Chemoinformatics-Driven Approaches for Kinase Drug Discovery

Kumulative Dissertation

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

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Filip Miljković

aus Niš, Serbien

Bonn

September, 2019

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

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Michael Gütschow
Tag der Promotion: Dec 17, 2019
Erscheinungsjahr: 2020

Abstract

Given their importance for the majority of cell physiology processes, protein kinases are among the most extensively studied protein targets in drug discovery. Inappropriate regulation of their basal levels results in pathophysiological disorders. In this regard, small-molecule inhibitors of human kinome have been developed to treat these conditions effectively and improve the survival rates and life quality of patients. In recent years, kinase-related data has become increasingly available in the public domain. These large amounts of data provide a rich knowledge source for the computational studies of kinase drug discovery concepts.

This thesis aims to systematically explore properties of kinase inhibitors on the basis of publicly available data. Hence, an established "selectivity versus promiscuity" conundrum of kinase inhibitors is evaluated, close structural analogs with diverging promiscuity levels are analyzed, and machine learning is employed to classify different kinase inhibitor binding modes. In the first study, kinase inhibitor selectivity trends are explored on the kinase pair level where kinase structural features and phylogenetic relationships are used to explain the obtained selectivity information. Next, selectivity of clinical kinase inhibitors is inspected on the basis of cell-based profiling campaign results to consolidate the previous findings. Further, clinical candidates are mapped to medicinal chemistry sources and promiscuity levels of different inhibitor subsets are estimated, including designated chemical probes. Additionally, chemical probe analysis is extended to expert-curated representatives to correlate the views established by scientific community and evaluate their potential for chemical biology applications. Then, large-scale promiscuity analysis of kinase inhibitor data combining several public repositories is performed to subsequently explore promiscuity cliffs (PCs) and PC pathways and study structure-promiscuity relationships. Furthermore, an automated extraction protocol prioritizing the most informative pathways is proposed with focus on those containing promiscuity hubs. In addition, the generated promiscuity data structures including cliffs, pathways, and hubs are discussed for their potential in experimental and computational follow-ups and subsequently made publicly available. Finally, machine learning methods are used to develop classification models of kinase

inhibitors with distinct experimental binding modes and their potential for the development of novel therapeutics is assessed.

Својој породици и пријатељима, на пруженом разумевању и подршци током докторских студија.

У сећање на Бисерку Стојановић и све наше драгоцене тренутке проведене заједно...

Contents

Chapter 1

Introduction

1.1 Protein Kinases

Protein kinases (or simply kinases) are a large family of enzymes responsible for the catalysis of protein phosphorylation processes.[1–3] This enzyme family belongs to the class of transferases (subclass: phosphotransferases). They mediate the transfer of the γ-phosphate group from adenosine triphosphate (ATP) to amino acid residues of protein substrates with a free hydroxyl group - serine (Ser), threonine (Thr), and tyrosine (Tyr). The products of this enzymatic reaction are adenosine diphosphate (ADP) and phospho-protein.

The covalently attached phosphate group transforms the substrate protein in different ways. These include the change in its activity, location within a cell, protein turnover or interaction with other macromolecules.[4,5] In addition, kinases regulate their own activity through the process of autophosphorylation. Phosphorylation processes can be terminated or reversed by a separate class of enzymes called phosphatases.[6,7]

Kinase targets are of great importance for drug discovery. Hence, their classification, structural context, importance in (patho)physiology and drug development will be discussed in the following.

1.1.1 Classification

In 2002, Manning *et al.* reported that the kinase family consists of 478 typical and 40 atypical enzymes, amounting to a total of 518 members.[1] As one of the largest protein families in the human proteome, kinases constitute ~2%

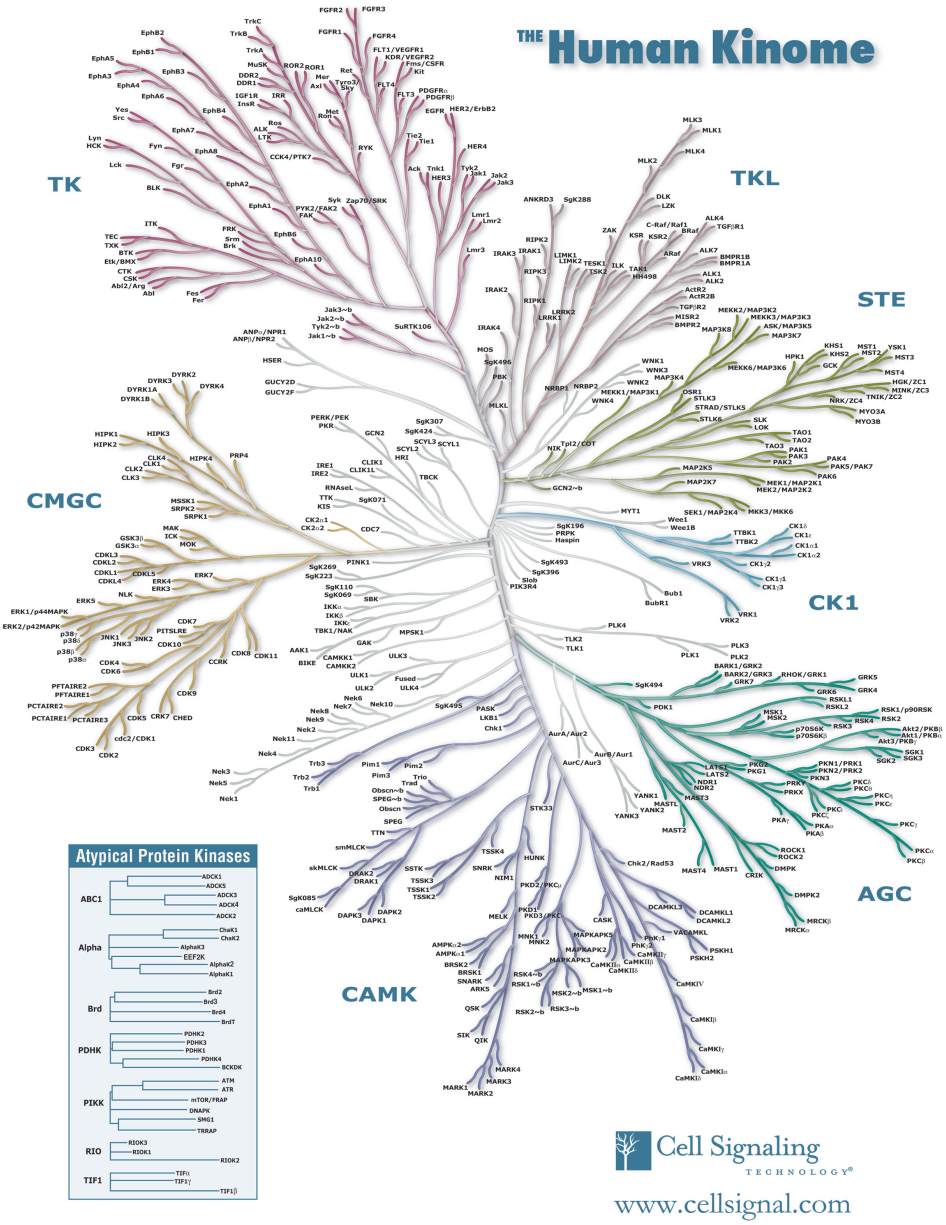


Figure 1.1: Human kinome. A phylogenetic tree representation of the human kinome is shown. Clustering of the leaf nodes denotes their structural and functional similarity, where branches containing neighboring nodes associate them in smaller classes (families and subfamilies) and larger branches of the same part of the tree depict their wider classification (group). Each kinase group is marked with capital blue letters whereas individual kinases are given in black letters. Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com).

of the human genome.[1,2] This assembly of 518 kinases is often referred to as the human kinome.

As stated, kinases catalyze the transfer of a phosphate group to Ser, Thr, and Tyr residues of substrate proteins. This forms the basis for their classification as protein-serine/threonine kinases (385 representatives), protein-tyrosine kinases (90), and protein-tyrosine kinase-like group (43). Moreover, protein-tyrosine kinases can be divided into receptor (58) and non-receptor (32) kinases.[1] Thus, a majority of the kinases act on both serine and threonine residues, whereas the others act only on tyrosine. In addition, only a small number of kinases acts on all three residues (dual-specificity kinases).[1]

Although the majority of human kinases share a common catalytic domain, sequence analysis showed substantial variations with distinct and ancient functions. In order to quantify this diversity, a standard kinase classification scheme was proposed. This classification took into consideration the evolutionary history, functions, sequence, and structural similarity. Accordingly, kinases were assigned to kinase groups consisting of many families, whereas some families were further divided into multiple subfamilies. Groups display broad specificity of substrate sites, families classify kinases by both broad biological function and sequence similarity, while subfamilies signify finer functional and sequence similarity. This classification was first published by Manning *et al.* It extended the previous classification scheme proposed by Hanks and Hunter.[8] In 1995, Hanks and Hunter clustered kinases into 5 groups, 44 families, and 51 subfamilies.[8] Seven years later, Manning *et al.* extended this classification to 9 groups, 134 families, and 196 subfamilies.[1] Kinase classification has further been refined over the years, mainly by Manning and coworkers at Salk Institute. Schematic representation of the phylogenetic tree of human kinome is shown in **Figure** 1.1.

1.1.2 Molecular Structure

Given that kinases possess a high level of structural complexity, their study from the aspect of structural biology requires special attention. Herein, the basic structural features will be outlined, as well as their importance for catalytic activity and drug research.

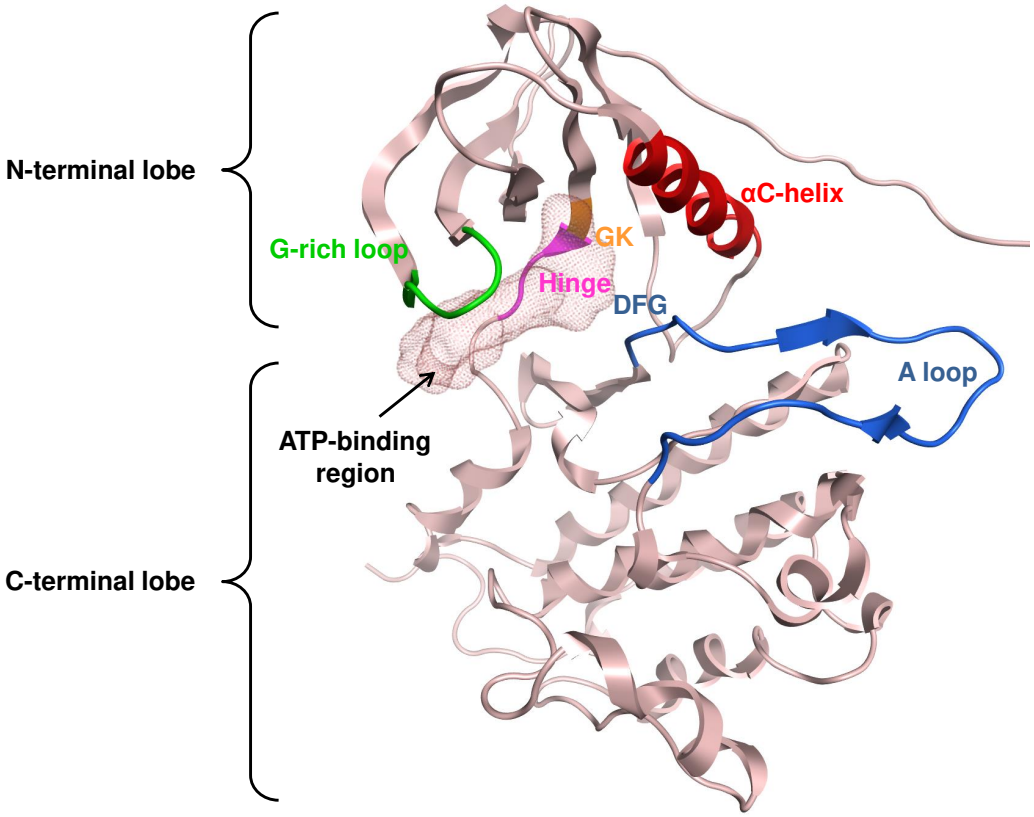


Figure 1.2: Kinase structure. Three-dimensional structure of EGFR kinase (PDB ID: 1M17) is schematically represented. Some of the discussed structural features are depicted as follows: N-terminal lobe, C-terminal lobe, G-rich loop (green), hinge (pink), GK residue (orange), αC-helix (red), and A loop with DFG motif (blue). Molecular surface of co-crystalized small-molecule inhibitor erlotinib marks the location of ATP-binding region.

Kinases are composed of two lobes of different sizes: a smaller amino(N)- and a larger carboxy(C)-terminal lobe (**Figure** 1.2).[9] The N-terminal lobe is largely composed of an antiparallel β-sheet consisting of five strands (β1 - β5).[10] Additionally, it holds a regulatory αC-helix that acquires either active or inactive orientation. A conserved glycine(G)-rich loop, sometimes called P-loop, is found between strands β1 and β2. It is responsible for binding phosphate group of ATP. Conserved Lys residue of β3 strand forms a salt bridge with Glu of αC-helix. This is a prerequisite for the active "αC_{in}" conformation of kinases. The absence of this salt bridge drives the kinase into the inactive state ("αC_{out}"). However, the presence of the salt bridge is not sufficient for the expression of full kinase activity.[11,12]

Opposite to N-terminal lobe, C-terminal lobe is mainly composed of helices, eight in total (αD - αI, αEF1, and αEF2).[13] Four conserved β-strands (β6 - β9) are also present in the larger lobe of active kinases. In inactive kinases, this structural composition is disrupted. In addition, the larger lobe contains residues of the catalytic loop. The catalytic loop participates in phosphate transfer between ATP and substrate protein. Key regulatory element of kinases present in the C-terminal lobe is the activation segment. This 35-40 residues long segment begins with the DFG (Asp - Phe - Gly) motif in nearly all kinases.[14] It controls both catalytic efficiency of an enzyme and binding of a protein substrate. The central part of the activation segment is known as the activation(A) loop. In all active kinases, the activation segment adopts an open or extended conformation, whereas the inactive form of a kinase contains the closed conformation. In the former case, the DFG motif is positioned towards the ATP-binding site, and is termed "DFG_{in}". In the latter, the motif is extended in the opposite direction, and is named "DFG_{out}". In addition, the presence of two Mg^{2+} ions is important for the catalytic activity of most kinases.

Both lobes assemble to form the catalytic spine (C-spine) and the regulatory spine (R-spine).[15,16] The C-spine is essential for ATP positioning, whereas the R-spine interacts with a substrate to allow catalysis. The R-spine consists of residues from both the activation segment and αC-helix. The segment of kinases connecting the small and the large lobes is called the "hinge". The hinge consists of several conserved residues that provide essential hydrogen bond interactions with the adenine moiety of ATP. The majority of current kinase inhibitors bind competitively to the ATP-binding site, forming hydrogen bonds with hinge residues.[17]

Adjacent to the hinge region, at the very end of the β5 strand, is positioned the "gatekeeper"(GK) residue. The terminology behind the name signifies its role in limiting the access to the hydrophobic pocket adjacent to the ATP-binding site.[18,19] This is particularly important for the design of small-molecule kinase inhibitors that exploit interactions with the residues of that region. Different sizes of GK residues in kinases, as well as their mutations, are of great importance for inhibitor selectivity and therapeutic efficiency, respec-

tively. About 77% of kinases contain a relatively large GK residue (e.g., Leu, Met, Phe), whereas others contain smaller residues (e.g., Thr, Val).[20]

1.1.3 Clinical Significance

Members of the kinase family assume a wide range of roles in cell physiology. Primarily, kinases mediate most of the signal transduction processes in cells of eukaryotic organisms, including human cells. In addition, they control transcription, metabolism, cytoskeletal rearrangement and cell movement, cell cycle progression, differentiation, and apoptosis. A crucial point of kinase-directed phosphorylation is found in homeostasis and physiological responses, intercellular communication during the development stage, as well as the performance of immune and nervous systems. Evidently, the catalytic activity of kinases is crucial to the maintenance of majority of cell physiology processes.[21]

Consequently, mutation, overexpression, and dysregulation of kinases is the foundation of pathogenesis in multiple diseases. These include autoimmune, asthma, cardiovascular, inflammatory, metabolic, and neurological disorders.[22–27] However, the widely explored therapeutic indications in kinase drug discovery are oncology-related as demonstrated by the number of designed therapeutics.[4,28,29] Despite their widespread potential, the clinical progress of therapeutic indications involving kinases has been uneven.

Development of small-molecule modulators, in particular inhibitors, has revolutionized the treatment of certain diseases, such as gastrointestinal stromal tumors and chronic myeloid leukaemia. For these, and other conditions, kinase inhibitors increased survival rates in patients tremendously.[30–32] On the other hand, smaller but significant responses were observed for cancer types that highly depend on angiogenesis. These forms of cancer, such as renal cell carcinoma, are in particular sensitive to inhibitors that target signaling pathways of vascular endothelial growth factor (VEGF).[33–35] By far, the least therapeutic efficiency has been achieved for breast, colorectal, lung, pancreatic, and prostate cancer. These oncological disorders still retain the highest mortality rates. Kinase inhibitors targeting these forms are successful in prolonging the survival rate of patients by only a few months.[36–38]

Besides oncology, equally exciting opportunities for kinase drug discovery exist for the other diseases.[39] For example, pan-Janus kinase (pan-JAK) inhibitor tofacitinib is approved for the treatment of rheumatoid arthritis.[40] Bruton tyrosine kinase (BTK) inhibitors such as evobrutinib[41] and MSC2364447C[42] are clinical candidates for rheumatoid arthritis and systemic lupus erythematosus, respectively. In addition, a combined inhibitor of both phosphoinositide 3-kinase (PI3K) δ and γ, duvelisib is currently under Phase II of clinical development for mild asthma.[43] Considerable attention is paid for inhibitors of VEGF receptor (VEGFR) and serine/arginine-rich protein-specific kinase 1 (SRPK1), such as X-82[44] and SPHINX31[45] respectively, in the treatment of wet age-related macular degeneration. Clearly, the importance of kinase targets for the contemporary drug discovery is multifaceted.[39]

1.1.4 Inhibitors of Human Kinome

The most common mechanism of action of kinase-directed therapeutics is inhibition. The development of kinase inhibitors has taken two paths. On one hand, monoclonal antibodies such as bevacizumab and cetuximab have been developed as extracellular inhibitors of protein-tyrosine kinases VEGF and epidermal growth factor receptor (EGFR), respectively.[46] On the other, small-molecule inhibitors such as crizotinib and dabrafenib, have been introduced as inhibitors of anaplastic lymphoma kinase (ALK) and serine/threonine-protein kinase B-Raf (BRAF) Val600Glu mutant, respectively.[46] The following sections will focus on small-molecule inhibitors of human kinome.[4,5,28]

As of April 12, 2019, the U.S. Food and Drug Administration (FDA)[47] has approved 49 small-molecule kinase inhibitors for the market.[48] Nearly all of them are orally effective, except for netarsudil[49] (given as an eye drop) and temsirolimus[50] (given intravenously). Among the approved drugs, 26 inhibit receptor and 10 non-receptor protein-tyrosine kinases. The remaining 13 inhibitors target protein-serine/threonine kinases. The majority of FDA-approved kinase inhibitors (43) are used to treat malignancies (36 against solid tumors and seven against non-solid tumors). Eight of 49 inhibitors are used for non-oncological therapeutic indications. Two kinase inhibitors, ibrutinib and sirolimus, are used in the treatment of both malignant and non-malignant

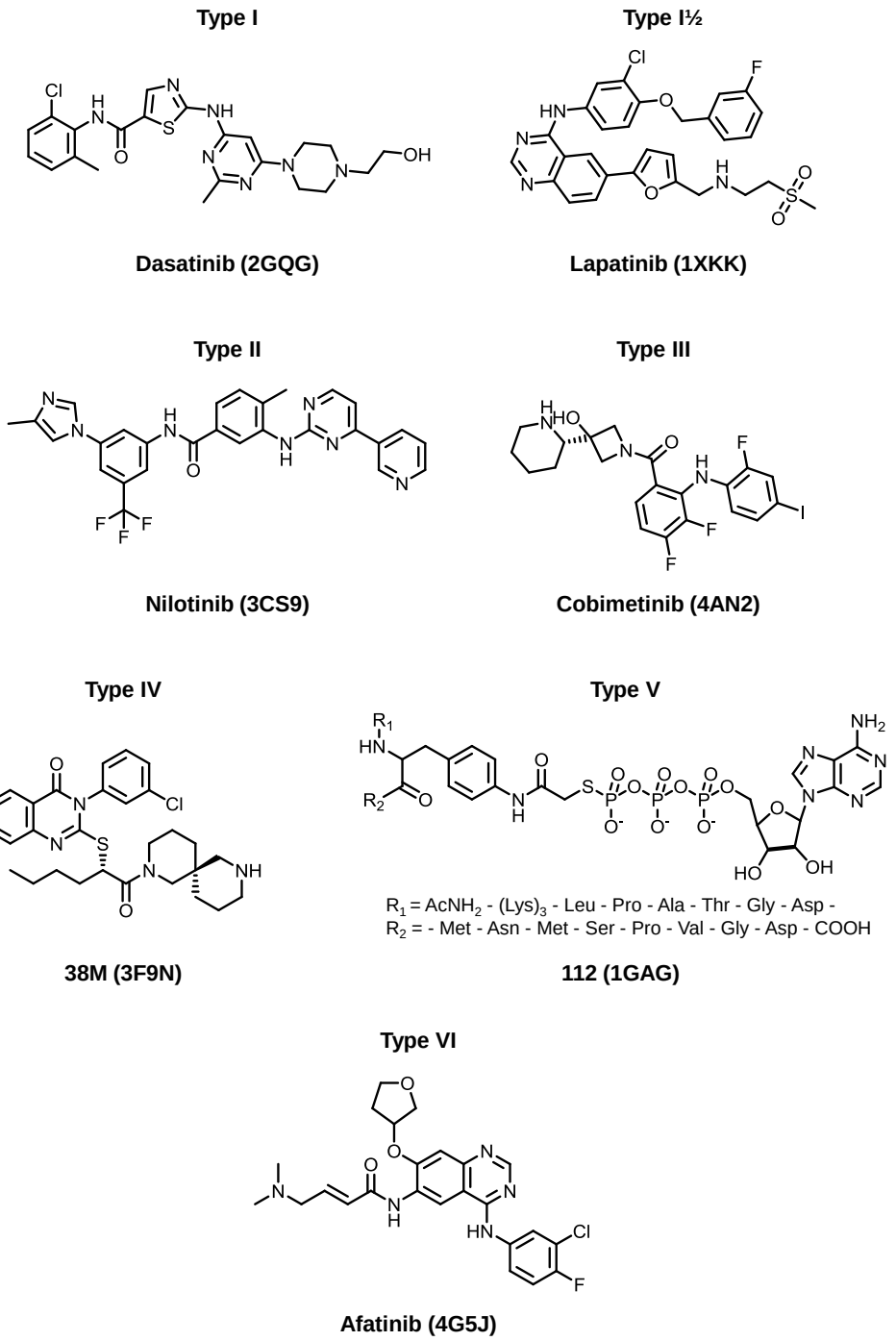


Figure 1.3: Types of kinase inhibitors. Kinase inhibitor examples of each discussed inhibitor type are shown. Below each inhibitor, its clinical name or three-character PDB name is given. In the brackets, PDB IDs are given for each crystal structure used to extract the corresponding ligand.

disorders. Non-malignant therapeutic indications targeted by approved kinase inhibitors are rheumatoid arthritis (baricitinib, tofacitinib), chronic immune thrombocytopenia (fostamatinib), myelofibrosis and polycythemia vera (ruxolitinib), idiopathic pulmonary fibrosis (nintedanib), renal graft versus host disease (sirolimus and ibrutinib), glaucoma (netarsudil), and psoriatic arthritis and ulcerative colitis (tofacitinib).[48]

Most commonly targeted kinases by FDA-approved drugs are ALK, Bcr-Abl, B-Raf, EGFR, and VEGFR. Direct interactions with the kinase domain are observed for 46 of 49 inhibitors. The remaining three, everolimus, sirolimus, and temsirolimus, bind to FK506-binding protein 12 (FKBP-12) to form a complex which inhibits mammalian target of rapamycin (mTOR). At least 18 inhibitors are known to be multi-kinase inhibitors. Covalent inhibition of kinase targets is achieved by six drugs, namely acalabrutinib, afatinib, dacomitinib, ibrutinib, neratinib, and osimertinib.[48]

The precise number of kinase inhibitors undergoing clinical development is not known. Carles *et al.* reported at least 180 inhibitors taking part in phase 0 to 4 of clinical trials worldwide.[51] On the other hand, Klaeger *et al.* explored selectivity profiles of 243 clinical candidates at various stages of clinical development.[52] The number of marketed drugs has almost doubled in size in the course of three years (2016: 27 drugs; 2019: 49 drugs).[17,48] This increase over a short period of time clearly denotes their importance for drug discovery.

Kinase inhibitors explore distinct kinase pockets as part of their inhibition mechanism. This provides basis for their classification into different inhibitor types.[17] First, Dar and Shokat divided small-molecule kinase inhibitors into three classes, which they labeled as type I, II, and III.[19] Type I inhibitor was defined as a small molecule that binds in the ATP pocket of active kinases (DFG_{in} and αC_{in}). Type II was introduced as compound binding to an inactive DFG_{out} conformation. It explores ATP-binding site and adjacent hydrophobic pocket that opens when the DFG motif assumes the "out" state. A subtype of type I, called type I$^1/_2$, was introduced by Zuccotto *et al.*[53] These inhibitors form hydrogen bonds with the hinge region and extend towards the back cavity, interacting with residues that bind type II pharmacophores. They bind to inactive kinases with DFG_{in} and αC_{out} conformations. Type III was described as an allosteric inhibitor, or a non-ATP competitive inhibitor. Allosteric compounds

bind to regions distinct from commonly explored active sites.[54] For kinase inhibitors, this refers to regions outside the ATP-binding pocket. Furthermore, Gavrin and Saiah divided type III allosteric inhibitors into type III and type IV.[55] According to them, type III inhibitors bind to the pocket adjacent to the ATP-binding site, whereas type IV bind outside the phosphoacceptor region. Type V inhibitors were defined by Lamba and Gosh as bivalent compounds that span two separate regions of the kinase domain.[56] Covalent inhibitors are defined as a separate class of type VI inhibitors.[17] Examples of the discussed inhibitor types are given in **Figure** 1.3.

1.1.5 Kinase Inhibitor Data in Public Domain

With the advent of genomic technologies, high-throughput screens, and computational infrastructures, kinase inhibitor experimental data has become increasingly available in the public domain.[57] In 2014, Hu *et al.* obtained 18,951 kinase inhibitors with high-confidence activity annotations from ChEMBL (release 18), a major repository of data from medicinal chemistry literature.[58,59] As defined previously, high-confidence data from ChEMBL is characterized by activity information with highest assay and highest measurement reliability.[60] These compounds were annotated against 266 kinases. In the course of five years (ChEMBL 24.1 in 2019), the number of available kinase inhibitors almost tripled in size. Using identical extraction criteria, 53,220 kinase inhibitors active against 311 kinases were found in ChEMBL.[59] Furthermore, as of June 15, 2019, more than 4450 structures of human catalytic kinase domains are available in Protein Data Bank (PDB).[61] In spite of the different origins, this data presents a rich source of knowledge for computationally-driven efforts in kinase drug discovery.

For example, large-scale analysis of kinase inhibitors can reveal their structure-property relationships to complement the current chemical optimization efforts for specific kinases. On the other hand, multi-kinase activities of inhibitors can be systematically explored to assess their potential for different therapeutic areas. Additionally, activity information of multi-kinase inhibitors may be used in *in silico* selectivity profiling studies.[62–64] In turn, this may reveal compound features participating in the differentiation of kinase targets. This

is of particular interest in the area of chemical biology where small-molecule chemical probes are used to explore biological functions of protein targets.[65]

These and many other analyses would not be possible without the increasing volumes of compound information from medicinal chemistry. Hence, chemoinformatics approaches can be used to explore the trends of kinase inhibitor data and their potential use for drug discovery. In the following, the chemoinformatics concepts used for large-scale analysis of kinase inhibitor activity data are discussed.

1.2 Molecular Representations

Computer-friendly representations of compounds are required for efficient storage and manipulation. Two-dimensional (2D) structures of molecules can be understood as graphs, where nodes correspond to heavy atoms and edges to bond relationships.[66] In such graphs, nodes store atom characteristics, such as charge or hybridization state, whereas edges capture bond information, such as bond order or stereochemistry. Moreover, the information about hydrogen atoms is treated as a special attribute of heavy atom nodes. Correspondingly, a graph provides topological information about each molecule, usually stored as a connectivity table. However, further simplification is possible with linear representations, which in computer code are treated as strings. Some of the most popular linear representations of molecular structures are Simplified Molecular-Input Line-Entry System (SMILES)[67] and International Chemical Identifier (InChI).[68]

SMILES representations were established to efficiently store chemical information and facilitate their retrieval and modeling.[67,69,70] On the basis of predefined rules, graphs of molecular structures are transformed into strings of ASCII characters. Atoms are represented as their atomic symbols and branching is enclosed in parentheses in SMILES notation. Special symbols are defined for aromaticity, chirality, isotope presence, and stereochemistry of a molecule. A typical SMILES annotation will usually take 50% to 70% less space compared to graph-based connectivity tables. Because of the ability to canonicalize their representation, SMILES notations are highly popular and practical. This allows efficient data storage and retrieval, which is not possible without canon-

icalization. Therefore, the consistent use of canonicalized SMILES presents a cornerstone for chemoinformatics approaches.

1.2.1 Descriptors

In addition to graph- and string-based representations, compounds can be characterized by numerical molecular descriptors. These describe properties and structural features of molecules. Some molecular descriptors, such as physico-chemical properties, can be derived from experimental measurements (e.g., dipole moment, partition coefficient, and molar refractivity), whereas others can be obtained *in silico* from mathematical models using an appropriate molecular representation (e.g., SMILES). Molecular descriptors account for a variety of physico-chemical, surface, topological, and other properties describing small molecules.[71–73] Their retrieval depends on the dimensionality (D) information of a molecule. In this regard, molecular descriptors can be classified as 1D, 2D, and 3D descriptors.[74] 1D descriptors are calculated directly from the molecular formula or linear annotation and are commonly known as constitutional descriptors (e.g., number of atoms, bond count, molecular weight). 2D descriptors are based on molecular topology and thus require molecular structure for their calculation (e.g., topological indices, fragment counts). Lastly, 3D descriptors require a specific three-dimensional conformation of a molecule that provides geometrical parameters (e.g., molecular surfaces and fields, parameters calculated in quantum chemistry programs). As an extension, 4D descriptors are described as representations merging several conformations (an ensemble of conformations) that combine both conformational flexibility and alignment freedom.[75] Descriptors range from those that are relatively simple to calculate to those having much higher complexity. The values of some descriptors may be shared among many compounds, whereas others are more discriminatory in nature. Therefore, the selection of an appropriate descriptor set is often pivotal to success.

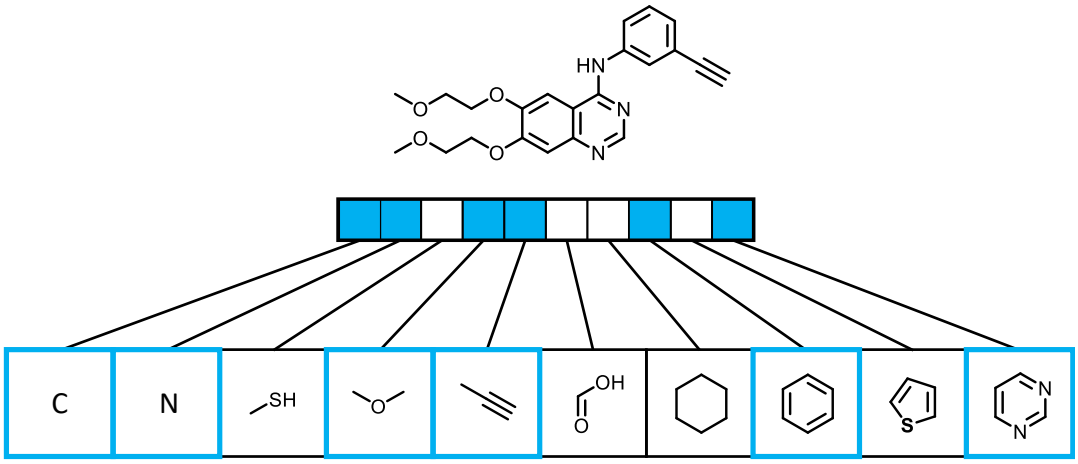


Figure 1.4: Substructure-based fingerprint. A substructure-based fingerprint of length ten is shown for a kinase inhibitor erlotinib. When a feature is present, a corresponding bit is set to "1" (blue), otherwise it is set to "0" (white). Substructural features comprising the fingerprint are shown as molecular patterns in an enlarged bit vector representation. The figure is adapted from Stumpfe *et al.*[76]

1.2.2 Fingerprints

Molecular fingerprints are among the most widely used group of descriptors for chemoinformatics applications. They are defined as bit string representations of molecular properties or structural features. In binary fingerprints, each bit position encodes the presence or absence of a feature thus taking either of the two values: "1" (feature is present) or "0" (feature is absent). Computationally, this array of bits can be manipulated and compared rather efficiently. This establishes the basis for their extensive use in areas such as similarity searching and machine learning.

Fingerprints can vary substantially in their design, length, and complexity.[76,77] One of the most commonly applied fingerprints are substructure-based fingerprints. These fingerprints are fixed in their length, where each position corresponds to a predefined substructural pattern. A prime example of substructure-based fingerprints is given by Molecular ACCess System (MACCS) keys that consist of 166 bit positions.[78] **Figure** 1.4 shows an example of a substructure-based fingerprint with arbitrary length.

Another popular fingerprint design are combinatorial fingerprints. In contrast to substructure-based design, combinatorial fingerprints have a flexible length as their features are not predefined. Instead, all possible substructural

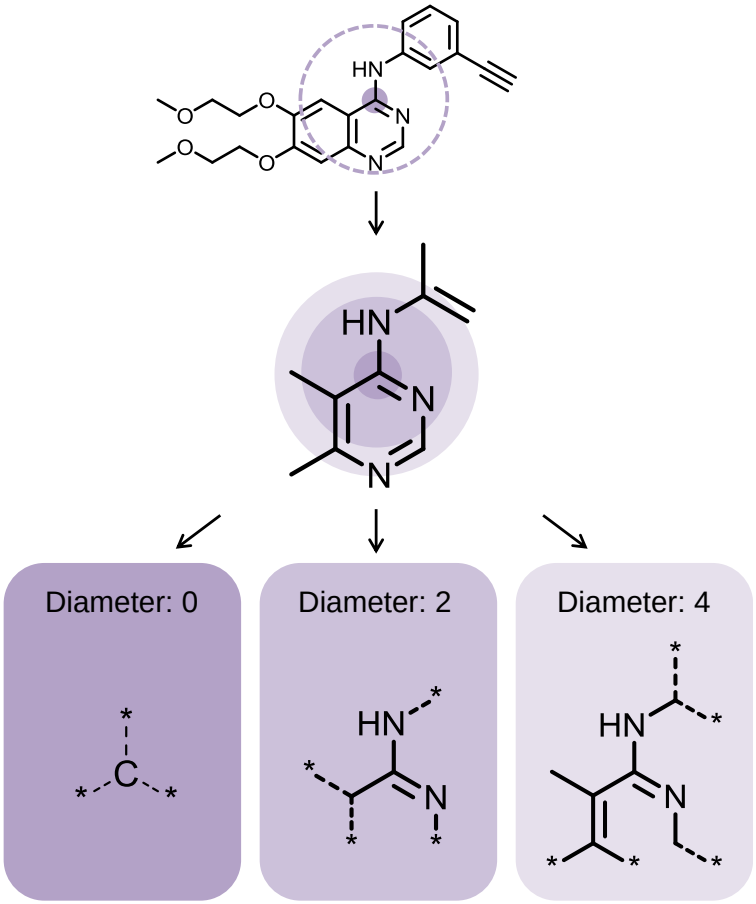


Figure 1.5: ECFP fingerprint. Calculation of an ECFP fingerprint with bond diameter four is shown for a representative carbon atom (deep purple) of erlotinib. Each increasing diameter representation is shown in a lighter shade of purple. The resulting topological environment per diameter is shown as structural pattern. The connectivity information outside observed diameter is given by dashed line, with dummy atoms represented by asterisks (*). The figure is adapted from Stumpfe *et al.*[76]

graphs up to a given size are extracted and then hashed to obtain numbers. The most prominent representative of this group is the Extended Connectivity FingerPrint (ECFP).[79] ECFP captures layered atom environments (topology) around each non-hydrogen atom up to a predefined bond diameter. For example, ECFP4 limits the bond diameter to four to delineate atom neighborhoods of growing size. An example of how ECFP fingerprints of increasing diameter are derived is shown in **Figure** 1.5. With the use of a hash function, features of ECFP fingerprints can be folded to obtain a specified number of bits, most commonly 1024-bit or 2048-bit representations. This however may introduce a

problem for the feature selection process because structural patterns of different origin might correspond to the same bit position.[80]

1.3 Structure-Property Relationships

Simple molecular representations such as SMILES can be used to extract a wide array of compound properties. However, not all compound properties are equally considered during the chemical optimization. Increasing volumes of compound data can be used to analyze structure-property relationships on a large scale in order to improve our understanding of properties that require optimization. Structure-property relationship studies of compounds are one of the key concepts of chemoinformatics. It relies on the "similarity property principle", which states that structurally similar molecules share similar properties.[81] For the design of compounds with desirable therapeutic efficacy, biological activity (or simply activity) is one of the most important properties to consider. Accordingly, the properties derived from well-defined activity values, such as selectivity and promiscuity, are important for drug discovery applications. An increasing number of experimental profiling campaigns is frequently supported by *in silico* studies, which aim to further advance our knowledge of compound selectivity and promiscuity. This holds true for kinase drug discovery, where the "selectivity versus promiscuity" balance plays a leading role for a number of therapeutic applications.

1.3.1 Activity and Selectivity

In accordance with drug discovery objectives, many chemoinformatics approaches rely on the exploration of structure-activity relationships (SARs) of compounds. SAR analysis evaluates how structural modifications in compounds affect their binding characteristics towards a specific target.[82] In particular, SAR analysis aims to explore sets of compounds sharing a high level of structural similarity, such as analog series. This enables us to study subtle structural differences responsible for high (or low) levels of activity. Although SAR analysis seeks to identify compounds with favorable activity values, this is often not sufficient to satisfy drug discovery efforts.

The development of active compounds that are selective for a target of interest has been, and continues to be, a central goal of drug discovery.[83] The compound selectivity requirement was initially propagated by the compound specificity paradigm in the 1980s. This paradigm states that a compound should be highly specific for its target (single-target compounds). With the advent of screening technologies, it was shown that many of the approved drugs elicit multiple target activity (multi-target compounds) as part of their mechanism of action.[84–86] Although many compounds elicit their therapeutic effect through modulation of multiple targets, compound selectivity will remain relevant for many research areas, including medicinal chemistry and chemical biology.[65,87]

Kinase inhibitors are a prime example of compounds for which the interplay between selectivity and multi-target activity plays a decisive role in determining their therapeutic efficacy.[22–28] Given that the majority of kinase inhibitors bind to the largely conserved ATP-binding site (type I inhibitors), these compounds are expected to have multi-kinase activity. This has been demonstrated by a number of experimental studies.[88–91] The multi-kinase activity resembles compound promiscuity, where small molecules specifically interact with multiple targets (kinases).[87] However, other kinase profiling studies provided contrasting views where many kinase inhibitors showed desirable selectivity patterns.[52,92,93] In addition to the active site, type II inhibitors explore an adjacent hydrophobic pocket that opens when the DFG motif assumes the "out" conformation.[19] As this pocket is structurally less conserved than the ATP-binding site, type II inhibitors are expected to be more selective than type I inhibitors.[94] However, some profiling experiments provided opposing evidence finding that type II inhibitors were frequently promiscuous.[94] Hence, the typically assumed selectivity differences of these two inhibitor types require further investigation.[95,96]

To support this, extensive structural studies of type I and II kinase inhibitor binding modes were conducted.[58,94,97] It was found that many type II inhibitors contained a typical "type I head" fragment and a set of moieties specific for the type II group. These type II-specific features included a combination of a hydrogen bond donor acceptor pair (e.g., amide or urea functional group) and a hydrophobic tail. Thus, a set of 70 fragment pairs (seven hydrogen bonding linkers and 10 hydrophobic tails) were established as a guideline for rational design of type II inhibitors.[94] However, the structural analysis confirmed that

the often assumed selectivity advantage of type II inhibitors could not be fully supported.[94,97] On the other hand, type I1/2 inhibitors explore kinase-specific subpockets and binding interactions common to both type I and type II inhibitors.[53] Thus, type I1/2 inhibitors may in fact be the most selective compared to the inhibitors with closely related type I and II binding modes.

On the other hand, type III and IV inhibitors are claimed to be more selective as they exploit binding pockets and regulatory mechanisms unique to particular kinases.[55] Only few have been reported to date as they are mainly the product of empirical research. Considerable success has been achieved in the field of targeted covalent inhibitors which show advantages over reversible counterparts.[98] First and foremost, their residence time is considerably increased. This allows application of lower doses which minimize potential off-target activity.[99] In addition, covalent inhibitors target residues unique to their targets (e.g., Cys) which additionally improves their selectivity. To this date, six of them have been approved to the market by the FDA.[48] Clearly, substantial efforts have been made so far to design selective kinase inhibitors serving as clinical candidates.

Moreover, compound selectivity plays a decisive role in other research areas such as chemical biology.[65] Here, small-molecule chemical probes are used as indispensable tools to interrogate biological consequences of target modulation.[100,101] Naturally, chemical probes should be subjected to stringent requirements as part of their development. In particular, they should be capable of selectively binding to designated targets and modulating their functions in physiological context.[65,100–102] Although many rigorous requirements are imposed on their development, many successful chemical probes for a number of kinase targets have been developed over the years.[103–107]

1.3.2 Promiscuity and Polypharmacology

Increasing volumes of compound activity data continue to provide an immense support for large-scale analyses of compound promiscuity.[60,108] Among these, multi-kinase inhibitors receive most attention, given their potential value for different therapeutic areas, primarily oncology.[109–111] For these inhibitors, multi-kinase activity forms the basis of polypharmacology. Polypharmacol-

ogy is an emerging paradigm in drug discovery according to which compounds elicit their therapeutic effects through multi-target interactions.[84–86] Compound promiscuity forms the basis for polypharmacology. It explores specific compound binding to multiple targets.[87] However, compound promiscuity may also originate from assay artifacts or nonspecific interactions.[112–115] In that case, compounds designated as kinase inhibitors would be discarded from further consideration, or handled with care.

At least a third of FDA-approved kinase inhibitors are known to bind multiple kinases.[48] For example, the orally administered multi-kinase inhibitor sunitinib is one of the first cancer drugs approved for two therapeutic indications simultaneously: imatinib-resistant gastrointestinal stromal tumor and renal cell carcinoma.[116,117] This multi-kinase inhibitor targets members of the protein-tyrosine kinase group, in particular VEGFR1, VEGFR2, fetal liver tyrosine kinase receptor 3 (FLT3), mast/stem cell growth factor receptor (KIT), platelet-derived growth factor receptor α (PDGFRα) and PDGFRβ in both biochemical and cell-based assays. The simultaneous inhibition of these targets reduces the tumor vascularization and initiates apoptosis of cancer cells, ultimately resulting in tumor shrinkage.[117] In 2009, numerous regulatory administrations worldwide approved pazopanib for advanced soft tissue sarcoma and metastatic renal cell carcinoma.[118] Similarly to sunitinib, pazopanib is a multi-receptor protein-tyrosine kinase inhibitor of VEGFR1, VEGFR2, VEGFR3, PDGFRα, PDGFRβ, and c-Kit.[119] Hence, pazopanib shows antiangiogenic and antitumor effects. These and many other examples such as CHIR-258[120] and MK-2461[121] support the purpose of multi-kinase inhibitors in a plethora of oncological indications.

In addition to experimental profiling studies, promiscuity can be estimated computationally through systematic data mining of large repositories of compound activity data.[122,123] In this case, data integrity and varying confidence levels need to be carefully considered to provide reliable estimates of compound promiscuity levels.[124] Compound promiscuity can be quantified using promiscuity degrees (PDs). This simple measure collects the number of targets against which a compound is active.[125] Extensive analysis of high-confidence compound activity data from ChEMBL suggested that bioactive compounds inhibited on average one or two targets, whereas the average for the most promiscuous com-

pounds was two to seven targets of the same protein family.[122,126] Similar data extraction criteria were applied to kinase inhibitor data which suggested that 76% of the publicly available kinase inhibitors were annotated with a single kinase.[127] When confidence criteria were iteratively relaxed, no notable increase in promiscuity of kinase inhibitors was detected.[111] This was in contrast to a general assumption that ATP site-directed kinase inhibitors tend to be promiscuous. Computational studies of kinase inhibitor promiscuity are often questioned due to data sparseness, because all compounds were not tested against all kinases.[128] However, given the consistency of promiscuity results that originate from large data sets, these observations should be statistically meaningful and cannot be simply attributable to data incompleteness.

There are several computational approaches used to study kinase inhibitor promiscuity on a molecular level. For example, structurally related compounds with alternating levels of promiscuity, known as promiscuity cliffs,[129–132] can be studied for structural features distinguishing between promiscuity and selectivity.[133] Additionally, network representations of chemically similar compounds can be used to establish novel target relationships that infer compound promiscuity.[134–136] Moreover, data coming from X-ray crystallography can be used to study binding site similarities of targets that contain promiscuous compounds,[137,138] as well as binding interactions underlining promiscuity.[139,140] Evidently, a multitude of computational approaches for promiscuity analysis exist for which kinases are a representative target family to explore.

1.4 Structural Similarity

1.4.1 Fingerprint Similarity

Analysis of structure-property relationships is based on the evaluation of compound similarity. For sets of structurally similar compounds, property relationships (e.g., activity, promiscuity) between compounds need to be estimated. Similarity indices were developed to quantify chemical similarity on the basis of bit string representations (e.g., fingerprints). Although many similarity indices can be employed for this purpose, the Tanimoto coefficient (or Jaccard index) is the most commonly used.[141] It is calculated as the ratio between intersection

and union of the two sets of patterns.[142] The obtained ratio between 0 (not similar) and 1 (identical fingerprints) is a measure of similarity for a pair of compounds. If compound A has a chemical patterns, and compound B has b chemical patterns, with value c representing the number of shared chemical patterns, the Tanimoto coefficient (Tc) is calculated using the following equation:

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

For some chemical descriptor-based approaches, similarity indices require the definition of a threshold value to consider compounds as similar. The use of a threshold value is not always straightforward, given that it relies on fingerprint design and evaluated data set. On the other hand, substructure-based approaches establish direct structural relationships using a set of predefined rules (on the basis of molecular graphs).[143] Some of the methods employing substructure-based approaches will be discussed in the following.

1.4.2 Matched Molecular Pairs

Matched molecular pairs (MMPs) are defined as pairs of compounds that differ by a chemical change at a single site.[144,145] The common substructure to both compounds is termed key fragment or MMP core, while the exchanged fragments constitute a chemical transformation.[146] Following a number of computationally expensive MMP approaches, Hussain and Rea introduced an algorithm that was more efficient and reliable.[146] The algorithm consists of two steps: molecule fragmentation and generation of MMPs.

First, each molecule is fragmented along the non-ring single bonds. To keep the connectivity information, an attachment point is added to mark where each bond was broken. Fragmentation can occur simultaneously at one, two, or three bonds, thus generating single, double, or triple "cuts", respectively. Frequently considered single-cut fragments consist of a larger fragment that is used as a key fragment, and a smaller one that is used to define the chemical transformation. Following the systematic fragmentation of all available compounds, compound pairs that present MMPs are identified. As several possible key fragments may exist, only the largest one is retained whereas the others are removed.

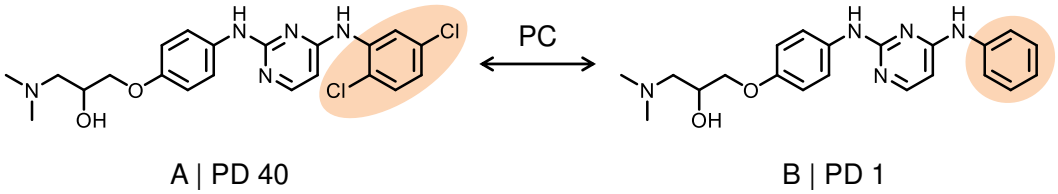


Figure 1.6: Matched molecular pair as promiscuity cliff. Shown are two compounds (A and B) forming a matched molecular pair (MMP) that conforms to promiscuity cliff (PC) requirements. The promiscuity degree (PD) is given for each compound. Structural modifications are highlighted in orange.

Transformation size-restricted MMPs are most commonly used to define compound relationships of chemically relevant analogs.[147] The original algorithm describes exhaustive fragmentation of non-ring single bonds. However, other fragmentation rules such as REtrosynthetic Combinatorial Analysis Procedure (RECAP)[148] rules can be applied to generate RECAP-MMPs[149] that resemble chemically more meaningful transformations.

The MMP concept provides a basis for the definition of activity cliffs (ACs)[150] and promiscuity cliffs (PCs).[129–132] ACs are defined as pairs of structurally similar compounds (e.g., MMPs) with large differences in activity for a particular target.[150] An activity ratio of 100-fold is usually taken as threshold for ACs. In addition, PCs are defined as pairs of structural analogs (usually MMPs) with large differences in promiscuity (ΔPD).[129–132] The ΔPD is defined as the difference of the PD values of two compounds. A number of different ΔPD values have been defined for different purposes so far.[129–132] PCs formed by inhibitors of human kinome are of particular relevance for the study of structure-promiscuity relationships.[133]

Moreover, highly promiscuous compounds can be used to infer new target hypotheses for sets of close structural analogs. Clearly, the application domain of kinase inhibitor PCs is multifaceted and will likely expand with growing compound activity data. An exemplary MMP, conforming to PC requirements, is shown on **Figure** 1.6.

1.4.3 Scaffolds and Compound Cores

Scaffolds are commonly referred to as core structures or structural backbones of molecules. Two or more molecules that contain the same scaffold also share a common substructure. The scaffold concept is widely used to cluster bioactive compounds and relate their property relationships.[151] However, the definition of a scaffold can be viewed very differently in both medicinal chemistry and chemoinformatics. In medicinal chemistry, the scaffold definition relies on the subjective perception of a chemist and structural context of analyzed analog series to which synthetic rules can be applied. In chemoinformatics, the extraction of scaffolds needs to be algorithmically efficient and generally applicable to large data sets. Computational approaches are continuously revising and introducing new scaffold definitions in order to satisfy the needs of medicinal chemistry.[152]

The Bemis-Murcko (BM) scaffold is most widely applied scaffold definition in chemoinformatics applications.[153] BM scaffolds are generated by the removal of all non-ring R-groups from compounds while retaining ring structures and linkers between them. Accordingly, BM scaffolds are obtained for molecules containing ring systems, which commonly applies to bioactive compounds. The combination of rings and retained linker fragments is also known as a "framework". BM scaffolds can further be simplified by disregarding atom type and bond information. It is usually done by transforming framework atoms to carbons and setting all bond orders to one.[154] This representation, called cyclic skeleton (CSK), further abstracts the chemical information while preserving the molecular topology. The representative BM scaffold and CSK derived from a compound set are shown in **Figure** 1.7.

Although highly popular, the concept of BM scaffolds comes with certain limitations. In medicinal chemistry, ring structures are also used as substituents during the optimization process and are not considered part of the modified scaffold. However, the BM concept treats ring structures as part of the framework, where an introduction of a new ring to the original molecule yields a new BM scaffold.[155] This modification would classify the compounds as dissimilar due to the structural differences between generated BM scaffolds. On the other hand, two molecules might still share the same BM scaffold while containing substituents of different sizes and complexity. To overcome these shortcomings

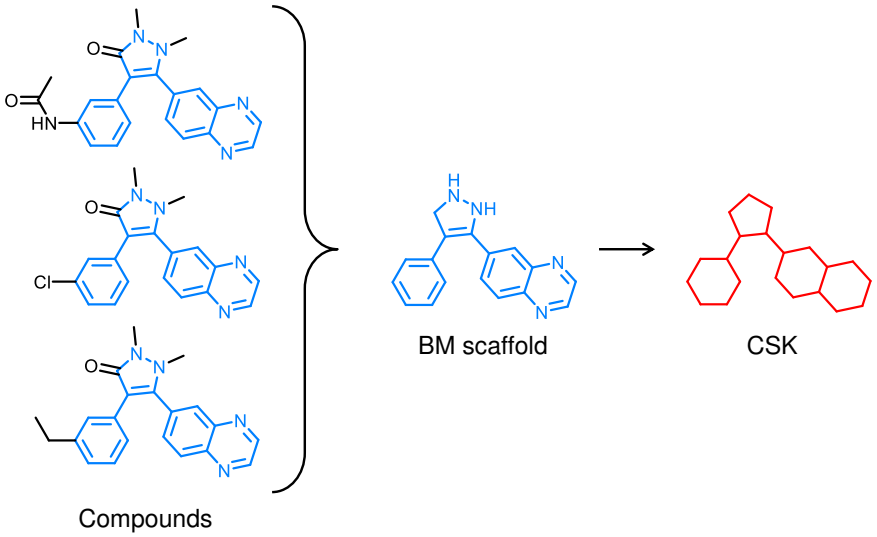


Figure 1.7: Hierarchy of chemical representations. Three exemplary compounds and their chemical abstractions (BM scaffold and CSK) are shown. The extracted BM scaffold (blue) is further simplified to obtain a CSK (red).

several new definitions of MMP-based scaffolds (cores) and analog series have been introduced in recent years.[156,157]

Among recent developments is the compound-core relationship (CCR) method, which systematically identifies analog series having a common core structure.[157] The CCR method relies on three sequential steps: generation of cores, exploration of CCRs, and identification of analog series. For each compound in a data set, all possible combinations of one to five bonds are systematically cleaved using RECAP rules. Thus, each combination of RECAP rules corresponding to the elimination of single or multiple bonds results in a potential core. Moreover, cores and substituents need to meet predefined size ratio. Cores that are identical except for the location of their substitution sites are not distinguished. Then, compounds are assigned to each of their derived cores, where the compound itself is also considered to be a core with no substitutions. At the end, an analog series is formed if two compounds share the same core. On the basis of CCRs, a compound can belong to different analog series. Thus, compounds are uniquely assigned to a single analog series in a disambiguation step, preferentially assigning compounds to larger series or in case of a tie to the larger core and series with fewer substitution sites.[157] This

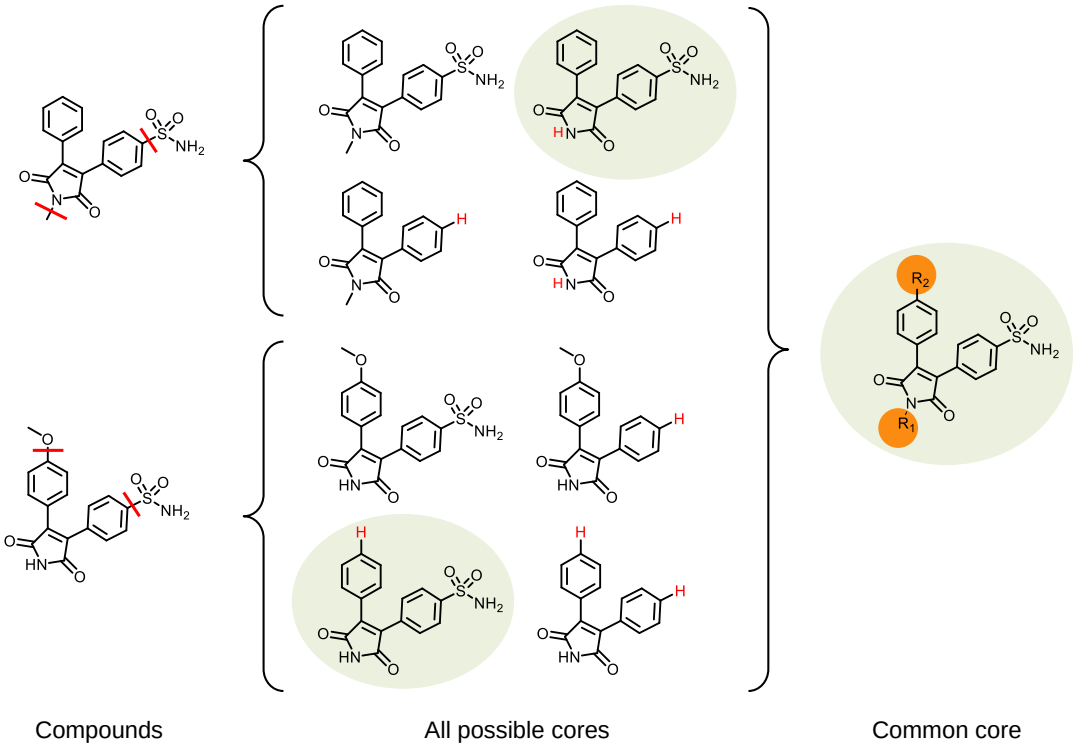


Figure 1.8: Concept of compound-core relationship method. Identification of analog series using the CCR method is schematically represented. For two compounds, all possible cores are generated following bond fragmentation (red line). For the two analogs, the largest common core is identified (encircled in green) and isolated. In the common core, substitution sites are encircled in orange. The figure is adapted from Naveja *et al.*[157]

methodology is conceptually simple, yet attractive for organic and medicinal chemists as it allows the organization of analog series in R-group tables. This further simplifies their use in SAR analyses.[158] **Figure** 1.8 shows a schematic representation of CCR method.

1.5 Machine Learning

Machine learning methods are used to develop computational models that are able to learn patterns or rules from provided data to classify objects (predict class labels). Machine learning has become widely popular in chemoinformatics and drug discovery applications for a number of classification and regression tasks.[159] In drug discovery applications, they are often used to predict novel active compounds and compound properties, and to perform compound

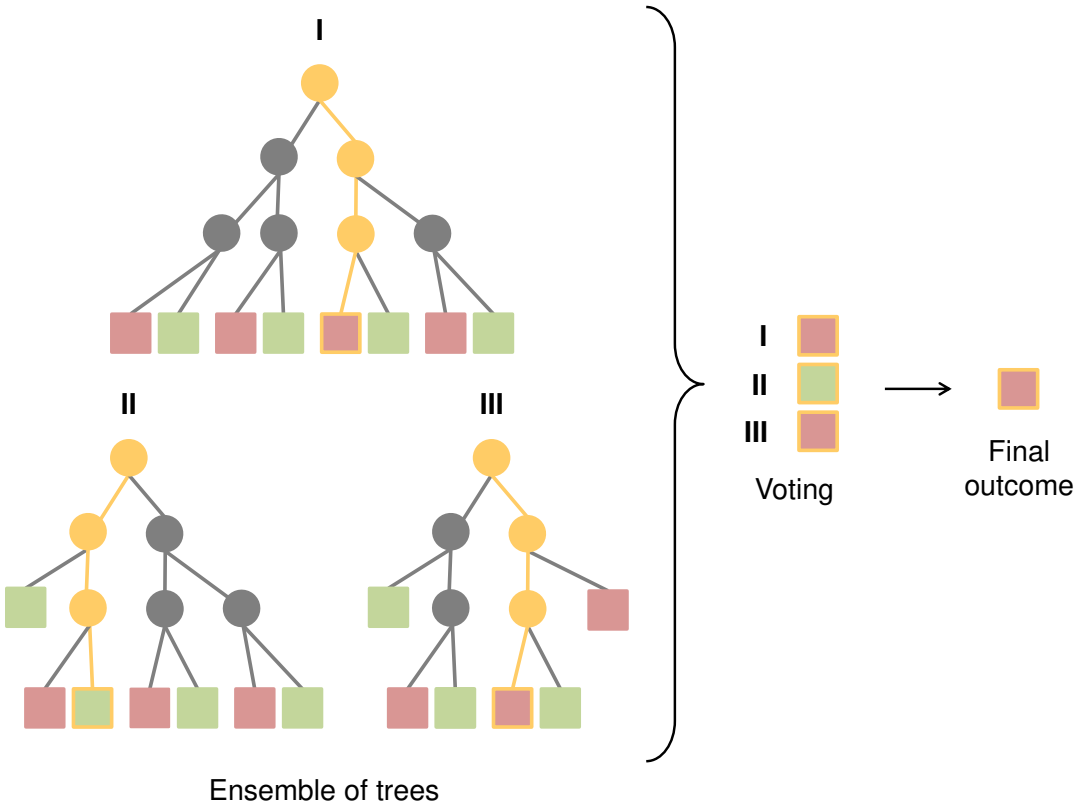


Figure 1.9: Random forest. An ensemble of three decision trees is shown forming an RF. The outcome of each leaf node is colored either red or green, depending on the predicted class. Prediction of a single compound instance is tracked in each tree with a yellow path. Two out of three trees predict a red class for the observed instance. Therefore, the majority vote predicts the final outcome to be "red". The figure is adapted from Mitchell.[159]

classification and ligand-based virtual screening.[160] In chemoinformatics, machine learning models typically use compound fingerprints, such as MACCS[78] or ECFP4.[79] A number of machine learning methods exist, of which random forest (RF),[161] support vector machine (SVM),[162] and deep neural network (DNN)[163] are widely applied and considered to be state-of-the-art methods. These will be discussed in the following regarding their application in classification tasks.

1.5.1 Random Forests

RF is a machine learning technique which uses an ensemble or "forest" of decision trees.[164] Each individual tree is trained on a bootstrapped sample of data using a stochastic recursive partitioning method. For each tree, a random subset of features is considered for node partitioning to avoid generation of

correlated trees promoting feature dominance. The partitioning aims to increase the homogeneity of groups in each terminal node. At the end, a number of different decision trees built from bootstrapped data samples is obtained. This model is then used to predict test data by means of consensus vote. For example, in binary classification problems where compounds can be predicted either as "active" or "inactive", if the majority of decision trees predicts a compound to be "active" then the final prediction classifies it as "active". **Figure** 1.9 depicts an example of an RF with an ensemble of three decision trees and an exemplary majority vote for one compound instance. RF calculations require relatively low computational power even when large number of trees and fingerprints of different size and complexity are included. RF has been used as a method of choice in many chemoinformatics-driven machine learning applications.[20,165,166]

1.5.2 Support Vector Machines

SVM aims to derive a separating hyperplane H that maximizes the distance, so called margin, between the objects with different class labels.[162] During the learning process, SVM projects the training data of different class labels into a high-dimensional space. If the data is linearly separable in this space, the number of hyperplanes that correctly classifies the data is infinite. However, only a unique H that optimizes the margin between the closest points of each label (support vectors) is chosen. An example of an SVM model is shown in **Figure** 1.10. The labels of the test data are predicted on basis of which side of H the instance is found. The H is defined by the normal vector w and bias b as follows:

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

where $\langle ., . \rangle$ is a scalar product. The following conditions must be satisfied to ensure the correct classification of all training instances:

$$y_{\mathrm{i}}\big(\langle x_{\mathrm{i}}, w \rangle + b\big) \geq 1 \quad \forall i$$

where x_{i} are the training instances and $y_{\mathrm{i}} \in \{-1, 1\}$ is the class label (negative or positive) for each training instance. The distance between the support vectors and H is given by $\frac{1}{\|w\|}$, which an optimal hyperplane maximizes. When training

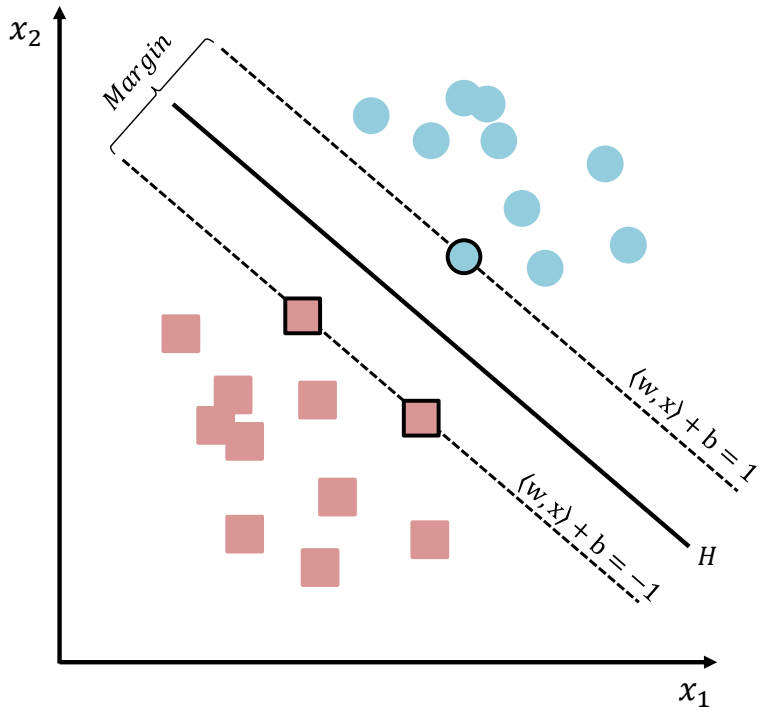


Figure 1.10: Support vector machine. A linear separation between positive (blue circle) and negative (red square) classes is shown. Here, a separating hyperplane H with a maximum margin is shown with a solid black line. The parallel lines defined by support vectors (instances with black outline) are shown with a dashed black line.

data cannot be linearly separated, direct minimization of $\|w\|$ is not possible and thus "kernel trick" is applied, as discussed below. In this case and in order to improve the generalization of models, slack variables ξ_i are introduced to permit errors. This allows some of the training instances to fall within the margin, or on the wrong side of the hyperplane.[167] However, with the increase of slack variable values the potential for training instances to be misclassified increases. Thus, the misclassification of data is penalized by introducing the cost or regularization parameter C. This optimization problem can be expressed using Lagrangian multipliers λ_i,[168] resulting in a representation of the H as a linear combination of training vectors:

$$w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

A nonzero value of Lagrangian multipliers is obtained only for the training examples that are misclassified or fall onto the margin. The latter case of training examples are called support vectors. Finally, a test instance can be classified as "positive" or "negative", depending on the side of H the instance is projected. Moreover, the test data can be ranked using the real value.[169]

The scalar product $\langle ., . \rangle$ requires a vector representation of the instances. Often given vector representations are not suitable for linear separation of the data. A strength of SVMs is that the scalar function can be replaced by a kernel function $K\langle ., . \rangle$, fulfilling certain criteria. Kernel function can be used to replace calculation of the scalar product in an implicit high-dimensional space that improves separability. This technique is known as the "kernel trick".[170] One of the most widely used kernel functions for fingerprint representations is the Tanimoto kernel:

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

where u and v present two compound fingerprints.[171] As one of the most popular machine learning methods in chemoinformatics, SVMs have been used in many applications including binary classification tasks,[172] multi-target predictions,[173] and compound ranking.[174]

1.5.3 Deep Neural Networks

DNNs are increasingly used in drug discovery as exemplified by a number of applications such as bioactivity prediction, *de novo* design, biological image analysis, and synthesis prediction.[175] DNNs are a class of machine learning methods that use artificial neural networks (ANNs) built with multilayered nonlinear processing units to learn provided data representations.[175] Each layer consists of a set of nodes (neurons) which are connected with nodes of the neighboring layers. The three basic layers are defined in DNN: the input layer, one or more hidden layers, and the output layer. Thus, the input variables are taken by the nodes of the input layer, transformed through the nodes of the hidden layer(s), and processed as the final output values of the output nodes. Each node of the hidden layer(s) and the output layer accepts the input from the

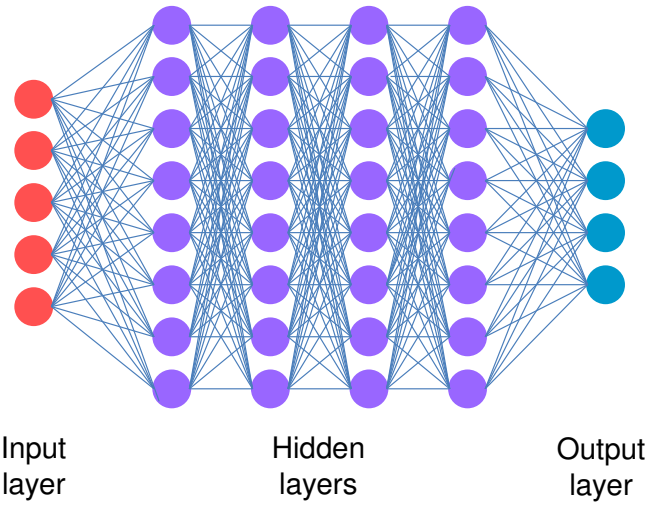


Figure 1.11: Deep neural network. A fully connected, multilayered DNN is schematically represented. The DNN consists of an input layer (red nodes), four hidden layers (purple nodes), and an output layer (blue nodes). Each node presents a single neuron.

previous layer and transforms it via an activation function (usually nonlinear) to yield an output value. For a node i the output value Y_{i} is calculated as follows:

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

where g is generally a nonlinear function, W_{ij} is the weight of input node j on node i, and a_{j} refers to input variables. DNNs are trained by an iterative modification of weight values, so that the errors between predicted and true values are optimized.[176] They usually contain multiple hidden layers, with each layer comprising hundreds of nonlinear process units. With current computational power, DNNs are able to take a large number of input features and neurons in multilayered architectures to automatically extract features at different hierarchical levels.[177] An example of a DNN involving an input layer, several hidden layers, and an output layer is given in **Figure** 1.11.

1.6 Thesis Outline

This thesis consists of eight studies organized in individual chapters. Selectivity and promiscuity of kinase inhibitors are explored in detail to assess the potential of these compounds for kinase drug discovery. Moreover, structurally related kinase inhibitors with alternating levels of promiscuity are investigated. At the end, machine learning methods are employed to classify kinase inhibitors with different binding modes.

- *Chapter 2* systematically explores selectivity of multi-kinase inhibitors on the basis of data from medicinal chemistry sources. Therefore, compound-based kinase pairs with increasing phylogenetic distances are formed. Furthermore, kinase binding regions are evaluated for known selectivity determinants and their correlation with kinase inhibitor data is studied.

- In *Chapter 3*, cell-based data from a major profiling campaign is used to study selectivity of clinical kinase inhibitors. Approaches similar to those described in *Chapter 2* are applied. Moreover, new categories of selectivity profiles are defined for kinase pairs carrying inhibitors with diverse selectivity potential.

- *Chapter 4* assesses selectivity of clinical kinase inhibitors on the basis of data from medicinal chemistry sources. Different data confidence criteria are applied. In addition, different inhibitor subsets are assembled for the estimation of selectivity profiles. These include the most and the least selective inhibitors from the profiling experiment, type I and II inhibitors, and chemical probes.

- In *Chapter 5*, selectivity analysis of chemical probes is extended to explore publicly available chemical probes. Different confidence criteria are applied to calculate their promiscuity values. Moreover, potential for off-target activities is evaluated.

- *Chapter 6* describes large-scale promiscuity analysis of kinase inhibitors collected from several public sources. These inhibitors are used to extract PCs following a predefined set of rules. Network representations of PC relationships reveal many disjoint PC clusters. These clusters are further

explored by tracing linear PC sequences termed PC pathways which are used to interpret structure-promiscuity relationships.

- In *Chapter 7*, a computational method is developed that systematically extracts PC pathways from PC clusters. This computational approach facilitates identification and ranking of the most interesting PC pathways in large and complex clusters. Pathways containing promiscuity hubs are identified.

- *Chapter 8* investigates the data from studies reported in *Chapter 6* and *Chapter 7* and makes them publicly available. Promiscuity hub analysis is extended and high-priority hubs are defined.

- In *Chapter 9*, machine learning methods are implemented to distinguish between kinase inhibitors with different binding modes coming from X-ray crystallography.

Chapter 10 summarizes the major findings of the studies and discusses their relevance for kinase drug discovery.

Chapter 2

Exploring Selectivity of Multi-Kinase Inhibitors across the Human Kinome

Introduction

Many currently available inhibitors of the human kinome bind to the largely conserved ATP-binding site. Although allosteric mechanisms of inhibition should be the most selective, only a handful of these compounds are reported. Hence, the evaluation of kinase inhibitor selectivity still largely depends on the ATP site-directed representatives.

Experimental profiling studies of these inhibitors revealed compounds with varying selectivity patterns. In addition, the substantially conserved ATP-binding site contains sequence variations which might influence inhibitor selectivity.

In this study, we systematically analyzed selectivity of multi-kinase inhibitors on the basis of high-confidence data coming from ChEMBL. Compound-based kinase pairs were formed and classified into categories with increasing phylogenetic distances. For classified kinase pairs, pair- and

compound-based selectivity profiles were generated and selectivity trends were further evaluated.

ACS OMEGA

Article

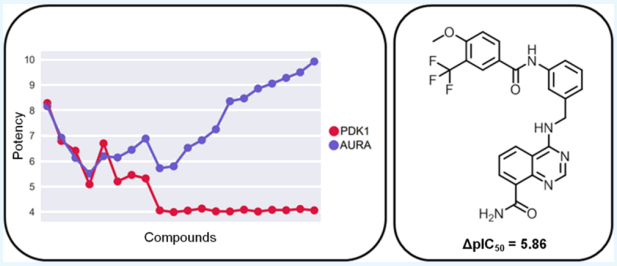
Cite This: ACS Omega 2018, 3, 1147−1153

Exploring Selectivity of Multikinase Inhibitors across the Human Kinome

Filip Miljković and Jürgen Bajorath*

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

ABSTRACT: Selectivity of kinase inhibitors, or the lack thereof, continues to be an intensely debated topic in drug discovery research. Especially, type I inhibitors, which represent most of the currently available kinase inhibitors, are often thought to lack selectivity because they target the largely conserved adenosine triphosphate-binding site in kinases. Herein, we present a large-scale analysis of potential selectivity among multikinase inhibitors, covering 141 human kinases and more than 10 000 qualifying compounds. By design, the analysis was focused on type I inhibitors and carried out at the level of systematically generated kinase pairs sharing inhibitors. Kinase pair category- and compound-based selectivity profiles identified in part highly selective inhibitors for many kinases. Sets of inhibitors associated with kinase pairs frequently contained nonselective as well as increasingly selective compounds. Selectivity of inhibitors did not result from gatekeeper residues settings or phylogenetic distance of kinases. Rather, it was most likely attributable to subtle differences between binding regions in kinases. Taken together, the results of our study reveal that many multikinase inhibitors are more selective than one might assume.



1. INTRODUCTION

Inhibitors of human kinases are among the most intensely investigated compounds in drug development.[1−5] Most currently available kinase inhibitors target the adenosine triphosphate (ATP) (cofactor)-binding site that is largely conserved across the human kinome.[6,7] Accordingly, ATP-site-directed kinase inhibitors are expected to be promiscuous and lack selectivity, as indicated by a number of kinase inhibitor profiling studies.[8−11] Therefore, attempts have been made to discover other types of inhibitors that target different regions in kinases and act by different mechanisms.[12,13] ATP-site-directed (type I) inhibitors bind to the so-called "DFG-in" conformation of the activation loop near the catalytic site, i.e., the active form of the kinase. In addition, type II inhibitors bind to the inactive "DFG-out" conformation of the activation segment, occupying pockets adjacent to the ATP-binding site that are less conserved.[13] Thus, type II inhibitors are expected to be more selective than type I inhibitors. Furthermore, there are type III and IV inhibitors that bind to regions outside the ATP-binding site and act by allosteric mechanisms.[13] Only a limited number of allosteric kinase inhibitors has been reported thus far, but these types of inhibitors might indeed be most selective.[14−16]

However, the often assumed lack of selectivity of type I inhibitors continues to be debated[17] and expected selectivity differences between type I and II inhibitors are subject to further investigation. For example, profiling experiments using type II inhibitors have shown that these inhibitors are often active against many kinases.[13] Furthermore, although subsets of highly promiscuous type I inhibitors have been identified[18] and promiscuity of kinase inhibitors has become a hallmark for successful cancer treatment,[2] there is also evidence for selectivity of ATP-site-directed inhibitors. For example, although a number of kinase inhibitor profiling experiments have indicated a lack of selectivity of type I inhibitors,[8−11] others have revealed selectivity patterns.[19,20] In addition, type I inhibitors are also capable of acting by different mechanisms.[21] Furthermore, on the basis of high-confidence activity data, 76% of publicly available kinase inhibitors were found to be annotated with a single kinase.[22] When activity data confidence criteria were iteratively lowered, no notable increase in kinase inhibitor promiscuity was detected,[23] suggesting that promiscuity was not a general rule. Of course, it has long been known that the ATP-binding site in kinases has some sequence variation, in particular, at the "gatekeeper" position,[7] where the presence of smaller or larger residues differentiates between classes of type I inhibitors. However, whether or not the gatekeeper is the only factor responsible for inhibitor differentiation within the ATP-binding site is currently unknown. Other subtle differences might also play a role. Clearly, the issue of kinase inhibitor selectivity is still not fully explored.

Herein, we present a systematic analysis of selectivity among multikinase inhibitors on the basis of currently available activity data. Selectivity profiles were generated for sets of inhibitors shared by kinases. The profiles revealed significant potency

Received: December 8, 2017
Accepted: January 18, 2018
Published: January 26, 2018

variations of subsets of inhibitors and identified compounds with selectivity for given kinases over others.

2. MATERIALS AND METHODS

2.1. Compounds, Targets, and Activity Data. Inhibitors of human protein kinases were assembled from ChEMBL version 23.[24] Compounds with activity in assays detecting direct interactions (target relationship type "D") with human protein kinases at the highest confidence level (confidence score 9) were selected. As potency measurements, IC_{50} values were considered. The amount of available K_i values was too small for a meaningful statistical analysis. If multiple IC_{50} values were available for a compound, the final potency annotation was calculated as the geometric mean of these values, provided all fell within the same order of magnitude (otherwise, the compound was disregarded). Approximate measurements associated with ">", "<", or "~" were not taken into account. On the basis of these criteria, 40 627 inhibitors with activity against 274 human kinases were obtained. From this compound pool, inhibitors were selected that were active against at least two kinases, yielding a final set of 10 367 inhibitors with activity against 266 human kinases. ChEMBL target identifiers of these kinases were mapped to UniProt,[25] and kinases were assigned to families and groups (of families) according to Manning et al.[6] and Miranda-Saavedra et al.[26]

2.2. Protein Kinase Pairs. The selected multikinase inhibitors were used to systematically form compound-based target pairs. Two kinases were paired if they shared at least 10 inhibitors. Given this constraint, a total of 596 pairs were obtained that included 141 kinases and 10 060 inhibitors. Kinase pairs were assigned to three different categories: same family, i.e., both kinases belonged to the same family (132 pairs); different families, i.e., both kinases belonged to different families within the same kinase group (262 pairs); and different groups, i.e., both kinases belonged to different groups (202 pairs). Kinases in pairs from the same family, different families, and different groups were increasingly distant (unrelated). For each pair, compound selectivity was assessed by calculating the logarithmic potency difference (ΔpIC_{50}) for each inhibitor.

2.3. Gatekeeper Residue and Binding-Site Comparison. The kinase−ligand interaction fingerprints and structures (KLIFS)[27,28] database defines a kinase "binding pocket" for type I−IV inhibitors as a set of 85 discontinuous residues. This sequence segment, which contains the gatekeeper residue at position 45, can be extracted for human kinases from KLIFS on the basis of UniProt identifiers using the 3D-e-Chem-VM engine.[29] For kinase pairs, gatekeeper residues were compared and sequence identity over the 85-residue segment was calculated as an indicator of binding-site resemblance. Phylogenetic trees of the human kinome were drawn with Kinome Render.[30]

3. RESULTS AND DISCUSSION

3.1. Qualifying Kinase Inhibitors. Figure 1 shows the distribution of inhibitors over all 596 pairs of kinases sharing at least 10 compounds, yielding a median value of 18 inhibitors per pair. Hence, kinase pairs were associated with sufficient numbers of inhibitors for a systematic assessment of selectivity profiles. The pairs involved 141 kinases distributed across the human kinome and 10 060 multikinase inhibitors from ChEMBL.

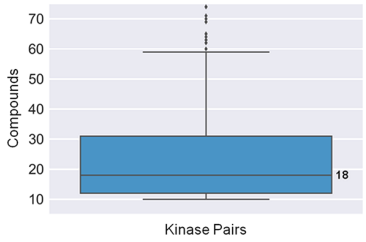


Figure 1. Distribution of compounds over kinase pairs. The boxplot reports the distribution of inhibitors over kinase pairs, yielding a median value of 18 inhibitors per pair. Boxplots report the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), largest value (top line), and outliers (points below the smallest or above the largest value).

Mapping of type II kinase inhibitor signature fragments[13] indicated that less than 1% of kinase inhibitors available in ChEMBL were type II inhibitors.[18] Thus, although it is not exactly known how many type II, or rare type III/IV, inhibitors are currently available in ChEMBL, for all practical considerations, our analysis was focused on type I multikinase inhibitors.

3.2. Global Selectivity. Potency differences of inhibitors against kinases forming pairs were calculated as a measure of selectivity. The larger the potency difference was, the more selective an inhibitor was for one kinase over the other. Initially, the global potency difference distribution was determined. Figure 2 (left) shows that average potency differences for all inhibitors associated with a pair were rather small, with a median ΔpIC_{50} value of 0.64 (i.e., well within 1 order of magnitude). At a first glance, this was what one might expect for largely nonselective inhibitors. However, the picture changed when only the inhibitor with largest potency difference from each pair was considered, as also shown in Figure 2 (right). In this case, the distribution yielded a median ΔpIC_{50} of 2.37, a difference of more than 2 orders of magnitude (100-fold), and a third quartile difference of 3 orders of magnitude. Thus, for individual inhibitors, a global tendency of selectivity emerged. Systematically enumerating pairs of kinases sharing inhibitors ensured that all possible selectivity relationships were taken into account. The union of pairwise relationships was expected to reveal general selectivity trends, if they existed.

The global selectivity tendency was also observed at the level of different kinase pair categories. Figure 3a shows the distribution of potency differences for the three pair categories in different formats. In all three cases, the median difference for all compounds fell within the same order of magnitude and exceeded 2 orders of magnitude for the most selective compounds.

3.3. Pair Category-Based Selectivity Profiles. The global selectivity tendency was further corroborated by pair category-based selectivity profiles shown in Figure 3b. These profiles were generated by recording the largest inhibitor potency difference for each pair and ordering the pairs by increasing ΔpIC_{50} values. In each case, more than half of the kinase pairs had one or more inhibitors with a potency difference exceeding 2 orders of magnitude. Furthermore, in each case, potency differences exceeding 4 or even 5 orders of magnitude were observed for multiple pairs. For kinases from different groups, 55% of the pairs had inhibitor(s) with potency

DOI: 10.1021/acsomega.7b01960
ACS Omega 2018, 3, 1147−1153

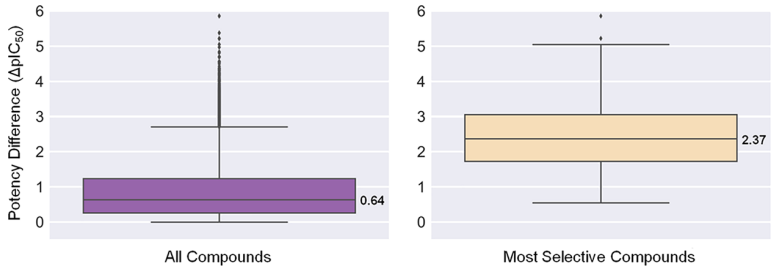


Figure 2. Compound potency differences for kinase pairs. Boxplots report the distribution of potency differences of inhibitors for paired kinases as the mean potency difference of all inhibitors (left) or the largest potency difference (most selective compounds; right). The distributions yield ΔpIC_{50} median values of 0.64 (left) and 2.37 (right).

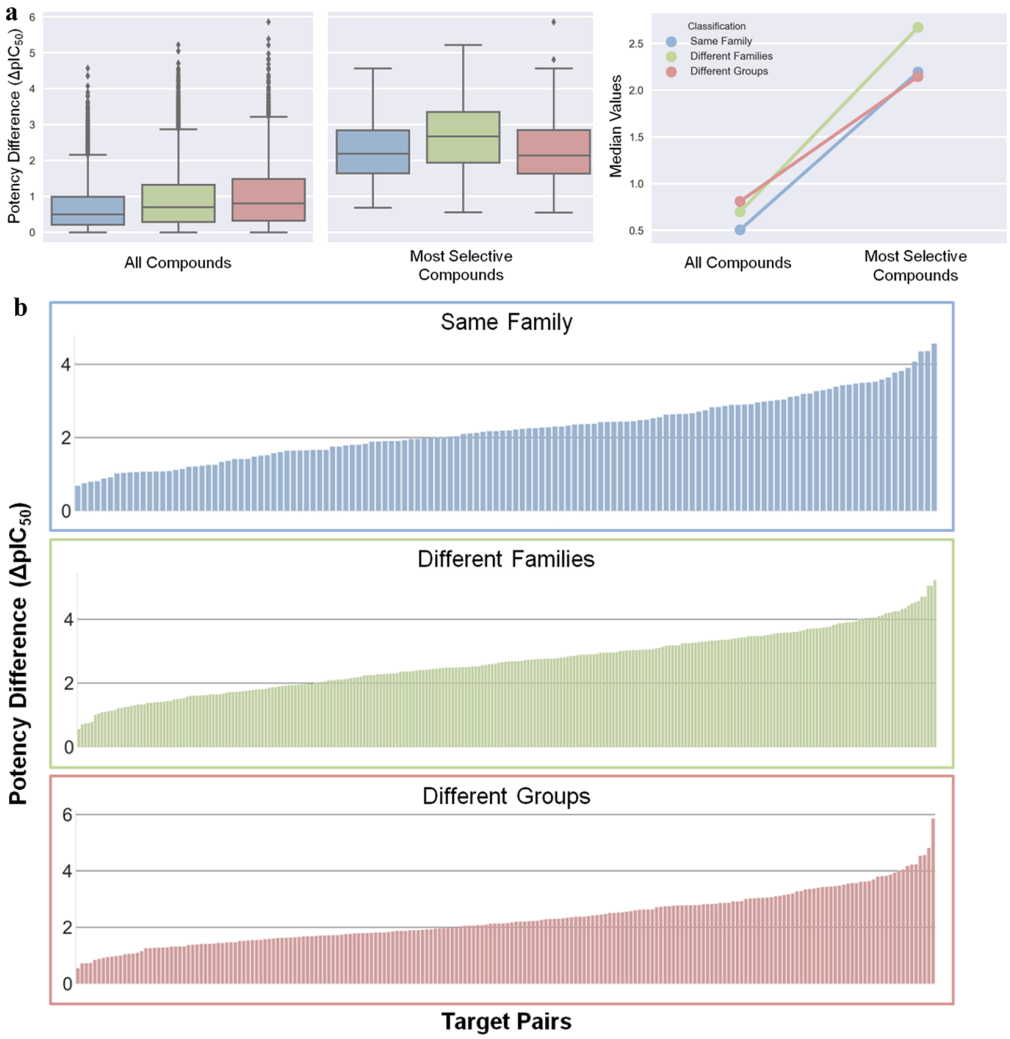


Figure 3. Compound potency differences for pair categories. (a) Distributions of ΔpIC_{50} values (left) for all versus the most selective inhibitors according to Figure 2 for kinase pairs from the same family (blue, 132 pairs), different families (green, 262 pairs), and different groups (red, 202 pairs). In addition, a comparison of ΔpIC_{50} median values is shown (right). (b) Selectivity profiles for the three pair categories that record the potency differences of the most selective inhibitor for each pair (in the order of increasing potency differences from left to right).

differences of more than 2 orders of magnitude and 22% of more than 3 orders of magnitude.

3.4. Compound-Based Selectivity Profiles. Detailed views of inhibitor selectivity were provided by compound-based selectivity profiles. Figure 4 (left) shows exemplary profiles for kinase pairs from the same family, different families, and different groups. Kinases from each pair had the same gatekeeper residue. In these profiles, potency values of all inhibitors are compared for kinases of a pair and inhibitors are ordered according to increasing potency differences. In addition, Figure 4 shows the least and most selective inhibitor for each pair (middle) and the location of paired kinases on a phylogenetic tree representing the human kinome (right). For

DOI: 10.1021/acsomega.7b01960
ACS Omega 2018, 3, 1147−1153

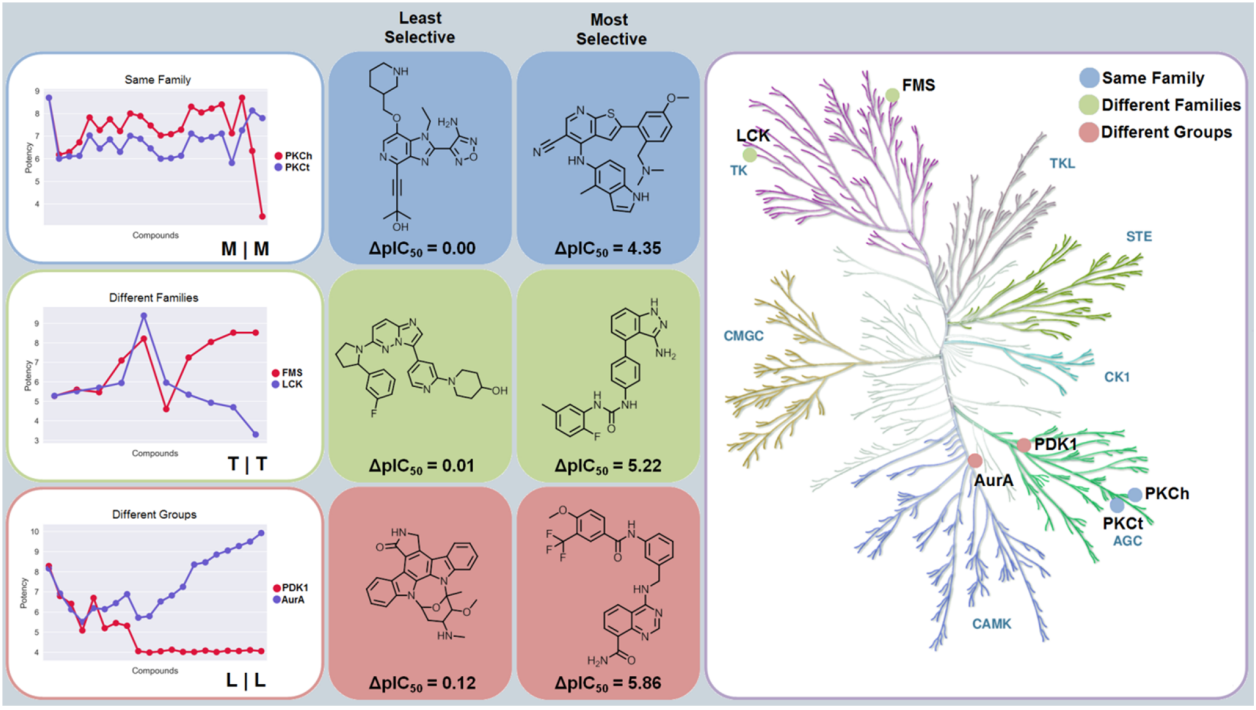


Figure 4. Compound-based selectivity profiles. Left: exemplary compound selectivity profiles for kinase pairs belonging to different categories. For each inhibitor, the potency against the two kinases is compared. From the left to the right, inhibitors are ordered according to increasing potency differences. On the lower right of each graph, gatekeeper residues of the kinase pair are reported (e.g., "M|M"). Middle: comparison of the least and most selective inhibitors for each pair. Right: kinases forming each pair are mapped onto a phylogenetic tree of the human kinome to illustrate their category relationships.

each pair, the most selective inhibitor displayed a potency difference of more than 4 or 5 orders of magnitude.

The selectivity profiles revealed in part striking differences in relative potencies between inhibitors. Compounds shared by the closely related protein kinase C eta type (PKCh) and protein kinase C theta type (PKCt) were generally slightly more potent against PKCh, preserving relative potency differences. However, two notable exceptions were detected, where potency against PKCh decreased sharply. In one of these cases, the inhibitor was essentially inactive against PKCh but retained high potency against PKCt, resulting in high selectivity for PKCt. The profile for macrophage colony-stimulating factor 1 receptor kinase (FMS) and tyrosine-protein kinase Lck (LCK) contained six inhibitors with comparable potency and four others with increasing potency differences and selectivity for FMS over LCK. Moreover, for the distantly related 3-phosphoinositide-dependent protein kinase 1 (PDK1) and aurora kinase A (AurA), there were five inhibitors with the same potency against both kinases, three with relatively small potency differences, and 12 others that were essentially inactive against PDK1 but increasingly potent against AurA, yielding a subset of selective AurA inhibitors. The most selective compound had a potency difference of nearly 6 orders of magnitude. Many other profiles revealing similar selectivity relationships were obtained. Thus, many inhibitors shared by pairs of 141 human kinases were highly selective, a rather unexpected finding.

3.5. Comparison of Gatekeeper Residues, Binding Regions, and Compound Selectivity. In light of these findings, we further investigated whether there might be straightforward explanations for the observed selectivity trends. Therefore, for all kinase pairs, combinations of gatekeeper residues were determined. For each gatekeeper combination, the number of pairs associated with inhibitor(s) having a ΔpIC_{50} of at least 2 orders of magnitude (selectivity criterion) was identified and compared to the number of pairs not meeting this selectivity criterion. The results are shown in Figure 5a. For most gatekeeper combinations, including conserved and different residues, more pairs with selective than nonselective inhibitors were available. Hence, conservation of gatekeeper residues did not preclude compound selectivity, as also illustrated in Figure 4, and for all gatekeeper combinations represented by multiple kinase pairs, selective inhibitors were available.

Furthermore, binding pocket similarity was calculated for all kinase pairs with selective inhibitors and others, as shown in Figure 5b. As expected, the similarity of binding regions decreased with increasing phylogenetic distances of paired kinases. However, pairs with selective and nonselective inhibitors were widely distributed over the entire similarity range, including all three pair categories. Hence, there was no detectable correlation between similarities of binding regions and the presence or absence of selective inhibitors. As shown in Figure 5b, even kinases with highly similar binding regions shared inhibitors that were selective. In addition, for each category, the percentage of kinase pairs for which selective inhibitors were available is provided. More than half of the kinase pairs in each category had selective inhibitors. However, there was no detectable correlation between the frequency of pairs with selected inhibitors and phylogenetic distance. Taken together, these findings indicated that rather subtle structural

DOI: 10.1021/acsomega.7b01960
ACS Omega 2018, 3, 1147−1153

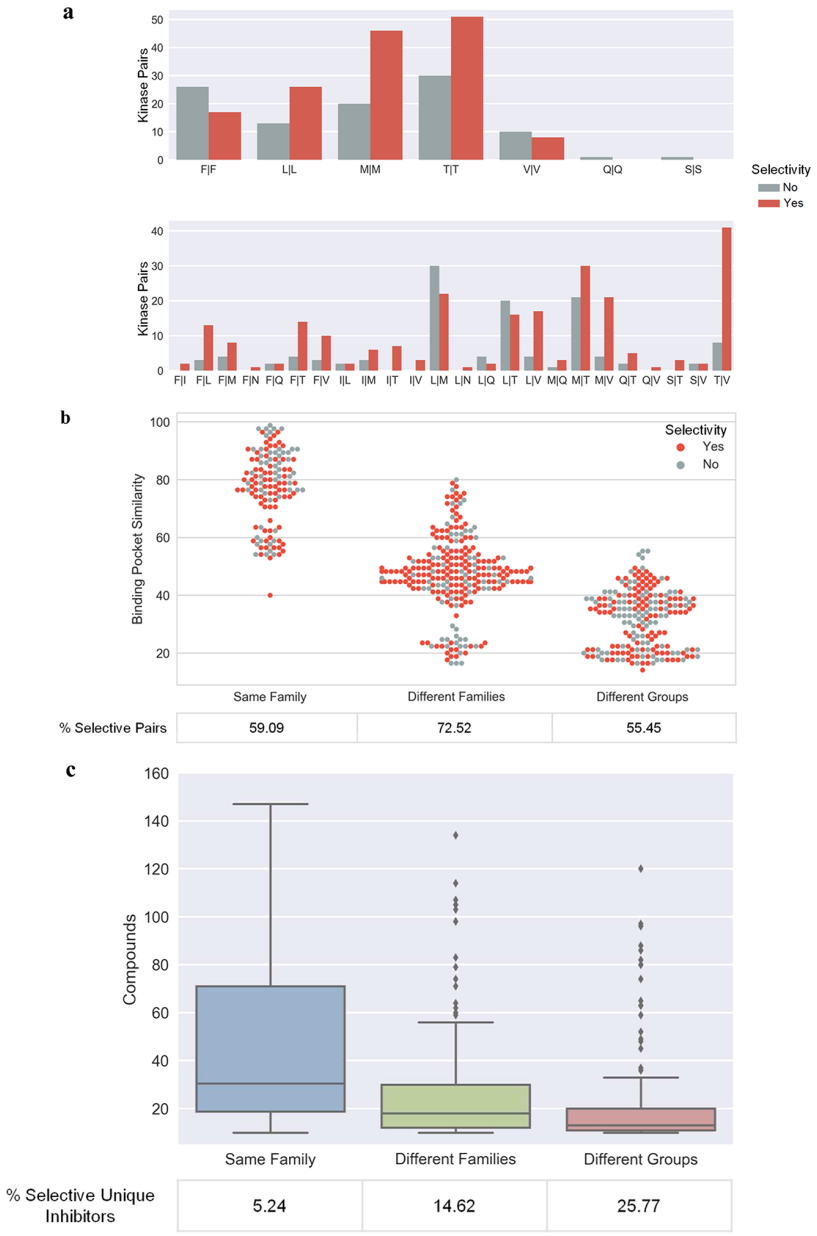


Figure 5. Gatekeeper residues, binding pocket similarity, and compound selectivity. (a) Histograms compare the number of kinase target pairs for each observed combination of gatekeeper residues (top, conserved residues; bottom, different residues), for which one or more selective (red) or no selective (gray) inhibitors were available. As a selectivity criterion, a potency difference of at least 2 orders of magnitude ($\Delta pIC_{50} \geq 2$) was applied. (b) Swarm plot (i.e., a boxplot in which all individual data points are displayed) capturing distributions of binding pocket similarity (sequence identity over the 85-residue segment) of kinases in pairs belonging to different categories to the presence (red) or absence (gray) of selective inhibitors. Individual data points on the *X*-axis are centered on the not displayed boxplot whisker for each category and depart from the central position if additional points have the same binding pocket similarity value. The percentage of kinase pairs with selective inhibitors ("selective pairs") is given for each category. (c) Distribution of compounds over kinase pairs in different categories. In addition, the proportion of selective inhibitors is given.

and/or property differences between kinases were largely responsible for the selectivity of shared inhibitors.

Figure 5c shows the distribution of shared inhibitors over kinase pairs from different categories. The number of shared inhibitors decreased with increasing phylogenetic distance between kinases in pairs. For each category, the proportion of selective unique inhibitors was also calculated. As expected, the percentage of selective inhibitors increased with increasing phylogenetic distance, as also shown in Figure 5c.

4. CONCLUSIONS

In this study, we have analyzed potential selectivity of multikinase inhibitors on a large scale based on currently available compound activity data. Previous studies have focused on kinase inhibitor selectivity profiling to identify new chemical probes for orphan receptors or compounds active against still little explored therapeutically relevant kinases.[31,32] Our analysis was facilitated by systematically generating pairs of 141 qualifying human kinases with increasing phylogenetic

DOI: 10.1021/acsomega.7b01960
ACS Omega 2018, 3, 1147−1153

distances that shared 10 or more inhibitors, providing a new reference frame for selectivity analysis. Contrary to our initial expectations, pair category- and compound-based selectivity profiles introduced herein revealed the presence of subsets of in part highly selective inhibitors for the majority of kinase pairs, providing extensive kinase coverage. Because the analysis was based on a statistically significant sample of more than 10 000 multikinase inhibitors, the detected selectivity trends were sound. Some striking observations were made at the level of compound-based selectivity profiles. In many instances, sets of inhibitors associated with kinase pairs contained subsets of nonselective compounds and others that were increasingly selective. These observations were of particular interest because the analysis was intrinsically focused on type I kinase inhibitors, which are often (but not always) thought to lack selectivity. We have also shown that observed inhibitor selectivity was not attributable to well-known kinase features, such as gatekeeper constellations or phylogenetic distances. It follows that selectivity determinants in kinases are likely to result from subtle differences that are far from being obvious, which should provide ample opportunities for future research. Clearly, although much progress has been made in recent years in rationalizing kinase inhibition and underlying mechanisms of actions, especially at the structural level, the jury on kinase inhibitor selectivity and its possible molecular origins is still out there. To support further exploration of kinase inhibitor selectivity, our kinase pair and inhibitor data set is made freely available as an open access deposition.[33]

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-2699-306.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

The study was carried out and the manuscript was 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 authors thank A. Kooistra for additional information about 3D-e-Chem-VM, R. Kunimoto for support with kinome representations, and M. Vogt for helpful discussions.

REFERENCES

(1) Cohen, P. Protein Kinases-the Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309−315.

(2) Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130−137.

(3) *Kinase Drug Discovery*; Ward, R. A., Goldberg, F. W., Eds.; RSC: Cambridge, U.K., 2011.

(4) Simmons, D. L. Targeting Kinases: A New Approach to Treating Inflammatory Rheumatic Diseases. *Curr. Opin. Pharmacol.* **2013**, *13*, 426−434.

(5) Laufer, S.; Bajorath, J. New Frontiers in Kinases: Second Generation Inhibitors. *J. Med. Chem.* **2014**, *57*, 2167−2168.

(6) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.

(7) Noble, M. E.; Endicott, J. A.; Johnson, L. N. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science* **2004**, *303*, 1800−1805.

(8) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A Small Molecule-Kinase Interaction Map for Clinical Kinase Inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.

(9) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(10) Cheng, A. C.; John Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity Using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53*, 4502−4510.

(11) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the Kinome. *Nat. Chem. Biol.* **2011**, *7*, 200−202.

(12) Gavrin, L. K.; Saiah, E. Approaches to Discover Non-ATP Site Kinase Inhibitors. *Med. Chem. Commun.* **2013**, *4*, 41−51.

(13) Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9*, 1230−1241.

(14) Ohren, J. F.; Chen, H.; Pavlovsky, A.; Whitehead, C.; Zhang, E.; Kuffa, P.; Yan, C.; McConnell, P.; Spessard, C.; Banotai, C.; Mueller, W. T.; Delaney, A.; Omer, C.; Sebolt-Leopold, J.; Dudley, D. T.; Leung, I. K.; Flamme, C.; Warmus, J.; Kaufman, M.; Barrett, S.; Tecle, H.; Hasemann, C. A. Structures of Human MAP Kinase Kinase 1 (MEK1) and MEK2 Describe Novel Noncompetitive Kinase Inhibition. *Nat. Struct. Mol. Biol.* **2004**, *11*, 1192−1197.

(15) Adrián, F. J.; Ding, Q.; Sim, T.; Velentza, A.; Sloan, C.; Liu, Y.; Zhang, G.; Hur, W.; Ding, S.; Manley, P.; Mestan, J.; Fabbro, D.; Gray, N. S. Allosteric Inhibitors of Bcr-Abl-Dependent Cell Proliferation. *Nat. Chem. Biol.* **2006**, *2*, 95−102.

(16) Ashwell, M. A.; Lapierre, J. M.; Brassard, C.; Bresciano, K.; Bull, C.; Cornell-Kennon, S.; Eathiraj, S.; France, D. S.; Hall, T.; Hill, J.; Kelleher, E.; Khanapurkar, S.; Kizer, D.; Koerner, S.; Link, J.; Liu, Y.; Makhija, S.; Moussa, M.; Namdev, N.; Nguyen, K.; Nicewonger, R.; Palma, R.; Szwaya, J.; Tandon, M.; Uppalapati, U.; Vensel, D.; Volak, L. P.; Volckova, E.; Westlund, N.; Wu, H.; Yang, R. Y.; Chan, T. C. Discovery and Optimization of a Series of 3-(3-Phenyl-3H-imidazo-[4,5-b]pyridin-2-yl)pyridin-2-amines: Orally Bioavailable, Selective, and Potent ATP-Independent Akt Inhibitors. *J. Med. Chem.* **2012**, *55*, 5291−5310.

(17) Levitzki, A. Tyrosine Kinase Inhibitors: Views of Selectivity, Sensitivity, and Clinical Performance. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 161−185.

(18) Hu, Y.; Furtmann, N.; Bajorath, J. Current Compound Coverage of the Kinome. *J. Med. Chem.* **2015**, *58*, 30−40.

(19) Anastassiadis, 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.

(20) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046−1051.

(21) Müller, S.; Chaikuad, A.; Gray, N. S.; Knapp, S. The Ins and Outs of Selective Kinase Inhibitor Development. *Nat. Chem. Biol.* **2015**, *11*, 818−821.

(22) Dimova, D.; Bajorath, J. Assessing Scaffold Diversity of Kinase Inhibitors Using Alternative Scaffold Concepts and Estimating the Scaffold Hopping Potential for Different Kinases. *Molecules* **2017**, *22*, No. 730.

DOI: 10.1021/acsomega.7b01960
ACS Omega 2018, 3, 1147−1153

(23) Stumpfe, D.; Tinivella, A.; Rastelli, G.; Bajorath, J. Promiscuity of Inhibitors of Human Protein Kinases at Varying Data Confidence Levels and Test Frequencies. *RSC Adv.* **2017**, *7*, 41265−41271.
(24) 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.
(25) UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43*, D204−D212.
(26) Miranda-Saavedra, D.; Barton, G. J. Classification and Functional Annotation of Eukaryotic Protein Kinases. *Proteins* **2007**, *68*, 893−914.
(27) van Linden, O. P. J.; Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Knowledge-Based Structural Database To Navigate Kinase-Ligand Interaction Space. *J. Med. Chem.* **2014**, *57*, 249−277.
(28) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Res.* **2016**, *44*, D365−D371.
(29) McGuire, R.; Verhoeven, S.; Vass, M.; Vriend, G.; de Esch, I. J. P.; Lusher, S. J.; Leurs, R.; Ridder, L.; Kooistra, A. J.; Ritschel, T.; de Graaf, C. 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *J. Chem. Inf. Model.* **2017**, *57*, 115−121.
(30) Chartier, M.; Chénard, T.; Barker, J.; Najmanovich, R. Kinome Render: A Stand-Alone and Web-Accessible Tool to Annotate the Human Protein Kinome Tree. *PeerJ* **2013**, *1*, No. e126.
(31) Elkins, J. M.; Fedele, V.; Szklarz, M.; Abdul Azeez, K. R.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X. P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive Characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2016**, *34*, 95−103.
(32) Miduturu, C. V.; Deng, X.; Kwiatkowski, N.; Yang, W.; Brault, L.; Filippakopoulos, P.; Chung, E.; Yang, Q.; Schwaller, J.; Knapp, S.; King, R. W.; Lee, J. D.; Herrgard, S.; Zarrinkar, P.; Gray, N. S. High-Throughput Kinase Profiling: A More Efficient Approach toward the Discovery of New Kinase Inhibitors. *Chem. Biol.* **2011**, *18*, 868−879.
(33) https://doi.org/10.5281/zenodo.1148959.

DOI: 10.1021/acsomega.7b01960
ACS Omega 2018, 3, 1147−1153

Summary

A large-scale analysis of multi-kinase inhibitor selectivity was carried out on the basis of currently available high-confidence data. We systematically extracted 10,060 multi-kinase inhibitors from ChEMBL annotated for 141 qualifying human kinases. These kinases were organized in compound-based pairs of increasing phylogenetic distances. In total 596 kinase pairs sharing at least 10 inhibitors were isolated. This provided a solid framework for computationally-driven selectivity analysis.

Subsets of in part highly selective inhibitors across the majority of kinase pairs were obtained using pair- and compound-based selectivity profiles. Many instances showed subsets of nonselective kinase inhibitors, as well as others that were increasingly selective. In addition, the known selectivity determinants of the kinase binding region did not support the observed selectivity trends of kinase inhibitors. This indicated that the selectivity features are far from obvious, providing opportunities for subsequent analysis. Taken together, multi-kinase inhibitors are more selective than one might anticipate.

The analysis was based on high-confidence activity data from ChEMBL. This implies that only activity points coming from single-target assays were taken in consideration. However, cell-based assays may provide a differentiating view on inhibitor selectivity, as imposed by cell culture conditions. In order to consolidate the obtained findings, data coming from a cell-based profiling study needs to be evaluated.

In the next chapter, we focus on selectivity analysis of multi-kinase clinical candidates derived from a comprehensive cell-based profiling study.

Chapter 3

Evaluation of Kinase Inhibitor Selectivity Using Cell-Based Profiling Data

Introduction

The cell-based chemoproteomic profiling study by Klaeger *et al.*[52] presents the most comprehensive kinase inhibitor experiment up to date. In this work, 243 kinase inhibitors at different stages of clinical development were found to interact with a total of 253 kinases. The study revealed compounds with different inhibition characteristics, including both highly promiscuous as well as highly selective inhibitors. As a result, the profiling study data were made publicly available.

Herein, we further explored selectivity profiles of clinical kinase inhibitors using the available cell-based data. This yielded insights into the kinase inhibitor selectivity under cell-based assay conditions. In order to reconcile the findings with previous results, compound-based kinase pairs were formed and their selectivity profiles were evaluated.

Communication

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201800024

Evaluation of Kinase Inhibitor Selectivity Using Cell-based Profiling Data

Filip Miljković[a] and Jürgen Bajorath*[a]

Abstract: Kinases are among the most heavily investigated drug targets and inhibition of kinases and kinase-dependent signaling has become a paradigm for therapeutic intervention. Kinase inhibitors and associated activity data have increasing 'big data' character, which presents challenges for computational analysis, but also unprecedented opportunities for learning from compound data and for data-driven medicinal chemistry. Herein, publicly available kinase inhibitor data are evaluated and a number of characteristics are discussed. In addition, selectivity of clinical kinase inhibitors is explored computationally on the basis of recently reported cell-based profiling data. For inhibitors shared by pairs of kinases, selectivity profiles were generated and a variety of selective inhibitors were identified. Uni-directional selectivity profiles revealed inhibitors that were selective for one kinase over the other, while bi-directional profiles uncovered compounds with inverted selectivity for paired kinases.

Keywords: Protein kinases · drug targets · kinase inhibitors · selectivity · profiling data · compound data mining

In the first part of this contribution, we comment on emerging big data characteristics of compound activity data in general and kinase inhibitors in particular. In the second part, we present a selectivity analysis for clinical kinase inhibitors on the basis of cell-based profiling data.

Kinases are among the most popular targets in drug discovery.[1–4] In oncology, kinase inhibitors have taken center stage over the past decade.[2] In other therapeutic areas such as immunology and inflammation, second generation kinase inhibitors are on the rise.[4] While many kinase inhibitors used in cancer treatment act on multiple targets and are efficacious through polypharmacology,[2,5–7] kinase inhibitors considered for other therapeutic applications, for example, the treatment of chronic inflammatory diseases, must have a high degree of target selectivity.[4]

Compound activity data in pharmaceutical research have increasing 'big data' character,[8] which also applies to kinase inhibitors. The volume of compound activity data is steadily increasing. Of course, millions of small molecules with associated activity measurements now available are still a small sample of "data points" compared to other fields such as biology, particles physics, or – even more so – telecommunication and social networks. However, data volumes need to be considered within the particular context of a scientific discipline and – from this viewpoint – the big data era is looming in medicinal chemistry. However, there is more to big data than increasing volumes.[8] Additional criteria need to be taken into consideration such as increasing heterogeneity of compound data across different repositories, an unprecedented variety of data entries, and their increasing complexity.[8] These criteria clearly apply to compound activity data, supporting increasing big data character. Thus, the big data era does not only provide substantial opportunities for learning from data and for data-driven research and development but also challenges data analysis and requires new concepts and strategies for large-scale data exploration and learning.

For inhibitors of the human kinome, which comprises 518 kinases,[9] big data trends are evident. In 2015, we detected 18,951 kinase inhibitors in ChEMBL[10] release 18 for which high-confidence activity data were available. In this context, high-confidence data require highest possible assay and measurement confidence on the basis of ChEMBL records.[11] These 18,951 inhibitors were active against a total of 266 human kinases belonging to 10 different kinase groups. In 2017, we identified 45,728 kinase inhibitors in ChEMBL release 23, which were active against 286 human kinases from 12 groups.[11] Hence, over the course of only two years the number of publicly available kinase inhibitors with high-confidence activity data more than doubled (with a factor of 2.41), while coverage of the human kinome only moderately increased. However, when all available kinase annotations were taken into consideration, regardless of data confidence levels, 128,260 putative inhibitors were identified in ChEMBL release 23 that were annotated with a total of 439 human kinases,[11] representing more than 80% of the kinome. Profiling experiments in which inhibitors are tested under varying conditions against large numbers of kinases are a particularly rich source of activity data.[12–15]

[a] *F. Miljković, J. Bajorath*
Department of Life Science Informatics
Bonn-Aachen International Center for Information Technology
Rheinische Friedrich-Wilhelms-Universität Bonn
Endenicher Allee 19c, D-53115 Bonn (Germany)
Tel: +49-228-7369-100
Fax: +49-228-7369-101
E-mail: bajorath@bit.uni-bonn.de

However, most profiling experiments are currently carried out in pharmaceutical companies and are rarely published.

Given the relevance of polypharmacology for cancer treatment,[2] the promiscuity of kinase inhibitors continues to be a much debated issue. Without doubt, highly promiscuous kinase inhibitors exist,[13,15] but there are also highly selective inhibitors.[13,15,16] Prevalent among kinase inhibitors are so-called type I inhibitors, which bind to the ATP site in the active form of kinases. The ATP site is largely conserved across the kinome.[17] By contrast, type II inhibitors bind to the inactive form of kinases to regions adjacent to the ATP site that are less conserved.[17] Therefore, type II inhibitors are often thought to be more selective than type I inhibitors. In addition, small numbers of allosteric type III and type IV inhibitors have been discovered that bind to non-conserved regions in individual kinases distant from the ATP site and are thus expected to be most selective or even specific.[17,18]

In light of ongoing discussions about binding characteristics, it is important to note that currently available activity data do not support the presence of generally high degrees of promiscuity among kinase inhibitors. Neither can the absence of detectable promiscuity be entirely attributed to data sparseness, given the large volumes of inhibitors and activity measurements that are already available. For example, more than 95% of kinase inhibitors currently available in ChEMBL are most likely type I inhibitors.[19] For kinase inhibitors with high-confidence activity data in ChEMBL release 23, the mean promiscuity degree (number of kinase targets) was merely 1.36.[11] Even if no confidence criteria were applied and all kinase annotations taken into account, the mean promiscuity degree of 128,260 putative kinase inhibitors only moderately increased to 3.86.[11] On the other hand, neither inhibitor profiling[15] nor systematic compound data mining[20] confirmed assumed selectivity differences between type I and II kinase inhibitors with experimentally confirmed binding modes. Thus, the topic of kinase inhibitor promiscuity versus selectivity remains to be further investigated.

In the following, we report selectivity analysis for a set of clinical kinase inhibitors.

Recently, we have systematically investigated inhibitor selectivity at the level of kinase pairs.[16] Pair-based analysis was carried out to account for all possible selectivity relationships. From ChEMBL release 23, inhibitors with high-confidence activity data for two or more human kinases were selected und used to form inhibitor-based kinase pairs. Two kinases formed a pair if they shared at least 10 inhibitors. On the basis of 10,060 qualifying multi-kinase inhibitors, 596 pairs were obtained that involved 141 kinases distributed across the human kinome. For each pair, compound selectivity was assessed by calculating the logarithmic potency difference (ΔpIC_{50}) for each inhibitor. On the basis of this analysis, more than half of all kinase pairs were associated with one or more inhibitors having a potency difference of more than two orders of magnitude,[16] reflecting notable selectivity. Although one might anticipate that selectivity might increase for pairs of kinases with increasing phylogenetic distances, interestingly, similar proportions of pairs with selective inhibitors were detected for kinases from the same family, different families, and different groups.[16]

The recent publication of the currently most comprehensive kinase inhibitor profiling study has been an important advance for the field. For 243 kinase inhibitors at different stages of clinical development, cell-based chemoproteomic profiling was carried out.[15] Kinobead assays using immobilized non-specific kinase inhibitors were used to screen lysates of cancer cell lines and extract bound target proteins. Binding was detected using quantitative mass spectrometry. Using loaded kinobeads, dose-dependent competition assays with the set of 243 clinical kinase inhibitors were then carried out to identify their targets. The inhibitors interacted with a total of 253 kinases, revealing different inhibition characteristics and highly promiscuous as well as selective inhibitors.[15] These findings correlated well with conclusions drawn from systematic mining of kinase inhibitor data from ChEMBL.[20]

Herein we have applied kinase pair-based analysis of inhibitor selectivity[16] to cell-based profiling data of Klaeger et al.[15] Therefore, clinical kinase inhibitors studied by Klaeger et al. and their kinase annotations were assembled from ProteomicsDB.[21] Only kinase information associated with the "4 cell line mix" lysate type and the "kinobeads" type were considered. Activity-related target classification was set to "high confidence" yielding apparent dissociation constants (K_d values) as activity measurements for inhibitors. K_d values were converted to pK_d values by calculating the negative decadic logarithm of reported nanomolar concentrations. On the basis of these criteria, 216 clinical kinase inhibitors were obtained that were annotated with a total of 225 human kinases.

On the basis of the selected activity data, pairs of kinases were systematically assembled that shared at least 10 inhibitors, which resulted in a total of 2369 pairs. These pairs involved 137 kinases and 190 clinical inhibitors. Figure 1 shows the distribution of these kinases over the human kinome. For all compounds associated with pairs, potency differences between the kinases were determined on the basis of pK_d values. Figure 2 (top) shows that potency differences between inhibitors were overall small, with a median value of 0.67 pK_d units, well within an order of magnitude. When only the inhibitor with largest potency difference per pair was considered, a different distribution was observed, as shown in Figure 2 (bottom). In this case, a median value of 2.24 was obtained, indicating that there were individual inhibitors across many pairs with a strong tendency of selectivity.

On the basis of these findings, we further analyzed the most selective inhibitors for both kinases forming a pair. Hence, in each case, the inhibitors most selective for kinase *A* over *B* and *B* over *A* were identified, yielding two

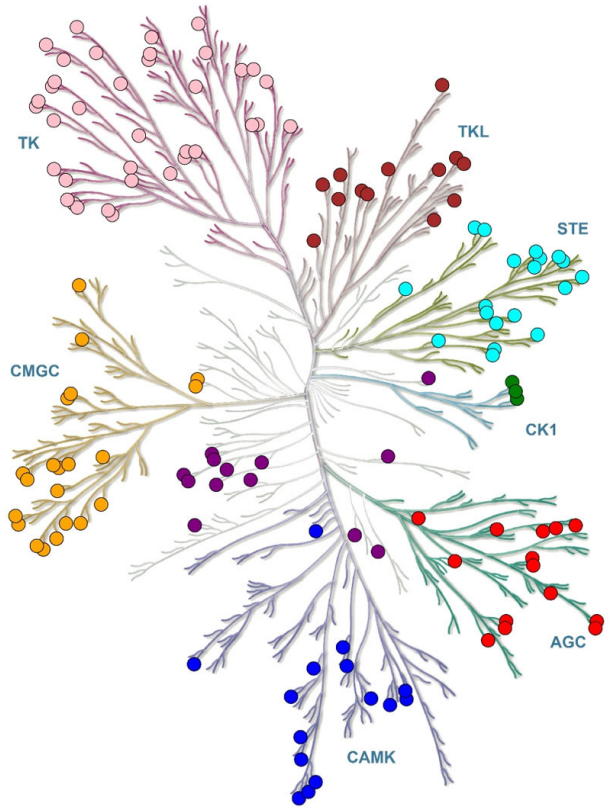


Figure 1. Kinase distribution. Kinases forming inhibitor-based pairs with more than 10 inhibitors are mapped onto a phylogenetic tree of the human kinome (drawn with KinMap[22]). Kinases are represented as dots and color-coded by kinase groups.

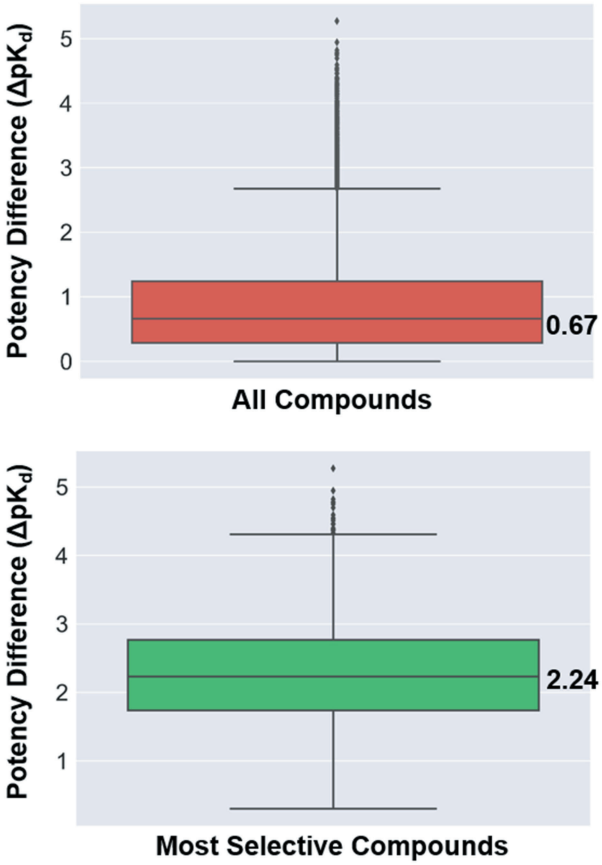


Figure 2. Compound potency differences. Boxplots report distribution of potency differences of inhibitors over kinase pairs as the mean potency difference of all inhibitors per pair (top) and the largest potency difference (bottom, representing most selective compounds). For each distribution, the median value is given. 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), largest value (top line), and outliers (points below the smallest or above the largest value).

inhibitors per pair. Then, a *selectivity range* was calculated as follows:

$$\textit{Selectivity range} = \Delta pK_{d\,(A\,over\,B)} + \Delta pK_{d\,(B\,over\,A)}$$

where $\Delta pK_{d\,(A\,over\,B)}$ represents potency difference for inhibitor selective for kinase *A* over *B* and $\Delta pK_{d\,(B\,over\,A)}$ represents potency difference for inhibitor selective for kinase *B* over *A*. For example, if an inhibitor most selective for kinase *A* had a potency difference of 0.5 and another most selective for kinase *B* a potency difference of 2.0, the selectivity range for the kinase pair was 2.5 orders of magnitude.

Figure 3 shows the distribution of selectivity ranges over kinase pairs revealing the presence of many large selectivity ranges, with a median of 3.42 pK_d units. Thus, many pair sets contained inhibitors with selectivity for each kinase.

For kinase pair sets, compound-based selectivity profiles[16] were generated. These profiles record potency differences for inhibitors. For each inhibitor, the potency against the two kinases is plotted. Thus, each inhibitor is represented by two corresponding data points. Compounds are then ordered according to increasing potency difference for each kinase from the left to right and *vice versa* (Figure 4a).

For profile analysis, a potency difference-based selectivity criterion of at least 2 orders of magnitude ($pK_d \geq 2$) was applied, i.e., inhibitors meeting this criterion were classified as selective. A compound set might contain one or more inhibitors selective for kinase *A* over *B* but none selective for *B* over *A*. These selectivity profiles, which we term *uni-directional*, were originally generated for kinase inhibitors from ChEMBL.[16] An example for an uni-directional profile is shown in Figure 4a (top) where inhibitors become increasingly selective from the right to the left for kinase FLT3 over MAPK9. Furthermore, it is also possible that selectivity profiles capture inhibitors that are selective for kinase *A* over *B* and others that are selective for *B* over *A*. These *bi-directional* profiles are of particular interest because they contain inhibitors with inverted selectivity. Three exemplary bi-directional profiles are shown below the uni-directional profile in Figure 4a. Since compounds are ordered accord-

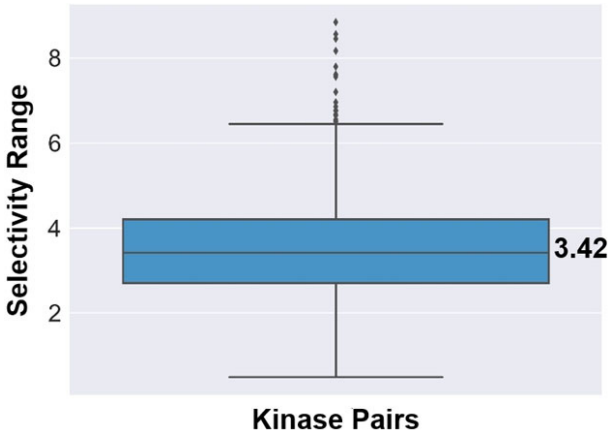


Figure 3. Selectivity ranges. Reported is the distribution of selectivity ranges over kinase pairs. The boxplot is represented according to Figure 2.

ing to increasing potency differences with respect to both kinases, these profiles contain inhibitors with inverted kinase selectivity on the left and on the right. For example, the bi-directional profile for the PDGFRB and YES1 kinase pair comprises four inhibitors in the central section that have essentially the same potency against both kinases and others that are increasingly selective for PDGFRB (left) or YES1 (right). Inhibitors with inverted selectivity for kinases merit close inspection because they might reveal selectivity-conferring core structures or other selectivity determinants. Selective inhibitors from exemplary profiles are shown in Figure 4b.

The 2369 kinase pair sets included 1453 sets (~61%) with selective clinical inhibitors. Thus, selective inhibitors were widely distributed over kinase pairs. Sets with selective inhibitors yielded 1229 uni-directional and 224 bi-directional selectivity profiles (~10%). The bi-directional profiles were associated with 90 kinases and 157 inhibitors. Hence, selectivity trends revealed by bi-directional profiles involved a significant number of kinases and clinical inhibitors.

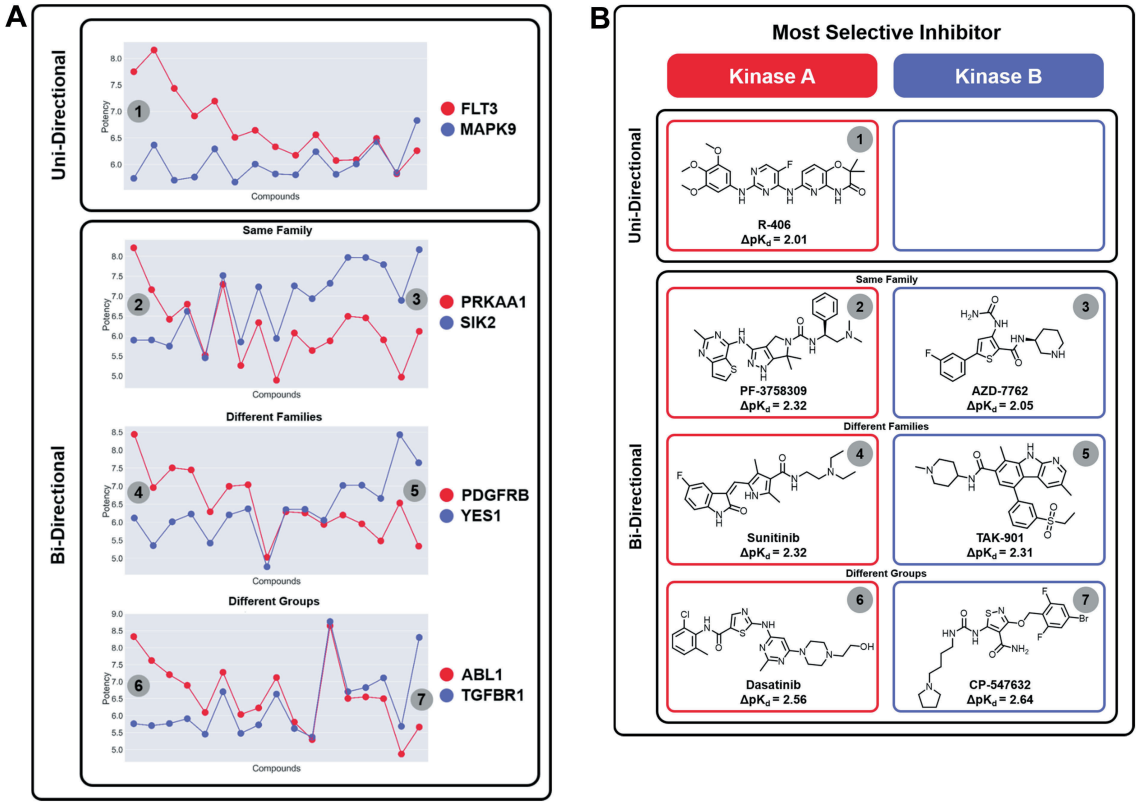


Figure 4. Selectivity profiles and representative inhibitors. (a) Shown are one uni-(top) and three bi-directional compound-based selectivity profiles (for pairs of kinases from the same family, different families, and different groups, respectively). For each inhibitor, the potency against the two kinases forming a pair is compared. Compounds are ordered according to increasing potency differences for each kinase from the left to right and vice versa. The most selective inhibitors are labeled. In (**b**), names and structures of most selective inhibitors meeting the $pK_d \geq 2$ selectivity criterion are shown and their potency differences are reported. Kinases are abbreviated as follows:[15] FLT3, receptor-type tyrosine-protein kinase FLT3; MAPK9, mitogen-activated protein kinase 9; PRKAA1, 5′-AMP-activated protein kinase catalytic subunit alpha-1; SIK2, serine/threonine-protein kinase SIK2; PDGFRB, platelet-derived growth factor receptor beta kinase; YES1, tyrosine-protein kinase Yes; ABL1, tyrosine-protein kinase ABL1; TGFBR1, TGF-beta receptor type-1 kinase.

We also revisited previously reported kinase pairs from ChEMBL release 23[16] and searched for bi-directional selectivity profiles. In this case, 10,060 multi-kinase inhibitors were annotated with 141 human kinases, yielding 596 pairs sharing at least 10 inhibitors. The corresponding pair sets included 380 sets with selective inhibitors ($pK_d \geq 2$), which yielded 288 uni- and 92 bi-directional (~24%) selectivity profiles.

Thus, cell-based profiling using a small set of clinical inhibitors resulted in larger kinome coverage than collection of a large set of inhibitors and associated activity data from medicinal chemistry. However, in both cases, similar selectivity trends were detected.

Herein we have discussed characteristic features of inhibitors targeting the human kinome and carried out kinase pair-based selectivity analysis using cell-based profiling data for clinical kinase inhibitors. Our findings suggest that many currently available kinase inhibitors have the potential to differentiate between kinase targets and that there are substantial differences in selectivity among these inhibitors.

Conflict of Interest

None declared.

Acknowledgement

The authors gratefully acknowledge Klaeger et al. for making their kinase inhibitor profiling data publicly available.

References

[1] P. Cohen *Nat. Rev. Drug Discovery* **2002**, *1*, 309–315.
[2] Z. A. Knight, H. Lin, K. M. Shokat, *Nat. Rev. Cancer* **2010**, *10*, 130–137.
[3] D. L. Simmons, *Curr. Opin. Pharmacol.* **2013**, *13*, 426–434.
[4] S. Laufer, J. Bajorath, *J. Med. Chem.* **2014**, *57*, 2167–2168.
[5] Y. Hu, J. Bajorath, *Drug Discov. Today* **2013**, *18*, 644–650.
[6] J.-U. Peters, *J. Med. Chem.* **2013**, *56*, 8955–8971.
[7] A. Anighoro, J. Bajorath, G. Rastelli, *J. Med. Chem.* **2014**, *57*, 7874–7887.
[8] Y. Hu, J. Bajorath, *Future Science OA* **2017**, *3*, FSO179.
[9] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, *Science* **2002**, *298*, 1912–1934.
[10] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
[11] D. Stumpfe, A. Tinivella, G. Rastelli, J. Bajorath, *RSC Adv.* **2017**, *7*, 41265–41271.
[12] M. W. Karaman, S. Herrgard, D. K. Treiber, P. Gallant, C. E. Atteridge, B. T. Campbell, K. W. Chan, P. Ciceri, M. I. Davis, P. T. Edeen, R. Faraoni, M. Floyd, J. P. Hunt, D. J. Lockhart, Z. V. Milanov, M. J. Morrison, G. Pallares, H. K. Patel, S. Pritchard, L. M. Wodicka, P. P. Zarrinkar, *Nat. Biotechnol.* **2008**, *26*, 127–132.
[13] M. I. Davis, J. P. Hunt, S. Herrgards, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, P. P. Zarrinkar, *Nat. Biotechnol.* **2011**, *29*, 1046–1052.
[14] J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, P. J. Hajduk, *Nat. Chem. Biol.* **2011**, *7*, 200–202.
[15] S. Klaeger, S. Heinzlmeir, M. Wilhelm, H. Polzer, B. Vick, P. A. Koenig, M. Reinecke, B. Ruprecht, S. Petzoldt, C. Meng, J. Zecha, K. Reiter, H. Qiao, D. Helm, H. Koch, M. Schoof, G. Canevari, E. Casale, S. R. Depaolini, A. Feuchtinger, Z. Wu, T. Schmidt, L. Rueckert, W. Becker, J. Huenges, A. K. Garz, B. O. Gohlke, D. P. Zolg, G. Kayser, T. Vooder, R. Preissner, H. Hahne, N. Tõnisson, K. Kramer, K. Götze, F. Bassermann, J. Schlegl, H. C. Ehrlich, S. Aiche, A. Walch, P. A. Greif, S. Schneider, E. R. Felder, J. Ruland, G. Médard, I. Jeremias, K. Spiekermann, B. Kuster, *Science* **2017**, *358*, eaan4368.
[16] F. Miljković, J. Bajorath, *ACS Omega* **2018**, *3*, 1147–1153.
[17] Z. Zhao, H. Wu, L. Wang, Y. Liu, S. Knapp, Q. Liu, N. S. Gray, *ACS Chem. Biol.* **2014**, *9*, 1230–1241.
[18] L. K. Gavrin, E. Saiah, *MedChemComm* **2013**, *4*, 41–51.
[19] Y. Hu, N. Furtmann, J. Bajorath, *J. Med. Chem.* **2015**, *58*, 30–40.
[20] F. Miljković, J. Bajorath, *ACS Omega* **2018**, *3*, 3113–3119..
[21] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz et al., *Nature* **2014**, *509*, 582–587; http://www.proteomicsdb.org/.
[22] S. Eid, S. Turk, A. Volkamer, F. Rippmann, S. Fulle, *BMC Bioinf.* **2017**, *18*, e16.

Received: March 2, 2018
Accepted: March 10, 2018
Published online on March 30, 2018

Summary

Extensive evaluation of kinase inhibitor selectivity was carried out on the basis of recently reported cell-based profiling data. A total of 2369 kinase pairs were formed by 190 clinical inhibitors, targeting 137 kinases. Compared to previous results on the basis of medicinal chemistry data, a significantly smaller set of inhibitors yielded a larger number of qualifying kinase pairs. However, both studies resulted in similar selectivity trends.

Furthermore, compound-based selectivity profiles with compounds meeting the selectivity criteria were divided into two major types. Uni-directional selectivity profiles uncovered inhibitors selective for one kinase over another, while bi-directional profiles revealed compounds with inverted selectivity for pair-forming kinases. These selectivity profiles were detected across all categories with increasing phylogenetic distances. Many currently available clinical candidates were found to differentiate between kinase targets.

In the following study, we explore selectivity trends of clinical kinase inhibitors on the basis of currently available medicinal chemistry data. Kinase inhibitors classified as chemical probes were of special interest in this study.

Chapter 4

Reconciling Selectivity Trends from a Comprehensive Kinase Inhibitor Profiling Campaign with Known Activity Data

Introduction

Previously, selectivity trends of clinical kinase inhibitors were evaluated using the cell-based activity data. Herein, we aimed to correlate these findings with the selectivity of clinical candidates on the basis of medicinal chemistry data.

The importance of different data confidence criteria when interpreting the findings was discussed. For the set of most and least selective inhibitors found in the profiling study, their PD values were estimated. Similarly, a number of inhibitors designated as type I and type II were compared to evaluate their selectivity. In addition, for clinical candidates designated as chemical probes target profiles were generated.

ACS OMEGA

Article

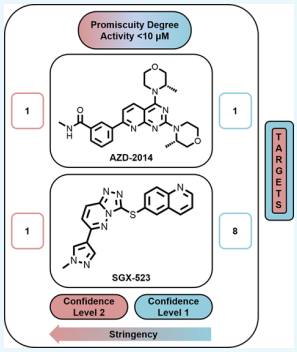
Cite This: ACS Omega 2018, 3, 3113−3119

Reconciling Selectivity Trends from a Comprehensive Kinase Inhibitor Profiling Campaign with Known Activity Data

Filip Miljković and Jürgen Bajorath*

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

ABSTRACT: Kinase inhibitors are among the most intensely investigated compounds in medicinal chemistry and drug development. Profiling experiments and kinome screens reveal binding characteristics of kinase inhibitors and lead to better understanding of selectivity and promiscuity patterns. However, only limited amounts of profiling data are publicly available. By contrast, a large body of activity data for inhibitors of human kinases has become available from medicinal chemistry. In this study, we have correlated selectivity assessment of clinical kinase inhibitors from the most comprehensive profiling campaign reported to date with systematic mining of activity data from other sources. The results of our comparative analysis reveal consistency of orthogonal approaches in the study of kinase inhibitor selectivity versus promiscuity and stress the importance of taking alternative data confidence criteria into account. Moreover, it is also shown that there are little if any detectable differences in selectivity between type I and II kinase inhibitors and that inhibitors designated as chemical probes have very different target profiles.



1. INTRODUCTION

Kinase inhibitors are prime candidates for drug development in different therapeutic areas such as oncology, inflammatory diseases, and so forth.[1−4] To better understand their binding characteristics and target profiles, kinase inhibitors have been—and continue to be—subjected to profiling experiments including various panel assays and kinome screens.[5−12] Because the majority of current kinase inhibitors bind to the conserved adenosine 5′-triphosphate (ATP) site in kinases, or regions proximal to this site,[12−14] selectivity versus promiscuity of kinase inhibitors is still an intensely debated issue,[2−4,12−16] with important implications for therapeutic applications and clinical performance.[2,3,17]

Recently, Klaeger et al. have reported the most comprehensive kinase inhibitor profiling study available to date,[18] yielding a variety of binding, functional, and structural data for a set of 243 kinase inhibitors at different stages of clinical evaluation and development, including marketed drugs. The authors primarily applied a chemoproteomics approach. "Kinobeads", that is, nonspecific kinase inhibitors immobilized on the solid phase, were used to extract bound target proteins from mixed lysates of different cancer cell lines. Target binding was then determined using quantitative mass spectrometry. Using loaded kinobeads from lysates, dose-dependent competition assays with clinical kinase inhibitors were carried out to identify their targets and determine apparent dissociation constants.[18] The set of clinical kinase inhibitors was found to interact with a total of 253 kinases, comprising nearly half of the human kinome. A key finding of this study has been that the investigated clinical kinase inhibitors covered a wide spectrum of binding characteristics ranging from selective to highly promiscuous compounds.[18]

Given this extensive in vivo-oriented target identification effort for clinical kinase inhibitors and the variety of target profiles that were observed, we were interested in evaluating how some of the findings of Klaeger et al. might relate to the promiscuity assessment of kinase inhibitors on the basis of currently available activity data. We reasoned that comparison with literature data from medicinal chemistry might often provide complementary or orthogonal views of inhibitor selectivity versus promiscuity, given the many different assays these compounds were tested in. Klaeger et al. also searched the kinase inhibitor literature and retrieved biological activity annotations from ChEMBL,[19] the major public repository of compounds and activity data from medicinal chemistry sources. They noted that no bioactivity records were deposited for 35 of the clinical kinase inhibitors under investigation.[18]

We have systematically collected all activity data available in ChEMBL for the clinical kinase inhibitors studied by Klaeger et al. and organized these data according to different confidence criteria. Then target annotations were identified and promiscuity degrees (PDs) of inhibitors were calculated at different data confidence levels. We also identified kinase inhibitors that were most and least selective on the basis of the data of Klaeger et al. and separately determined the target profiles for these inhibitors. The comparison of our findings with results of Klaeger et al. is reported herein.

Received: February 8, 2018
Accepted: March 5, 2018
Published: March 14, 2018

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

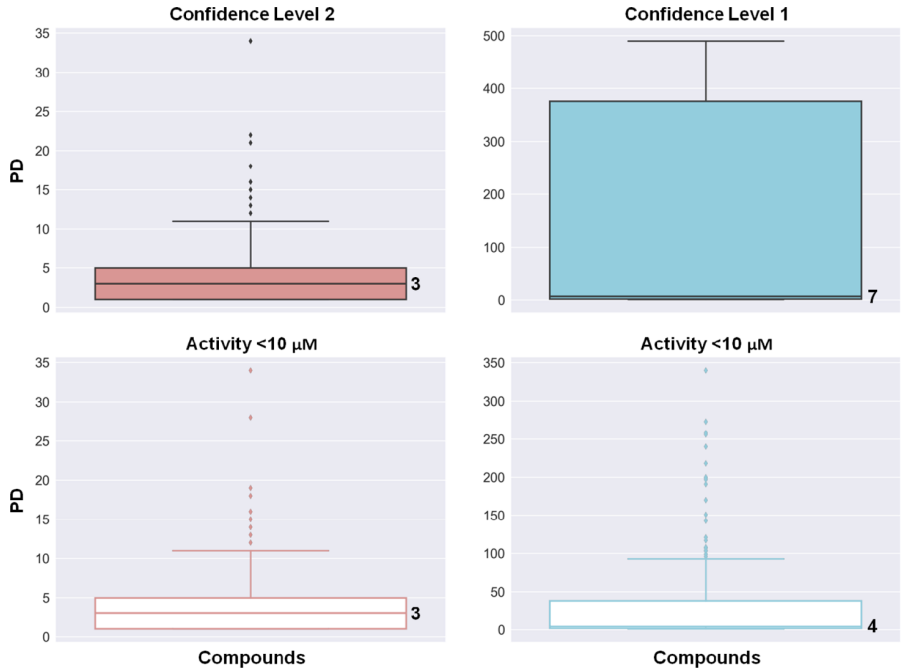


Figure 1. PDs on the basis of known activities. Box plots report the distribution of the PD of clinical kinase inhibitors on the basis of activity data from ChEMBL at different confidence levels (top: level 2, 166 inhibitors; level 1, 185) and after applying the <10 μM activity threshold (bottom: level 2, 164; level 1, 172). Box plots contain the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary), largest value (top line), and outliers (points below the bottom or above the top line).

2. MATERIALS AND METHODS

2.1. Clinical Kinase Inhibitors and Data Confidence Level Criteria. ChEMBL identifiers and SMILES representations of clinical kinase inhibitors were taken from the supplementary information of Klaeger et al.[18] and mapped to ChEMBL (release 23, accessed in Jan 2018).[19] Only human targets were considered to conform with Klaeger et al. For inhibitors annotated with human targets, activity data were collected at two different confidence levels including levels 1 (intermediate) and 2 (high). For confidence level 1, the highest assay confidence was required for ChEMBL data, and for level 2, the highest assay and, in addition, highest measurement confidence were required for ChEMBL data. Accordingly, activity data were only selected from direct inhibition assays (assay relationship type "D") for single targets with the highest assay confidence score ("9"). In addition, only unambiguously specified K_i or IC_{50} measurements with standard activity unit ("nM") and consistent "activity comments" were considered. To address the compound concentration dependence of target annotations and identify weak inhibitory interactions, we also generated results for comparison after applying a <10 μM activity (potency) threshold to both data confidence levels.

2.2. Selectivity Scores and Promiscuity Degrees. Klaeger et al. introduced the "Concentration- And Target-Dependent Selectivity" (CATDS) score for the analysis of their experiments.[18] The CATDS score quantifies the *reduction in binding of a given target to kinobeads at a particular inhibitor concentration relative to the summed reduction in binding of all available targets*. As defined, CATDS scores of inhibitors are target-dependent. Scores approaching 1 are characteristic of a *selective* kinase inhibitor (i.e., the compound almost exclusively inhibits a single target), whereas scores close to 0 are usually indicative of a *nonselective* (highly promiscuous) inhibitor. For each clinical kinase inhibitor found in ChEMBL, we selected the largest available CATDS score as a selectivity measure. Furthermore, we calculated the "promiscuity degree" (PD) of an inhibitor as the number of its unique targets on the basis of the activity records from ChEMBL that qualified for confidence levels 1 or 2 in the presence or absence of the <10 μM activity threshold.

2.3. Kinase Inhibitor Types and Chemical Probes. A subset of clinical kinase inhibitors was assigned by Klaeger et al. to type I or type II inhibitor category on the basis of the available structural data and binding mode information.[18] Type I kinase inhibitors bind to the conserved ATP site in the active form of the kinase, whereas type II inhibitors bind to the inactive form and less conserved regions adjacent to the ATP site.[13,14] Therefore, type II inhibitors are often expected to be more selective than type I inhibitors. We separately analyzed PDs for inhibitors with type I or II binding mode.

Furthermore, a subset of clinical kinase inhibitors were designated chemical probes[18] on the basis of reports from the Chemical Probes Portal.[20] For such probe compounds used in chemical biology, a high degree of selectivity is generally required. Therefore, we also separately analyzed PD values of the designated chemical probes among clinical kinase inhibitors.

3. RESULTS AND DISCUSSION

3.1. Clinical Kinase Inhibitors and Data Confidence Levels. Structures of 3 of the 243 clinical kinase inhibitors were not found in ChEMBL. In addition, 38 inhibitors were not annotated with human targets, leaving 202 inhibitors with at least one known human target for activity data confidence analysis. For confidence level 1, the highest assay confidence was required, and for the more stringent level 2, the highest assay plus highest measurement confidence were required.

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

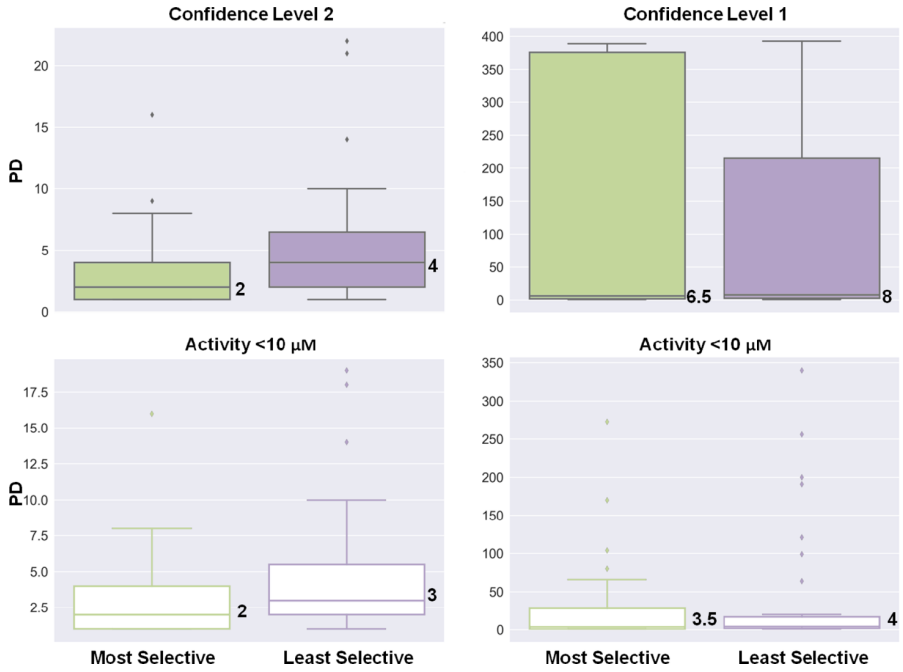


Figure 2. PDs for different selectivity categories. Box plots report the distribution of PD values of the subsets of the most and least selective clinical kinase inhibitors according to CATDS scores (see Materials and Methods) at different confidence levels of ChEMBL data (top: level 2, most selective: 36 inhibitors, least selective: 31; level 1, most selective: 38, least selective: 33) and after applying the <10 μM activity threshold (bottom: level 2, most selective: 35 inhibitors, least selective: 31; level 1, most selective: 36, least selective: 31).

These data confidence levels were established to exclude target annotations from the analysis that were only weakly supported experimentally (e.g., target annotations from cell-based assays lacking confirmation of direct target engagement).

Confidence level 1 was met by activity data of 185 of 202 clinical kinase inhibitors available in ChEMBL. These 185 inhibitors were active against a total of 394 human kinases and 218 nonkinase targets. After applying the <10 μM activity threshold, 172 inhibitors were available for confidence level 1 that were active against 379 kinases and 64 nonkinase targets. Furthermore, 166 of the 185 inhibitors qualified for confidence level 2, which were active against 122 human kinases and 66 nonkinase targets. After applying the <10 μM activity threshold to confidence level 2, 164 inhibitors were obtained with activity against 122 kinases and 52 nonkinase targets. Hence, there was a sharp decline in target numbers at increasing data confidence. For comparison, at confidence level 1, clinical inhibitors were active against a total of 394 human kinases on the basis of the currently available data (379 human kinases after applying the activity threshold), while Klaeger et al. identified 253 kinase targets. The human kinome comprises 518 kinases.[21]

Although a significant number of nonkinase targets were identified, the majority of the clinical kinase inhibitors were predominantly active against kinases. After applying the <10 μM activity threshold to both data confidence levels, the number of nonkinase target annotations notably reduced by 154 targets for confidence level 1 and by 14 targets for confidence level 2, much more so than the number of human kinase annotations (with 15 kinases for confidence level 1 and 0 for confidence level 2). For statistical considerations, we also calculated fractional kinase PD values, defined as (#kinase targets/#targets). On average, these values were very close to 1. Therefore, for the purpose of our statistical analysis, it was not required to further distinguish between kinase and nonkinase targets.

3.2. Global Promiscuity Degrees. Figure 1 shows the distribution of PD values for the inhibitor subsets at confidence levels 1 and 2. At level 1, a broad distribution was observed with inhibitors in the upper quartile having hundreds of target annotations. However, although supported by in vitro assay confidence, many of these PDs were most likely artificially high because it is hardly conceivable that a clinical compound might indeed act in vivo on hundreds of targets. Of course, at high—or artificially high—compound concentrations, more activities might be detected. When the activity threshold was applied to level 1, the distribution became much more narrow, and the median PD was reduced from 7 to 4, whereas the distribution for level 2 remained nearly unchanged. In this context, it should be noted that Klaeger et al. detected 494 transcribed kinases including mutant forms in their experiments and 363 translated kinases, 253 of which were bound to kinobeads.[18]

Hence, on the basis of medicinal chemistry data, the 185 inhibitors qualifying for confidence level 1 (172 after applying the activity threshold) were annotated with a larger fraction of the human kinome (394 kinases, 379 after applying the threshold) than that was accessible to the 243 inhibitors during proteomics profiling.

However, Figure 1 also shows that many inhibitors at confidence level 1 had only low PD values, especially after applying the activity threshold. Taken together, these findings were consistent with the identification of selective to highly promiscuous inhibitors by Klaeger et al.

At confidence level 2, the distribution of the PD values was narrow, with a median of 3, and an upper quartile range of 3−5, with only a limited number of statistical outliers having PD values larger than 10 (there were only little differences when applying the activity threshold). Thus, the comparison in Figure

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

1 revealed a strong influence of data confidence criteria on the global distribution of PDs. Hence, analyzing activity data and target annotations at different confidence levels yielded a differentiated view of the target space of kinase inhibitors charted under varying experimental stringency. It also provides a meaningful framework for evaluating the results of profiling experiments.

3.3. Most and Least Selective Kinase Inhibitors. Next, we analyzed the distribution of CATDS scores reported by Klaeger et al. to determine subsets of the most and least selective inhibitors according to this scoring scheme. Therefore, a score histogram was generated, and the resulting distribution was fitted to a normal distribution, yielding a mean value m and standard deviation σ of 0.510 and 0.285, respectively. Then subsets of the most and least selective inhibitors were defined by applying score thresholds of 1 σ above and below the mean, respectively. The resulting most selective (CATDS $\geq m + \sigma$; CATDS $\geq$ 0.795) and least selective (CATDS $\leq m - \sigma$; CATDS $\leq$ 0.225) subsets contained 39 and 36 inhibitors, respectively.

Figure 2 shows the distribution of PD values for these subsets at confidence levels 1 and 2. At confidence level 1, broad distributions were observed for both subsets, similar to that of Figure 1, with median PD values of 6.5 and 8 for the most and least selective inhibitors, respectively. Thus, differences in selectivity between these subsets were only small. The distributions became very narrow after applying the activity threshold, and the median PD values were reduced. However, the distribution for the most selective inhibitors contained compounds with hundreds of target annotations, more so than the distribution for least selective inhibitors, indicating that this data confidence level was inappropriate to reconcile differences in selectivity suggested by CATDS scoring. A different picture emerged for distributions generated at confidence level 2. In this case, the distributions were narrow, in the presence or absence of the activity threshold, similar to that of Figure 1, yielding mean PD values of 2 and 4 (or 3) for the most and least selective inhibitors, respectively. Thus, at high activity data confidence, differences in selectivity between these subsets were also small, taking into account that the most selective inhibitor subset was defined by a CATDS score threshold of nearly 0.8, and the least selective subset was defined by a CATDS score threshold of less than 0.23. Thus, these observations suggested that similar target profiles might yield CATDS scores of different magnitudes, dependent on the relative binding contributions of different targets and that CATDS scoring and PDs might reflect selectivity in different ways.

Examples are given in Figure 3 that shows two clinical kinase inhibitors, capmatinib and lapatinib, which both belonged to the subset of most selective inhibitors. At confidence levels 2 and 1, capmatinib was only active against its primary kinase target on the basis of the literature data, also reflecting high selectivity. By contrast, lapatinib was active against 5 and 389 targets at confidence levels 2 and 1, respectively. After applying the activity threshold, lapatinib was annotated with against 3 and 13 targets at confidence levels 2 and 1, respectively. Thus, in this case, application of the activity threshold balanced the view of lapatinib promiscuity at data confidence level 1.

3.4. Different Binding Modes. Clinical kinase inhibitors available in ChEMBL included 85 compounds that were categorized as type I and 27 as type II inhibitors on the basis of the binding mode information. Figure 4 shows the PD value distributions of type I and II inhibitors. Because the number of

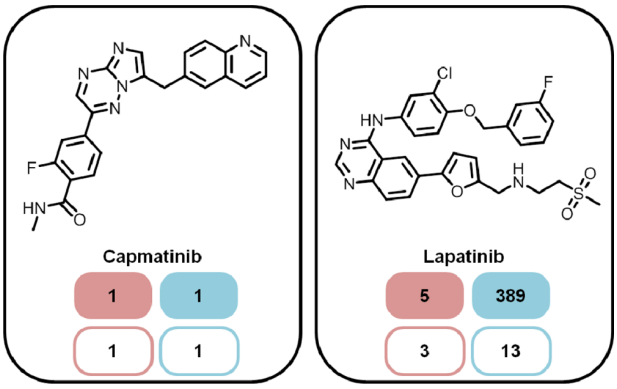


Figure 3. Examples of the most selective clinical kinase inhibitors. Two kinase inhibitors belonging to the most selective CATDS score-based subset are shown. For each inhibitor, the PD value on the basis of ChEMBL data is reported at confidence levels 2 (red background) and 1 (blue background) and after applying the <10 μM activity threshold (level 2, red outline; level 1, blue outline).

type II inhibitors was much smaller than that of type I inhibitors, statistical assessment was limited in this case, and it was difficult to directly compare the distributions. However, at confidence level 1, at least half of the designated type II inhibitors were highly promiscuous, with a median PD of 295.5, which was much larger than the median PD of 48 obtained for type I inhibitors. Similar trends were observed after applying the activity threshold, with PD median values of 9 and 26 for type I and type II inhibitors, respectively. At confidence level 2, the results were similar for type I and II inhibitors, with PD median values of 4 and 3, respectively, and outliers present in both cases. While only a limited number of type II were available, these findings did not provide evidence for often assumed greater selectivity of type II versus type I inhibitors, consistent with the results and conclusions of Klaeger et al. and earlier proposals.[14]

It should also be noted that 16 type I and 4 type II inhibitors belonged to the most selective inhibitor subset according to Figure 2, whereas 22 type I and 4 type II inhibitors belonged to the least selective subset. Hence, there was no notable relative enrichment of the designated type II over type I inhibitors in the most selective subset.

3.5. Chemical Probes. Clinical kinase inhibitors classified as chemical probes represented another interesting subset for our analysis, given that compounds used as probes typically have rather stringent requirements for selectivity. The 164 clinical kinase inhibitors meeting data confidence level 2 in the presence of the activity threshold were found to contain 13 designated chemical probes that are shown in Figure 5. For each of these inhibitors, the CATDS score is provided, revealing the presence of a large scoring range for these putative probes. In fact, only two of these compounds belonged to the most selective subset (having a CATDS score of 1), whereas two others belonged to the least selective subset (with scores of 0.15 and 0.22, respectively). However, at high data confidence (level 2), all 13 probes were selective (with one or two targets) or at least moderately selective (with four, six, or nine targets). By contrast, at confidence level 1, a clear separation was observed, as also shown in Figure 5. In this case, only four inhibitors retained PD values of 2, and three others had PD values of 14, 16, and 38, whereas the remaining six inhibitors were each annotated with more than 370 or 380

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

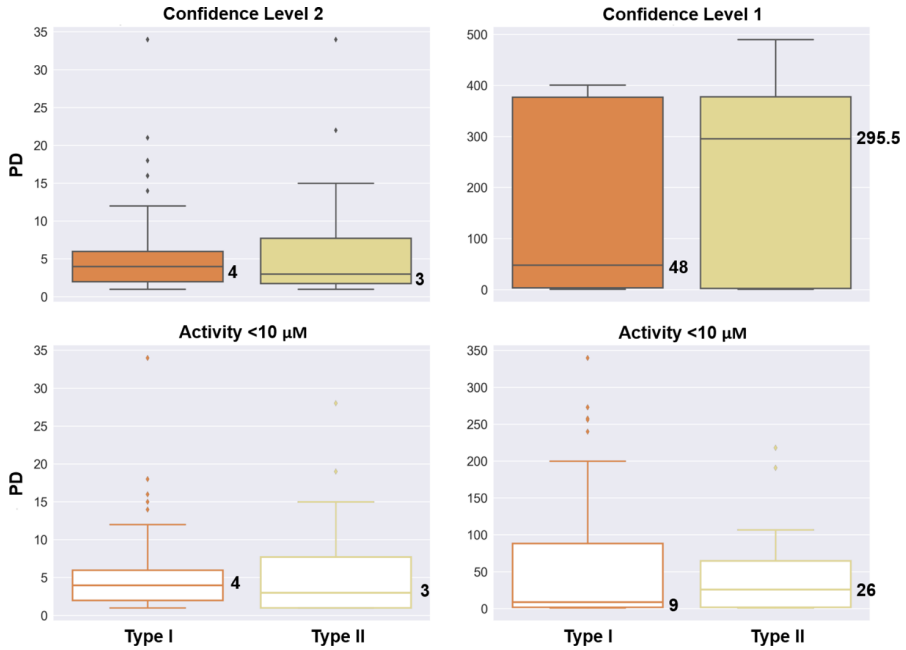


Figure 4. PDs for type I and type II kinase inhibitors. Box plots report the distribution of PD values for type I and type II inhibitors on the basis of ChEMBL data at confidence levels 1 and 2 (top: level 2, type I: 75 inhibitors; type II: 24; level 1, type I: 81; type II: 26). In addition, the PD values are reported for levels 2 and 1 after applying the <10 μM activity threshold (bottom: level 2, type I: 75; type II: 24; level 1, type I: 79; type II: 25).

targets, thus calling probe characteristics into question. There were only 4 of 13 inhibitors with PD values of 2 at both confidence levels 1 and 2, which could be considered meaningful chemical probes applying stringent criteria. However, when the activity threshold was applied to confidence level 1, the number of target annotations for chemical probes was significantly reduced, resulting in four highly promiscuous inhibitors (with 32, 37, 81, and 121 annotations) and nine others with less than 10 targets per probe. Taken together, these observations also corroborated findings by Klaeger et al. that clinical inhibitors with assumed selectivity were often promiscuous. Figure 6 shows the two designated chemical probes having the maximal CATDS score of 1, AZD-2014 and SGX-523. Inhibitor AZD-2014 was only active against two targets at data confidence levels 1 and 2 (only one after applying the activity threshold) and belonged to the group of four preferred chemical probes referred to above. By contrast, SGX-523 was active against a single kinase at confidence level 2, but annotated with 376 targets at confidence level 1, making it difficult to support its use as a probe on the basis of available activity data. After applying the activity threshold at level 1, SGX-523 was left with eight targets. However, weak activities against a variety of targets were likely in this case. This comparison illustrates the importance of comprehensive activity data analysis for evaluating putative chemical probes.

4. CONCLUSIONS

In this study, we have—to our knowledge for the first time—correlated results of an extensive cell-based kinase inhibitor profiling campaign with those obtained by systematic mining of compound activity data from different sources. Given the limited availability of profiling data in the public domain, this analysis was of high interest to us, especially considering the exploration of kinase inhibitor selectivity versus promiscuity. The analysis was focused on kinase inhibitors at different stages of clinical development, which are typically well characterized experimentally. At varying activity data confidence levels substantial differences in inhibitor promiscuity were observed. The clinical inhibitors covered a wide spectrum of target profiles, ranging from selective to highly promiscuous compounds, as revealed by both chemoproteomics profiling and data mining. A subset of inhibitors was annotated with more kinases on the basis of the activity data than were expressed under the conditions of the profiling experiment. In some instances, in vitro assays yielded hundreds of target annotations for kinase inhibitors, which could not possibly translate into in vivo settings for clinically viable compounds, thus highlighting the likely limitations of assay relevance. It was also of interest to determine the target profiles of kinase inhibitors with different binding modes thought to cause differences in selectivity. However, neither experimental profiling nor activity data mining revealed notable differences between type I and II kinase inhibitors. Moreover, we analyzed kinase inhibitors that were considered chemical probes, which also complemented the results of experimental profiling. For putative chemical probes, very different target profiles were observed and the majority of these compounds were non-selective at different data confidence levels. Main findings of the analysis can be summarized as follows:

(i) Cell-based kinase inhibitor profiling and mining of available kinase activity data from medicinal chemistry was complementary and revealed similar trends.
(ii) In part, significant differences in promiscuity were detected for clinical kinase inhibitors.
(iii) The analysis revealed the importance of considering activity data extracted from databases at different confidence levels.
(iv) At data confidence level 1, application of an activity threshold significantly reduced PDs and balanced the view of kinase inhibitor promiscuity.

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

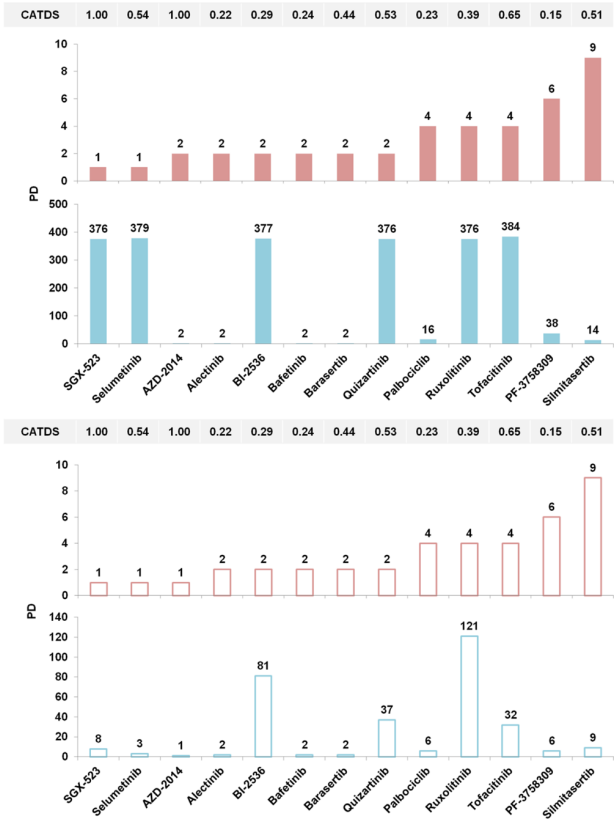


Figure 5. PD of chemical probes. Histograms show PD values (ChEMBL data) of a subset of 13 clinical kinase inhibitors (bottom) designated as chemical probes (top: confidence level 2, red background; level 1, blue background). In addition, histograms at the bottom show PD values of chemical probes after applying the <10 μM activity threshold (level 2, red outline; level 1, blue outline). For each inhibitor, the CATDS score is reported (gray background).

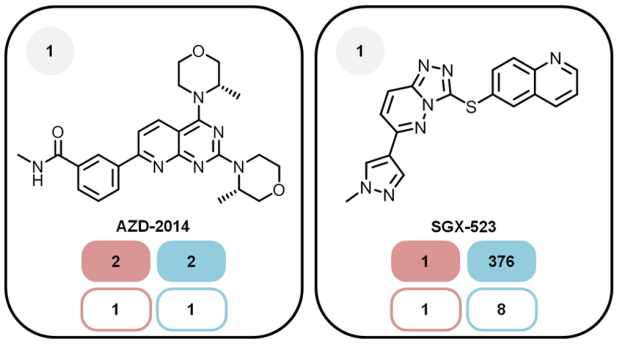


Figure 6. Exemplary chemical probes. Shown are two chemical probes (AZD-2014 and SGX-523) having the maximal CATDS score of 1 (gray background). For each inhibitor, the PD value on the basis of ChEMBL data is reported at confidence levels 2 (red background) and 1 (blue background) and after applying the <10 μM activity threshold (level 2, red outline; level 1, blue outline).

(v) Often assumed differences in selectivity between type I and II kinase inhibitors could not be confirmed by cell-based profiling or systematic compound data mining.

(vi) Even clinical kinase inhibitors regarded as chemical probes showed notable difference in PDs and contained a subset of highly promiscuous.

In summary, correlating the results of experimental profiling and compound data mining has further advanced our understanding of binding characteristics of currently most advanced kinase inhibitors, clearly showing that there are no simple relationships between clinical performance and selectivity versus promiscuity of these compounds. We conclude by emphasizing that the data made available by Klaeger et al. provide a rich source for different types of follow-up analysis. Herein, we have focused on compound selectivity, given the applicability domain of compound data mining. However, there are many more functional data provided by Klaeger et al. that can be further explored via other computational or experimental approaches.

AUTHOR INFORMATION

Corresponding Author

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

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Klaeger et al. are gratefully acknowledged for making their data publicly available.

REFERENCES

(1) *Kinase Drug Discovery*; Ward, R. A., Goldberg, F. W., Eds.; RSC: Cambridge, U.K., 2011.

(2) Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130−137.

(3) Simmons, D. L. Targeting Kinases: A New Approach to Treating Inflammatory Rheumatic Diseases. *Curr. Opin. Pharmacol.* **2013**, *13*, 426−434.

(4) Laufer, S.; Bajorath, J. New Frontiers in Kinases: Second Generation Inhibitors. *J. Med. Chem.* **2014**, *57*, 2167−2168.

(5) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, J.-M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A Small Molecule-Kinase Interaction Map for Clinical Kinase Inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329−336.

(6) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(7) Anastassiadis, 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.

(8) Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046−1051.

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

(9) Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the Kinome. *Nat. Chem. Biol.* **2011**, *7*, 200−202.
(10) Miduturu, C. V.; Deng, X.; Kwiatkowski, N.; Yang, W.; Brault, L.; Filippakopoulos, P.; Chung, E.; Yang, Q.; Schwaller, J.; Knapp, S.; King, R. W.; Lee, J.-D.; Herrgard, S.; Zarrinkar, P.; Gray, N. S. High-Throughput Kinase Profiling: A More Efficient Approach toward the Discovery of New Kinase Inhibitors. *Chem. Biol.* **2011**, *18*, 868−879.
(11) Elkins, J. M.; Fedele, V.; Szklarz, M.; Azeez, K. R. A.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X.-P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive Characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2016**, *34*, 95−103.
(12) Hu, Y.; Furtmann, N.; Bajorath, J. Current Compound Coverage of the Kinome. *J. Med. Chem.* **2015**, *58*, 30−40.
(13) Gavrin, L. K.; Saiah, E. Approaches to Discover Non-ATP Site Kinase Inhibitors. *Med. Chem. Commun.* **2013**, *4*, 41−51.
(14) Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9*, 1230−1241.
(15) Stumpfe, D.; Tinivella, A.; Rastelli, G.; Bajorath, J. Promiscuity of Inhibitors of Human Protein Kinases at Varying Data Confidence Levels and Test Frequencies. *RSC Adv.* **2017**, *7*, 41265−41271.
(16) Miljković, F.; Bajorath, J. Exploring Selectivity of Multikinase Inhibitors across the Human Kinome. *ACS Omega* **2018**, *3*, 1147−1153.
(17) Levitzki, A. Tyrosine Kinase Inhibitors: Views of Selectivity, Sensitivity, and Clinical Performance. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 161−185.
(18) Klaeger, S.; Heinzlmeir, S.; Wilhelm, M.; Polzer, H.; Vick, B.; Koenig, P.-A.; Reinecke, M.; Ruprecht, B.; Petzoldt, S.; Meng, C.; Zecha, J.; Reiter, K.; Qiao, H.; Helm, D.; Koch, H.; Schoof, M.; Canevari, G.; Casale, E.; Depaolini, S. R.; Feuchtinger, A.; Wu, Z.; Schmidt, T.; Rueckert, L.; Becker, W.; Huenges, J.; Garz, A.-K.; Gohlke, B.-O.; Zolg, D. P.; Kayser, G.; Vooder, T.; Preissner, R.; Hahne, H.; Tõnisson, N.; Kramer, K.; Götze, K.; Bassermann, F.; Schlegl, J.; Ehrlich, H.-C.; Aiche, S.; Walch, A.; Greif, P. A.; Schneider, S.; Felder, E. R.; Ruland, J.; Médard, G.; Jeremias, I.; Spiekermann, K.; Kuster, B. The Target Landscape of Clinical Kinase Inhibitors. *Science* **2017**, *358*, No. eaan4368.
(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) Arrowsmith, C. H.; Audia, J. E.; Austin, C.; Baell, J.; Bennett, J.; Blagg, J.; Bountra, C.; Brennan, P. E.; Brown, P. J.; Bunnage, M. E.; Buser-Doepner, C.; Campbell, R. M.; Carter, A. J.; Cohen, P.; Copeland, R. A.; Cravatt, B.; Dahlin, J. L.; Dhanak, D.; Edwards, A. M.; Frederiksen, M.; Frye, S. V.; Gray, N.; Grimshaw, C. E.; Hepworth, D.; Howe, T.; Huber, K. V. M.; Jin, J.; Knapp, S.; Kotz, J. D.; Kruger, R. G.; Lowe, D.; Mader, M. M.; Marsden, B.; Mueller-Fahrnow, A.; Müller, S.; O'Hagan, R. C.; Overington, J. P.; Owen, D. R.; Rosenberg, S. H.; Roth, B.; Ross, R.; Schapira, M.; Schreiber, S. L.; Shoichet, B.; Sundström, M.; Superti-Furga, G.; Taunton, J.; Toledo-Sherman, L.; Walpole, C.; Walters, M. A.; Willson, T. M.; Workman, P.; Young, R. N.; Zuercher, W. J. The Promise and Peril of Chemical Probes. *Nat. Chem. Biol.* **2015**, *11*, 536−541.
(21) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.

DOI: 10.1021/acsomega.8b00243
ACS Omega 2018, 3, 3113−3119

Summary

Systematic evaluation of clinical kinase inhibitors was performed on the basis of publicly available data. Our comparative analysis revealed consistent selectivity observations in cell-based and medicinal chemistry-driven approaches. Significant promiscuity differences were detected for a subset of clinical candidates. Consistent with this finding, sets of the most and least selective candidates from cell-based profiling studies revealed similar promiscuity trends, taking different data sources into account. Varying confidence levels of activity data strongly influenced selectivity profiles. Comparison of type I and type II inhibitors revealed no significant differences in selectivity. However, clinical candidates classified as chemical probes contained a subset of highly promiscuous representatives.

In the next study, we explore selectivity and off-target activities of designated chemical probes using activity data from different sources.

Chapter 5

Data-Driven Exploration of Selectivity and Off-Target Activities of Designated Chemical Probes

Introduction

Chemical probes must meet stringent selectivity requirements. The scientific community works frequently to revise requirements for high-quality probes. In spite of these efforts, their improper use and poor characterization represents an ongoing problem. For example, experts at Chemical Probes Portal establish ranks and commentaries to guide external investigators who select probes for targets of interest. Such recommendations would benefit from data-driven evaluation.

To complement expert views, we comprehensively analyzed highly curated probes from Chemical Probes Portal using the compound activity data from ChEMBL. Promiscuity of chemical probes was explored by applying activity data confidence levels of increasing stringency. Results were compared to those reported by Chemical Probes Portal. In addition, scaffold analysis and analog relationships were used to evaluate potential off-target activities.

molecules

MDPI

Article

Data-Driven Exploration of Selectivity and Off-Target Activities of Designated Chemical Probes

Filip Miljković and Jürgen Bajorath *

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; miljkovi@bit.uni-bonn.de

* Correspondence: bajorath@bit.uni-bonn.de; Tel.: +49-228-7369-100

Received: 31 August 2018; Accepted: 21 September 2018; Published: 23 September 2018

Abstract: Chemical probes are of central relevance for chemical biology. To unambiguously explore the role of target proteins in triggering or mediating biological functions, small molecules used as probes should ideally be target-specific; at least, they should have sufficiently high selectivity for a primary target. We present a thorough analysis of currently available activity data for designated chemical probes to address several key questions: How well defined are chemical probes? What is their level of selectivity? Is there evidence for additional activities? Are some probes "better" than others? Therefore, highly curated chemical probes were collected and their selectivity was analyzed on the basis of publicly available compound activity data. Different selectivity patterns were observed, which distinguished designated high-quality probes.

Keywords: chemical biology; bioactive compounds; chemical probes; target selectivity; promiscuity; molecular scaffold; off-target activity

1. Introduction

In 2003, the Human Genome Project was completed [1,2], which catalyzed biomedical research in an unprecedented manner. Among others, the emerging field of chemical biology was spurred on through the availability of annotated gene sequences, the development of new screening techniques, and advances in computational biology [3,4]. Many newly sequenced genes and their products became available for further exploration, which also opened the door for the identification of new targets for drug discovery [5–7]. Studying the function(s) of candidate targets in physiological environments and under pathological conditions became a primary objective for chemical biology. For this purpose, small molecules were used as chemical probes to specifically assess consequences of target intervention for biological functions and processes [8–10].

Chemical probes have stringent requirements. They must be capable of selectively binding to targets and modulating protein functions in their physiological context; ideally, probes should be target-specific. Furthermore, if they are used in a phenotypic context, it must ultimately be possible to deconvolute and identify targets that are responsible for an interesting biological readout [8,11–14]. These are challenging tasks. Not surprisingly, the scientific community is continuously revising and updating requirements for high-quality probes. One example is provided by a proposal of the Structural Genomics Consortium [15], according to which a high-quality probe should have higher than 100 nM potency against its designated target and exhibit greater than 30-fold selectivity for its primary target over other proteins belonging to the same family. Moreover, for phenotypic applications, a probe should display significant cellular on-target activity at 1 μM concentration. However, more often than not, such requirements are not met by candidate compounds considered as probes [13].

In addition to defining key requirements for chemical probes, the scientific community has set out to evaluate and rate probes and validate their use in cellular or in vivo model systems. For example, this is attempted through submission of candidate compounds to the Chemical Probes Portal [13,14], which recommends probes on the basis of ratings provided by their Scientific Advisory Board (SAB), consisting of experts in medicinal chemistry, chemical biology, or pharmacology. Guided by such ratings, an external investigator should be capable of choosing the best available probe for a target of interest and acquiring it through listed vendors [14].

Despite ongoing efforts to set high standards for chemical probes, the scientific community is still facing problems due to poor characterization of small molecule modulators, improper use of probes, and outdated recommendations [8,12,13]. The characterization and dissemination of probes will likely benefit from further support. For example, from a scientific viewpoint, a data-driven assessment of chemical probes is expected to complement expert views and experimental case studies, especially since compound activity data currently grow in an unprecedented manner.

In this study, we report a comprehensive analysis of highly curated chemical probes from the Chemical Probes Portal on the basis of compound activity data available in ChEMBL [16], the major public repository of data from medicinal chemistry. Promiscuity of probes was calculated at different data confidence levels and potency thresholds. Compound selectivity was investigated and compared to reports of the Chemical Probes Portal. Applying a scaffold concept [17], activities of chemical probes and structurally analogous bioactive compounds were compared and potential off-target activities of probes were further explored via network analysis [18]. Our findings are reported in the following.

2. Results and Discussion

2.1. Qualifying Chemical Probes

Table 1 shows target classes of chemical probes as reported by the Chemical Probes Portal. A total of 67 probes are listed, for which high-confidence activity data were available in ChEMBL and at least one activity annotation for a human target with a potency of $\leq$10 μM. Data confidence criteria and potency thresholds applied in our analysis are detailed in the Materials and Methods section. The 67 probes were assigned to six target classes. Almost half of these probes (33) were directed against protein kinases, followed by epigenetic probes (16).

Table 1. The table reports designated target classes of chemical probes from the Chemical Probes Portal, for which qualifying activity data were available in ChEMBL.

Target Class	Chemical Probes	Target-Based Categories
Protein kinases	33	Chemical probes for kinase targets
Lipid kinases	1	
Epigenetics	16	Chemical probes for non-kinase targets
Other post-translation modification proteins	13	
Other proteins	3	
Structural proteins	1	

Based on the classification in Table 1, the compounds were broadly divided into probes for kinases (34) and non-kinase targets (33). This was done because kinase inhibitors are of particular interest in chemical biology (as well as drug discovery), given the key role kinases play in many signaling pathways.

Protein kinases share an adenosine triphosphate (ATP) (cofactor)-binding site that is highly conserved across the human kinome [19]. The majority of currently available kinase inhibitors are type I inhibitors directed against the conserved ATP site, making target promiscuity among such inhibitors likely [20–22]. However, the presence of promiscuity cannot be assumed a priori because many type I inhibitors also display apparent selectivity for a given kinase over others [20,22]. Hence,

for characterizing kinase inhibitors used as chemical probes, exploring the interplay between assumed selectivity and potential promiscuity is of particular interest.

2.2. Selectivity Trends of Chemical Probes

For each chemical probe, the promiscuity degree (PD) was defined as the number of its unique targets on the basis of ChEMBL activity records or target annotations of the Chemical Probes Portal. Activity of a compound against multiple targets including unrelated targets is generally rationalized as promiscuity, whereas specificity implies exclusive activity against a single target. Furthermore, selectivity is best understood as activity against very few related targets, for example, a primary target and one or two others from the same family. Hence, formally it is difficult to draw a line between low levels of compound promiscuity and selectivity. However, since promiscuity also applies to increasingly large numbers of targets, it is advantageous to introduce the promiscuity degree as a measure of multi-target activity, rather than selectivity degree.

PDs were calculated on the basis of medium-confidence activity data (level 1; see Materials and Methods) and high-confidence data (level 2) applying two different potency thresholds ($\leq$10,000 nM and $\leq$100 nM; see Materials and Methods). In the following, the $\leq$10,000 nM threshold is referred to as $\leq$10μM. For each probe, four PDs were obtained by combining data confidence level 1 and 2 with the two potency thresholds, and a fifth value was calculated on the basis of target annotations provided by the Chemical Probes Portal. The results obtained for the 67 probes are reported in Figure 1.

Based on the information provided by the Chemical Probes Portal, probes were active against one to four targets, with on average 1.6 targets per probe. The majority of probes only had a single target annotation, consistent with proposed high-quality probe characteristics. On the basis of activity data from ChEMBL, a somewhat different picture emerged. For the data confidence level 2/$\leq$100 nM threshold combination, most probes retained their Portal-based PD value, yielding a similar average of 1.7 targets per probe. For the confidence level 2/$\leq$10 μM threshold combination, nearly half of the probes retained their PD. However, for others, an increase in promiscuity was detected, yielding an average of 2.6 targets per probe. In some cases, a PD > 5 was observed. These findings indicated that a subset of probes were at least weakly active against multiple targets.

Next, data confidence criteria were relaxed and PD values calculated at confidence level 1 applying a potency threshold $\leq$10 μM. In this case, about half of the qualifying probes also retained their Portal-based PD value. However, for other probes a substantial increase in promiscuity was observed, resulting in an average PD of 6.3 targets per probe. When applying the more rigorous $\leq$100 nM threshold at level 1, the mean PD decreased again to 2.2 targets per probe, revealing that reported weak activities at medium data confidence were largely responsible for the significant increase in the average PD.

Taken together, the results in Figure 1 show that proposed target selectivity of about 50% of the designated high-quality probes was not altered by taking medicinal chemistry data at varying confidence levels and potency thresholds into account. This was an encouraging finding, which also applied to kinase probes having an intrinsic likelihood of multi-kinase activity. By contrast, a significant increase in promiscuity was observed for another subset of probes. Table 2 reports 10 kinase probes that were annotated with both kinase and non-kinase targets when applying the confidence level 1/$\leq$10 μM threshold combination. Two exemplary kinase probes with a different degree of selectivity are shown in Figure 2, NVS-PAK1-1 [23] and ruxolitinib [24].

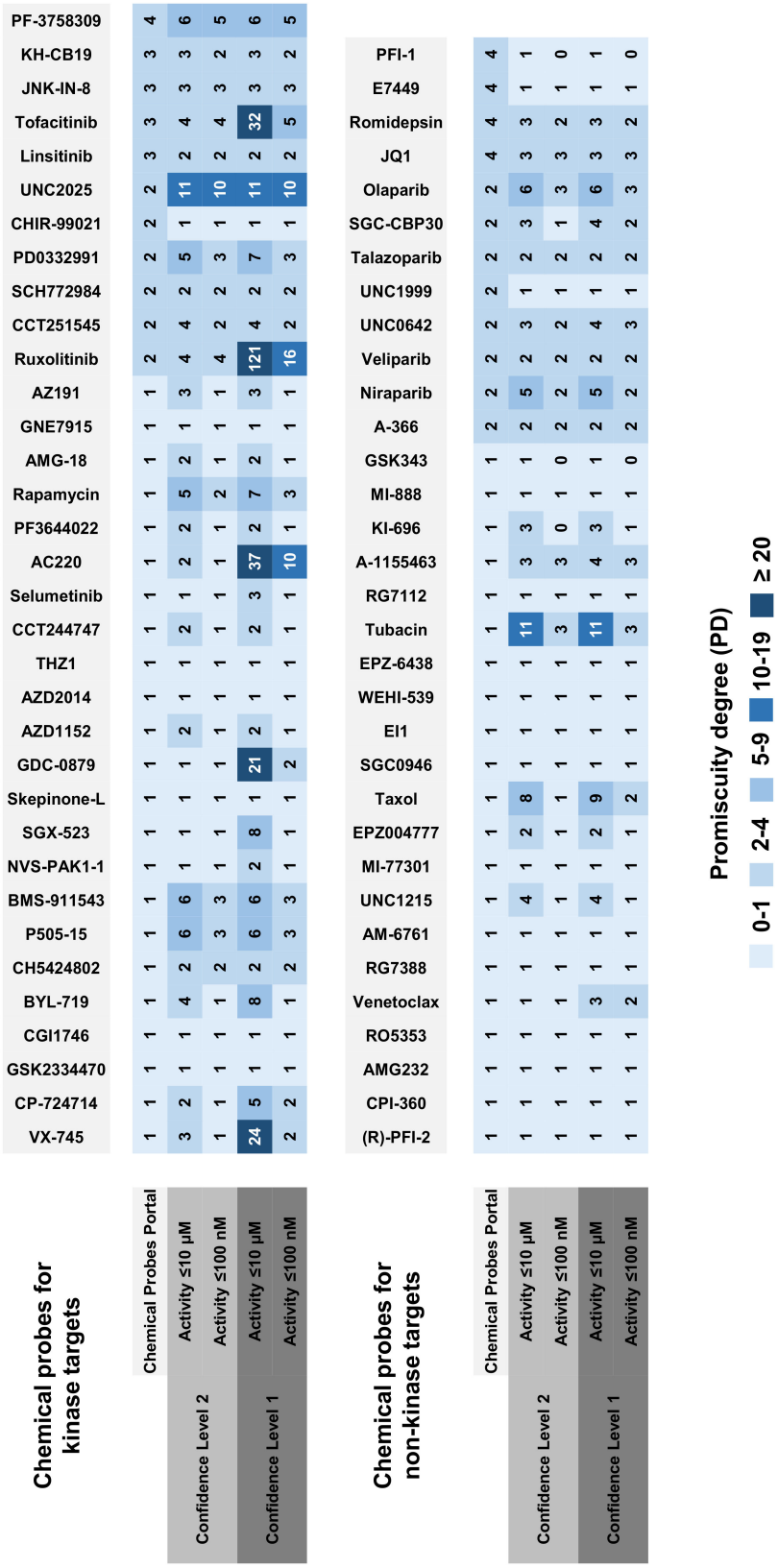


Figure 1. Promiscuity degree (PD) values of chemical probes are calculated on the basis of activity records from ChEMBL at two data confidence levels and for two potency thresholds and on the basis of target information from the Chemical Probes Portal. Probes for kinase and non-kinase targets are distinguished. Matrix cells represent PD values and are color-coded using a continuous spectrum ranging from light blue (0–1) to dark blue (≥20). "0" means that no target annotation is available for a given combination.

Table 2. Reported are kinase probes that are annotated in ChEMBL, with both kinase and non-kinase targets.

Chemical Probe	Confidence Level 2		Confidence Level 1	
	≤10 μM	≤100 nM	≤10 μM	≤100 nM
BMS-911543	6 (2) [1]	3	6 (2)	3
BYL-719	4 (2)	1	8 (3)	1
CCT244747	2 (1)	1	2 (1)	1
CCT251545	4 (1)	2 (1)	4 (1)	2 (1)
P505-15	6 (1)	3	6 (1)	3
PF3644022	2 (1)	1	2 (1)	1
Rapamycin	5 (4)	2 (1)	7 (6)	3 (2)
Ruxolitinib	4	4	121 (1)	16
SGX-523	1	1	8 (1)	1
VX-745	3 (2)	1	24 (2)	2

[1] Probes were selected if non-kinase targets were available for at least one confidence level/threshold combination. For each combination, promiscuity degree (PD) values are reported. If non-kinase targets are available, their number is given in parentheses. For example, "2 (1)" means that the probe is annotated with two targets, including one kinase target and "2" that the probe is annotated with two kinases (and no non-kinase target).

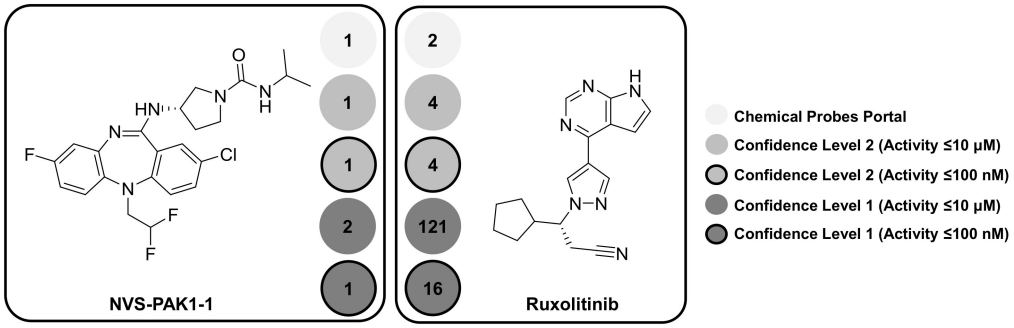


Figure 2. Shown are two exemplary kinase probes. For each probe, five PD values are reported applying different data selection criteria.

NVS-PAK1-1 [23] is an allosteric inhibitor of serine/threonine-protein kinase PAK1, as reported by the Chemical Probes Portal. Importantly, allosteric kinase inhibitors bind to regions outside the conserved ATP site and are thus expected to be more target-selective than type I inhibitors. At data confidence level 2, PAK1 was the only target of NVS-PAK1-1, regardless of the potency threshold. For the confidence level 1/≤10 μM threshold combination, only the closely related serine/threonine-protein kinase PAK2 was detected as an additional target. However, this was not the case when the ≤100 nM threshold was applied. Thus, on the basis of activity data analysis, NVS-PAK1-1 was a highly selective chemical probe, consistent with its allosteric mode-of-action and the Portal assessment.

Ruxolitinib [24] is an ATP-competitive pan-JAK inhibitor with JAK1 and JAK2 as designated primary targets. In this case, a different picture emerged. At confidence level 2 and both potency thresholds, activity of ruxolitinib was reported against two other members of the Janus kinase family, JAK3 and TYK2. Moreover, at confidence level 1, drastic increases in promiscuity were detected. At the ≤100 nM potency threshold, 16 kinase annotations were obtained and at the ≤10 μM threshold, a total of 121 targets were detected. These findings have two implications. First, promiscuity must be strictly considered in light of data confidence and potency criteria. For example, comparing the PD values obtained at confidence level 1 and 2, it is unlikely that ruxolitinib would be weakly active against more than 100 targets, and thus some of these annotations might well be false positive. Second, there was a clear difference in selectivity between NVS-PAK1-1 and ruxolitinib. Not unexpectedly, given its classification as an ATP site-directed pan-JAK inhibitor, ruxolitinib exhibited target promiscuity, and this well beyond the Janus kinase family, as revealed by our activity data analysis. Hence, the use

of ruxolitinib as a pan-JAK probe might be called into question, even if a number of target annotations detected at medium data confidence and, especially, low potency are false positive.

2.3. Chemical Probes and Historic Compounds

Chemical Probes Portal also reports a class of small molecules termed "historic compounds" [14]. As the name implies, many of these compounds were previously used as chemical probes, but considered to be obsolete or inferior to others at some point. Typically, historic compounds were found to be non-selective or not sufficiently potent to meet high-quality probe standards. For each historic compound, the Portal provides a rationale as to why it should not be further considered as a probe [14]. We reasoned that these historic compounds might present an interesting case for comparison with current probes.

For the confidence level 2 and 1/≤10 μM threshold combinations, activity annotations for 127 of the 164 historic compounds of Chemical Probes Portal were identified in ChEMBL, applying the same criteria as for chemical probes. For the level 2 and 1/≤100 nM threshold combinations, activity annotations were detected for 94 historic compounds. Figure 3 compares the distribution of PD values for chemical probes and historic compounds for all four combinations.

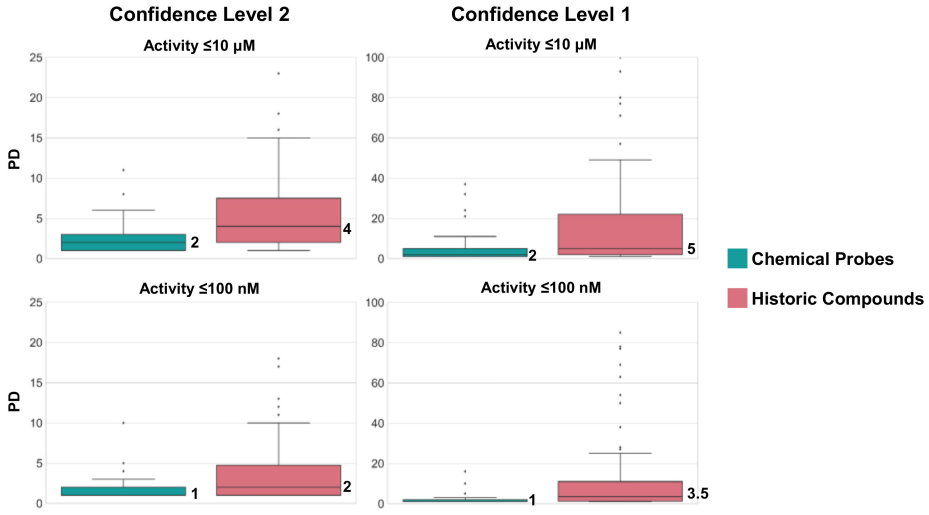


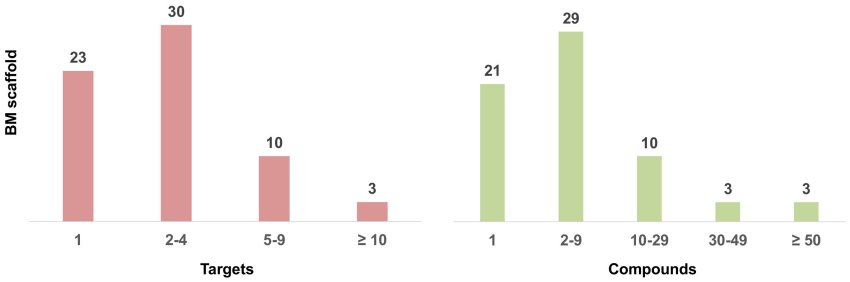
Figure 3. Boxplots report the distribution of PD values of qualifying chemical probes (green) and historic compounds (red) on the basis of ChEMBL data (top: Level 2 and 1, threshold ≤10 μM, chemical probes: 67, historic compounds: 127; bottom: Level 2 and 1, threshold ≤100 nM, chemical probes: 64; historic compounds: 94). Boxplots contain the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary), largest value (top line), and outliers (points below the bottom or above the top line).

In general, historic compounds had higher PD values than current chemical probes on the basis of available activity data, consistent with the Portal assessment. Hence, shortcomings of historic compounds were attributable to limited target selectivity.

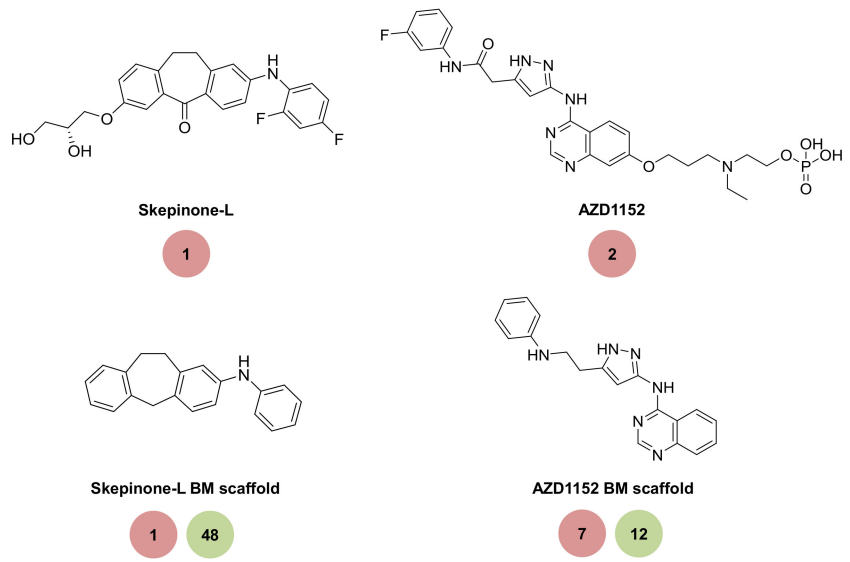
2.4. Scaffold Analysis of Chemical Probes

Applying the confidence level 2/≤10 μM threshold combination, 67 chemical probes, combined with qualifying 233,675 bioactive compounds from ChEMBL and Bemis-Murcko (BM) scaffolds [17] (see Materials and Methods), were extracted from this compound set. BM scaffolds represent molecular core structures and compounds sharing the same scaffold from a series of analogs. The 67 chemical probes yielded 66 unique BM scaffolds. For each probe, bioactive compounds sharing the same scaffolds were collected and their target annotations recorded. Target annotations of structural analogs assigned to probe scaffolds provide additional target hypotheses for probes (i.e., hints at off-target activities).

Figure 4a shows the distribution of target annotations (red) and ChEMBL compounds (green) over the 66 BM scaffolds extracted from probes. Only small numbers of bioactive compounds contained probe scaffolds. For 21 scaffolds, no additional ChEMBL compounds were identified (i.e., these scaffolds exclusively represented the probe). For 29 other scaffolds, a total of two to nine analogs (including the probe) were identified. Only six probe scaffolds represented 30 or more analogs. Thus, chemical probes frequently contained unique core structures. Since only limited numbers of analogs were detected for the majority of probes, the number of cumulative target annotations per probe scaffold was overall also small. The majority of scaffolds (53 of 66) were associated with one to four targets and only three scaffolds with 10 or more targets. Thus, "meta-level" promiscuity of chemical probe scaffolds was also low.

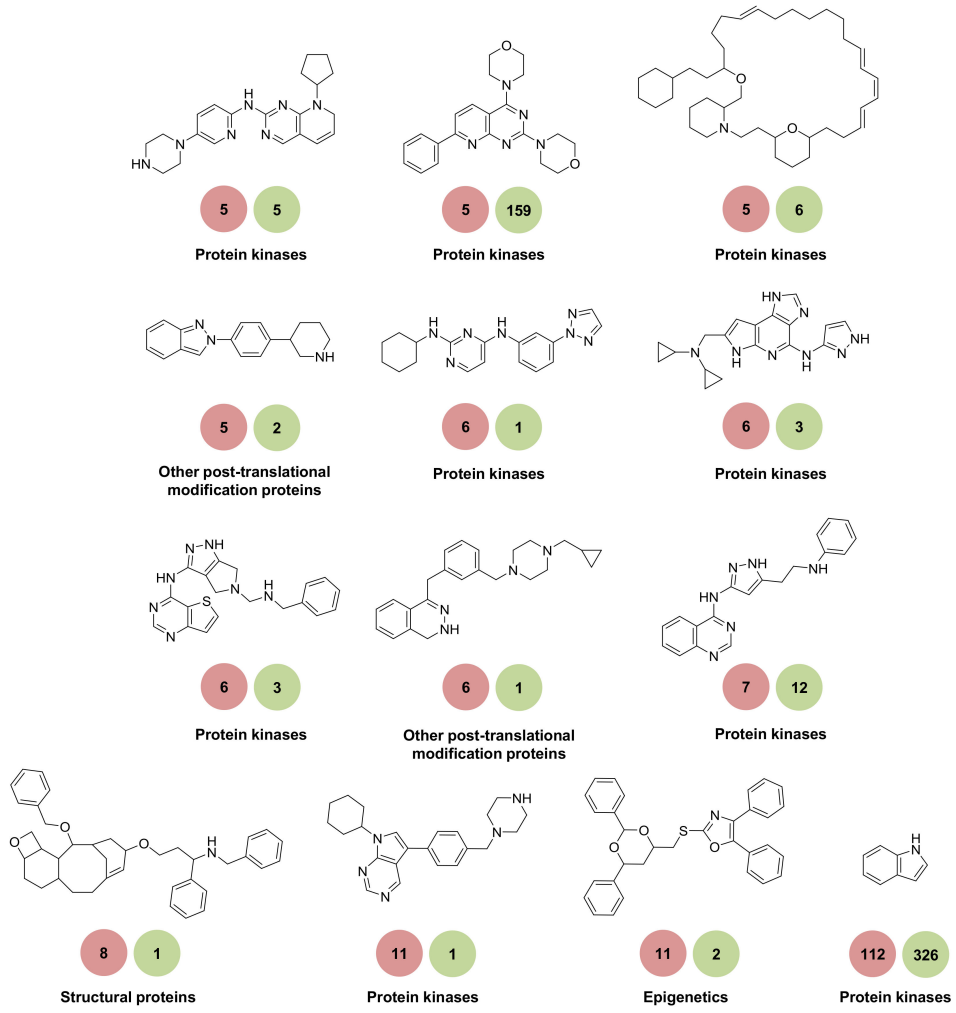


(**a**)



(**b**)

Figure 4. *Cont.*



(**c**)

Figure 4. (**a**) Bar graphs report the distributions of compounds and targets over probe scaffolds. (**b**) Kinase probes skepinone-L and AZD1152 are shown together with their BM scaffolds. For these probes and their scaffolds, promiscuity degrees are compared (red background). In addition, the total number of analogs represented by each probe scaffold is given (green). (**c**) Shown are 13 chemical probe scaffolds with highest "meta-level" promiscuity (i.e., associated with five or more targets). For these scaffolds, the promiscuity degrees (red background) and number of analogs (green) representing them are reported. In addition, classes of primary targets of compounds containing the scaffolds are given as reported by Chemical Probes Portal.

Figure 4b shows two kinase chemical probes and their BM scaffolds. Skepinone-L [25,26] is an ATP-competitive MAP kinase p38 alpha inhibitor with an unusual binding mode. Its scaffold represented a total of 48 analogs, all of which were exclusively annotated with MAP kinase p38 alpha. Hence, skepinone-L is another highly selective kinase probe. AZD1152 [27,28] is a phosphate-containing pro-drug that is rapidly converted in vivo into an active alcohol. Chemical Probes Portal reports the active form of AZD1152 as a selective inhibitor of serine/threonine kinase Aurora-B. In ChEMBL, an additional target was found for AZD1152. Moreover, although the scaffold of AZD1152 only represented 12 analogs—much less than the skepinone-L scaffold—these analogs were active against a total of seven targets, indicating that AZD1152 was likely less selective than proposed. Thus, skepinone-L and AZD1152 represent another example of designated high-quality probes with notable differences in selectivity revealed by activity data analysis, similar to NVS-PAK1-1 and ruxolitinib shown in Figure 2. The set of 13 structurally diverse scaffolds with highest "meta-level"

promiscuity is shown in Figure 4c. Compounds containing these scaffolds were active against a total of five or more targets.

2.5. Off-Target Activity Assessment in Networks

Going beyond scaffold analysis, similarity relationships between chemical probes and other bioactive compounds can also be explored on the basis of matched molecular pair (MMP) analysis [29]. An MMP is defined as a pair of compounds that are only distinguished by a chemical change at a single site. MMPs can be efficiently generated algorithmically. As another criterion of structural similarity [30], a chemical probe and bioactive compounds were classified as similar if they formed an MMP. Such structural relationships can be conveniently displayed in molecular networks in which nodes represent compounds and edges account for pairwise MMPs. If one assigns different node types to chemical probes and other bioactive compounds, a bipartite network is obtained, which can be further extended to a tripartite design by adding targets as a third node category. In the resulting tripartite network, edges between probes and analogs represent similarity relationships and edges between compounds and targets activity relationships. For chemical probes with bioactive analogs (connected by edges) additional targets associated with these analogs can be considered since it can be assumed that probes are likely to also be active against targets of structurally closely related compounds.

MMPs were obtained for 49 of 67 chemical probes and 738 bioactive analogs from ChEMBL. These compounds were active against a total of 135 targets. A tripartite network was constructed to capture all structural and activity relationships. The network revealed 47 previously unobserved probe-target associations involving a subset of 16 chemical probes and 40 targets from ChEMBL. New relationships can be studied in a structure-activity context by focusing on network neighborhoods of chemical probes. An example is provided in Figure 5, which shows the network neighborhood of CH5424802 [31], an ATP-competitive inhibitor of ALK tyrosine kinase, as reported by the Chemical Probes Portal, for which an additional target, RET tyrosine kinase, was identified in ChEMBL. CH5424802 had two close structural analogs with reported activity against the same and other kinases, thus providing additional target hypotheses for the chemical probe.

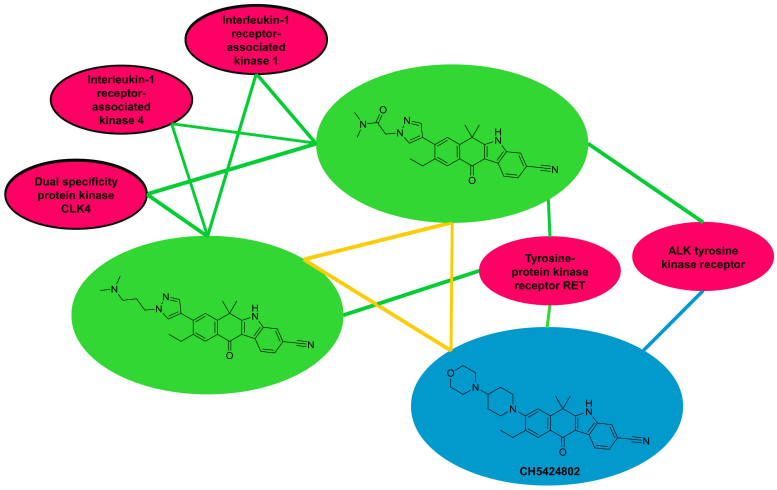


Figure 5. Shown is the neighborhood of a chemical probe (CH5424802) in a tripartite network. Blue and green nodes represent probes and bioactive analogs, respectively, and red nodes targets. Blue edges represent interactions between probes and targets from the Chemical Probes Portal, and green edges interaction between all compounds and targets from ChEMBL. In addition, yellow edges indicate MMP (similarity) relationships. Targets of analogs that provide additional hypotheses for the probe are encircled. This network should be considered "pseudo-tripartite" because the formation of edges (MMP relationships) is also permitted here between nodes belonging to the same category, departing from fundamental network theory. The network was drawn with Cytoscape 3.6.1. using the "organic layout" function [32].

2.6. Summary

Small molecular probes are of central importance to chemical biology. However, many currently investigated probes remain to be fully characterized. To these ends, important contributions are made by the Chemical Probes Portal, which carefully assesses candidate probes and prioritizes a set of highly curated chemical probes. Herein, we have further investigated designated high-quality probes by systematic analysis of available activity data for probes and closely related bioactive compounds. Our analysis adds another layer to the characterization of probes for chemical biology. Taking different data confidence and potency criteria into account, we show that ~50% of designated high-quality probes are target-selective when all available activity data are considered, consistent with expert curation; an encouraging finding. This applies to chemical probes directed against kinase or non-kinase targets. On the other hand, activity data analysis also differentiates between probes and identifies a subset of putative high-quality probes for which selectivity cannot be supported on the basis of currently available data, as summarized in Figure 1. These chemical entities might be deprioritized and should be used with caution when exploring biological functions and their origins. However, the analysis also emphasizes the presence of a variety of probes with striking selectivity—including kinase inhibitors—indicating that further progress in generating high-quality chemical probes can be anticipated, which will be exciting to follow.

3. Materials and Methods

3.1. Chemical Probes

Chemical probes were extracted from Chemical Probes Portal (accessed in July 2018) [13,14], which reports 189 small molecule modulators for applications in biomedical research. From this collection, only probes were selected that (i) were classified as inhibitors of a primary target, (ii) had non-ambiguous SMILES representations, (iii) were associated with ChEMBL identifiers (ChEMBL IDs) [16], and (iv) had a sufficiently high rating. The last selection criterion requires further explanation. Members of the Chemical Probes Portal SAB assign priority star ratings (1–4 stars) to probe candidates for their application in cellular and/or in vivo models. A star rating of 1 indicates that a candidate cannot be recommended as a probe, whereas a rating of 4 represents a high recommendation. Expert ratings for a candidate compound are averaged to obtain a final consensus rating. The Chemical Probes Portal endorses compounds as chemical probes only if their final rating reaches at least 3 stars [14]. Therefore, only candidate probes with a final rating of at least 3 stars were selected for our analysis. On the basis of selection criteria (i)–(iv), 80 highly curated probes qualified for our analysis.

3.2. Activity Data, Confidence Levels, and Historic Compounds

For selected chemical probes, available activity data for human targets were extracted from ChEMBL (release 24). Activity data were evaluated at two different confidence levels including level 1 (medium confidence) and level 2 (high confidence) according to [21]. For level 1, activity data with highest assay confidence were required, i.e., activity annotations were only selected from direct inhibition assays (ChEMBL assay relationship type "D") for single targets at the highest assay confidence level ("9"). For level 2, activity data with highest assay confidence plus highest measurement confidence were selected. Highest measurement confidence required the availability of numerically specified standard activity measurements (K_i or IC_{50} values with "=" standard relation), use of the nanomolar ("nM") activity unit, and presence of fully consistent "activity comments" in ChEMBL. To investigate selectivity characteristics of chemical probes, two potency thresholds were applied to activity data at confidence levels 1 and 2, i.e., $\leq$10,000 nM ($\leq$10μM) and $\leq$100 nM. For each chemical probe, PD values were calculated for each combination of a confidence level and potency threshold, yielding four PD values per probe from ChEMBL data. An additional PD value was calculated on the basis of target annotations reported by the Chemical Probes Portal. At confidence level

2 applying the ≤10 μM potency threshold, at least one activity record for a human target was obtained for 67 of 80 pre-selected chemical probes. These 67 probes provided our basis set for subsequent analysis. We also searched ChEMBL for 164 historic compounds designated by the Chemical Probes Portal on the basis of the same criteria applied to chemical probes.

3.3. Bioactive Compounds, Scaffold Analysis, and Off-Target Predictions

Scaffold analysis of chemical probes was performed applying the Bemis-Murcko (BM) scaffold concept [17]. BM scaffolds are extracted from compounds by eliminating all R-groups while retaining ring systems and linker moieties connecting rings. So-defined scaffolds were derived from all chemical probes and bioactive compounds for which target annotation(s) were available at confidence level 2 applying the ≤10 μM potency threshold. Potential off-target activities of chemical probes were also analyzed using a tripartite network data structure [18].

Author Contributions: F.M. and J.B. conceived the study and designed the experiments; F.M. performed the experiments; F.M. and J.B. analyzed the data and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Human Genome Sequencing Consortium. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409*, 860–921. [CrossRef] [PubMed]
2. International Human Genome Sequencing Consortium. Finishing the Euchromatic Sequence of the Human Genome. *Nature* **2004**, *431*, 931–945. [CrossRef] [PubMed]
3. Schenone, M.; Dančík, V.; Wagner, B.K.; Clemons, P.A. Target Identification and Mechanism of Action in Chemical Biology and Drug Discovery. *Nat. Chem. Biol.* **2013**, *9*, 232–240. [CrossRef] [PubMed]
4. Cornish, P.V.; Ha, T. A Survey of Single-Molecule Techniques in Chemical Biology. *ACS Chem. Biol.* **2007**, *2*, 53–61. [CrossRef] [PubMed]
5. Knowles, J.; Gromo, G. Target Selection in Drug Discovery. *Nat. Rev. Drug Discov.* **2003**, *2*, 63–69. [CrossRef] [PubMed]
6. Edwards, A.M.; Isserlin, R.; Bader, G.D.; Frye, S.V.; Willson, T.M.; Yu, F.H. Too Many Roads Not Taken. *Nature* **2011**, *470*, 163–165. [CrossRef] [PubMed]
7. Oprea, T.I.; Bologa, C.G.; Brunak, S.; Campbell, A.; Gan, G.N.; Gaulton, A.; Gomez, S.M.; Guha, R.; Hersey, A.; Holmes, J.; et al. Unexplored Therapeutic Opportunities in the Human Genome. *Nat. Rev. Drug Discov.* **2018**, *17*, 317–332. [CrossRef] [PubMed]
8. Bunnage, M.E.; Chekler, E.L.P.; Jones, L.H. Target Validation Using Chemical Probes. *Nat. Chem. Biol.* **2013**, *9*, 195–199. [CrossRef] [PubMed]
9. Jones, L.H.; Bunnage, M.E. Applications of Chemogenomic Library Screening in Drug Discovery. *Nat. Rev. Drug Discov.* **2017**, *16*, 285–296. [CrossRef] [PubMed]
10. Simon, G.M.; Niphakis, M.J.; Cravatt, B.F. Determining Target Engagement in Living Systems. *Nat. Chem. Biol.* **2013**, *9*, 200–205. [CrossRef] [PubMed]
11. Frye, S.V. The Art of the Chemical Probe. *Nat. Chem. Biol.* **2010**, *6*, 159–161. [CrossRef] [PubMed]
12. Workman, P.; Collins, I. Probing the Probes: Fitness Factors for Small Molecule Tools. *Chem. Biol.* **2010**, *17*, 561–577. [CrossRef] [PubMed]
13. Arrowsmith, C.H.; Audia, J.E.; Austin, C.; Baell, J.; Bennett, J.; Blagg, J.; Bountra, C.; Brennan, P.E.; Brown, P.J.; Bunnage, M.E.; et al. The Promise and Peril of Chemical Probes. *Nat. Chem. Biol.* **2015**, *11*, 536–541. [CrossRef] [PubMed]
14. Chemical Probes Portal. Available online: http://www.chemicalprobes.org/ (accessed on 14 July 2018).
15. Structural Genomics Consortium. Available online: https://www.thesgc.org/ (accessed on 22 August 2018).
16. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef] [PubMed]
17. Bemis, G.W.; Murcko, M.A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. [CrossRef] [PubMed]

18. Kunimoto, R.; Bajorath, J. Design of a Tripartite Network for the Prediction of Drug Targets. *J. Comput. Aided. Mol. Des.* **2018**, *32*, 321–330. [CrossRef] [PubMed]
19. Manning, G.; Whyte, D.B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934. [CrossRef] [PubMed]
20. Miljković, F.; Bajorath, J. Exploring Selectivity of Multikinase Inhibitors across the Human Kinome. *ACS Omega* **2018**, *3*, 1147–1153. [CrossRef] [PubMed]
21. Miljković, F.; Bajorath, J. Reconciling Selectivity Trends from a Comprehensive Kinase Inhibitor Profiling Campaign with Known Activity Data. *ACS Omega* **2018**, *3*, 3113–3119. [CrossRef] [PubMed]
22. Miljković, F.; Bajorath, J. Evaluation of Kinase Inhibitor Selectivity Using Cell-Based Profiling Data. *Mol. Inform.* **2018**, *37*, 1800024. [CrossRef] [PubMed]
23. Karpov, A.S.; Amiri, P.; Bellamacina, C.; Bellance, M.-H.; Breitenstein, W.; Daniel, D.; Denay, R.; Fabbro, D.; Fernandez, C.; Galuba, I.; et al. Optimization of a Dibenzodiazepine Hit to a Potent and Selective Allosteric PAK1 Inhibitor. *ACS Med. Chem. Lett.* **2015**, *6*, 776–781. [CrossRef] [PubMed]
24. Quintás-Cardama, A.; Vaddi, K.; Liu, P.; Manshouri, T.; Li, J.; Scherle, P.A.; Caulder, E.; Wen, X.; Li, Y.; Waeltz, P.; et al. Preclinical Characterization of the Selective JAK1/2 Inhibitor INCB018424: Therapeutic Implications for the Treatment of Myeloproliferative Neoplasms. *Blood* **2010**, *115*, 3109–3117. [CrossRef] [PubMed]
25. Koeberle, S.C.; Fischer, S.; Schollmeyer, D.; Schattel, V.; Grütter, C.; Rauh, D.; Laufer, S.A. Design, Synthesis, and Biological Evaluation of Novel Disubstituted Dibenzosuberones as Highly Potent and Selective Inhibitors of p38 Mitogen Activated Protein Kinase. *J. Med. Chem.* **2012**, *55*, 5868–5877. [CrossRef] [PubMed]
26. Koeberle, S.C.; Romir, J.; Fischer, S.; Koeberle, A.; Schattel, V.; Albrecht, W.; Grütter, C.; Werz, O.; Rauh, D.; Stehle, T.; et al. Skepinone-L is a Selective p38 Mitogen-activated Protein Kinase Inhibitor. *Nat. Chem. Biol.* **2012**, *8*, 141–143. [CrossRef] [PubMed]
27. Mortlock, A.A.; Foote, K.M.; Heron, N.M.; Jung, F.H.; Pasquet, G.; Lohmann, J.-J.M.; Warin, N.; Renaud, F.; De Savi, C.; Roberts, N.J.; et al. Discovery, Synthesis, and In Vivo Activity of a New Class of Pyrazoloquinazolines as Selective Inhibitors of Aurora B Kinase. *J. Med. Chem.* **2007**, *50*, 2213–2224. [CrossRef] [PubMed]
28. Yang, J.; Ikezoe, T.; Nishioka, C.; Tasaka, T.; Taniguchi, A.; Kuwayama, Y.; Komatsu, N.; Bandobashi, K.; Togitani, K.; Koeffler, H.P.; et al. AZD1152, a Novel and Selective Aurora B Kinase Inhibitor, Induces Growth Arrest, Apoptosis, and Sensitization for Tubulin Depolymerizing Agent or Topoisomerase II Inhibitor in Human Acute Leukemia Cells In Vitro and In Vivo. *Blood* **2007**, *110*, 2034–2040. [CrossRef] [PubMed]
29. Kenny, P.W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Wiley-Blackwell: Hoboken, NJ, USA, 2005; pp. 271–285.
30. 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. [CrossRef] [PubMed]
31. Sakamoto, H.; Tsukaguchi, T.; Hiroshima, S.; Kodama, T.; Kobayashi, T.; Fukami, T.A.; Oikawa, N.; Tsukuda, T.; Ishii, N.; Aoki, Y. CH5424802, a Selective ALK Inhibitor Capable of Blocking the Resistant Gatekeeper Mutant. *Cancer Cell* **2011**, *19*, 679–690. [CrossRef] [PubMed]
32. Smoot, M.E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T. Cytoscape 2.8: New Features for Data Integration and Network Visualization. *Bioinformatics* **2011**, *27*, 431–432. [CrossRef] [PubMed]

Sample Availability: Samples of the compounds are not available from the authors.

Summary

We carried out a data-driven evaluation of designated chemical probes from Chemical Probes Portal by taking ChEMBL activity data into the account. Different data confidence levels suggested that ∼50% of high-quality probes were selective for their targets. On the other hand, a subset of compounds was detected for which selectivity requirements cannot be supported. These modulators may be deprioritized or used with caution in chemical biology settings.

Previous findings show that kinase inhibitors, including clinical candidates and chemical probes, often contain different selectivity profiles. Kinase inhibitor analogs with significant differences in promiscuity were not yet explored on large scale.

Our next goal was the identification and analysis of pairs of structural analogs with large differences in promiscuity.

Chapter 6

Computational Analysis of Kinase Inhibitors Identifies Promiscuity Cliffs across the Human Kinome

Introduction

PCs are defined as pairs of structurally analogous compounds with large differences in number of annotated targets. This data structure is useful to explore structure-promiscuity relationships and derive additional target hypotheses for close structural analogs of extensively tested compounds. PCs were previously detected for inhibitors of the human kinome. We collected and curated kinase-related activity data from seven public databases and consolidated them into a new kinase data set for subsequent analyses.

This data set was used to perform promiscuity analysis and systematically search for structural analogs forming PCs. Obtained PCs were organized in network representations and PC pathways connecting individual PC pairs were further evaluated.

ACS OMEGA

Article

Cite This: ACS Omega 2018, 3, 17295−17308

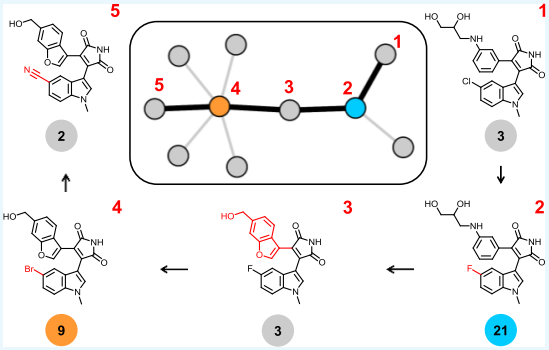
http://pubs.acs.org/journal/acsodf

Computational Analysis of Kinase Inhibitors Identifies Promiscuity Cliffs across the Human Kinome

Filip Miljković and Jürgen Bajorath*

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

ABSTRACT: Kinase inhibitors are high-priority drug candidates for a variety of therapeutic applications. Accordingly, there has been a rapid growth in the number of kinase inhibitors and volumes of associated activity data. A paradigm for the use of kinase inhibitors in oncology is that these compounds have multitarget activities and elicit their therapeutic effects through polypharmacology. An analysis of kinase inhibitors and associated activity data from medicinal chemistry has so far only identified small subsets of highly promiscuous kinase inhibitors. In this study, we have collected inhibitors of human kinases and their activity data from seven public repositories, curated, and combined these data, yielding more than 112 000 inhibitors with well-defined activity measurements from which qualitative target annotations were derived. An analysis of these unprecedentedly large data sets revealed that nearly 40% of human kinase inhibitors have multikinase activities but that only 4% are known to be active against five or more kinases. However, structurally analogous inhibitors often displayed significant differences in the number of kinase annotations, leading to the formation of nearly 16 000 "promiscuity cliffs". Moreover, 2236 promiscuity cliffs (14.03%) were formed by kinase inhibitors at different stages of clinical development. Overall, these cliffs suggested many target hypotheses for kinase inhibitors, taking data incompleteness into consideration, as well as hypotheses for structural modifications leading to kinase selectivity. Furthermore, from network representations, pathways comprising sequences of promiscuity cliffs were extracted that revealed unexpected structure−promiscuity relationships. To enable follow-up investigations, all promiscuity cliffs formed by human kinase inhibitors will be made freely available.



INTRODUCTION

Kinase inhibitors play a major role in drug discovery.[1,2] Originally, kinase inhibitors were successfully applied in oncology, where their therapeutic efficacy was found to be largely due to polypharmacology.[3,4] However, the clinical use of kinase inhibitors has been further expanded to other therapeutic areas such as immunology and inflammation or metabolic diseases, where target selectivity of inhibitors plays an important role.[5−7] It is thus not surprising that the topic of kinase inhibitor selectivity versus promiscuity has been intensely investigated over the past decade and continues to be a much debated issue.[8−13] Selectivity analysis is far from being a simple task, given the binding characteristics of kinase inhibitors and the many experimental variables that need to be considered. Furthermore, apparent promiscuity of compounds including kinase inhibitors is often associated with undesired effects such as artifacts resulting from assay interference.[14−16] However, promiscuity also refers to the presence of true multitarget activities of compounds that represent the molecular basis of polypharmacology.[17,18]

A mapping of signature fragments of kinase inhibitors adopting different binding modes revealed by X-ray crystallography[19,20] has shown that more than 95% of the currently available inhibitors of human kinases are type I inhibitors.[21,22] These inhibitors block the adenosine triphosphate (ATP) cofactor binding site that is largely conserved across the kinome, so they are expected to be promiscuous.[21] Subsets of highly promiscuous kinase inhibitors have indeed been identified including anticancer drugs,[22] consistent with their polypharmacology. Promiscuous kinase inhibitors include approved drugs as well as inhibitors at different stages of clinical development. A representative example is provided by sunitinib, a multikinase type I inhibitor, whose targets include vascular endothelial growth factor receptor 2 and platelet-derived growth factor receptor β kinases. Polypharmacology associated with sunitinib and various other promiscuous kinase inhibitors has proven essential for their efficacy in cancer treatment.[2−4] On the other hand, analyses of the available kinase inhibitors and associated activity data have not supported the often assumed general promiscuity of these inhibitors. For example, in the first large-scale analysis, 18 653 publicly available inhibitors with activity against 266 human kinases were identified for which high-confidence activity data were available.[22] On the basis of K_i and IC_{50} measurements, 68 and 77% of all the inhibitors were only

Received: October 29, 2018
Accepted: December 3, 2018
Published: December 13, 2018

annotated with a single human kinase, respectively, and only ∼1% of the inhibitors were active against five or more kinases.[22] Two years later, the number of human kinase inhibitors with available high-confidence data had more than doubled and 43 331 inhibitors with activity against 286 human kinases were available.[23] However, despite the rapid growth in kinase inhibitors, 76.5% of all the inhibitors were only annotated with a single kinase on the basis of combined K_i and IC_{50} measurements; again, only ∼1% of all the inhibitors had reported activity against five or more kinases.[23] There is the possibility that the dominance of kinase inhibitors with single target annotations is at least partly due to data incompleteness[24] because only a confined subset of kinase inhibitors have been subjected to kinome profiling. On the other hand, the mean promiscuity degree (PD) of ∼1.5 determined by activity data analyses[22,23] did not significantly increase when the activity data confidence criteria were gradually relaxed and increasing numbers of activity measurements considered or primary kinase screening assays were analyzed. The PD is defined as the number of unique kinase annotations available for an inhibitor and serves as a qualitative measure of compound promiscuity. The PD is derived on the basis of well-defined activity measurements including, among others, (assay-dependent) IC_{50} and (assay-independent) K_i values. Although potency values reported on the basis of different types of measurements should not be directly compared, they are qualified to serve as sources for target annotations.

In the light of the findings discussed above, large-scale activity data analysis did not provide support for the view that ATP site-directed kinase inhibitors might generally be promiscuous. Moreover, kinase inhibitor activity profiles are multifaceted. For example, within the subset of promiscuous kinase inhibitors, in part strong target selectivity tendencies for individual kinases were detected, resulting from differential potency for multiple kinases.[25−27] In this context, it should also be noted that large-scale analyses of kinase inhibitor activity data reported so far were exclusively[22,23,26] or mostly[25,27] based on ChEMBL,[28] the major public repository for compounds and activity data from medicinal chemistry. Hence, one might consider revisiting kinase inhibitor analysis by integrating data from different repositories that have become available over time. The number of available kinase inhibitors and volumes of associated activity data steadily grow, which reflects intense efforts to advance inhibitors to preclinical and clinical development in different therapeutic areas. However, there is no simple correlation between increasing amounts of available data and clinical advancements in the kinase inhibitor field, especially because requirements for kinase inhibitors considered for different therapeutic applications depart from standards established in drug discovery including, first and foremost, kinase promiscuity and ensuing polypharmacology.[5,6]

The promiscuity cliff (PC) data structure was introduced previously to explore the structural basis of multitarget activities of small molecules.[29−31] A PC is defined as a pair of structurally analogous compounds that have a large difference in the number of target annotations.[29] PCs have been identified in screening libraries[29] and compound sets from ChEMBL including kinase inhibitors.[30] The PC data structure is useful for exploring structure−promiscuity relationships and deriving additional target hypotheses for structural analogues of extensively tested kinase inhibitors, especially those that have advanced to the clinic or have been approved as drugs. Structurally related compounds have often not been extensively tested. Therefore, PCs involving such inhibitors immediately suggest follow-up experiments, given likely data incompleteness.

Herein, we extend the systematic analysis of kinase inhibitors and their promiscuity on the basis of inhibitors and activity data that were selected from different source databases and combined, yielding unprecedented coverage of the human kinome. The analysis was combined with a systematic assessment of the PCs formed by human kinase inhibitors and PC pathways extracted from network representations.

MATERIALS AND METHODS

General Compound Selection Criteria. Compounds were represented as canonical SMILES.[32] The following selection criteria were generally applied:

(1) Only inhibitors of human kinases having UniProt[33] IDs were selected.
(2) The permitted molecular weight range was [200, 900] Da.
(3) Potency had to be reported using a standard concentration or constant (such as IC_{50}, K_i, or K_d) and a numerically specified value with standard unit (such as units μM, nM, or pM). All potency measurements were recorded as the negative decadic logarithm.
(4) A potency threshold of 10 μM was applied (pPOT ≥ 5).
(5) If multiple potency values were reported for the same kinase, the highest value was selected.
(6) Each kinase annotation of an inhibitor was recorded as a separate "interaction".

Source Databases and Data Curation. Databases were accessed in September 2018, except PubChem, which was accessed in June 2017. The following database-specific curation and selection criteria were applied.

ChEMBL. From ChEMBL[28] release 24, human kinase inhibitors were selected if inhibition of single kinases (target type "SINGLE PROTEIN") in direct interaction assays (relationship type "D") at the highest level of confidence (confidence score "9") was reported using the standard activity relationship "=". In addition, consistent activity records were required (e.g., excluding compounds designated as "active", "inactive", and/or "inconclusive" in the same record).

PubChem. From PubChem,[34,35] primary, confirmatory, and panel assays for human kinases were obtained that reported potency measurements with μM or nM activity units. PubChem's target GI numbers were mapped to the corresponding UniProt IDs. Only compounds with a consistent designation as active with standard relationship "=" for a human kinase assay or across different assays for the same kinase were considered.

Probes and Drugs Portal. The Probes and Drugs Portal combines activity data from ∼50 different sources.[36] Human kinases from the Portal data were mapped to UniProt IDs. Kinase inhibitors with potency measurements such as pIC_{50}, pK_i, or pK_d were selected.

BindingDB. From BindingDB,[37] inhibitors of human kinases with available pIC_{50}, pK_i, pK_d, or pEC_{50} were selected.

PDBbind. Protein−ligand complexes in PDBbind[38] were filtered for human kinases with UniProt IDs and associated PDB[39] codes for single targets. Reported compound activity measurements included IC_{50}, K_d, and K_i values with standard relationship. Because PDBbind only provides PDB codes for complexes, these codes were searched for matches in the KLIFS database[40] from which the corresponding inhibitors were obtained.

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

ProteomicsDB. Results of a profiling study of clinical kinase inhibitors[41] have been made available in ProteomicsDB.[42] From this data set, measurements designated as "high confidence" were selected, given as K_d values with standard relationship, yielding 215 human kinase inhibitors.

Drug Target Commons. Compound annotations with human kinases in the Drug Target Commons database[43] were filtered for UniProt IDs, standard relationship "=", and single-target assays. The database refers to compounds using their ChEMBL IDs. Therefore, qualifying kinase inhibitors were retrieved from ChEMBL.

Unifying Kinase Inhibitor Data from Different Sources. To combine data from different sources, assemble unique inhibitors, and evaluate compound sourcing, ChEMBL was used as a reference database. Compounds selected from all databases were mapped to ChEMBL and the overlap was determined. Then, inhibitors not contained in other databases were extracted from ChEMBL by applying the selection criteria specified above. Finally, it was determined how many unique human kinase inhibitors were obtained from each database. Unification of kinase data from different sources is generally hindered by the application of different assay systems, activity detection technologies, and experimental conditions such as varying ATP concentrations in the assays. These variables typically lead to different activity read-outs. Moreover, inconsistencies in data curation may lead to further bias in judging and comparing activity profiles. Therefore, we analyzed our data selection for potential inconsistencies in activity data across different data sources. Moreover, only 4.9% of the interactions had activity variations exceeding one order of magnitude, thus lending credence to the data curation and selection process. In this limited number of cases, for formal consistency, the highest reported activity value was selected and recorded to establish an interaction.

Alerts for Pan-Assay Interference Compounds (PAINS). For PC analysis, kinase inhibitors were screened for pan-assay interference compounds (PAINS)[14,15] using three public filters available in ChEMBL,[28] RDKit,[44] and ZINC.[45] Although it is by no means certain that compounds containing PAINS substructures will cause assay interference and activity artifacts,[46,47] excluding potential false-positives is of critical relevance for defining PCs. This is the case because single-assay interference compounds with artificial target annotations might give rise to many incorrect PCs. Therefore, kinase inhibitors with PAINS alerts were excluded from PC analysis.

Promiscuity Cliffs. PCs formed by human kinase inhibitors were identified by systematically searching for transformation size-restricted matched molecular pairs (MMPs).[48] An MMP is defined as a pair of compounds that are only distinguished by a chemical modification at a single site,[49,50] termed a transformation.[50] The MMPs were then screened for a participating inhibitor with a PD of 1−4 and another inhibitor with a larger PD value, yielding a PD difference (ΔPD) of 5 or more. The MMPs meeting these PD/ΔPD conditions were classified as PCs.

PC networks in which nodes represent PC compounds and edges pairwise PCs were generated with Cytoscape.[51] Furthermore, phylogenetic trees of the human kinome[52] were drawn with KinMap.[53]

RESULTS AND DISCUSSION

Human Kinase Inhibitors. A total of 112 624 unique inhibitors were identified that were active against 426 human kinases (pPOT ≥ 5). These inhibitors covered 82.2% of the human kinome (518 kinases)[52] and formed a total of 234 740 unique compound−kinase interactions. For 97.2% of these interactions, only one type of potency measurement (e.g., K_i) was available. Furthermore, IC_{50}, K_i, and K_d values represented 96.4% of all potency measurements, with IC_{50} representing two thirds of the data (67.5%), followed by K_i (22.3%) and K_d (6.2%) values. When a more stringent potency threshold of 100 nM (pPOT ≥ 7) was applied to this set, 69 774 inhibitors were obtained that covered 408 human kinases.

Our previous analysis of kinase inhibitors[23] was exclusively based on ChEMBL and ChEMBL-specific data selection criteria. Here, the scope of compound and activity data analysis was expanded and data curation and selection criteria were balanced to cover seven databases. To evaluate compound selection, ChEMBL release 24 was used as a reference database and found to contain 83 647 of the 112 624 inhibitors (74.3%). These compounds had at least one human kinase annotation in ChEMBL, not taking data confidence criteria into consideration.

Table 1 reports the number of inhibitors and interactions that were uniquely contributed by individual databases. A subset of 3457 inhibitors was only present in ChEMBL, but no other database. With 27 277 compounds, BindingDB provided by far the largest fraction of unique inhibitors, which formed 44 547 interactions. BindingDB was followed by ChEMBL and PubChem (444 unique inhibitors). In total, 31 753 inhibitors originated from a single source database. In addition to uniquely contributed compounds, 681 inhibitors not contained in ChEMBL were shared by two or more other databases. The consolidated set of 112 624 human kinase inhibitors provided the basis for our subsequent analysis.

Table 1. Unique Inhibitors and Interactions Originating from Different Databases[a]

no	database	unique inhibitors	unique interactions
1	ChEMBL	3457	6807
2	PubChem	444	471
3	Probes and Drugs Portal	188	1971
4	BindingDB	27 277	44 547
5	PDBbind	365	380
6	ProteomicsDB	6	126
7	Drug Target Commons	16	16
	∑	31 753	54 318

[a]The table reports the number of inhibitors and compound−kinase interactions that were uniquely contributed by each database. Taken together, 31 753 inhibitors originated from only one of the source databases. The remaining 80 871 of the total of 112 624 qualifying inhibitors were shared by two or more databases.

Promiscuity Analysis. For promiscuity analysis, each defined compound−kinase interaction yielded an individual target annotation for an inhibitor, whose sum gave its PD. In addition, for each inhibitor with multikinase activity, the "nanomolar ratio" was determined as the proportion of nM relative to (nM + μM) potency measurements. The so-defined nM ratio served as a measure of the strength and relevance of interactions involving promiscuous kinase inhibitors.

Figure 1 shows the distribution of PD values of the 112 624 human kinase inhibitors. Majority of inhibitors (61%) only had a single kinase annotation. More than a third of the inhibitors had known activity against two to four kinases, whereas only 4% were active against five or more kinases (4510 inhibitors). Among

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

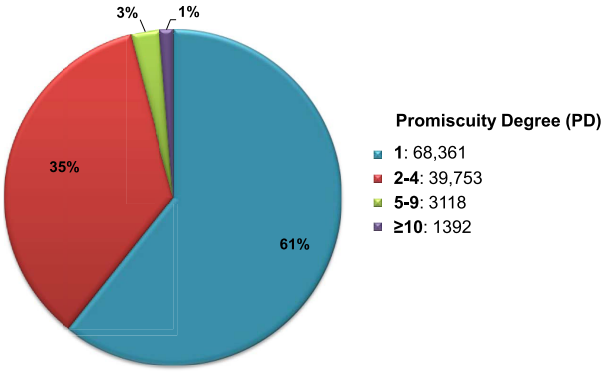


Figure 1. Distribution of promiscuity degrees. A pie chart shows the distribution of PD values for the set of 112 624 human kinase inhibitors.

these, 1% (1392 inhibitors) had 10 or more kinase annotations, thus representing the subset of most promiscuous inhibitors across the human kinome. With mean and median PD of 2.1 and 1.0, respectively, kinase inhibitor promiscuity was overall only slightly higher across 426 human kinases than indicated by a mean PD of 1.5, which was previously determined on the basis of 43 331 inhibitors with high-confidence activity data for 286 human kinases that exclusively originated from ChEMBL.[23] Thus, although our current analysis was based on many more compounds and a much larger kinome coverage, the assessment of global promiscuity among kinase inhibitors remained consistent with earlier findings. For promiscuous kinase inhibitors (PD $\geq$ 2), a mean and median PD of 3.7 and 2.0 was obtained, respectively. Only a small subset of the inhibitors had a high promiscuity. In addition to assessing kinase inhibitor promiscuity, it was also of interest to determine which kinase groups might form the largest numbers of inhibitor interactions. For the subset of promiscuous kinase inhibitors, the highest recorded number of interactions was found for tyrosine kinases (group TK) (66 011 interactions; 42.8% of all interactions), followed by CMGC (22 816; 14.8%) and AGC (16 097; 10.4%) kinases.

Figure 2 reports the distribution of nM ratios for decreasing numbers of promiscuous inhibitors with increasing PD values. The widest distribution was observed for inhibitors with at least

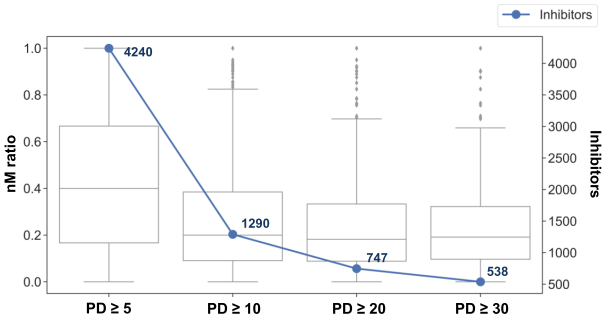


Figure 2. Distribution of nanomolar ratios for inhibitors with increasing PD values. Boxplots monitor distributions of nM ratios (vertical axis on the left) for subsets of inhibitors at different PD thresholds (horizontal axis). The blue curve reports the number of inhibitors in each set (vertical axis on the right). 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), largest value (top line), and outliers (points below the smallest or above the largest value).

five target annotations, yielding a median value of 0.4 corresponding to 40% nM potency values. For inhibitors with a minimum of 10 target annotations, nM ratios were significantly reduced, yielding a median of 0.2. However, with further increase in PD thresholds, the distributions remained essentially constant. About half of 538 inhibitors with a PD of at least 30 had nM ratios of 0.2 or greater. Hence, highly promiscuous inhibitors were frequently active in the nanomolar range against multiple kinases. Moreover, promiscuity patterns of inhibitors greatly varied. Figure 3 shows four representative examples of inhibitors with more than 50 kinase annotations and the distributions of their activities across the kinome. As can be seen, highly promiscuous inhibitors were either active against kinases from different groups with similar frequency, corresponding to a wide distribution of activities across the human kinome, or predominantly targeting individual groups such as tyrosine kinases. In addition, the distribution of μM vs nM potencies substantially varied. In some instances, nM potencies of inhibitors were largely confined to single kinase groups, in others they were distributed over different groups. Thus, inhibitors displayed diversified promiscuity patterns, which revealed differential activities across the kinome, even for highly promiscuous inhibitors.

Promiscuity Cliffs. Next, we systematically searched for structural analogous kinase inhibitors forming PCs, which required the consideration of additional analysis criteria. First, inhibitors with PAINS alerts were excluded from PC analysis to minimize the risk of false-positive PC assignments. Second, a data-driven PD difference (ΔPD) criterion for cliff formation was established.

A total of 7132 inhibitors with PAINS alerts were detected (6.3%), thus only a small proportion, which included 4177 PAINS among 68 361 inhibitors with single kinase annotations and 2955 PAINS among 44 263 promiscuous inhibitors. Following the removal of PAINS, the PD value distribution was re-calculated for the remaining 41 308 promiscuous inhibitors, which again yielded a mean and median of 3.7 and 2.0, respectively, the same as for all promiscuous inhibitors including PAINS, indicating that inhibitors with PAINS alerts were in general not highly promiscuous.

We then determined the PD distribution for the subset of promiscuous inhibitors with five or more target annotations, which yielded a median PD of 6. Hence, on the PD scale, the top ~2% of kinase inhibitors had PD values of 6 or greater. These compounds were considered as candidates for highly promiscuous PC partners. Therefore, we set the ΔPD threshold for PC formation to 5. Accordingly, the PC of smallest magnitude involving an inhibitor with a single kinase annotation was formed with a qualifying structural analogue having a PD of 6. In addition, we set the criterion that weakly promiscuous cliff partners were limited to inhibitors with PD values ranging from 1 to 4. Application of this criterion ensured that no PCs were formed by pairs of highly promiscuous inhibitors.

PC analysis was then based on a total of 105 492 inhibitors without PAINS alerts. A large number of 15 939 PCs was identified that involved 10 741 inhibitors (10.2%) including 1653 compounds with PD values of 6−295. We also determined that 2236 PCs (14.0%) were formed by 129 kinase inhibitors at different stages of clinical development and close structural analogues. Nearly all (i.e., 126) clinical inhibitors forming PCs were highly promiscuous cliff partners (PD $\geq$ 6), and 68 of the clinical inhibitors formed at least 10 PCs and thus served as "promiscuity hubs" in a PC network, as further discussed below.

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

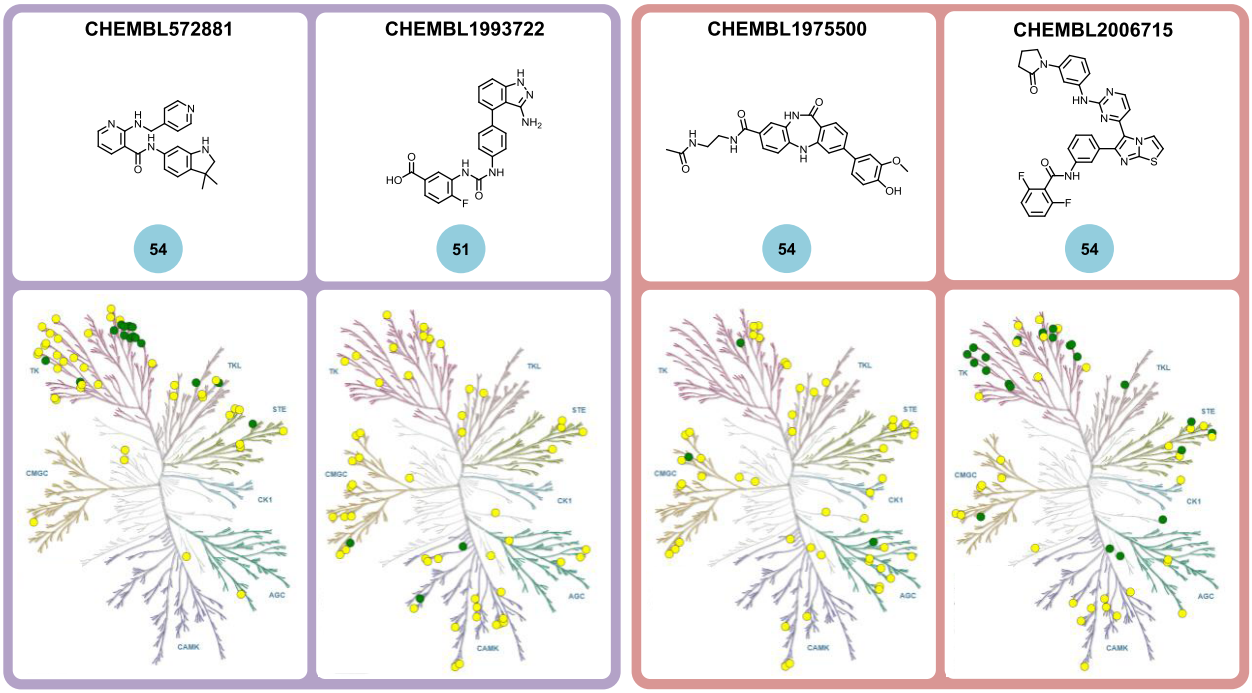


Figure 3. Promiscuity patterns. Shown are pairs of highly promiscuous inhibitors displaying different promiscuity patterns. ChEMBL IDs are reported above the compounds and their PD values below in blue circles. For each inhibitor, a phylogenetic tree of the human kinome is shown onto which its kinase annotations are mapped. Each dot represents a kinase the inhibitor is active against. Dots are color-coded according to compound potency (green, nanomolar; yellow, micromolar).

Figure 4 shows the distribution of ΔPD values over all PCs. More than half of the PCs (55%) had ΔPD values of 10 or more

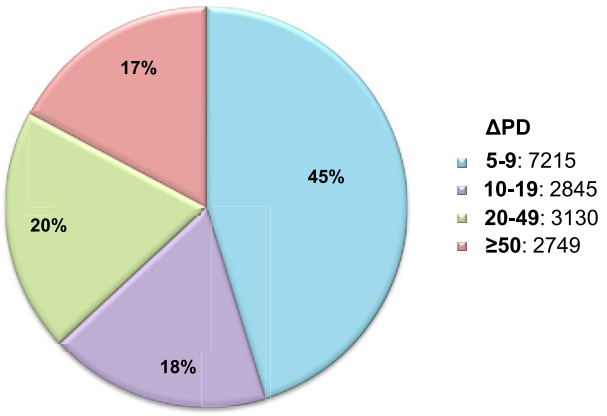


Figure 4. Distribution of ΔPD values for promiscuity cliffs. A pie chart shows the distribution of ΔPD values for the set of 15 939 PCs.

and 5879 PCs (37%) ΔPD values of 20 or more. Thus, significant numbers of large-magnitude PCs were identified. In Figure 5, exemplary PCs of increasing magnitude are shown, which reveal small structural changes that distinguish inhibitors with increasingly large differences in potency. As such, each PC encodes (i) additional target hypotheses for weakly or nonpromiscuous inhibitors (taking data incompleteness into consideration) and (ii) hypotheses for structural changes that might be responsible for achieving target selectivity or trigger promiscuity. Accordingly, computationally identified PCs provide a wealth of opportunities for follow-up investigations.

Promiscuity Cliff Network. We then generated a global network from all 15 939 PCs in which nodes represented inhibitors and edges pairwise PC relationships. The global PC network was found to consist of a total 622 clusters with two to 633 inhibitors per cluster, with a mean of 17.3 and median of 6.5 inhibitors. These clusters contained between one and 1351 PCs, with a mean and median of 25.6 and 6.0 PCs per cluster. Thus, PCs were typically formed by groups of structurally related inhibitors, similar to what has been observed for activity cliffs,[54,55] the majority of which are formed in a "coordinated" manner[55] as well as for "interaction cliffs", which take protein−ligand interaction similarity into account, in addition to structural similarity.[56] As discussed below, coordination of PCs further increased the structural context information for promiscuity analysis. Figure 6a shows exemplary PC clusters from the global network. As can be seen, these clusters vary greatly in their size, topology, and complexity.

Promiscuity Cliff Pathways. PC clusters served as a source of "PC pathways" (PCPs), as also illustrated in Figure 6a. A PCP represents a linear substructure (subgraph) of a PC cluster and a data structure for the extraction of structure−promiscuity relationships from the clusters. Figure 6b−g show a variety of PCPs of increasing length that are traced in clusters. A simple PCP is depicted in Figure 6b, which was isolated from a PC cluster with a "star" topology, resulting from a central highly promiscuous inhibitor forming PCs with many others. This PCP consists of only two PCs of smallest possible magnitude (i.e., PD values of 1 and 6). Figure 6c shows a PCP from a cluster containing a highly promiscuous inhibitor (PD 56) and a number of weakly promiscuous analogues. The highly promiscuous inhibitor has a substituent that is chemically distinct from those of its analogues, which might be responsible for its high promiscuity, providing experimentally testable

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

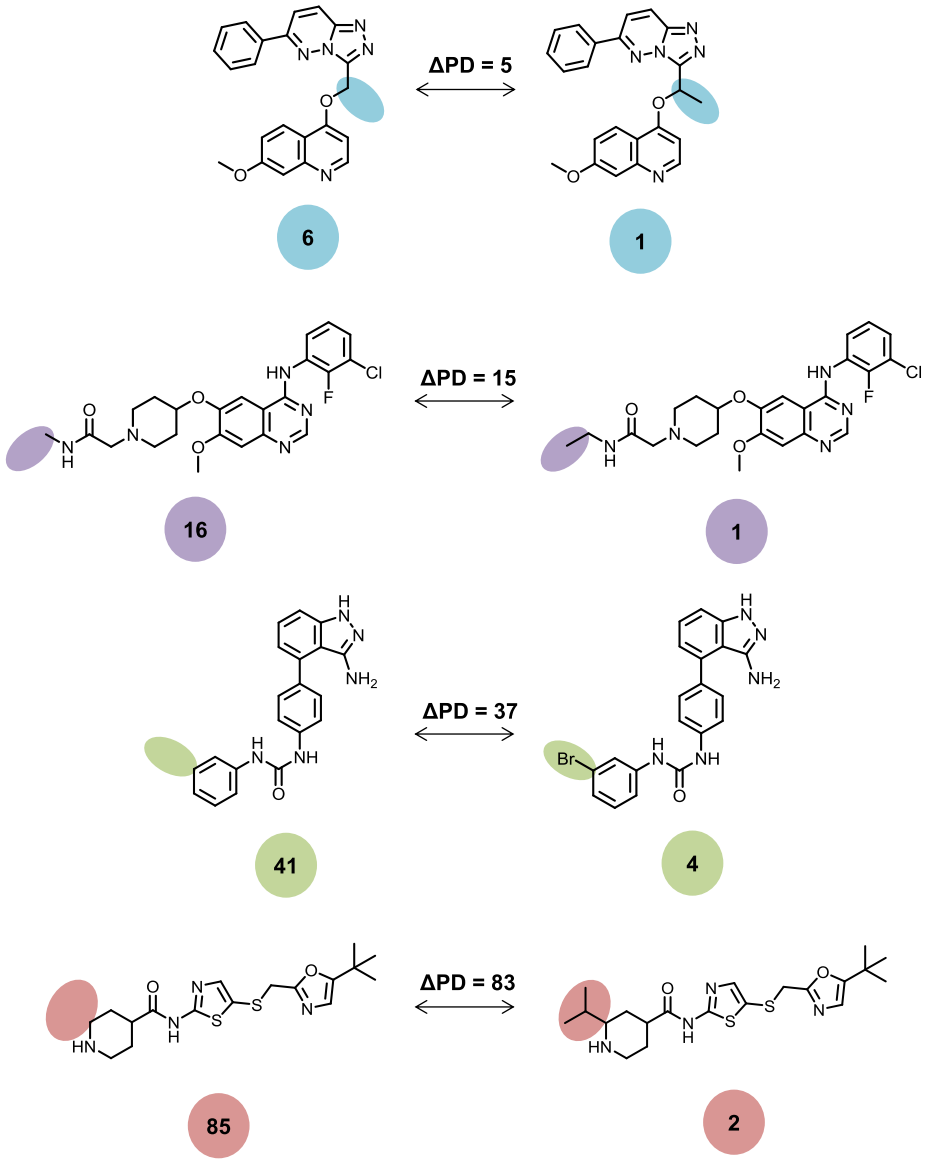


Figure 5. Exemplary promiscuity cliffs. For each ΔPD category in Figure 4, an exemplary PC is given. For each inhibitor, the PD value is reported and chemical modifications distinguishing PC partners are color-coded.

hypotheses. The PCP depicted in Figure 6d combines inhibitors with single kinase annotations and others with varying degrees of promiscuity including one of the most promiscuous inhibitors identified (compound 6, PD 165). It is striking to observe how small chemical modifications along the PCP relate non-promiscuous and highly promiscuous inhibitors to each other, for example, compounds 3 (PD 1) and 4 (PD 14) or compounds 6 (PD 165) and 7 (PD 1). A characteristic feature of PC sequences forming PCPs is that they consist of alternating structural analogues with low and high PD values. As such, the PCP uncovers multiple structure−promiscuity relationships that can be further investigated. Even a medium-sized PCP, such as the one shown in Figure 6d, provides many additional target hypotheses for inhibitors as well as hypotheses for structural modifications altering promiscuity. The information provided by a single PCP would be sufficient for initiating an experimental program to further explore a kinase inhibitor analogue series. As shown in Figure 6e−g, PCPs of increasing lengths can be isolated from increasingly large and complex clusters to extract structure−promiscuity relationship information from them. Both PCPs in Figure 6e,f are characterized by the presence of inhibitors with significantly varying PD values. Furthermore, the PCP in Figure 6g organizes a number of large-magnitude PCs involving inhibitors with single target annotations and highly promiscuous ones. It also illustrates that large series of overlapping PCs might encompass inhibitors of varying size and structural complexity. In cluster VI (Figure 6a) from which this PCP was extracted, compounds 11 and in particular 13 represent promiscuity hubs (with a PD value of 16 and 34, respectively), which have many weakly promiscuous or nonpromiscuous near neighbors. Such promiscuity hubs and their neighbors represent prime candidates for exploring the role of data incompleteness in subsequent kinase profiling assays as well as molecular origins of experimentally confirmed differences in promiscuity. Figure 7 shows exemplary clinical kinase inhibitors and their PC network neighborhoods.

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

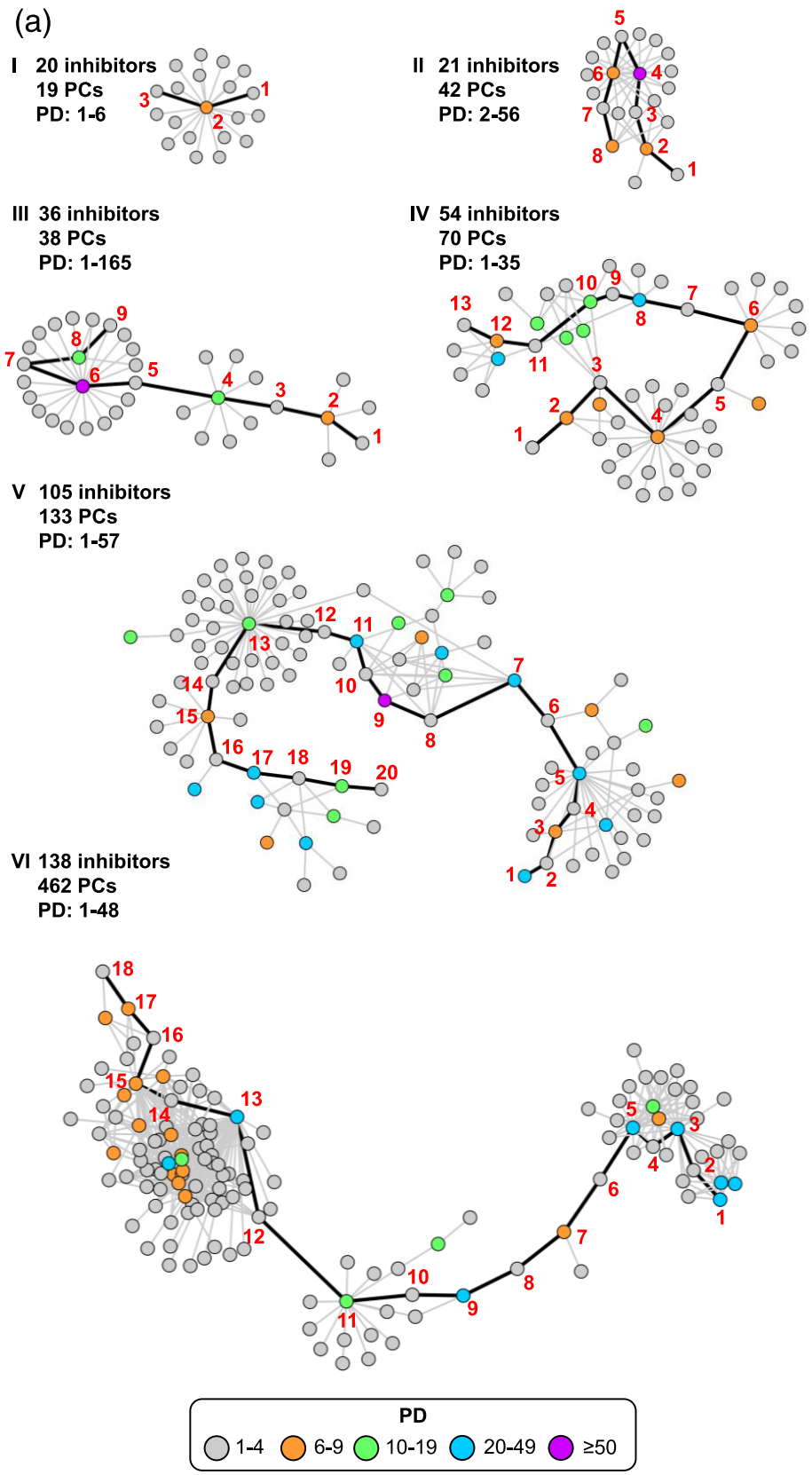


Figure 6. continued

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

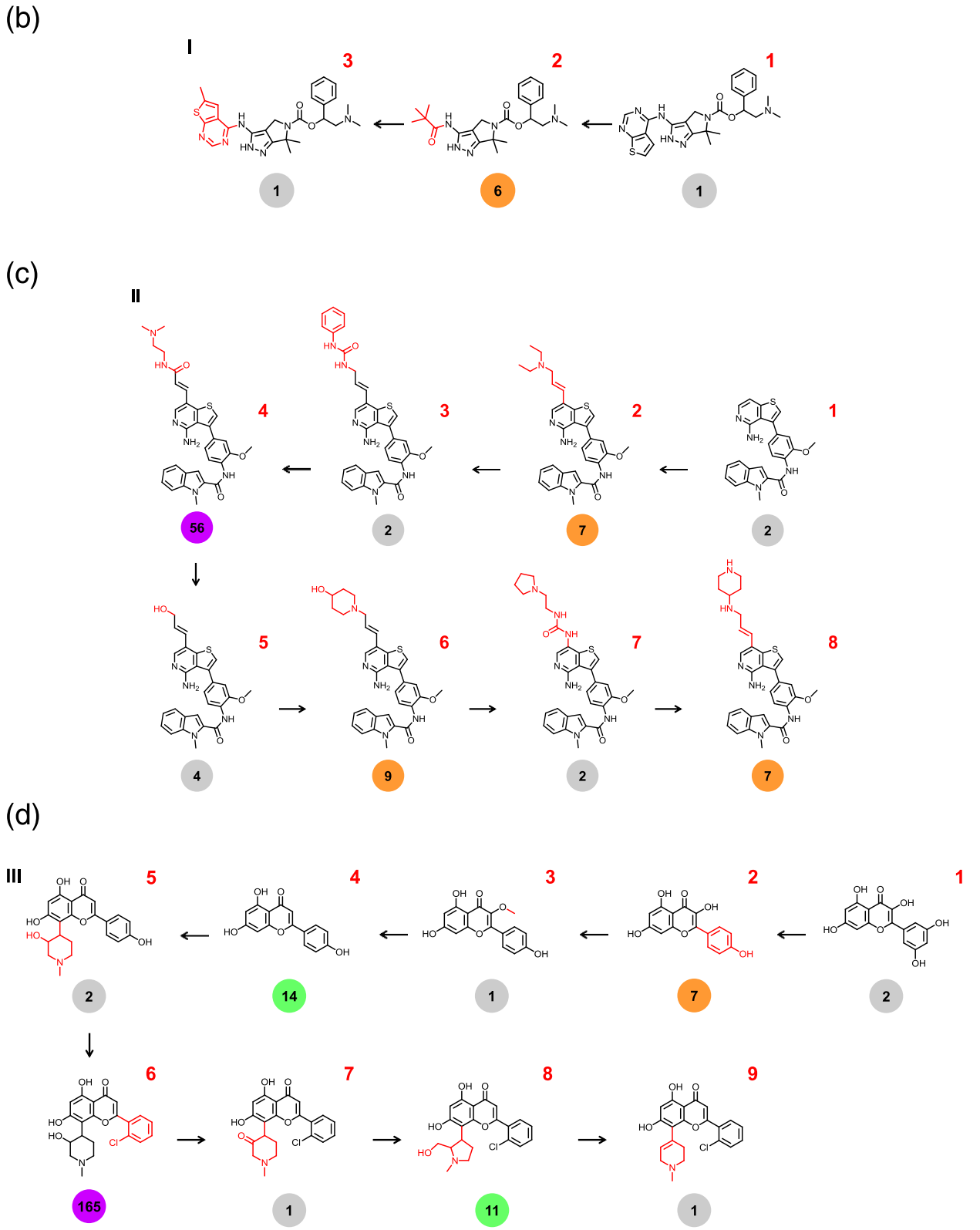


Figure 6. continued

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

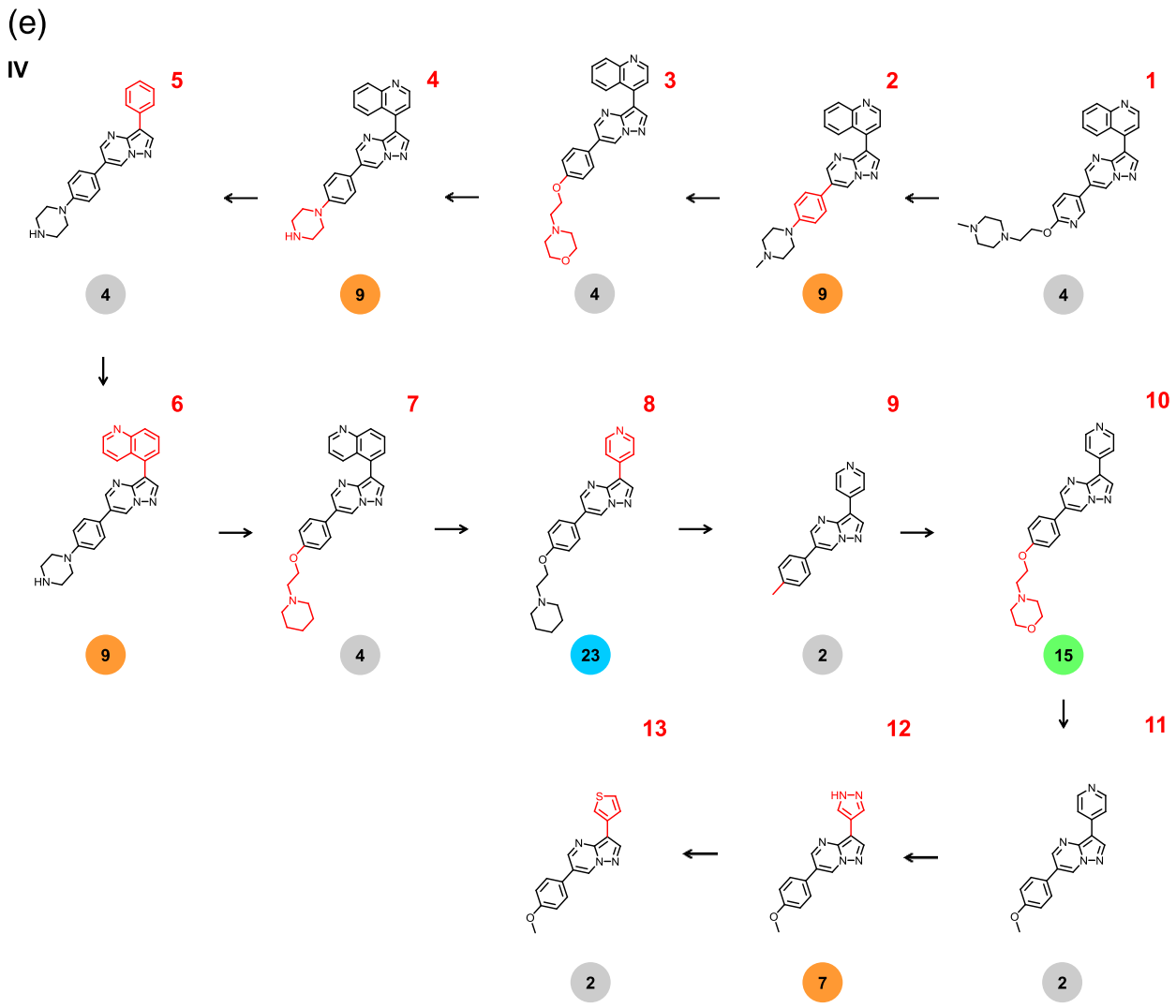


Figure 6. continued

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

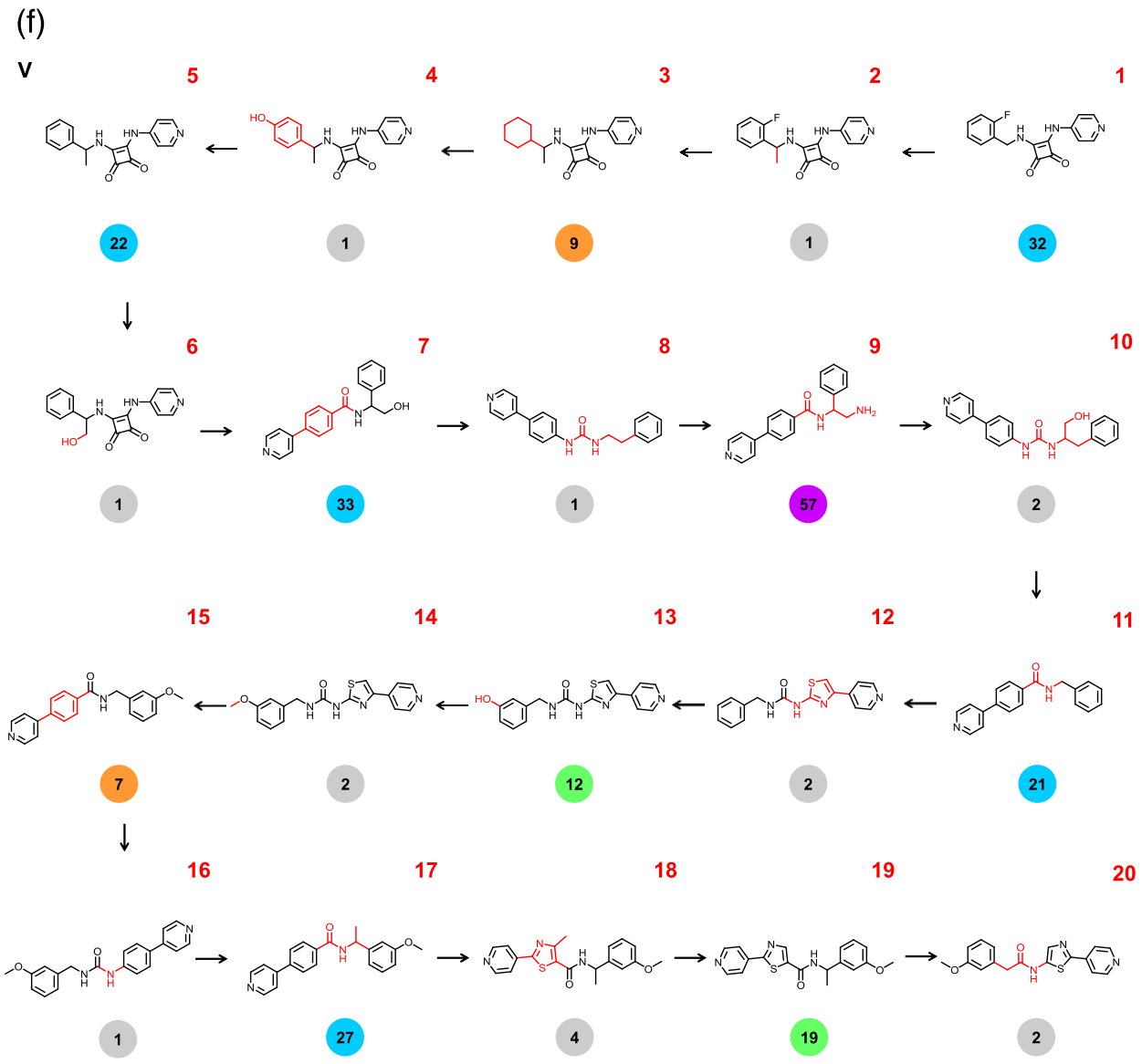


Figure 6. continued

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

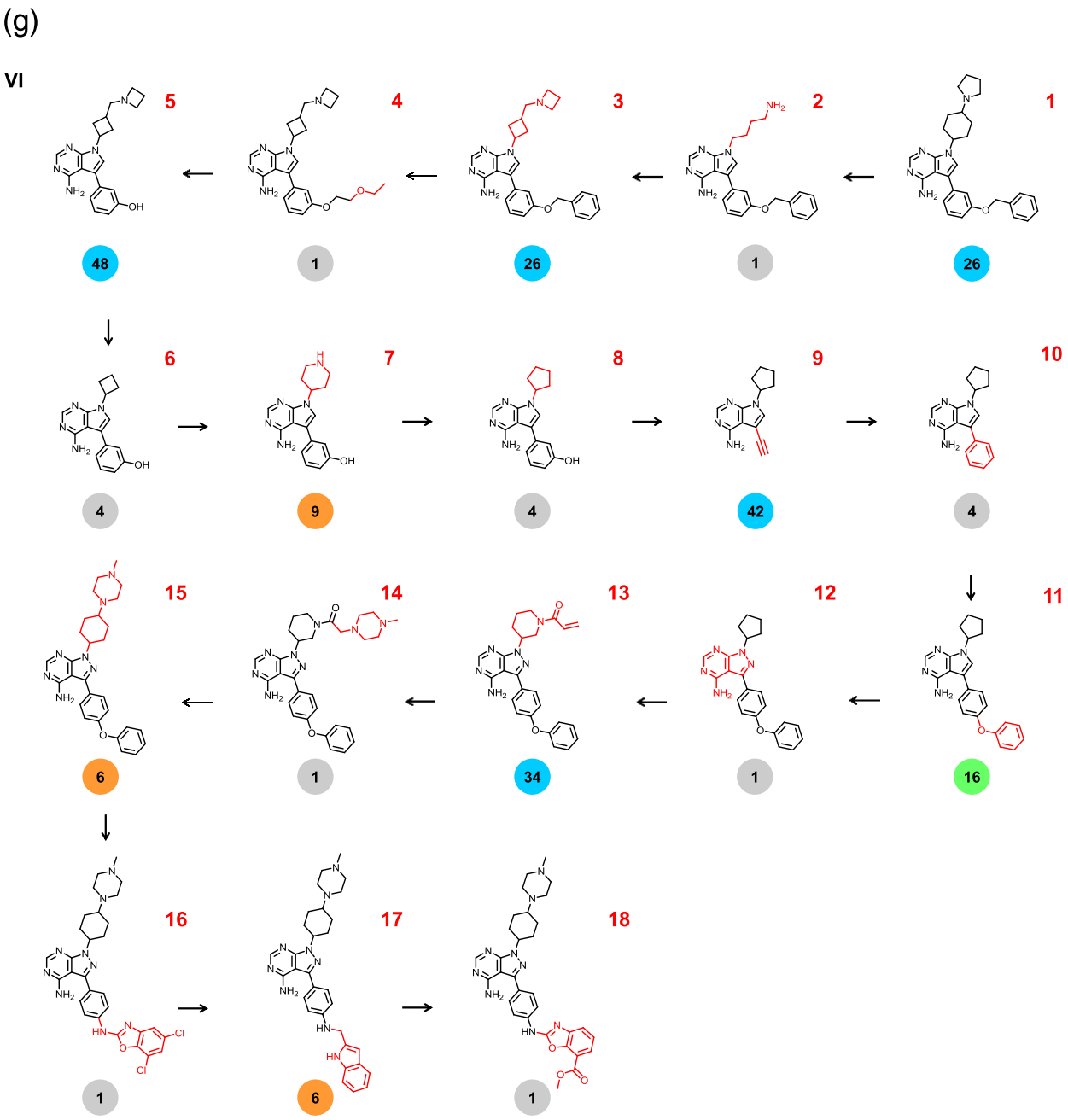


Figure 6. Promiscuity cliff pathways. (a) Selected clusters (I−VI) from the global PC network and their composition. Nodes represent inhibitors and edges pairwise PCs. Nodes are color-coded according to different PD ranges. Pathways formed by sequences of PCs are traced using thick black edges and selected compounds are numbered. In (b)−(g), pathways from clusters I−VI in (a) are depicted in detail. The PD values of the participating inhibitors are reported applying the same color code as in (a) and iterative structural modifications that distinguish inhibitors along the paths are shown in red.

Concluding Discussion. In this work, we have combined kinase inhibitors and associated activity data from different public repositories, yielding an unprecedentedly large collection of over 112 000 inhibitors with well-defined potency measurements and achieving 82% coverage of the human kinome. On the basis of these data, the much debated issue of kinase inhibitor promiscuity was revisited. As in a previous analyses focused on ChEMBL data, low global promiscuity was detected for human kinase inhibitors, with mean and median PD values of 2.1 and 1.0, respectively. Only 4% of the currently available inhibitors were active against five or more human kinases. However, a small subset (∼1 to 2%) of highly promiscuous inhibitors was identified that did not contain PAINS substructures. Interestingly, highly promiscuous inhibitors displayed different nM ratios and activity distributions across the kinome.

Special emphasis was put on systematically identifying PCs and PCPs. The PC concept was first introduced when analyzing compound array experiments.[29] Another previous investigation established a promiscuity ontology of structurally related

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

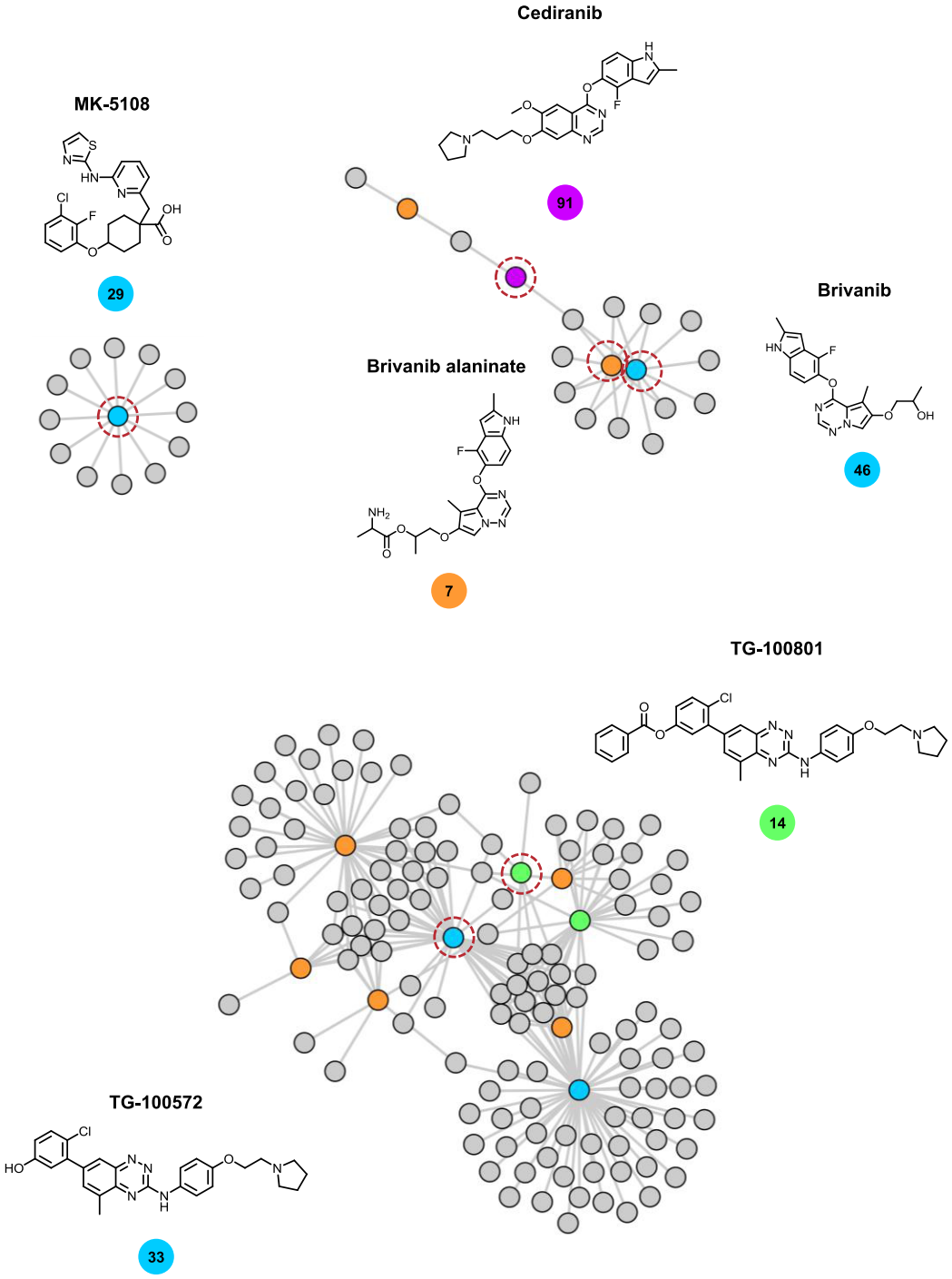


Figure 7. Network environment of clinical kinase inhibitors. Shown are exemplary clinical kinase inhibitors together with their neighborhoods in the global PC network. The node corresponding to clinical kinase inhibitor is encircled and the PD values given below the inhibitor. The representation is according to Figure 6.

compounds on the basis of calculated fingerprint similarity and distinguished promiscuous from selective compounds.[57] In our current analysis, we have systematically analyzed PCs, PC clusters, and PCPs to explore the structural modifications associated with large promiscuity differences. Therefore, data-driven PC criteria were established. A large number of ∼16 000 PCs were identified that were predominantly formed in a coordinated manner, as revealed by network analysis. We introduced the PCP concept to extract structure−promiscuity relationships from PC clusters and organize them in an interpretable form. The analysis uncovered many structurally analogous inhibitors with large PD value differences and chemical modifications converting high into weak or non-promiscuous inhibitors and vice versa. Observed large differences in promiscuity between structural analogues were surprising and might be due to multiple reasons, as discussed.

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

Systematic analysis of PC clusters and PCPs revealed many structure−promiscuity relationships and additional target hypotheses for inhibitors. As such, our study provides an example for large-scale computational data analysis and generation of data structures that provide a basis for experimental design. Therefore, following publication of this work, our kinase inhibitor data, PCs, and PC clusters will be made freely available to enable follow-up investigations.

In summary, our analysis has yielded

1. a large kinase inhibitor collection from different sources, achieving 82% coverage of the human kinome;
2. a detailed view of kinase promiscuity across the kinome;
3. approximately 16 000 PCs, which suggests target hypotheses of kinase inhibitors for follow-up investigations;
4. a global PC network from which PC clusters can be extracted;
5. PC pathways that can be directly used to explore structure−promiscuity relationships in medicinal chemistry.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: +49-228-7369-100.

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Author Contributions

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Ctibor Škuta of the Probes and Drugs Portal for providing kinase inhibitor data. The authors also thank Chemical Computing Group for providing an academic licence for the Molecular Operating Environment used for descriptor calculations and OpenEye Scientific Software for providing an academic license for software toolkits used for compound standardization and MMP generation.

REFERENCES

(1) Cohen, P. Protein Kinases - the Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309−315.

(2) *Kinase Drug Discovery*; Ward, R. A.; Goldberg, F. W., Eds.; RSC: Cambridge, U.K., 2011.

(3) Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130−137.

(4) Gross, S.; Rahal, R.; Stransky, N.; Lengauer, C.; Hoeflich, K. P. Targeting Cancer with Kinase Inhibitors. *J. Clin. Invest.* **2015**, *125*, 1780−1789.

(5) Simmons, D. L. Targeting Kinases: A New Approach to Treating Inflammatory Rheumatic Diseases. *Curr. Opin. Pharmacol.* **2013**, *13*, 426−434.

(6) Laufer, S.; Bajorath, J. New Frontiers in Kinases: Second Generation Inhibitors. *J. Med. Chem.* **2014**, *57*, 2167−2168.

(7) Wu, P.; Nielsen, T. E.; Clausen, M. H. Small-Molecule Kinase Inhibitors: An Analysis of FDA-Approved Drugs. *Drug Discovery Today* **2016**, *21*, 5−10.

(8) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(9) Anastassiadis, 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.

(10) Cheng, A. C.; John Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity Using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53*, 4502−4510.

(11) Levitzki, A. Tyrosine Kinase Inhibitors: Views of Selectivity, Sensitivity, and Clinical Performance. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 161−185.

(12) Müller, S.; Chaikuad, A.; Gray, N. S.; Knapp, S. The Ins and Outs of Selective Kinase Inhibitor Development. *Nat. Chem. Biol.* **2015**, *11*, 818−821.

(13) Elkins, J. M.; Fedele, V.; Szklarz, M.; Abdul Azeez, K. R.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X. P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive Characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2016**, *34*, 95−103.

(14) 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.

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

(16) Baell, J. B.; Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017—Utility and Limitations. *ACS Chem. Biol.* **2018**, *13*, 36−44.

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

(18) 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.

(19) Liu, Y.; Gray, N. S. Rational Design of Inhibitors that Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2006**, *2*, 358−364.

(20) Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9*, 1230−1241.

(21) Gavrin, L. K.; Saiah, E. Approaches to Discover Non-ATP Site Kinase Inhibitors. *Med. Chem. Commun.* **2013**, *4*, 41−51.

(22) Hu, Y.; Furtmann, N.; Bajorath, J. Current Compound Coverage of the Kinome. *J. Med. Chem.* **2015**, *58*, 30−40.

(23) Dimova, D.; Bajorath, J. Assessing Scaffold Diversity of Kinase Inhibitors Using Alternative Scaffold Concepts and Estimating the Scaffold Hopping Potential for Different Kinases. *Molecules* **2017**, *22*, No. e730.

(24) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. Data Completeness − The Achilles Heel of Drug-Target Networks. *Nat. Biotechnol.* **2008**, *26*, 983−984.

(25) Stumpfe, D.; Tinivella, A.; Rastelli, G.; Bajorath, J. Promiscuity of Inhibitors of Human Protein Kinases at Varying Data Confidence Levels and Test Frequencies. *RSC Adv.* **2017**, *7*, 41265−41271.

(26) Miljković, F.; Bajorath, J. Exploring Selectivity of Multikinase Inhibitors across the Human Kinome. *ACS Omega* **2018**, *3*, 1147−1153.

(27) Miljković, F.; Bajorath, J. Reconciling Selectivity Trends from a Comprehensive Kinase Inhibitor Profiling Campaign with Known Activity Data. *ACS Omega* **2018**, *3*, 3113−3119.

(28) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

(29) Dimova, D.; Hu, Y.; Bajorath, J. Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity. *J. Med. Chem.* **2012**, *55*, 10220−10228.
(30) Dimova, D.; Gilberg, E.; Bajorath, J. Identification and Analysis of Promiscuity Cliffs Formed by Bioactive Compounds and Experimental Implications. *RSC Adv.* **2017**, *7*, 58−66.
(31) Dimova, D.; Bajorath, J. Rationalizing Promiscuity Cliffs. *ChemMedChem* **2018**, *13*, 490−494.
(32) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(33) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699.
(34) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(35) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A. B.; Shoemaker, A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955−D963.
(36) Skuta, C.; Popr, M.; Muller, T.; Jindrich, J.; Kahle, M.; Sedlak, D.; Svozil, D.; Bartunek, P. Probes & Drugs Portal: An Interactive, Open Data Resource for Chemical Biology. *Nat. Methods* **2017**, *14*, 759−760.
(37) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.
(38) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.
(39) 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.
(40) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Res.* **2016**, *44*, D365−D371.
(41) Klaeger, S.; Heinzlmeir, S.; Wilhelm, M.; Polzer, H.; Vick, B.; Koenig, P. A.; Reinecke, M.; Ruprecht, B.; Petzoldt, S.; Meng, C.; Zecha, J.; Reiter, K.; Qiao, H.; Helm, D.; Koch, H.; Schoof, M.; Canevari, G.; Casale, E.; Depaolini, S. R.; Feuchtinger, A.; Wu, Z.; Schmidt, T.; Rueckert, L.; Becker, W.; Huenges, J.; Garz, A. K.; Gohlke, B. O.; Zolg, D. P.; Kayser, G.; Vooder, T.; Preissner, R.; Hahne, H.; Tõnisson, N.; Kramer, K.; Götze, K.; Bassermann, F.; Schlegl, J.; Ehrlich, H. C.; Aiche, S.; Walch, A.; Greif, P. A.; Schneider, S.; Felder, E. R.; Ruland, J.; Médard, G.; Jeremias, I.; Spiekermann, K.; Kuster, B. The Target Landscape of Clinical Kinase Inhibitors. *Science* **2017**, *358*, No. eaan4368.
(42) Schmidt, T.; Samaras, P.; Frejno, M.; Gessulat, S.; Barnert, M.; Kienegger, H.; Krcmar, H.; Schlegl, J.; Ehrlich, H. C.; Aiche, S.; Kuster, B.; Wilhelm, M. ProteomicsDB. *Nucleic Acids Res.* **2018**, *46*, D1271−D1281.
(43) Tang, J.; Tanoli, Z.-U.-R.; Ravikumar, B.; Alam, Z.; Rebane, A.; Vähä-Koskela, M.; Peddinti, G.; van Adrichem, A. J.; Wakkinen, J.; Jaiswal, A.; Karjalainen, E.; Gautam, P.; He, L.; Parri, E.; Khan, S.; Gupta, A.; Ali, M.; Yetukuri, L.; Gustavsson, A.-L.; Seashore-Ludlow, B.; Hersey, A.; Leach, A. R.; Overington, J. P.; Repasky, G.; Wennerberg, K.; Aittokallio, T. Drug Target Commons: A Community Effort to Build a Consensus Knowledge Base for Drug-Target Interactions. *Cell Chem. Biol.* **2018**, *25*, 224−229.
(44) *RDKit: Cheminformatics and Machine Learning Software*. http://www.rdkit.org, 2013.
(45) Sterling, T.; Irwin, J. J. ZINC 15−Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(46) 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.
(47) 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.
(48) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.
(49) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.
(50) 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.
(51) Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P. L.; Ideker, T. Cytoscape 2.8: New Features for Data Integration and Network Visualization. *Bioinformatics* **2011**, *27*, 431−432.
(52) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912−1934.
(53) Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: A Web-Based Tool for Interactive Navigation through Human Kinome Data. *BMC Bioinformatics* **2017**, *18*, No. 16.
(54) Maggiora, G. M. On Outliers and Activity Cliffs − Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(55) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and their Utility in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 18−28.
(56) Méndez-Lucio, O.; Kooistra, A. J.; de Graaf, C.; Bender, A.; Medina-Franco, J. L. Analyzing Multitarget Activity Landscapes Using Protein−Ligand Interaction Fingerprints: Interaction Cliffs. *J. Chem. Inf. Model.* **2015**, *55*, 251−262.
(57) Yongye, A. B.; Medina-Franco, J. L. Data Mining of Protein-Binding Profiling Data Identifies Structural Modifications That Distinguish Selective and Promiscuous Compounds. *J. Chem. Inf. Model.* **2012**, *52*, 2454−2461.

DOI: 10.1021/acsomega.8b02998
ACS Omega 2018, 3, 17295−17308

Summary

Combining inhibitors from different databases, we obtained 112,624 kinase inhibitors with well-defined activity measurements covering 82% of the human kinome. Only 4% of the reported inhibitors were found to be active against five or more kinases.

Altogether, ∼16,000 PCs were formed in a predominantly coordinated manner as shown by the network analysis. Moreover, 2236 PCs (∼14%) were formed by clinical kinase inhibitors. This finding suggested many target hypotheses for inhibitors of the human kinome and provided a basis for studying structural modifications determining inhibitor selectivity and promiscuity. From PC networks, PC pathways comprising sequences of PCs were isolated that revealed unexpected structure-promiscuity relationships.

Around ∼600 disjoint PC clusters were found in the network. PC clusters of increasing size and complexity are rich in structure-promiscuity information and difficult to interpret interactively. Therefore, the development of an automated protocol for identifying and extracting PC pathways from clusters was essential.

In the next chapter, a computational method for systematic identification of PC pathways is introduced.

Chapter 7

Systematic Computational Identification of Promiscuity Cliff Pathways Formed by Inhibitors of the Human Kinome

Introduction

Nearly 16,000 PCs were extracted from network representations, forming 626 disjoint PC clusters. PC pathways were introduced as a promiscuity data structure connecting compounds with alternating low and high promiscuity. Automated extraction and systematic evaluation of PC pathways was required. In this study, we developed a computational method to systematically identify PC pathways and extract them from PC clusters. A set of pathway parameters was defined and rank fusion was used to prioritize the most informative paths. PC pathways capturing network nodes characterized as promiscuity hubs were studied.

Journal of Computer-Aided Molecular Design (2019) 33:559–572
https://doi.org/10.1007/s10822-019-00198-9

Systematic computational identification of promiscuity cliff pathways formed by inhibitors of the human kinome

Filip Miljković[1] · Martin Vogt[1] · Jürgen Bajorath[1]

Received: 12 December 2018 / Accepted: 12 March 2019 / Published online: 26 March 2019

Abstract
The ability of a small molecule to interact with multiple target proteins provides the molecular basis of polypharmacology. So-defined compound promiscuity is intensely investigated in drug discovery. For example, for kinase inhibitors, the interplay between target selectivity and promiscuity plays a decisive role for different therapeutic applications. The "promiscuity cliff" (PC) concept was introduced previously to aid in promiscuity analysis. A PC is defined as a pair of structurally similar compounds with a large difference in promiscuity. Accordingly, PCs can reveal small structural modifications that might be responsible for selectivity or multi-target activity. In network representations, PCs form clusters of varying size and complexity that are difficult to analyze interactively. Herein, we introduce a computational method to systematically identify PC pathways, which are particularly rich in structure-promiscuity information, and extract them from PC clusters. PC pathways provide informative templates for experimental design. In a proof-of-concept investigation, we have applied the new computational approach to systematically identify pathways in more than 600 PC clusters formed by inhibitors of the human kinome, demonstrating the utility of the method and revealing many interesting promiscuity patterns.

Keywords Compound promiscuity · Structure-promiscuity relationships · Promiscuity cliffs · Promiscuity cliff pathways · Computational analysis · Automated pathway identification · Human kinome · Kinase inhibitors

Introduction

Possible origins of compound promiscuity continue to be debated in the drug discovery community. Promiscuity is often due to non-specific binding resulting from aggregation effects and other assay artifacts and thus highly undesirable [1–5]. On the other hand, compound promiscuity may originate from true binding events when a small molecule interacts with multiple targets in a defined way. Such multi-target activities form the basis of polypharmacology with its associated functional effects [6–8]. The polypharmacology concept has gradually revised and further extended the long-standing single-target specificity paradigm in drug discovery [9, 10]. However, achieving target specificity of small molecules will continue to be a guiding principle for many therapeutic applications including, among others, the treatment of chronic diseases or development of anti-infective agents. Target specificity is also of critical relevance in other areas such as chemical biology where the development of high-quality chemical probes to interrogate target-dependent functional effects is a major focal point [11, 12]. By contrast, compounds with multi-kinase activity have been successfully applied in oncology [13, 14]. Other multi-target compounds show promise in therapeutic areas such as neurological disorders [15].

There are several computational and experimental avenues to explore molecular promiscuity. Compounds with multi-target activity can be identified through computational analysis of curated activity data [6, 7, 16] from medicinal chemistry that is available in major repositories such as ChEMBL [17] or biological screening data available in PubChem [18]. In addition, compound profiling and array experiments are a major source of multi-target activity information [18–23].

The "promiscuity cliff" (PC) concept [24–26] was originally introduced to bridge between computational and experimental approaches and aid in the analysis of

✉ Jürgen Bajorath
bajorath@bit.uni-bonn.de

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

compound array data [24]. A PC is defined as a pair of structurally analogous compounds, i.e., compounds that are only distinguished by a single substitution (R-group replacement), having a significant difference in the number of targets they are active against [24–26]. Accordingly, PCs reveal small chemical modifications that are implicated in causing promiscuity [25, 26]. Furthermore, differences in apparent promiscuity between PC compounds might be influenced by varying test (assay) frequencies. Thus, PCs also suggest additional target hypotheses for structural analogs of highly promiscuous compounds [26]. For meaningful applications of the PC concept, it must be ensured that compounds with assay liabilities and resulting frequent hitter characteristics are excluded from consideration [26, 27]. Going beyond the analysis of compound array experiments, PCs were identified on a large scale in publicly available active compounds from different sources [27, 28].

Kinase inhibitors are a prime target for promiscuity analysis because the vast majority of currently available inhibitors target the adenosine triphosphate (ATP) (cofactor)-binding site that is largely conserved across the human kinome [29, 30]. Hence, these inhibitors are expected to be promiscuous [31]. However, general promiscuity and lack of selectivity of kinase inhibitors is neither supported by profiling experiments [22, 23], nor compound activity data analysis [32–34].

To quantify differences in kinase activities of ATP site-directed compounds, a systematic search for PCs was carried out in a large collection of more than 112,000 inhibitors of 426 human kinases (82% of the human kinome) that were assembled from several public compound databases [28]. Nearly 16,000 PCs were identified. In a global network representation, these PCs formed more than 600 clusters of varying composition [28].

PC clusters represent a rich source of information for promiscuity analysis. For example, from clusters, PC pathways (PCPs) can be isolated that represent sequences of compounds with alternating low promiscuity -or selectivity- and high promiscuity. Hence, inspection of PCPs makes it possible to follow stepwise structural modifications that strongly influence apparent promiscuity levels [28]. However, the large number of increasingly complex PC clusters quickly limits manual analysis of PCPs and makes it essentially impossible to comprehensively study pathways in an interactive manner. Hence, there is a need to automate this process and enable systematic analysis of PC clusters and PCPs.

Herein, we present a computational approach to systematically identify PCPs in clusters, prioritize most informative PCPs, and extract them. In addition, an entropy-based measure is applied to assess the distribution of pathway-associated kinase activities across the kinome.

Materials and methods

Data set

The previously reported set of kinase inhibitor PC clusters [28] was taken for method development and subjected to systematic analysis. Transformation size-restricted matched molecular pairs (MMPs) [35, 36] were calculated to generate pairs of structurally analogous kinase inhibitors. An MMP is defined as a pair of compounds that are only distinguished by a chemical modification (transformation) at a single site [36, 37]. For inhibitors forming MMPs, the promiscuity degree (PD) was determined as the number of kinase annotations on the basis of curated activity data, applying a potency threshold of 10 μM to IC_{50}, K_i, or K_d values. An MMP was considered a PC if the absolute difference of inhibitor PD values (ΔPD) was at least 5, i.e., if one inhibitor was active against five more kinases than the other. In addition, the PD value of the less promiscuous inhibitor was required to be between 1 and 4 such that PCs could not be formed by pairs of highly promiscuous inhibitors. Accordingly, the smallest possible PC involved an inhibitor with PD = 1 and a structural analog with PD = 6. Applying these criteria, a total of 15,939 PCs were obtained that involved 10,741 kinase inhibitors, including 1653 inhibitors with PD values between 6 and 295. These inhibitors were capable of participating in PCs as highly promiscuous cliff partners. The global network representation of the 15,939 PCs (nodes: compounds, edges: pairwise PC relationships) contained 622 disjoint PC clusters [28].

Computational extraction of PC pathways

For computational analysis, PCP was defined as the shortest path between two nodes from a PC cluster. When multiple shortest paths existed between two nodes, ΔPD of the edges was considered and the path yielding the largest cumulative ΔPD value was chosen. In addition, to eliminate path redundancy, only a single path was retained if multiple shortest paths contained the same set of promiscuous compounds (PD ≥ 6). So-defined PCPs were systematically generated for all pairs of promiscuous non-terminal nodes (i.e., inhibitors forming at least two PC relationships with others). For each qualifying path, three parameters were calculated:

1. Length (number of nodes)
2. Total number of PCs involving promiscuous inhibitors with PD ≥ 6
3. Cumulative ΔPD of edges of the path.

We note that the application of criterion 2 makes it possible to prioritize PCPs that contain "promiscuity hubs", i.e., pathway compounds that form large numbers of PCs with others outside the PCP. Pathway hubs are further discussed below.

In addition to applying criteria for PCP prioritization, a frequency model for n kinase groups [29] associated with a path is obtained by counting the frequency of occurrence of kinases belonging to each represented group. From frequency counts, the Shannon entropy (SE) [38] was calculated:

$$SE = -\sum_{i=1,p_i>0}^{n} p_i \log_2 p_i$$

Here, the p_i is the relative frequency of occurrence of each kinase group. Low SE values indicate that kinases associated with a path belong to a single group while increasing values indicate that associated kinases belong to multiple (and increasing numbers of) groups.

PCPs were ranked separately in decreasing order according to criteria 1–3 specified above. Then, rank fusion was applied. Therefore, the three ranks of each path were sorted in ascending order yielding a tuple (r_a, r_b, r_c) with $r_a \leq r_b \leq r_c$. The PCPs were ranked according to the lexicographic order of the tuples. Initially, only the highest rank r_a was considered and only in case of a tie, the second best rank $r_b$was used; if there was a tie for both ranks, r_c was taken into consideration. Lexicographic ranking ensured that the highest ranked pathways according to each criterion appeared near the top of the final ranking.

All calculations were carried out using the Python-implemented NetworkX package [39]. Shortest path calculations of the unweighted network were performed using a breadth-first search strategy similar to Dijkstra's algorithm [40]. The method organizes nodes of a network in layers of increasing distance around a source node. Each node in a layer represents a target node that contains pointers to all nodes of the previous layer, which extend possible shortest paths to the target node. Thus, all shortest paths from a source node to an arbitrary target node can be determined and prioritized according to the criteria outlined above.

Pathway visualization

Highly-ranked PCPs in PC clusters were visualized. Clusters were drawn using NetworkX [39] applying the Kamada–Kawai force-directed layout algorithm [41]. Cluster nodes were color-coded according to PD value ranges. In clusters, selected PCPs were traced using a thick black line. In addition, PCP compounds forming hubs with other nodes were identified. For kinases associated with PCP nodes, the frequency of occurrence was counted. For each selected PCP, a phylogenetic tree was drawn using KinMap [42], in which each dot represented a kinase associated with a PCP compound. Dots were scaled in size according to the frequency of kinase annotations.

Results and discussion

The new methodology for PCP extraction from PC clusters was tested on kinase inhibitor PCs identified on the basis of medicinal chemistry data. For these active compounds, no test frequencies were available. We note that PCs have also been identified on the basis of publicly available screening compounds for which test frequencies were available [43]. These PCs also extensively formed clusters [43], similar to the kinase inhibitor PCs used herein. For the development of our method, the source of PCs (medicinal chemistry data or biological screening) made no difference.

Promiscuity cliff clusters

The 15,939 PCs formed by 10,741 kinase inhibitors were organized in a PC network in which nodes represented inhibitors and edges pairwise PC relationships. This network contained 622 isolated clusters. Figure 1 reports the distribution of inhibitors, PCs, and mean ΔPD values for the clusters. About half of the clusters contained small numbers of compounds and PCs, with median values of 6.5 and 6.0, respectively. However, about 25% of the clusters contained 20 or more compounds and PCs, representing increasingly large and complex clusters. The median ΔPD value for PC clusters was close to 10 and the third quartile value was close to 25. Thus, PC clusters captured large differences in compound promiscuity.

Promiscuity cliff pathways

An exemplary PCP is shown in Fig. 2a. The PCP data structure is particularly attractive for the analysis of promiscuity patterns because PCPs consist of sequences of PC compounds with alternating large and small PD values. Hence, along a path iterative structural modifications can be examined that lead to large differences in promiscuity between structurally analogous compounds. In addition, as also shown in Fig. 2a, promiscuous PCP compounds frequently represent promiscuity hubs forming multiple PCs with other structural analogs outside the path that are only

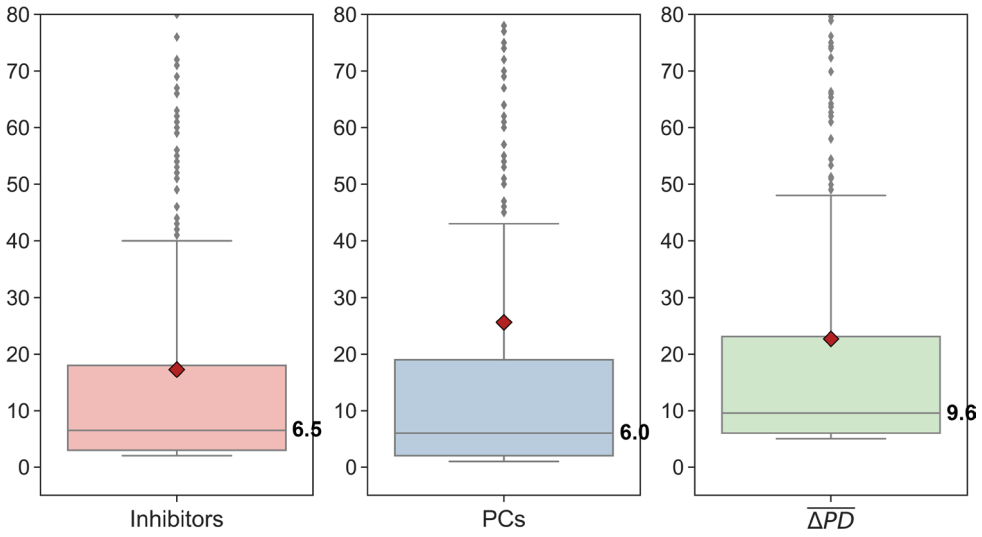


Fig. 1 Distribution of inhibitors, PCs, and mean ΔPD values for PC clusters. Boxplots report distribution of compounds, PCs, and mean ΔPD values for 622 kinase inhibitor PC clusters. Median values are reported and red diamond markers indicate the mean values of the distributions. Boxplots report the smallest value (bottom line), first quartile (lower boundary of the box), median value (thick line), third quartile (upper boundary of the box), largest value (top line), and outliers (points below the smallest or above the largest value)

Table 1 Cluster and pathway statistics

Cluster	Inhibitors	PCs	PCPs	PCP size	Cumulative ΔPD
A	132	234	42	3–7	23–230
B	117	261	21	3–5	10–175

For two exemplary PC clusters, the number of kinase inhibitors, PCs, and PCPs is reported. For PCPs, the size range (number of inhibitors) and cumulative ΔPD range are provided

weakly promiscuous or non-promiscuous, which provides additional information. Thus, for the exploration of structure-promiscuity relationships, PCPs represent an informative data structure.

Computational identification of pathways

Manually tracing PCPs is cumbersome and becomes essentially impossible when PC clusters grow in size beyond a few compounds such as the exemplary cluster shown in Fig. 2a. PC clusters contain many possible PCPs that need to be systematically examined to identify most informative paths. To these ends, computational analysis is essential and we introduce a new computational method for systematically identifying PCPs and extracting them from clusters. The approach relies on shortest path calculations between nodes in networks using breadth-first search akin to the Dijkstra's algorithm [40]. Application of this approach makes it possible to exhaustively mine PC clusters for PCPs and automate their extraction, guided by criteria to prioritize PCPs according to their structure-promiscuity relationship information content. PCPs were extracted from all kinase inhibitor PC clusters containing at least two promiscuous compounds with PD ≥ 6. In the following, exemplary cases are presented.

Pathway analysis

Table 1 reports the composition of two representative clusters A and B from the global PC network and their pathway statistics resulting from computational analysis. Cluster A contained 132 kinase inhibitors and 42 computationally identified PCPs meeting the criteria specified above and cluster B contained 117 inhibitors and 21 PCPs. The comparison illustrates that the number of PCPs does not necessarily scale with the number of compounds. Rather, the topology of clusters and content of hubs are major factors determining the number of PCPs. For cluster A and B, PCPs with up to seven and five inhibitors were identified, respectively. Figure 2a depicts cluster A and the top ranked PCP identified by computational analysis. It consists of seven structural analogs with substitutions at three sites. The PCP compounds include two densely connected hubs (compounds 1 and 5) and have striking difference in promiscuity including four in part highly promiscuous inhibitors, especially compound 1 (PD = 62), and three others with single kinase annotations. Large differences in promiscuity along the path are accompanied by confined structural modifications. In Fig. 2b, a part of the hub configuration around highly promiscuous compound 1 is displayed, which forms PCs with numerous inhibitors having mostly single kinase annotations. These analogs are distinguished from the highly promiscuous inhibitor by only minor chemical modifications leading to very large differences in apparent

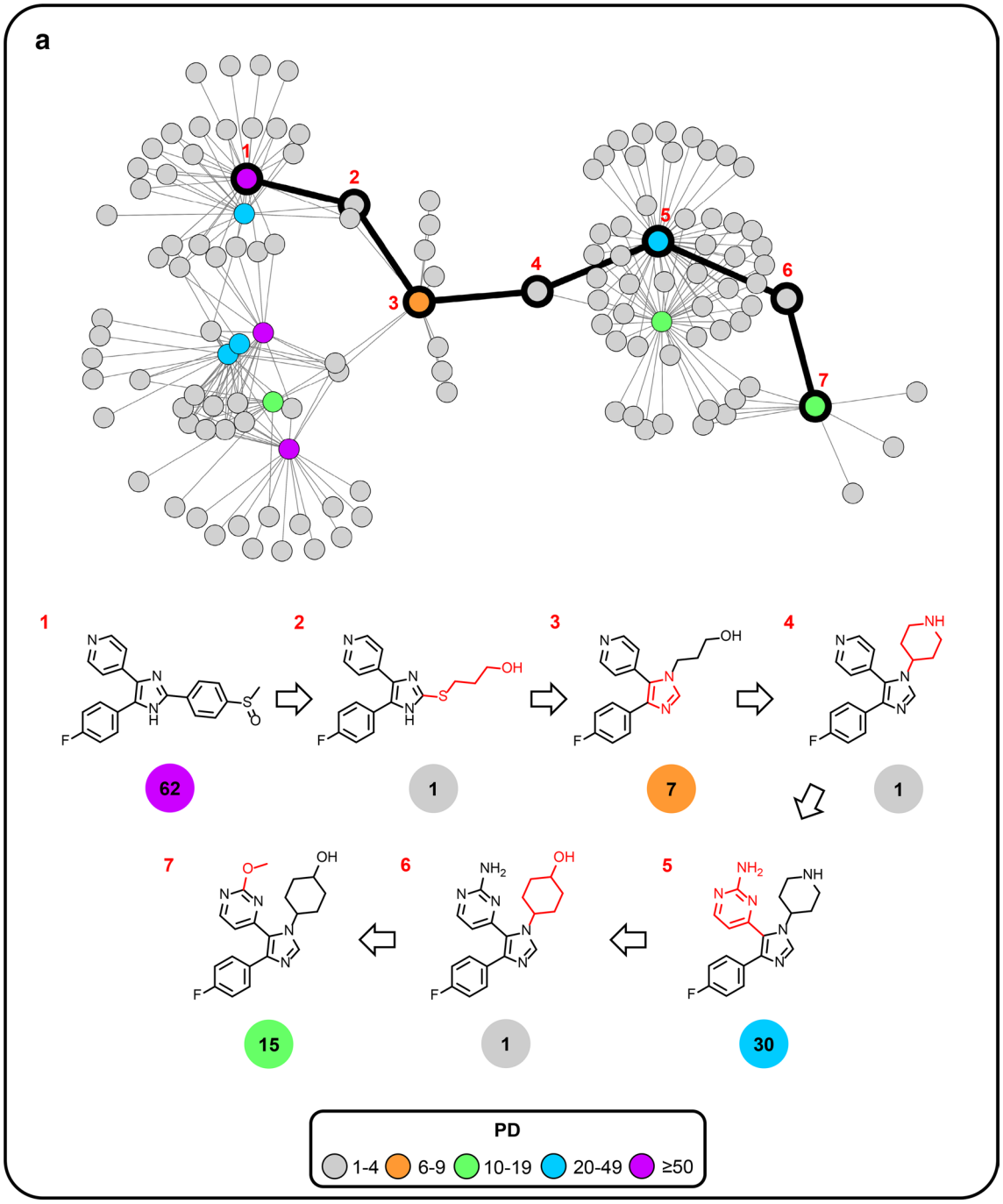


Fig. 2 Promiscuity cliff pathways from cluster A. In **a**, the top ranked PCP is traced. Nodes are color-coded according to PD value ranges and nodes of PCP compounds are numbered. Below the cluster, structures of PCP compounds are shown and their PD values are reported in corresponding nodes. Structural modifications distinguishing pairs of inhibitors along the path are colored red. In **b**, a promiscuity hub from the PCP is depicted that forms multiple PCs to other inhibitors with one or two kinase annotations. Structures of exemplary analogs are shown. In **c**, mapping of kinase annotations from the top ranked PCP onto a phylogenetic tree of the human kinome is shown. Each kinase associated with the PCP is represented as a red dot. The dots are scaled in size according to the number of kinase annotations along the path. In **d**, a lower ranked PCP from cluster A is traced

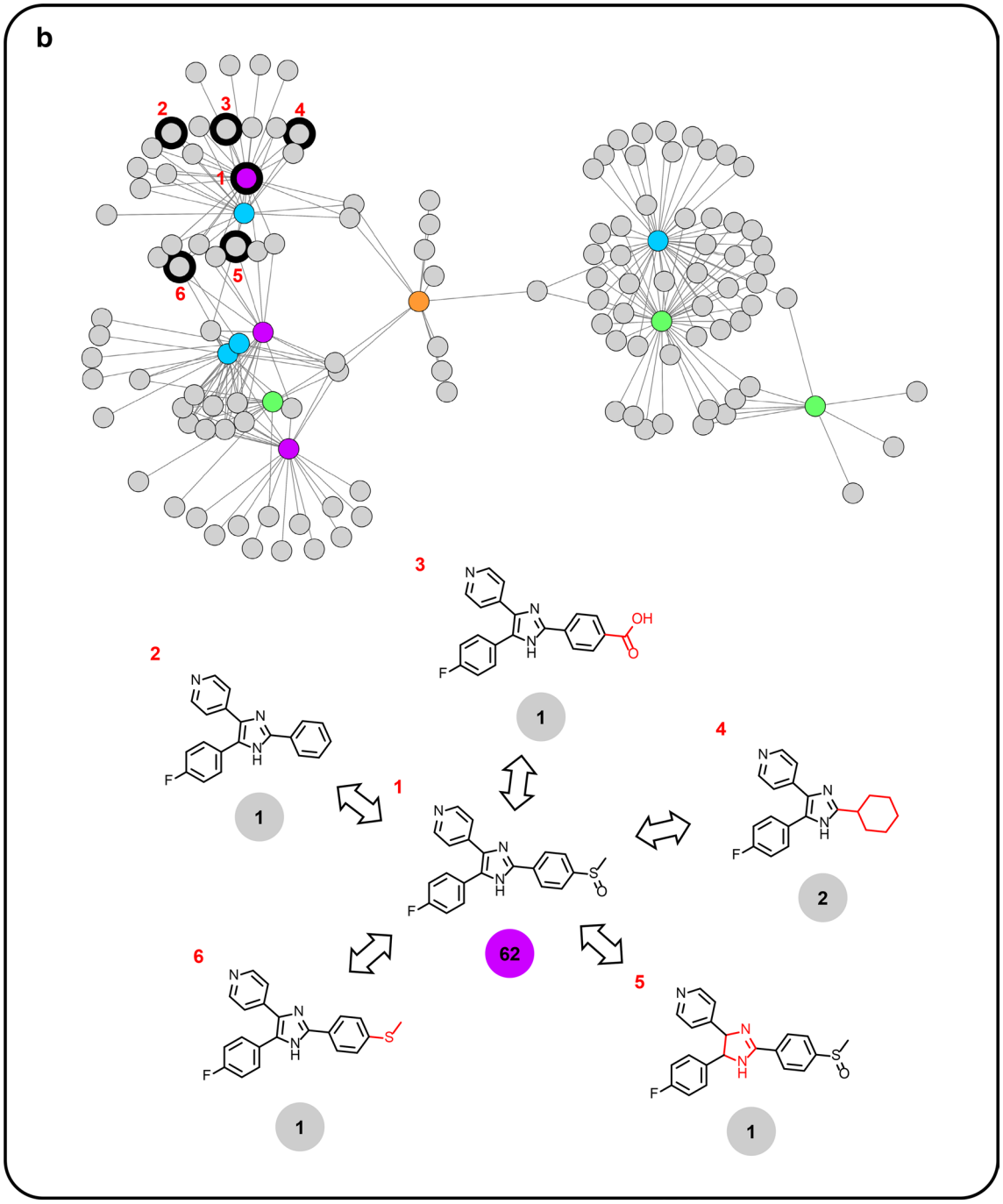


Fig. 2 (continued)

promiscuity. These observations are puzzling and this PCP alone would provide a basis for extensive follow-up experiments to better understand possible origins of large-magnitude differences in promiscuity. For example, inhibitors with apparent specificity (PD = 1) might be tested against other PCP-associated kinases and/or additional analogs might be generated to probe the influence of selected and combined chemical modifications on promiscuity. Without the identification and analysis of PCPs, many of these puzzling structure-promiscuity relationships would most likely

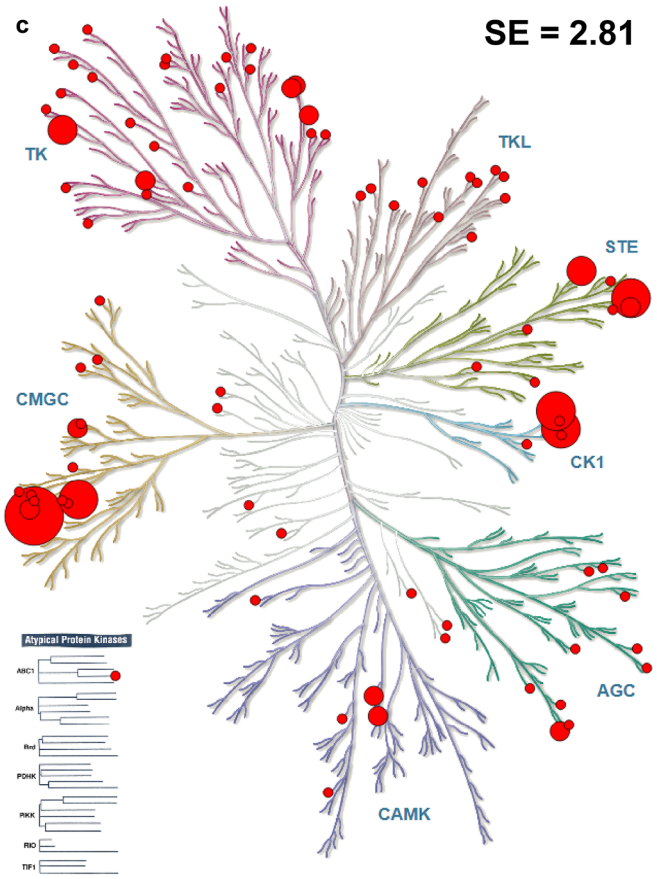


Fig. 2 (continued)

remain unnoticed, illustrating the utility of the PCP data structure.

Figure 2c shows that kinase annotations of inhibitors forming the top ranked PCP are widely distributed across the human kinome. The distribution of large dots indicates that a variety of distantly related kinases have multiple annotations originating from inhibitors of the PCP, suggesting additional target hypotheses for PCP compounds and hub analogs.

Figure 2d depicts a lower ranked PCP from cluster A that overlaps with the top ranked path. This PCP consists of five inhibitors including two densely connected hubs (compound 1 and 5) and one highly promiscuous inhibitor (compound 1, PD = 47). The lower rank of this PCP compared to the top ranked path is mainly due to its smaller size and lower cumulative ΔPD value. The kinome coverage of kinase annotations from both PCPs is comparable. Despite its lower rank, this PCP also reveals a variety of structure-promiscuity patterns and represents another informative template for experimental design.

Figure 3a depicts cluster B and its top ranked PCP. It consists of five inhibitors including three promiscuity hubs and two inhibitors with dual kinase activity. With 140 kinase annotations, PCP compound 1 is one of the most promiscuous kinase inhibitors we have identified. The PCP contains a close structural analog of this inhibitor with dual kinase activity (compound 2) that only differs by a hydroxyl to fluoro substitution. In addition, as shown in Fig. 3b, the hub environment of compound 1 also contains a variety of close analogs with only two or three kinase annotations. Thus, at a first glance, one might hypothesize that many analogs of compound 1 would also be more promiscuous but might have not been sufficiently tested. However, this immediate and plausible assumption of data sparseness as a cause of apparent differences in promiscuity is called

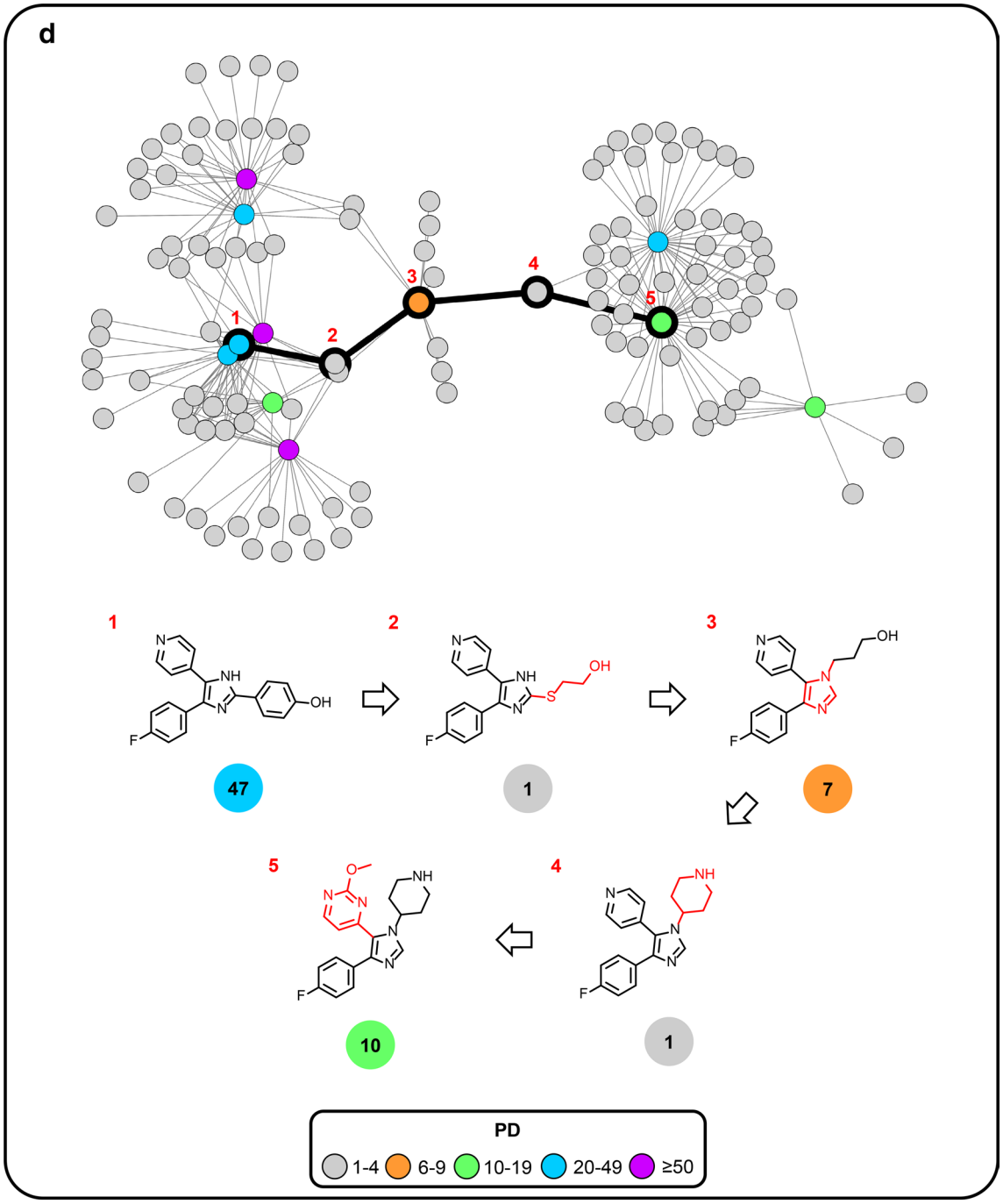


Fig. 2 (continued)

into question when analyzing the kinome distribution of PCP kinase annotations, shown in Fig. 3c. In this case, kinome-wide activities only result from the pan-kinase inhibitor (compound 1), whereas activities of the other PCP compounds and PCP-associated inhibitors are strongly focused on the Src family within the tyrosine kinase (TK) group. This is a characteristic of inhibitors comprising cluster B, as also illustrated by considering another lower

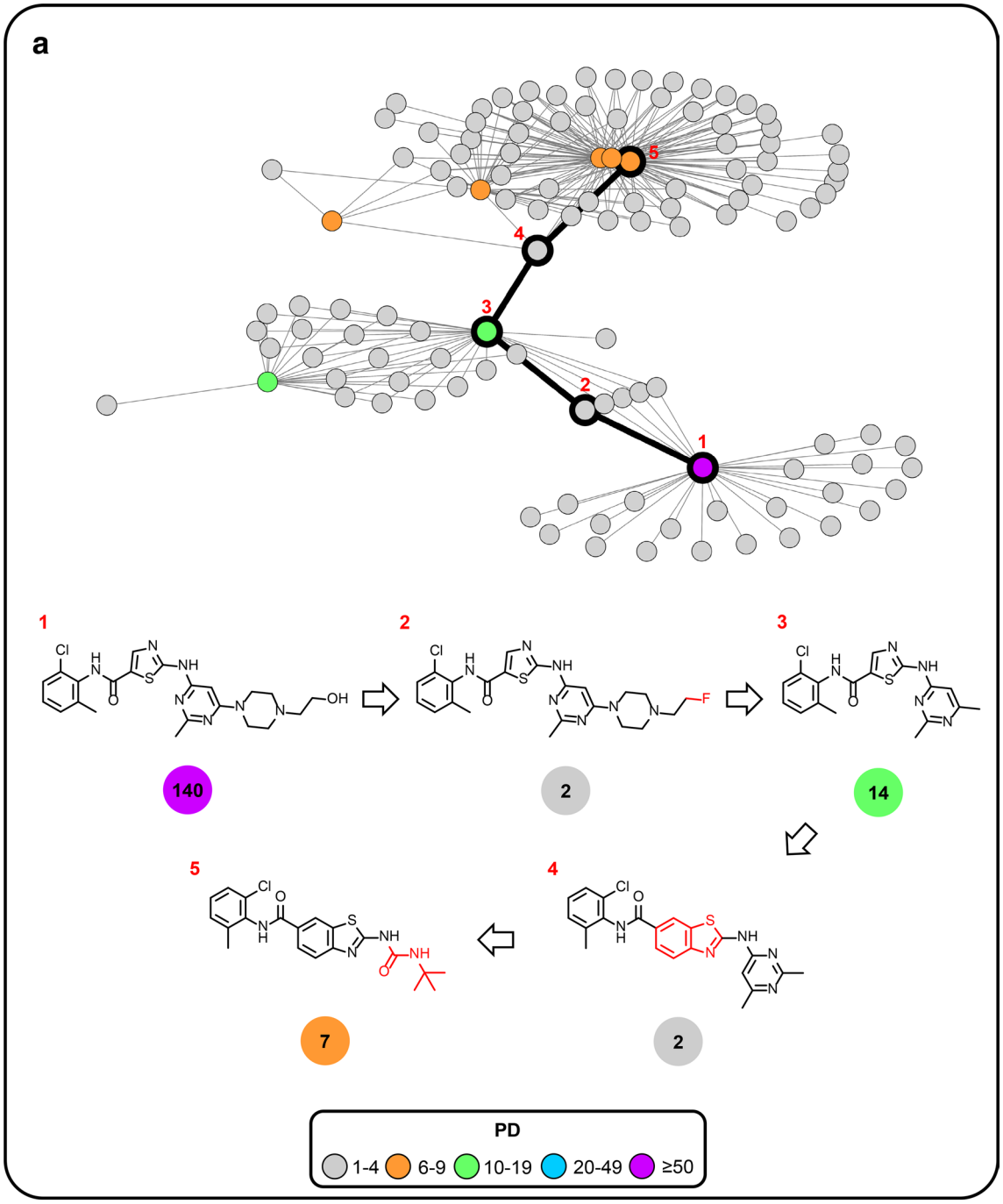


Fig. 3 Promiscuity cliff pathways from cluster B. In **a**, the top ranked PCP is traced. In **b**, a promiscuity hub is shown in detail. In **c**, the phylogenetic tree representation of kinase annotations associated with the top ranked PCP is depicted. In **d**, a lower ranked PCP from cluster B is shown. In **e**, the phylogenetic tree representation of the lower ranked PCP is displayed. The representation is according to Fig. 2

ranked PCP from this cluster, depicted in Fig. 3d. This PCP comprises five inhibitors and includes three promiscuity hubs (with a maximum of 14 kinase annotations). As revealed in Fig. 3e, these inhibitors are exclusively active against members of the TK group. Taken together, these observations suggest that it is unlikely that data sparseness alone would account for the apparent difference in promiscuity between compound 1 in Fig. 3a and other inhibitors in

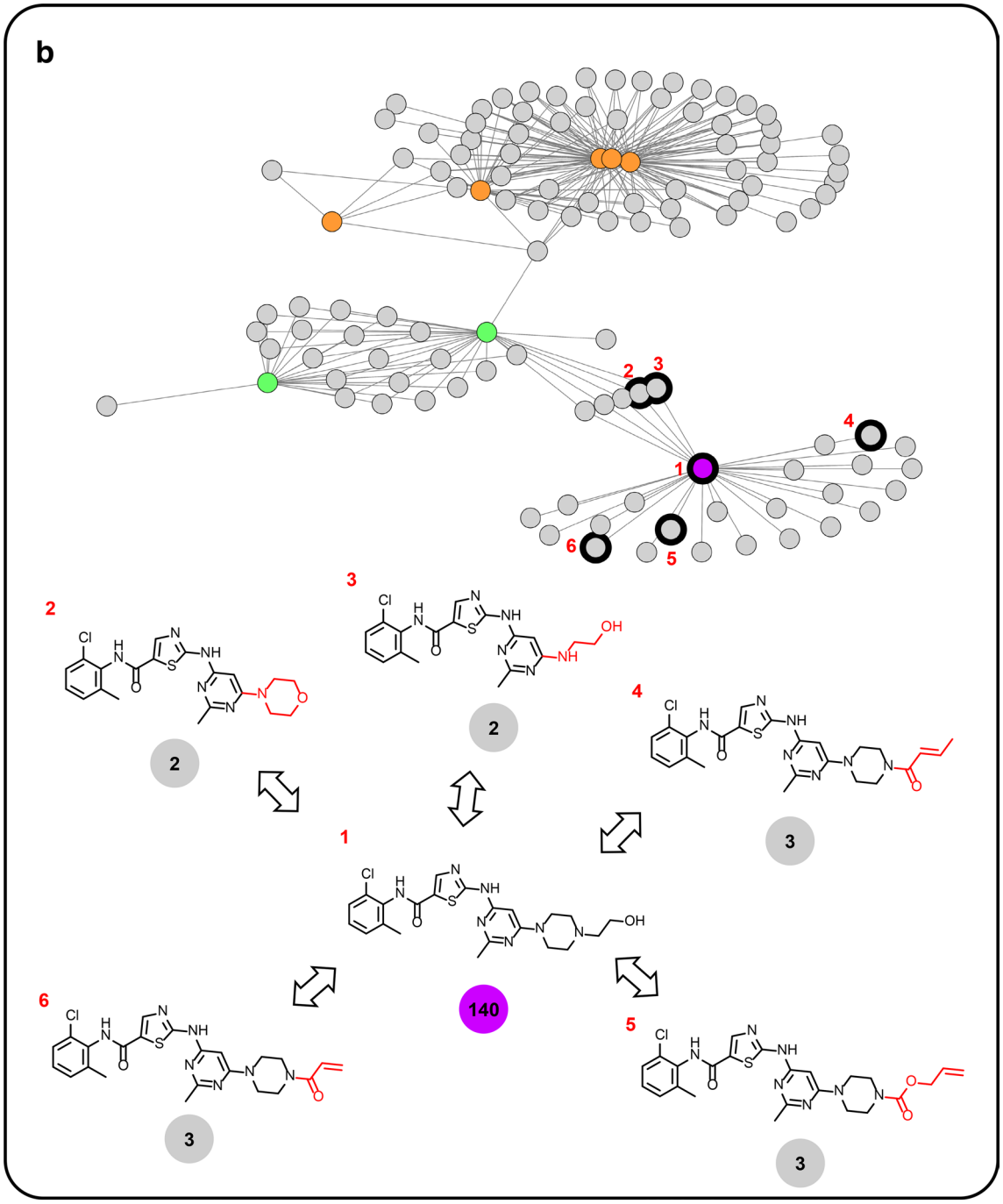


Fig. 3 (continued)

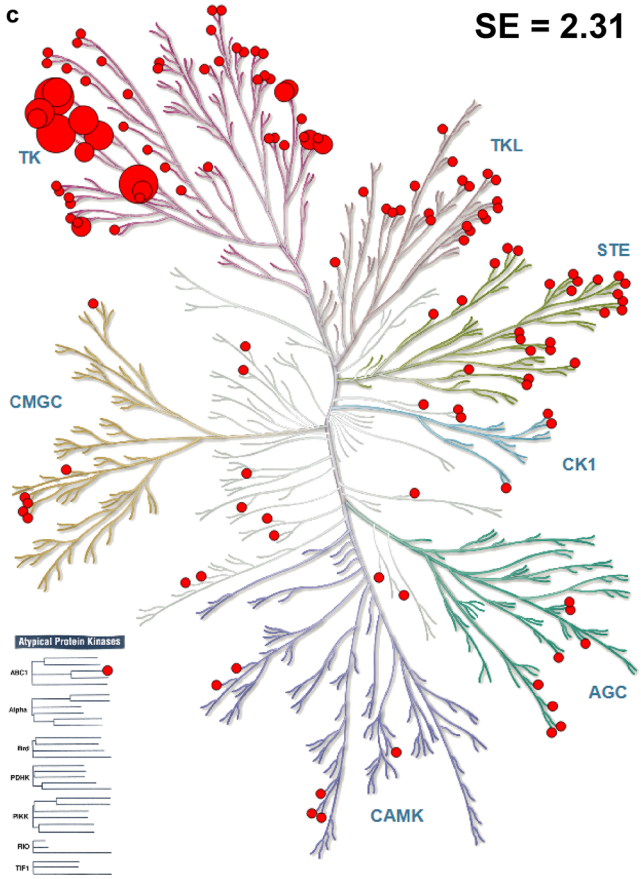


Fig. 3 (continued)

cluster B. Accordingly, exploring possible structural origins of pan-kinase versus TK promiscuity should, in this case, also be an attractive opportunity for follow-up investigation. Cluster A and B are representative of many PC clusters formed by kinase inhibitors that can be studied in detail on the basis of computational PCP analysis.

For the promiscuity hub examples reported in Figs. 2b and 3b, no comparative X-ray data are available to further investigate promiscuity differences. However, other examples of promiscuous compounds have recently been discussed on the basis of structural data [44], which are well worth considering in the context of PC analysis.

Conclusions

PC clusters from network representations represent a rich source of structure-promiscuity relationship information. The PCP data structure is particularly informative for promiscuity analysis and suitable to aid in experimental design. However, interactive graphical analysis of PC clusters and manual delineation of PCPs is difficult and limits PC analysis. Therefore, we have introduced a new computational approach to systematically extract and organize PCPs from PC clusters. The methodology makes it possible to exhaustively identify PCPs in data sets, as exemplified by our analysis of PC clusters formed by inhibitors of the human kinome. Systematically identified PCPs reveal many structure-promiscuity relationships that would be difficult, if not impossible to detect on the basis of interactive case-by-case analysis. PCPs provide a basis for exploring structural modifications that are implicated in triggering promiscuity versus selectivity and identify compound subsets in which apparent differences in promiscuity are likely due to data sparseness. Accordingly, the computational approach

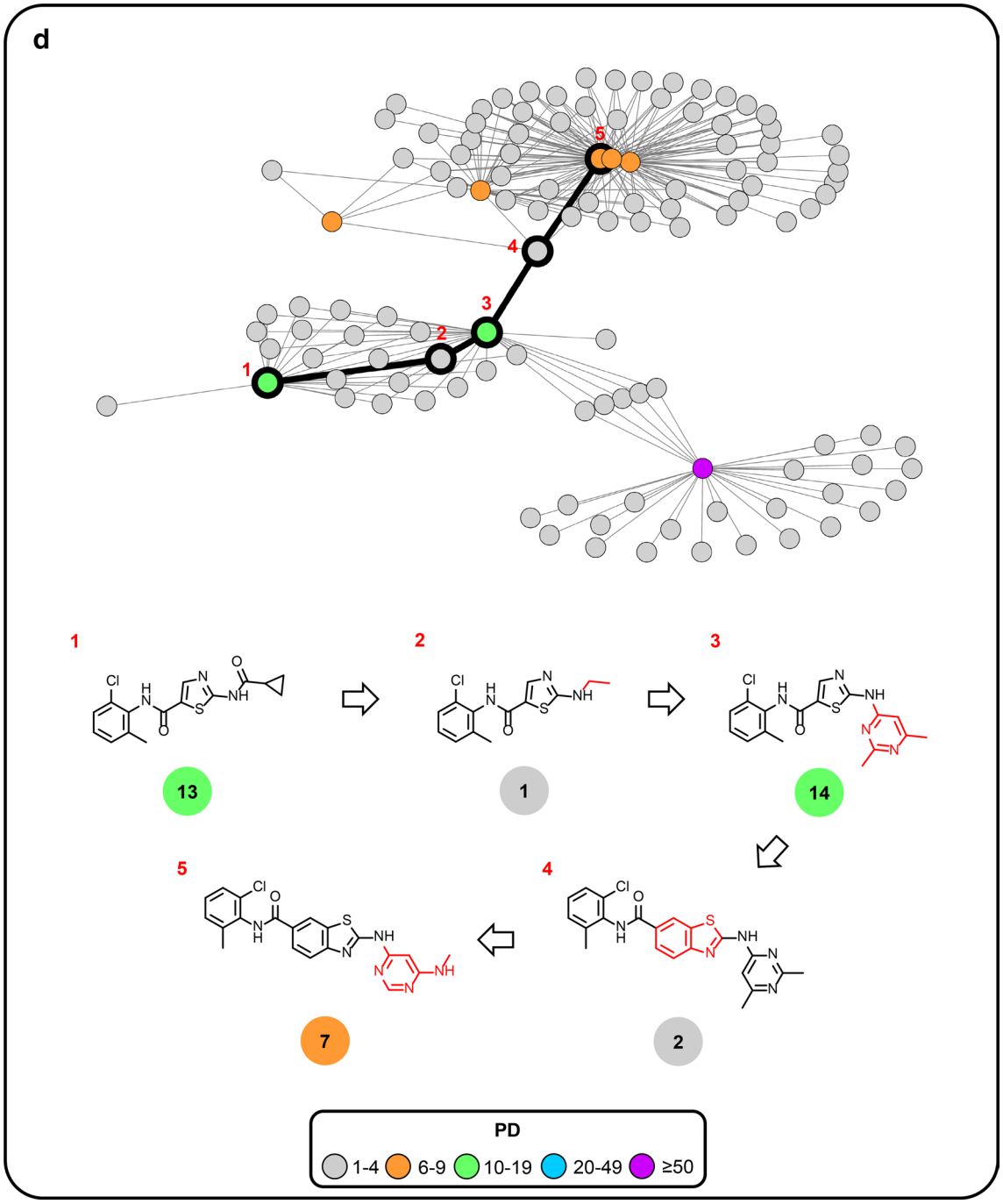


Fig. 3 (continued)

introduced herein enables a thorough investigation of promiscuity patterns on the basis of PCPs and associated promiscuity hubs. PCPs covering the human kinome we have identified as a part of our study will be made freely available for follow-up investigations as an open access deposition on the ZENODO platform [45].

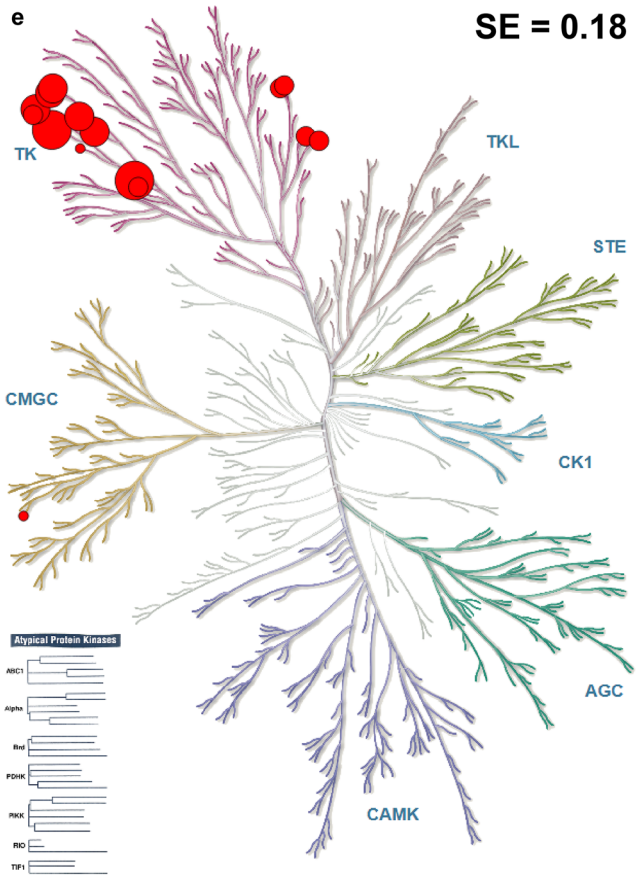


Fig. 3 (continued)

References

1. McGovern SL, Caselli E, Grigorieff N, Shoichet BK (2002) A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. J Med Chem 45:1712–1722
2. Feng BY, Shelat A, Doman TN, Guy RK, Shoichet BK (2005) High-throughput assays for promiscuous inhibitors. Nat Chem Biol 1:146–148
3. Shoichet BK (2006) Screening in a spirit haunted world. Drug Discov Today 11:607–615
4. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740
5. Baell J, Walters MA (2014) Chemistry: chemical con artists foil drug discovery. Nature 513:481–483
6. Hu Y, Bajorath J (2013) Compound promiscuity: what can we learn from current data? Drug Discov Today 18:644–650
7. Hu Y, Bajorath J (2013) High-resolution view of compound promiscuity. F1000Research 2:e144
8. Anighoro A, Bajorath J, Rastelli G (2014) Polypharmacology: challenges and opportunities in drug discovery. J Med Chem 57:7874–7887
9. Nurse P (1997) The ends of understanding. Nature 387:657–657
10. Roukos DH (2011) Networks medicine: from reductionism to evidence of complex dynamic biomolecular interactions. Pharmacogenomics 12:695–698
11. Arrowsmith CH, Audia JE, Austin C, Baell J, Bennett J, Blagg J, Bountra C, Brennan PE, Brown PJ, Bunnage ME, Doepner-Buser C, Campbell RM, Carter AJ, Cohen P, Copeland RA, Cravatt B, Dahlin JL, Dhanak D, Edwards AM, Frederiksen M, Frye SV, Gray N, Grimshaw CE, Hepworth D, Howe T, Huber KVM, Jin J, Knapp S, Kotz JD, Kruger RG, Lowe D, Mader MM, Marsden B, Mueller-Fahrnow A, Müller S, O'Hagan RC, Overington JP, Owen DR, Rosenberg SH, Ross R, Roth B, Schapira M, Schreiber SL, Shoichet B, Sundström M, Superti-Furga G, Taunton J, Toledo-Sherman L, Walpole C, Walters MA, Willson TM, Workman P, Young RN, Zuercher WJ (2015) The promise and peril of chemical probes. Nat Chem Biol 11:536–541
12. Miljković F, Bajorath J (2018) Data-driven exploration of selectivity and off-target activities of designated chemical probes. Molecules 23:e2434
13. Knight ZA, Lin H, Shokat KM (2010) Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10:130–137

14. Gross S, Rahal R, Stransky N, Lengauer C, Hoeflich KP (2015) Targeting cancer with kinase inhibitors. J Clin Invest 125:1780–1789
15. Bolognesi ML, Cavalli A (2016) Multitarget drug discovery and polypharmacology. ChemMedChem 11:1190–1192
16. Stumpfe D, Tinivella A, Rastelli G, Bajorath J (2017) Promiscuity of inhibitors of human protein kinases at varying data confidence levels and test frequencies. RSC Adv 7:41265–41271
17. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR (2017) The ChEMBL database in 2017. Nucleic Acids Res 45(Database issue):D945–D954
18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem substance and compound databases. Nucleic Acids Res 44(Database issue):D1202–D1213
19. Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT, Faraoni R, Floyd M, Hunt JP, Lockhart DJ, Milanov ZV, Morrison MJ, Pallares G, Patel HK, Pritchard S, Wodicka LM, Zarrinkar PP (2008) A quantitative analysis of kinase inhibitor selectivity. Nat Biotechnol 26:127–132
20. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol 29:1039–1045
21. Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, Koehler AN, Schreiber SL (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. Proc Natl Acad Sci USA 107:18787–18792
22. Elkins JM, Fedele V, Szklarz M, Abdul Azeez KR, Salah E, Mikolajczyk J, Romanov S, Sepetov N, Huang XP, Roth BL, Al Haj Zen A, Fourches D, Muratov E, Tropsha A, Morris J, Teicher BA, Kunkel M, Polley E, Lackey KE, Atkinson FL, Overington JP, Bamborough P, Müller S, Price DJ, Willson TM, Drewry DH, Knapp S, Zuercher WJ (2016) Comprehensive characterization of the published kinase inhibitor set. Nat Biotechnol 34:95–103
23. Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig PA, Reinecke M, Ruprecht B, Petzoldt S, Meng C, Zecha J, Reiter K, Qiao H, Helm D, Koch H, Schoof M, Canevari G, Casale E, Depaolini SR, Feuchtinger A, Wu Z, Schmidt T, Rueckert L, Becker W, Huenges J, Garz AK, Gohlke BO, Zolg DP, Kayser G, Vooder T, Preissner R, Hahne H, Tõnisson N, Kramer K, Götze K, Bassermann F, Schlegl J, Ehrlich HC, Aiche S, Walch A, Greif PA, Schneider S, Felder ER, Ruland J, Médard G, Jeremias I, Spiekermann K, Kuster B (2017) The target landscape of clinical kinase drugs. Science 358:eaan4368
24. Dimova D, Hu Y, Bajorath J (2012) Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. J Med Chem 55:10220–10228
25. Bajorath J (2017) From activity cliffs to promiscuity cliffs. Fut Sci OA 3:FSO227
26. Dimova D, Bajorath J (2018) Rationalizing promiscuity cliffs. ChemMedChem 13:490–494
27. Dimova D, Gilberg E, Bajorath J (2017) Identification and analysis of promiscuity cliffs formed by bioactive compounds and experimental implications. RSC Adv 7:58–66
28. Miljković F, Bajorath J (2018) Computational analysis of kinase inhibitors identifies promiscuity cliffs across the human kinome. ACS Omega 3:17295–17308, 2018
29. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298:1912–1934
30. Noble MEM, Endicott JA, Johnson LN (2004) Protein kinase inhibitors: insights into drug design from structure. Science 303:1800–1805
31. Levitzki A (2013) Tyrosine kinase inhibitors: views of selectivity, sensitivity, and clinical Performance. Annu Rev Pharmacol Toxicol 53:161–185
32. Miljković F, Bajorath J (2018) Exploring selectivity of multikinase inhibitors across the human kinome. ACS Omega 3:1147–1153
33. Miljković F, Bajorath J (2018) Reconciling selectivity trends from a comprehensive kinase inhibitor profiling campaign with known activity data. ACS Omega 3:3113–3119
34. Miljković F, Bajorath J (2018) Evaluation of kinase inhibitor selectivity using cell-based profiling data. Mol Inform 37:e1800024
35. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 52:1138–1145
36. Kenny PW, Sadowski J (2004) Structure modification in chemical databases. In: Oprea TI (ed) Chemoinformatics in drug discovery. Wiley-VCH, Weinheim, pp 271–285
37. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50:339–348
38. Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27:379–423
39. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds) Proceedings of the 7th python in science conference (SciPy 2008), Pasadena, CA, Aug 19–24, pp 11–15
40. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271
41. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inf Process Lett 31:7–15
42. Eid S, Turk S, Volkamer A, Rippmann F, Fulle S (2017) KinMap: a web-based tool for interactive navigation through human kinome data. BMC Bioinform 18:e16
43. Hu Y, Jasial S, Gilberg E, Bajorath J (2017) Structure-promiscuity relationship puzzles—extensively assayed analogs with large differences in target annotations. AAPS J 19:856–864
44. Gilberg E, Bajorath J (2019) Recent progress in structure-based evaluation of compound promiscuity. ACS Omega 4:2758–2765
45. https://www.zenodo.org

Summary

We introduced a new computational method to systematically identify, prioritize and extract most informative PC pathways from PC clusters. Computationally identified pathways revealed many structure-promiscuity relationships that would be difficult, if not impossible, to manually detect using case-by-case investigation. Promiscuity hubs and their structural analogs present a good starting point for exploring structural modifications responsible for alternating promiscuity levels.

With this in mind, we decided to make our data structures for promiscuity analysis of kinase inhibitors publicly available. In the next chapter we further analyze PC pathways and hubs.

Chapter 8

Data Structures for Compound Promiscuity Analysis: Cliffs, Pathways, and Hubs Formed by Inhibitors of the Human Kinome

Introduction

Data structures described in *Chapter 6* and *Chapter 7* provide ample opportunities for the study of structure-promiscuity relationships among kinase inhibitors. Thus, we were determined to make them publicly available to enable further exploration.

In this data note, the applications and limitations of these data structures were discussed. In addition, promiscuity hub analysis was extended and a subset of high-priority hubs were defined. Hub neighborhoods were analyzed to identify structural analogs of clinical candidate hubs. Furthermore, to assess their structural relationships, promiscuity hubs were organized into analog series using the recently developed compound-core relationship (CCR) method.

Data Note

For reprint orders, please contact: reprints@future-science.com

Future Science OA

Data structures for compound promiscuity analysis: promiscuity cliffs, pathways and promiscuity hubs formed by inhibitors of the human kinome

Filip Miljković[1] & Jürgen Bajorath*,[1]

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

*Author for correspondence: Tel.: +49 228 736 9100; Fax: +49 228 736 9101; bajorath@bit.uni-bonn.de

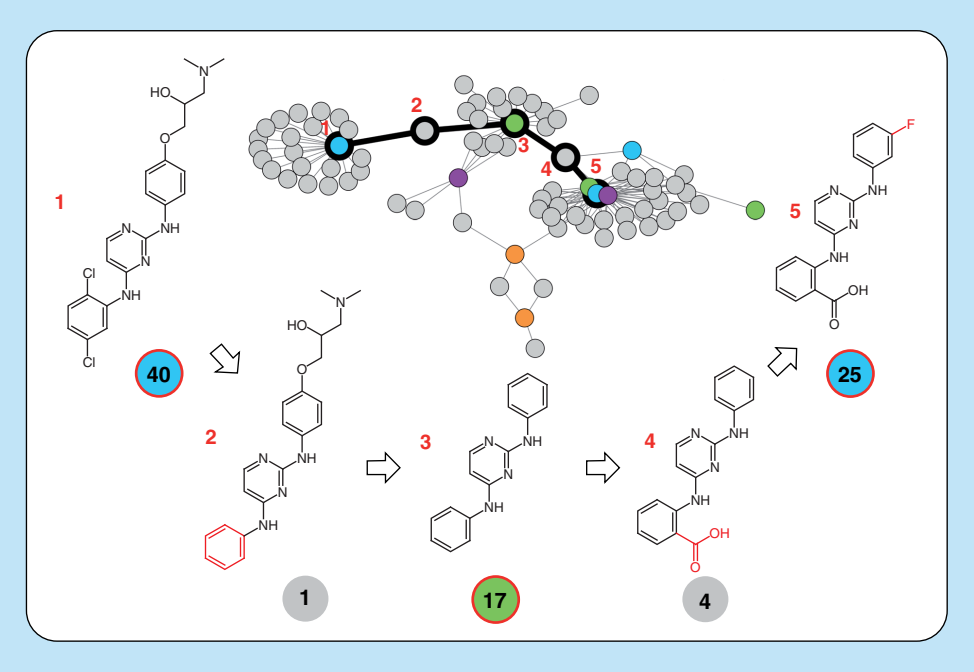
Aim: A large collection of promiscuity cliffs (PCs), PC pathways (PCPs) and promiscuity hubs (PHs) formed by inhibitors of human kinases is made freely available. **Methodology:** Inhibitor PCs were systematically identified and organized in network representations, from which PCPs were extracted. PH compounds were classified and their neighborhoods analyzed. **Data & exemplary results:** Nearly 16,000 PCs covering the human kinome were identified, which yielded more than 600 PC clusters and 8900 PCPs. Moreover, 520 PHs were obtained. **Limitations & next steps:** PC and PCP data structures capture structure–promiscuity relationships. Promiscuity assessment is also affected by data sparseness. Given the rapid growth of kinase inhibitor data, the relevance of PC/PCP/PH information for medicinal chemistry and chemical biology applications will further increase.

Lay abstract: Promiscuity cliffs (PCs) are formed by structurally very similar (analogous) compounds with large differences in the number of targets they are active against. Inhibitors of human kinases are of high interest in drug discovery and so are PCs they form. This is the case because these PCs reveal structural modifications of inhibitors that strongly influence promiscuity (multitarget activity). Sequences of overlapping PCs form pathways that are rich in structure–promiscuity relationship information. PCs and PCPs of inhibitors covering the human kinome have been systematically identified and these data are made freely available as a basis for further investigations.

First draft submitted: 27 March 2019; Accepted for publication: 28 May 2019; Published online: 25 July 2019

Keywords: compound promiscuity • human kinome coverage • kinase inhibitors • open access data • promiscuity cliff pathways • promiscuity cliffs • promiscuity hubs • structure–promiscuity relationships

Graphical abstract:



A cluster from a promiscuity cliff (PC) network is shown here in which compounds are represented as nodes and PCs as edges. Nodes are color-coded according to different promiscuity degrees (number of kinase annotations). A PC pathway formed by five kinase inhibitors (1–5) is highlighted (black) and their promiscuity degrees are reported in color-coded circles (red borders indicate promiscuity hubs). Structural modifications distinguishing pairs of compounds along the pathway are colored red.

Compound promiscuity refers to the ability of small molecules to specifically bind to multiple targets [1]. Promiscuity provides the basis for ligand-based polypharmacology [1,2], an emerging concept in drug discovery [2] that represents a departure from the single-target specificity paradigm that has for long dominated drug-discovery efforts [3]. It should be noted that the term promiscuity is often also used with a negative connotation, when referring to compound aggregation- or reactivity-based assay artifacts [4,5]. However, herein promiscuity exclusively refers to genuine multitarget activity of small molecules.

Inhibitors of human kinases are a good example for the interplay between drug polypharmacology and target specificity or selectivity. The efficacy of kinase inhibitor drugs used in oncology clearly depends on multikinase engagement and ensuing polypharmacology [6], whereas the use of kinase inhibitors in other therapeutic areas such as immunology and inflammation or metabolic diseases mostly depends on kinase selectivity [7]. Experimental and computational approaches have been used to analyze promiscuity and selectivity of kinase inhibitors [8–10].

The promiscuity cliff (PC) concept was introduced to aid in the analysis of structure–promiscuity relationships [11,12], in other words, to identify small chemical changes that lead to large apparent differences in promiscuity between structurally analogous compounds. Accordingly, a PC was formally defined as a pair of analogs with a large difference in promiscuity [11]. PCs have been identified among compounds with activity against many therapeutic targets [12] including protein kinases [13]. A large-scale analysis of currently available kinase inhibitors covering more than 80% of the human kinome yielded nearly 16,000 PCs [13].

The formation of PCs can be visualized in network representations where compounds are nodes and edges pairwise PC relationships between nodes [13]. In such networks, PCs form clusters of varying size and complexity. PC clusters are disjoint subgraphs in a PC network. These clusters are rich in structure–promiscuity relationship information, but difficult to analyze. Therefore, as an extension of the PC concept, the PC pathway (PCP) data structure was introduced [13,14]. PCPs are formed in PC clusters and consist of sequences of PCs with overlapping compounds. PCP compounds have alternating high and low promiscuity (or are highly promiscuous and nonpromiscuous). They can be systematically extracted from PC clusters using a computational search method [14]. Nearly 16,000 PCs formed by inhibitors of human kinases were organized in more than 600 separate network clusters [13] and

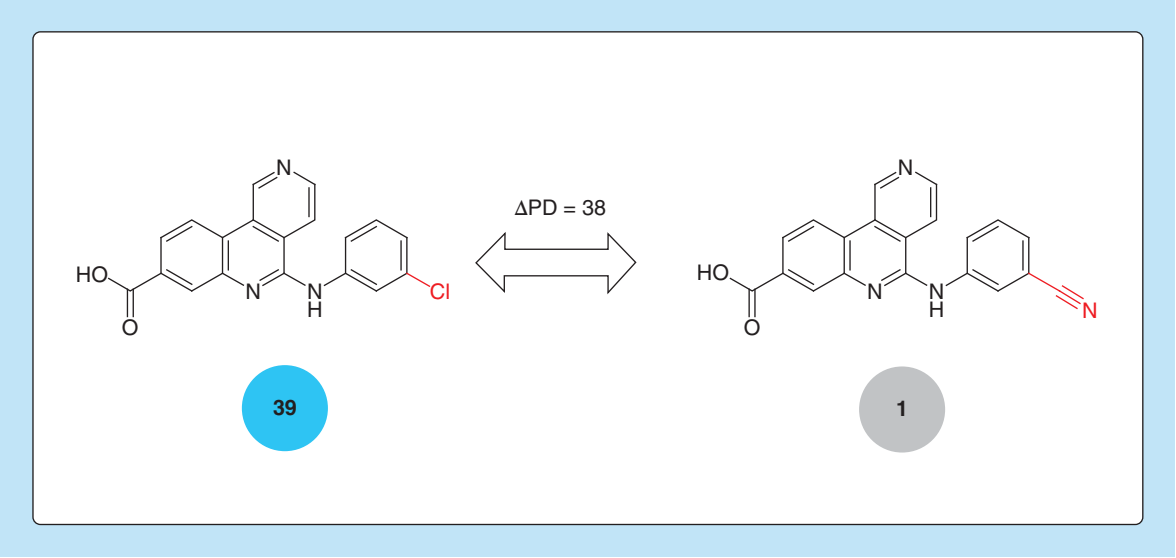


Figure 1. Promiscuity cliff. Shown is an exemplary promiscuity cliff formed by a highly promiscuous and a nonpromiscuous kinase inhibitor. The promiscuity degree of each compound is reported in a color-coded circle. The structural modification distinguishing the cliff compounds is colored red.

from these clusters, 8900 PCPs were isolated [14]. PCPs often contain so-called promiscuity hubs (PHs). Following network terminology, hubs are densely connected nodes. PHs are highly promiscuous PCP compounds that form large numbers of PCs with weakly or nonpromiscuous structural analogs outside the PCP [14]. PHs also occur in network regions outside PCPs.

PCs, PCPs and PHs provide a wealth of hypotheses for structural determinants of promiscuity and also for additional targets of weakly or nonpromiscuous compounds. For example, structural analogs of a PH might not have been tested against many confirmed PH targets and thus additional targets might be inferred for individual analogs. Taken together, PCs, PCPs and PHs provide valuable information for medicinal chemistry or chemical biology projects. This data note details an open access deposition of PCs identified across the human kinome [13], PCPs extracted from their network clusters [14] and PHs formed by individual kinase inhibitors including clinical compounds. These data are made freely available in an organized and easily accessible form.

Methodology

Kinase inhibitor data

Inhibitors of human kinases were collected from several public data sources and activity data were curated [13]. A total of 112,624 unique inhibitors with well-defined activity measurements were obtained that were active against 426 human kinases, corresponding to 82.2% of the human kinome. After removal of potential assay interference compounds [4,5], 105,492 inhibitors remained for PC analysis. For each inhibitor, its promiscuity degree (PD) was calculated as a total number of kinases it was active against.

Matched molecular pairs

For human kinase inhibitors, matched molecular pairs (MMPs) were generated by systematic fragmentation of exocyclic single bonds [15]. An MMP represents a pair of compounds that are only distinguished by a single chemical modification, termed transformation [15]. Transformation size restrictions were introduced as follows [16]: the MMP core shared by two compounds was required to have at least twice the size (number of nonhydrogen atoms) of the transformation substructures. In addition, each substructure was permitted to consist of at most 13 nonhydrogen atoms and their size difference was limited to at most eight atoms. An MMP with these transformation size restrictions represents a pair of structural analogs [16]. An exemplary MMP is shown in Figure 1. The transformation substructures are highlighted in red and the common core structure of the compound (MMP core) is shown in black.

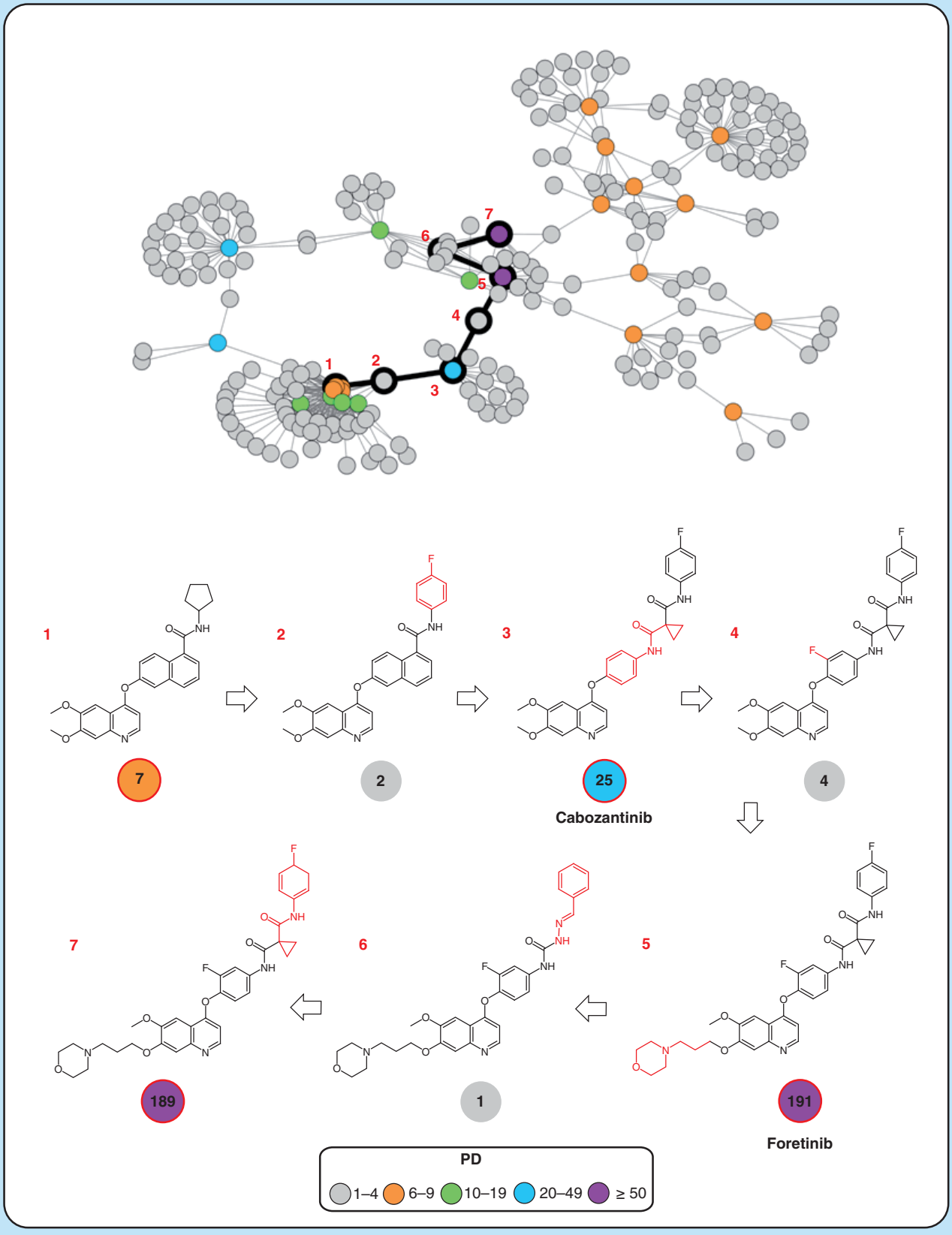


Figure 2. Promiscuity cliff cluster and pathway. Shown is a promiscuity cliff cluster in which a promiscuity cliff pathway formed by seven kinase inhibitors (1–7) is highlighted (black). Nodes are color-coded according to different promiscuity degree values. Structures of promiscuity cliff pathway compounds are shown and their promiscuity degrees are reported in color-coded circles. Structural modifications distinguishing pairs of compounds along the pathway are colored red. Inhibitors 1, 3 (cabozantinib), 5 (foretinib) and 7 are promiscuity hubs. Cabozantinib and foretinib are clinical kinase inhibitors.
PD: Promiscuity degree.

Promiscuity cliffs

The definition of PCs requires the consideration of two criteria including the *similarity* criterion and *promiscuity difference* criterion. For PC analysis of kinase inhibitors, these criteria were set as follows [13]:

- *Similarity:* formation of transformation size-restricted MMPs.
- *Promiscuity difference:* ΔPD ≥5; PD range of weakly promiscuous MMP inhibitor: [1,4].

We deliberately restrict the PD range of weakly promiscuous compounds in PCs to low promiscuity values (PD <5). This restriction avoids the generation of PC pairs consisting of two highly promiscuous inhibitors (e.g., PD = 40 and PD = 20), which would not be meaningful. Accordingly, this definition ensures that a PC is consistently formed by a highly and weakly or nonpromiscuous (PD = 1) inhibitor. Accordingly, a kinase inhibitor PC with smallest possible PD sum is formed by a pair of inhibitors with PD values of 6 and 1, respectively. Figure 1 shows an exemplary PC.

Promiscuity cliff pathways

PCPs are defined as linear subgraphs of PC clusters and consist of compounds with alternating high and low promiscuity [13]. From PC network clusters, PCPs are systematically extracted using an algorithm based on breadth-first search for shortest paths [14]. In breadth-first search, edges between neighboring nodes have equal length. Therefore, the shortest path between two nodes is determined as the path containing the smallest number of edges. For visualization, PC clusters in which PCPs are traced are drawn using the Kamada–Kawai force-directed layout algorithm [17].

Promiscuity hubs

PHs are densely connected nodes in a PC network. For our current analysis, PHs are defined as inhibitors forming at least 10 PCs with structural analogs having a PD value of 1–4 (corresponding to a PH node degree ≥10). As a reference, in the global kinase inhibitor PC network, the mean node degree was approximately 3. We note that PHs may or may not participate in the formation of PCPs. Special attention was paid to kinase inhibitors at different stages of clinical development (clinical kinase inhibitors) [9] that qualified as PHs. Furthermore, PHs are organized into analog series (ASs) using the compound–core relationship algorithm [18]. This MMP-based method systematically extracts ASs with single or multiple substitution sites from compound collections [18]. For compound–core relationship calculations, transformation size-restricted MMPs are applied (as described above).

Data & exemplary results

PCs & clusters

The 105,492 kinase inhibitors yielded 15,939 PCs that were formed by 10,741 unique inhibitors including 1653 compounds with PD ≥6. These PCs had ΔPD values ranging from 5 to 294. In a global kinase inhibitor PC network, the 15,939 PCs were organized in 622 separate clusters that contained 2 to 633 inhibitors forming 1 to 1351 PCs [13]. Figure 2 shows an exemplary PC cluster.

Promiscuity cliff pathways & promiscuity hubs

From the 622 PC clusters, a total of 8900 PCPs were algorithmically extracted via breadth-first search (see above). For further methodological details, the interested reader is referred to the original publication [14]. These PCPs consisted of 3 to 17 nodes. The characteristic feature of PCPs is their sequence of alternating highly and weakly promiscuous (or nonpromiscuous) compounds. For each PCP, the cumulative ΔPD was calculated over all pairs of nodes. These ΔPD values ranged from 10 to 869. In the PC cluster in Figure 2, a PCP is traced that consists of seven inhibitors including clinical kinase inhibitors cabozantinib (compound 3; 25 kinase annotations) and foretinib (compound 5), a pan-kinome inhibitor with 191 kinase annotations.

In the global PC network, a total of 520 inhibitors (4.8%) qualified as PHs on the basis of the criteria given above. Most PCPs (7749; 87.1%) contained at least one PH. The 520 PHs were involved in the formation of 12,131 PCs (76.1%) with 7278 weakly or nonpromiscuous structural analogs. The 12,131 PCs included 6997 PCs that involved 4300 inhibitors having a single kinase annotation. Thus, more than half of the PCs with highly promiscuous PHs involved nonpromiscuous compounds. These findings emphasized the high information content of PH network neighborhoods, revealing many possible structure–promiscuity relationships and suggesting a wealth of kinase target hypotheses for structural analogs of PHs.

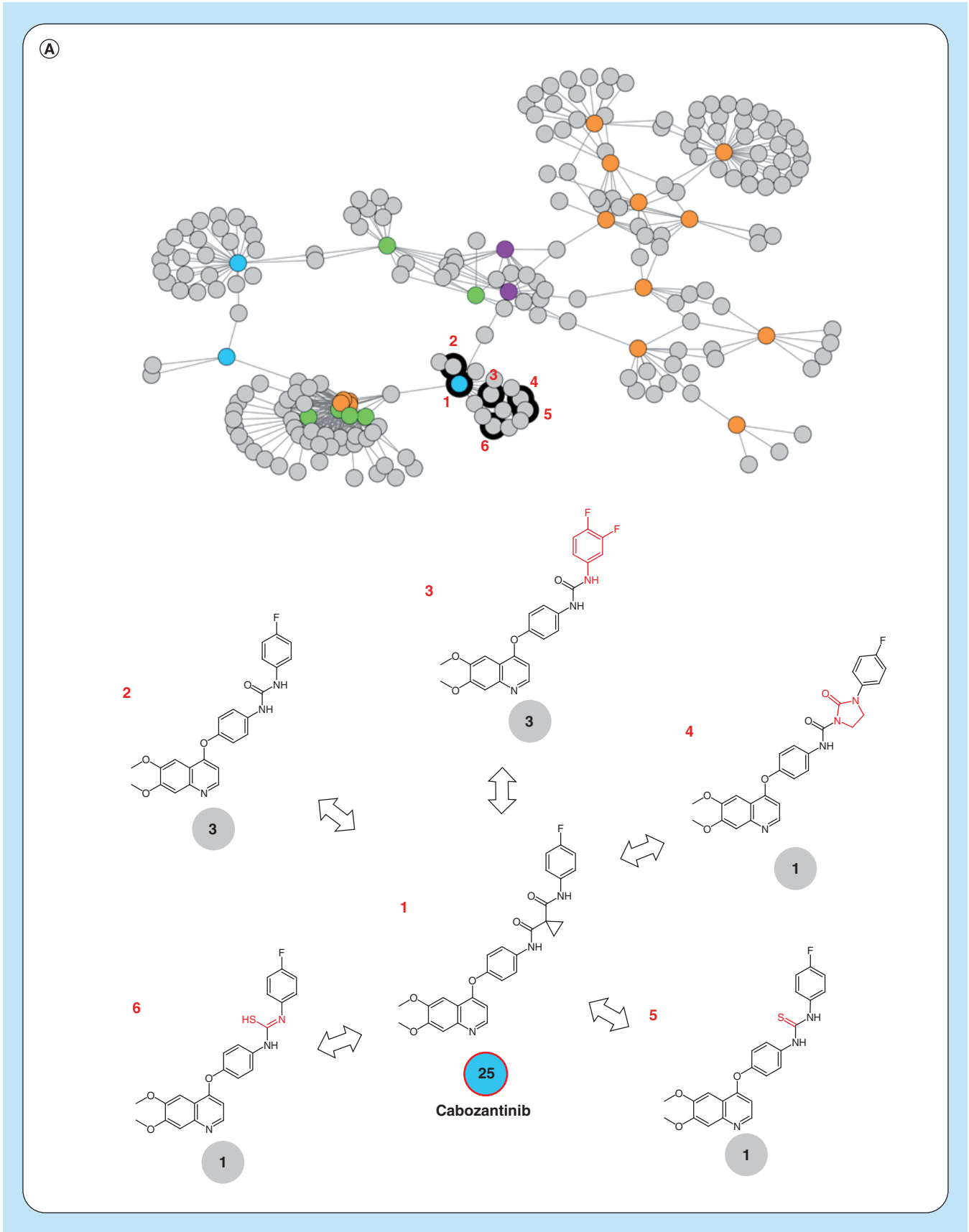


Figure 3. Promiscuity hub neighborhoods. Shown are promiscuity cliffs from the network neighborhoods of **(A)** cabozantinib and **(B)** foretinib. The presentation is according to Figure 2. Exemplary inhibitors forming promiscuity cliffs with the two clinically relevant promiscuity hubs are numbered and their structures are shown.

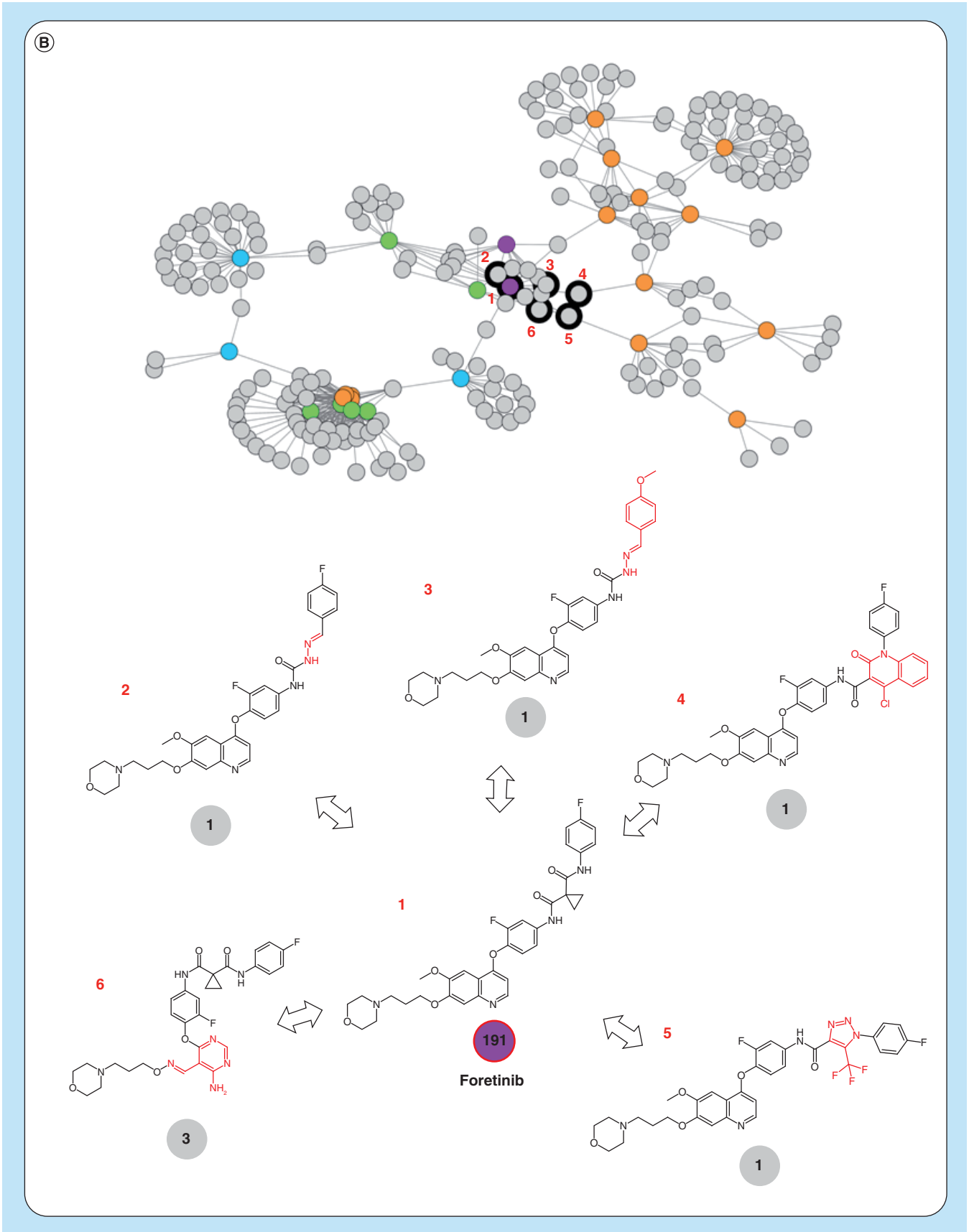


Figure 3. Promiscuity hub neighborhoods (cont.). Shown are promiscuity cliffs from the network neighborhoods of **(A)** cabozantinib and **(B)** foretinib. The presentation is according to Figure 2. Exemplary inhibitors forming promiscuity cliffs with the two clinically relevant promiscuity hubs are numbered and their structures are shown.

PHs formed by clinical kinase inhibitors

The 520 PHs contained 68 clinical kinase inhibitors, 37 of which were also found in PCPs. Cabozantinib and foretinib, shown in Figure 2, were two of these clinically relevant PHs occurring in PCPs. Their neighborhoods are depicted in Figure 3A and B, respectively. The 68 clinical PHs formed more than 90% of all PCs involving 129 clinical kinase inhibitors contained in the global network, thus highlighting their central role for promiscuity exploration.

Analog relationships between PHs

To further explore structural relationships between PHs, we also investigated whether they might form ASs. A subset of 334 of the 520 PHs was found to form 88 ASs that consisted of 2 to 25 analogs. The remaining 186 PHs were not involved in analog relationships. Thus, PHs were not only structurally closely related inhibitors but also included a variety of other compounds. However, for each of these highly promiscuous kinase inhibitors, 10 or more weakly or nonpromiscuous structural analogs were available. Therefore, PHs and their neighborhoods provide many opportunities for experimental follow-up investigations.

Data deposition

The collection of kinase inhibitor PCs, PCPs and PHs is made available in three separate, tab-delimited text files. In addition, a readme.txt file specifies all entries and abbreviations in the PC, PCP and PH data files.

For each PC, the Simplified Molecular Input Line Entry Systems (SMILES) [19] representation of the inhibitors, SMILES pattern of the transformation, common MMP core, compound identifiers and PD value of the inhibitors are provided.

For each PCP, the pathway identifier, pathway length, list of compounds forming the PCP, available PHs and cumulative ΔPD value are given.

Furthermore, all PHs are listed with their compound identifiers from the PC file and SMILES representations and clinical kinase inhibitors are identified.

The PC, PCP, PH and readme files are provided in an open access deposition on the ZENODO platform [20].

Limitations & next steps

Data sparseness is likely to affect promiscuity analysis. Sparseness refers to the situation that not all inhibitors might have been extensively tested against all kinases. Importantly, PCs, PCPs and PHs uncover all detectable promiscuity patterns, regardless of whether they reveal structural determinants of promiscuity or provide additional target hypotheses. Thus, these data structures make it possible to further investigate structure–promiscuity relationships and their potential origins in detail. Increasing availability of x-ray structures enables further exploration of PCs. Potential origins of PC formation can be investigated on the basis of protein–ligand interactions taking active site characteristics into consideration. Structural analysis of PCs on a larger scale than currently possible is expected to provide new insights into structural patterns that are responsible for promiscuous versus selective binding events.

Kinase inhibitor data have rapidly grown in recent years, more so than could have been anticipated, and there is no end in sight. Therefore, we will continue to search for PCs, PCPs and PHs and periodically update our collections to further support promiscuity exploration.

Author contributions

J Bajorath conceived the study. F Miljković carried out the analysis. J Bajorath and F Miljković analyzed the results and prepared the manuscript.

Acknowledgments

The authors thank OpenEye Scientific Software and Chemical Computing Group for free academic licenses.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background
- Compound promiscuity is defined.
- The promiscuity cliff (PC) and promiscuity cliff pathway (PCP) concepts are introduced.
- Promiscuity hubs (PHs) are introduced in the context of PC networks.

Methodology
- PC criteria are specified.
- PC cluster and PCP analysis are described.
- PH neighborhoods are discussed.

Data & exemplary results
- PC, PCP and PH statistics are reported.
- Exemplary PC clusters, PCPs and PHs are presented.
- Clinical kinase inhibitors forming PHs are analyzed.
- Structural relationships between PHs are systematically detected.
- An open access deposition of PCs, PCPs and PHs is described.

Limitations & next steps
- The potential influence of data sparseness on promiscuity is discussed.
- Further growth of kinase inhibitor data is anticipated.
- Data growth motivates continued promiscuity exploration.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Hu Y, Bajorath J. Compound promiscuity: what can we learn from current data? *Drug Discov. Today* 18(13–14), 644–650 (2013).

• **Evaluation of compound promiscuity as a basis for polypharmacology.**

2. Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery. *J. Med. Chem.* 57(19), 7874–7887 (2014).
3. Nurse P. The ends of understanding. *Nature* 387(6634), 657–657 (1997).
4. McGovern SL, Caselli E, Grigorieff N, Shoichet BK. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* 45(8), 1712–1722 (2002).
5. Baell J, Walters MA. Chemical con artists foil drug discovery. *Nature* 513(7519), 481–483 (2014).
6. Knight ZA, Lin H, Shokat KM. Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* 10(2), 130–137 (2010).
7. Simmons DL. Targeting kinases: a new approach to treating inflammatory rheumatic diseases. *Curr. Opin. Pharmacol.* 13(3), 426–434 (2013).

• **Kinase inhibition for different therapeutic indications.**

8. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* 29(11), 1039–1045 (2011).
9. Klaeger S, Heinzlmeir S, Wilhelm M *et al.* The target landscape of clinical kinase drugs. *Science* 358(6367), 4368 (2017).

•• **Extensive cell-based profiling of clinical kinase inhibitors.**

10. Miljković F, Bajorath J. Exploring selectivity of multikinase inhibitors across the human kinome. *ACS Omega* 3(1), 1147–1153 (2018).
11. Dimova D, Hu Y, Bajorath J. Matched molecular pair analysis of small molecule microarray data identifies promiscuity cliffs and reveals molecular origins of extreme compound promiscuity. *J. Med. Chem.* 55(22), 10220–10228 (2012).

•• **Introduction of the promiscuity cliff (PC) concept.**

12. Dimova D, Gilberg E, Bajorath J. Identification and analysis of promiscuity cliffs formed by bioactive compounds and experimental implications. *RSC Adv.* 7(1), 58–66 (2017).
13. Miljković F, Bajorath J. Computational analysis of kinase inhibitors identifies promiscuity cliffs across the human kinome. *ACS Omega* 3(12), 17295–17308 (2018).

• **Large-scale analysis of PCs formed by human kinase inhibitors.**

14. Miljković F, Vogt M, Bajorath J. Systematic computational identification of promiscuity cliff pathways formed by inhibitors of the human kinome. *J. Comput. Aided Mol. Des.* 33(6), 559–572 (2019).

- **Extraction of PC pathways from kinase inhibitor PC network clusters.**

15. Hussain J, Rea C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* 50(3), 339–348 (2010).
16. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J. MM P-cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* 52(5), 1138–1145 (2012).
17. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* 31(1), 7–15 (1989).
18. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* 4(1), 1027–1032 (2019).
19. Weininger D SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28(1), 31–36 (1988).
20 . Miljković F, Bajorath J. Promiscuitycliffs (PCs), promiscuity cliff pathways (PCPs), and promiscuity hubs (PHs) formed by inhibitors of human kinases (open access data deposition). doi: 10.5281/zenodo.2611184 (2019).

Summary

A set of ∼16,000 PCs organized in ∼600 clusters yielded 8900 PC pathways using the automated extraction method. Moreover, 520 promiscuity hubs were obtained that had at least 10 weakly or non-promiscuous analogs per hub. A subset of 334 hubs was found to form 88 analog series. The remaining 186 hubs were not involved in any analog relationships. As part of this study, PCs, PC pathways, and promiscuity hubs were made publicly available.

Classification of kinase inhibitors on the basis of different binding modes is made possible using X-ray crystallography data. Machine learning methods can be applied to investigate various classification tasks.

In the next chapter, we report machine learning models for the classification of kinase inhibitors with different binding modes.

Chapter 9

Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes

Introduction

X-ray structures of kinase-inhibitor complexes revealed different inhibitor binding modes and well-defined conformational states of kinase binding sites. Various combinations of "in" and "out" states of the DFG motif and αC-helix defined these binding modes.

We employed machine learning methods to classify kinase inhibitors on the basis of specific binding modes using molecular graph representations. Kinase inhibitors were extracted from Protein Data Bank (PDB) and divided into type I (DFG_{in} / αC_{in}), type I$^{1}/_{2}$ (DFG_{in} / αC_{out}), type II (DFG_{out} / αC_{out}), and allosteric inhibitors using the KLIFS database. Global and balanced models for binary classification were built using three machine learning methods: random forest, support vector machine, and deep neural network. In addition, multi-task learning was used to simultaneously classify kinase inhibitors according to the four binding modes.

Journal of Medicinal Chemistry

Article

Cite This: J. Med. Chem. XXXX, XXX, XXX−XXX

pubs.acs.org/jmc

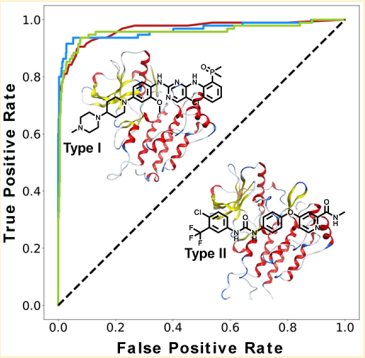
Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes

Filip Miljković,[†] Raquel Rodríguez-Pérez,[†,‡] and Jürgen Bajorath*[,†]

[†]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

[‡]Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Strasse 65, 88397 Biberach/Riß, Germany

ABSTRACT: Noncovalent inhibitors of protein kinases have different modes of action. They bind to the active or inactive form of kinases, compete with ATP, stabilize inactive kinase conformations, or act through allosteric sites. Accordingly, kinase inhibitors have been classified on the basis of different binding modes. For medicinal chemistry, it would be very useful to derive mechanistic hypotheses for newly discovered inhibitors. Therefore, we have applied different machine learning approaches to generate models for predicting different classes of kinase inhibitors including types I, $I^1/_2$, and II as well as allosteric inhibitors. These models were built on the basis of compounds with binding modes confirmed by X-ray crystallography and yielded unexpectedly accurate and stable predictions without the need for deep learning. The results indicate that the new machine learning models have considerable potential for practical applications. Therefore, our data sets and models are made freely available.



INTRODUCTION

Tyrosine and serine/threonine kinases are major drug targets,[1] and kinase inhibitors are among the most intensely investigated drug candidates in oncology and beyond.[1−3] Nearly 115000 kinase inhibitors with well-defined activity measurements have accumulated in the public domain,[4] making these inhibitors also preferred compound classes for large-scale activity data analysis[4] or the evaluation of computational screening methods.[5] Experimental efforts to identify and characterize kinase inhibitors continue to expand. For example, kinase profiling experiments and kinome scans have become major sources of kinase inhibitor activity and selectivity data.[6−8] Furthermore, kinases and their complexes with many different inhibitors have been extensively studied by X-ray crystallography,[9,10] providing essential insights into structural features of kinases and binding characteristics of their inhibitors.

X-ray structures of kinase−inhibitor complexes have revealed different binding modes of inhibitors that correlate with defined conformational changes in binding sites.[11−16] Conformational determinants of different binding modes include the activation loop with the DFG tripeptide motif,[13] which opens or closes the ATP binding site region, and the αC-helix.[14] The activation loop is located at the entrance of the ATP binding site proximal to the catalytic site and the αC-helix adjacent to the ATP site. If the activation loop is closed, adopting the so-called "DFG in" conformation, the kinase is active. By contrast, if the loop opens ("DFG out") the kinase becomes inactive.[13] In addition, the αC-helix forms a conserved E−K salt bridge ("αC-helix in" conformation) that is involved in coordinating phosphate groups of bound ATP. If this salt bridge is disrupted, the helix moves out of its position ("αC-helix out"), which renders the kinase inactive.[14] Hence, the fully active form of a kinase is characterized by the "DFG in/αC-helix in" conformational state combination, whereas the inactive form is characterized by the "DFG out/αC-helix out" combination.

The majority of currently available kinase inhibitors competitively bind to the ATP cofactor binding site in the active form of kinases and are designated type I inhibitors.[13] The ATP binding site is largely conserved across the kinome. By contrast, type II inhibitors bind to the inactive form of a kinase and are accommodated in an induced pocket adjacent to the ATP binding site (often called "back pocket") that opens up as a consequence of the "out" movements of the DFG motif and αC-helix.[11,15] This region is less conserved across kinases and type II inhibitors were thus originally expected to be more selective than type I inhibitors. Moreover, type $I^1/_2$ inhibitors were found to bind to an intermediate "DFG in/αC-helix out" conformational combination, which sets them apart from type I and II inhibitors.[12] In addition to active site-directed inhibitors, other types of noncovalent inhibitors have been discovered that bind to (induced) allosteric sites in kinases and are frequently designated type III or IV inhibitors.[13] Type III inhibitors bind to an allosteric pocket proximal to the ATP site, whereas type IV inhibitors

Special Issue: Artificial Intelligence in Drug Discovery

Received: May 29, 2019

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

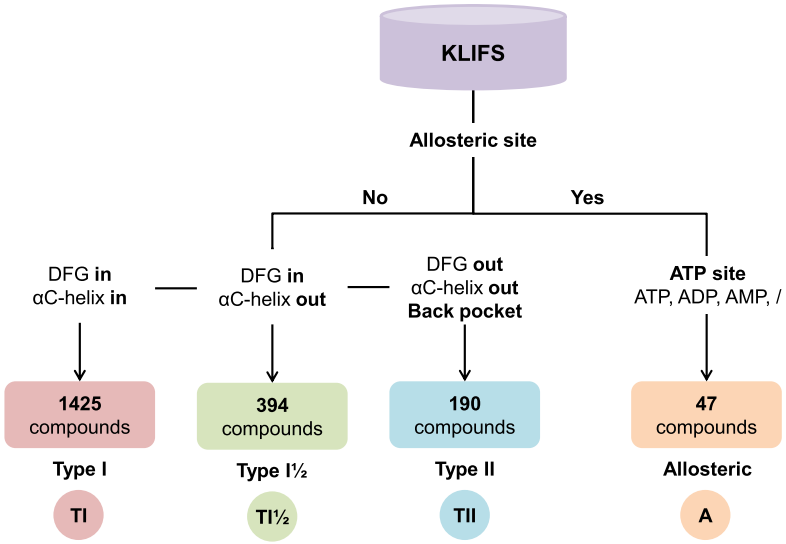


Figure 1. Compound selection. The structure-based inhibitor selection scheme is summarized. "T" stands for "Type" and "/" means "no ligand".

occupy allosteric pockets distant from the ATP binding site region.

Although type II inhibitors were originally anticipated to have higher selectivity than type I inhibitors, such differences have not been confirmed experimentally,[15] as both selective and nonselective types I and II inhibitors have frequently been found. However, some $I^1/_2$ inhibitors have shown high selectivity for individual kinases.[16] Furthermore, kinase binding profiles of limited (but growing) numbers of allosteric inhibitors discovered so far indicate that these types of inhibitors are more selective than active site-directed inhibitors,[13,16] as one might expect.

While structural biology has yielded many insights into inhibitor binding modes and conformational determinants, there are already many more kinase inhibitors available than could possibly be characterized structurally. Moreover, although structural features including a hydrogen bond donor−acceptor function such as an amide or urea group and hydrophobic moieties binding into the DFG pocket below the αC-helix have been identified to characterize some type II inhibitors,[11,15] distinguishing between different types of kinase inhibitors on the basis of molecular structure is not straightforward. Kinase inhibitors represent much more of a structural continuum than discrete subsets, and differences between them are often subtle and difficult to relate to alternative binding modes. For example, many type II inhibitors were found to contain a type I head fragment and similar core structures[17] and also fragments with hydrogen bonding capacity or hydrophobic character that were more characteristic for type II inhibitors.[17,18] Such hydrogen bonding and hydrophobic tail fragments yielded 70 different combinations,[18] which together with shared type I/II fragments resulted in a structural continuum among these inhibitors.

We have asked the question whether it might be possible to differentiate between kinase inhibitors with different crystallographically confirmed binding modes only on the basis of compound structure without taking additional interaction information into account. Therefore, current state-of-the-art machine learning methods were applied to generate a variety of predictive models.

Kinases and their inhibitors have previously been subjected to machine learning exercises beyond virtual screening. For example, given the popularity of kinase profiling campaigns, computational models have been generated to predict kinase activity profiles of inhibitors[19] or systematically evaluate potential kinase−inhibitor interactions.[20] Furthermore, multitask learning strategies were applied to distinguish between highly and weakly potent kinase inhibitors.[21] Moreover, in an interesting application on kinases (rather than inhibitors), random forest models were generated to predict activity-relevant conformational states of kinases.[22]

However, to our knowledge, it has thus far not been attempted to distinguish between kinase inhibitors adopting different binding modes on the basis of molecular graph representations via machine learning. Deriving such predictive models is also relevant for the practice of medicinal chemistry to develop binding mode hypotheses for newly identified inhibitors, providing a basis for the design of inhibitor type-specific optimization strategies. In the following, we report the derivation of various machine learning models including multitask learning that predict different types of kinase inhibitors with high accuracy. These models were built using existing machine learning approaches in the absence of new experimental data. However, their surprisingly high accuracy and notable potential for practical medicinal chemistry applications inspired us to present this work to a medicinal chemistry audience and make our data sets and models freely available.

RESULTS

Different Types of Kinase Inhibitors. Type I, $I^1/_2$, II, and allosteric kinase inhibitors were extracted from X-ray structures of kinase−inhibitor complexes contained in the KLIFS database,[9,10] a specialized repository for kinase structures. A total of 4365 X-ray structures of catalytic domains of 287 human kinases in complex with 2969 unique inhibitors were obtained. Figure 1 summarizes the compound selection approach. Initially, structures were divided into complexes that contained or did not contain inhibitors in allosteric binding sites. Complexes with allosteric inhibitors were further considered only if they contained ATP, ADP, AMP (or

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

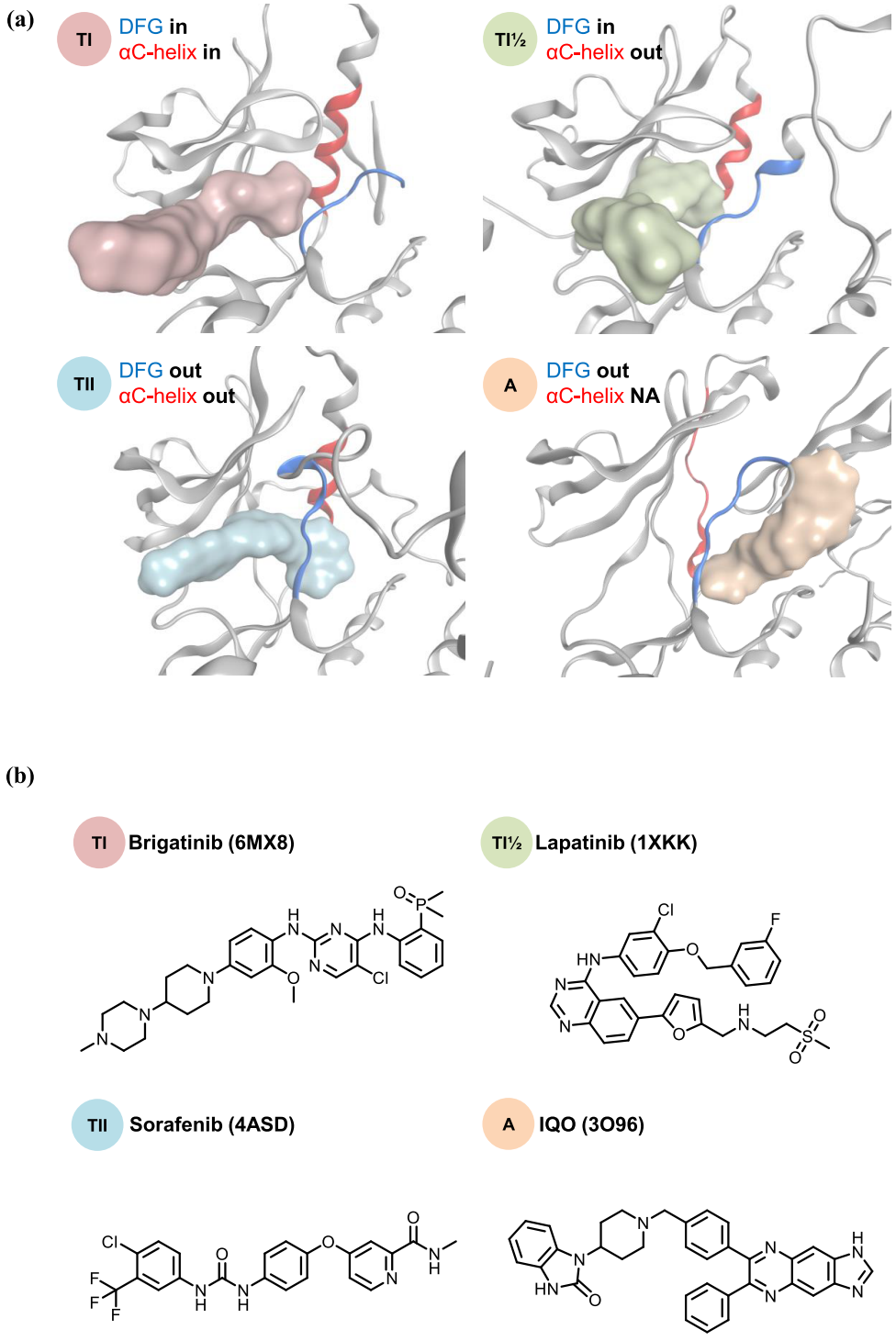


Figure 2. Representative kinase structures and inhibitors. Shown are X-ray structures of kinase−inhibitor complexes that represent different active site conformations and compound binding modes. (a) For each type of inhibitor, a representative complex is shown. For each kinase, the activation loop containing the DFG motif and the αC-helix are colored blue and red, respectively. Bound inhibitors are shown in surface representation. "NA" means "not applicable". (b) The structure of each kinase inhibitor is shown and the PDB identifier of the corresponding complex is given.

analogues thereof), or no ligand in the ATP cofactor binding site (but no other small molecule). From qualifying complexes, 47 allosteric inhibitors were obtained. All allosteric inhibitors were combined into one set in order to obtain a sufficient number of compounds for model building. Structures without allosteric inhibitors were then surveyed for combinations of conformational states of the DFG motif and the αC-helix that were characteristic of different types of kinase inhibitors. This systematic analysis led to the identification of 1425 type I ("DFG in/αC-helix in"), 394 type $I^1/_2$ ("DFG in/αC-helix out"), and 190 type II ("DFG out/αC-helix out") inhibitors (Figure 1). Designated type II inhibitors were also required to occupy the "back pocket" proximal to the ATP binding site, which becomes accessible when the "DFG out/αC-helix out"

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

conformation is adopted. Furthermore, inhibitors available in multiple X-ray structures were only selected and classified if only one conformational combination was consistently observed. If there were indications of inconsistent binding modes across different kinases, an inhibitor was omitted from further consideration.

The 2969 crystallographic inhibitors contained 2410 ATP site-directed inhibitors, 1296 of which were also available as designated kinase inhibitors in ChEMBL.[23] However, only 87 of these 2410 X-ray inhibitors (3.6%) were found to adopt alternative binding modes in different structures. Hence, there was only very little ambiguity in the structure-based assignment of kinase inhibitor types. In addition, on the basis of high-confidence compound activity data from ChEMBL,[23] the number of kinase annotations per compound (promiscuity degree) was determined for a large number of inhibitors with a single binding mode and small number of inhibitors with more than one observed binding mode. For inhibitors with single and multiple binding modes, the mean promiscuity degrees were 2.2 and 5.5, respectively, hence providing an indication that multiple binding modes of inhibitors correlated with increasing promiscuity.

The 2056 inhibitors that were assigned to different types originated from X-ray structures with 191 different human kinases. For 144 of these kinases, all inhibitors with determined structures belonged to the same type (mostly I or II). The set of classified crystallographic inhibitors was structurally diverse, as revealed by the presence of 1672 distinct Bemis–Murcko scaffolds,[24] 20 of which were found in inhibitors with different binding modes.

Moreover, computational identification of analogue series[25] showed that the 2056 X-ray inhibitors contained 163 analogue series containing a total of only 405 compounds, with on average ~2.5 analogues per series. The remaining inhibitors were singletons, consistent with the large number of distinct Bemis–Murcko scaffolds extracted from the X-ray inhibitors. For example, the 190 classified type II inhibitors contained nine small and structurally distinct analogue series comprising a total of 22 compounds. The remaining 168 type II inhibitors were singletons. Thus, potential structural bias due to the presence of large individual analogue series that might dominate inhibitor sets and classification calculations could be excluded.

Figure 2 shows representative structures and compounds for all four types of inhibitors that were considered. For type I, $I^1/_2$, and II, clinical kinase inhibitors are shown. These structures illustrate different conformational states of the DFG motif and αC-helix whose "in"/"out" combinations give rise to different binding site architectures and compound binding modes.

Inhibitor Classification via Machine Learning. Predictive models were generated for classifying kinase inhibitors according to different binding modes deduced from X-ray structures. The classification tasks are summarized in Figure 3. Models were built to distinguish between type I vs II (TI/TII), type I vs $I^1/_2$, (TI/$TI^1/_2$), type II vs $I^1/_2$ (TII/$TI^1/_2$), and A vs (I, $I^1/_2$, II) (A/(TI + $TI^1/_2$ + TII)) inhibitors. It is important to note that type I, $I^1/_2$, and II inhibitors have overlapping binding sites and act by related yet distinct mechanisms, delineating a spectrum from type I over $I^1/_2$ to type II inhibitors. Therefore, the TI/TII, TI/$TI^1/_2$, and TII/$TI^1/_2$ classification models were generated to address challenging pairwise prediction tasks. The last task A/(TI + $TI^1/_2$ + TII)

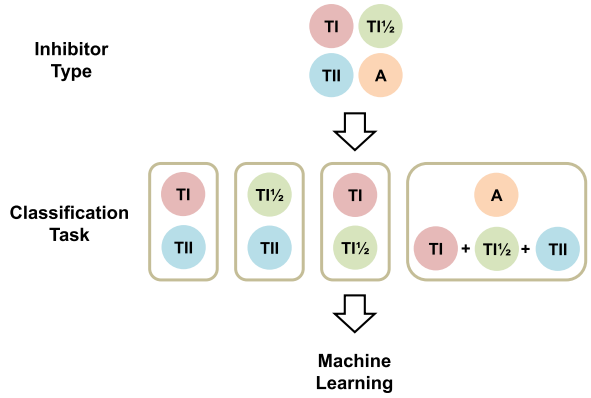


Figure 3. Derivation of classification models. Model building strategies for distinguishing between different types of kinase inhibitors are summarized.

aimed to distinguish allosteric from nonallosteric kinase inhibitors having similar yet distinct mechanisms-of-action. Therefore, to distinguish allosteric inhibitors with completely distinct mechanisms from nonallosteric inhibitors with similar mechanisms, nonallosteric inhibitors were combined in this case. As state-of-the-art machine learning approaches, random forest (RF), support vector machine (SVM), and deep neural network (DNN) algorithms were applied. For each classification task and method, global and balanced models were generated. Global models were derived on the basis of imbalanced training sets using all available inhibitors and balanced models on the basis of sets containing the same number of inhibitors for different classes (see the Experimental Section for further details).

Figure 4 summarizes the generation of training, test, and validation sets. Two different strategies were applied. Following strategy 1, compounds were divided into evenly sized training and test sets and 10 independent trials were carried out. Following strategy 2, 20% of the compounds were excluded from modeling as an external validation set and the remaining 80% were used to train and test models in 10

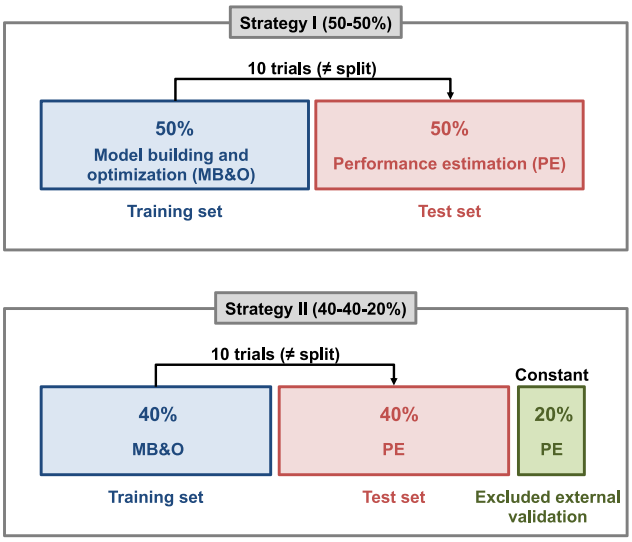


Figure 4. Training, test, and validation sets. The generation of different training, test, and external validation sets is summarized. Two different strategies (I and II) were applied.

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX–XXX

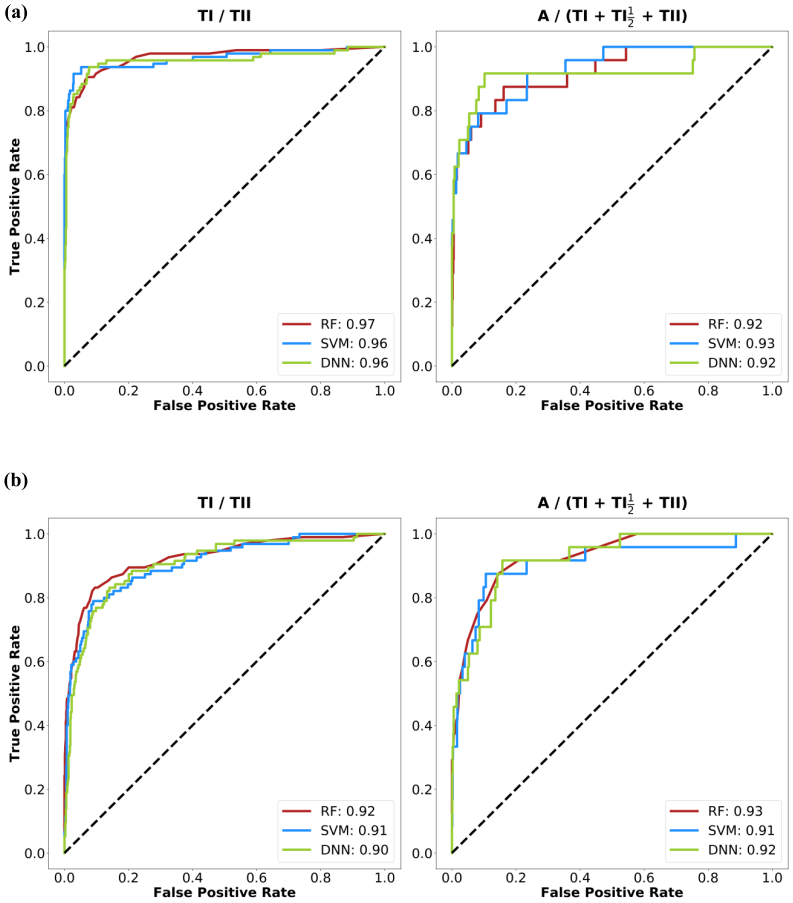


Figure 5. Exemplary ROC curves. For two classification tasks, TI/TII (left) and A/(TI + $TI^1/_2$ + TII) (right), representative ROC curves for a randomly chosen trial using global models are shown. Each graph contains three curves for RF (red), SVM (blue), and DNN (green) calculations, respectively. Inserts report AUROC values. As a molecular representation, the (a) ECFP4 and (b) MACCS fingerprint was used.

Table 1. Performance of Global Models (Strategy I)[a]

classification task	metric	RF	SVM	DNN
TI/TII	BA	0.75 ± 0.02	0.85 ± 0.02	0.88 ± 0.02
	F1	0.66 ± 0.04	0.80 ± 0.03	0.80 ± 0.02
	MCC	0.68 ± 0.03	0.79 ± 0.03	0.77 ± 0.02
TII/$TI^1/_2$	BA	0.84 ± 0.01	0.89 ± 0.02	0.88 ± 0.02
	F1	0.80 ± 0.02	0.86 ± 0.02	0.85 ± 0.03
	MCC	0.75 ± 0.02	0.80 ± 0.03	0.78 ± 0.04
TI/$TI^1/_2$	BA	0.72 ± 0.02	0.81 ± 0.02	0.81 ± 0.01
	F1	0.60 ± 0.03	0.74 ± 0.03	0.67 ± 0.02
	MCC	0.57 ± 0.03	0.69 ± 0.03	0.57 ± 0.03
A/(TI + $TI^1/_2$ + TII)	BA	0.63 ± 0.08	0.71 ± 0.07	0.70 ± 0.09
	F1	0.33 ± 0.16	0.52 ± 0.12	0.47 ± 0.19
	MCC	0.40 ± 0.10	0.53 ± 0.13	0.53 ± 0.10

[a]Reported are the mean and standard deviation (mean ± SD) of BA, F1, and MCC values for 10 independent trials using RF, SVM, and DNN global models based upon ECFP4 fingerprints.

independent trials according to strategy 1 (further details are provided in the Experimental Section). Hence, in contrast to test sets, the external validation set was kept constant and consisted of compounds that were never encountered during training and testing.

Global Models. Initially, global models derived following strategy I in Figure 4 were evaluated, and the predictions were monitored in ROC curves. Figure 5 shows representative examples of individual trials of RF, SVM, and DNN models generated for different prediction tasks using two alternative molecular representations, the extended connectivity finger-

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 2. Performance of Balanced Models (Strategy I)[a]

classification task	metric	RF	SVM	DNN
TI/TII	BA	0.88 ± 0.01	0.90 ± 0.02	0.84 ± 0.07
	F1	0.87 ± 0.02	0.90 ± 0.02	0.86 ± 0.05
	MCC	0.76 ± 0.03	0.80 ± 0.04	0.70 ± 0.11
TII/TI$^1/_2$	BA	0.87 ± 0.01	0.90 ± 0.02	0.86 ± 0.03
	F1	0.86 ± 0.01	0.90 ± 0.02	0.87 ± 0.03
	MCC	0.76 ± 0.02	0.81 ± 0.04	0.74 ± 0.06
TI/TI$^1/_2$	BA	0.79 ± 0.02	0.81 ± 0.01	0.70 ± 0.03
	F1	0.78 ± 0.02	0.81 ± 0.02	0.75 ± 0.02
	MCC	0.58 ± 0.04	0.62 ± 0.03	0.43 ± 0.04
A/(TI + TI$^1/_2$ + TII)	BA	0.79 ± 0.04	0.75 ± 0.07	0.73 ± 0.05
	F1	0.79 ± 0.05	0.70 ± 0.13	0.76 ± 0.04
	MCC	0.59 ± 0.09	0.53 ± 0.13	0.49 ± 0.10

[a]Reported are the mean and standard deviation (mean ± SD) of BA, F1, and MCC values for 10 independent trials using RF, SVM, and DNN balanced models based upon ECFP4 fingerprints.

Table 3. Performance of Global Models on Test Sets (Strategy II)[a]

classification task	metric	RF	SVM	DNN
TI/TII	BA	0.77 ± 0.08	0.84 ± 0.03	0.87 ± 0.03
	F1	0.65 ± 0.10	0.79 ± 0.03	0.79 ± 0.02
	MCC	0.66 ± 0.07	0.78 ± 0.03	0.77 ± 0.03
TII/TI$^1/_2$	BA	0.83 ± 0.03	0.88 ± 0.02	0.88 ± 0.03
	F1	0.79 ± 0.04	0.85 ± 0.02	0.84 ± 0.04
	MCC	0.74 ± 0.04	0.80 ± 0.02	0.77 ± 0.06
TI/TI$^1/_2$	BA	0.73 ± 0.02	0.79 ± 0.01	0.80 ± 0.01
	F1	0.61 ± 0.04	0.71 ± 0.02	0.66 ± 0.03
	MCC	0.59 ± 0.04	0.66 ± 0.01	0.56 ± 0.04
A/(TI + TI$^1/_2$ + TII)	BA	0.63 ± 0.07	0.64 ± 0.01	0.56 ± 0.13
	F1	0.28 ± 0.16	0.39 ± 0.21	0.11 ± 0.24
	MCC	0.32 ± 0.13	0.52 ± 0.11	0.55 ± 0.13

[a]Reported are the mean and standard deviation (mean ± SD) of BA, F1, and MCC values for 10 independent trials using RF, SVM, and DNN global models based upon ECFP4 fingerprints. Predictions are reported for test sets according to strategy II.

print with bond diameter 4 (ECFP4) and MACCS structural keys. High prediction accuracy was consistently observed, with area under the ROC curve (AUROC) values of 0.9 and above. There were only small differences between calculations using the alternative fingerprints. Overall, ECFP4-based calculations displayed marginally better performance in some cases, but there were no significant differences. Therefore, in the following, results obtained for ECFP4 are presented.

Given the consistently observed high AUROC values, alternative performance measures were considered for model comparison. Table 1 summarizes the results obtained for all classification tasks and methods on the basis of balanced accuracy (BA), F1 score, and Matthews correlation coefficient (MCC) (see the Experimental Section for details). MCC is particularly well suited for evaluating predictions on unbalanced data sets. It ranges from −1 to +1 (with +1 indicating perfect, 0 random, and −1 completely incorrect predictions).

The results in Table 1 confirm generally high prediction accuracy for all prediction tasks and methods. BA values mostly ranged from ca. 70−90%. Furthermore, MCC values consistently exceeded 0.75 for the TI/TII and TII/TI$^1/_2$ prediction tasks and ranged from 0.40 to 0.69 for the TI/TI$^1/_2$ and A/(TI + TI$^1/_2$ + TII) tasks. Overall lowest MCC values were obtained for RF (between 0.40 and 0.75 for all tasks). In addition, F1 scores mostly had values >0.6, with A/(TI + TI$^1/_2$ + TII) tasks having the lowest F1 scores (max. 0.52 for SVM). On the basis of the different performance measures, type II inhibitors were generally distinguished from type I and I$^1/_2$ inhibitors with slightly higher accuracy than type I inhibitors were distinguished from I$^1/_2$ and allosteric inhibitors from others. Notably, only 23 allosteric inhibitors were available for training, a smaller number than typically used for machine learning, especially for DNNs. Nonetheless, the resulting models were predictive. Considering all prediction tasks, SVM produced the best performing models, followed by DNN and RF, although performance differences were small.

Balanced Models. Next, models were generated on the basis of balanced training sets following strategy I, which are often more predictive in machine learning than models derived from imbalanced data. The results obtained for balanced classification models are summarized in Table 2. Again, high prediction accuracy was observed across all classification tasks and models. BA yielded high values of ∼0.9 for the first two classification tasks and slightly lower values of ∼0.8 for the other two. MCC values were consistently high (>0.75) for the TI/TII and TII/TI$^1/_2$ classification tasks. The remaining two tasks yielded lower MCC values, as observed for global models,

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

Table 4. External Validation of Global Models (Strategy II)[a]

classification task	metric	RF	SVM	DNN
TI/TII	BA	0.76 ± 0.08	0.78 ± 0.02	0.81 ± 0.02
	F1	0.63 ± 0.10	0.71 ± 0.03	0.67 ± 0.04
	MCC	0.63 ± 0.06	0.70 ± 0.04	0.63 ± 0.05
TII/TI$^1/_2$	BA	0.78 ± 0.03	0.82 ± 0.02	0.79 ± 0.03
	F1	0.71 ± 0.05	0.77 ± 0.03	0.72 ± 0.04
	MCC	0.65 ± 0.04	0.69 ± 0.03	0.61 ± 0.05
TI/TI$^1/_2$	BA	0.70 ± 0.02	0.74 ± 0.02	0.66 ± 0.05
	F1	0.56 ± 0.03	0.58 ± 0.04	0.46 ± 0.05
	MCC	0.51 ± 0.03	0.47 ± 0.05	0.27 ± 0.08
A/(TI + TI$^1/_2$ + TII)	BA	0.63 ± 0.06	0.63 ± 0.07	0.56 ± 0.12
	F1	0.34 ± 0.15	0.36 ± 0.18	0.12 ± 0.26
	MCC	0.41 ± 0.10	0.48 ± 0.09	0.60 ± 0.09

[a]Reported are the mean and standard deviation (mean ± SD) of BA, F1, and MCC values for 10 independent trials using RF, SVM, and DNN global models based upon ECFP4 fingerprints. Predictions are reported for the external validation set according to strategy II.

with lowest values (<0.5) obtained for DNN. F1 scores further increased when assessing predictions on balanced sets, with F1 > 0.7 for all models. Taken together, the results for balanced models corresponded to those obtained for global models. Both categories of models yielded accurate and stable predictions with low standard deviations across different trials. Overall, highest prediction accuracy under balanced conditions was observed for SVM models, followed by RF, and DNN models. Again, performance differences were only small.

External Validation. Strategy II was implemented to provide a constant external validation set. First, predictions for different test sets were compared. Table 3 reports results for global models according to strategy II. For the first three classification tasks, BA, F1, and MCC scores were very similar to the results obtained for global models following strategy I. For the remaining A/(TI + TI$^1/_2$ + TII) classification task, a maximal reduction of 0.3 was observed for the F1 score of DNN models.

Table 4 summarizes the performance of global models in predicting the external validation set. Overall, these predictions were of high accuracy comparable to test set predictions. The F1 score was only reduced for DNN and the A/(TI + TI$^1/_2$ + TII) classification task and the MCC value only for DNN and the TI/TI$^1/_2$ task compared to global models. We note that the largest performance differences between test and validation set predictions were found for F1 scores, in particular, for DNN models. To assess the statistical significance of such differences, nonparametric Mann−Whitney tests were performed and p-values ≪0.001 were obtained for TI/TII, TII/TI$^1/_2$, and TI/TI$^1/_2$. Therefore, for these three classification tasks, the F1 value distributions were statistically different (albeit of relatively small magnitude). In the case of A/(TI + TI$^1/_2$ + TII), the p-value was 0.48, which indicated no statistically significant differences.

The application of strategy II made it also possible to include the multitask DNN (MT-DNN) methodology into the comparison. In MT-DNNs, all predictive tasks are simultaneously modeled. These prediction tasks represent a multiclass learning problem aiming to predict mutually exclusive classes. Global MT-DNN models were built and evaluated on the basis of a 40−40−20% split of training, test, and external validation compounds.

Table 5 reports MT-DNN model performance on the test sets and the external validation set. We note that results for

Table 5. Performance of Multitask Models on Test and Validation Sets (Strategy II)[a]

class	metric	test sets	validation set
TI	BA	0.82 ± 0.02	0.73 ± 0.02
	F1	0.90 ± 0.01	0.78 ± 0.02
	MCC	0.66 ± 0.03	0.42 ± 0.04
TII	BA	0.85 ± 0.03	0.79 ± 0.01
	F1	0.77 ± 0.04	0.67 ± 0.04
	MCC	0.76 ± 0.04	0.65 ± 0.04
TI$^1/_2$	BA	0.81 ± 0.02	0.68 ± 0.02
	F1	0.70 ± 0.03	0.46 ± 0.03
	MCC	0.62 ± 0.04	0.31 ± 0.04
A	BA	0.69 ± 0.08	0.66 ± 0.08
	F1	0.49 ± 0.14	0.40 ± 0.18
	MCC	0.54 ± 0.12	0.44 ± 0.18

[a]For MT-DNN global models, the mean and standard deviation (mean ± SD) of BA, F1, and MCC values are reported for 10 independent trials predicting the test sets and external validation set.

MT-DNN and single-task ML models are not directly comparable because the performance of binary classification tasks is only assessed on the basis of compounds that are assigned to one of the two classes. The results in Table 5 show that multitask learning also produced a model that predicted test instances with high accuracy and the external validation set with only slightly reduced accuracy on the basis of MCC and BA values. We also note that the variance of test set results originated from training and testing with different subsets, whereas the validation set remained constant and only the training set used for model building differed. Because neither the variable test sets nor the constant validation set were used for model hyper-parameter optimization and originated from the same compound pool, differences in performance were only due to random selection of different test compounds for performance evaluation. The external validation was more

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

difficult to predict, as reflected by statistically significant performance differences on the basis of F1 scores for DNN models. Taken together, these results showed that predictions were accurate for independently assembled sets and also reflected the relevance of cross-validation approaches for evaluating classifier performance.

Table 6 reports the mean confusion matrix for the external validation set over 10 independent trials using MT-DNN.

Table 6. Mean Confusion Matrix for Multitask Models[a]

		observed			
		TI	TII	TI$^1/_2$	A
predicted	TI	205	4	27	3
	TII	5	23	2	1
	TI$^1/_2$	74	10	50	3
	A	1	1	0	3

[a]The mean confusion matrix over 10 independent trials is reported for predictions of the external set using global MT-DNN models.

Results across different trials were stable and revealed correct predictions of the majority of compounds per class, except for allosteric inhibitors, which represented the minority class and were overall most difficult to predict. In addition, type I$^1/_2$ inhibitors were frequently confused with type I and II inhibitors. While the majority of type I$^1/_2$ inhibitors were correctly detected, 34% were incorrectly assigned to type I inhibitors, representing the majority class.

As an additional control, we also generated SVM classifiers to distinguish between kinase inhibitors and nonkinase compounds. Initially, classification models were built using a set of 10000 randomly drawn compounds from ZINC[26] as negative training instances. This set contained 7666 Bemis–Murcko scaffolds,[24] and the compounds had an average molecular weight of 340.91 Da. Thus, ZINC compounds were chemically diverse and within the weight range of kinase inhibitors. Four different balanced models were generated to address the TI/ZINC, TI$^1/_2$/ZINC, TII/ZINC, and A/ZINC classification tasks. These models were highly predictive with largest MCC values of 0.95, 0.92, 0.91, and 0.71, respectively, and largest F1 scores of 0.97, 0.96, 0.95, and 0.86, respectively, for the different tasks. As expected, the performance of these control models was much higher compared to those generated for predicting different types of kinase inhibitors. A second series of SVM models used bioactive compounds from ChEMBL (release 24) that were not annotated with kinases as negative training instances. This set contained 202034 bioactive compounds with 78101 different scaffolds and an average weight of 458.9 Da. Again, balanced SVM models were generated for TI/ChEMBL, TI$^1/_2$/ChEMBL, TII/ChEMBL, and A/ChEMBL. The performance of all ChEMBL-based models was high (and only slightly lower than for models based on ZINC compounds). For the different tasks, the largest reported MCC values were 0.88, 0.82, 0.89, and 0.59, respectively, and the largest F1 scores were 0.94, 0.91, 0.94, and 0.81, respectively.

Finally, as another form of external validation, we have used our SVM models to screen all designated high-confidence protein kinase inhibitors (52614 inhibitors) available in ChEMBL (release 24). On the basis of these calculations, ca. 93% of all currently available kinase inhibitors were predicted to be type I inhibitors, consistent with other assessments.[27]

DISCUSSION AND CONCLUSIONS

In this study, we have investigated machine learning approaches for predicting kinase inhibitors with different binding modes. Distinguishing between type I, I$^1/_2$, II, and allosteric (III/IV) inhibitors and exploring their activity and selectivity profiles are topical issues in medicinal chemistry. For all inhibitors used for modeling, binding modes were confirmed by X-ray crystallography. However, for machine learning, compounds were only represented using molecular fingerprints, without taking other information into account. Different prediction tasks were defined to distinguish between different types of inhibitors.

When designing this study, we anticipated that these predictions would be rather challenging, given the often only subtle structural modifications of kinase inhibitors with different modes of action and the structural continuum they represent. However, we were taken by surprise. For all prediction tasks, methods, and molecular representations, the accuracy of classification models was consistently high, although there were differences between individual methods and models. Comparably high levels of accuracy were consistently achieved by global and balanced models.

Given the generally high performance levels, DNN did not offer an advantage over RF and SVM binary models. In this regard, it should also be noted that only limited amounts of training data from X-ray crystallography were available for this study, which restricted the capacity of DNN training. Moreover, our current and many other machine-learning exercises in compound classification typically make use of well-defined molecular representations such as fingerprints or arrays of numerical descriptors. The use of such representations does not play into strengths of deep learning. This is the case because performance increases of deep learning architectures over other machine learning approaches are often attributable to initial deep representation learning such as in image analysis or natural language processing. However, representation learning is not required, and hence not making an impact, if well-defined canonical representations are used, as is typically the case in compound classification and activity prediction.

On the other hand, the deep learning MT-DNN architecture enabled the implementation of a multiclass model to predict the inhibitor types, making it possible to use all available training data in concert and hence further improve the basis for deep learning. Taken together, the results of our study show that the machine learning models we have derived for predicting different types of kinase inhibitors are robust and accurate. Accordingly, these models should have considerable potential for a variety of practical applications such as, for example, the ChEMBL kinase inhibitor survey reported above. The derived models can be used to predict the type of any new kinase inhibitor. In practical applications, parallel or sequential use of predictive models can serve as a consistency check, for example, by initially predicting test compounds as type I vs II inhibitors, followed by prediction of type I vs I$^1/_2$ or II vs type I$^1/_2$ inhibitors. In addition, our findings also suggest that allosteric kinase inhibitors can be accurately distinguished from others. In this case, only small numbers of inhibitors were required to generate predictive models.

As a part of our study, we make RF, SVM, and DNN global and balanced models freely available. In addition, for each classification task, the compound data sets and precomputed ECFP4 fingerprints are provided. Models and data are made

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

available in an open access deposition on the Zenodo platform.[28] It is hoped that these models will be helpful in the practice of medicinal chemistry to characterize new kinase inhibitors.

EXPERIMENTAL SECTION

Compound Selection. Different types of kinase inhibitors were selected from KLIFS,[9,10] which collects and organizes X-ray structures of kinase−inhibitor complexes from the Protein Data Bank (PDB).[29] Information about conformational states of the DFG motif and the αC-helix and bound inhibitors was obtained from KLIFS using the open source virtual machine 3D-e-Chem-VM.[30] To exclude fragments, only inhibitors with a molecular weight of at least 250 Da were considered. For selected inhibitors, SMILES representations were generated and standardized using the OpenEye OEChem toolkit.[31]

Training and Test Sets for Global and Balanced Models. For model building, two different validation strategies were implemented, as shown in Figure 4. In the first strategy, inhibitors of each type were randomly divided into equally sized training and test subsets (i.e., 50−50% split). For each binary classification task, subsets from two different classes were combined to yield the final training and test sets. For global models, all compounds were used and thus training and test sets contained different numbers of inhibitors of each type. For balanced models, the number of randomly selected inhibitors of different type in training sets was adjusted to the smaller subset. Hence, in this case, the same numbers of training compounds with different class labels were used. Table 7 details the composition of these training sets.

Table 7. Training Set Ratios for Global and Balanced Models (Strategy I)[a]

	training set ratio	
classification task	global	balanced
TI/TII	712/95	95/95
TII/TI$^1/_2$	95/197	95/95
TI/TI$^1/_2$	712/197	197/197
A/(TI + TI$^1/_2$ + TII)	23/1004	23/23

[a]For different prediction tasks, the ratios of kinase inhibitors from different sets used to train global and balanced models following strategy I are reported.

The second strategy applied a 40−40−20% data split for training, testing, and external validation, respectively. Accordingly, 20% of the inhibitors in each set were excluded from training and testing and reserved for external validation. For predicting the external validation set, MT-DNN models were also trained and tested. For both validation strategies, internal validation was carried out using the test sets. RF and SVM model hyper-parameters were optimized using internal 10-fold cross-validation, whereas hyper-parameter optimization for DNN and MT-DNN models was carried out on the basis of an internal 80−20% training/test data split.

Molecular Representations. As molecular representations for machine learning, ECFP4[32] and MACCS[33] were used. ECFP4 is a feature set fingerprint which enumerates layered atom environments that are encoded as integers using a hashing function. MACCS is a fragment fingerprint where each of 166 bit positions encodes the presence or absence of a specific structural pattern. ECFP4 was generated using OEChem-based[31] and MACCS using RDKit-based[34] Python scripts.

Machine Learning Methods. Classification models were generated using RF, SVM, and DNN algorithms. For model building, training instances were represented by a feature vector $x \in \chi$ and associated with a class label $y \in \{0,1\}$.

Random Forest. RF is constructed from an ensemble of decision trees, each of which is built from a bootstrapped[35] sample of training data.[36] During node splitting, a random subset of features is considered for the construction of individual trees.[36] For each RF trial, the number of trees was set to 100 and class weights were considered. The minimum number of samples at a leaf node (min_samples_leaf) was optimized via 10-fold cross validation using candidate values of 1, 5, and 10. In addition, the number of features to search for the best data split (max_features) was set to the square root of the total number of features. RF calculations were performed using the Python-implemented Scikit-learn package.[37]

Support Vector Machine. SVM is a supervised machine learning algorithm that constructs a hyper-plane H to best separate two classes of objects by maximizing the distance between objects having different class labels (margin).[38] This hyper-plane is defined by a weight vector $\mathbf{w}$ and a bias b such that $H = \{\mathbf{x}|\langle \mathbf{w},\mathbf{x}\rangle + b = 0\}$. To generalize the models, slack variables are included to permit a limited number of errors for training instances that fall within the margin or on the incorrect side of H. Regularization or cost hyper-parameter C controls the relation between the training errors and margin size. Hyper-parameter C was optimized via 10-fold cross validation using values of 0.01, 0.1, 1, 10, 100, and 1000. Moreover, different class weights were also considered during SVM training to increasingly penalize errors in the minority class. If linear separation of training classes is not possible in a given feature, space kernel functions are applied.[39] The scalar product $\langle .,. \rangle$ is replaced by a kernel function $K(.,.)$ projecting the data into a higher dimensional space where linear separation is possible.[39] Herein, Tanimoto kernel[40] was used as a preferred kernel function for fingerprint representations:

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

SVM calculation protocols were implemented in Python using Scikit-learn.[32]

Deep Neural Network. A feedforward DNN derives a function that maps an input x to a class y, where $y = f(x;w)$. The function learns the value of parameter w to provide the best approximation.[41] The DNN architecture is composed of different layers of computational neurons, including an input layer, several hidden layers, and an output layer.[42] Each neuron of a hidden and the output layer accepts an n-dimensional input x and transforms it into a linear m-dimensional vector $y = W^T x + b$, where W and b are parameters of dimension (m, n) and m, respectively. Then, a nonlinear activation function $h(y)$ is applied to the weighted sum of its inputs. The weights from the network are iteratively adjusted during training on the basis of a cost function to minimize (gradient descent). Hyper-parameters were either set to constant values or optimized using internal validation with 80% vs 20% data splits.[42−45] For binary DNN models, learning rates values of 0.01 and 0.001 were evaluated. A set of network architectures (represented as the values of the output features in hidden layers) were investigated: [100, 100], [250, 250], [250, 500], [500, 250], [100, 500], [500, 100], [500, 250, 100], [100, 250, 500], and [250, 100, 250]. Thus, pyramidal, rectangular, and autoencoder architectures were used during hyper-parameter optimization. The epoch number was set to 30, and the batch size was set to 50.

Multitask (MT) learning aims to simultaneously model different classification outcomes and DNN can be extended for this purpose. In this case, an MT-DNN was implemented using an output layer with a one-hot encoded categorical outcome that consisted of four elements (one per inhibitor type). Categorical cross-entropy was used as the loss function to minimize. For MT-DNN, the candidate values for the learning rate were 0.001, 0.0001, and 0.00001. Moreover, the following architectures (nodes per hidden layer) were evaluated: [200,100], [2000,1000], [1000, 100, 100], and [2000,1000,100]. The number of epochs was set to 20 and 100 for internal and external validation, respectively, and the batch size was 64. For both single-task (ST) and MT-DNNs, the drop-out rate was set to 25%. The Adam optimization algorithm[40] was chosen as the optimization function and the "rectified linear unit" (ReLU)[46] as the activation function. For output nodes, the "softmax" activation function was used. DNN architectures were implemented using TensorFlow[47] and Keras.[48]

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

Hyperparameter Optimization. Training of models under hyper-parameter optimization revealed the best performance levels were generally already achieved using standard parameter (or close to standard parameter) settings. The observations indicated that model performance were overall stable and not dependent on very specific parameter settings for RF and SVM methods. However, some preferred parameters were identified through optimization including an architecture with three hidden layers, with following combination of nodes [100, 250, 500] and learning rate of 0.001 for DNN binary classification models.

For most of the trials, MT-DNN models provided the best mean MCC performance in internal validation with an architecture of two hidden layers and the following number of nodes: [2000, 1000], and a learning rate of 0.01. Average MCC values across all the classes ranged from 0.44 to 0.73, reflecting the importance of hyper-parameter optimization.

Performance Measures. In addition to generating ROC curves and calculating AUROC values, model performance was assessed using three different measures including balanced accuracy (BA), regular F1 score, and Matthews correlation coefficient (MCC). BA, F1, and MCC are defined as

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

$$\mathrm{F1} = 2 \times \frac{\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

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

where TP means "true positives", TN "true negatives", FP "false positives", and FN "false negatives".

AUTHOR INFORMATION

Corresponding Author

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

ORCID

Jürgen Bajorath: 0000-0002-0557-5714

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the OpenEye Scientific Software, Inc., for providing a free academic license of the OpenEye toolkit. Furthermore, we thank Swarit Jasial and Tomoyuki Miyao for helpful discussions.

ABBREVIATIONS USED

ADP, adenosine diphosphate; AMP, adenosine monophosphate; ATP, adenosine triphosphate; AUROC, area under the receiver operating characteristic; BA, balanced accuracy; DNN, deep neural network; ECFP, extended connectivity fingerprint; FP, false positives; FN, false negatives; MACCS, molecular access system; MCC, Matthews correlation coefficient; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; TP, true positives; TN, true negatives

REFERENCES

(1) Cohen, P. Protein Kinases - the Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309−315.

(2) Laufer, S.; Bajorath, J. New Frontiers in Kinases: Second Generation Inhibitors. *J. Med. Chem.* **2014**, *57*, 2167−2168.

(3) Wu, P.; Nielsen, T. E.; Clausen, M. H. Small-Molecule Kinase Inhibitors: An Analysis of FDA-Approved Drugs. *Drug Discovery Today* **2016**, *21*, 5−10.

(4) Miljković, F.; Bajorath, J. Computational Analysis of Kinase Inhibitors Identifies Promiscuity Cliffs across the Human Kinome. *ACS Omega* **2018**, *3*, 17295−17308.

(5) Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20*, 2839−2860.

(6) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127−132.

(7) Anastassiadis, 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.

(8) Elkins, J. M.; Fedele, V.; Szklarz, M.; Abdul Azeez, K. R.; Salah, E.; Mikolajczyk, J.; Romanov, S.; Sepetov, N.; Huang, X. P.; Roth, B. L.; Al Haj Zen, A.; Fourches, D.; Muratov, E.; Tropsha, A.; Morris, J.; Teicher, B. A.; Kunkel, M.; Polley, E.; Lackey, K. E.; Atkinson, F. L.; Overington, J. P.; Bamborough, P.; Müller, S.; Price, D. J.; Willson, T. M.; Drewry, D. H.; Knapp, S.; Zuercher, W. J. Comprehensive Characterization of the Published Kinase Inhibitor Set. *Nat. Biotechnol.* **2016**, *34*, 95−103.

(9) van Linden, O. P. J.; Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Knowledge-Based Structural Database to Navigate Kinase−Ligand Interaction Space. *J. Med. Chem.* **2014**, *57*, 249−277.

(10) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Res.* **2016**, *44*, D365−D371.

(11) Liu, Y.; Gray, N. S. Rational Design of Inhibitors that Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2006**, *2*, 358−364.

(12) Koeberle, S. C.; Romir, J.; Fischer, S.; Koeberle, A.; Schattel, V.; Albrecht, W.; Gruetter, C.; Werz, O.; Rauh, D.; Stehle, T.; Laufer, S. A. Skepinone-L Is a Selective p38 Mitogen-activated Protein Kinase Inhibitor. *Nat. Chem. Biol.* **2012**, *8*, 141−143.

(13) Gavrin, L. K.; Saiah, E. Approaches to Discover Non-ATP Site Kinase Inhibitors. *MedChemComm* **2013**, *4*, 41−51.

(14) Palmieri, L.; Rastelli, G. αC Helix Displacement as a General Approach for Allosteric Modulation of Protein Kinases. *Drug Discovery Today* **2013**, *18*, 407−414.

(15) Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9*, 1230−1241.

(16) Müller, S.; Chaikuad, A.; Gray, N. S.; Knapp, S. The Ins and Outs of Selective Kinase Inhibitor Development. *Nat. Chem. Biol.* **2015**, *11*, 818−821.

(17) Liu, Y.; Gray, N. S. Rational Design of Inhibitors that Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2006**, *2*, 358−364.

(18) Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9*, 1230−1241.

(19) Merget, B.; Turk, S.; Eid, S.; Rippmann, F.; Fulle, S. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J. Med. Chem.* **2017**, *60*, 474−485.

(20) Janssen, A. P.; Grimm, S. H.; Wijdeven, R. H.; Lenselink, E. B.; Neefjes, J.; van Boeckel, C. A.; van Westen, G. J.; van der Stelt, M. Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome−Inhibitor Interaction Landscapes. *J. Chem. Inf. Model.* **2019**, *59*, 1221−1229.

(21) Rodríguez-Pérez, R.; Bajorath, J. Multitask Machine Learning for Classifying Highly and Weakly Potent Kinase Inhibitors. *ACS Omega* **2019**, *4*, 4367−4375.

(22) Ung, P. M. U.; Rahman, R.; Schlessinger, A. Redefining the Protein Kinase Conformational Space with Machine Learning. *Cell Chem. Biol.* **2018**, *25*, 916−924.

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

(23) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.
(24) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(25) Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound−Core Relationship Method. *ACS Omega* **2019**, *4*, 1027−1032.
(26) Sterling, T.; Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(27) Hu, Y.; Furtmann, N.; Bajorath, J. Current Compound Coverage of the Kinome. *J. Med. Chem.* **2015**, *58*, 30−40.
(28) Miljković, F.; Rodríguez-Pérez, R.; Bajorath, J. *Machine Learning Models for Predicting Kinase Inhibitors with Different Binding Modes*; Zenodo, 2019; https://zenodo.org/record/3370478.
(29) 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.
(30) McGuire, R.; Verhoeven, S.; Vass, M.; Vriend, G.; de Esch, I. J. P.; Lusher, S. J.; Leurs, R.; Ridder, L.; Kooistra, A. J.; Ritschel, T.; de Graaf, C. 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *J. Chem. Inf. Model.* **2017**, *57*, 115−121.
(31) *OEChem, TK*, version 2.0.0; OpenEye Scientific Software, Santa Fe, NM, 2015.
(32) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(33) *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.
(34) *RDKit: Cheminformatics and Machine Learning Software*, 2013; http://www.rdkit.org (accessed Apr 25, 2019).
(35) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1−26.
(36) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(38) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
(39) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods: Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT Press: Cambridge, 1999; pp 169−184.
(40) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw* **2005**, *18*, 1093−1110.
(41) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, 2016.
(42) Nielsen, M. A. *Neural Networks and Deep Learning*; Determination Press, 2015.
(43) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.
(44) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.
(45) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *3rdInternational Conference on Learning Representations, San Diego, CA, May 7−9*, 2015; arXiv.org e-Print archive, arXiv:1412.6980. https://arxiv.org/abs/1412.6980 (accessed Apr 25, 2019).
(46) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*; Omnipress: 2010; pp 807−814.
(47) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M., Yu, Y.; Zheng, X. TensorFlow: A System for Large-scale Machine Learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Nov 2−4, 2016*; USENIX Association: Savannah, GA2016, 2016.
(48) Chollet, F. *Keras*, version 2.1.3; GitHub, 2015; https://github.com/keras-team/keras (accessed Apr 25, 2019).

DOI: 10.1021/acs.jmedchem.9b00867
J. Med. Chem. XXXX, XXX, XXX−XXX

Summary

Herein, we explored different machine learning approaches to classify kinase inhibitors on the basis of binding modes. Different molecular representations were used to represent inhibitors.

Prediction was consistently high for all tasks, methods, and molecular representations. Comparable results were obtained for both global and balanced binary models, as well as for multi-task models. The differences between machine learning methods were minor across different binary models.

These machine learning models were accurate and robust, providing opportunities for practical applications.

Chapter 10

Conclusion

With 518 members, protein kinases present one of the largest protein families in the human proteome. Dysregulation of their basal levels results in multiple diseases. In this regard, the design of small-molecule inhibitors of the human kinome has progressed therapeutic applications in oncology, immunology, cardiology, and neurology. Increasing availability of kinase inhibitor data provides opportunities for large-scale computational studies. In this thesis, selectivity and promiscuity of kinase inhibitors were extensively explored. In addition, a set of promiscuity data structures was defined to evaluate structure-promiscuity relationships of close structural analogs. Finally, machine learning methods were employed to classify kinase inhibitors on the basis of different binding modes as revealed by X-ray crystallography. This chapter summarizes the major findings of this thesis and draws final conclusions.

In the first study (*Chapter 2*), a systematic analysis of selectivity of multi-kinase inhibitors was performed on the basis of single-protein assay data from ChEMBL. As the majority of kinase inhibitors target the largely conserved active site of kinases, they are expected to be promiscuous and lack selectivity. A total of 10,060 multi-kinase inhibitors were annotated with 141 kinases, yielding 596 compound-based kinase pairs of increasing phylogenetic distances. This new reference frame for selectivity analysis showed that multi-kinase inhibitors were more selective than anticipated. Given the statistically significant sample of kinase inhibitors, clear selectivity trends were observed. Selectivity trends of multi-kinase inhibitors were further explored using activity data from cell-based assays. In *Chapter 3*, the most comprehensive kinase inhibitor profiling study

reported to date provided a basis for the systematic exploration of multi-kinase inhibitor selectivity. An approach similar to the one described in *Chapter 2* was used to explore selectivity patterns. In total 2369 compound-based kinase pairs were formed by 190 inhibitors at different stages of clinical development. Similar selectivity trends were observed with inhibitors widely distributed across kinase pairs. Furthermore, selectivity profiles were categorized into two major groups: uni-directional and bi-directional. Selectivity profiles suggested that many kinase inhibitors were able to differentiate between kinase targets with substantial differences in selectivity. Next, clinical candidates from the cell-based profiling study were associated with activity annotations from ChEMBL (*Chapter 4*). Subsets of highly selective and nonselective candidates were detected across different confidence levels with increasing stringency. No selectivity differences were detected on the basis of compound activity data for subsets of the most and least selective candidates from the profiling study and clinical candidates classified as type I and II inhibitors. A number of clinical candidates designated as chemical probes were found to be highly promiscuous. Thus, selectivity trends of clinical candidates were complementary for cell-based and medicinal chemistry-based data. In *Chapter 5*, the analysis was further extended to include expert-curated chemical probes from the Chemical Probes Portal. The PD values of chemical probes were calculated on the basis of activity annotations from ChEMBL. In addition, Portal-based PD values were calculated for the comparison. For ∼50% of well-defined probes, promiscuity levels were consistent with those reported by the Portal. Thus, highly selective chemical probes were detected across different activity data confidence levels. In addition, analog and scaffold relationships were explored to evaluate potential off-target activities. However, off-target hypotheses could not be inferred for the majority of probes. Furthermore, sets of close structural analogs with large differences in promiscuity were systematically analyzed and new promiscuity data structures were introduced. Therefore, kinase inhibitors were collected from several public databases and assembled into a curated data set to perform large-scale promiscuity analysis (*Chapter 6*). A comprehensive kinase inhibitor set was obtained consisting of 112,624 inhibitors active against 426 kinases (82% of the human kinome). Different promiscuity profiles were detected providing a basis for PC analysis. Following stringent PC criteria, ∼16,000 PCs

were formed in a predominantly coordinated manner in network representations. Therefore, a new promiscuity data structure termed PC pathway was defined to trace linear PC sequences in PC clusters.

As PC clusters become increasingly large and complex, manual identification and extraction of PC pathways becomes difficult. In the next study (*Chapter 7*), a computational method was developed to systematically identify PC pathways in clusters. Pathway parameters were defined to rank the most informative pathways. PC pathways containing promiscuity hubs were of particular interest for this study and revealed many unexpected structure-promiscuity relationships. To enable further computational and experimental follow-up analyses, PCs and PC pathways formed by kinase inhibitors were made publicly available. In *Chapter 8* the potential applications and limitations of these data structures were discussed. Approximately 16,000 PCs were organized in over 600 clusters, from which 8900 PC pathways were extracted. Furthermore, promiscuity hub analysis was extended and 520 high-quality hubs were identified. Application of recently reported CCR method revealed that the majority of hubs were not structurally related. In addition, hub neighborhoods provided many opportunities for promiscuity studies.

Different binding modes of kinase inhibitors were elucidated on the basis of X-ray structure data. In *Chapter 9*, classification of kinase inhibitors according to binding modes was investigated using a variety of machine learning approaches. Both binary and multi-task models were built that displayed high performance levels. A deep neural network did not provide an advantage over random forest and support vector machine models except for multi-task learning. Accurate and robust classification models identified structural patterns that distinguished between kinase inhibitors with different binding modes.

In conclusion, novel computational approaches for selectivity and promiscuity analysis of kinase inhibitors were introduced. The elucidation of therapeutic and biological roles of these inhibitors largely relies on the study of their selectivity/promiscuity profiles. In addition, close structural analogs with large differences in promiscuity were identified to explore structure-promiscuity relationships of kinase inhibitors and derive new target hypotheses for non-promiscuous analogs. To these ends, new data structures were introduced to facilitate promiscuity analysis in network representations. Finally, machine

learning models were generated to classify kinase inhibitors into different binding modes. Taken together, these findings provided a sound basis for further computational analyses of kinase inhibitors and application in kinase drug discovery.

Bibliography

[1] Manning, G.; Whyte, D. B.; Martinez, R.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934.

[2] Miranda-Saavedra, D.; Barton, G. J. Classification and Functional Annotation of Eukaryotic Protein Kinases. *Proteins* **2007**, *68*, 893–914.

[3] Ward, R. A., Goldberg, F. W., Eds. *Kinase Drug Discovery*; Royal Society of Chemistry, 2018.

[4] Cohen, P. Protein Kinases - The Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discovery* **2002**, *1*, 309–315.

[5] Roskoski, R. A Historical Overview of Protein Kinases and Their Targeted Small Molecule Inhibitors. *Pharmacol. Res.* **2015**, *100*, 1–23.

[6] Cohen, P. The Structure and Regulation of Protein Phosphatases. *Annu. Rev. Biochem.* **1989**, *58*, 453–508.

[7] Hunter, T. Protein Kinases and Phosphatases: The Yin and Yang of Protein Phosphorylation and Signaling. *Cell* **1995**, *80*, 225–236.

[8] Hanks, S. K.; Hunter, T. Protein Kinases 6. The Eukaryotic Protein Kinase Superfamily: Kinase (Catalytic) Domain Structure and Classification. *FASEB J.* **1995**, *9*, 576–596.

[9] Knighton, D.; Zheng, J.; Ten Eyck, L.; Ashford, V.; Xuong, N.; Taylor, S.; Sowadski, J. Crystal Structure of the Catalytic Subunit of Cyclic Adenosine Monophosphate-Dependent Protein Kinase. *Science* **1991**, *253*, 407–414.

[10] Taylor, S. S.; Kornev, A. P. Protein Kinases: Evolution of Dynamic Regulatory Proteins. *Trends Biochem. Sci.* **2011**, *36*, 65–77.

[11] Vijayan, R. S. K.; He, P.; Modi, V.; Duong-Ly, K. C.; Ma, H.; Peterson, J. R.; Dunbrack, R. L.; Levy, R. M. Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. *J. Med. Chem.* **2015**, *58*, 466–479.

[12] Kooistra, A. J.; Volkamer, A. *Annu. Rep. Med. Chem.*; Elsevier, 2017; pp 197–236.

[13] Roskoski, R. Cyclin-Dependent Protein Serine/Threonine Kinase Inhibitors as Anticancer Drugs. *Pharmacol. Res.* **2019**, *139*, 471–488.

[14] Hari, S. B.; Merritt, E. A.; Maly, D. J. Sequence Determinants of a Specific Inactive Protein Kinase Conformation. *Chem. Biol.* **2013**, *20*, 806–815.

[15] Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Ten Eyck, L. F. Surface Comparison of Active and Inactive Protein Kinases Identifies a Conserved Activation Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 17783–17788.

[16] Kornev, A. P.; Taylor, S. S.; Ten Eyck, L. F. A Helix Scaffold for the Assembly of Active Protein Kinases. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 14377–14382.

[17] Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacol. Res.* **2016**, *103*, 26–48.

[18] Liu, Y.; Shah, K.; Yang, F.; Witucki, L.; Shokat, K. M. A Molecular Gate which Controls Unnatural ATP Analogue Recognition by the Tyrosine Kinase v-Src. *Bioorg. Med. Chem.* **1998**, *6*, 1219–1226.

[19] Dar, A. C.; Shokat, K. M. The Evolution of Protein Kinase Inhibitors from Antagonists to Agonists of Cellular Signaling. *Annu. Rev. Biochem.* **2011**, *80*, 769–795.

[20] Ung, P. M.-U.; Rahman, R.; Schlessinger, A. Redefining the Protein Kinase Conformational Space with Machine Learning. *Cell Chem. Biol.* **2018**, *25*, 916–924.e2.

[21] Huse, M.; Kuriyan, J. The Conformational Plasticity of Protein Kinases. *Cell* **2002**, *109*, 275–282.

[22] Kumar, S.; Boehm, J.; Lee, J. C. p38 MAP Kinases: Key Signalling Molecules as Therapeutic Targets for Inflammatory Diseases. *Nat. Rev. Drug Discovery* **2003**, *2*, 717–726.

[23] Patterson, H.; Nibbs, R.; McInnes, I.; Siebert, S. Protein Kinase Inhibitors in the Treatment of Inflammatory and Autoimmune Diseases. *Clin. Exp. Immunol.* **2014**, *176*, 1–10.

[24] Simmons, D. L. Targeting Kinases: A New Approach to Treating Inflammatory Rheumatic Diseases. *Curr. Opin. Pharmacol.* **2013**, *13*, 426–434.

[25] Force, T.; Pombo, C. M.; Avruch, J. A.; Bonventre, J. V.; Kyriakis, J. M. Stress-Activated Protein Kinases in Cardiovascular Disease. *Circ. Res.* **1996**, *78*, 947–953.

[26] Mueller, B. K.; Mack, H.; Teusch, N. Rho Kinase, a Promising Drug Target for Neurological Disorders. *Nat. Rev. Drug Discovery* **2005**, *4*, 387–398.

[27] Duan, W.; Wong, W. S. F. Targeting Mitogen-Activated Protein Kinases for Asthma. *Curr. Drug Targets* **2006**, *7*, 691–698.

[28] Knight, Z. A.; Lin, H.; Shokat, K. M. Targeting the Cancer Kinome Through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130–137.

[29] Laufer, S.; Bajorath, J. New Frontiers in Kinases: Second Generation Inhibitors. *J. Med. Chem.* **2014**, *57*, 2167–2168.

[30] Druker, B. J.; Talpaz, M.; Resta, D. J.; Peng, B.; Buchdunger, E.; Ford, J. M.; Lydon, N. B.; Kantarjian, H.; Capdeville, R.; Ohno-Jones, S.; Sawyers, C. L. Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *N. Engl. J. Med.* **2001**, *344*, 1031–1037.

[31] Heinrich, M. C. et al. Kinase Mutations and Imatinib Response in Patients With Metastatic Gastrointestinal Stromal Tumor. *J. Clin. Oncol.* **2003**, *21*, 4342–4349.

[32] Druker, B. J. et al. Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *N. Engl. J. Med.* **2006**, *355*, 2408–2417.

[33] Motzer, R. J.; Escudier, B.; Oudard, S.; Hutson, T. E.; Porta, C.; Bracarda, S.; Grünwald, V.; Thompson, J. A.; Figlin, R. A.; Hollaender, N.; Urbanowitz, G.; Berg, W. J.; Kay, A.; Lebwohl, D.; Ravaud, A. Efficacy of Everolimus in Advanced Renal Cell Carcinoma: A Double-Blind, Randomised, Placebo-Controlled Phase III Trial. *The Lancet* **2008**, *372*, 449–456.

[34] Escudier, B. et al. Sorafenib in Advanced Clear-Cell Renal-Cell Carcinoma. *N. Engl. J. Med.* **2007**, *356*, 125–134.

[35] Motzer, R. J.; Hutson, T. E.; Tomczak, P.; Michaelson, M. D.; Bukowski, R. M.; Rixe, O.; Oudard, S.; Negrier, S.; Szczylik, C.; Kim, S. T.; Chen, I.; Bycott, P. W.; Baum, C. M.; Figlin, R. A. Sunitinib versus Interferon Alfa in Metastatic Renal-Cell Carcinoma. *N. Engl. J. Med.* **2007**, *356*, 115–124.

[36] Shepherd, F. A. et al. Erlotinib in Previously Treated Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **2005**, *353*, 123–132.

[37] Geyer, C. E. et al. Lapatinib plus Capecitabine for HER2-Positive Advanced Breast Cancer. *N. Engl. J. Med.* **2006**, *355*, 2733–2743.

[38] Moore, M. J. et al. Erlotinib Plus Gemcitabine Compared With Gemcitabine Alone in Patients With Advanced Pancreatic Cancer: A Phase III Trial of the National Cancer Institute of Canada Clinical Trials Group. *J. Clin. Oncol.* **2007**, *25*, 1960–1966.

[39] Ferguson, F. M.; Gray, N. S. Kinase inhibitors: The Road Ahead. *Nat. Rev. Drug Discovery* **2018**, *17*, 353–377.

[40] Fleischmann, R.; Kremer, J.; Cush, J.; Schulze-Koops, H.; Connell, C. A.; Bradley, J. D.; Gruben, D.; Wallenstein, G. V.; Zwillich, S. H.;

Kanik, K. S. Placebo-Controlled Trial of Tofacitinib Monotherapy in Rheumatoid Arthritis. *N. Engl. J. Med.* **2012**, *367*, 495–507.

[41] Haselmayer, P.; Camps, M.; Liu-Bujalski, L.; Nguyen, N.; Morandi, F.; Head, J.; O'Mahony, A.; Zimmerli, S. C.; Bruns, L.; Bender, A. T.; Schroeder, P.; Grenningloh, R. Efficacy and Pharmacodynamic Modeling of the BTK Inhibitor Evobrutinib in Autoimmune Disease Models. *J. Immunol.* **2019**, *202*, 2888–2906.

[42] Norman, P. Investigational Bruton's Tyrosine Kinase Inhibitors for the Treatment of Rheumatoid Arthritis. *Expert Opin. Invest. Drugs* **2016**, *25*, 891–899.

[43] Rodrigues, D. A.; Sagrillo, F. S.; Fraga, C. A. M. Duvelisib: A 2018 Novel FDA-Approved Small Molecule Inhibiting Phosphoinositide 3-Kinases. *Pharmaceuticals* **2019**, *12*, e69.

[44] Kudelka, M. R.; Grossniklaus, H. E.; Mandell, K. J. Emergence of Dual VEGF and PDGF Antagonists in the Treatment of Exudative Age-Related Macular Degeneration. *Expert Rev. Ophthalmol.* **2013**, *8*, 475–484.

[45] Batson, J. et al. Development of Potent, Selective SRPK1 Inhibitors as Potential Topical Therapeutics for Neovascular Eye Disease. *ACS Chem. Biol.* **2017**, *12*, 825–832.

[46] Gharwan, H.; Groninger, H. Kinase Inhibitors and Monoclonal Antibodies in Oncology: Clinical Implications. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 209–227.

[47] U.S. Food & Drug Administration. Drugs@FDA. https://www.accessdata.fda.gov/scripts/cder/daf/, Accessed May 21, 2019.

[48] Roskoski, R. Properties of FDA-Approved Small Molecule Protein Kinase Inhibitors. *Pharmacol. Res.* **2019**, *144*, 19–50.

[49] Ren, R.; Li, G.; Le, T. D.; Kopczynski, C.; Stamer, W. D.; Gong, H. Netarsudil Increases Outflow Facility in Human Eyes Through Multiple Mechanisms. *Investig. Opthalmology Vis. Sci.* **2016**, *57*, 6197–6209.

[50] Hudes, G. et al. Temsirolimus, Interferon Alfa, or Both for Advanced Renal-Cell Carcinoma. *N. Engl. J. Med.* **2007**, *356*, 2271–2281.

[51] Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23*, e908.

[52] Klaeger, S. et al. The Target Landscape of Clinical Kinase Drugs. *Science* **2017**, *358*, eaan4368.

[53] Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation. *J. Med. Chem.* **2010**, *53*, 2681–2694.

[54] Monod, J.; Changeux, J.-P.; Jacob, F. Allosteric Proteins and Cellular Control Systems. *J. Mol. Biol.* **1963**, *6*, 306–329.

[55] Gavrin, L. K.; Saiah, E. Approaches to Discover Non-ATP Site Kinase Inhibitors. *Med. Chem. Commun.* **2013**, *4*, 41–51.

[56] Lamba, V.; Ghosh, I. New Directions in Targeting Protein Kinases: Focusing Upon True Allosteric and Bivalent Inhibitors. *Curr. Pharm. Des.* **2012**, *18*, 2936–2945.

[57] Hu, Y.; Bajorath, J. Entering the 'Big Data' Era in Medicinal Chemistry: Molecular Promiscuity Analysis Revisited. *Futur. Sci. OA* **2017**, *3*, FSO179.

[58] Hu, Y.; Furtmann, N.; Bajorath, J. Current Compound Coverage of the Kinome. *J. Med. Chem.* **2014**, *58*, 30–40.

[59] Gaulton, A. et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2016**, *45*, D945–D954.

[60] Hu, Y.; Jasial, S.; Bajorath, J. Promiscuity Progression of Bioactive Compounds Over Time. *F1000Research* **2015**, *4*, e118.

[61] 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.

[62] Martin, E.; Mukherjee, P. Kinase-Kernel Models: Accurate *in Silico* Screening of 4 Million Compounds Across the Entire Human Kinome. *J. Chem. Inf. Model.* **2012**, *52*, 156–170.

[63] Urich, R.; Wishart, G.; Kiczun, M.; Richters, A.; Tidten-Luksch, N.; Rauh, D.; Sherborne, B.; Wyatt, P. G.; Brenk, R. De Novo Design of Protein Kinase Inhibitors by *in Silico* Identification of Hinge Region-Binding Fragments. *ACS Chem. Biol.* **2013**, *8*, 1044–1052.

[64] Schmidt, F.; Matter, H.; Hessler, G.; Czich, A. Predictive *in Silico* Off-Target Profiling in Drug Discovery. *Future Med. Chem.* **2014**, *6*, 295–317.

[65] Frye, S. V. The Art of the Chemical Probe. *Nat. Chem. Biol.* **2010**, *6*, 159–161.

[66] Leach, A. R.; Gillet, V. J. *An Introduction To Chemoinformatics*; Springer Netherlands, 2007; pp 1–25.

[67] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

[68] Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015**, *7*, e23.

[69] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29*, 97–101.

[70] Weininger, D. SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures. *J. Chem. Inf. Model.* **1990**, *30*, 237–243.

[71] Brugger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Model.* **1976**, *16*, 105–110.

[72] Glen, R. C.; Rose, V. S. Computer Program Suite for the Calculation, Storage and Manipulation of Molecular Property and Activity Descriptors. *J. Mol. Graphics* **1987**, *5*, 79–86.

[73] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH Verlag GmbH: Weinheim, Germany, 2000.

[74] Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.

[75] Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.

[76] Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 260–282.

[77] Bajorath, J. Molecular Crime Scene Investigation - Dusting for Fingerprints. *Drug Discov. Today Technol.* **2013**, *10*, e491–e498.

[78] *MACCS Structural Keys*; Accelrys: San Diego, CA, 2011.

[79] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[80] Liu, Y. A Comparative Study on Feature Selection Methods for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823–1828.

[81] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.

[82] Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Med. Chem.* **2009**, *1*, 451–466.

[83] Hu, Y.; Bajorath, J. Systematic Assessment of Molecular Selectivity at the Level of Targets, Bioactive Compounds, and Structural Analogues. *ChemMedChem* **2015**, *11*, 1362–1370.

[84] Hopkins, A. L. Network Pharmacology: The Next Paradigm in Drug Discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.

[85] Boran, A. D. W.; Iyengar, R. Systems Approaches to Polypharmacology and Drug Discovery. *Curr. Opin. Drug Discov. Devel.* **2010**, *13*, 297–309.

[86] Anighoro, A.; Bajorath, J.; Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.* **2014**, *57*, 7874–7887.

[87] Hu, Y.; Bajorath, J. Compound Promiscuity: What Can We Learn from Current Data? *Drug Discov. Today* **2013**, *18*, 644–650.

[88] Fabian, M. A. et al. A Small Molecule-Kinase Interaction Map for Clinical Kinase Inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.

[89] Karaman, M. W. et al. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

[90] Cheng, A. C.; Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53*, 4502–4510.

[91] Metz, J. T.; Johnson, E. F.; Soni, N. B.; Merta, P. J.; Kifle, L.; Hajduk, P. J. Navigating the Kinome. *Nat. Chem. Biol.* **2011**, *7*, 200–202.

[92] Anastassiadis, 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.

[93] Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046–1051.

[94] Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.* **2014**, *9*, 1230–1241.

[95] Levitzki, A. Tyrosine Kinase Inhibitors: Views of Selectivity, Sensitivity, and Clinical Performance. *Annu. Rev. Pharmacol. Toxicol.* **2013**, *53*, 161–185.

[96] Müller, S.; Chaikuad, A.; Gray, N. S.; Knapp, S. The Ins and Outs of Selective Kinase Inhibitor Development. *Nat. Chem. Biol.* **2015**, *11*, 818–821.

[97] Liu, Y.; Gray, N. S. Rational Design of Inhibitors that Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2006**, *2*, 358–364.

[98] Chaikuad, A.; Koch, P.; Laufer, S. A.; Knapp, S. The Cysteinome of Protein Kinases as a Target in Drug Development. *Angew. Chem. Int. Ed.* **2018**, *57*, 4372–4385.

[99] Baillie, T. A. Targeted Covalent Inhibitors for Drug Design. *Angew. Chem. Int. Ed.* **2016**, *55*, 13408–13421.

[100] Skuta, C.; Popr, M.; Muller, T.; Jindrich, J.; Kahle, M.; Sedlak, D.; Svozil, D.; Bartunek, P. Probes & Drugs Portal: An Interactive, Open Data Resource for Chemical Biology. *Nat. Methods* **2017**, *14*, 759–760.

[101] Arrowsmith, C. H. et al. The Promise and Peril of Chemical Probes. *Nat. Chem. Biol.* **2015**, *11*, 536–541.

[102] Workman, P.; Collins, I. Probing the Probes: Fitness Factors For Small Molecule Tools. *Chem. Biol.* **2010**, *17*, 561–577.

[103] Mortlock, A. A. et al. Discovery, Synthesis, and *in Vivo* Activity of a New Class of Pyrazoloquinazolines as Selective Inhibitors of Aurora B Kinase. *J. Med. Chem.* **2007**, *50*, 2213–2224.

[104] Quintás-Cardama, A. et al. Preclinical Characterization of the Selective JAK1/2 inhibitor INCB018424: Therapeutic Implications for the Treatment of Myeloproliferative Neoplasms. *Blood* **2010**, *115*, 3109–3117.

[105] Koeberle, S. C.; Fischer, S.; Schollmeyer, D.; Schattel, V.; Grütter, C.; Rauh, D.; Laufer, S. A. Design, Synthesis, and Biological Evaluation of Novel Disubstituted Dibenzosuberones as Highly Potent and Selective

Inhibitors of p38 Mitogen Activated Protein Kinase. *J. Med. Chem.* **2012**, *55*, 5868–5877.

[106] Koeberle, S. C.; Romir, J.; Fischer, S.; Koeberle, A.; Schattel, V.; Albrecht, W.; Grütter, C.; Werz, O.; Rauh, D.; Stehle, T.; Laufer, S. A. Skepinone-L is a Selective p38 Mitogen-Activated Protein Kinase Inhibitor. *Nat. Chem. Biol.* **2011**, *8*, 141–143.

[107] Karpov, A. S. et al. Optimization of a Dibenzodiazepine Hit to a Potent and Selective Allosteric PAK1 Inhibitor. *ACS Med. Chem. Lett.* **2015**, *6*, 776–781.

[108] Jasial, S.; Hu, Y.; Bajorath, J. Determining the Degree of Promiscuity of Extensively Assayed Compounds. *PLOS ONE* **2016**, *11*, e0153873.

[109] Hu, Y.; Bajorath, J. Exploring the Scaffold Universe of Kinase Inhibitors. *J. Med. Chem.* **2015**, *58*, 315–332.

[110] Hu, Y.; Kunimoto, R.; Bajorath, J. Mapping of Inhibitors and Activity Data to the Human Kinome and Exploring Promiscuity from a Ligand and Target Perspective. *Chem. Biol. Drug Des.* **2017**, *89*, 834–845.

[111] Stumpfe, D.; Tinivella, A.; Rastelli, G.; Bajorath, J. Promiscuity of Inhibitors of Human Protein Kinases at Varying Data Confidence Levels and Test Frequencies. *RSC Advances* **2017**, *7*, 41265–41271.

[112] 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.

[113] Shoichet, B. K. Screening in a Spirit Haunted World. *Drug Discov. Today* **2006**, *11*, 607–615.

[114] 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.

[115] Baell, J.; Walters, M. A. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.

[116] Kim, D. W.; Jo, Y. S.; Jung, H. S.; Chung, H. K.; Song, J. H.; Park, K. C.; Park, S. H.; Hwang, J. H.; Rha, S. Y.; Kweon, G. R.; Lee, S.-J.; Jo, K.-W.; Shong, M. An Orally Administered Multitarget Tyrosine Kinase Inhibitor, SU11248, Is a Novel Potent Inhibitor of Thyroid Oncogenic RET/Papillary Thyroid Cancer Kinases. *J. Clin. Endocrinol. Metab.* **2006**, *91*, 4070–4076.

[117] Faivre, S.; Delbaldo, C.; Vera, K.; Robert, C.; Lozahic, S.; Lassau, N.; Bello, C.; Deprimo, S.; Brega, N.; Massimini, G.; Armand, J.-P.; Scigalla, P.; Raymond, E. Safety, Pharmacokinetic, and Antitumor Activity of SU11248, a Novel Oral Multitarget Tyrosine Kinase Inhibitor, in Patients With Cancer. *J. Clin. Oncol.* **2006**, *24*, 25–35.

[118] Sonpavde, G.; Hutson, T. E. Pazopanib: A Novel Multitargeted Tyrosine Kinase Inhibitor. *Curr. Oncol. Rep.* **2007**, *9*, 115–119.

[119] Kumar, R.; Harrington, L. E.; Hopper, T. M.; Miller, C. G.; Onori, J. A.; Cheung, M.; Stafford, J. A.; Epperly, A. H.; Gilmer, T. M. Correlation of Anti-Tumor and Anti-Angiogenic Activity of VEGFR Inhibitors with Inhibition of VEGFR2 Phosphorylation in Mice. *J. Clin. Oncol.* **2005**, *23*, 9537–9537.

[120] Trudel, S.; Li, Z. H.; Wei, E.; Wiesmann, M.; Chang, H.; Chen, C.; Reece, D.; Heise, C.; Stewart, A. K. CHIR-258, a novel, multitargeted tyrosine kinase inhibitor for the potential treatment of t(4;14) multiple myeloma. *Blood* **2005**, *105*, 2941–2948.

[121] Pan, B.-S. et al. MK-2461, a Novel Multitargeted Kinase Inhibitor, Preferentially Inhibits the Activated c-Met Receptor. *Cancer Res.* **2010**, *70*, 1524–1533.

[122] Hu, Y.; Bajorath, J. High-Resolution View of Compound Promiscuity. *F1000Research* **2013**, *2*, e144.

[123] Jalencas, X.; Mestres, J. On the Origins of Drug Polypharmacology. *Med. Chem. Commun.* **2013**, *4*, 80–87.

[124] Hu, Y.; Bajorath, J. Influence of Search Parameters and Criteria on Compound Selection, Promiscuity, and Pan Assay Interference Characteristics. *J. Chem. Inf. Model.* **2014**, *54*, 3056–3066.

[125] Hu, Y.; Bajorath, J. Analyzing Compound Activity Records and Promiscuity Degrees in Light of Publication Statistics. *F1000Research* **2016**, *5*, e1227.

[126] Hu, Y.; Bajorath, J. Promiscuity Profiles of Bioactive Compounds: Potency Range and Difference Distributions and the Relation to Target Numbers and Families. *MedChemComm* **2013**, *4*, 1196–1201.

[127] Dimova, D.; Bajorath, J. Assessing Scaffold Diversity of Kinase Inhibitors Using Alternative Scaffold Concepts and Estimating the Scaffold Hopping Potential for Different Kinases. *Molecules* **2017**, *22*, e730.

[128] Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. Data Completeness-The Achilles Heel of Drug-Target Networks. *Nat. Biotechnol.* **2008**, *26*, 983–984.

[129] Dimova, D.; Hu, Y.; Bajorath, J. Matched Molecular Pair Analysis of Small Molecule Microarray Data Identifies Promiscuity Cliffs and Reveals Molecular Origins of Extreme Compound Promiscuity. *J. Med. Chem.* **2012**, *55*, 10220–10228.

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

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

[132] Dimova, D.; Bajorath, J. Rationalizing Promiscuity Cliffs. *ChemMedChem* **2017**, *13*, 490–494.

[133] Blaschke, T.; Miljković, F.; Bajorath, J. Prediction of Different Classes of Promiscuous and Nonpromiscuous Compounds Using Machine Learning and Nearest Neighbor Analysis. *ACS Omega* **2019**, *4*, 6883–6890.

[134] Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

[135] Keiser, M. J. et al. Predicting New Molecular Targets for Known Drugs. *Nature* **2009**, *462*, 175–181.

[136] Kunimoto, R.; Bajorath, J. Design of a Tripartite Network for the Prediction of Drug Targets. *J. Comput. Aided Mol. Des.* **2018**, *32*, 321–330.

[137] Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* **2013**, *8*, e65894.

[138] 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*, e10102.

[139] Gilberg, E.; Gütschow, M.; Bajorath, J. Promiscuous Ligands from Experimentally Determined Structures, Binding Conformations, and Protein Family-Dependent Interaction Hotspots. *ACS Omega* **2019**, *4*, 1729–1737.

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

[141] Horvath, D.; Marcou, G.; Varnek, A. Do Not Hesitate to Use Tversky-and Other Hints for Successful Active Analogue Searches with Feature Count Descriptors. *J. Chem. Inf. Model.* **2013**, *53*, 1543–1562.

[142] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

[143] Stumpfe, D.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Advancing the Activity Cliff Concept, Part II. *F1000Research* **2014**, *3*, e75.

[144] Kenny, P. W.; Sadowski, J. *Methods and Principles in Medicinal Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA, 2005; pp 271–285.

[145] 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.

[146] 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.

[147] Hu, Y.; de la Vega de León, A.; Zhang, B.; Bajorath, J. Matched Molecular Pair-Based Data Sets for Computer-Aided Medicinal Chemistry. *F1000Research* **2014**, *3*, e36.

[148] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP - Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

[149] de la Vega de León, A.; Bajorath, J. Matched Molecular Pairs Derived by Retrosynthetic Fragmentation. *Med. Chem. Commun.* **2014**, *5*, 64–67.

[150] Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932–2942.

[151] Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Mol. Inf.* **2011**, *30*, 646–664.

[152] Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.

[153] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

[154] Xu, Y.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.

[155] Katritzky, A. R.; Kiely, J. S.; Hébert, N.; Chassaing, C. Definition of Templates within Combinatorial Libraries. *J. Comb. Chem.* **2000**, *2*, 2–5.

[156] Dimova, D.; Stumpfe, D.; Hu, Y.; Bajorath, J. Analog Series-Based Scaffolds: Computational Design and Exploration of a New Type of Molecular Scaffolds for Medicinal Chemistry. *Futur. Sci. OA* **2016**, *2*, FSO149.

[157] Naveja, J. J.; Vogt, M.; Stumpfe, D.; Medina-Franco, J. L.; Bajorath, J. Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound–Core Relationship Method. *ACS Omega* **2019**, *4*, 1027–1032.

[158] Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *J. Med. Chem.* **2015**, *59*, 4189–4201.

[159] Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481.

[160] Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

[161] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

[162] Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer New York, 2000.

[163] Nielsen, M. A. *Neural Networks and Deep Learning*; Determination Press, 2015.

[164] Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discov. Today* **2015**, *20*, 318–331.

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

[166] Rodríguez-Pérez, R.; Bajorath, J. Prediction of Compound Profiling Matrices, Part II: Relative Performance of Multitask Deep Learning and Random Forest Classification on the Basis of Varying Amounts of Training Data. *ACS Omega* **2018**, *3*, 12033–12040.

[167] Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273–297.

[168] Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167.

[169] Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

[170] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of The Fifth Annual Workshop on Computational Learning Theory* **1992**, 144–152.

[171] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093–1110.

[172] Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767–779.

[173] Heikamp, K.; Bajorath, J. Prediction of Compounds with Closely Related Activity Profiles Using Weighted Support Vector Machine Linear Combinations. *J. Chem. Inf. Model.* **2013**, *53*, 791–801.

[174] Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-Directed Similarity Searching Using Support Vector Machines. *Chem. Biol. Drug Des.* **2010**, *77*, 30–38.

[175] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1241–1250.

[176] Goodfellow, I.; Bengio, Y.; A, C. *Deep Learning*; The MIT Press, 2017.

[177] Lee, H.; Grosse, R.; Ranganath, R.; Ng, A. Y. Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks. *Commun. ACM* **2011**, *54*, 95–103.

Additional Publications

Miljković, F.; Kunimoto, R.; Bajorath, J. Identifying Relationships between Unrelated Pharmaceutical Target Proteins on the Basis of Shared Active Compounds. *Futur. Sci. OA* **2017**, *3*, FSO212.

Blaschke, T.; Miljković, F.; Bajorath, J. Prediction of Different Classes of Promiscuous and Nonpromiscuous Compounds Using Machine Learning and Nearest Neighbor Analysis. *ACS Omega* **2019**, *4*, 6883-6890.