>4</sup>. They were tested in more than 300 and 400 assays, respectively, yielding a relatively low assay overlap of 59%, and had hit rates of 2.2% and 2.6%, respectively, resulting in a low MMS hit rate of 2.9%. Thus, these

analog were far from being consistently active, as one might assume for strong aggregators. In Figure 4c, a pair of thieno[2,3-d]pyrimidine-2-acetic acid ethyl ester analogs is shown that were tested in 442 assays with large overlap. These compounds had high hit rates of 5.9% and 7.9%, respectively, resulting in a high MMS hit rate of 9.8%. Moreover, Figure 4d shows a triplet of sulfonylpyrimidines that were tested in 357–361 assays with 89.7% overlap, having very high hit rates of 8.7% (one analog) and more than 13% (two analogs). The analogs in



Figure 4c and Figure 4d have previously not been classified as interference candidates. The interference potential of sulfonylpyrimidines was assessed via a SciFinder substructure search for 2-[(phenylmethyl)sulfonyl]-pyrimidine. This substructure appeared in more than 100 publications related to biological studies and more than 1000 chemical reactions. Although the potential of sulfonylpyrimidines to undergo nucleophilic aromatic substitutions in organic synthesis is well established in the literature<sup>34,35</sup>, reactivities under assay conditions remain to be confirmed experimentally. Compounds forming each of the MMSs in Figure 4 displayed consistent hit rate characteristics, hence assigning confidence to their observed activity phenotype. Taken together, these examples of analog pairs and triplets (i.e., minimally sized MMSs) are indicative of the potential of well characterized MMSs for follow-up investigations focusing on assay interference and multi-target activities.

## Conclusions

Herein, a detailed analysis of hit rates of nearly 440,000 extensively assayed screening compounds has been presented. On the basis of hit rate distributions, 12.7% of the compounds with highest hit rates were selected. From these compounds, analog series with single substitution sites were systematically extracted to complement hit rate statistics with the assessment of structural relationships between active compounds. A total of 6941 unique MMSs were obtained comprising 14,646 compounds. These MMSs were characterized using different parameters prioritizing high-confidence series for activity analysis. A major goal of our study has been the data-driven generation of a pool of analog series for the evaluation of assay

interference potential and multi-target activities. More than 5000 MMSs did not contain known interference candidates, providing an opportunity to evaluate compounds with interference potential on a large scale. In the next step, analog series will be evaluated from a medicinal chemistry perspective to complement and further extend statistical considerations. Annotated series and associated assay/target information will then be made freely available. The statistics and selection steps reported herein also make it possible to regenerate compound subsets at different hit rate levels and subject them to further analysis. In addition, large numbers of compounds with high hit rates that were not part of MMSs are also available. For reasons discussed, our preferred approach is taking compound series information into account when judging assay promiscuity.

## Data availability

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

## Competing interests

No competing interests were declared.

## Grant information

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

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

## References

- Aldrich C, Bertozzi C, Georg GI, *et al.*: **The Ecstasy and Agony of Assay Interference Compounds.** *ACS Cent Sci.* 2017; 3(3): 143–147.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McGovern SL, Caselli E, Grigorieff N, *et al.*: **A Common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening.** *J Med Chem.* 2002; 45(8): 1712–1722.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Shoichet BK: **Screening in a spirit haunted world.** *Drug Discov Today.* 2006; 11(13–14): 607–615.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Irwin JJ, Duan D, Torosyan H, *et al.*: **An Aggregation Advisor for Ligand Discovery.** *J Med Chem.* 2015; 58(17): 7076–7087.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Feng BY, Simeonov A, Jadhav A, *et al.*: **A high-throughput screen for aggregation-based inhibition in a large compound library.** *J Med Chem.* 2007; 50(10): 2385–2390.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Jadhav A, Ferreira RS, Klumpp C, *et al.*: **Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease.** *J Med Chem.* 2010; 53(1): 37–51.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ferreira RS, Simeonov A, Jadhav A, *et al.*: **Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors.** *J Med Chem.* 2010; 53(13): 4891–4905.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Baell JB, Holloway GA: **New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.** *J Med Chem.* 2010; 53(7): 2719–2740.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Baell J, Walters MA: **Chemistry: Chemical con artists foil drug discovery.** *Nature.* 2014; 513(7519): 481–483.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Dahlin JL, Nissink JW, Strasser JM, *et al.*: **PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS.** *J Med Chem.* 2015; 58(5): 2091–2113.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dahlin JL, Nissink JW, Francis S, *et al.*: **Post-HTS case report and structural alert: Promiscuous 4-aryl-1,5-disubstituted-3-hydroxy-2H-pyrrol-2-one actives verified by ALARM NMR.** *Bioorg Med Chem Lett.* 2015; 25(21): 4740–4752.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gilberg E, Jasial S, Stumpfe D, *et al.*: **Highly Promiscuous Small Molecules from Biological Screening Assays Include Many Pan-Assay Interference Compounds but Also Candidates for Polypharmacology.** *J Med Chem.* 2016; 59(22): 10285–10290.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Baell JB: **Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS).** *J Nat Prod.* 2016; 79(3): 616–628.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bisson J, McAlpine JB, Friesen JB, *et al.*: **Can Invalid Bioactives Undermine Natural Product-Based Drug Discovery?** *J Med Chem.* 2016; 59(5): 1671–1690.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nelson KM, Dahlin JL, Bisson J, *et al.*: **The Essential Medicinal Chemistry of Curcumin.** *J Med Chem.* 2017; 60(5): 1620–1637.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang Y, Xiao J, Suzek TO, *et al.*: **PubChem's BioAssay Database.** *Nucleic Acids Res.* 2012; 40(Database issue): D400–D412.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

17. Capuzzi SJ, Muratov EN, Tropsha A: **Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay Interference Compounds**. *J Chem Inf Model*. 2017; **57**(3): 417–427.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Jasial S, Hu Y, Bajorath J: **How Frequently Are Pan-Assay Interference Compounds Active? Large-Scale Analysis of Screening Data Reveals Diverse Activity Profiles, Low Global Hit Frequency, and Many Consistently Inactive Compounds**. *J Med Chem*. 2017; **60**(9): 3879–3886.  
[PubMed Abstract](#) | [Publisher Full Text](#)
19. Gilberg E, Stumpfe D, Bajorath J: **Activity profiles of analog series containing pan assay interference compounds**. *RSC Adv*. 2017; **7**(57): 35638–35649.  
[Publisher Full Text](#)
20. Hu Y, Bajorath J: **Compound promiscuity: what can we learn from current data?** *Drug Discov Today*. 2013; **18**(13–14): 644–650.  
[PubMed Abstract](#) | [Publisher Full Text](#)
21. Paolini GV, Shapland RH, van Hoorn WP, *et al.*: **Global mapping of pharmacological space**. *Nat Biotechnol*. 2006; **24**(7): 805–815.  
[PubMed Abstract](#) | [Publisher Full Text](#)
22. Boran AD, Iyengar R: **Systems approaches to polypharmacology and drug discovery**. *Curr Opin Drug Discov Devel*. 2010; **13**(3): 297–309.  
[PubMed Abstract](#) | [Free Full Text](#)
23. Jasial S, Hu Y, Bajorath J: **Determining the Degree of Promiscuity of Extensively Assayed Compounds**. *PLoS One*. 2016; **11**(4): e0153873.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Griffen E, Leach AG, Robb GR, *et al.*: **Matched molecular pairs as a medicinal chemistry tool**. *J Med Chem*. 2011; **54**(22): 7739–7750.  
[PubMed Abstract](#) | [Publisher Full Text](#)
25. Hussain J, Rea C: **Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets**. *J Chem Inf Model*. 2010; **50**(3): 339–348.  
[PubMed Abstract](#) | [Publisher Full Text](#)
26. Wawer M, Bajorath J: **Local structural changes, global data views: graphical substructure-activity relationship trailing**. *J Med Chem*. 2011; **54**(8): 2944–2951.  
[PubMed Abstract](#) | [Publisher Full Text](#)
27. Lewell XQ, Judd DB, Watson SP, *et al.*: **RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry**. *J Chem Inf Comput Sci*. 1998; **38**(3): 511–522.  
[PubMed Abstract](#) | [Publisher Full Text](#)
28. de la Vega de León A, Bajorath J: **Matched molecular pairs derived by retrosynthetic fragmentation**. *Med Chem Commun*. 2014; **5**(1): 64–67.  
[Publisher Full Text](#)
29. Hu X, Hu Y, Vogt M, *et al.*: **MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs**. *J Chem Inf Model*. 2012; **52**(5): 1138–1145.  
[PubMed Abstract](#) | [Publisher Full Text](#)
30. Berthold MR, Cebron N, Dill F, *et al.*: **KNIME: The Konstanz Information Miner**. In *Studies in Classification, Data Analysis, and Knowledge Organization*. Preisach C, Burkhardt H, Schmidt-Thieme L and Decker R, Eds.; Springer: Berlin, Germany, 2008; 319–326.  
[Publisher Full Text](#)
31. OEChem TK: **OpenEye Scientific Software**. Inc.: Santa Fe, NM, U.S., 2012.
32. Sterling T, Irwin JJ: **ZINC 15--Ligand Discovery for Everyone**. *J Chem Inf Model*. 2015; **55**(11): 2324–2337.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Gaulton A, Bellis LJ, Bento AP, *et al.*: **ChEMBL: a large-scale bioactivity database for drug discovery**. *Nucleic Acids Res*. 2012; **40**(Database issue): D1100–D1107.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Font D, Linden A, Heras A, *et al.*: **A simple approach for the regioselective synthesis of imidazo[1,2-a]pyrimidones and pyrimido[1,2-a]pyrimidinones**. *Tetrahedron*. 2006; **62**(7): 1433–1443.  
[Publisher Full Text](#)
35. Krueger AC, Madigan DL, Beno DW, *et al.*: **Novel hepatitis C virus replicon inhibitors: synthesis and structure-activity relationships of fused pyrimidine derivatives**. *Bioorg Med Chem Lett*. 2012; **22**(6): 2212–2215.  
[PubMed Abstract](#) | [Publisher Full Text](#)



## Open Peer Review

Current Referee Status:   

---

### Version 2

Referee Report 10 October 2017

doi:10.5256/f1000research.13934.r26808



**John A. Lowe III** 

Jl3pharma LLC, Stonington, CT, USA

**Competing Interests:** No competing interests were disclosed.

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

Referee Report 10 October 2017

doi:10.5256/f1000research.13934.r26809



**José L Medina-Franco** , **Fernanda I. Saldívar-González**

Department of Pharmacy, Faculty of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico

The authors have addressed the minor points raised by the referees.

**Competing Interests:** No competing interests were disclosed.

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

---

### Version 1

Referee Report 02 October 2017

doi:10.5256/f1000research.13396.r25426



**Michael Walters** 

Department of Medicinal Chemistry, University of Minnesota, Minneapolis, MN, USA

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

Notes:

1. Inconsistent is misspelled in Panel B of Figure 3.
2. Though this will need to be experimentally-verified, the sulfonylpyrimidines are almost certainly interfering by  $S_NAr$  reactions. A literature search to gauge this potential reactivity of the core structure retrieved >100 articles and >1000 reactions. This group is certainly not in the PAINS definitions, but its assay interference potential can readily be ascertained by a quick reaction substructure search. Most importantly, this observation reinforces that lack of PAINS "flagging" is not sufficient to ensure that compounds don't have assay interference potential. The authors may wish to highlight this in their discussion.

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

Yes

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

Yes

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

Yes

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

Yes

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

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** assay interference, medicinal chemistry

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

Referee Report 12 September 2017

doi:[10.5256/f1000research.13396.r25674](https://doi.org/10.5256/f1000research.13396.r25674)



**José L Medina-Franco** , **Fernanda I. Saldívar-González**

Department of Pharmacy, Faculty of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico

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

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

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

Minor suggestions to further improve the quality of the manuscript:

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

Other suggestions (per section):

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

## References

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

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

Yes

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

Yes

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

Yes

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

Yes

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

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

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

Referee Report 18 August 2017

doi:10.5256/f1000research.13396.r25140



**John A. Lowe III** 

Jl3pharma LLC, Stonington, CT, USA

This article makes an important contribution to addressing a serious complication in screening for new biological activity, especially in the context of new drug discovery. Several researchers continue to point out the waste of time and money in following up on false positive hits from biological screening programs. While paradigms to filter out false positives, such as aggregators or other “frequent hitters” such as PAINs, are becoming mainstream techniques, these may not be sufficient, either because they miss some false positives or because they mistakenly classify legitimate hits as false positives. By carrying out a statistical analysis of a large set of screening data for hit rates using matched molecular pairs and series, this article offers a valuable perspective on this issue.

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

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

Yes

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

Yes

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

Yes

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

Yes

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

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

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

---

## Discuss this Article

Version 1

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

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

Thank you for your comment and interest. We are still investigating the systematic extraction of analog series from data sets of any composition and size and intend finalizing this methodological work before addressing code release issues. For the time being, the interested reader is referred to a variety of partly related implementations we have made freely available on the Zenodo open access platform.

On a more general note, although we are among the computational groups promoting open science by public release of many in-house generated data sets and software tools, our experiences with free data and tool sharing have not been entirely positive; another point of consideration going forward. Perhaps it might make sense to (re-)consider other collaborative models and release procedures.

**Competing Interests:** None

Reader Comment 06 Sep 2017

**Greg Landrum**, T5 Informatics GmbH, Switzerland

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

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

**Competing Interests:** No competing interests.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

F1000Research

## Conclusions

A data set of MMSs for the systematic assessment of assay interference and multitarget activities was generated. Based on compounds that showed highest activity in primary biological screening assays, 6941 unique analog series were systematically extracted that comprised 14,646 compounds. To ensure statistically sound and chemically meaningful results, series hit rates, assay overlap, and inconsistent experimental activities between analogs were considered for the prioritization of MMSs. Data sets were made freely available for follow-up investigations. Notably, in the generation of MMSs, PAINS alerts were omitted. Thus, by circumventing previously discussed shortcomings of PAINS filters, a solely data-driven assessment of assay interference and multitarget activities was made possible.

The provision of a dataset containing exceptionally promiscuous compounds and series completed the analysis of assay interference and multitarget activity based on biological screening data. To provide confirmatory evidence for true multitarget activity that is not caused by chemical interference, a structure-based approach is applied in the next chapter. Based on structurally confirmed promiscuity, template structures for multitarget and multifamily ligand design are generated.





# 7 X-ray-Structure-Based Identification of Compounds with Activity Against Targets from Different Families and Generation of Templates for Multitarget Ligand Design

## Introduction

The interpretation of compound promiscuity on the basis of assay data is a challenging task considering the difficulty of discriminating artificial results from real multitarget activities. One way to receive confirmatory evidence for true multitarget binding events is provided by the analysis of X-ray structures of ligand-target complexes containing promiscuous small molecules.

In this chapter, crystallographic ligands are identified that form multiple complexes with distantly or unrelated targets. For these multifamily ligands, analogs and additional target annotations are systematically explored in the medicinal chemistry literature. From analog series of multifamily ligands, molecular scaffolds are derived to provide references for the design of multitarget ligands.

My contribution to this work was the identification and analysis of crystallographic multifamily ligands and the generation of analog series based scaffolds.

Reprinted with permission from 'E. Gilberg, D. Stumpfe, J. Bajorath. X-ray-Structure-Based Identification of Compounds with Activity against Targets from Different Families and Generation of Templates for Multitarget Ligand Design. *ACS Omega* **2018**, 3, 106–111'. Copyright 2018 American Chemical Society



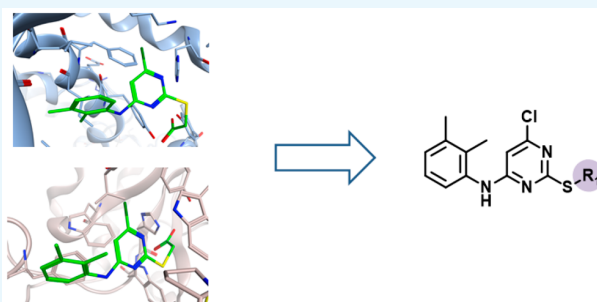
# X-ray-Structure-Based Identification of Compounds with Activity against Targets from Different Families and Generation of Templates for Multitarget Ligand Design

Erik Gilberg,<sup>†,‡</sup> Dagmar Stumpfe,<sup>†</sup> and Jürgen Bajorath<sup>\*,†,‡</sup>

<sup>†</sup>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

<sup>‡</sup>Pharmaceutical Institute, Rheinische Friedrich-Wilhelms-Universität, An der Immenburg 4, D-53121 Bonn, Germany

**ABSTRACT:** Compounds with multitarget activity (promiscuity) are increasingly sought in drug discovery. However, promiscuous compounds are often viewed controversially in light of potential assay artifacts that may give rise to false-positive activity annotations. We have reasoned that the strongest evidence for true multitarget activity of small molecules would be provided by experimentally determined structures of ligand–target complexes. Therefore, we have carried out a systematic search of currently available X-ray structures for compounds forming complexes with different targets. Rather unexpectedly, 1418 such crystallographic ligands were identified, including 702 that formed complexes with targets from different protein families (multifamily ligands). About half of these multifamily ligands originated from the medicinal chemistry literature, making it possible to consider additional target annotations and search for analogues. From 168 distinct series of analogues containing one or more multifamily ligands, 133 unique analogue-series-based scaffolds were isolated that can serve as templates for the design of new compounds with multitarget activity. As a part of our study, all of the multifamily ligands we have identified and the analogue-series-based scaffolds are made freely available.



## 1. INTRODUCTION

Over the past decade, the interest in small molecules with multitarget activity has been steadily on the rise,<sup>1–3</sup> especially in the context of polypharmacology.<sup>4–7</sup> This concept refers to increasing evidence that the efficacy of drugs frequently depends on engagement of multiple therapeutic targets.<sup>4–7</sup> Accordingly, the molecular foundation of polypharmacology, which also includes undesired side effects, is provided by specific interactions of compounds with multiple targets.<sup>8</sup> However, while multitarget drug discovery is given prime consideration in therapeutic areas such as neurodegenerative diseases<sup>3</sup> and oncology,<sup>9</sup> compound promiscuity per se is often viewed controversially.<sup>8</sup> This is the case because it is generally difficult to draw the line between true multitarget activity of small molecules<sup>8</sup> and aggregation effects or potential reactivity under assay conditions,<sup>10–13</sup> which may or may not<sup>14,15</sup> lead to artifacts and false-positive assay signals.<sup>13,16,17</sup> Hence, differentiating between multitarget activity and assay interference has become a major task in biological screening and medicinal chemistry.<sup>17</sup> In addition to their drug discovery relevance, small molecules with true multitarget activity are also of high interest for basic research in order to explore why and how such chemical entities are capable of forming specific interactions with multiple targets, especially if these targets are only distantly related or unrelated and have different functions.

We have been interested in identifying compounds that are active against target proteins from different families. In light of potential caveats associated with promiscuity analysis (vide supra), we have reasoned that particularly strong evidence and support for multitarget activity would be provided by structural data confirming that compounds are indeed bound to active sites of different target proteins. Therefore, we have carried out a systematic search for X-ray structures of ligands bound to multiple target proteins from different families. This search was complemented by identifying and analyzing series of analogues involving such ligands, thereby bridging between structural biology and medicinal chemistry.

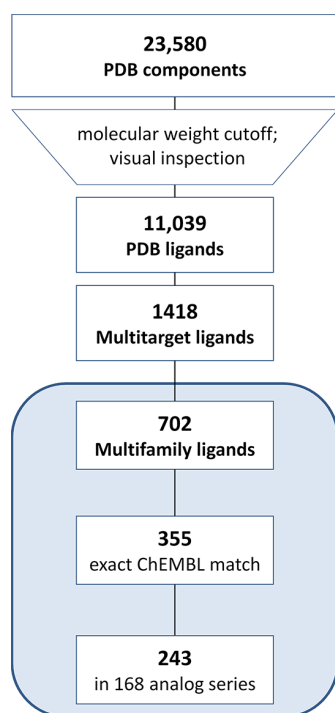
## 2. RESULTS AND DISCUSSION

**2.1. Crystallographic Ligands.** From 102 625 entries in the RCSB Protein Data Bank (PDB),<sup>18</sup> 23 580 crystallographic ligands were extracted, which included 11 039 organic compounds with a molecular weight of at least 300 Da and unique structures. This subset of PDB ligands provided the basis for our analysis. The complete selection protocol is summarized in Figure 1.

**Received:** November 24, 2017

**Accepted:** December 18, 2017

**Published:** January 5, 2018



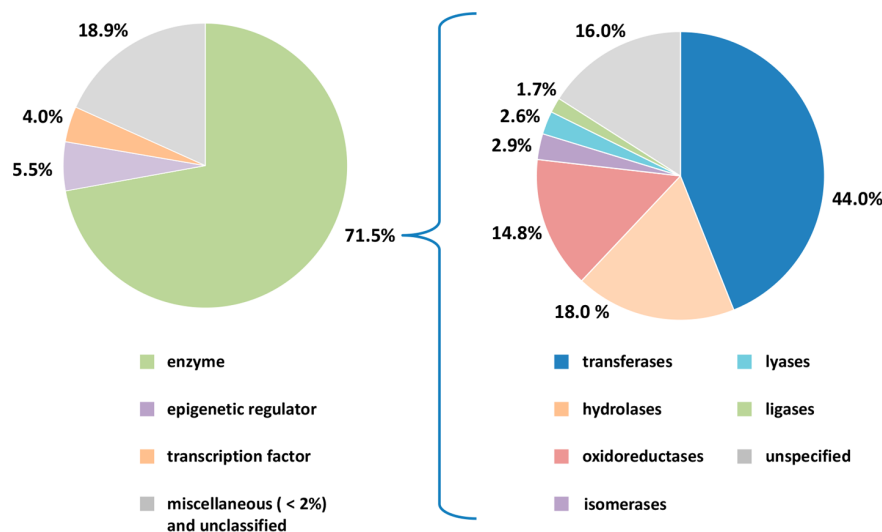
**Figure 1.** Compound selection. The protocol applied to select crystallographic ligands, multitarget and multifamily ligands, and analogues from medicinal chemistry is summarized.

**2.2. Multitarget and Multifamily Ligands.** The selected PDB ligands were found to contain 1418 compounds from X-ray structures of complexes with at least two different target proteins (i.e., multitarget ligands; [Figure 1](#)). We then determined that these multitarget ligands contained a subset of 702 compounds whose crystallographic targets originated from different families (i.e., multifamily ligands; [Figure 1](#)). For this subset, the median value was three targets per ligand. Multifamily ligands were most interesting to us because their structurally confirmed targets were only distantly related (if not

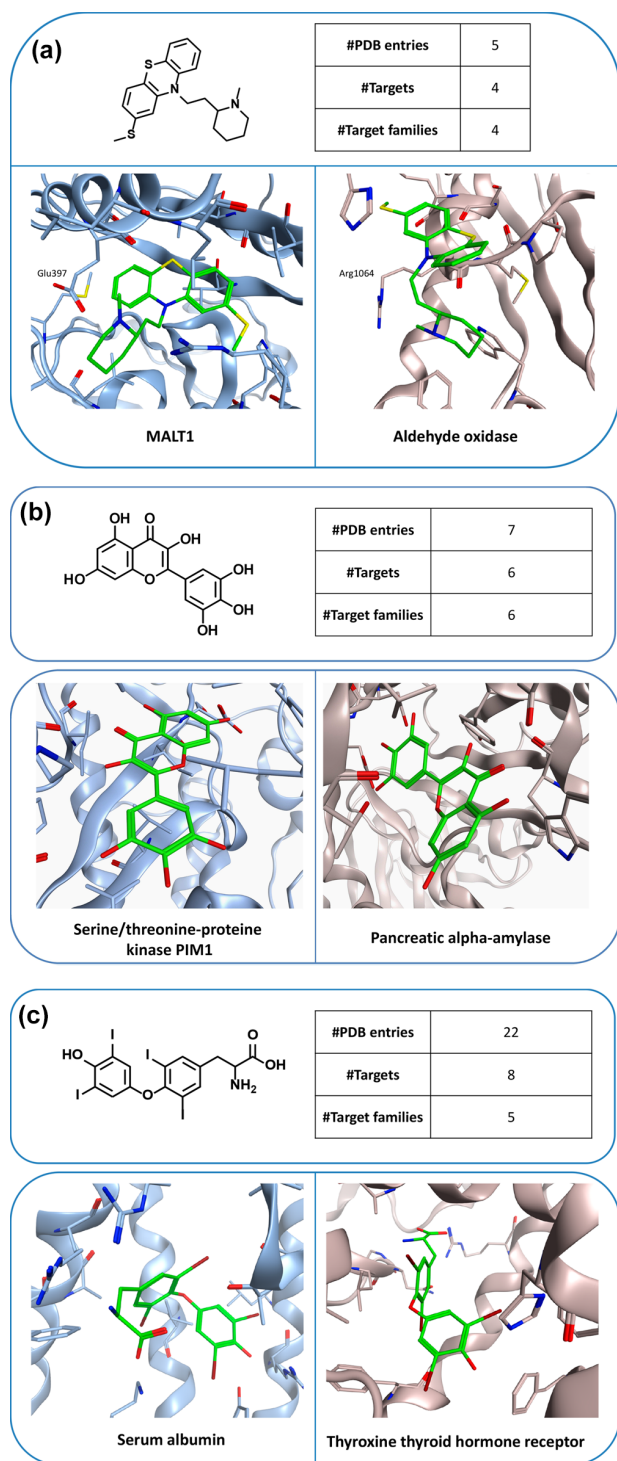
unrelated). Targets of multifamily ligands included 488 human proteins, which were distributed across different families as shown in [Figure 2](#). The majority of targets were enzymes. Among these, transferases were prevalent. This observation can be explained by considering that the composition of the PDB is biased toward targets that are straightforward to crystallize (such as many cytoplasmic enzymes). Consequently, some major classes of pharmaceutical targets such as G-protein-coupled receptors and other membrane proteins continue to be under-represented in the PDB. It is possible to compensate this inherent target bias in part by mapping of multifamily ligands from the PDB to ChEMBL and searching for additional target annotations of these ligands and available structural analogues from medicinal chemistry, as further discussed below.

**2.3. Exemplary Ligands and X-ray Structures.** [Figure 3](#) shows X-ray structures of ligands in complex with targets from different families. Comparison of X-ray structures of the same ligand in complex with different targets frequently revealed differences in binding modes. For instance, the phenothiazine derivative thioridazine shown in [Figure 3a](#) was found in five X-ray complexes with four targets from four different families. As an exemplary comparison, the binding mode of thioridazine observed in mucosa-associated lymphoid tissue lymphoma translocation protein 1 (MALT1),<sup>19</sup> a cysteine protease, clearly differs from the one in aldehyde oxidase,<sup>20</sup> an unrelated enzyme. While the tricyclic ring system of thioridazine is located in a hydrophobic pocket of MALT1, it is partially solvent-exposed in the X-ray complex with aldehyde oxidase. In addition, the positively charged *N*-methylpiperidinyl moiety forms charge-assisted hydrogen bonds with Glu397 of MALT1, whereas the tertiary amine of the ligand forms backbone interactions with the carbonyl oxygen of Arg1064 in the active site of aldehyde oxidase.

[Figure 3b](#) shows an example of an inverted ligand binding mode in two different active sites. The flavonoid myricetin was found in seven complex structures involving six targets from six different families. It displays opposite head-to-tail orientations when bound to human pancreas amylase<sup>21</sup> and the ATP-binding site of PIM1 kinase.<sup>22</sup>



**Figure 2.** Distribution of human targets of multifamily ligands. The pie chart on the left reports the distribution of human targets from complex X-ray structures with multifamily ligands. For enzymes, the distribution of catalytic functions is shown in the pie chart on the right.



**Figure 3.** Multifamily ligands and X-ray structures. In (a–c), exemplary ligands and X-ray structures of their complexes with targets from different families are shown. For each ligand, the total number of complex X-ray structures, the number of PDB targets, and the number of families from which these targets originated are reported. In the X-ray structures, bound ligands are shown in stick representation with standard atom coloring.

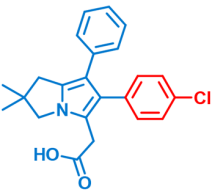
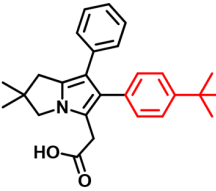
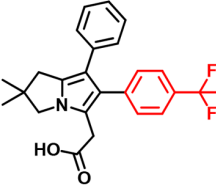
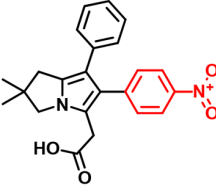
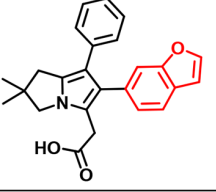
Binding modes can also be compared for multifamily ligands when interactions with different targets lead to desired or undesired functional effects. An example is shown in Figure 3c,

where the thyroid hormone thyroxine (T4) is bound to the IIa subdomain of human serum albumin<sup>23</sup> or the ligand binding domain of thyroxine thyroid hormone receptor beta (TR), its natural receptor.<sup>24</sup> Binding to serum albumin causes hyperthyroxinemia.<sup>23</sup> Notably, T4 reaches deep into the TR binding pocket, where it interacts with three arginine residues via charge-assisted hydrogen bonds. In addition, the iodine atoms of T4 are accommodated in small subsites mostly formed by the side chains of Phe459 and Phe455. By contrast, T4 binds to human serum albumin in a surface-directed manner and predominantly interacts with residues that are partially solvent-exposed.

**2.4. Multifamily Ligands from Medicinal Chemistry.** A subset of 355 of the 702 multifamily ligands were detected in the ChEMBL database,<sup>25</sup> the major public repository of compounds and activity data from the medicinal chemistry literature. For these ligands, ChEMBL target annotations from high-confidence direct binding/inhibition assays were collected. Taking these additional annotations into account represented an expansion into medicinal chemistry target space and increased the median value from three PDB (vide supra) to 17 unique PDB/ChEMBL targets per multifamily ligand. Thus, crystallographic multifamily ligands were generally promiscuous on the basis of medicinal chemistry data. Although it cannot be excluded that some target annotations from assays might be false positives, the availability of multiple X-ray structures of these ligands in complex with different targets lends credence to their promiscuous nature, strongly suggesting their relevance for the study of multitarget activity and polypharmacology.

**2.5. Analogues of Multifamily Ligands.** For the 355 multifamily ligands available in ChEMBL, a systematic search for analogue series (ASs) was carried out. For 243 of these ligands, analogues were detected, yielding 168 unique ASs. Each AS consisted of at least one X-ray ligand and varying numbers of noncrystallographic analogues from ChEMBL. An exemplary AS is depicted in Figure 4. This AS contains an X-ray ligand and several ChEMBL compounds with multitarget annotations, providing corroborating evidence for the promiscuity of the multifamily ligand from the PDB.

**2.6. Scaffolds and Design Templates.** From ASs containing multifamily ligands, analogue series-based (ASB) scaffolds<sup>26,27</sup> were derived. By design, ASB scaffolds take retrosynthetic criteria into account and capture chemical information on compound series, including the conserved substructure and substitution sites where analogues are distinguished.<sup>26,27</sup> For 133 of the 168 ASs with multifamily ligands ASB scaffolds could be derived. Exemplary scaffolds are shown in Figure 5. Since ASs were associated with multiple targets, further extending the set of PDB targets of multifamily ligands, the corresponding ASB scaffolds also represent templates for the design of compounds with different multitarget activities. On the basis of each scaffold, different target combinations can be explored. The ASB scaffolds also make it possible to differentiate between template structures with different degrees of promiscuity. For example, scaffolds from highly promiscuous analogue series, as shown in Figure 5, might be deprioritized as template structures for the design of compounds with desired activity against a few targets, even if these targets are contained in the scaffold-associated target profiles. Instead, scaffolds from other less promiscuous series with desired targets might be considered. Furthermore, for ASB scaffolds with target combinations of interest, it is advisable to inspect the target annotations of individual analogues to

Analogues	Target annotations (Target families)		
	PDB	ChEMBL	Union
	2 (2 <sup>a</sup> )	11 (6 <sup>b</sup> )	13 (6)
	0	3 (2 <sup>c</sup> )	3 (2)
	0	3 (2 <sup>c</sup> )	3 (2)
	0	3 (2 <sup>c</sup> )	3 (2)
	0	3 (2 <sup>c</sup> )	3 (2)

<sup>a</sup>) Phospholipase A2 family, Transferrin family

<sup>b</sup>) Lipoxigenase family, Prostaglandin G/H synthase family, MAPEG family,

Sigma family, Heme-copper respiratory oxidase family, Cytochrome C oxidase subunit 2 family

<sup>c</sup>) Prostaglandin G/H synthase family, MAPEG family

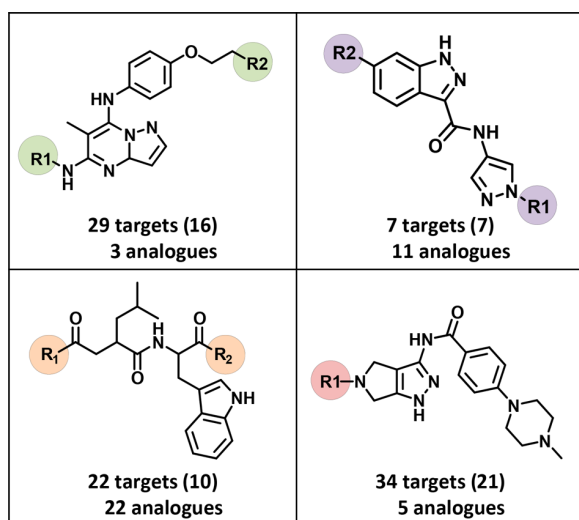
**Figure 4.** Analogue series. Shown is an exemplary AS including a multifamily ligand (blue core). For the crystallographic ligand, the number of PDB targets, the number of targets reported in ChEMBL, and the number of unique targets are given. For each ChEMBL analogue, the number of targets from ChEMBL is provided. In each case, the corresponding number of target families is given in parentheses. ChEMBL analogues have no PDB target annotations. Substituents that distinguish analogues are colored red.

rationalize the series-based target profile in more detail. Analogues can be easily obtained by substructure searching using ASB scaffolds.

**2.7. Conclusions.** We have systematically searched for crystallographic ligands bound to multiple targets from different families. Such X-ray data were thought to provide firm evidence for true multitarget activity of compounds. An unexpectedly large number of qualifying ligands (702) were identified that covered targets from a variety of families. Approximately half of these ligands originated from the medicinal chemistry literature, which yielded additional target annotations. Moreover, a total of 168 distinct series of analogues that contained X-ray ligands were identified. From these, 133 analogue-series-based scaffolds

were extracted that captured chemical and target information on individual series. Crystallographic multifamily ligands represent a large, high-confidence knowledge base for multitarget activity. Scaffolds derived from ASs containing such ligands can be considered as templates for compound design. Therefore, multifamily ligands, scaffolds, and associated target information are made freely available as a part of this study. We also note that a variety of computational methods are available to predict targets of test compounds. The uncertainties associated with target predictions go much beyond experimental uncertainties associated with compound data. However, searching for compounds with true multitarget activities is difficult on the basis of experimental activity data, taking assay-





**Figure 5.** Exemplary scaffolds. Shown are examples of ASB scaffolds representing series of promiscuous structural analogues, including multifamily ligands. For each scaffold, the total number of unique targets against which the analogues were active and (in parentheses) the number of corresponding target families are reported. Substitution sites in ASB scaffolds are highlighted.

dependent activity readouts and potential artifacts into account. For these reasons, X-ray structures of ligand–target complexes provided the initial focal point of our analysis and were complemented by taking medicinal chemistry data into account. By contrast, possible computational predictions were deliberately avoided, given the motivation and scope of our analysis.

### 3. MATERIALS AND METHODS

All calculations were carried out using in-house Perl and Python scripts with the aid of the OpenEye chemistry toolkit,<sup>28</sup> KNIME protocols,<sup>29</sup> and RStudio.<sup>30</sup> X-ray structures were graphically analyzed using the Molecular Operating Environment.<sup>31</sup>

**3.1. Ligands from X-ray Structures.** X-ray structures and associated compound data were extracted from the Ligand Expo section<sup>32</sup> of the PDB.<sup>18</sup> Salts and other buffer components were removed, and ligands with a molecular weight of at least 300 Da yielding unique aromatic nonstereo SMILES<sup>33</sup> representations were retained. Application of the molecular weight cutoff ensured that small organic components and fragments were excluded from further consideration. All of the selected complex X-ray structures were visually inspected.

**3.2. Compounds and Activity Data.** From ChEMBL (release 23)<sup>25</sup> a total of 853 533 unique compounds were extracted for which activity data from direct binding/inhibition assays (target relationship type “D”) were available.

**3.3. Target Family Distribution.** For crystallographic targets of human origin, family assignments were obtained by combining the classification schemes of UniProt<sup>34</sup> and ChEMBL. In addition, known targets of all of the selected ChEMBL compounds were determined on the basis of unique UniProt identifiers.

**3.4. Analogue Series and Scaffolds.** From combined PDB and ChEMBL compounds, ASs were systematically extracted using a recently developed algorithm<sup>35</sup> utilizing the matched molecular pair (MMP) formalism.<sup>36</sup> An MMP is

defined as a pair of compounds that are distinguished only by a structural change at a single site,<sup>36</sup> often termed a chemical transformation.<sup>37</sup> To generate MMPs, compounds were systematically fragmented<sup>37</sup> according to retrosynthetic rules,<sup>38</sup> yielding RECAP-MMPs.<sup>39</sup> From ASs, recently introduced ASB scaffolds<sup>26,27</sup> were extracted, which capture the conserved substructure of a series and all substitution sites.

**3.5. Data Deposition.** All of the multifamily ligands have been made available, together with their crystallographic targets, PDB identifiers, and total numbers of targets, including annotations from ChEMBL (if available). In addition, all of the ASB scaffolds derived from ASs containing multifamily ligands are provided. The collection of ligands and scaffolds is freely available in a deposition on the Zenodo open access platform.<sup>40</sup>

### ■ AUTHOR INFORMATION

#### Corresponding Author

\*Phone: 49-228-2699-306. E-mail: [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de).

#### ORCID

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

#### Author Contributions

The study was carried out and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.

#### Notes

The authors declare no competing financial interest.

### ■ ACKNOWLEDGMENTS

We are grateful to OpenEye Scientific Software, Inc., for the free academic license of the OpenEye Toolkits. D.S. was supported by Sonderforschungsbereich 704 of the Deutsche Forschungsgemeinschaft.

### ■ REFERENCES

- Zimmermann, G. R.; Lehar, J.; Keith, C. T. Multi-Target Therapeutics: When the Whole is greater than the Sum of the Parts. *Drug Discovery Today* **2007**, *12*, 34–42.
- 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.
- Geldenhuis, W. J.; Van der Schyf, C. J. Designing Drugs with Multi-Target Activity: The Next Step in the Treatment of Neurodegenerative Disorders. *Expert Opin. Drug Discovery* **2013**, *8*, 115–129.
- Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874–7887.
- Bolognesi, M. L. Polypharmacology in a Single Drug: Multitarget Drugs. *Curr. Med. Chem.* **2013**, *20*, 1639–1645.
- Bolognesi, M. L.; Cavalli, A. Multitarget Drug Discovery and Polypharmacology. *ChemMedChem* **2016**, *11*, 1190–1192.
- Rosini, M. Polypharmacology: The Rise of Multitarget Drugs over Combination Therapies. *Future Med. Chem.* **2014**, *6*, 485–487.
- Hu, Y.; Bajorath, J. Compound Promiscuity - What Can We Learn From Current Data. *Drug Discovery Today* **2013**, *18*, 644–650.
- Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130–137.
- 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.
- Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
- 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.

(13) Baell, J. B.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.

(14) 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.

(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) 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.

(17) 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.

(18) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(19) Schlauderer, F.; Lammens, K.; Nagel, D.; Vincendeau, M.; Eitelhuber, A. C.; Verhelst, S. H.; Kling, D.; Chrusciel, A.; Ruland, J.; Krappmann, D.; Hopfner, K. P. Structural Analysis of Phenothiazine Derivatives as Allosteric Inhibitors of the MALT1 Paracaspase. *Angew. Chem., Int. Ed.* **2013**, *52*, 10384–10387.

(20) Coelho, C.; Foti, A.; Hartmann, T.; Santos-Silva, T.; Leimkühler, S.; Romao, M. J. Structural Insights into Xenobiotic and Inhibitor Binding to Human Aldehyde Oxidase. *Nat. Chem. Biol.* **2015**, *11*, 779–783.

(21) Williams, L. K.; Li, C.; Withers, S. G.; Brayer, G. D. Order and Disorder: Differential Structural Impacts of Myricetin and Ethyl Caffate on Human Amylase, an Antidiabetic Target. *J. Med. Chem.* **2012**, *55*, 10177–10186.

(22) Holder, S.; Zemskova, M.; Zhang, C.; Tabrizid, M.; Bremer, R.; Neidigh, J. W.; Lilly, M. B. Characterization of a Potent and Selective Small-Molecule Inhibitor of the PIM1 Kinase. *Mol. Cancer Ther.* **2007**, *6*, 163–172.

(23) Petitpas, I.; Petersen, C. E.; Ha, C. E.; Bhattacharya, A. A.; Zunszain, P. A.; Ghuman, J.; Bhagavan, N. V.; Curry, S. Structural Basis of Albumin-Thyroxine Interactions and Familial Dysalbuminemic Hyperthyroxinemia. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 6440–6445.

(24) Sandler, B.; Webb, P.; Apriletti, J. W.; Huber, B. R.; Togashi, M.; Cunha Lima, S. T.; Juric, S.; Nilsson, S.; Wagner, R.; Fletterick, R. J.; Baxter, J. D. Thyroxine-Thyroid Hormone Receptor Interactions. *J. Biol. Chem.* **2004**, *279*, 55801–55808.

(25) 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.

(26) Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. ASB Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. *Future Science OA* **2016**, *2*, FSO149.

(27) Dimova, D.; Stumpfe, D.; Bajorath, J. Computational Design of New Molecular Scaffolds for Medicinal Chemistry, Part II: Generalization of Analog Series-Based Scaffolds. *Future Sci. OA* **2017**, FSO267.

(28) OEChem, version 1.7.7; OpenEye Scientific Software: Santa Fe, NM, 2012.

(29) Berthold, M. R.; Cebon, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, 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, 2008; pp 319–326.

(30) RStudio: *Integrated Development Environment for R*; RStudio, Inc.: Boston, MA, 2016.

(31) *Molecular Operating Environment (MOE)*, version 2014.09; Chemical Computing Group: Montreal, QC, 2017.

(32) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155. <http://ligand-expo.rcsb.org/> (accessed Sept 28, 2017).

(33) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(34) The UniProt Consortium. UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

(35) 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. *J. Med. Chem.* **2016**, *59*, 7667–7676.

(36) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.

(37) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739–7750.

(38) 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 Application in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

(39) De la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *MedChemComm* **2014**, *5*, 64–67.

(40) <https://zenodo.org/record/1116185> (accessed Nov 24, 2017).



## Conclusions

Templates for a polypharmacologically oriented ligand design based on compounds with structurally confirmed multifamily activities were presented. 702 ligands were demonstrated to interact with various targets from different families. The target space of these ligands was extended by considering target annotations from high-confidence data of binding and inhibition assays. For 168 multifamily ligands, ASs were generated by considering structural analogs that were found in the medicinal chemistry literature. From these series, 133 ASB scaffolds were isolated.

This study provides a starting point for the use of crystallographic data to design compounds that can be used for polypharmacology. In this context, it would be of interest to better understand the molecular basis for true multitarget activity. In the last chapter, the concept of ligand promiscuity was further analyzed on the basis of experimental structures to rationalize and compare multifamily interactions in detail.



# 8 Promiscuous Ligands from Experimentally Determined Structures, Binding Conformations, and Protein Family Dependent Interaction Hotspots

## Introduction

True multitarget activity provides the basis for polypharmacology and is of increasing relevance for drug discovery. However, promiscuity as the molecular basis of polypharmacology requires careful consideration. In light of the complex nature of polypharmacology, rational design of multitarget ligands is a challenging task. Although structural data are limited compared to assay data, the investigation of promiscuous ligands based on crystallographic complexes provides a major advantage because these binding events are confirmed at the molecular level and can be explored as such.

In the following, binding modes and characteristics of multifamily ligands are systematically explored to understand ligand promiscuity across different target families. Molecular properties of promiscuous ligands are determined and compared to other PDB compounds. Moreover, binding conformations of multifamily ligands are analyzed and family dependent interaction hotspots are identified.

Reprinted with permission from 'E. Gilberg, M. Gütschow, J. Bajorath. Promiscuous Ligands from Experimentally Determined Structures, Binding Conformations, and Protein Family-Dependent Interaction Hotspots. *ACS Omega* **2019**, *4*, 1729–1737.' Copyright 2019 American Chemical Society



# Promiscuous Ligands from Experimentally Determined Structures, Binding Conformations, and Protein Family-Dependent Interaction Hotspots

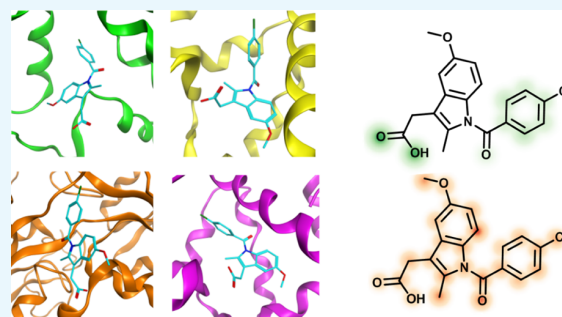
Erik Gilberg,<sup>†,‡</sup> Michael Gütschow,<sup>‡,§</sup> and Jürgen Bajorath<sup>\*,†,§</sup>

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

<sup>‡</sup>Pharmaceutical Institute, Rheinische Friedrich-Wilhelms-Universität, An der Immenburg 4, D-53121 Bonn, Germany

## Supporting Information

**ABSTRACT:** Compound promiscuity is often attributed to nonspecific binding or assay artifacts. On the other hand, it is well-known that many pharmaceutically relevant compounds are capable of engaging multiple targets in vivo, giving rise to polypharmacology. To explore and better understand promiscuous binding characteristics of small molecules, we have searched X-ray structures (and very few qualifying solution structures) for ligands that bind to multiple distantly related or unrelated target proteins. Experimental structures of a given ligand bound to different targets represent high-confidence data for exploring promiscuous binding events. A total of 192 ligands were identified that formed crystallographic complexes with proteins from different families and for which activity data were available. These “multifamily” compounds included endogenous ligands and were often more polar than other bound compounds and active in the submicromolar range. Unexpectedly, many promiscuous ligands displayed conserved or similar binding conformations in different active sites. Others were found to conformationally adjust to binding sites of different architectures. A comprehensive analysis of ligand–target interactions revealed that multifamily ligands frequently formed different interaction hotspots in binding sites, even if their bound conformations were similar, thus providing a rationale for promiscuous binding events at the molecular level of detail. As a part of this work, all multifamily ligands we have identified and associated activity data are made freely available.



## 1. INTRODUCTION

Compound optimization efforts in medicinal chemistry traditionally aim to develop drug candidates that are highly selective and potent toward a specific biological target. This principle is based upon the assumption that therapeutic effects following drug administration solely result from interactions with a single target. However, this paradigm was called into question and revised when it became evident that the efficacy of drugs, but also side effects, frequently depended on multitarget activities and associated functional consequences, a concept referred to as “polypharmacology”.<sup>1–6</sup>

Despite the relevance of polypharmacology for drug efficacy, compounds with promiscuous binding behavior are often viewed controversially.<sup>7,8</sup> This is the case because high hit rates of small molecules in biological assays are frequently not the result of multiple binding events.<sup>9</sup> Rather, aggregation effects and potential chemical reactivities under assay conditions can lead to false positive assay signals.<sup>9–12</sup> In light of concerns about such artifacts, studying multitarget activities of ligands and differentiating between false positive and true positive interactions have become important tasks in medicinal chemistry and biological screening.<sup>13–17</sup>

In addition to their relevance for drug development, the study of promiscuous small molecules is also of high interest in basic research. Importantly, physiological effects of endogenous chemical entities such as coenzymes, substrates, or transmitters are often elicited because of their ability to interact with distantly related or unrelated proteins having diverse functions.<sup>18,19</sup> Hence, “true” promiscuity represents an evolutionary principle for physiologically relevant ligands. However, the molecular basis of promiscuous binding events remains to be further explored.

Although the ligand specificity paradigm will continue to play an important role in drug discovery, there are many opportunities to utilize polypharmacology.<sup>3</sup> For example, multitarget compounds used for the treatment of a given pathology might be repositioned for other therapeutic applications that require engagement of different targets.<sup>20</sup> A text book example of such repurposing efforts is methotrexate, a drug used for many years in cancer treatment, which has

**Received:** December 12, 2018

**Accepted:** January 10, 2019

**Published:** January 22, 2019

recently found alternative low-dose applications in the treatment of inflammatory disorders like psoriasis and rheumatoid arthritis.<sup>21</sup> Notably, polypharmacology has high potential for treatment of diseases that result from perturbation of target networks and associated signaling pathways. Promiscuous kinase inhibitors successfully used in oncology are prime examples for compounds that interfere with target networks and their signaling cascades.<sup>22</sup>

Given the complex nature of polypharmacology, rational design of multitarget ligands is an equally challenging and attractive area of research.<sup>3,7,23–25</sup> To this end, several studies have attempted to determine structure–activity relationship profiles of multitarget compounds. For example, on the basis of publicly available activity data, compounds with multitarget activity were identified and similarity relationships between them were explored.<sup>25–27</sup> Furthermore, X-ray structures were used to associate multitarget drugs with proteins having similar functions,<sup>28</sup> relate multitarget activities of ligands to protein binding site similarity,<sup>29</sup> or identify compounds bound to targets from different families (multifamily ligands).<sup>30</sup> Although structural data are limited, studying multitarget and multifamily ligands on the basis of complex X-ray structures, rather than assay data, has the intrinsic advantage that these binding events are confirmed at the molecular level of detail and can be investigated as such.

Herein, we have searched for multifamily ligands with available X-ray [or nuclear magnetic resonance (NMR)] structures to better understand origins of ligand promiscuity across different target families. Therefore, we have carried out a systematic search for experimental structures of small molecules bound to multiple targets from different protein families. A set of structure-based multifamily compounds was identified that included endogenous ligands as well as approved drugs. Molecular properties and bound conformations of these multifamily ligands were systematically analyzed and interaction hotspots in different protein binding sites were identified. Taken together, the results of our analysis shed light on the ability of small molecules to interact with distantly or unrelated targets.

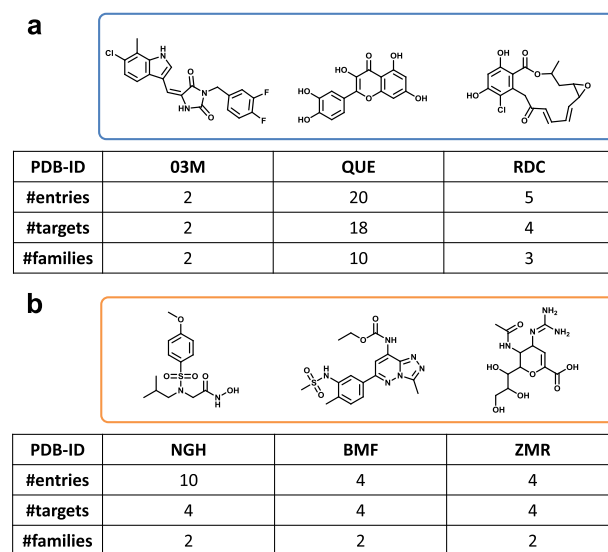
## 2. RESULTS AND DISCUSSION

**2.1. Identification and Characterization of Multifamily Ligands.** From 112 212 structures (entries) available in the Protein Data Bank (PDB),<sup>31</sup> 26 073 bound ligands were extracted. These ligands included 6496 organic compounds with a molecular weight of at least 300 Da and one or more reported activity values (in original references) of at least 10  $\mu\text{M}$  ( $\text{pIC}_{50}$ ,  $\text{pK}_i$ , or  $\text{pK}_d \geq 5$ ). This set of PDB ligands provided the basis of our study.

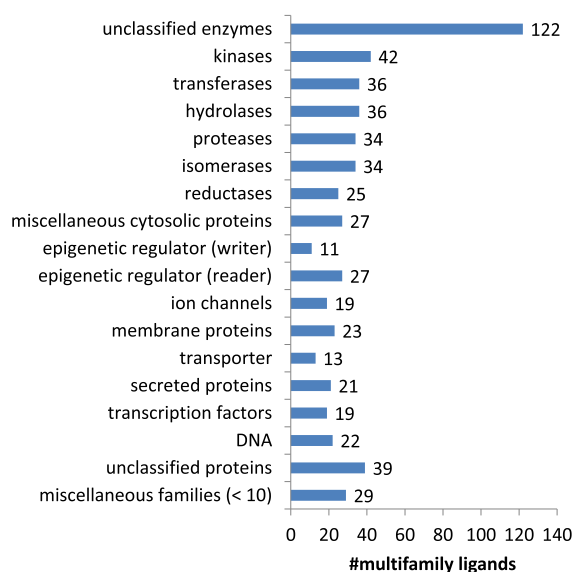
The preselected ligands were subjected to a two-stage analysis. First, target family assignments were computationally carried out in a consistent manner (without subjective intervention) to identify ligands that were active against different target families and ensure reproducibility of the analysis (see [Materials and Methods](#)). Second, for each designated multifamily ligand, assigned targets and binding domains were carefully compared to examine similarities between targets from different families and prioritize multifamily ligands for promiscuity analysis, as further discussed below.

Computational analysis of the preselected PDB ligands identified 192 compounds that formed complexes with a variety of target proteins from 2 to 16 different families. These

192 compounds were designated multifamily ligands and further analyzed. [Figure 1](#) shows exemplary compounds and



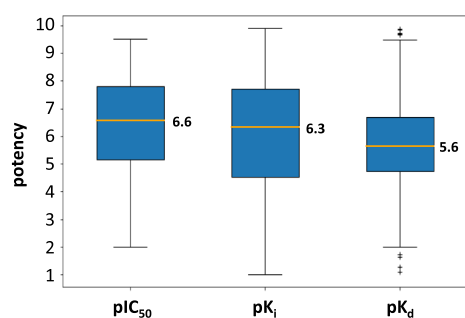
**Figure 1.** Exemplary multifamily ligands. For each ligand, the PDB Ligand Expo identifier (PDB-ID) is given and the number of qualifying complex X-ray structures (entries), targets, and target families is reported. Shown are exemplary ligands from (a) subset IV and (b) subset III.



**Figure 2.** Distribution of multifamily ligands over target families. The bar plot reports the distribution of multifamily ligands over families of crystallographic targets according to the ChEMBL protein family classification scheme.

[Figure 2](#) shows the distribution of multifamily ligands over protein families. Kinases and other transferases formed the largest number of complexes with multifamily ligands (with 42 and 36 ligands, respectively). The majority of complex structures involved cytosolic enzymes, which are over-represented in the PDB because of ease of crystallization.

Multifamily ligands were available in complexes with 2 to 131 crystallographic targets, with a median value of 3 unique targets per ligand. The 192 ligands were represented by a total of 3398 complex structures. These structures only included 20 solution (NMR) structures of ligand–protein complexes and 34 NMR structures of ligand–DNA complexes (for completeness, DNA was included as a biological target). The small number of solution structures only entered the initial statistical analysis of multifamily ligands. Subsequent analysis was focused on X-ray structures. Distributions of potency ( $pIC_{50}$ ,  $pK_i$ , or  $pK_d$ ) values of the 192 multifamily ligands for X-ray targets are shown in Figure 3. The distributions were broad and interquartile ranges spanned several orders of magnitude, with median values in the low micromolar to submicromolar range.



**Figure 3.** Potency values. For multifamily ligands, the distributions of different logarithmic potency values ( $pIC_{50}$ ,  $pK_i$ , and  $pK_d$ ) are reported in box plots. The yellow horizontal line indicates the median value of each distribution (reported next to the line).

In stage two of our analysis, targets of all multifamily ligands were compared individually and the ligands were assigned to 4 different subsets:

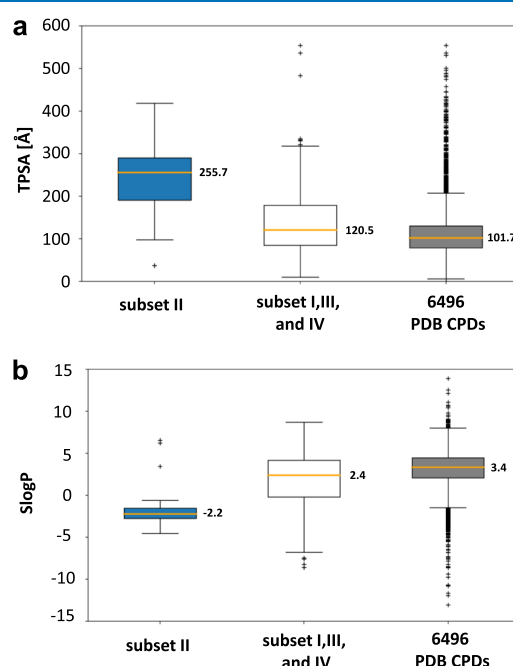
Subset I: ligands whose multifamily assignment depended on complexes with metabolizing enzymes or serum proteins (10 ligands); II: endogenous ligands (40); III: ligands binding to similar proteins from different families or to similar binding domains (51); and IV: multifamily ligands interacting with distinct targets (91).

The 10 ligands from subset I were omitted from further consideration because binding to serum proteins or metabolizing enzymes such as cytochromes is not relevant for polypharmacology (for all remaining ligands, complexes with such proteins were not included in subsequent analysis steps). Endogenous ligands such as adenosine 5'-triphosphate (ATP) or nucleoside derivatives have evolved to interact with different proteins. As such, these naturally occurring ligands are set apart from synthetic compounds and should best be separately considered. Furthermore, proteins from different families distinguished by established classification schemes might partly be structurally related and have similar biological functions. Therefore, subset III captured multifamily ligands for which at least some of the participating proteins had similar enzymatic functions or similar binding domains. By contrast, subset IV contained ligands that interacted exclusively with unrelated or distantly related targets (both in terms of structure and function). Figure 1a shows representative examples of subset IV ligands such as QUE that interacts with numerous distinct targets. Figure 1b shows subset III ligands. For example, NGH

inhibits metalloproteases from 2 different families and BMF binds to bromodomains in proteins from 4 different families.

On the basis of our analysis, the 91 multifamily ligands belonged to subset IV having highest priority for promiscuity analysis, given that they interacted with unrelated targets. Therefore, specific examples discussed below were taken from subset IV.

For the initial set of 192 multifamily ligands and all other preselected PDB, different molecular properties were calculated and compared, revealing some interesting differences in the topological polar surface area (TPSA) and  $S \log P$  values. Multifamily ligands had overall large TPSA (with a median of 145.5 Å) and low  $S \log P$  values (median 1.7), indicating that multifamily ligands were generally polar. The apparent increase in hydrophilicity among multifamily ligands was further investigated by calculating TPSA (Figure 4a) and  $S \log P$

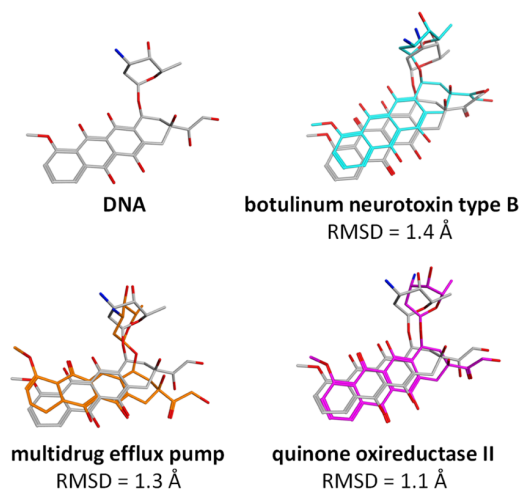


**Figure 4.** Molecular properties. The distributions of (a) TPSA and (b)  $S \log P$  values for endogenous multifamily ligands (subset II, blue), remaining multifamily ligands (white), and all preselected PDB ligands (PDB CPDs, gray) are reported in box plots. The yellow horizontal line indicates the median value of each distribution.

values (Figure 4b) for individual ligand subsets. With a median TPSA of 255.7 Å and  $S \log P$  value of  $-2.2$ , endogenous ligands were partly—but not exclusively—responsible for the relative increase in hydrophilicity because they included a variety of nucleosides with phosphate groups. However, even after removal of all subset II ligands, the remaining multifamily ligands had detectably higher hydrophilic character than other PDB compounds, with a median TPSA of 120.5 Å vs. 101.7 Å and median  $S \log P$  value of 2.4 versus 3.4, respectively. Thus, the multifamily activity of ligands was not attributable to hydrophobic “stickiness”. Rather, they were more hydrophilic in nature than many other PDB compounds, even when endogenous ligands were excluded. Furthermore, only 17 multifamily ligands (<10%) were found to contain substructures implicated in assay interference effects. We also searched for structural analogues and analogue series among

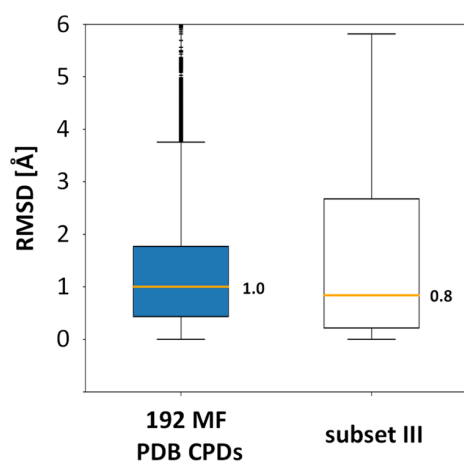
multifamily ligands. Only a single series containing 3 analogues was identified. Thus, multifamily ligands were not dominated by individual compound classes but were structurally diverse.

**2.2. Binding Conformations.** Next, bound conformations of each multifamily ligand were systematically superposed and compared. Figure 5 shows exemplary pairwise superpositions



**Figure 5.** Binding conformations of a multifamily ligand. Shown are exemplary pairwise superpositions of crystallographic conformations of doxorubicin. As a reference conformation, doxorubicin bound to DNA (Ligand Expo ID: DM2, PDB entry ID: 1DA9) is used (gray carbon atoms) onto which bound conformations of doxorubicin extracted from complex structures with diverse targets are superposed. For each superposition, the RMSD value is reported.

of target-dependent conformations. Figure 6 shows the distribution of Root-Mean-Square Deviation (RMSD) values resulting from exhaustive comparison of binding conformations of all 192 multifamily ligands. Pairwise rmsd values ranged from close to 0 to 6.0 Å, with a median RMSD of 1.0 Å. Thus, approximately half of the comparisons identified similar

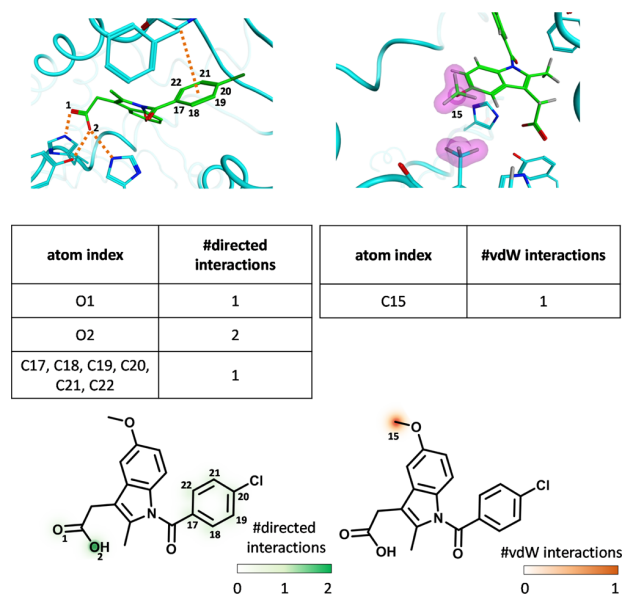


**Figure 6.** Comparison of binding conformations. For all 192 multifamily ligands (MF PDB CPDs, blue) and the subset of prioritized multifamily ligands (subset III, white), the distribution of RMSD values for all pairwise superpositions of bound conformations is reported in a box plot. The yellow horizontal line indicates the median value of the distributions.

binding conformations in different structural environments. The third quartile reached a value of 1.8 Å. At this level, conformations of typical ligands become dissimilar. Therefore, approximately a quarter of the comparisons indicated target-dependent conformational differences. However, overall most bound conformations of multifamily ligands were similar, regardless of the conformational space available to ligands and differences in the geometry and shape of binding sites. Figure 6 also reports the corresponding distribution of RMSD values for the 91 high-priority ligands from subset IV. In this case, the median RMSD value was only 0.8 Å, thus even lower, despite interactions with unrelated targets.

Hence, it remained to be determined how similar ligand conformations were accommodated in different structural environments.

**2.3. Target–Ligand Interactions.** Therefore, a systematic analysis of intermolecular interactions was carried out (details are provided in Materials and Methods). Directed polar interactions including hydrogen bonds, ligand–metal contacts, ionic, and  $\pi$ -interactions were accounted for and, in addition, van der Waals (vdW) contacts between ligand atoms and nonpolar amino acids. These vdW contacts were quantified as an indicator of hydrophobic interactions and shape complementarity. Figure 7 illustrates target–ligand interaction analysis using indomethacin as an exemplary ligand bound to human peroxisome proliferator-activated receptor  $\gamma$  (PPAR $\gamma$ ).<sup>32</sup> Atoms involved in directed and/or vdW interactions were uniquely indexed and individual atomic contacts were counted. Then, contacts were mapped onto ligand atoms and color-coded according to their frequency.



**Figure 7.** Identification of target–ligand interaction hotspots. For an exemplary bound ligand (green carbon atoms, Ligand Expo ID: IMN, PDB entry 3ADS), the atom-based number of directed interactions (hydrogen bonds, ligand–metal contacts, ionic, and  $\pi$ -interactions) and vdW interactions with hydrophobic protein residues are reported. Atoms involved in interactions are indexed. In the X-ray structure, directed interactions are shown as dotted lines and vdW contacts are presented as magenta spheres. In the corresponding 2D representations, ligand atoms are color-coded according to the number of their interactions (directed: light to dark green, vdW: light to dark orange).



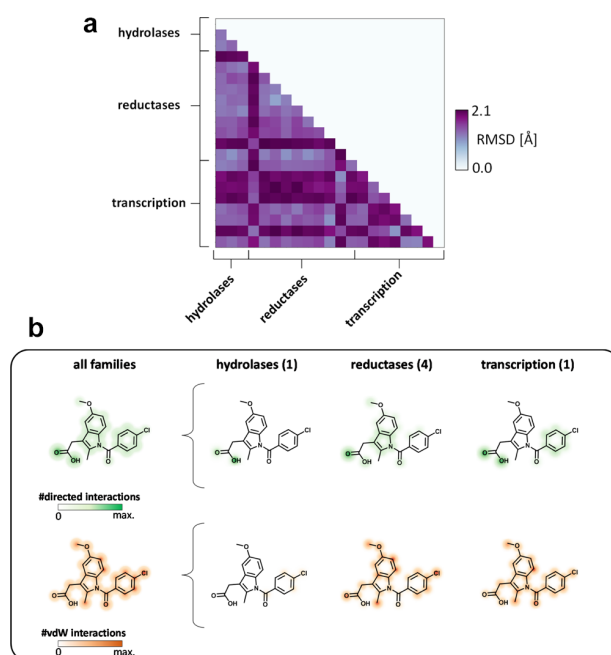
Accordingly, dark green and dark orange atoms, or groups of atoms, indicated centers of polar and vdW interactions, respectively, as illustrated in Figure 7. For each multifamily ligand, interaction patterns were then monitored separately for targets belonging to different families and compared. The analysis revealed that multifamily ligands mostly formed different “interaction hotspots” with targets belonging to different families, even if bound conformations were similar, as discussed in the following.

For the examples presented, binding site similarity between participating protein families was also calculated (see Materials and Methods) and binding sites reaching a threshold for detectable similarity were identified.

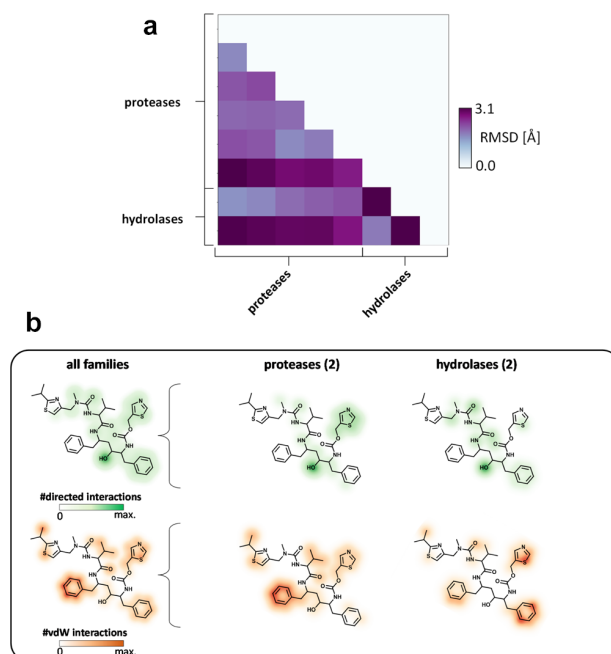
#### 2.4. Interaction Hotspots of Multifamily Ligands.

Interaction hotspots were defined as ligand atoms most frequently involved in specific ligand–target interactions. They were calculated by mapping detectable ligand–target interactions on participating ligand atoms and determining their frequency on a per-atom basis (see Materials and Methods). Thus, so-defined hotspots revealed centers of interactions in ligands and other regions that did not participate in such interactions. Combining the analysis of binding conformations and target–ligand interactions made it possible to rationalize different multitarget binding events. For example, indomethacin represents a well-characterized poly-pharmacological drug<sup>33</sup> that is known to interact with unrelated targets including cyclooxygenases,<sup>34</sup> phospholipase A2,<sup>35</sup> and PPAR $\gamma$ ,<sup>32</sup> and also serum albumin.<sup>36</sup> As revealed by its RMSD matrix in Figure 8a, indomethacin belongs to the subset of multifamily ligands that display target-dependent differences in binding conformations with largest RMSD values exceeding 2.0 Å. Largest conformational variations were observed for transcription factor binding compared to hydrolases, reductases, and secreted proteins. Hence, indomethacin conformationally adapted to different structural environments. Figure 8b compares the interactions between indomethacin and targets from different families. The aliphatic carboxylic acid group of indomethacin was a conserved hotspot for polar interactions across all 3 protein families. On the other hand, aromatic interactions of the central indole ring moiety were only observed in binding sites of reductases. However, vdW interactions involving this moiety were mostly found in reductases and the transcription factor. By contrast, in the active site of hydrolases, no interactions with the central part of indomethacin were detectable. Thus, binding of this drug across different target families involved both conserved and distinct interaction patterns, which was a recurrent theme among multifamily ligands.

The HIV protease inhibitor ritonavir<sup>37</sup> is an example of a multifamily ligand with different binding conformations, yielding largest RMSD values exceeding 3.0 Å (Figure 9a). Among other targets, ritonavir is known to bind to cytochrome P450 enzymes, which causes undesirable side effects and drug interactions.<sup>38</sup> The peptidomimetic nature of ritonavir with a large number of rotatable bonds supports flexibility of binding conformations. Accordingly, as shown in Figure 9b, this ligand displayed different polar interaction patterns when bound to proteases and hydrolases that were closely related and had significant binding site similarity. Polar interactions in proteases and hydrolases were centered on a hydroxyl group, while distinct polar hotspots were identified for the thiazole ring. Moreover, this ligand displayed overlapping yet distinct vdW interactions in different binding sites with three hotspots,



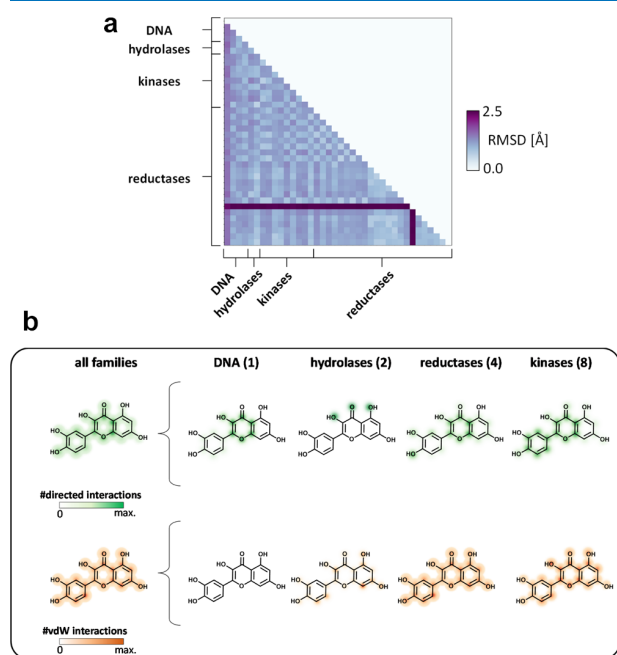
**Figure 8.** Multifamily binding of indomethacin. (a) Shows an RMSD value matrix for comparison of indomethacin conformations bound to targets from different families. Each cell represents an RMSD value for a pairwise superposition. Cells are color-coded according to RMSD values (from light blue (0.0 Å) to purple (maximum rmsd)). (b) Shows interaction hotspots of indomethacin for different protein families. The number of targets per family is given in parentheses. The representation of interaction hotspots is according to Figure 7. A low binding site similarity was detected for transcription factors, secreted proteins, and hydrolases.



**Figure 9.** Multifamily binding of ritonavir. (a) RMSD matrix. (b) Interaction hotspots for different protein families. The presentation is according to Figure 8. A high binding site similarity was detected for hydrolases and proteases.

only one of which (the central phenyl moiety) was shared by the two target families. Hence, ritonavir provided an intuitive example of a multifamily ligand where conformational adaptability was accompanied by the formation of different interaction hotspots.

Because ligand binding across different protein families was not only attributable to conformational variability and resulting differences in interaction hotspots, we reasoned that differences in interaction patterns should also be present for multifamily ligands that bound with similar conformations to different targets. For example, quercetin is a relatively small and rigid compound that belongs to the large subset of multifamily ligands with essentially conserved binding conformations across different target families (with only one exception), as illustrated by its RMSD matrix in Figure 10a.

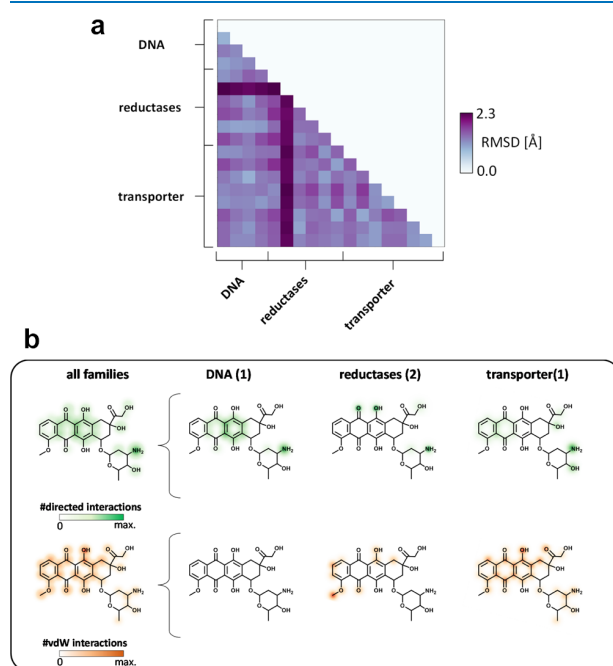


**Figure 10.** Multifamily binding of quercetin. (a) RMSD matrix. (b) Interaction hotspots for different protein families. No binding site similarity was detected.

Quercetin contains a polyphenolic flavonoid scaffold. Notably, flavonoids were considered privileged substructures in drug discovery<sup>39</sup> capable of forming interactions with kinases,<sup>40</sup> DNA,<sup>41</sup> or hydrolases.<sup>42</sup> In addition, there is crystallographic evidence for the oxidative cleavage of quercetin by quercetinase.<sup>43</sup> However, polyphenols such as quercetin were also implicated in reactivity under assay conditions and other potential liabilities such as membrane perturbation, adding them to the spectrum of interference compounds.<sup>12,13</sup> X-ray structures of quercetin in complex with DNA and targets from three protein families also revealed both conserved and family-dependent interaction hotspots, as shown in Figure 10b. The 3-hydroxy group of quercetin was consistently involved in target–ligand interactions, whereas the carbonyl oxygen formed a hydrolase-specific interaction hot spot. The C-ring of quercetin was involved in  $\pi$ -interactions in all complexes except when bound to hydrolases. Especially in kinases, all three rings were involved in aromatic interactions. In addition, extensive vdW contacts were formed when quercetin was

bound to reductases and kinases, which were largely absent in hydrolases (and DNA).

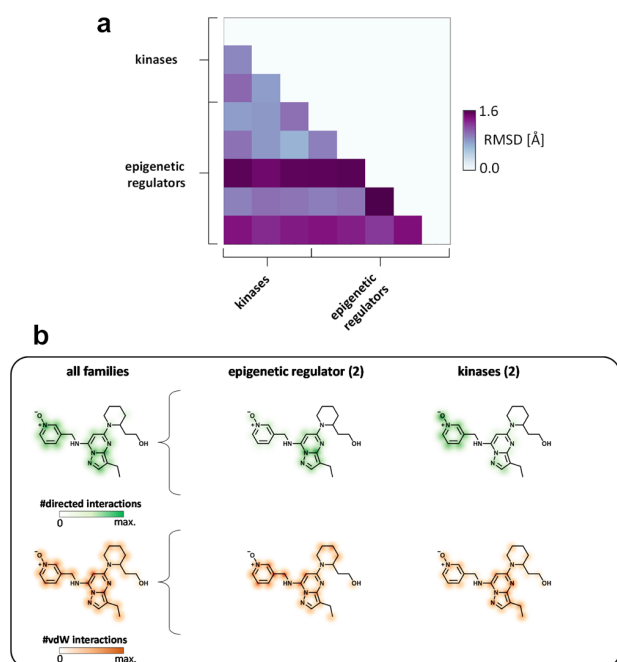
Comparable conformational invariance was also observed for the chemotherapeutic agent doxorubicin, given its rigid structure. The presumed mechanism of action of doxorubicin involves the intercalation of the planar anthracycline core with the DNA double helix.<sup>44</sup> Similar to quercetin, doxorubicin contains structural elements that contribute to ligand–target interactions but also cause assay liabilities and potentially adverse pharmacological effects.<sup>45–47</sup> In light of the complex pharmacokinetics of anthracyclines, interactions of doxorubicin with different target proteins were analyzed in a number of crystallographic investigations including complexes with efflux pumps<sup>48</sup> and cytosolic reductases.<sup>49</sup> Because of the rigidity of the anthracycline core, limited conformational flexibility was due to bond rotation in the terminal carboxylic acid and aminoglycoside moiety, respectively (Figure 11a). Rather



**Figure 11.** Multifamily binding of doxorubicin. (a) RMSD matrix. (b) Interaction hotspots for different protein families. No binding site similarity was detected.

unexpectedly, interaction analysis of doxorubicin revealed that  $\pi$ -interactions of the aromatic core were only dominant when binding to DNA but that vdW contacts involving this moiety were preferentially observed in complexes with reductases and a transporter (Figure 11b). By contrast, the primary amine of the aminoglycoside was found to be a conserved interaction hotspot across 3 protein families. On the other hand, carbonyl and hydroxyl oxygens of the central anthracycline core only formed polar contacts when binding to reductases. Thus, polar and vdW interactions distinguished binding of doxorubicin in different structural environments.

The ATP-competitive kinase inhibitor dinaciclib has much more conformational freedom than quercetin and doxorubicin. However, it also bound with similar conformations to cyclin-dependent kinases<sup>50</sup> and epigenetic regulators,<sup>51</sup> as shown in Figure 12a, with a maximum RMSD value of 1.6 Å. Cyclin-dependent kinases and epigenetic regulators had detectable



**Figure 12.** Multifamily binding of dinaciclib. (a) RMSD matrix. (b) Interaction hotspots for different protein families. Binding site similarity between epigenetic regulators and kinases was detected.

binding site similarity. Figure 12b compares interaction hotspots of dinaciclib with these 2 protein families. In both cases, extensive vdW interactions with essentially all parts of the inhibitor were observed, reflecting a high degree of shape complementarity in these binding sites. By contrast, distinct hotspots for polar interactions emerged. For epigenetic regulators, directed interactions with the central pyrazol[1,5-a]pyrimidine scaffold were detected. On the other hand, charge-assisted interactions and  $\pi$ -interactions involving the pyridine-*N*-oxide moiety were prevalent in complexes with kinases. Accordingly, dinaciclib was also representative of many multifamily ligands that had largely conserved binding conformations across different targets but formed different interaction hotspots in changing protein environments.

### 3. CONCLUSIONS

In this work, we have systematically identified ligands with available experimental structures of complexes with targets from different families. These structures of multifamily ligands provided firm evidence for the presence of true binding events. Properties and binding characteristics of multifamily ligands were analyzed in detail to better understand the molecular basis of their promiscuous binding behavior. Multifamily ligands also included drugs with known polypharmacology. Surprisingly, multifamily ligands were overall slightly more hydrophilic than other PDB compounds. Moreover, many—but not all—multifamily ligands had similar binding conformations when interacting with targets from different families. In some instances, conformational variability in different binding sites was expectedly accompanied by the formation of different interaction hotspots. In other cases, conserved binding conformations of rigid or flexible ligands revealed overlapping yet distinct interaction hotspots across different target families. The formation of target family-dependent interaction hotspots in the presence of variable or

conserved binding conformations emerged as a recurrent theme across multifamily ligands. The ligands interacted similarly with targets from the same family, leading to family-dependent hotspots, but interaction hotspots clearly differed between families. These observations provided a rationale for the promiscuous binding capacity of ligands at the molecular level of detail.

### 4. MATERIALS AND METHODS

All calculations were carried out using in-house Python scripts with the aid of RDKit<sup>52</sup> and the OpenEye's chemistry toolkit,<sup>53</sup> KNIME protocols,<sup>54</sup> and the molecular operating environment (MOE).<sup>55</sup>

**4.1. Ligands from X-ray Structures.** X-ray structures and associated compound data were extracted from the Ligand Expo section<sup>56</sup> of the PDB and complemented with experimental binding affinity data from the PDBbind database.<sup>57</sup> Ligands were considered for further analysis if they had a minimum molecular weight of 300 Da and if at least one activity value of 10  $\mu$ M ( $pIC_{50}$ ,  $pK_i$ , or  $pK_d$ ) or better was available. Application of the molecular weight and activity cutoff ensured that salts, small organic components, and molecular fragments were excluded. Molecular descriptors of PDB ligands were calculated using RDKit. PDB ligands were screened in silico for structures containing Pan Assay Interference Compounds (PAINS)<sup>11</sup> utilizing SMARTS<sup>58</sup> strings obtained from three publicly available filters (ZINC,<sup>59</sup> RDKit, and ChEMBL<sup>60</sup>).

**4.2. Target Family Distribution.** For crystallographic targets, family assignments were obtained by matching UniProt<sup>61</sup> target identifiers to ChEMBL identifiers and applying the ChEMBL protein family classification scheme. In addition, the number of targets per compound was determined on the basis of unique UniProt identifiers.

**4.3. Searching for Structural Analogues.** A systematic search for analogues among multifamily ligands was carried out using a matched molecular pair-based computational method.<sup>62</sup>

**4.4. Analysis of Binding Conformations.** For each multifamily ligand, bound conformations were extracted from the corresponding X-ray complexes and superposed. From these superpositions, pairwise RMSD values of ligand conformations were calculated using MOE.

**4.5. Analysis of Target–Ligand Interactions.** For multifamily ligands, crystallographic target–ligand interactions were systematically analyzed using a KNIME implementation of MOE if two or more complexes representing a protein family were available. Crystallographic water molecules were removed from X-ray structures to avoid overestimation of water contacts in complexes.<sup>63</sup> Nonbonded interactions involving ligand atoms were determined within a radius of 4.5 Å. Hydrogen bonds, ligand–metal contacts, ionic, and  $\pi$ -interactions were identified with the aid of an empirical geometry-based scoring function.<sup>64</sup> In addition, vdW contacts between ligand atoms and hydrophobic protein residues were determined by applying a maximal interaction energy of  $-0.5$  kcal/mol. The sum of all polar and vdW interactions was calculated for each multifamily ligand atom over all binding sites in X-ray structures of a given protein family. For each family, the atom-based number of interactions was mapped onto a 2D representation of the ligand<sup>64</sup> using the chemistry toolkit of RDKit. Atom positions were color-coded according to the number of mapped interactions.



**4.6. Binding Site Similarity.** Similarity of binding sites from different protein families was analyzed using ProBiS.<sup>65</sup> For pairwise comparison of nonredundant targets from different families, the lowest recommended similarity z-score of 1.0 was applied as a threshold for detectable binding site similarity.<sup>65</sup> If a pairwise comparison yielded a score of 1.0 or greater, the binding sites were classified as similar.

**4.7. Data Availability.** All multifamily ligands, family assignments, available affinity data, and the ligand subset classification are made available as Table S1 of the [Supporting Information](#).

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: [10.1021/acsomega.8b03481](https://doi.org/10.1021/acsomega.8b03481).

Multifamily ligands, family assignments, available affinity data, and the ligand subset classification ([PDF](#))

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de). Phone: 49-228-2699-306 (J.B.).

### ORCID

Michael Gütschow: [0000-0002-9376-7897](https://orcid.org/0000-0002-9376-7897)

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

### Author Contributions

The study was carried out and the manuscript written with contributions of all authors. All authors have approved the final version of the manuscript.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The use of OpenEye's toolkits was made possible by their free academic licensing program.

## ■ REFERENCES

- Zimmermann, G. R.; Lehar, J.; Keith, C. T. Multi-Target Therapeutics: When the Whole is greater than the Sum of the Parts. *Drug Discovery Today* **2007**, *12*, 34–42.
- Hopkins, A. L. Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.
- Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874–7887.
- Bolognesi, M. L. Polypharmacology in a Single Drug: Multitarget Drugs. *Curr Med Chem* **2013**, *20*, 1639–1645.
- Bolognesi, M. L.; Cavalli, A. Multitarget Drug Discovery and Polypharmacology. *ChemMedChem* **2016**, *11*, 1190–1192.
- Rosini, M. The Rise of Multitarget Drugs over Combination Therapies. *Future Med. Chem.* **2014**, *6*, 485–487.
- Hu, Y.; Bajorath, J. Compound Promiscuity - What Can We Learn From Current Data. *Drug Discovery Today* **2013**, *18*, 644–650.
- Kuhn, M.; Banchaabouchi, M. A.; Campillos, M.; Jensen, L. J.; Gross, C.; Gavin, A.-C.; Bork, P. Systematic Identification of Proteins that Elicit Drug Side Effects. *Mol Syst Biol* **2013**, *9*, 663.
- Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
- 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.
- 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.
- Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.
- Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 – Utility and Limitations. *ACS Chem. Biol.* **2017**, *13*, 36–44.
- 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.
- 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.
- 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.
- 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.
- Nath, A.; Atkins, W. M. A Quantitative Index of Substrate Promiscuity. *Biochemistry* **2008**, *47*, 157–166.
- Srinivasan, B.; Marks, H.; Mitra, S.; Smalley, D. M.; Skolnick, J. Catalytic and Substrate Promiscuity: Distinct Multiple Chemistries Catalysed by the Phosphatase Domain of Receptor Protein Tyrosine Phosphatase. *Biochem. J.* **2016**, *473*, 2165–2177.
- Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijter, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L. H.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181.
- Cronstein, B. N. Low-Dose Methotrexate: A Mainstay in the Treatment of Rheumatoid Arthritis. *Pharmacol. Rev.* **2005**, *57*, 163–172.
- Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130–137.
- Hopkins, A.; Mason, J.; Overington, J. Can We Rationally Design Promiscuous Drugs? *Curr. Opin. Struct. Biol.* **2006**, *16*, 127–136.
- Morphy, R.; Rankovic, Z. Designed Multiple Ligands. An Emerging Drug Discovery Paradigm. *J. Med. Chem.* **2005**, *48*, 6523–6543.
- Gupta-Ostermann, D.; Hu, Y.; Bajorath, J. Systematic Mining of Analog Series with Related Core Structures in Multi-target Activity Space. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 665–674.
- Hu, Y.; Bajorath, J. SAR Matrix Method for Large-scale Analysis of Compound Structure-Activity Relationships and Exploration of Multi-Target Activity Spaces. *Methods Mol. Biol.* **2018**, *1825*, 339–352.
- de la Vega de León, A.; Bajorath, J. Design of a Three-Dimensional Multi-Target Activity Landscape. *J. Chem. Inf. Model.* **2012**, *52*, 2876–2883.
- Moya-García, A.; Adeyelu, T.; Kruger, F. A.; Dawson, N. L.; Lees, J. G.; Overington, J. P.; Orengo, C.; Ranea, J. A. G. Structural and Functional View of Polypharmacology. *Sci. Rep.* **2017**, *7*, 10102.
- Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS One* **2013**, *8*, No. e65894.
- Gilberg, E.; Stumpfe, D.; Bajorath, J. X-Ray Structure Based Identification of Compounds with Activity against Targets from

Different Families and Generation of Templates for Multitarget Ligand Design. *ACS Omega* **2018**, *3*, 106–111.

(31) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(32) Waku, T.; Shiraki, T.; Oyama, T.; Maebara, K.; Nakamori, R.; Morikawa, K. The Nuclear Receptor PPAR $\gamma$  Individually Responds to Serotonin- and Fatty Acid-Metabolites. *EMBO J.* **2010**, *29*, 3395–3407.

(33) Song, J.; Liu, X.; Rao, T. S.; Chang, L.; Meehan, M. J.; Blevitt, J. M.; Wu, J.; Dorrestein, P. C.; Milla, M. E. Phenotyping Drug Polypharmacology via Eicosanoid Profiling of Blood. *J. Lipid Res.* **2015**, *56*, 1492–1500.

(34) Blanco, F. J.; Guitian, R.; Moreno, J.; de Toro, F. J.; Galdo, F. Effect of Antiinflammatory Drugs on COX-1 and COX-2 Activity in Human Articular Chondrocytes. *J. Rheumatol.* **1999**, *26*, 1366–1373.

(35) Singh, N.; Kumar, R. P.; Kumar, S.; Sharma, S.; Mir, R.; Kaur, P.; Srinivasan, A.; Singh, T. P. Simultaneous Inhibition of Anti-Coagulation and Inflammation: Crystal Structure of Phospholipase A2 Complexed with Indomethacin at 1.4 Å Resolution Reveals the Presence of the New Common Ligand-Binding Site. *J. Mol. Recognit.* **2009**, *22*, 437–445.

(36) Bogdan, M.; Pirnau, A.; Floare, C.; Bugeac, C. Binding Interaction of Indomethacin with Human Serum Albumin. *J. Pharm. Biomed. Anal.* **2008**, *47*, 981–984.

(37) Rock, B. M.; Hengel, S. M.; Rock, D. A.; Wienkers, L. C.; Kunze, K. L. Characterization of Ritonavir-Mediated Inactivation of Cytochrome P450 3A4. *Mol. Pharmacol.* **2014**, *86*, 665–674.

(38) Foy, M.; Sperati, C. J.; Lucas, G. M.; Estrella, M. M. Drug Interactions and Antiretroviral Drug Monitoring. *Curr. HIV/AIDS Rep.* **2014**, *11*, 212–222.

(39) Reis, J.; Gaspar, A.; Milhazes, N.; Borges, F. Chromone as a Privileged Scaffold in Drug Discovery: Recent Advances. *J. Med. Chem.* **2017**, *60*, 7941–7957.

(40) Yokoyama, T.; Kosaka, Y.; Mizuguchi, M. Structural Insight into the Interactions between Death-Associated Protein Kinase 1 and Natural Flavonoids. *J. Med. Chem.* **2015**, *58*, 7400–7408.

(41) Srivastava, S.; Somasagara, R. R.; Hegde, M.; Nishana, M.; Tadi, S. K.; Srivastava, M.; Choudhary, B.; Raghavan, S. C. Quercetin, a Natural Flavonoid Interacts with DNA, Arrests Cell Cycle and Causes Tumor Regression by Activating Mitochondrial Pathway of Apoptosis. *Sci. Rep.* **2016**, *6*, 24049.

(42) Xue, G.; Gong, L.; Yuan, C.; Xu, M.; Wang, X.; Jiang, L.; Huang, M. A Structural Mechanism of Flavonoids in Inhibiting Serine Proteases. *Food Funct.* **2017**, *8*, 2437–2443.

(43) Jeoung, J.-H.; Nianios, D.; Fetzner, S.; Dobbek, H. Quercetin 2,4-Dioxygenase Activates Dioxygen in a Side-on O<sub>2</sub>-Ni Complex. *Angew. Chem., Int. Ed. Engl.* **2016**, *55*, 3281–3284.

(44) Howerton, S. B.; Nagpal, A.; Dean Williams, L. Surprising Roles of Electrostatic Interactions in DNA-Ligand Complexes. *Biopolymers* **2003**, *69*, 87–99.

(45) 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.

(46) Mordente, A.; Meucci, E.; Silvestrini, A.; Martorana, G.; Giardina, B. New Developments in Anthracycline-Induced Cardiotoxicity. *Curr. Med. Chem.* **2009**, *16*, 1656–1672.

(47) Motlagh, N. S. H.; Parvin, P.; Ghasemi, F.; Atyabi, F. Fluorescence Properties of Several Chemotherapy Drugs: Doxorubicin, Paclitaxel and Bleomycin. *Biomed. Opt. Express* **2016**, *7*, 2400–2406.

(48) Eicher, T.; Cha, H.-j.; Seeger, M. A.; Brandstatter, L.; El-Delik, J.; Bohnert, J. A.; Kern, W. V.; Verrey, F.; Grutter, M. G.; Diederichs, K.; Pos, K. M. Transport of Drugs by the Multidrug Transporter AcrB Involves an Access and a Deep Binding Pocket that are Separated by a Switch-Loop. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 5687–5692.

(49) Leung, K. K. K.; Shilton, B. H. Binding of DNA-Intercalating Agents to Oxidized and Reduced Quinone Reductase 2. *Biochemistry* **2015**, *54*, 7438–7448.

(50) Chen, P.; Lee, N. V.; Hu, W.; Xu, M.; Ferre, R. A.; Lam, H.; Bergqvist, S.; Solowiej, J.; Diehl, W.; He, Y.-A.; Yu, X.; Nagata, A.; VanArsdale, T.; Murray, B. W. Spectrum and Degree of CDK Drug Interactions Predicts Clinical Performance. *Mol. Cancer Ther.* **2016**, *15*, 2273–2281.

(51) Ember, S. W. J.; Zhu, J.-Y.; Olesen, S. H.; Martin, M. P.; Becker, A.; Berndt, N.; Georg, G. I.; Schönbrunn, E. Acetyl-lysine Binding Site of Bromodomain-Containing Protein 4 (BRD4) Interacts with Diverse Kinase Inhibitors. *ACS Chem. Biol.* **2014**, *9*, 1160–1171.

(52) RDKit. Cheminformatics and Machine Learning Software, 2013. <http://www.rdkit.org> (accessed Aug 01, 2018).

(53) OEChem. TK, version 2.0.0; OpenEye Scientific Software: Santa Fe, NM, 2015.

(54) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinel, 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., Burkhardt, H., Schmidt Thieme, L., Decker, R., Eds.; Springer: Berlin, 2008; pp 319–326.

(55) *Molecular Operating Environment (MOE), 2018.01*; Chemical Computing Group ULC: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2018.

(56) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153–2155. <http://ligand-expo.rcsb.org/> (accessed September 16, 2018).

(57) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(58) James, C. A.; Weininger, D.; Delany, J. *SMARTS Theory. Daylight Theory Manual; Daylight Chemical Information Systems; Laguna Niguel, CA, 2000.*

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

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

(61) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.

(62) 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. *J. Med. Chem.* **2016**, *59*, 7667–7676.

(63) Lu, Y.; Wang, R.; Yang, C.-Y.; Wang, S. Analysis of Ligand-Bound Water Molecules in High-Resolution Crystal Structures of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 668–675.

(64) Clark, A. M.; Labute, P. 2D Depiction of Protein-Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 1933–1944.

(65) Konc, J.; Česnik, T.; Konc, J. T.; Penca, M.; Janežič, D. ProBiS-Database: Precalculated Binding Site Similarities and Local Pairwise Alignments of PDB Structures. *J. Chem. Inf. Model.* **2012**, *52*, 604–612.



## Conclusions

This study completes the structure-based evaluation of compound promiscuity provided in *chapters 3 and 7*. Molecular properties and binding modes of multifamily ligands were analyzed in detail to explore the molecular basis of their promiscuous binding behavior.

Surprisingly, multifamily ligands polar, which was especially the case for a subset of endogenous ligands. Moreover, bound conformations of multifamily ligands were overall similar, regardless of the conformational space available to ligands and differences between binding sites in distantly related or unrelated targets. In some cases, when conformational variability in different binding site occurred, it was also accompanied by the formation of different interaction hotspots. Across multifamily ligands, the formation of such family dependent interaction hotspots emerged as a recurrent theme. These observations rationalized the capacity of promiscuous ligands to bind multiple targets from different families at the molecular level of detail.





# Conclusion

Biological screening data carry the inherent risk of being compromised by artificial readouts resulting from numerous assay interference mechanisms. Therefore, it is a major concern in biological screening, medicinal chemistry, and chemoinformatics to provide confirmatory evidence for the biological activity of hit compounds. This evidence is needed to promote candidates for hit-to-lead optimization or to identify truly promiscuous compounds that can be used in polypharmacology efforts. In this thesis, chemoinformatic methods are utilized to elucidate the PAINS actions and provide a reference frame for judging assay interference in medicinal chemistry. Additionally, on the basis of structurally confirmed binding events, promiscuity is studied at the molecular level of detail.

The first representative study (*chapter 2*) provided initial insight into the uncertainties associated with biological screening data. Of 466 extensively tested promiscuous screening compounds, more than half were identified as PAINS or aggregators. In addition, visual inspection of the remaining molecules revealed that only a subset of 30 compounds were not associated with evident chemical liabilities. This indicated that substructure filters did not comprehensively cover interference mechanisms and had intrinsic limitations. For example, the study demonstrated that specific PAINS SMARTS patterns did not take tautomerism into account. It was shown that a confined number of truly promiscuous compounds that qualify for polypharmacology existed. However, the identification of such compounds required careful analysis of biological screening data beyond the sole use of structural filters. In *chapter 3*, X-ray structures of target-ligand complexes containing PAINS were identified and explored. An unexpectedly high number of 1107 unique ligands that contained a PAINS motif was present in 2874 X-ray complexes. By considering structurally confirmed binding events, these findings added a new dimension to the challenging assessment of assay interference activi-

ties. For example, for notorious PAINS such as catechols and aminoacridines, the presence of specific ligand-target interactions and reactivity under assay conditions were not mutually exclusive. Notably, complexes were identified in which distinct structural modifications prevented interference reactions to take place. This was considered a particularly interesting finding because the structural environment in which a PAINS substructure was presented determined its activity. Expanding the hypothesis of structural context dependence, ASs of extensively tested screening compounds containing PAINS were identified and analyzed in *chapter 4*. In this study, the systematic assessment of assay hit rates revealed varying activity profiles among 177 individual PAINS classes present in the ASs. Thus, distinct PAINS motifs, such as catechols or quinones, were predominantly found in series with high rates, while others often displayed low activity or inactivity, including aniline-derivatives. Additionally, the ASs enabled a SAR analysis of interference activities. For example, structural modifications of rhodanines and Mannich bases were identified that had a significant influence on hit rates of these PAINS. These studies provided a novel reference frame for the knowledge-based assessment of assay interference. Subsequently, machine-learning models were introduced in *chapter 5* to enable a systematic classification between highly promiscuous and consistently inactive PAINS, thus bypassing challenges of case-by-case analysis by medicinal chemists. Classification models were successfully applied to balanced sets of PAINS represented by structural fingerprints. ECFP4 features that favored successful predictions of SVM models were identified and mapped onto correctly classified PAINS. This allowed addressing the 'black-box' character of machine learning models and rationalize the classification process, thereby revealing novel insights into the structural context dependency of PAINS. For example, electron withdrawing substituents in the vicinity of electrophilic Michael acceptors favored the correct classification by SVM models. In *chapter 6* the SAR analysis of interference compounds was complemented by the provision of MMSs containing compounds with high assay promiscuity. These series were not submitted to filtering steps, thereby circumventing previously discussed shortcomings of substructure alerts. Hence, they can be utilized for an unrestricted systematic assessment of assay interference taking structural context information into account.

In *chapter 7*, structure-based confirmatory evidence for true binding events

was revisited and expanded. In this study, 702 crystallographic multifamily ligands were identified that bound to multiple distantly related or unrelated crystallographic targets. Of these, 355 multifamily ligands were also present in the medicinal chemistry literature and were generally promiscuous. The availability of these ligands in complex with multiple targets lend credence to their promiscuous nature. Therefore, 133 ASB scaffolds were isolated from series containing multifamily ligands and their analogs from medicinal chemistry, hence providing reference templates for multitarget and polypharmacological ligand design. In *chapter 8*, properties and binding characteristics of multifamily ligands were analyzed in detail to understand the molecular basis of their promiscuous binding behavior. Multifamily ligands included drugs with known pharmacology such as indomethacin and were overall more hydrophilic than other PDB compounds. Many, but not all, multifamily ligands adopted similar binding conformations when interacting with distinct targets. Notably, after systematic assessment of target-ligand interactions, family dependent interaction hotspots were identified. Typically, ligands interacted similarly with targets from the same family, but formed different interaction hotspots in binding sites of targets from another family.

In conclusion, this thesis introduced and explored the structural context dependence of interference mechanisms, thereby providing a reference frame for the confirmation of compound integrity in medicinal chemistry. It has also distinguished promiscuity from artificial multitarget activity and utilized three-dimensional target structures to provide a rationale for promiscuous binding at the molecular level of detail.



# Additional Publications

## Reviews and Perspectives

J. Schmitz, E. Gilberg, R. Löser, J. Bajorath, U. Bartz, M. Gütschow. Cathepsin B: Active Site Mapping with Peptidic Substrates and Inhibitors. *Bioorganic & Medicinal Chemistry* **2018**, *10*, 2745–2761.

E. Gilberg, J. Bajorath. Recent Progress in Structure-Based Evaluation of Compound Promiscuity. *ACS Omega* **2019**, *4*, 2758–2765.

## Original Research Publications

N. Furtmann, E. Gilberg, M. Spütz, M. Gütschow. Oxidation of Disulfides to Taurine and Sulfanilic Acid Derivatives *Synthesis* **2015**, *47*, 2609–2616.

A. M. Beckmann, E. Gilberg, S. Gattner, T. L. Huang, J. J. Vanden Eynde, J. Bajorath, M. Stirnberg, M. Gütschow. Evaluation of Bisbenzamidines as Inhibitors for Matriptase-2. *Bioorganic & Medicinal Chemistry Letters* **2016**, *26*, 3741–3745.

D. Dimova, E. Gilberg, J. Bajorath. Identification and Analysis of Promiscuity Cliffs Formed by Bioactive Compounds and Experimental Implications. *RSC Advances* **2017**, *7*, 58–66.

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

D. Häußler, A. C. Schulz-Fincke, A. M. Beckmann, A. Kelis, E. Gilberg, M. Mangold, J. Bajorath, M. Stirnberg, T. Steinmetzer, M. Gütschow. A Fluorescent-Labeled Phosphono Bisbenzguanidine As an Activity-Based Probe for Matriptase. *Chemistry* **2017**, *23*, 5205–5209.

S. Kayastha, D. Horvarth, E. Gilberg, M. Gütschow, J. Bajorath, A. Varnek. Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps. *Journal of Chemical Information and Modeling* **2017**, *57*, 1218–1232.

D. Stumpfe, E. Gilberg, J. Bajorath. Series of Screening Compounds with High Hit Rates for the Exploration of Multi-Target Activities and Assay Interference. *Future Science OA* **2018**, FSO279.

A. C. Schulz-Fincke, A. S. Tikhomirov, A. Braune, T. Girbl, E. Gilberg, J. Bajorath, M. Blaut, S. Nourshargh, M. Gütschow. Design of an Activity-Based Probe for Human Neutrophil Elastase: Implementation of the Lossen Rearrangement to Induce Förster Resonance Energy Transfers. *Biochemistry* **2018**, *57*, 742–752.

# Bibliography

- [1] J. Drews. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960–1964.
- [2] A. W. Jones. Early Drug Discovery and the Rise of Pharmaceutical Chemistry. *Drug Testing and Analysis* **2011**, *3*, 337–344.
- [3] D. A. Dias, S. Urban, U. Roessner. A Historical Overview of Natural Products in Drug Discovery. *Metabolites* **2012**, *2*, 303–336.
- [4] E. A. Ratti, D. G. Trist. The Continuing Evolution of the Drug Discovery Process in the Pharmaceutical Industry. *Farmaco* **2001**, *56*, 13–19.
- [5] J. Hughes, S. Rees, S. Kalindjian, K. Philpott. Principles of Early Drug Discovery. *British Journal of Pharmacology* **2011**, *162*, 1239–1249.
- [6] R. C. Mohs, N. H. Greig. Drug Discovery and Development: Role of Basic Biological Research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **2017**, *3*, 651–657.
- [7] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, J. P. Overington. A Comprehensive Map of Molecular Drug Targets. *Nature Reviews Drug Discovery* **2017**, *16*, 19–34.
- [8] C. Smith. Drug Target Validation: Hitting the Target. *Nature* **2003**, *422*, 341–347.
- [9] M. Schenone, V. Dančák, B. K. Wagner, P. A. Clemons. Target Identification and Mechanism of Action in Chemical Biology and Drug Discovery. *Nature Chemical Biology* **2013**, *9*, 232–240.
- [10] J. E. Neggers, B. Kwanten, T. Dierckx, H. Noguchi, A. Voet, L. Bral, K. Minner, B. Massant, N. Kint, M. Delforge, T. Vercruysse, E. Baloglu, W. Senapedis, M. Jacquemyn, D. Daelemans. Target Identification of Small Molecules Using Large-Scale CRISPR-Cas Mutagenesis Scanning of Essential Genes. *Nature Communications* **2018**, *9*, 502.

- [11] J. Knowles, G. Gromo. A Guide to Drug Discovery: Target Selection in Drug Discovery. *Nature Reviews Drug Discovery* **2003**, *2*, 63–69.
- [12] M. M. Frigault, J. C. Barrett. Is Target Validation All We Need? *Current Opinion in Pharmacology* **2014**, *17*, 81–86.
- [13] W. P. Walters, M. Namchuk. Designing Screens: How to Make Your Hits a Hit. *Nature Reviews Drug Discovery* **2003**, *2*, 259–266.
- [14] A. Smith. Screening for Drug Discovery: The Leading Question. *Nature* **2002**, *418*, 453–459.
- [15] N. Favalli, G. Bassi, J. Scheuermann, D. Neri. DNA-Encoded Chemical Libraries – Achievements and Remaining Challenges. *FEBS Letters* **2018**, *592*, 2168–2180.
- [16] J. A. Frearson, I. T. Collie. HTS and Hit Finding in Academia – from Chemical Genomics to Drug Discovery. *Drug Discovery Today* **2009**, *14*, 1150–1158.
- [17] G. M. Keserú, G. M. Makara. Hit Discovery and Hit-to-Lead Approaches. *Drug Discovery Today* **2006**, *11*, 741–748.
- [18] J. L. Dahlin, M. A. Walters. The Essential Roles of Chemistry in High-Throughput Screening Triage. *Future Medicinal Chemistry* **2014**, *6*, 1265–1290.
- [19] B. K. Shoichet. Screening in a Spirit Haunted World. *Drug Discovery Today* **2006**, *11*, 607–615.
- [20] M. S. Lipsky, L. K. Sharp. From Idea to Market: The Drug Approval Process. *The Journal of the American Board of Family Practice* **2001**, *14*, 362–367.
- [21] C. L. Meinert. Clinical Trials: Design, Conduct and Analysis. Oxford University Press, **2012**.
- [22] J. G. Lombardino, J. A. Lowe. The Role of the Medicinal Chemist in Drug Discovery - Then and Now. *Nature Reviews Drug Discovery* **2004**, *3*, 853–862.
- [23] A. Mullard. New Drugs Cost US\$2.6 Billion to Develop. *Nature Reviews Drug Discovery* **2014**, *13*, 877.
- [24] J. W. Scannell, J. Bosley. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PLOS ONE* **2016**, *11*, e0147215.
- [25] J. Baell, M. A. Walters. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature News* **2014**, *513*, 481–483.
- [26] S. S. Ou-Yang, J. Y. Lu, X. Q. Kong, Z. J. Liang, C. Luo, H. Jiang. Computational Drug Discovery. *Acta Pharmacologica Sinica* **2012**, *33*, 1131–1140.



- [27] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe. Computational Methods in Drug Discovery. *Pharmacological Reviews* **2014**, *66*, 334–395.
- [28] A. Hillisch, N. Heinrich, H. Wild. Computational Chemistry in the Pharmaceutical Industry: From Childhood to Adolescence. *ChemMedChem* **2015**, *10*, 1958–1962.
- [29] J. Bajorath. Rational Drug Discovery Revisited: Interfacing Experimental Programs with Bio- and Chemo-Informatics. *Drug Discovery Today* **2001**, *6*, 989–995.
- [30] T. Engel. Basic Overview of Chemoinformatics. *Journal of Chemical Information and Modeling* **2006**, *46*, 2267–2277.
- [31] S. Frye, M. Crosby, T. Edwards, R. Juliano. US Academic Drug Discovery. *Nature Reviews Drug Discovery* **2011**, *10*, 409–410.
- [32] C. J. Tralau-Stewart, C. A. Wyatt, D. E. Kleyn, A. Ayad. Drug Discovery: New Models for Industry–Academic Partnerships. *Drug Discovery Today* **2009**, *14*, 95–101.
- [33] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Research Database issue* **2019**, *47*, D1102–D1109.
- [34] P. D. Leeson, B. Springthorpe. The Influence of Drug-like Concepts on Decision-Making in Medicinal Chemistry. *Nature Reviews Drug Discovery* **2007**, *6*, 881–890.
- [35] C. A. Lipinski. Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technologies* **2004**, *1*, 337–341.
- [36] S. L. McGovern, E. Caselli, N. Grigorieff, B. K. Shoichet. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *Journal of Medicinal Chemistry* **2002**, *45*, 1712–1722.
- [37] J. J. Irwin, D. Duan, H. Torosyan, A. K. Doak, K. T. Ziebart, T. Sterling, G. Tumanian, B. K. Shoichet. An Aggregation Advisor for Ligand Discovery. *Journal of Medicinal Chemistry* **2015**, *58*, 7076–7087.
- [38] J. B. Baell, G. A. Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry* **2010**, *53*, 2719–2740.
- [39] R. M. Eglen, T. Reisine, P. Roby, N. Rouleau, C. Illy, R. Bossé, M. Bielefeld. The Use of AlphaScreen Technology in HTS: Current Status. *Current Chemical Genomics* **2008**, *1*, 2–10.

- [40] C. Aldrich, C. Bertozzi, G. I. Georg, L. Kiessling, C. Lindsley, D. Liotta, K. M. Merz, A. Schepartz, S. Wang. The Ecstasy and Agony of Assay Interference Compounds. *ACS Central Science* **2017**, *3*, 143–147.
- [41] J. L. Dahlin, J. W. M. Nissink, J. M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang, M. A. Walters. 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.
- [42] C. Maxim, T. D. Pasatoiu, V. C. Kravtsov, S. Shova, C. A. Muryn, R. E. P. Winpenny, F. Tuna, M. Andruh. Copper(II) and Zinc(II) Complexes with Schiff-Base Ligands Derived from Salicylaldehyde and 3-Methoxysalicylaldehyde: Synthesis, Crystal Structures, Magnetic and Luminescence Properties. *Inorganica Chimica Acta* **2008**, *361*, 3903–3911.
- [43] A. Golcu, M. Tumer, H. Demirelli, R. A. Wheatley. Cd(II) and Cu(II) Complexes of Polydentate Schiff Base Ligands: Synthesis, Characterization, Properties and Biological Activity. *Inorganica Chimica Acta* **2005**, *358*, 1785–1797.
- [44] H. E. Latuasan, W. Berends. On the Origin of the Toxicity of Toxoflavin. *Biochimica Et Biophysica Acta* **1961**, *52*, 502–508.
- [45] H. I. Ingólfsson, P. Thakur, K. F. Herold, E. A. Hobart, N. B. Ramsey, X. Periole, D. H. de Jong, M. Zwama, D. Yilmaz, K. Hall, T. Maretzky, H. C. Hemmings, C. Blobel, S. J. Marrink, A. Koçer, J. T. Sack, O. S. Andersen. Phytochemicals Perturb Membranes and Promiscuously Alter Protein Function. *ACS Chemical Biology* **2014**, *9*, 1788–1798.
- [46] C. Manivannan, K. M. Sundaram, R. Renganathan, M. Sundararaman. Investigations on Photoinduced Interaction of 9-Aminoacridine with Certain Catechols and Rutin. *Journal of Fluorescence* **2012**, *22*, 1113–1125.
- [47] J. B. Baell. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *Journal of Natural Products* **2016**, *79*, 616–628.
- [48] J. B. Baell, J. W. M. Nissink. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017-Utility and Limitations. *ACS Chemical Biology* **2018**, *13*, 36–44.
- [49] T. Tomasić, L. P. Masic. Rhodanine as a Privileged Scaffold in Drug Discovery. *Current Medicinal Chemistry* **2009**, *16*, 1596–1629.

- [50] K. M. Nelson, J. L. Dahlin, J. Bisson, J. Graham, G. F. Pauli, M. A. Walters. The Essential Medicinal Chemistry of Curcumin. *Journal of Medicinal Chemistry* **2017**, *60*, 1620–1637.
- [51] J. Reis, A. Gaspar, N. Milhazes, F. Borges. Chromone as a Privileged Scaffold in Drug Discovery: Recent Advances. *Journal of Medicinal Chemistry* **2017**, *60*, 7941–7957.
- [52] S. J. Capuzzi, E. N. Muratov, A. Tropsha. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS. *Journal of Chemical Information and Modeling* **2017**, *57*, 417–427.
- [53] P. W. Kenny. Comment on The Ecstasy and Agony of Assay Interference Compounds. *Journal of Chemical Information and Modeling* **2017**, *57*, 2640–2645.
- [54] S. Jasial, Y. Hu, J. Bajorath. 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.
- [55] T. Mendgen, C. Steuer, C. D. Klein. Privileged Scaffolds or Promiscuous Binders: A Comparative Study on Rhodanines and Related Heterocycles in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 743–753.
- [56] T. I. Oprea, C. G. Bologa, S. Boyer, R. F. Curpan, R. C. Glen, A. L. Hopkins, C. A. Lipinski, G. R. Marshall, Y. C. Martin, L. Ostopovici-Halip, G. Rishton, O. Ursu, R. J. Vaz, C. Waller, H. Waldmann, L. A. Sklar. A Crowdsourcing Evaluation of the NIH Chemical Probes. *Nature Chemical Biology* **2009**, *5*, 441–447.
- [57] M. S. Lajiness, G. M. Maggiora, V. Shanmugasundaram. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *Journal of Medicinal Chemistry* **2004**, *47*, 4891–4896.
- [58] P. S. Kutchukian, N. Y. Vasilyeva, J. Xu, M. K. Lindvall, M. P. Dillon, M. Glick, J. D. Coley, N. Brooijmans. Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery. *PLOS ONE* **2012**, *7*, e48476.
- [59] J. Drews. Case Histories, Magic Bullets and the State of Drug Discovery. *Nature Reviews Drug Discovery* **2006**, *5*, 635–640.
- [60] P. Nurse. Reductionism: The Ends of Understanding. *Nature* **1997**, *387*, 657.

- [61] G. R. Zimmermann, J. Lehár, C. T. Keith. Multi-Target Therapeutics: When the Whole Is Greater than the Sum of the Parts. *Drug Discovery Today* **2007**, *12*, 34–42.
- [62] D. H. Roukos. Networks Medicine: From Reductionism to Evidence of Complex Dynamic Biomolecular Interactions. *Pharmacogenomics* **2011**, *12*, 695–698.
- [63] A. L. Hopkins. Network Pharmacology: The next Paradigm in Drug Discovery. *Nature Chemical Biology* **2008**, *4*, 682–690.
- [64] A. Anighoro, J. Bajorath, G. Rastelli. Polypharmacology: Challenges and Opportunities in Drug Discovery. *Journal of Medicinal Chemistry* **2014**, *57*, 7874–7887.
- [65] M. L. Bolognesi. Polypharmacology in a Single Drug: Multitarget Drugs. *Current Medicinal Chemistry* **2013**, *20*, 1639–1645.
- [66] M. L. Bolognesi, A. Cavalli. Multitarget Drug Discovery and Polypharmacology. *ChemMedChem* **2016**, *11*, 1190–1192.
- [67] K. M. Giacomini, R. M. Krauss, D. M. Roden, M. Eichelbaum, M. R. Hayden, Y. Nakamura. When Good Drugs Go Bad. *Nature* **2007**, *446*, 975–977.
- [68] A. S. Reddy, S. Zhang. Polypharmacology: Drug Discovery for the Future. *Expert Review of Clinical Pharmacology* **2013**, *6*.
- [69] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Williams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, M. Pirmohamed. Drug Repurposing: Progress, Challenges and Recommendations. *Nature Reviews Drug Discovery* **2019**, *18*, 41–58.
- [70] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kujer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, B. L. Roth. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181.
- [71] W. A. Silverman. The Schizophrenic Career of a “Monster Drug”. *Pediatrics* **2002**, *110*, 404–406.
- [72] B. N. Cronstein. Low-Dose Methotrexate: A Mainstay in the Treatment of Rheumatoid Arthritis. *Pharmacological Reviews* **2005**, *57*, 163–172.
- [73] Z. A. Knight, H. Lin, K. M. Shokat. Targeting the Cancer Kinome through Polypharmacology. *Nature Reviews Cancer* **2010**, *10*, 130–137.
- [74] S. Gross, R. Rahal, N. Stransky, C. Lengauer, K. P. Hoeflich. Targeting Cancer with Kinase Inhibitors. *The Journal of Clinical Investigation* **2015**, *125*, 1780–1789.

- [75] M. J. Millan. On 'Polypharmacy' and Multi-Target Agents, Complementary Strategies for Improving the Treatment of Depression: A Comparative Appraisal. *International Journal of Neuropsychopharmacology* **2014**, *17*, 1009–1037.
- [76] Y. Hu, J. Bajorath. Compound Promiscuity: What Can We Learn from Current Data? *Drug Discovery Today* **2013**, *18*, 644–650.
- [77] J. Mestres, E. Gregori-Puigjané, S. Valverde, R. V. Solé. The Topology of Drug-Target Interaction Networks: Implicit Dependence on Drug Properties and Target Families. *Molecular bioSystems* **2009**, *5*, 1051–1057.
- [78] J. Mestres, E. Gregori-Puigjané, S. Valverde, R. V. Solé. Data Completeness—the Achilles Heel of Drug-Target Networks. *Nature Biotechnology* **2008**, *26*, 983–984.
- [79] D. Stumpfe, A. Tinivella, G. Rastelli, J. Bajorath. Promiscuity of Inhibitors of Human Protein Kinases at Varying Data Confidence Levels and Test Frequencies. *RSC Advances* **2017**, *7*, 41265–41271.
- [80] S. Jasial, J. Bajorath. Dark Chemical Matter in Public Screening Assays and Derivation of Target Hypotheses. *MedChemComm* **2017**, *8*, 2100–2104.
- [81] Y. Hu, J. Bajorath. Entering the 'Big Data' Era in Medicinal Chemistry: Molecular Promiscuity Analysis Revisited. *Future Science OA* **2017**, *3*, FSO179.
- [82] S. A. Canny, Y. Cruz, M. R. Southern, P. R. Griffin. PubChem Promiscuity: A Web Resource for Gathering Compound Promiscuity Data from PubChem. *Bioinformatics* **2012**, *28*, 140–141.
- [83] A. L. Hopkins, J. S. Mason, J. P. Overington. Can We Rationally Design Promiscuous Drugs? *Current Opinion in Structural Biology*. Protein-nucleic acid interactions/Folding and binding **2006**, *16*, 127–136.
- [84] A. de la Vega de León, J. Bajorath. Design of a Three-Dimensional Multitarget Activity Landscape. *Journal of Chemical Information and Modeling* **2012**, *52*, 2876–2883.
- [85] E. Proschak, H. Stark, D. Merk. Polypharmacology by Design: A Medicinal Chemist's Perspective on Multitargeting Compounds. *Journal of Medicinal Chemistry* **2019**, *62*, 420–444.
- [86] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet. Relating Protein Pharmacology by Ligand Chemistry. *Nature Biotechnology* **2007**, *25*, 197–206.

- [87] A. Moya-García, T. Adeyelu, F. A. Kruger, N. L. Dawson, J. G. Lees, J. P. Overington, C. Orengo, J. A. G. Ranea. Structural and Functional View of Polypharmacology. *Scientific Reports* **2017**, *7*, 10102.
- [88] V. J. Haupt, S. Daminelli, M. Schroeder. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLOS ONE* **2013**, *8*, e65894.
- [89] L. Peltason, J. Bajorath. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Medicinal Chemistry* **2009**, *1*, 451–466.
- [90] E. X. Esposito, A. J. Hopfinger, J. D. Madura. Methods for Applying the Quantitative Structure-Activity Relationship Paradigm. In: *Cheminformatics: Concepts, Methods, and Tools for Drug Discovery*. Ed. by J. Bajorath. Methods in Molecular Biology™. Totowa, NJ: Humana Press, **2004**, 131–213.
- [91] M. A. Johnson, G. M. Maggiora. Concepts and Applications of Molecular Similarity. New York: Wiley, **1990**.
- [92] G. M. Maggiora. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *Journal of Chemical Information and Modeling* **2006**, *46*, 1535–1535.
- [93] D. Stumpfe, J. Bajorath. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 2932–2942.
- [94] D. Dimova, E. Gilberg, J. Bajorath. Identification and Analysis of Promiscuity Cliffs Formed by Bioactive Compounds and Experimental Implications. *RSC Advances* **2016**, *7*, 58–66.
- [95] A. Bender, R. C. Glen. Molecular Similarity: A Key Technique in Molecular Informatics. *Organic & Biomolecular Chemistry* **2004**, *2*, 3204–3218.
- [96] G. Maggiora, M. Vogt, D. Stumpfe, J. Bajorath. Molecular Similarity in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2014**, *57*, 3186–3204.
- [97] Y. Takaoka, Y. Endo, S. Yamanobe, H. Kakinuma, T. Okubo, Y. Shimazaki, T. Ota, S. Sumiya, K. Yoshikawa. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1269–1275.
- [98] R. Todeschini, V. Consonni. Handbook of Molecular Descriptors. Wiley, **2008**.

- [99] D. Weininger. 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.
- [100] D. Weininger, J. Delany. SMARTS Theory. In: *Daylight Theory Manual*. Daylight Chemical Information Systems, Laguna Niguel, CA, **2000**.
- [101] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences* **1992**, *32*, 244–255.
- [102] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.
- [103] W. E. Brugger, A. J. Stuper, P. C. Jurs. Generation of Descriptors from Molecular Structures. *Journal of Chemical Information and Computer Sciences* **1976**, *16*, 105–110.
- [104] R. C Glen, V. S Rose. Computer Program Suite for the Calculation, Storage and Manipulation of Molecular Property and Activity Descriptors. *Journal of Molecular Graphics* **1987**, *5*, 79–86.
- [105] S. A. Wildman, G. M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873.
- [106] S. Prasanna, R. J. Doerksen. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. *Current Medicinal Chemistry* **2009**, *16*, 21–41.
- [107] S. Roehrig, A. Straub, J. Pohlmann, T. Lampe, J. Pernerstorfer, K.-H. Schlemmer, P. Reinemer, E. Perzborn. Discovery of the Novel Antithrombotic Agent 5-Chloro-N-((5S)-2-Oxo-3-[4-(3-Oxomorpholin-4-yl)phenyl]-1,3-oxazolidin-5-yl)methyl) Thiophene-2-Carboxamide (BAY 59-7939): An Oral, Direct Factor Xa Inhibitor. *Journal of Medicinal Chemistry* **2005**, *48*, 5900–5908.
- [108] K. Heikamp, J. Bajorath. Fingerprint Design and Engineering Strategies: Rationalizing and Improving Similarity Search Performance. *Future Medicinal Chemistry* **2012**, *4*, 1945–1959.
- [109] C. U. Symyx Software. *MACCS Structural Keys*. **2011**.
- [110] D. Rogers, M. Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

- [111] P. W. Kenny, J. Sadowski. Structure Modification in Chemical Databases. In: *Cheminformatics in Drug Discovery*. Ed. by T. I. Oprea. John Wiley & Sons, Ltd, **2005**, 271–285.
- [112] E. Griffen, A. G. Leach, G. R. Robb, D. J. Warner. Matched Molecular Pairs as a Medicinal Chemistry Tool. *Journal of Medicinal Chemistry* **2011**, *54*, 7739–7750.
- [113] A. M. Wassermann, P. Haebel, N. Weskamp, J. Bajorath. SAR Matrices: Automated Extraction of Information-Rich SAR Tables from Large Compound Data Sets. *Journal of Chemical Information and Modeling* **2012**, *52*, 1769–1776.
- [114] X. Hu, Y. Hu, M. Vogt, D. Stumpfe, J. Bajorath. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *Journal of Chemical Information and Modeling* **2012**, *52*, 1138–1145.
- [115] R. P. Sheridan. The Most Common Chemical Replacements in Drug-Like Compounds. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 103–108.
- [116] J. W. Raymond, I. A. Watson, A. Mahoui. Rationalizing Lead Optimization by Associating Quantitative Relevance with Molecular Structure Modification. *Journal of Chemical Information and Modeling* **2009**, *49*, 1952–1962.
- [117] J. Hussain, C. Rea. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *Journal of Chemical Information and Modeling* **2010**, *50*, 339–348.
- [118] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann. 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 Computer Sciences* **1998 May-Jun**, *38*, 511–522.
- [119] A. d. l. V. de León, J. Bajorath. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *MedChemComm* **2013**, *5*, 64–67.
- [120] N. M. O’Boyle, J. Boström, R. A. Sayle, A. Gill. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. *Journal of Medicinal Chemistry* **2014**, *57*, 2704–2713.
- [121] D. Stumpfe, D. Dimova, J. Bajorath. 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.
- [122] G. W. Bemis, M. A. Murcko. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.



- [123] D. Dimova, D. Stumpfe, Y. Hu, J. Bajorath. Analog Series-Based Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. *Future Science OA* **2016**, *2*, FSO149.
- [124] D. Dimova, D. Stumpfe, J. Bajorath. Computational Design of New Molecular Scaffolds for Medicinal Chemistry, Part II: Generalization of Analog Series-Based Scaffolds. *Future Science OA* **2017**, *4*, FSO267.
- [125] K. Lundstrom. Genomics and Drug Discovery. *Future Medicinal Chemistry* **2011**, *3*, 1855–1858.
- [126] B. A. Manjasetty, K. Büssow, S. Panjekar, A. P. Turnbull. Current Methods in Structural Proteomics and Its Applications in Biological Sciences. *3 Biotech* **2012**, *2*, 89–113.
- [127] T. L. Blundell, H. Jhoti, C. Abell. High-Throughput Crystallography for Lead Discovery in Drug Design. *Nature Reviews Drug Discovery* **2002**, *1*, 45–54.
- [128] G. T. Montelione, D. Zheng, Y. J. Huang, K. C. Gunsalus, T. Szyperski. Protein NMR Spectroscopy in Structural Genomics. *Nature Structural & Molecular Biology* **2000**, *7*, 982–985.
- [129] Y. Zhang, B. Sun, D. Feng, H. Hu, M. Chu, Q. Qu, J. T. Tarrasch, S. Li, T. Sun Kobilka, B. K. Kobilka, G. Skiniotis. Cryo-EM Structure of the Activated GLP-1 Receptor in Complex with a G Protein. *Nature* **2017**, *546*, 248–253.
- [130] D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nature Reviews Drug Discovery* **2004**, *3*, 935–949.
- [131] S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlić, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva, C. Zardecki. RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy. *Nucleic Acids Research* **2019**, *47*, D464–D474.

- [132] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, D. S. Wishart. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Research Database* issue **2014**, *42*, D1091–1097.
- [133] R. Wang, X. Fang, Y. Lu, S. Wang. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* **2004**, *47*, 2977–2980.
- [134] P. W. Rose, A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman, S. K. Burley. The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Research* **2017**, *45*, D271–D281.
- [135] V. Cherezov, E. Abola, R. C. Stevens. Toward Drug Design: Recent Progress in the Structure Determination of GPCRs, a Membrane Protein Family with High Potential as Pharmaceutical Targets. *Methods in Molecular Biology* **2010**, *654*, 141–168.
- [136] C. Bissantz, B. Kuhn, M. Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry* **2010**, *53*, 5061–5084.
- [137] R. E. Babine, S. L. Bender. Molecular Recognition of Protein-Ligand Complexes: Applications to Drug Design. *Chemical Reviews* **1997**, *97*, 1359–1472.
- [138] J. A. Caro, K. W. Harpole, V. Kasinath, J. Lim, J. Granja, K. G. Valentine, K. A. Sharp, A. J. Wand. Entropy in Molecular Recognition by Proteins. *Proceedings of the National Academy of Sciences* **2017**, *114*, 6563–6568.
- [139] N. Brooijmans, I. D. Kuntz. Molecular Recognition and Docking Algorithms. *Annual Review of Biophysics and Biomolecular Structure* **2003**, *32*, 335–373.
- [140] J. L. Medina-Franco, O. Méndez-Lucio, K. Martinez-Mayorga. Chapter One - The Interplay Between Molecular Modeling and Chemoinformatics to Characterize Protein-Ligand and Protein-Protein Interactions Landscapes for Drug Discovery. In: *Advances in Protein Chemistry and Structural Biology*. Ed. by T. Karabencheva-Christova. Vol. 96. Biomolecular Modelling and Simulations. Academic Press, **2014**, 1–37.

- [141] Y. Hu, J. Bajorath. Exploration of 3D Activity Cliffs on the Basis of Compound Binding Modes and Comparison of 2D and 3D Cliffs. *Journal of Chemical Information and Modeling* **2012**, *52*, 670–677.
- [142] N. Furtmann, Y. Hu, J. Bajorath. Comprehensive Analysis of Three-Dimensional Activity Cliffs Formed by Kinase Inhibitors with Different Binding Modes and Cliff Mapping of Structural Analogues. *Journal of Medicinal Chemistry* **2015**, *58*, 252–264.
- [143] K. Stierand, M. Rarey. Drawing the PDB: Protein-Ligand Complexes in Two Dimensions. *ACS Medicinal Chemistry Letters* **2010**, *1*, 540–545.
- [144] C. Da, D. Kireev. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *Journal of Chemical Information and Modeling* **2014**, *54*, 2555–2561.
- [145] Z. Deng, C. Chuaqui, J. Singh. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *Journal of Medicinal Chemistry* **2004**, *47*, 337–344.
- [146] J. Desaphy, E. Raimbaud, P. Ducrot, D. Rognan. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *Journal of Chemical Information and Modeling* **2013**, *53*, 623–637.
- [147] A. M. Clark, P. Labute. 2D Depiction of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling* **2007**, *47*, 1933–1944.
- [148] I. Muegge, Y. C. Martin. A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach. *Journal of Medicinal Chemistry* **1999**, *42*, 791–804.
- [149] Molecular Operating Environment (MOE). *MOE2018.01*. **2018**.
- [150] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinel, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel. KNIME: The Konstanz Information Miner. In: *Data Analysis, Machine Learning and Applications*. Ed. by C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg, **2008**, 319–326.
- [151] 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, M. Glick. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nature Chemical Biology* **2015**, *11*, 958–966.

- [152] S. Jasial, Y. Hu, J. Bajorath. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLOS ONE* **2016**, *11*, e0153873.
- [153] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, J. Zhang. PubChem BioAssay: 2017 Update. *Nucleic Acids Research* **2017**, *45*, D955–D963.
- [154] RDKit: Cheminformatics and Machine Learning Software. *RDKit*. **2018**.
- [155] T. Sterling, J. J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.
- [156] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Research* **2012**, *40*, D1100–D1107.
- [157] V. Vapnik. The Nature of Statistical Learning Theory. 2nd ed. Information Science and Statistics. New York: Springer-Verlag, **2000**.
- [158] J. Balfer, J. Bajorath. Visualization and Interpretation of Support Vector Machine Activity Predictions. *Journal of Chemical Information and Modeling* **2015**, *55*, 1136–1147.
- [159] S. Jasial, E. Gilberg, T. Blaschke, J. Bajorath. 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.
- [160] C. Ferroud, P. Rool, J. Santamaria. Singlet Oxygen Mediated Alkaloid Tertiary Amines Oxidation by Single Electron Transfer. *Tetrahedron Letters* **1998**, *39*, 9423–9426.
- [161] E. E. Carlson, J. F. May, L. L. Kiessling. Chemical Probes of UDP-Galactopyranose Mutase. *Chemistry & Biology* **2006**, *13*, 825–837.
- [162] J. T. Metz, J. R. Huth, P. J. Hajduk. Enhancement of Chemical Rules for Predicting Compound Reactivity towards Protein Thiol Groups. *Journal of Computer-Aided Molecular Design* **2007**, *21*, 139–144.

- [163] M. E. Voss, P. H. Carter, A. J. Tebben, P. A. Scherle, G. D. Brown, L. A. Thompson, M. Xu, Y. C. Lo, G. Yang, R.-Q. Liu, P. Strzemienski, J. G. Everlof, J. M. Trzaskos, C. P. Decicco. 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- Antagonists. *Bioorganic & Medicinal Chemistry Letters* **2003**, *13*, 533–538.
- [164] M. M. McCallum, P. Nandhikonda, J. J. Temmer, C. Eyermann, A. Simeonov, A. Jadhav, A. Yasgar, D. Maloney, A. L. Arnold. High-Throughput Identification of Promiscuous Inhibitors from Screening Libraries with the Use of a Thiol-Containing Fluorescent Probe. *Journal of Biomolecular Screening* **2013**, *18*, 705–713.
- [165] K. M. Soares, N. Blackmon, T. Y. Shun, S. N. Shinde, H. K. Takyi, P. Wipf, J. S. Lazo, P. A. Johnston. Profiling the NIH Small Molecule Repository for Compounds That Generate H<sub>2</sub>O<sub>2</sub> by Redox Cycling in Reducing Environments. *Assay and Drug Development Technologies* **2010**, *8*, 152–174.