KinFragLib: Exploring the Kinase Inhibitor Space Using Subpocket-Focused Fragmentation and Recombination

Dominique Sydow,[†,¶] Paula Schmiel,[†,¶] Jérémie Mortier,[‡] and Andrea Volkamer[*,†]

[†]*In Silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany*

[‡]*Bayer AG, Digital Technologies, Computational Molecular Design, 13342 Berlin, Germany*

[¶]*Authors contributed equally to this paper.*

E-mail: andrea.volkamer@charite.de

Abstract

Protein kinases play a crucial role in many cell signaling processes, making them one of the most important families of drug targets. In this context, fragment-based drug design strategies have been successfully applied to develop novel kinase inhibitors, usually following a knowledge-driven approach to optimize a focused set of fragments to a potent kinase inhibitor. Alternatively, KinFragLib is a new method that allows to explore and extend the chemical space of kinase inhibitors using data-driven fragmentation and recombination, built on available structural kinome data from the KLIFS database for over 2,500 kinase DFG-in complexes. The computational fragmentation method splits the co-crystallized non-covalent kinase inhibitors into fragments with respect to their 3D proximity to six predefined functionally relevant subpocket centers. The resulting fragment library consists of six subpocket pools with over 7,000 fragments, available at https://github.com/volkamerlab/KinFragLib. KinFragLib offers two main applications: (i) In-depth analyses of the chemical space of known kinase inhibitors, subpocket characteristics and connections, as well as (ii) subpocket-informed recombination of fragments to generate potential novel inhibitors. The latter showed that recombining only a subset of 624 representative fragments generated a combinatorial library of 6.7 million molecules, containing, besides some known kinase inhibitors, more than 99% novel chemical matter compared to ChEMBL and 63% molecules compliant with Lipinski's rule of five.

Keywords

kinases, fragment-based drug design, kinase inhibitors, computational drug design

Introduction

Protein kinases and kinase inhibitors

Kinase function and dysregulation. Protein kinases constitute one of the largest protein families, with roughly 518 kinases encoded in the human genome.[1] Kinases share a catalytic domain for adenosine triphosphate (ATP) binding, catalyzing the transfer of its γ-phosphate group to serine, threonine, or tyrosine residues of themselves (autophosphorylation) or of other proteins. Protein phosphorylation is the major mechanism through which protein function is regulated and is fundamental to most aspects of cell life. A variety of diseases including cancer, inflammation, and autoimmune disorders, are associated with aberrant regulation of protein kinases. Dysregulation can occur due to mutations, overexpression, chromosomal rearrangements, or gene amplification. Thus over the past 20 years, protein kinases have become one of the most important classes of drug targets, especially in the field of oncology.[2–5]

Kinase groups. Protein kinases are generally divided into eukaryotic protein kinases, which share a similar sequence and structure, and atypical protein kinases, which have biochemical kinase activity, but lack sequence similarity to the typical kinase domain. Furthermore, eukaryotic protein kinases can be classified based on their sequence identity into 8 main kinase groups: AGC, CAMK, CK1, CMGC, STE, TK, TKL, and Other.[1,6]

Kinase structure. Protein kinase structures consist of two domains, i.e. the N- and C-lobes, connected via a hinge region. The majority of kinase inhibitors target the catalytic cleft between these lobes, which contains the highly conserved ATP binding site. Based on over 1,200 kinase-ligand crystal structures, van Linden et al.[7] have defined the binding site to comprise 85 residues and 19 well-defined regions/motifs, covering a (i) front cleft, (ii) gate area, and (iii) back cleft (important regions shown in Figure 1.B and their KLIFS[7,8] numbering in brackets in the following): (i) ATP solely occupies the front cleft, which contains the hinge region (46-48), linker (49-52), glycine-rich loop (4-9), and catalytic loop (68-75). ATP's adenine group as well as most kinase inhibitors form hydrogen bonds to the hinge region. (ii) The gate area contains the conserved DFG motif (81-83), the conserved lysine residue K17 (17), and the gatekeeper residue (45), often used for inhibitor selectivity, preceding the hinge region. (iii) The back cleft contains amongst others the αC-helix (20-30), including the conserved glutamine residue E24 (24), which forms the conserved K17-E24 salt bridge in the so-called αC-in (as opposed to the αC-out) conformations. Furthermore, the DFG motif can undergo a significant conformational change, which results in an inactive state of the kinase (DFG-out instead of DFG-in conformation). This DFG-flip opens a hydrophobic region in the back cleft targeted by inhibitors stabilizing the inactive state.[7,9] The KLIFS database[7,8] has made this and further information about the available kinases structures, their bound ligands, and the interactions between them freely available.

Kinase inhibitors. Kinase inhibitors are classified by their binding modes.[10] Type I and II inhibitors occupy mainly the front cleft and form hydrogen bonds with the hinge region: Type I and I½ inhibitors bind to the active and inactive DFG-in conformation, respectively, whereas type II inhibitors stabilize the inactive DFG-out conformation. Allosteric inhibitors bind only next to the ATP binding site (type III) or outside of the catalytic cleft (type IV). Type V inhibitors are bivalent, i.e. binding different regions simultaneously. While type I-V inhibitors bind reversibly, covalent inhibitors are classified as type VI.

Fragmentation and recombination of kinase inhibitors

Fragment-based drug discovery (FBDD) has been successfully applied to develop novel and selective compounds, including kinase inhibitors.[11,12] Fragments are low molecular weight compounds targeting a specific subpocket within the active site of a protein. They usually bind to their target with weaker activity than traditional drug-like molecules but with a good binding efficiency, i.e. a higher proportion of the atoms is interacting with the protein.[13,14]

In drug design, molecules can be viewed as combinations of multiple fragments. Linking, replacing, or recombining fragments is the essence of FBDD. Fragments can be generated computationally by decomposing larger compounds. Clearly, the choice of the fragmentation technique will have an impact on the resulting fragment library. For example, RECAP (REtrosynthetic Combinatorial Analysis Procedure)[15] and BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures)[16] aim to cut only synthetically meaningful chemical bonds. *e*MolFrag[17] builds on top of BRICS to generate a set of (larger) bricks and (smaller) connecting linkers.

Typically, FBDD starts with the screening of a fragment library to identify binders to specific targets, and only these hits are optimized into larger compounds by fragment linking or fragment growing. The screening step can be done experimentally or *in silico*.[18] In the context of kinase inhibition, Urich et al.[19] extracted ∼6,000 fragments with hinge-binding motifs from a kinase-unfocused library of 2.3 million compounds and docked them against 46 kinase structures to identify potential hinge binders. Fragment expansion of promising hits yielded a number of potent kinase inhibitors. Rachman et al.[20] reported a potent hinge binding fragment, selected from a kinase-unfocused fragment library (624 fragments) via docking against the JAK2 ATP binding site, filtered by (i) pharmacophoric restraints (restrained docking) at the hinge region and (ii) interaction strength measured by the work necessary to break a defined hinge hydrogen bond (dynamic undocking). However, it is also possible to start off directly with a kinase-focused library of fragments that provide optimal interaction patterns with the ATP binding site. For instance, Mukherjee et al.[21] report the Kinase Crystal Miner to extract the smallest possible fragment in each kinase-ligand crystal structure with hydrogen bonds to the hinge region, yielding about 1,000 fragments from 2,250 ligands. Substructure searches for these fragments in large molecule databases supplied molecules with kinase binding potential. Note that all the aforementioned approaches make use of 3D structural information and focus on hinge-binding fragments to be used for fragment expansion or substructure searches in compound libraries.

An alternative approach is to decompose a compound library based on kinase-focused criteria and recombine the resulting fragments to a kinase-focused molecule library. Recently, Yang et al.[22] reported a ligand-based fragmentation and recombination strategy, which was applied on both a kinase-focused (194 kinase inhibitors from PKIDB[23]) and a kinase-unfocused library of ∼4.6 million compounds. The fragments were assigned to three different fragment pools representing three designated parts of a kinase inhibitor, i.e. the core, connecting, and modifying fragments. Without using 3D structural information, fragments were assigned to the core fragment pool if a donor-acceptor hinge recognition pattern could be found. Enumerating different combinations of core-connector-modifying fragments yielded two virtual kinase-focused recombined molecule libraries (∼500,000 and ∼40 million recombined molecules), based on the aforementioned kinase-focused and kinase-unfocused input data.

KinFragLib methodology

KinFragLib, which is introduced here, takes advantage of the large amount of structural data on kinase ligands from KLIFS for subpocket-based fragmentation and recombination (Figure 1). Organizing fragments from kinase ligands by subpocket allows not only to perform a detailed subpocket-specific analysis of their fragment space, but also to better understand the composition and spatial arrangement of reported kinase-ligand complexes. Moreover, this kinase-focused fragment library organized by subpocket allows for a specific and controlled fragment recombination, unveiling a completely unexplored territory in the chemical space of kinase inhibitors.

Data and Methods

The following sections describe the procedure for (1) collecting and preprocessing the dataset of kinase complex structures, (2) defining subpockets, (3) fragmenting each of the co-crystallized ligands in the dataset, (4) analyzing the fragment library, (5) recombining fragments, and (6) studying the combinatorial library.

(1) Data collection and preprocessing

Structures of kinase-inhibitor complexes were collected from the KLIFS database[7,8] (downloaded on 2019-11-06), which offers superimposed kinase structures from the PDB[24] with 85 residues defined as kinase binding site. In KLIFS, several entries can exist for one PDB code, since crystal structures were split into all existing alternate location models and all kinase-domain-containing chains of heteromeric protein complexes. Each KLIFS entry comes with details on the species, kinase, kinase group, PDB code of the complex and the ligand, sequence alignment of the 85 binding site residues, DFG conformation (in, out, or out-like), ligand position (within or outside the main pocket), and KLIFS quality score. The latter ranges from 0 (bad) to 10 (flawless) and describes the quality of the alignment as well as structure based on each structure's alignment to a reference as well as its number of missing residues and atoms, respectively.

The structural data is preprocessed as summarized in Table 1. (A.1) Only human kinases in the DFG-in conformation and with a ligand lying within the main pocket (type I and

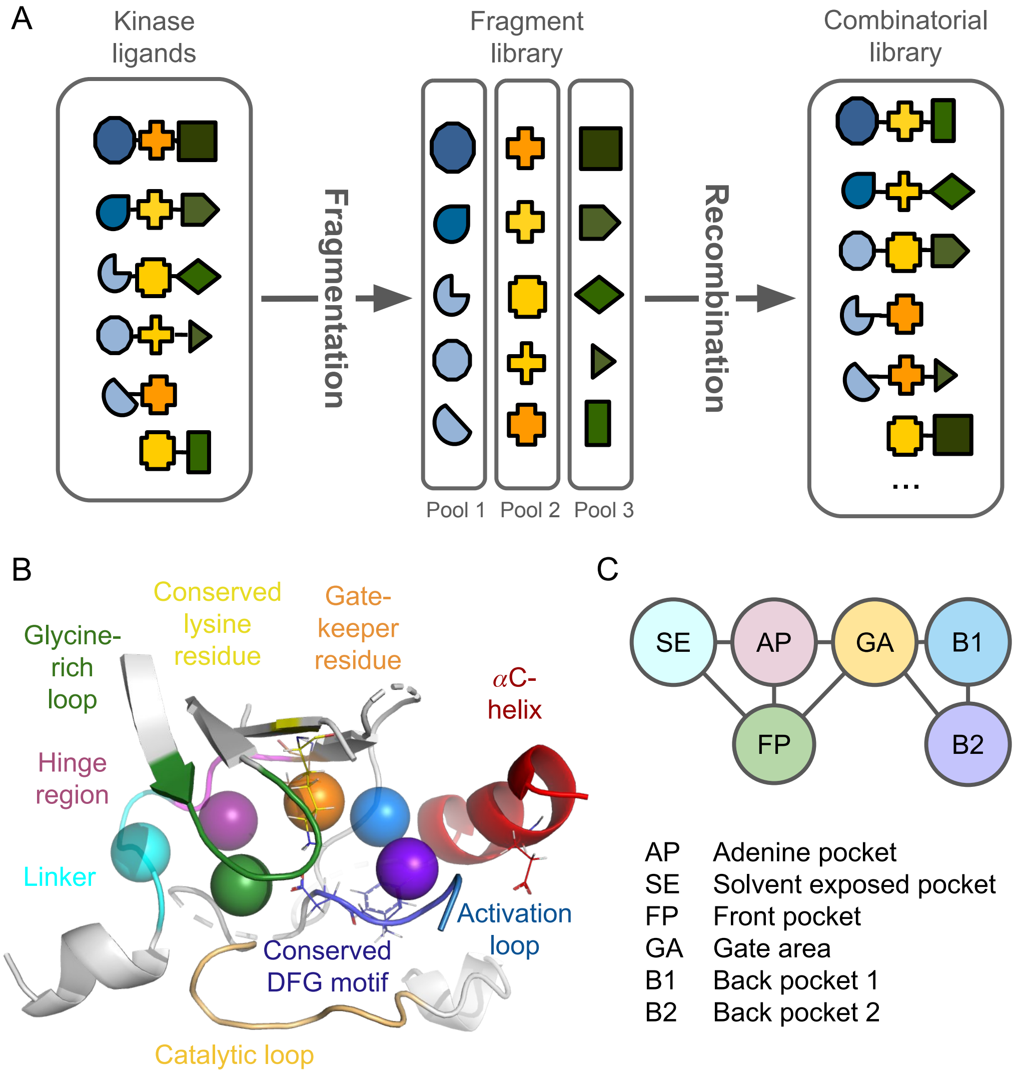


Figure 1: (A) Simplified schematic depiction of the KinFragLib approach. Based on their location in the binding site, known kinase ligands are fragmented and placed into subpocket pools, which can than be used to generate a combinatorial library. (B) The kinase binding site is shown with important regions and the six defined subpocket centers as spheres (PDB:3W2S (EGFR)). (C) Schematic depiction of the six subpockets and the predefined allowed connections between these subpockets. Colors of the subpockets are matching in B and C.

Table 1: Dataset filtering steps during preprocessing as well as additional filtering steps during the fragmentation procedure, including the number of discarded and remaining structures after each step.

Preprocessing steps	**Discarded structures**	**Remaining structures**
(A.1) KLIFS download (human, DFG-in, ligand within main pocket)		**7,370**
(A.2) Discard atypical kinases	216	7,154
(A.3) Choose best quality entry for each PDB	3,775	3,379
(A.4) Discard mol2 files not readable with RDKit	22	3,357
(A.5) Discard substrates and substrate derivatives	429	2,928
(A.6) Discard complexes with multiple ligands	17	2,911
(A.7) Discard covalent inhibitors	110	**2,801**
Additional filtering steps		
(B.1) Discard structures with important atoms missing	7	2,794
(B.2) Discard ligands with large BRICS fragments	134	2,660
(B.3) Discard ligands not occupying AP	100	2,560
(B.4) Discard ligands with unwanted subpocket connections	7	**2,553**

$\mathrm{I}^{1}/_{2}$) were selected for download from the KLIFS website,[25] yielding a starting set of 7,370 complex structures. (A.2) Atypical kinases were discarded because of the large difference in the binding site compared to eukaryotic kinases. (A.3) For each PDB code, the KLIFS entry (specified by PDB code, chain identifier, and alternative location) with the best quality score, or the first entry if there were multiple structures with an equal score, was extracted. (A.4) Mol2 files containing the binding site and the ligand of each chosen structure were loaded into RDKit.[26] Due to inconsistencies in the supported mol2 formats, some files were not readable and, thus, discarded. (A.5) Kinase structures in complex with adenine or any molecule containing a phosphate group or a ribose substructure were discarded (covering amongst others PDB ligand IDs AMP, ADP, ATP, ACP, ANP, ADN, and ADE). These are kinase substrate or substrate analogues and were therefore less in the focus for the design of novel kinase inhibitors. (A.6) Some kinase structures in the database are in complex with multiple disconnected molecules in the ATP binding site. If one of these ligands is a substrate or substrate analogue, the complete structure is discarded, since the ligand binding is not substrate-competitive. If multiple ligands consisting of more than 14 heavy atoms exist, the structure was also discarded. Otherwise, only the largest ligand was extracted. (A.7)

Finally, as the current approach focuses on the discovery of reversible inhibitors, covalent ligands were also excluded. These were identified by downloading the PDB file corresponding to the KLIFS structure and checking the CONECT records for any connection between the kinase and the ligand. Note that after personal communication with A. Kooistra,[27] two PDB entries were excluded manually (2clx, 4cfn), since the ligand was found to be not covalently bound; and three PDB entries (4d9t, 4hct, 4kio) were added, because the ligands bind covalently but the CONECT entries were missing (see full list of removed structures with covalent ligands in the SI). The dataset after preprocessing consists of 2,801 kinase-ligand structures. Further filtering steps during the fragmentation procedure as described in "(3) Molecule fragmentation" result in a final dataset of 2,553 complex structures (see Table 1).

(2) Subpocket definition and allowed connections

In this work, the kinase binding site was divided into six subpockets, which were selected based on their location and function in known kinase-inhibitor structures. Each subpocket is described by the geometric center of the Cα atoms of newly identified anchor residues from the 85 binding site residues defined by KLIFS.[7] The respective subpocket spanning anchor residues (Table 2) were selected manually after visual inspection of several structures with the aim to define a location that overlays with important parts of known kinase ligands and to provide a good distribution of centers within the pocket. As one example, the subpocket centers within the binding site of the epidermal growth factor receptor (EGFR) kinase are shown in Figure 1.B. Later, fragments will be assigned to the closest subpocket, by measuring their distance to the subpocket centers, and stored in subpocket-specific library pools (*subpocket pools*). In the following, the residue numbering refers to the numbering used in KLIFS.

Subpocket locations. The *adenine pocket* (AP), located at the geometric center of the spanning residues 15, 46, 51, and 75, lies next to the hinge region. It is usually occupied by adenine in the ATP-bound state of a kinase and allows to anchor substrate or other compounds by forming up to three hydrogen bonds. The *solvent-exposed pocket* (SE), defined here by the single residue 51, at the entrance of the binding site adjacent to AP was also called the selectivity entrance by Zhao et al.[28], as it shows diverse characteristics in different kinases and can therefore be used to achieve improved selectivity. The *front pocket* (FP), here represented by the geometric center of residues 10, 51, 72, and 81, is occupied by the ribose and phosphate groups of ATP and is partially solvent-exposed.[9] The *gate area* (GA) acts as a gate between the front cleft (containing AP, FP, and SE) and the back cleft. The GA pocket is defined by the region between the gatekeeper (residue 45), the conserved lysine (residue 17) and the aspartic acid (residue 81) in the DFG motif. The back cleft was split into two subpockets, *back pocket I and II* (B1 and B2), both lying next to the αC-helix, spanned by residues 28, 38, 43 and 81 as well as 18, 24, 79, and 83, respectively. In addition to the six subpocket pools, a seventh pool X was created to hold fragments that cannot be assigned clearly to a subpocket because the distance to their closest subpocket center exceeds 8 Å.

Table 2: Subpockets of the kinase binding site as defined in this work. Each subpocket is described by the geometric center of its anchor residues' Cα atoms (KLIFS residue numbering). For comparison, the corresponding KLIFS subpockets[7] are annotated (approximate manual assignment).

Subpocket	**Abbreviation**	**Anchor residues**	**KLIFS correspondence**
Adenine pocket	AP	15, 46, 51, 75	AP
Solvent-exposed pocket	SE	51	none
Front pocket	FP	10, 51, 72, 81	FP-I & FP-II
Gate area	GA	17, 45, 81	BP-I-A & BP-I-B
Back pocket I	B1	28, 38, 43, 81	BP-II-A, BP-II-in & BP-II-B
Back pocket II	B2	18, 24, 70, 83	

Exceptions for anchor residue definition. The definition of the 85 binding site residues in the KLIFS database is based on a multiple sequence alignment, which can have gaps. It was therefore avoided to set residues with a high gap rate among the structures as anchor residue. Furthermore, some coordinates of an amino acid or a single atom may be missing because they could not be resolved by crystallography. If the coordinates of an anchor residue's Cα atom was missing, the following procedure was applied: If possible, the coordinates were replaced with the geometric center of the two neighboring residues' Cα atoms. If one of those was absent as well, the coordinates of the other neighboring residue were used instead. If both adjacent Cα atoms were missing, the structure was discarded (see Table 1 (B.1)).

Allowed subpocket connections. In order to set up the fragment library, first, the connections between the above defined subpockets were investigated. After manual inspection of the typical structure of known kinase inhibitors (type I and I$^1/_2$ only), eight allowed subpocket connections were identified as schematically depicted in Figure 1.C. A first investigation of the generated fragments revealed that 95.2% of the molecules comply with this scheme. The remaining 4.8% ligands exhibiting unexpected subpocket connections were manually inspected: For some cases, special rules could be applied to direct fragmentation towards the defined subpocket connections, others were discarded during this analysis step (see "(3) Molecule fragmentation" and Table 1 (B.4)).

(3) Molecule fragmentation

A fragmentation algorithm was implemented to generate fragments from a given ligand in complex with a kinase structure, assign them to subpockets, and thereby populate the fragment library's subpocket pools (see Figure 1). The fragmentation algorithm is schematically depicted in Figure 2. Each kinase-ligand complex is processed successively in the following way (steps (3.1)-(3.4)):

(3.1) Subpocket center calculation. The six subpocket centers are calculated for the binding site of the respective kinase structure (see "(2) Subpocket definition and allowed connections").

(3.2) Initial BRICS fragmentation. The BRICS algorithm[16] was chosen for fragmentation. Generally, BRICS employs 16 rules to cleave bond types by taking the chemical environment and neighboring substructures into account. Thereby, BRICS ensures that

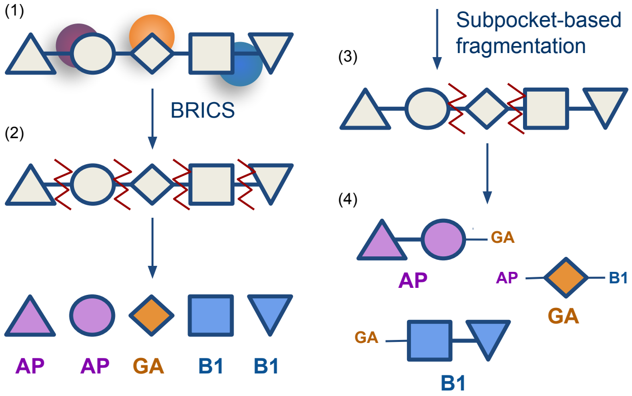


Figure 2: The implemented fragmentation algorithm splits a given kinase ligand based on the subpockets that it occupies. (1) The subpocket centers within the kinase binding site are calculated. (2) The BRICS algorithm is used for an initial fragmentation. The resulting BRICS fragments are then assigned to the closest subpocket. (3) Finally, the molecule is fragmented only at those bonds that separate fragments assigned to two subpockets.(4) At the fragmented bonds, the information on the originally adjacent subpocket is stored.

structural features of organic compounds stay intact, increasing the chance of synthetic accessibility of the recombined fragments.

To determine the potential cleaving positions, the co-crystallized ligand of the structure in hand is submitted to an initial fragmentation step, applying the RDKit implementation of the BRICS algorithm. Next, each of the resulting fragments needs to be assigned to a subpocket. Therefore, the geometric center of all atoms (including hydrogens) in the fragment, and its distance to all subpocket centers, is calculated. Then, the fragment is assigned to the subpocket with the closest subpocket center. However, if the closest subpocket to a fragment is more than 8 Å away, this fragment is considered as lying outside of the binding site and assigned to the outlier pool X. Note that the information on the BRICS environment type of each fragment is kept for later recombination.

Subsequently, the cleavage assignments are revised in order to avoid too small fragments in the final fragment library. For each fragment with less than three atoms the neighboring fragments are checked. If all neighboring fragments are assigned to the same subpocket, nothing needs to be done, because by default they will be merged in the next step. If the subpockets of the neighboring fragments differ, the current small fragment is assigned to the subpocket of the largest neighboring fragment. This procedure is repeated until all fragments with less than three atoms are reassigned.

Finally, for each bond between two BRICS fragments, the subpockets of the two fragments are compared. If the two subpockets differ, this bond is stored as a cleaving position for the final fragmentation.

(3.3) Final subpocket-based fragmentation. The original ligand is now fragmented only at bonds crossing two subpockets, while storing for each fragment the subpocket that it occupies. The subpockets of neighboring fragments are compared in order to detect unwanted subpocket connections (see section "(2) Subpocket definition and allowed connections"). (i) If a connection between subpockets FP and B1 or FP and B2 is detected, the distance of the FP fragment is calculated to the GA subpocket center. If this distance is smaller than 5 Å, this fragment is reassigned to GA instead (applied to only 15 cases). Else, the fragment in B1 or B2, respectively, is assigned to pool X. (ii) If any unwanted subpocket connection is still present after this procedure, the complete ligand is excluded from the fragment library (see Table 1 (B.4) and "Results and Discussion: Subpocket connections" for more detail).

(3.4) Fragment information storage. Each atom in the fragment is labeled with the name of its subpocket. The original attachment point(s) of each fragment is/are conserved and stored as *dummy atom(s)* and the subpocket of the former adjacent fragment is stored as a property. This allows to retrace the subpocket that the adjacent fragment was targeting in the original ligand (needed for the later recombination). Fragments are stored in structure-data files (SDF), one file for each subpocket pool as well as pool X. In addition to the structural information (3D coordinates, elements, and bonds), the following data are stored for each fragment: (i) PDB code of the original kinase-ligand complex and name of the ligand itself, (ii) chain and alternate model of this complex in KLIFS, (iii) kinase, kinase family, and kinase group, (iv) subpocket of each atom, including dummy atoms, and (v) BRICS environment type for each atom.

Summary of removed ligands during fragmentation. During the fragmentation procedure, some complexes were discarded due to the following reasons (Table 1): (B.1) A few kinase structures were missing required atom positions, thus, their subpocket centers could not be calculated. (B.2) Some ligands, such as staurosporine are not suitable for fragmentation, as they contain large, unfragmentable portions. Thus, structures with BRICS fragment(s) with more than 22 heavy atoms were discarded. (B.3) Ligands not occupying AP were excluded from the fragment library, as this work focuses on ligands targeting the ATP binding site and most kinase inhibitors developed so far bind in the AP subpocket.[7] (B.4) Ligands displaying unwanted subpocket connections were discarded. Consequently, 2,553 ligands remained and their fragments were included in the fragment library (available at `https://github.com/volkamerlab/KinFragLib`).

(4) Fragment analysis

The following paragraphs describe the different analyses that were performed on the fragment level.

Deduplicated fragments. Several fragments were contained more than once in a subpocket, therefore, a unified set was created for further analysis. First, fragments were simplified by replacing dummy atoms with hydrogens and removing all non-explicit hydrogens (*simplified fragments*). Second, fragments within one subpocket pool were deduplicated based on their canonical SMILES representation, i.e. in case of identical fragments only one was kept (*deduplicated fragments*).

Fragment similarity was calculated to allow to analyze the fragment diversity within subpockets as well as within and across kinase groups.

For the subpocket-based analysis, fragments were deduplicated per subpocket and similarities between all pairwise fragment combinations per subpocket were calculated. To this end, the topological RDKit molecular fingerprint[29] was generated for each fragment and the Tanimoto similarity metric was applied. Self-comparisons of fragments were omitted.

To analyze similarities within and across kinase groups, fragments were categorized by subpocket and kinase group (according to the structure they were bound to) and deduplicated per category. For each subpocket (excluding pool X), similarities between all pairwise fragment combinations within and across all kinase groups were calculated as described in the previous paragraph.

Common fragment motifs per subpocket. In order to identify the most common fragments in each subpocket (excluding pool X), the number of occurrences of each fragment was calculated before deduplication based on the *simplified fragments*. The 50 most common fragments in each subpocket were then clustered based on the Butina algorithm[30] using topological RDKit molecular fingerprints[29] and a distance threshold of 0.6. Note that subpockets B1 and B2 contain less than 50 deduplicated fragments and thus all fragments were chosen for clustering.

Furthermore, representative fragments were extracted manually for each subpocket in order to provide a visual overview on chemical differences and overlaps between subpockets. Each selected fragment represents a variety of common fragments with similar scaffolds and R-groups.

(5) Fragment recombination

Novel molecules can be created by recombining fragments from the fragment library. For a proof-of-concept study, only a subset of the fragment library was used. The individual steps for data reduction and fragment recombination are explained in this section.

Data reduction. The full fragment library contains 7,486 fragments. In order to reduce the combinatorial library size and run time, a diverse subset of fragments was chosen. (i) All fragments that are not suitable for recombination were removed, i.e. duplicates, fragments in pool X, fragments without dummy atoms (unfragmented ligands), and fragments with dummy atoms only connecting to pool X. Furthermore, only fragments complying with the rule of three,[31] a filter for fragment-likeness, and hinge-like AP fragments were kept. The latter filter checks for at least one hydrogen bond donor or acceptor in the AP fragment, together with at least one aliphatic or aromatic ring. The filtering steps in (i) result in 2,029 fragments. (ii) Per subpocket, a diverse set of fragments was selected for recombination to avoid enumerating highly similar fragments. The Butina algorithm[30] was applied to cluster each subpocket's filtered fragments using topological RDKit molecular fingerprints[29] and a distance threshold of 0.6. Per cluster, the most common fragments were selected. The larger the cluster the more fragments were chosen (one fragment per 10 cluster members, whereby clusters with less than 10 fragments are represented with one fragment). The final reduced fragment library consists of 624 fragments (AP: 145, FP: 192, SE: 140, GA: 93, B1: 24, and B2: 30).

Recombination procedure. All possible fragment combinations of the above described reduced set were enumerated, while preserving the original subpocket connections when connecting the fragmented bonds using the subpocket-labeled dummy atoms. Recombination started from AP fragments only, while fragments from other subpockets were consecutively added, thereby excluding any recombined molecules not occupying AP. Fragments were combined by adding a bond between two atoms adjacent to dummy atoms, while removing the dummy atoms. Thereby, two fragments were connected via a new bond between two atoms if the following conditions were fulfilled: (i) The first fragment's dummy atom was associated with the same subpocket as the second fragment and vice versa. (ii) The BRICS environment types of the atoms to be connected were matching according to the BRICS rules,[16] in order to preserve synthetic accessibility. The bond type (single or double bond) between dummy atoms was preserved when connecting the fragments. (iii) While connecting the fragments, it was ensured that the resulting molecule did not contain two fragments from the same subpocket, i.e. to occupy one subpocket multiple times.

Recombination was deemed complete if either the molecule had no dummy atoms left to another subpocket (excluding pool X), the molecule's remaining dummy atoms could not be replaced by any matching fragment, or the molecule consisted of 4 fragments. This upper limit of occupied subpockets was introduced, since the majority of kinase ligands occupies only up to 4 subpockets (see Figure 3.A) and molecules occupying more subpockets will mostly not fulfill the requirements of a drug-like molecule due to their size (e.g. Lipinski's rule of five[32]). Finally, if the resulting recombined molecule contained any remaining dummy atoms, they were replaced with hydrogen atoms. This recombination strategy produced over 6.7 million ligands based on 624 fragments.

(6) Recombined molecule analysis

The recombined molecules were compared against two sets of ligands: (i) the 542 ligands, from which the reduced set of 624 fragments originated (*reduced original ligands*), were searched for exact and substructure matches and (ii) the ChEMBL database[33] was screened for exact matches and the most similar molecules. From the latter, all 1,870,461 molecules from the ChEMBL 25[34] dataset were downloaded. If an entry contained a mixture, the largest molecule was extracted. Duplicates were dropped (based on canonical SMILES) and only molecules with more than 4 heavy atoms were kept. Molecules were standardized[35] and neutralized[36] using RDKit's rdMolStandardize module,[37] resulting in 1,782,229 standardized molecules stored as InChI[38] strings to be used for the exact matches search between the combinatorial library and ChEMBL. Futhermore, for each recombined ligand, a Tanimoto comparison based on topological RDKit fingerprints[29] yielded the most similar ChEMBL molecule.

Used software and libraries

The project was implemented in Python 3.6.8. RDKit[26] (2020.03.3) was used to perform most molecule-related calculations, matplotlib[39] (3.2.2) and seaborn[40] (0.10.1) to generate plots, and PyMol[41] (1.9.0.0) to visualize structures and subpocket centers.

Results and Discussion

The main objective of this work has been to decompose kinase ligands with respect to 3D information and to assign each resulting fragment to the kinase subpocket it binds to. Only kinase-ligand complex structures with molecules targeting the ATP binding site in the DFG-in conformation were selected, such as type I and I$^1/_2$ inhibitors, to reduce the conformational space of the kinase structures. After filtering the 7,370 starting structures assembled from the KLIFS database, 2,553 protein kinase-ligand structures were chosen for this study.

In a first step, inspired by the functional subpocket annotation in KLIFS, six functionally relevant subpockets were defined covering the ATP binding site. Note that KLIFS specifies eight subpockets, some of which describe relatively small subpockets that were combined into one subpocket in KinFragLib. Too small subpockets are algorithmically less desired in this case, because either too small fragments would be generated or larger fragments would span over several of these small subpockets. Additionally, a solvent-exposed pocket (SE) was introduced in KinFragLib, a region of the binding site occupied by many kinase inhibitors (see subpocket definitions in Table 2).

In a second step, the co-crystallized kinase ligands were fragmented with respect to the subpockets that they occupy, resulting in a kinase-focused *fragment library* with six subpocket pools (plus the pool X) and 7,486 fragments. An in-depth analysis of the six subpocket pools enabled novel insights into subpocket structural trends, chemical diversity, and typical kinase-binding motifs.

In the last step, a subset of this kinase-focused fragment library was used to create a *combinatorial library* by enumerating all feasible fragment combinations. The potential of the combinatorial library is demonstrated in comparison to the KLIFS ligands from which the subset of fragment originate and to the ChEMBL database.

The generated fragment and combinatorial libraries alongside Jupyter[42] notebooks illustrating library usage as well as the analyses for both libraries as discussed in the following are available on GitHub: `https://github.com/volkamerlab/KinFragLib`.

Subpockets and fragment library

The generated kinase-focused fragment library allows to analyze kinase-ligand interactions and explore the chemical space of kinase ligands. In total, 7,486 fragments (7,201 fragments without pool X) originating from 2,553 co-crystallized ligands were generated by the fragmentation procedure. After subpocket-based deduplication, 2,977 fragments remain (without pool X). In the following, this fragment library is analyzed with respect to the following aspects: Ligand occupancy and connectivity across subpockets, fragment occurrence, properties, and similarity per subpocket, fragment promiscuity, as well as common fragments and motifs per subpocket. This analysis aims to provide a better understanding of kinase-inhibitor binding and may serve as a valuable starting point for the design of novel kinase inhibitors.

Ligand occupancy across subpockets

The compiled fragment library enables an in-depth analysis of the number of subpockets occupied by the original ligands (Figure 3.A).

Ligands occupying 2-4 subpockets. The majority of ligands occupies two (28%) or three (53%) of the six subpockets. In another 13% of the cases, the ligand spans over four subpockets (examples of such can be seen in Figure 4.A, B and E) . This demonstrates that kinase ligands usually do not fully exploit the available space in the kinase binding site, but target only specific subpockets.

Ligands occupying 5-6 subpockets. Very few ligands (1%) occupy five subpockets, and only one visits all six subpockets. For instance, in the ALK kinase structure PDB:4FNZ,[43] the co-crystallized ligand (NZF, CHEMBL2023556[44]) covers all six subpockets (see Figure S1.A), and was indeed measured to be active on ALK (pIC_{50}=7.2) as well as IGF1R (pIC_{50}=6.9). An example of a ligand covering five subpockets is the co-crystallized active compound (W2R, CHEMBL2322330[45]) in the EGFR structure PDB:3W2S[46] (pIC_{50} = 8.2), as shown in Figure 4.D.

Ligands occupying 1 subpocket. Additionally, 127 ligands (5%) target only one subpocket and were left unfragmented during the fragmentation procedure. Since this study focuses on ligands covering the AP subpocket, all these unfragmented ligands are located in AP. They have an average number of 15 heavy atoms, which is higher than the average over all AP fragments (11 heavy atoms). As shown in Figure S1.B-D, these molecules represent either (i) small fragment-like molecules or (ii) large rigid molecules that contain a large fraction of rings, which are difficult to split for fragmentation most algorithms. An example for the former group (i) is the series of halogenated pyrazoles that stem from a fragment-based approach for druggability assessment and hit generation,[47] see Figure S1.B1-B8. The latter group (ii) contains complete drug-like molecules that either could not be divided because none of the BRICS rules applied or they had a potential BRICS cleavage bond in the initial fragmentation step, which was not broken because the two potential fragments were located in the same subpocket. Furthermore, there are rigid molecules that only contain fused rings with small decorations and, thus, do not apply to any fragmentation approach (such as quinalizarin, a CK2 inhibitor, and derivatives, see Figure S1.C1-C2). An example of a molecule that could not be fragmented by BRICS is the co-crystallized ligand HK4 (CHEMBL248396,[48] pIC_{50} = 8.3) bound to the CHK1 structure (PDB:4FST,[49] see Figure S1.D1). The two ring moieties clearly cover distinct subpockets (AP and GA), but could not be assigned to them since no rule exists that allows to split next to a triple bond between two carbon atoms.

Note that the unfragmented ligands cannot be used in the recombination algorithm (because no attachment point resulting from the fragmentation could be assigned). This could be seen as a restriction in available chemical space of the current approach, since each fragment-like molecule can be seen as a potential starting point for fragment growing. Nevertheless, roughly 28% of the unfragmented ligands were found to be substructures of other original ligands. More than half of these unfragmented ligands are fragment-like (i.e. fulfill the rule of three[31]). Thus, they are implicitly used in the introduced recombination approach. The remaining 72% unfragmented ligands are however not considered, a limitation which could be addressed by manually adding attachment points on relevant positions.

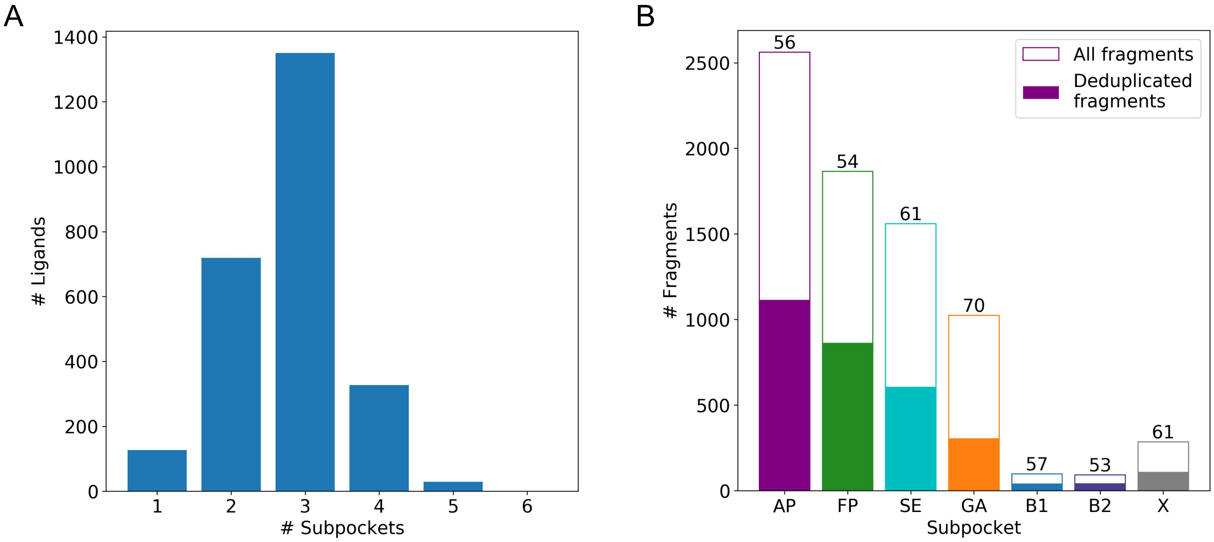


Figure 3: (A) Distribution of the number of subpockets (excluding pool X) occupied by the ligands. (B) Number of fragments and deduplicated fragments (fragments remaining after removal of duplicates) in each subpocket pool, percentage of deduplicated fragments on top of each subpocket's bar.

Ligand connectivity across subpockets

The fragmentation of existing kinase inhibitors yields an overview of how the fragments are arranged within the binding site and throughout the individual subpockets. This allows to analyze via which subpockets the fragments are most frequently connected.

Disallowed subpocket connections/special cases. As described in "Data and Methods", a few design choices were made to only allow the subpocket connections as depicted in Figure 1.C, defined based on prior investigation of known kinase inhibitors. 95.2% of the analyzed molecules follow this scheme, whereas (i) another 4.5% of the molecules could be rescued by the defined rules and (ii) the remaining 0.3% were discarded in this analysis as discussed in the following.

(i) In 113 cases, FP-B2 connections were detected initially. Manual inspection revealed two different methodological drawbacks that could be resolved by the introduced rules: First, in some cases a fragment was assigned to FP because its centroid was slightly closer to FP than GA, although visual inspection showed that the fragment acts as a gate from the front to the back cleft, and should therefore belong to GA (14 cases). The molecules containing these fragments could thus be included by reassigning them to GA (see "(3) Molecule fragmentation"). Second, the FP-B2 connection was observed when the FP fragment was relatively large. While part of it pointed mostly into the solvent, the part was still close enough to B2 and, thus, was assigned to this subpocket. Furthermore, very rare cases were manually observed where the fragment actually covered B2. Since the latter two cases could not be distinguished algorithmically, and the FP-B2 connection is rather unexpected, these B2 fragments were reassigned to pool X (99 cases). The same applies for FP-B1 connections, where each of the two cases described above occurred once.

(ii) Connections between non-adjacent subpockets (e.g. SE-GA, AP-B1) usually occur when one of the two subpockets contains a large BRICS fragment (that cannot be further fragmented), which also spans the respective subpocket in between. This happened only rarely, i.e. for AP-B1 and AP-B2 connections in 4 and 3 cases, respectively. Note that potential SE-GA connections were not counted as these ligands do not contain an AP fragment and were excluded from the study beforehand.

Subpocket connections and fragment arrangements. The fragment connectivity of the co-crystallized ligands was analyzed to identify the typical layout of kinase inhibitors. Examples of ligands representing different subpocket connections including their frequency are illustrated in Figure 4. The central connections starting from AP are observed most often. The AP-FP connection is present most frequently in 61.5% of the analyzed ligands, closely followed by the AP-SE and the AP-GA connections with 58.8% and 36.0%, respectively (see Figure 4.A). This agrees with the finding that subpocket pools AP, FP, SE, and GA contain the most fragments in descending order (Figure 3.B). FP-GA and FP-SE connections also occur in more than 7% of the ligands each (see Figure 4.B and C). Generally, the back pockets B1 and B2 are covered less often in the fragment set and they can only be reached through GA. Thus, the GA-B1 or GA-B2 connections appear only in 3.7% and 3.3% of the cases, respectively, while a GA-B1 connection happens slightly more often (see Figure 4.D). B1-B2 connections are present in only 10 ligands (0,4%, see Figure 4.E).

These findings seem to be in good agreement with the inhibitor binding modes reported in KLIFS (Table 5 in the original publication,[7] see also Table 2). The majority of ligands are in both approaches described to be front cleft binders, occupying mostly AP-GA and AP-FP subpockets (since SE is not defined in KLIFS, AP-SE and FP-SE connections are part of the KLIFS equivalent of the AP and FP subpockets). In contrast, back cleft binders that describe ligands that occupy the back pockets (AP-GA-B1/2 combinations) occur by far less often. While the KLIFS binding mode annotation is based on kinase-ligand interaction fingerprints, the analysis reported here shows that KinFragLib's automated subpocket-based procedure generates reasonable fragments.

Fragment occurrence per subpocket

The number of fragments per subpocket is reported in Figure 3.B and Table S1. Containing 35.6% of the 7,201 fragments (excluding pool X), AP is the most frequently occupied subpocket. Remember that by design this study focuses on ligands covering the AP subpocket, and all ligands not occupying AP were discarded beforehand. Also note that AP contains 8 fragments more than the actual number of fragmented ligands, i.e. 2,553. This is possible because the fragmentation algorithm allows that two not neighboring fragments of a ligand occupy the same subpocket, since not all ligands perfectly fit to the defined subpockets (this happens only rarely). The second most occupied subpocket is FP (25.9% of fragments), followed by SE (21.7%) and GA (14.2%). The back pockets B1 and B2 are occupied by only 2.6% of the fragments in total. According to this, known type I and I½ kinase ligands mostly target the same subpockets as the kinase substrate ATP (AP and

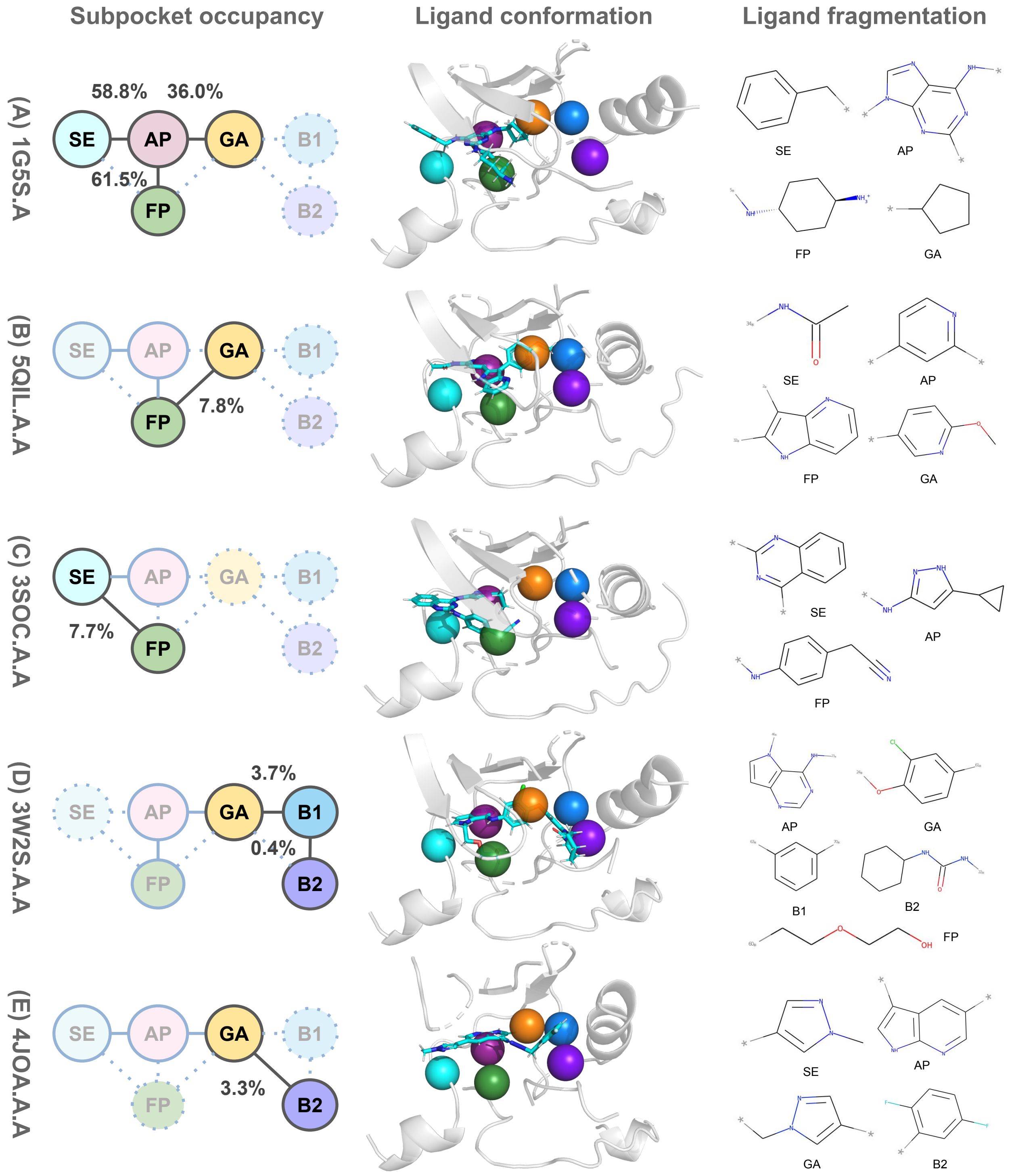


Figure 4: Subpocket connectivity for example ligands in KLIFS ([PDB code].[chain].[alternate model]): (Left) Subpockets and allowed connections with solid/dotted lines if present/not present in example ligand, including frequency of ligands with highlighted connection. (Middle) ligand conformation in kinase structure, including subpocket centers (spheres). (Right) Ligand fragmentation with assigned subpockets and dummy atoms (grey).

FP) to gain potency, followed by their neighboring subpockets, such as GA targeted to gain selectivity. In this dataset, the remote back pocket is targeted less frequently due to two reasons: First, the 69% of the underlying kinase structures show the αC-in conformation, limiting the available space for ligands in B1 and B2. Second, 73% of the front cleft binder, whereas only 25% of the back cleft binders target the αC-in conformation. Pool X contains 285 additional fragments, i.e. these fragments were classified as lying outside of the main binding site or showing not allowed subpocket connections.

Fragment properties per subpocket

In the following, the fragment pools were analyzed with respect to duplicate fragments and physicochemical properties across subpockets.

Duplicates. On average, 59% of the fragments in each subpocket were present in more than one structure (referred to as duplicates). This can be explained by the traditional medicinal chemistry approach to study a wide range of decorating groups around a shared molecular scaffold and thereby explore structure-activity relationships. Such approaches can result in the crystallization of multiple analogs from the same series. However, this finding also highlights the limited chemical diversity of the known kinase inhibitor space (considering molecules with available crystal structures only). The highest relative number of duplicates was identified in GA (70%), for the other subpockets the values do not differ largely from the average (Figure 3.B). The higher share of duplicates in GA could be explained by the generally smaller fragment size in this subpocket (compared to AP, FP, and SE, see Figure 5.D).

Physicochemical properties. In order to identify particularities in the chemical space of the different subpocket pools, standard chemical descriptors were calculated. These include (i) hydrogen bond donors and acceptors (HBD and HBA), (ii) logP values, and (iii) molecule size as in the number of heavy atoms. The distributions of these descriptors for each subpocket pool are displayed as boxplots in Figure 5.A, while excluding duplicates.

(i) AP fragments generally have a higher number of HBD and HBA, as this part of the inhibitor usually forms hydrogen bonds to the hinge region and acts as anchor to position the ligand.[9] (ii) The logP values vary widely in all subpocket pools. X, FP, and SE fragments have the lowest median logP, i.e. they tend to be more hydrophilic. For SE, this can be explained by the solvent-exposure of this part of the kinase binding site. The same holds for FP, which is also partially solvent-exposed.[9] While the AP fragments usually do provide the hydrogen bonds as anchor, they are often surrounded by a hydrophobic pocket, which could explain the comparatively high logP of these fragments. (iii) AP, FP, and SE fragments tend to be larger in terms of the number of heavy atoms, with AP having the highest median value. Note that most of the outliers in AP refer to unfragmented ligands as shown in Figure S1.C and S1.D, while outliers in FP mostly refer to large fragments that extent widely into the solvent.

This analysis reflects the general knowledge medicinal chemists have about kinase inhibitors: An HBD-HBA recognition motif is required for binding to the hinge region, the SE subpocket is used to attach functional groups that increase compound solubility, and the GA region accommodates small and hydrophobic moieties. This demonstrates the KinFragLib method's ability to automatically capture the chemical properties of kinase inhibitors.

Fragment similarity per subpocket

In the following, the fragment similarity was analyzed within each subpocket to assess if certain subpockets are occupied by more similar ligands than others. Overall, the intra-subpocket fragment similarity does not differ largely between the subpockets and is generally rather low (Figure 5.B, Table S1). The highest average intra-subpocket similarity was observed in AP with a mean of 0.14, the lowest in B1 (0.07), B2 (0.09), and FP (0.09). A higher similarity in AP can be explained by the lower flexibility of this kinase region and the targeted design of chemical moieties interacting specifically with the hinge region. The low average similarity within FP might be observed due to the larger space around the FP

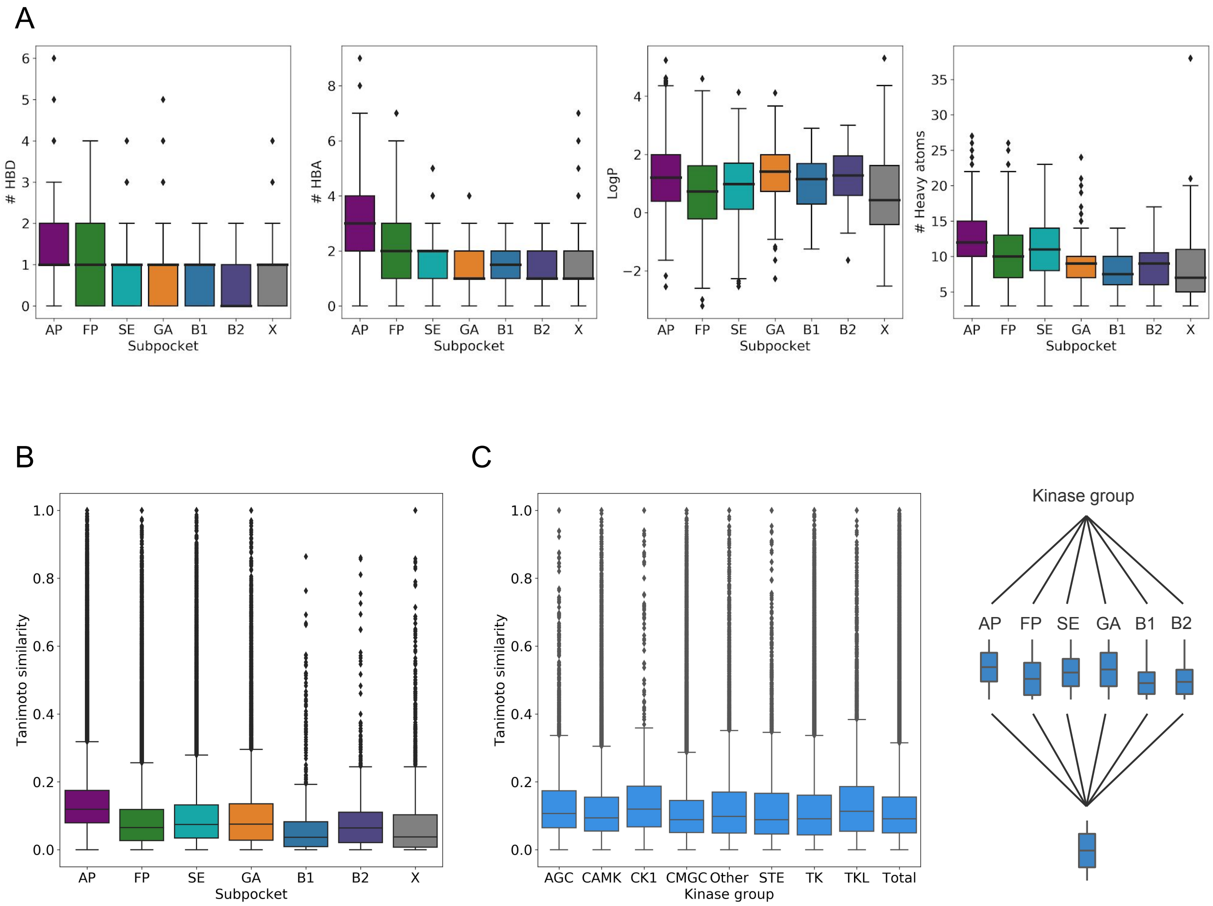


Figure 5: (A) Chemical descriptor statistics for each subpocket pool. Calculated descriptors are the number of hydrogen bond donors and acceptors (HBD and HBA), logP, and the number of heavy atoms, while excluding duplicate fragments. (B) Distribution of Tanimoto similarities between all pairwise fragment combinations per subpocket, while excluding duplicate fragments per subpocket. (C) Distribution of Tanimoto similarities between pairwise fragment combinations in each kinase group and across all kinase groups (Total), while excluding duplicate fragments within each kinase group and subpocket as well as comparing only fragments within the same subpocket.

center compared to the other subpockets, allowing a higher diversity in FP fragments. The low similarity in B1 and B2 is probably the result of the small amount of data available for these subpockets.

In general, this analysis indicates that after removing duplicates in each subpocket pool, a high diversity of chemical structure is present in the fragment pools, which underlines the potential of the KinFragLib to generate novel chemical matter.

Fragment promiscuity

Fragment promiscuity was addressed from two angles: (i) Are fragments more similar within kinase groups than across kinase groups? (ii) If fragments are observed multiple times in the same subpocket pool, are the respective ligands co-crystallized with different kinases (or kinases from the same group)?

(i) All fragments were grouped by subpockets (excluding pool X) and kinase groups. Within each of these subsets, fragments were deduplicated and similarities for all pairwise fragment combinations were calculated and pooled by kinase groups. This results in fragment similarities per kinase group, while in each kinase group only fragments were compared that occupy the same subpocket. If fragments were indeed selective for specific kinase groups, a higher fragment similarity would be observed within kinase groups compared to across all kinase groups (i.e. pooling all similarities from all subpockets). Nevertheless, no significant difference can be observed (Figure 5.C). This result indicates that the collected fragments are potentially useful for the design of an inhibitor of any target kinase.

(ii) All fragments were grouped by and deduplicated within subpockets (excluding pool X), while the number of duplicates was kept per deduplicated fragment: 67% represent singletons (appear only once per subpocket) and 12% originate from different molecules that were bound to the exact same kinase and subpocket. One interpretation of this result can be that 79% of the collected fragments have the potential to be part of a molecule that specifically inhibits one kinase. This is in line with the arguments by Xing et al.[50] and Hu and Bajorath[51] after exploring kinase hinge binding scaffolds. Another interpretation can be that 4 out of 5 fragments have never been explored on kinase targets from a different family. Using this information to create kinase-focused chemical matter could therefore be extremely useful. The remaining 21% of the fragments were bound to more than one kinase. More than three quarter of this fragment set even co-crystallized with kinases from more than one kinase group. This result supports the conclusion that fragments can be promiscuous, i.e. identical fragments can interact with multiple different kinase targets. Instead, the combination of different fragments could be the key for kinase selectivity.

Common fragments and motifs per subpocket

In order to illustrate the chemical nature of the fragments within each subpocket pool and demonstrate differences and similarities across them, representative fragments are shown in Figure 6.

The AP subpocket binds mainly heteroaromatic systems based on single or fused 5- or 6-membered rings, mostly showing the prominent donor-acceptor patterns for hinge binding. The SE subpocket is predominantly occupied by single aromatic rings, while the FP subpocket shows both single aromatic and non-aromatic rings with different substitutions. Both subpockets show residual groups rich in nitrogen, oxygen, and halogen. The GA subpocket binds mostly benzene rings with oxygen- and halogen-rich residual groups. Both GA and FP also accommodate smaller linear fragments, which are mostly terminal fragments, since a large fraction of molecules are front pocket binders and thus, do not extend further into the back pockets. For B1 and B2, much less data is available (about 90 vs. 1,000-2,500 molecules per pocket), thus, the fragments are less representative for the chemical matter that could be accommodated by these pockets: The B1 subpocket pool contains many sulfonyl groups and is rich in halogen substitutions (e.g. trifluoromethyl groups), whereas the B2 subpocket shows a quite diverse set of ligands. An overview of the 50 most common fragments per subpocket is shown in Figure S2-S7. The identified common fragments are in good agreement with the representative scaffolds reported for the different KLIFS subpockets by van Linden et al.[7] (Table 6 in the original publication).

In order to assess overlaps and differences in results from different approaches, hinge binding fragments from literature are compared to fragments from the hinge-equivalent subpocket in this study, i.e. the AP subpocket building the *AP subpocket pool*. Xing et al.[50] and Mukherjee et al.[21] both report their 10 most common hinge scaffolds/fragments (Figure 1 and Figure 7 in the original publications, respectively). Excluding adenine and staurosporine from the comparison which were removed from this library (see Table 1 (A.5) and (B.2)), all 8 fragments reported by Xing et al.[50] and 5 (out of 7) fragments reported by Mukherjee et al.[21] have exact matches in the AP subpocket pool reported in this work. When considering also highly similar (difference in one atom) AP fragments, all fragments from both studies are in the top 15 of the most common AP fragments in this study (Table S2). While both reported methods check for hydrogen bonding between the fragment and the hinge region in crystal structures, KinFragLib is able to retrieve hinge-contacting fragments without specifically searching for hinge contacts but by checking the position within the binding site. Furthermore, Yang et al.[22] report 15 examples of hinge-binding fragments extracted using hinge-like donor-acceptor patterns from kinase inhibitors (Figure 5 in the original publication). More than half of these fragments are similar to fragments in KinFragLib's top 21 AP fragments (only few exact matches), the remaining fragments were no substructures of KinFragLib's original ligands and thus are not part of the fragment library.

Recombined molecules

To exemplify the power of the combinatorial library, molecules were enumerated based on a reduced and diverse subset of the fragment library consisting of 624 fragments (see Subsection "(5) Fragment recombination"). The recombination algorithm generated 6,752,232 molecules, of which only 31,595 molecules were duplicates, yielding 6,720,637 distinct molecules. This means that only 0.005% of the library contains duplicates, i.e. molecules that were gen-

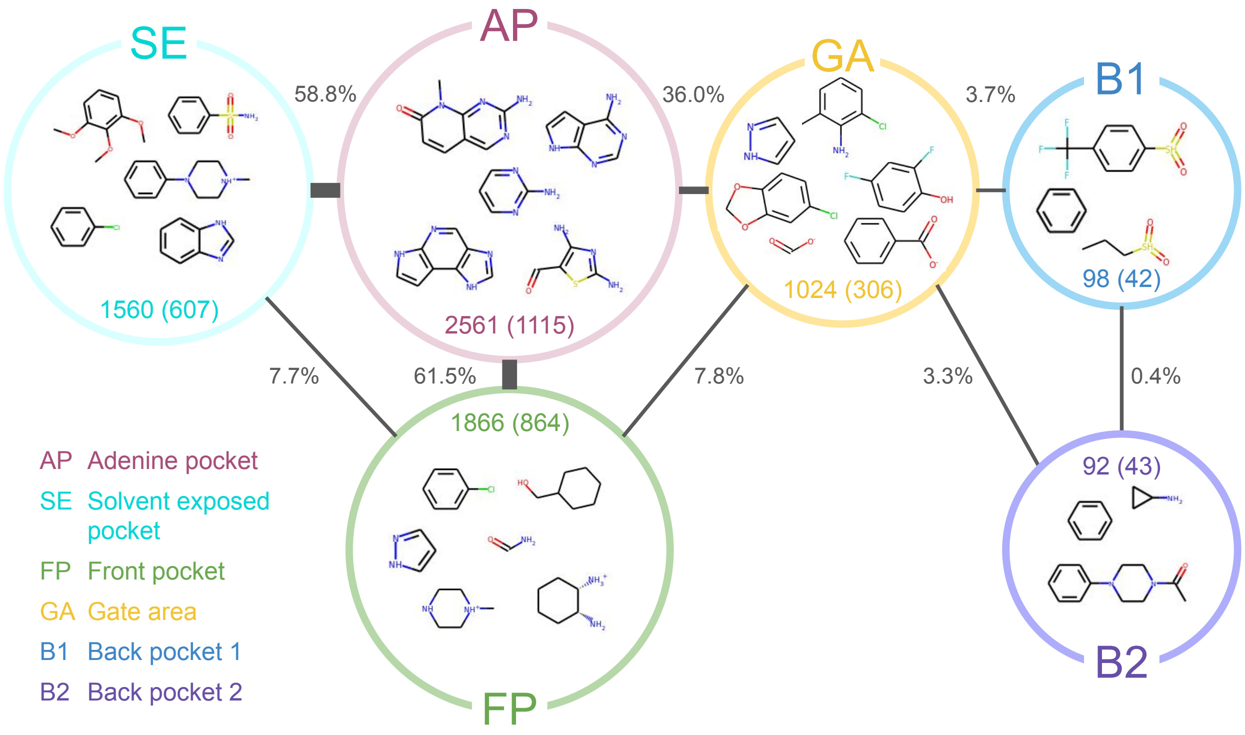


Figure 6: Representative kinase ligand composition: The representative fragments (manual selection) of the most common fragments are shown per subpocket. The subpockets' circle size illustrates the number of fragments (number of deduplicated fragments) per subpocket. Fragment connections between subpockets are shown as lines, including the percentage of ligands showing each connection. The full list of the top 50 most common fragments per subpocket is shown in Figure S2-S7. Note that dummy atoms were replaced by hydrogen atoms.

erated coincidentally from different fragment combinations.

Recombined original ligands from KLIFS

An important way to control the relevance of the generated chemical matter is to demonstrate this workflow's ability to reconstruct the ligands from which the reduced set of 624 fragments originate (reduced original ligands): 35 recombined molecules have exact matches and 324 recombined molecules are substructures. Note that only a subset of fragments (624 out of 2,977) was used for recombination, thus only a fraction of original ligands can be retrieved.

Recombined ChEMBL molecules

The search for exact matches in ChEMBL[33] (1,782,229 molecules) revealed that only 298 of the over 6.7 million recombined molecules have already been described in ChEMBL. Only 218 matching molecules remain after removing the exact and substructure matches in the "reduced original ligands" used for the fragmentation. Consulting bioactivity data available in ChEMBL, 47 out of these 218 molecules have been shown to be active against human target(s) (activity is here defined as $IC_{50} \leq 500\,\text{nM}$): 44 are active against kinases, two against cytochrome P450, and one against an voltage-gated ion channel. In total, 10 molecules show a high activity against kinases with an $IC_{50} \leq 5\,\text{nM}$ (see Figure 7). More details on the ChEMBL IDs and molecular structures are shown in Table S3 and Figure S8. This shows strong evidence that the library contains molecules with a high chance of exhibiting kinase activity.

Chemical novelty (with respect to KLIFS subset and ChEMBL)

Excluding the 359 original ligands (35 exact and 324 substructure matches in KLIFS) and the 218 exact matches in ChEMBL (without KLIFS matches), the recombination generated 6,720,058 novel molecules out of 6,720,637 deduplicated recombined molecules, i.e. 99.99 %

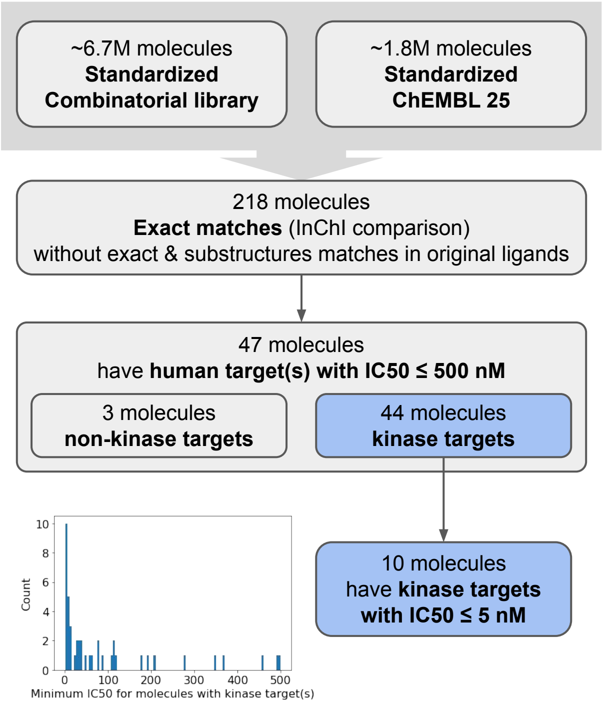


Figure 7: Exact matches (based on standardized InChI comparison) of recombined ligands in the ChEMBL 25 dataset, including targets that these molecules are active against (activity is here defined as $\mathrm{IC}_{50} \leq 500\,\mathrm{nM}$). The histogram shows the IC_{50} values for molecules that are active against kinases.

of chemical matter with no precedent in ChEMBL and in the 542 reduced original ligands in KLIFS. Furthermore, comparison of the recombined ligands with their most similar ChEMBL molecules revealed that the combinatorial library is not highly similar to the ChEMBL chemical space (mean similarity of 0.54 with a standard deviation of 0.07, see Figure S9).

At the same time, as discussed before, 35 original kinase inhibitors from KLIFS and 44 additional potent kinase inhibitors in ChEMBL could be recombined, while using only a subset of the fragment library. This indicates that the novel fragment library can generate large libraries of novel chemical matter, while being tailored for the design of kinase inhibitors.

Properties of recombined molecules

The majority of the 6.7 million recombined molecules include fragments able to occupy 4 subpockets (90%), whereas the majority of original ligands is smaller and occupies three (53%) or two (28%) subpockets only. This is a consequence of a choice made in order to illustrate the power of exhaustive *in silico* library enumeration (the linking of fragments reaching up to four subpockets was allowed in this case). But most importantly, the presented workflow allows for tailored library design that can easily be adapted to fulfill the requirements of a particular project.

While 86% of all kinase inhibitors in clinical trails (dataset from 2020-07-15 downloaded from PKIDB[23]) fulfill Lipinski's rule of five, still 63% of the combinatorial library (4.2 million molecules) comply with Lipinsik's rule of five (Figure 8), representing a large kinase-focused library to be used for virtual screening studies.

Note that only a subset of fragments was used to generate the recombined library, thus, even larger libraries could be generated by taking into account all fragments identified in this study.

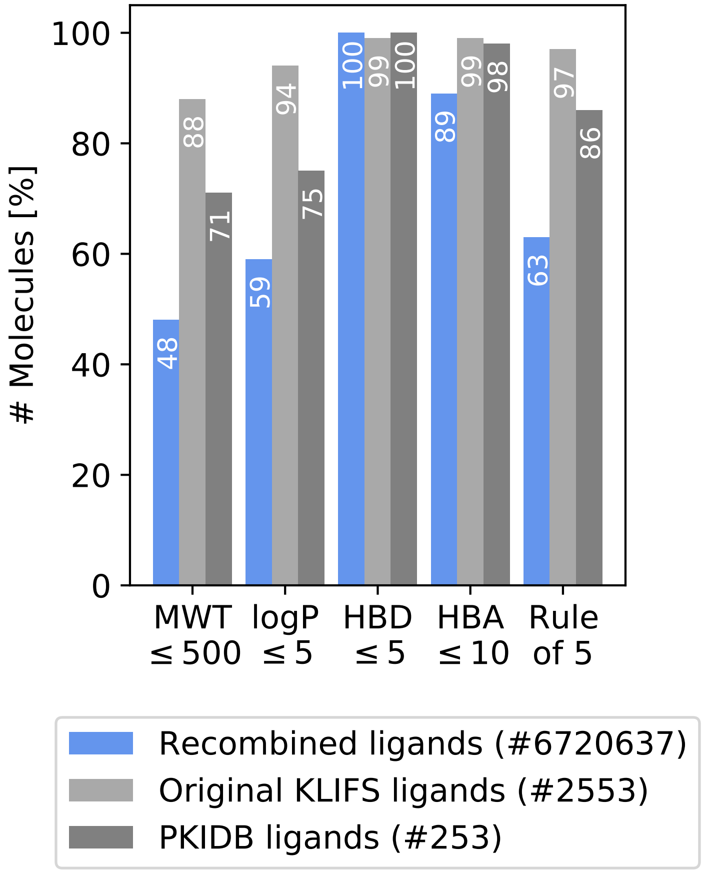


Figure 8: Lipinski's rule of five criteria applied to the 6.7 million molecules in the recombined library (blue) in comparison to (i) the 253 kinase inhibitors in clinical trials from PKIDB (light grey) and (ii) the 2,553 original KLIFS ligands used for building the kinase-focused fragment library (dark grey). The recombined molecules are overall molecules with a larger molecular weight (MWT) and more hydrogen bond acceptors (HBA), whereas the partition coefficient (logP) and the number of hydrogen bond donors (HBD) stay relatively the same. In total, 63% of the recombined library (4.2 million molecules) fulfill Lipinski's rule of five.

Conclusion

Kinases are one of the most studied protein families in medicinal chemistry, resulting in an amount of available data too large to be handled by a human brain. By combining a precise cartography of the ATP binding site and a tailored fragmentation method, KinFragLib allows to read, fragment, and organize by subpocket inhibitors co-crystallized with a kinase in the DFG-in conformation. The subsequent analysis of the chemical matter of the compiled fragments is in agreement with the general knowledge of medicinal chemists, identifying small and lipophilic fragments in the gatekeeper area, solubilizing fragments in the front pocket, and typical hinge binders for the adenine pocket. While this analysis is also in line with previous works conducted for the hinge binding fragments, this study provides for the first time a fragment library that is organized by subpocket, unveiling subpocket occupation and connection frequencies. It was found that chemically diverse fragments can bind the same subpocket. Furthermore, 79% of the identified fragments were only observed in one kinase structure, but the other 21% could bind the same subpocket of different kinase groups. This result indicates that a fragment binding one kinase subpocket is likely to bind the same region of other kinases. Therefore, the high chemical diversity of the generated fragment library is a rich source of inspiration for building novel kinase inhibitors. To investigate this possibility, a library of recombined fragments was enumerated *in silico* (using a diverse subset of the fragments only). The resulting virtual library containing over 6.7 million molecules was compared to the ChEMBL database (exact matches), indicating 99.99% of novel chemical matter. The rare exceptions of compounds with precedence in the literature include predominately known kinase inhibitors. These results clearly highlight the enormous potential of this fragment library for the design of novel kinase inhibitors.

The reported method focused on two types on kinase inhibitors (type I and I$^1/_2$), however other libraries could be generated by fragmenting other kinase inhibitor types. Similarly, the same protocol could be applied to a more specific set of ligands, e.g to design a library of fragments specific of a kinase group, or a different dataset of ligand-kinase 3D structures.

And finally, this workflow is also perfectly suited to support a fragment-growing approach after one novel fragment has been validated in a kinase subpocket.

Code and Data Availability

The generated fragments and recombined ligands, the full fragment and combinatorial library analysis, and a quick start notebook on how to access the data are freely available at `https://github.com/volkamerlab/KinFragLib` (v1.0.0, `10.5281/zenodo.3956638`).

Author Contributions

Conceptualization, A.V. and J.M.; Methodology, A.V., P.S., and D.S.; Software, P.S. and D.S.; Formal Analysis, D.S. and P.S.; Investigation, P.S., D.S., J.M. and A.V.; Writing - Original Draft, Review and Editing, D.S., P.S., A.V., and J.M.; Visualization, P.S. and D.S.; Supervision, A.V., J.M., and D.S.; Funding Acquisition, A.V.

Abbreviations

ATP, Adenosine triphosphate; GK, Gatekeeper residue; FBDD, Fragment-based drug discovery; BRICS, Breaking of Retrosynthetically Interesting Chemical Substructures; KLIFS, Kinase-Ligand Interaction Fingerprints and Structures database; PDB, RCSB Protein Data Bank; AP, Adenine pocket; SE, solvent-exposed pocket; FP, front pocket; GA, gate area; B1, back pocket I; B2, back pocket II; SDF, Structure-data file; MWT, Molecular weight; HBD, Hydrogen bond donor; HBA, Hydrogen bond acceptor.

Acknowledgement

A.V. and D.S. gratefully acknowledge funding from the Deutsche Forschungsgemeinschaft (Grant VO 2353/1-1). A.V. acknowledges support from the Bundesministerium für Bildung und Forschung (Grant 031A262C). The authors thank Albert Kooistra for many fruitful discussions about the fragmentation procedure and the ZEDAT HPC service (Freie Universität Berlin) for cluster time.

Supporting Information Available

- `kinfraglib_si.pdf`; contains supporting descriptions, figures, and tables.

References

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

(2) Cohen, P. Protein Kinases – The Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discov.* **2002**, *1*, 309.

(3) Cohen, P.; Alessi, D. R. Kinase Drug Discovery – What's Next in the Field? *ACS Chem. Biol.* **2012**, *8*, 96–104.

(4) Kooistra, A. J.; Volkamer, A. In *Annu. Rep. Med. Chem.*; Goodnow Jr, R. A., Ed.; Elsevier, 2017; Vol. 50; pp 197–236.

(5) Fabbro, D.; Cowan-Jacob, S. W.; Moebitz, H. Ten Things You Should Know About Protein Kinases: IUPHAR Review 14. *Br. J. Pharmacol.* **2015**, *172*, 2675–2700.

(6) Sakkiah, S.; Ping Cao, G.; P Gupta, S.; Woo Lee, K. Overview of the Structure and Function of Protein Kinases. *Curr. Enzym. Inhib.* **2017**, *13*, 81–88.

(7) van Linden, O. P.; Kooistra, A. J.; Leurs, R.; De Esch, I. J.; De Graaf, C. KLIFS: A Knowledge-Based Structural Database to Navigate Kinase-Ligand Interaction Space. *J. Med. Chem.* **2013**, *57*, 249–277.

(8) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P.; Leurs, R.; de Esch, I. J.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Res.* **2015**, *44*, D365–D371.

(9) Liao, J. J.-L. Molecular Recognition of Protein Kinase Binding Pockets for Design of Potent and Selective Kinase Inhibitors. *J. Med. Chem.* **2007**, *50*, 409–424.

(10) Roskoski Jr, R. Classification of Small Molecule Protein Kinase Inhibitors Based Upon the Structures of their Drug-Enzyme Complexes. *Pharmacol. Res.* **2016**, *103*, 26–48.

(11) Mortenson, P. N.; Berdini, V.; O'Reilly, M. Fragment-Based Approaches to the Discovery of Kinase Inhibitors. *Methods Enzymol.* **2014**, *548*, 69–92.

(12) Erlanson, D. A.; De Esch, I. J.; Jahnke, W.; Johnson, C. N.; Mortenson, P. N. Fragment-to-Lead Medicinal Chemistry Publications in 2018. *J. Med. Chem.* **2020**, *63*, 4430–4444.

(13) Rabal, O.; Urbano-Cuadrado, M.; Oyarzabal, J. Computational Medicinal Chemistry in Fragment-Based Drug Discovery: What, How and When. *Future Med. Chem.* **2011**, *3*, 95–134.

(14) Mortier, J.; Rakers, C.; Frederick, R.; Wolber, G. Computational Tools for in Silico Fragment-Based Drug Design. *Curr. Top. Med. Chem.* **2012**, *12*, 1935–1943.

(15) 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.

(16) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* **2008**, *3*, 1503–1507.

(17) Liu, T.; Naderi, M.; Alvin, C.; Mukhopadhyay, S.; Brylinski, M. Break Down in Order to Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. *J. Chem. Inf. Model.* **2017**, *57*, 627–631.

(18) Lamoree, B.; Hubbard, R. E. Current Perspectives in Fragment-Based Lead Discovery (FBLD). *Essays Biochem.* **2017**, *61*, 453–464.

(19) 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.

(20) Rachman, M.; Bajusz, D.; Hetényi, A.; Scarpino, A.; Merő, B.; Egyed, A.; Buday, L.; Barril, X.; Keserű, G. M. Discovery of a Novel Kinase Hinge Binder Fragment by Dynamic Undocking. *RSC Med Chem* **2020**, *11*, 552–558.

(21) Mukherjee, P.; Bentzien, J.; Bosanac, T.; Mao, W.; Burke, M.; Muegge, I. Kinase Crystal Miner: A Powerful Approach to Repurposing 3D Hinge Binding Fragments and Its Application to Finding Novel Bruton Tyrosine Kinase Inhibitors. *J. Chem. Inf. Model.* **2017**, *57*, 2152–2160.

(22) Yang, Y.; Zhang, Y.; Hua, Y.; Chen, X.; Fan, Y.; Wang, Y.; Liang, L.; Deng, C.; Lu, T.; Chen, Y.; Liu, H. In Silico Design and Analysis of a Kinase-Focused Combinatorial Library Considering Diversity and Quality. *J. Chem. Inf. Model.* **2020**, *60*, 92–107.

(23) 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*, 908.

(24) 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.

(25) Division of Medicinal Chemistry, Vrije Universiteit Amsterdam, KLIFS - Kinase-Ligand Interaction Fingerprints and Structures database. `https://klifs.vu-compmedchem.nl/` (accessed 2019-11-06).

(26) RDKit, RDKit Version 2020.03.3. `http://www.rdkit.org` (accessed 2020-04-05), 2018.

(27) Kooistra, A. Personal communication, 2019.

(28) Zhao, Z.; Xie, L.; Xie, L.; Bourne, P. E. Delineation of Polypharmacology Across the Human Structural Kinome Using a Functional Site Interaction Fingerprint Approach. *J. Med. Chem.* **2016**, *59*, 4326–4341.

(29) RDKit, RDKit Fingerprint Version 2020.03.3. `https://www.rdkit.org/docs/source/rdkit.Chem.rdFingerprintGenerator.html` (accessed 2020-04-05).

(30) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput.Sci.* **1999**, *39*, 747–750.

(31) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, *8*, 876–877.

(32) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.

(33) Gaulton, A. et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.

(34) ChEMBL, ChEMBL25 download. doi.org/10.6019/CHEMBL.database.25 (accessed 2019-07-04).

(35) RDKit, rdkit.Chem.MolStandardize.rdMolStandardize.StandardizeSmiles Version 2020.03.3. https://www.rdkit.org/docs/source/rdkit.Chem.MolStandardize.rdMolStandardize.html (accessed 2020-04-05).

(36) RDKit, rdkit.Chem.MolStandardize.rdMolStandardize.Uncharger.uncharge Version 2020.03.3. https://www.rdkit.org/docs/source/rdkit.Chem.MolStandardize.rdMolStandardize.html (accessed 2020-04-05).

(37) RDKit, rdkit.Chem.MolStandardize.rdMolStandardize Version 2020.03.3. https://www.rdkit.org/docs/source/rdkit.Chem.MolStandardize.rdMolStandardize.html (accessed 2020-04-05).

(38) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics* **2015**, *7*, 23.

(39) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95.

(40) Seaborn, Seaborn v0.9.0. https://seaborn.pydata.org/ (accessed 2020-04-05), 2018.

(41) DeLano, W. L., et al. PyMol: An Open-Source Molecular Graphics Tool (Version 1.9). *CCP4 Newsletter on protein crystallography* **2002**, *40*, 82–92.

(42) Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; Willing, C.; Jupyter Development Team, In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*; Loizides, F., Schmidt, B., Eds.; IOS Press: Amsterdam, The Netherlands, 2016; pp 87–90.

(43) Epstein, L. F.; Chen, H.; Emkey, R.; Whittington, D. A. The R1275Q Neuroblastoma Mutant and Certain ATP-competitive Inhibitors Stabilize Alternative Activation Loop Conformations of Anaplastic Lymphoma Kinase. *J. Biol. Chem.* **2012**, *287*, 37447–37457.

(44) ChEMBL, Compound ID CHEMBL2023556. `https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL2023556/` (accessed 2020-03-20).

(45) ChEMBL, Compound ID CHEMBL2322330. `https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL2322330/` (accessed 2020-03-20).

(46) Sogabe, S.; Kawakita, Y.; Igaki, S.; Iwata, H.; Miki, H.; Cary, D. R.; Takagi, T.; Takagi, S.; Ohta, Y.; Ishikawa, T. Structure-Based Approach for the Discovery of Pyrrolo[3,2-d]pyrimidine-Based EGFR T790M/L858R Mutant Inhibitors. *ACS Med. Chem. Lett.* **2013**, *4*, 201–205.

(47) Wood, D. J.; Lopez-Fernandez, J. D.; Knight, L. E.; Al-Khawaldeh, I.; Gai, C.; Lin, S.; Martin, M. P.; Miller, D. C.; Cano, C.; Endicott, J. A.; Hardcastle, I. R.; Noble, M. E. M.; Waring, M. J. FragLites—Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation. *Journal of Medicinal Chemistry* **2019**, *62*, 3741–3752, PMID: 30860382.

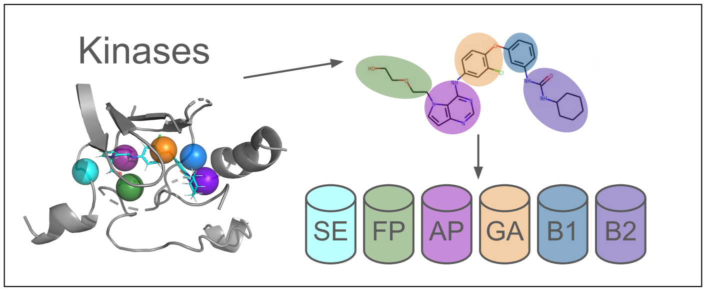
(48) ChEMBL, Compound ID CHEMBL248396. `https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL248396/` (accessed 2020-03-20).

(49) PDB, Entry ID 4FST. `https://www.rcsb.org/structure/4fst` (accessed 2020-04-05).

(50) Xing, L.; Klug-Mcleod, J.; Rai, B.; Lunney, E. A. Kinase Hinge Binding Scaffolds and Their Hydrogen Bond Patterns. *Bioorg. Med. Chem.* **2015**, *23*, 6520–6527.

(51) Hu, Y.; Bajorath, J. Exploring the Scaffold Universe of Kinase Inhibitors. *J. Med. Chem.* **2014**, *58*, 315–332.

Graphical TOC Entry


Kinases
SE
FP
AP
GA
B1
B2