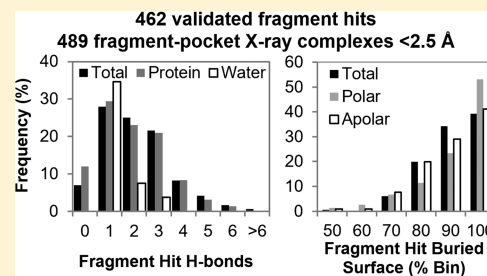


## Fragment Hits: What do They Look Like and How do They Bind?

Fabrizio Giordanetto,<sup>\*,†</sup> Chentian Jin,<sup>†</sup> Lindsay Willmore,<sup>†</sup> Miklos Feher,<sup>†</sup> and David E. Shaw<sup>\*,†,‡</sup><sup>†</sup>D. E. Shaw Research, New York, New York 10036, United States<sup>‡</sup>Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032, United States

## Supporting Information

**ABSTRACT:** A “fragment hit”, a molecule of low molecular weight that has been validated to bind to a target protein, can be an effective chemical starting point for a drug discovery project. Our ability to find and progress fragment hits could potentially be improved by enhancing our understanding of their binding properties, which to date has largely been based on tacit knowledge and reports from individual projects. In the work reported here, we systematically analyzed the molecular and binding properties of fragment hits using 489 published protein–fragment complexes. We identified a number of notable features that these hits tend to have in common, including preferences in buried surface area upon binding, hydrogen bonding and other directional interactions with the protein targets, structural topology, functional-group occurrence, and degree of carbon saturation. In the future, taking account of these preferences in designing and selecting fragments to screen against protein targets may increase the chances of success in fragment screening campaigns.



## INTRODUCTION

A major challenge in drug discovery research is the identification of a suitable small molecule that binds the target and can serve as a starting point for chemistry exploration and optimization. This critical early hit-identification step significantly influences the overall likelihood of success of a drug discovery project. Very small molecules (molecular weight <300 Da),<sup>1</sup> referred to as “fragments”, can be ideal hits in some respects, as they tend to have favorable physical properties and form high-quality interactions with the target protein. Screening a library of fragments by evaluating their binding to a target protein has proven to be an effective method for identifying hits. From its origins as an infrequently used NMR-based method,<sup>2</sup> fragment screening has evolved into an approach that has been widely adopted by industry and academia<sup>3–10</sup> and has played a central role in the discovery of two approved drugs to date<sup>11–13</sup> and in the identification of a number of clinical candidates.<sup>14</sup>

Fragment screening has a number of potential advantages over screening larger compounds as an approach to the identification of starting points for drug development. Although, in contrast to larger compounds, a fragment typically has fewer interactions with the target protein and thus lower affinity overall, thermodynamic<sup>15</sup> and probabilistic<sup>16</sup> models of small molecule–macromolecule binding suggest that these fragment–protein interactions are individually of greater energetic reward. Additionally, screening can more effectively sample the chemical space of smaller, less complex compounds,<sup>17</sup> which improves the odds of identifying binders that have high ligand efficiency.<sup>18</sup> Fragments may thus provide medicinal chemists with a greater number of promising opportunities and, owing to their smaller size, greater flexibility in the optimization process.

These advantages notwithstanding, fragment-based approaches have a number of limitations. It can be challenging to detect weak-affinity fragment hits<sup>19</sup> and effectively distinguish them from false positives.<sup>3</sup> Furthermore, structural information describing the atomic interactions between the fragment hit and its protein target is typically necessary for successful fragment optimization, but such information can be difficult to obtain.<sup>20</sup> These caveats hinder fragment-hit identification, evaluation, and prioritization, thus significantly limiting the reliability and success of fragment-based screening protocols.

Fragment-based approaches could potentially be improved with a deeper understanding of the molecular properties of fragment hits and how such fragments bind to their target proteins. To date, our understanding of fragment–protein binding has largely been derived from individual case studies rather than from a broad structural analysis of validated fragment hits. Here, we have collected and reviewed publicly available fragment-hit data, of which a substantial amount was deposited since 2012, from a variety of fragment-based screening campaigns.<sup>21–23</sup> We first describe the protein systems used in these fragment screens, followed by a cheminformatics analysis of the fragment hits to assess their properties (such as lipophilicity and structural topology). We then analyze how the fragments interact with their targets using various measures (such as the amount of buried polar surface area (SA) and the number of hydrogen bonds (H-bonds) between the fragments and the proteins). By studying a large set of fragment hits in this manner, we elucidate some of their salient properties, quantifying and adding to what is

Received: November 27, 2018

Published: March 15, 2019

typically tacit knowledge among fragment-based screening practitioners and medicinal chemists. Our findings could potentially be used to improve the chances of success for fragment-based screening methods, in particular by informing fragment design and the evaluation of fragment hits.

## METHODS

Structures of the complexes were extracted from the Protein Data Bank (PDB)<sup>24</sup> using the keyword “fragment” in a text-based query performed on Jan 13, 2019 and then filtered by keeping only the structures with a crystallographic resolution  $\leq 2.5$  Å and containing a ligand with a total number of nonhydrogen atoms  $\leq 20$ . The resulting list of 5115 complexes was further refined by removing all entries that did not relate to fragment screening, identification, and characterization by analysis of the primary literature citation associated with each structure as well as the examination of the components of the structure. PDB entries where the chemical component consisted only of typical crystallization buffer elements (e.g., glycerol), adjuvants (e.g., carboxylates), or native cofactors/cosubstrates (e.g., pyridoxal phosphate), for example, were not considered. Visual inspection of the resulting 1623 complexes supported by the corresponding literature sources served to distinguish bona fide fragment hits from optimized fragments or lead compounds and to identify the associated binding pockets. Since crystallization buffers and conditions vary greatly across the considered complexes, any chemical component from the crystallization buffer that did not have biological relevance to the system (e.g., dimethyl sulfoxide) was removed from the structure before further analysis.

Structural complexes where multiple fragments bound to the protein within close structural proximity ( $< 4$  Å for their shortest interatomic distance) were also removed. Structures in which the fragment was tethered to the protein through covalent linkages were discarded. Fragments bound to different binding sites on the same protein were treated independently in cases where the evidence supports the relevance of the alternate binding pockets, based on the primary literature source and electron density analysis. In cases where multiple structures were found of a single fragment bound to the same protein site, preference was given to the structure with the highest atomic resolution and clearest electron density, unless a significant change in the fragment-binding conformation was detected (root-mean-square deviation  $> 1$  Å for the fragment heavy atoms between binding conformations), in which case both structures were included. Fragments that displayed a real-space correlation coefficient of less than 0.8 (which is the threshold value proposed by Deller and Rupp for unambiguous fragment binding<sup>25</sup>) were not considered further. Fragment-binding sites located at the interface between the protein and its copies in the crystal lattice were not considered, to create a final data set in which every pocket involves just one copy of the protein. The 489 complexes that remained on the list after these filtering steps were then analyzed with respect to the observed protein–fragment interactions.

The ionization, tautomer, and rotamer states of the fragment and the amino acids in the binding pocket were generated using Protonate3D<sup>26</sup> and manually refined as needed based on typical geometric features for molecular interactions (e.g., H-bonds),<sup>27</sup> the local protein environment and potential interaction networks, the calculated  $pK_a$ <sup>28</sup> for the fragment, and the pH of the experimental crystallization conditions used. No heavy-atom coordinates were modified during the preparation process. Water molecules are only discussed for structures having a crystallographic resolution  $\leq 1.5$  Å,<sup>29</sup> and only if the water molecules appear to be structurally relevant and are clearly supported by the electron density maps (i.e., a real-space correlation coefficient  $\geq 0.8$ ). We define structurally relevant water molecules to be those with  $\geq 2$  H-bonds to the protein and  $\geq 1$  H-bonds to the fragment.

Molecular properties of the fragments, including the number of heavy atoms, rotatable bonds, chiral centers, formal charges, and H-bond donors and acceptors in their protonation state at pH 7, were calculated as implemented in MOE.<sup>30</sup> The *cxcalc* command line script

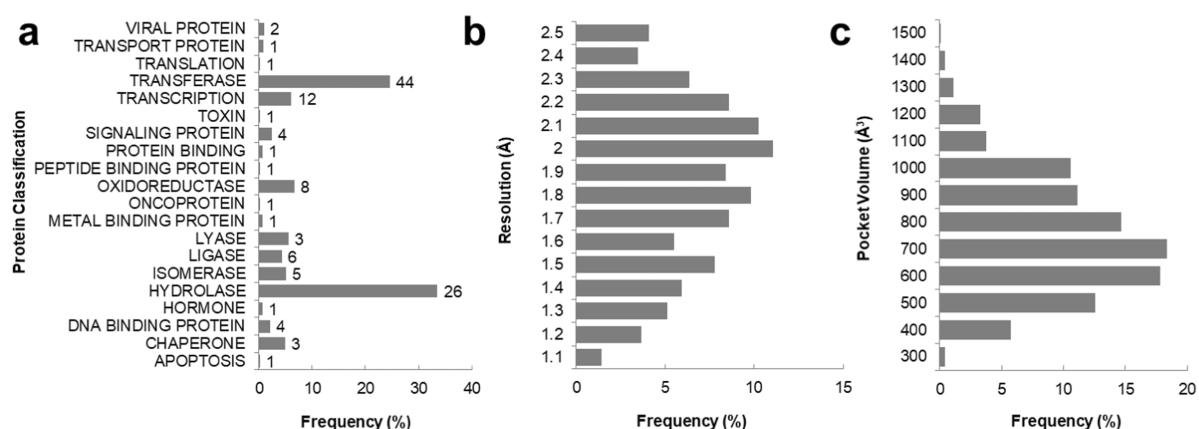
from ChemAxon<sup>28</sup> was used to compute the  $pK_a$  values for ionizable groups on the fragment hits. The dissimilarity distribution of fragments was evaluated using extended connectivity fingerprints (ECFP4)<sup>31</sup> with the *screenmd* command line script within JChem from ChemAxon.<sup>28</sup> Octanol–water partition coefficient (*clogP*) values for the fragments were calculated using the classic algorithm within the batch version of ACD/Percepta.<sup>32,33</sup> The number and identity of ring assemblies, Bemis–Murcko frameworks, and the fraction of  $sp^3$ -hybridized carbon atoms (*Fsp<sup>3</sup>*) were calculated using Vortex.<sup>34</sup> Protein pocket descriptors were computed using default parameters in *dpocket* and implemented as part of the  $\alpha$  spheres-based methodology available in *fpocket*.<sup>35</sup> Here, selection of the relevant fragment hit was used to explicitly define the associated binding pocket.

Surface-based descriptors including total, polar, and apolar solvent-accessible surface areas were calculated in MOE<sup>30</sup> using a 1.4 Å solvent radius probe. The differences in these values between the unbound and bound states of fragment hit and protein yielded the corresponding buried solvent-accessible surface areas. Molecular interaction counts between the fragment and corresponding protein, water, and metal ions were computed using a probabilistic receptor potential within MOE.<sup>30</sup> Here, H-bonds, metal coordination bonds, arene-based interactions, carbon–hydrogen bonds, halogen bonds, and sulfur-mediated contacts are scored with empirical type-based scoring functions using the extended Hückel theory. These functions are trained using statistics derived from contacts in the RCSB PDB, and each interaction is scored in terms of the percentage likelihood of being geometrically ideal. The default energy threshold of interaction ( $0.5 \text{ kcal mol}^{-1}$ ) was used to identify relevant interactions. The H-bond count only includes interactions involving oxygen or nitrogen atoms (whereas weak H-bonds, which are interactions in which the hydrogen is covalently bonded to a carbon atom or in which the acceptor is a halogen, are omitted). Water molecules are only discussed in cases where they are potentially relevant to the fragment–protein interaction, as defined above. The degree of burial of an H-bond was calculated as the difference between the solvent-accessible surface area in the unbound and bound states for the protein atom involved in the H-bond.

Atom types for the fragment hits used in the frequency of distribution analysis were assigned based on the MMFF94 force field<sup>36</sup> as implemented in MOE. The atom-type-based frequencies of molecular interactions were obtained by dividing the number of observed interactions by the total number of occurrences for a given atom type across the entire fragment set to control for over-representation of specific functional groups and thus provide useful background information for molecular design purposes.




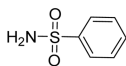
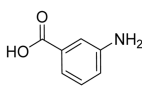


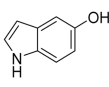

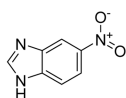
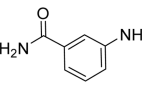

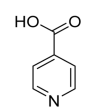
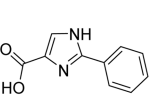
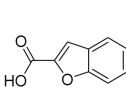
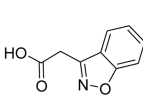
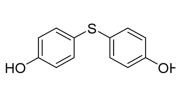
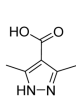
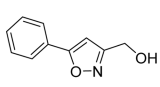
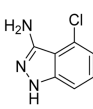
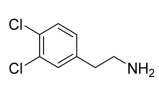
A procedure analogous to the one described above was used to generate a comparative data set in which larger ligands were bound to the protein pockets that were observed in the original (fragment hit) data set; the construction of this second data set was such that any given protein pocket has a similar frequency of occurrence in each data set. Structures of the complexes were extracted from the PDB<sup>24</sup> using the UniProtKB accession numbers<sup>38</sup> of the fragment–protein complexes previously identified in a protein-based query, and then filtered by keeping only the structures with a crystallographic resolution  $\leq 2.5$  Å and containing a ligand with a total number of nonhydrogen atoms  $\geq 25$ . This list was further refined by removing all entries that did not contain a chemical component that resulted from ligand screening and optimization, using analysis of the primary literature citation associated with each structure as well as examination of the components of the structure. Structures in which the ligand was covalently bound to the protein were discarded. Preparation of the ligand–protein complexes for calculation of ligand–protein interactions and ligand properties followed the same procedure described for the fragment–protein complexes.

To ensure that the analysis was not skewed by particular protein pockets that are over-represented in the data set, each data point was normalized by the number of occurrences of the bound protein pocket in the entire data set. A given observation for one of the 19 PDE10A–fragment complexes surveyed in this work, for example,



**Figure 1.** Characteristics of the proteins, structures, and binding pockets. (a) Distribution of protein classes, as codified by the PDB.<sup>24</sup> Numeric labels indicate the number of unique proteins belonging to each protein class, as indicated by their respective UniProt access codes.<sup>66</sup> (b) Distribution of crystallographic resolution values. (c) Distribution of fragment-binding pocket volumes using dpocket,<sup>35</sup> normalized by the occurrence of a given protein binding site.

**Table 1. Fragments Hits That Bind to More than One Protein Pocket<sup>a</sup>**

					
4B6E, 5FPO (3)	5CYQ, 5CXR (2)	5CYM, 5CW1 (2)	2WEJ, 5FZN, 5JAD	2PQF, 3FHB, 5CSV	2YED, 5DYX
					
3NUY, 4LM4	3FUH, 4B3C	2VTA, 4B2I	4MSA, 4N9C	3HKV, 3KCZ	2JJC, 5FPR
					
5FXV, 5FZ6	2XP4, 3GRJ	3IMG, 4OYP	3ZL6, 5I5S	5OSS, 5Z4H	5FPN (2)
					
3VQ4 (2)	6G8X (2)	5CLP (2)			

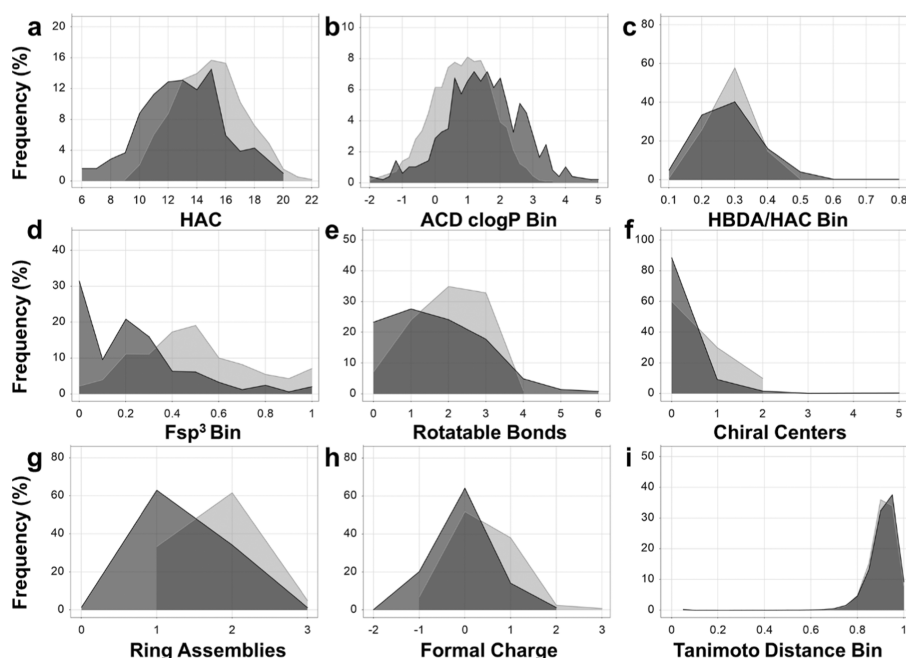
<sup>a</sup>The captions show the PDB IDs of the corresponding complexes. In cases where the fragment bound distinct pockets on the same protein, the number of pockets bound is indicated in parentheses.

carries a weight of 5.26% in the final analysis. When a data normalization based on a pocket-sequence-identity cutoff of 60%<sup>37</sup> was used, there were no significant changes in the results (Figures S1 and S2).

## RESULTS

**General Characteristics of Fragment Hits and Their Protein Targets.** Our data set consists of 489 structures of a fragment bound to a pocket of a protein or protein domain. The set contains 126 unique proteins spanning 20 different protein families and 79 structural domains (Figures 1 and S3). As shown in Figure 1, 67% of the structures of the complexes

have a crystallographic resolution of  $\leq 2$  Å, with 1.03 Å as the best-reported resolution (PDB ID: 4Y4J). The structural data is for the most part relatively recent; 79% of the structures were deposited in the PDB since the beginning of 2012 (Figure S1). Transferases and hydrolases account for 58% of the complexes, with 44 and 26 unique protein entries from each class, respectively. The remainder of the data set is distributed across 18 protein families, including DNA-binding proteins, oxidoreductases, isomerases, and viral proteins (Figure 1). Several important drug targets appear in the data set, including poly ADP-ribose polymerase, carbonic anhydrase,  $\beta$ -lactamase, estrogen receptor, DNA gyrase, Bruton's tyrosine kinase, and



**Figure 2.** Characteristics of the fragment hits ( $N = 462$ , black profiles). Distributions of the (a) number of heavy atoms, (b) ACD calculated  $\log P$ , (c) number of atoms capable of accepting or donating H-bonds expressed as a percentage of the fragment's heavy atoms, (d) fraction of  $sp^3$ -hybridized carbon atoms ( $F_{sp^3}$ ), (e) number of rotatable bonds, (f) number of chiral centers, (g) number of ring assemblies, and (h) formal charges are shown. (i) Frequency distribution of Tanimoto distances calculated based on ECFP4 for all possible fragment pairs. Distributions for a representative commercial fragment library<sup>46</sup> ( $N = 1794$ , gray profiles) are included.

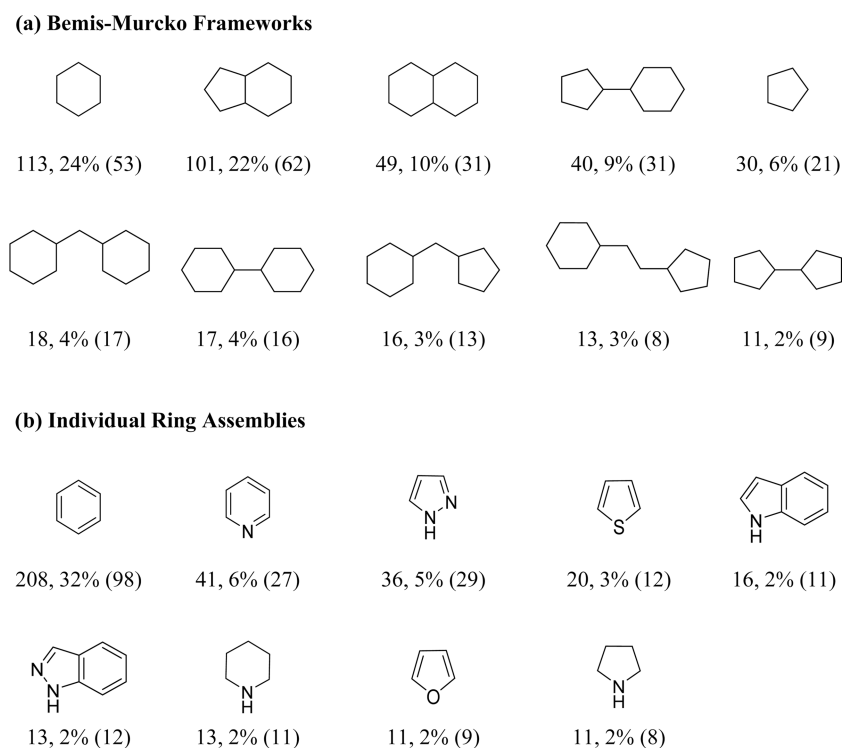
Janus kinase 2. The most frequent protein entries, representing 19% of the data surveyed here, are the aspartic protease endothiapepsin ( $N = 57$ ),<sup>39,40</sup> cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A2 (PDE10A2,  $N = 19$ ),<sup>41,42</sup> and heat-shock protein 90 (Hsp90,  $N = 18$ ).<sup>43,44</sup>

A total of 168 unique fragment-binding pockets were identified in the 126 proteins in the data set, according to the selection criteria described in the [Methods](#) section. To ensure that the analysis was not skewed by the over-representation of certain protein pockets, each binding data point was normalized by the number of occurrences of the bound protein pocket in the data set. 82% of the proteins ( $N = 103$ ) feature only a single binding site. The remaining 23 proteins bind fragments at more than one site, with HIV-1 reverse transcriptase containing 7 different sites ([Table S1](#)). The fragment-binding pockets span a 6-fold difference in size, as estimated using a Voronoi tessellation and  $\alpha$  sphere-based method<sup>35</sup> ([Figure 1](#)), ranging from a small cleft on human cyclophilin D accommodating pyrrolidine-1-carbaldehyde (PDB ID: 3R54) to a voluminous funnel-shaped pocket on hepatitis C virus polymerase NS5B bound to 4-(2-phenylhydrazinyl)-1*H*-pyrazolo[3,4-*d*]pyrimidine (PDB ID: 4IH5).

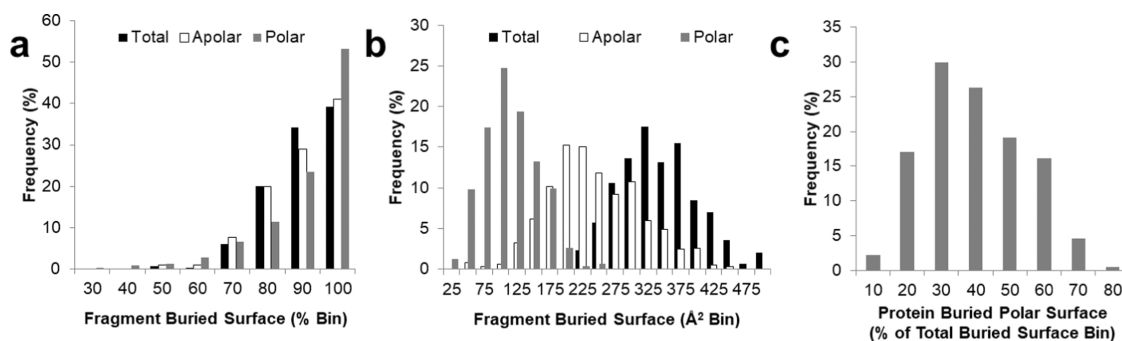
In total, 462 unique fragments are covered in the present analysis. 21 of these fragments occur more than once in the data set, either because they bind different proteins or because they bind to distinct pockets on the same protein ([Table 1](#)). None of these 21 fragments match the pan-assay interference compounds (PAINS) structural filters defined by Saubern et al.<sup>45</sup> Across the whole data set, only four fragments (pyrocathocol, PDB: 4K7I; 4-methylbenzene-1,2-diol, PDB: 4K7N; 4-(*tert*-butyl)benzene-1,2-diol, PDB: 4K7O; 4-(2-amino-1-hydroxyethyl)benzene-1,2-diol, PDB: 4Y4J) were flagged as potential PAINS hits, all with the polyphenolic structural alert. The fragments range in size from 6 to 20 heavy

atoms, with 81% of the fragment set containing between 10 and 16 heavy atoms ([Figure 2](#)). At the extremes, 2-chloro-1*H*-imidazole (with 6 heavy atoms) binds to the BAZ2B bromodomain (PDB ID: 5E9K), and *N*-[2-(morpholin-4-yl)phenyl]thiophene-3-carboxamide (with 20 heavy atoms) binds to soluble epoxide hydrolase (PDB ID: 3WKD). Calculated octanol–water partition coefficient ( $\log P$ ) values<sup>32</sup> range from  $-2.4$  to  $4.8$ , with 68% of the compounds displaying  $\log P < 2$ , as summarized in [Figure 2](#). Examples of the extremes of lipophilicity in this data set include 4-acetylpiperazin-2-one complexed with the bromodomain-containing protein 1 ( $\log P$ :  $-2.4$ , PDB ID: SAME) and 2-(5-chloro-3-methylbenzo[*b*]thiophen-2-yl)acetic acid complexed with farnesyl pyrophosphate synthase ( $\log P$ :  $4.8$ , PDB ID: 3N1V). 71% of the fragment hits have a significant proportion (20–30%) of their heavy atoms as nitrogen or oxygen atoms that are capable of accepting or donating H-bonds. More than half of the fragments (68%) display a net formal charge of 0, and for the remainder of the (charged) fragments, twice as many have a net negative formal charge (22%) as a net positive charge (11%). Only eight fragments could have zwitterionic character, with both strongly basic and acidic functionalities (predicted  $pK_a > 9$  and  $< 5$ , respectively).

In terms of structural complexity, the vast majority (>90%) of the fragment hits are achiral, with limited carbon saturation (fraction of  $sp^3$ -hybridized carbon atoms  $F_{sp^3} < 0.5$ ), and up to three rotatable bonds and two ring assemblies ([Figure 2](#)). The fragment hits are structurally diverse (extended connectivity fingerprint 4 (ECFP4)-based Tanimoto distance  $> 0.7$  for 97% of the set; [Figure 2](#)). The fragment hits presented here compare well to an established commercial fragment library;<sup>46</sup> the most notable difference is that the latter has a  $\sim 20\%$  greater proportion of fragments with a high degree of carbon saturation and a high number of ring assemblies ([Figure 2](#)).



**Figure 3.** (a) Bemis–Murcko frameworks occurring more than 10 times across the fragment hits. The number of observations and frequency of occurrence as a percentage of the total observed frameworks ( $N = 52$ , see Table S2 for the full list) are indicated in the caption. Values in parentheses indicate the number of unique proteins the framework was found bound to. (b) The most frequently occurring individual ring assemblies. The number of observations and frequency of occurrence as a percentage of the total observed individual ring assemblies ( $N = 138$ , Table S3) are indicated in the caption. Values in parentheses indicate the number of unique proteins bound by the ring assembly in this data set.



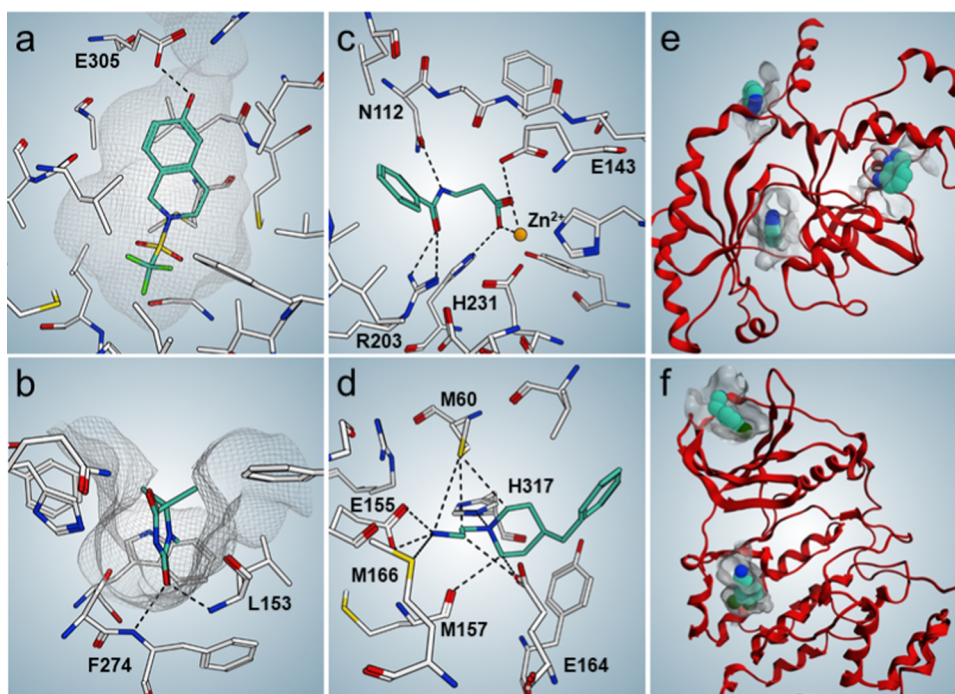
**Figure 4.** Surface characteristics of the fragment–pocket complexes. (a) Polar and apolar buried SA of fragments, expressed as a percentage of their total polar and apolar SA, respectively. The total buried SA of fragments is also indicated as a percentage of the whole-fragment SA. (b) Extents of the total, apolar, and polar SA that are buried by the fragments upon binding. (c) Buried polar SA of the protein.

The fragment hits in this work are described by 52 unique Bemis–Murcko frameworks.<sup>47</sup> Of the 52 unique molecular frameworks, 33 (63%) occur only once, whereas five frameworks account for 71% of the total. These five frameworks are: monocyclic 6- and 5-membered rings; bicyclic 5-6- and 6-6-fused ring systems; and 5-6 rings connected by one bond (Figure 3). When hybridization, heteroatoms, and exocyclic carbonyl groups are considered, the present set encompasses 138 unique ring assemblies, of which 78 (57%) appear only once. Benzene, pyridine, pyrazole, thiophene, and indole taken together account for 49% of the ring assemblies, as shown in Figure 3. Among the other ring assemblies with more than 10 representatives, piperidine and indazole each occur in several distinct binding pockets (Figure 3).

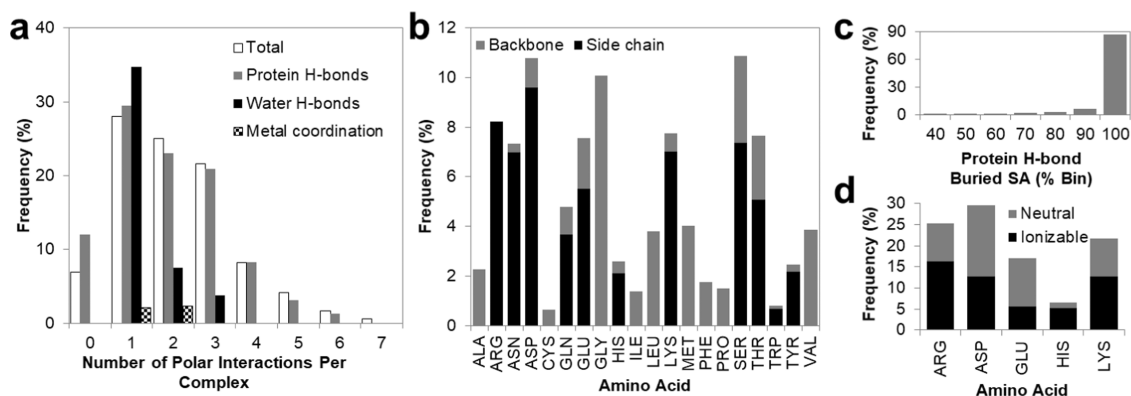
**Characteristics of Fragment Hits: Biological–Target Interactions.** We analyzed the fragment–pocket complexes to

assess the molecular interactions between the two binding partners. As molecular interactions are conceptual models that describe complex physical phenomena, general considerations regarding surface contacts and atomic contacts are summarized first, followed by more specific interaction models.

Most of the fragment hits (73%) bury more than 80% of their total solvent-accessible surface area (SA) upon binding. Even the two fragments that are the most exposed to solvent upon binding still hide 50 and 57% of their surface in the bound pose (PDB IDs: SJAN and SJ4H, respectively); these fragments are the only ones that bury <60% of their SA. In contrast, 21 fragments are completely engulfed by the protein (Figures 4 and 5a,b). Notably, 77% of fragment hits bury more than 80% of their polar SA, and 53% of hits bury over 90% of their polar SA. The apolar SA of fragments is also largely buried upon binding, although to a slightly lesser extent, with



**Figure 5.** Selected fragment hits bound to proteins. (a) Maximum and (b) minimum solvent exposure of fragments (PDB IDs: 3OMQ and 5JAH, respectively). (c) Maximum number of polar interactions (PDB ID: 3FGD). (d) Maximum number of additional directional interactions (PDB ID: SEGS). (e) and (f) Examples of fragments binding to multiple pockets within the same protein (PDB IDs: 5FPO and SCLP, respectively). Fragments are depicted as bold sticks (cyan carbon atoms). Relevant H-bonds and additional directional interactions are indicated as dashed black lines. Relevant protein pocket surfaces are displayed as gray mesh.



**Figure 6.** Polar interactions established by the fragment hits. (a) Frequency distribution of polar interaction counts per fragment complex, including H-bonds to protein or water and coordination bonds to metal ions. Fragment complexes that do not display any water H-bonds or metal coordination bonds (54 and 95% of the total, respectively) have been omitted from the histogram for clarity. (b) Frequency distribution of fragment H-bonds to protein amino acids at the side-chain and the backbone level. (c) Frequency distribution of buried surface area for protein atoms involved in H-bonds to fragments. (d) Frequency distribution of H-bonds between fragment atoms (neutral and ionizable) and protein amino acid side chains.

70 and 42% of the set burying >80 and >90% of the lipophilic surface, respectively. The polar fraction of the buried SA of a protein pocket varies significantly, ranging from 5 to 74% of the total protein buried surface. In terms of absolute values, fragments bury on average 336 Å<sup>2</sup> upon binding, with a substantial tendency to bury more apolar than polar SA (Figure 4). Accordingly, the ratio of apolar to polar SA buried has a mean of 3.1 and a median of 2.2 (Table S4).

Having analyzed the binding of fragment hits to their targets with surface-based descriptors, we next evaluated their molecular interactions. H-bonds to protein and water molecules, as well as metal coordination bonds, were identified

based on geometric criteria.<sup>27</sup> As shown in Figure 6, 92% of the fragment–protein complexes are stabilized by at least one H-bond to the protein or to a structural water or by a coordination bond to a structural metal ion. In one complex, seven such molecular interactions were noted (PDB ID: 3FGD, Figure 5c). A large majority of H-bonds between fragments and proteins (88%) are completely buried (Figure 6).

A group of 37 complexes (from 8 unique proteins) feature fragment hits bound to the structural metal ions present in the pockets, namely, zinc, manganese, and iron ions (Table S5). Negatively charged oxygen atoms from fragment carboxylic

acid groups are the atoms that most frequently establish coordination bonds to these metal ions (Table S5). A maximum of two metal coordination bonds was recorded for a given fragment-pocket entry in the current data set (PDB ID: SACW).

53 of the 116 X-ray structures with resolution  $\leq 1.5$  Å (46% of the total, with 20 unique binding sites) have structural water molecules with at least two H-bonds to the protein and one to the fragment hit. The maximum number of individual water-based H-bonds observed for a fragment was 3 (PDB ID: SMOH). Nitrogen fragment atoms have a higher occurrence of H-bonds to water than do oxygen fragment atoms (Table S6). Anilinic nitrogen atoms and  $sp^3$ -hybridized nitrogen atoms in aliphatic amine groups display the highest frequency of water H-bonds in the data set (0.33 and 0.52, respectively; Table S6).

Fragment H-bonds to proteins stabilize 89% of the bound complexes, with 74% of the entries displaying between one and three H-bonds. The highest number of fragment–protein H-bonds in a complex is 6 (PDB ID: 4Y4G) (Figure 6). Side-chain atoms account for 58% of the total protein–fragment interactions. H-bonds occurring between ionizable functional groups on both the fragment and the amino acid side chain represent 17% of the total number of H-bonds in the data set (Figure 6). As expected, the side chains of polar amino acids make the greatest contribution to H-bonds between proteins and fragments. Aspartic acid and serine each provide more than 10% of the total H-bonds observed in the fragment–protein complexes. Interestingly, glycine, which establishes H-bonds exclusively through its backbone atoms, also accounts for greater than 10% of total H-bonds observed, occurring in 27 unique protein pockets (Figure 6). Cysteine, isoleucine, phenylalanine, proline, and tryptophan each account for fewer than 2% of H-bonds. Histidine and glutamic acid have a preference for H-bonds to ionizable and neutral fragment atoms, respectively (Figure 6).

The occurrence of H-bonds to the protein based on specific fragment atom types is summarized in Table 2. After normalizing for the overall atom-type occurrence in our data set, nitrogen and oxygen atoms show a similar preference for H-bond formation, with 0.62 and 0.61 H-bonds per atom, respectively. Positively charged nitrogen atoms establish the highest number of H-bonds to protein residues, with aliphatic amines displaying a higher likelihood of forming such bonds than aromatic or conjugated ones (cf.,  $sp^2$   $NH^+$  and  $sp^3$   $NH^+$ , Table 2). The only negatively charged nitrogen atoms that form H-bonds to proteins are embedded in heterocyclic systems (Table S5). Tetrazole moieties, for example, were found to establish up to three H-bonds to serine and threonine residues of CTX-M-9 class A  $\beta$ -lactamase (PDB IDs: 3G2Y and 3G32). In contrast, the deprotonated nitrogen atoms of sulfonamide mainly coordinate metal ions (Table S5). H-bond donors in the form of neutral nitrogen atoms are represented by five different functional groups (Table 2). Among the functional groups with more than 50 observations, anilinic nitrogen atoms form the highest number of H-bonds per atom (0.80), followed by heterocyclic and amidic NHs (0.66 and 0.61, respectively). H-bond acceptors featuring a neutral, unprotonated nitrogen atom engage in H-bonds with the protein in one-third of cases. As shown in Table 2, 4 out of 10 nitrile functionalities form an H-bond in this data set (e.g., PDB ID: 4Y4T). Interestingly, one aliphatic amine may potentially accept an H-bond from the protein, based on the

**Table 2.** Fragment Atoms Involved in H-Bonds to the Protein<sup>a</sup>

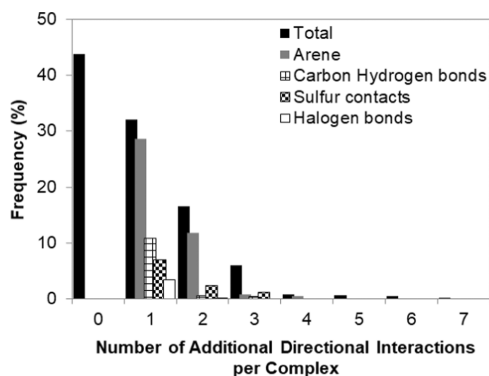
atom type	total occurrence	H-bond occurrence	ratio
<b>nitrogen</b>	833	517	0.62
<b>N (neutral)</b>	328	97	0.30
$sp^2$	317	92	0.29
$sp$	10	4	0.40
$sp^3$	1	1	1.00
<b>N<sup>-</sup></b>	29	7	0.24
heterocyclic	29	7	0.24
<b>N<sup>+</sup></b>	104	150	1.44
$sp^2$	41	50	1.22
$sp^3$	63	100	1.59
<b>NH</b>	372	263	0.71
amide	137	83	0.61
anilinic	138	110	0.80
heterocyclic	79	52	0.66
sulfonamide	11	8	0.73
hydrazide	7	10	1.43
<b>oxygen</b>	720	441	0.61
<b>O (neutral)</b>	363	171	0.47
carbonyl	214	142	0.66
sulfonamide	38	20	0.53
ether	78	5	0.06
aromatic	33	4	0.12
<b>O<sup>-</sup></b>	270	202	0.75
carboxylic acid	192	188	0.98
sulfonic acid	6	5	0.83
phenol	57	5	0.09
nitro	14	3	0.21
N-oxide	1	1	1.00
<b>OH</b>	87	68	0.78
aliphatic	30	46	1.53
aromatic	57	22	0.39

<sup>a</sup>The total occurrence of the various atomic types in the fragment set based on the MMFF94 force field definitions,<sup>36</sup> corresponding subclass based on atomic hybridization or associated functional group, and the calculated number of H-bonds and occurrence ratios are presented.

surrounding atomic environment, crystallization pH (7.5), and predicted  $pK_a$  (7.3) for the fragment species (PDB ID: SFYU).

Negatively charged oxygen atoms occur in five different chemical environments in this data set. Carboxylic acid groups, which occur the most frequently, form on average one H-bond per oxygen atom (Table 2). All other functional groups within this subclass occur much less frequently ( $N \leq 57$ ). Notably, there is one phenolic group in the data set that may capture two H-bonds (PDB ID: 3GVB), based on the crystallization pH (8.7) and calculated  $pK_a$  (8.7). Alcoholic groups on fragment hits secure on average 0.78 H-bonds, with aliphatic alcohols having a higher H-bond formation frequency than their phenolic counterparts (1.53 and 0.39, respectively). Unprotonated oxygen atoms as H-bond acceptors represent the most conspicuous oxygen atom class ( $N = 363$ , 50%), with an average of 0.47 H-bonds per atom. Carbonyl oxygen atoms from amide, urea, ester, and ketone moieties display, on average, 0.66 H-bonds each, compared to 0.53 H-bonds each for oxygen atoms in sulfonamide groups. Ether and aromatic oxygen atoms display H-bonds in only 6 and 12% of cases, respectively.

Only 7% ( $N = 34$ ) of the fragment–pocket complexes analyzed here do not contain any of the polar interactions described above. The distribution of the fragment buried surface values of this subset of fragments was not significantly skewed compared to the remainder of the set. In half of these complexes, the fragment establishes at least one H-bond to a water molecule, which in turn makes an H-bond to the protein, thus acting as a bridge between the fragment hit and the protein. Interestingly, when additional directional interactions including arene- and sulfur-mediated contacts,<sup>27,48</sup> halogen bonds,<sup>49</sup> and carbon–H-bonds<sup>50</sup> are also considered, all complexes displayed at least one such interaction between the fragment and protein. Figure 7 summarizes the distribution



**Figure 7.** Frequency distribution of additional directional interactions per fragment pocket including: arene-based interactions (i.e., arene–arene, arene–cation, and arene–hydrogen), carbon H-bonds, sulfur-mediated contacts, and halogen (i.e., iodine, bromine, and chlorine) bonds to protein atoms. The fractions of fragment complexes without arene (58%), carbon H-bonds (88%), sulfur contacts (89%), and halogen bonds (96%) have been omitted from the graph for clarity.

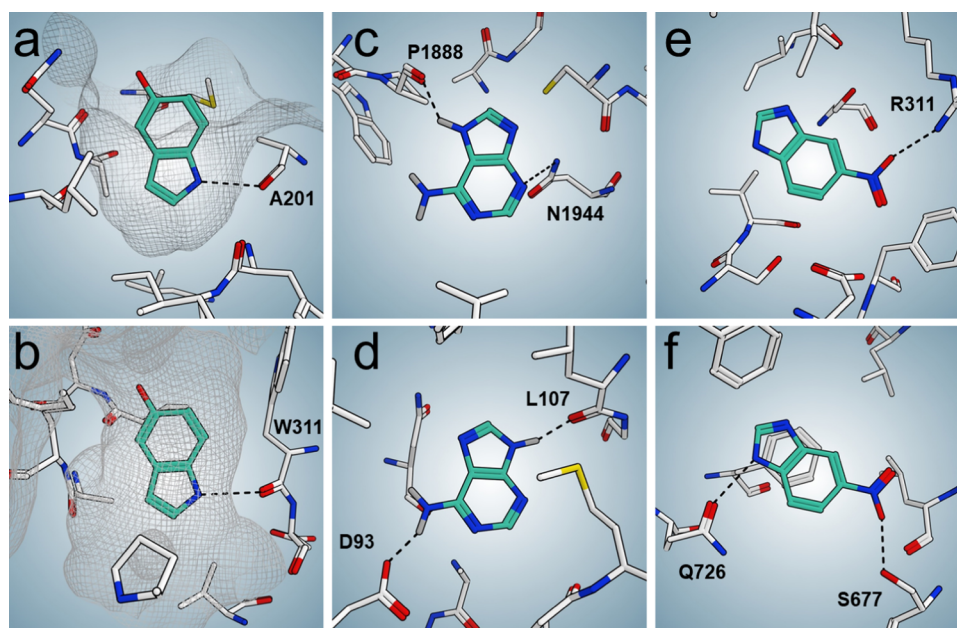
of additional directional interactions across the whole data set, as assessed using MOE's nonbonded-contacts detection algorithm<sup>30</sup> (Supporting Information). 56% of fragment hits establish at least one such interaction with the corresponding protein. Arene-based interactions occur most frequently (42% of cases), followed by carbon H-bonds (12%), sulfur-mediated contacts (11%), and halogen bonds (3%). As many as seven different non-H-bond interactions in a single complex have been found in the current set (PDB ID: SEGS, Figure 5d). Sulfur atoms in fragment hits establish, on average, 0.84 interactions, with sulfur–oxygen contacts the most frequent (0.4 per sulfur atom). Halogen bonds and carbon H-bonds occur markedly less often: 0.13 and 0.02 per atom, respectively (Table S7).

#### Comparison of Fragment Hits and Larger Ligands.

We constructed a comparative data set of 445 protein–ligand complexes (Table S8) such that any given protein pocket occurs with similar frequency as it does in the fragment data set. The 439 unique ligands are on average twice as large and twice as lipophilic as the fragment hits previously discussed and have a greater number of ring assemblies, rotatable bonds, and  $sp^3$ -hybridized atoms in their structures. There are no significant differences between the fragments and the ligands with respect to the distribution of formal charges or the ratios of atoms capable of establishing H-bonds (Figure S4).

The ligands are ~10% less likely than the fragments to bury their total SA, and they are equally less likely to bury their polar and apolar SA. In absolute terms, owing to their larger size and lipophilicity, ligands bury a greater amount of total SA, mainly of apolar nature, when compared to fragments (Figure S5).

On average, ligands establish one additional protein H-bond and twice as many arene interactions compared to fragments, whereas water H-bonds and additional directional interactions



**Figure 8.** Selected fragment hits bound to different proteins. (a, b) Different degrees of solvent exposure of 5-hydroxyindole bound to RadA (PDB ID: 4B3C) and leukotriene A-4 hydrolase (PDB ID: 3FUH), respectively. (c, d) Different tautomers of adenine bound to BAZ2B bromodomain (PDB ID: 5DYX) and Hsp90 (PDB ID: 2YED), respectively. (e, f) Different interactions of the nitro group of 5-nitro-benzimidazole bound to nicotinamide phosphoribosyltransferase (PDB ID: 4N9C) and PDE10A2 (PDB ID: 4MSA), respectively. Fragments are depicted as bold sticks (cyan carbon atoms) and relevant H-bonds as dashed black lines. Relevant protein pocket surfaces are displayed as gray mesh.



do not substantially change. Despite the higher number of protein H-bonds established, these bonds are 10% less likely to be fully buried, and the polar atoms of ligands show a reduced share of H-bonds compared to those of fragments (Figure S6).

## DISCUSSION

We compiled and analyzed a data set composed of 462 unique fragments bound to 168 different pockets on 126 individual proteins, resulting in a total of 489 fragment–pocket complexes (Figure 1). This work was made possible by a large number of crystallographic studies performed by scientists across numerous organizations in the past two decades (Table S9). Several systems are over-represented due to the large number of fragment-binding studies published for these proteins,<sup>39–44</sup> but the data set is nevertheless diverse in terms of structural domains and distinct protein families (Figure S1). From the analysis of this diverse data set, we observe a number of notable fragment features, discussed below, that can be used as a guide for the design, selection, and evaluation of fragments.

**Binding Versatility of the Fragment Hits.** 21 fragments bound to more than one protein pocket, in many cases on different proteins (Table 1). The ability of fragments to bind to multiple proteins<sup>16</sup> reinforces the appeal of using fragment-based methods to generate chemical starting points for drug discovery. In seven cases, a single fragment bound to different binding sites on its target protein (Table 1, Figure S5,f). Fragment screening is thus well suited to uncover and evaluate alternative binding sites and target interaction mechanisms of potential therapeutic relevance.<sup>51</sup> Interestingly, with the exception of saturated carbocyclic rings, these 21 fragments recapitulate most of the pharmacophoric elements typically exploited for molecular interactions. If aptly complemented with missing fragment pharmacophores and normalized for relative pharmacophoric occurrence, we recommend these fragments as a useful choice for a minimalistic, “first pass” library for pilot fragment screens, particularly for X-ray crystallography screening. Here, the so-called “promiscuity” of fragments in a well-defined structural context is understood as a practical advantage for mapping hotspots on a protein<sup>39,52,53</sup> and identifying fragment binders for further optimization. In the absence of relevant structural information, however, fragment promiscuity could be detrimental, especially when relying on biophysical screening.<sup>54</sup>

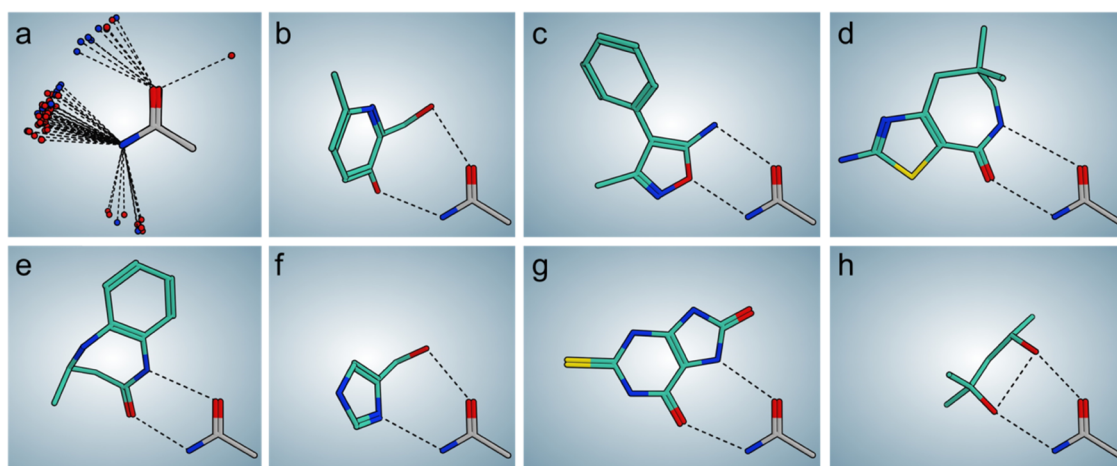
The fragment hits are remarkably versatile in their interactions with different binding pockets, as shown in Figure 8 for selected examples. 5-Hydroxyindole, for example, is fully engulfed by leukotriene A-4 hydrolase (PDB ID: 3FUH), in contrast to the complex this fragment forms with the DNA repair and recombination protein RadA, on which it binds to a highly solvent-exposed cleft (PDB ID: 4B3C). Interestingly, in both complexes, the fragment nitrogen atom (and not its 5-hydroxy group) is H-bonded to the protein. In another example of fragment versatility, adenine exploits several of its pharmacophoric elements when binding to the BAZ2B bromodomain and Hsp90 (PDB IDs: 5DYX and 2YED, respectively). The tautomerism of adenine’s imidazole moiety further adds to its adaptability; both the 7- and 9-position nitrogen atoms are independently H-bonded to backbone carbonyl groups on the two proteins. Another fragment, 5-nitro-benzimidazole, exemplifies how a structural element typically frowned upon in recent medicinal chemistry practice (i.e., the nitro group) can serve an important molecular-

recognition function in the early phases of drug discovery. Here, it secures H-bonds to both neutral (serine) and charged (arginine) side chains in the binding pockets of PDE10A2 (PDB ID: 4MSA) and nicotinamide phosphoribosyltransferase (PDB ID: 4N9C), respectively. These examples highlight the ability of fragments to effectively sample chemical space during screening.

**Properties of the Fragment Hits and Relevance to Fragment Design.** The fragment hits described here mostly comply with the rule-of-three (Ro3) guidelines,<sup>1</sup> with less than 5% of the set deviating from any of the Ro3 parameters. The propensity of fragment hits to display a quarter of their atoms as H-bond recognition elements (Figure 2) hints at a particular balance between exposed polarity and lipophilicity that is most conducive to productive interactions with different proteins, while ensuring that physicochemical properties are compatible with fragment-screening experiments. We thus strongly suggest favoring fragments with a polar atom fraction of  $\sim 0.25$  when evaluating novel fragment topologies and pharmacophores during fragment-library enrichment campaigns. Interestingly, in our independent data set of protein–ligand complexes, we found that the ligands have a similar fraction of polar atoms as the fragments (Figure S4), suggesting that this  $\sim 0.25$  polar atom fraction may be of general utility in ensuring favorable interactions with target proteins.

The limited degree of chirality and carbon  $sp^3$  saturation of the validated fragment hits is noteworthy, especially when coupled to their wide chemical diversity and over-reliance on a handful of topological skeletons (Figures 1–3). This might reflect historic trends in the first generation of fragment libraries, especially given the retrospective nature of the present study. Recently, there has been renewed interest in fragment structural complexity and three-dimensionality as driving forces in fragment design. This is borne out by the fact that the primary difference between the fragment hits presented here and those in a representative commercial fragment library from an established vendor in the fragment-based community<sup>46</sup> is the higher  $F_{sp^3}$  distribution of the latter (Figure 2). Nevertheless, it is noteworthy that achiral, heteroaromatic assemblies belonging to five topological frameworks can result in productive binding against a diverse set of pocket shapes and features. We suggest that the three-dimensional character of fragments (and molecules in general) is misrepresented by typical two-dimensional molecular descriptors such as  $F_{sp^3}$ . Several of the aromatic fragment hits in this study are indeed nonplanar both in their shape and, more importantly, in the way that they interact with the protein. This nonplanarity is afforded by virtue of ortho substituents (PDB ID: 5JAO), monoatomic linkers connecting individual rings (PDB ID: 2YE7), and a small amount of hydrocarbon saturation (PDB ID: 5FYU).

In our opinion, fragments should be kept relatively simple during fragment screening to maximize their potential interactions with proteins. To this end, we view the current set of 462 validated fragment hits as a relevant first approximation of a diverse fragment library with adequate structural complexity, which could then be subjected to further refinement based on, for example, target-specific hypotheses or diversity-optimization goals. In accordance with probabilistic interaction models,<sup>16</sup> the comparison in this work between the molecular properties of fragments and larger ligands (Figure S5) suggests that structural complexity can and should be built in at a later stage, during fragment optimization. The challenge



**Figure 9.** Fragment hits H-bonded to the side chain of asparagine. (a) Spatial distribution of the fragment atoms interacting with asparagine side chains ( $N = 70$ ). (b–h) Diverse selection of fragment hits engaged in paired H-bonds to the side chain of asparagine (PDB IDs: 4CUR, 4LR6, 4TZ8, 4YK0, 5DYU, 5E3G, and 5E9Y). Fragment hits and the asparagine side chain (backbone atoms omitted for clarity) are depicted as bold sticks (cyan and light gray carbon atoms, respectively) and relevant H-bonds as dashed black lines.

(and an underappreciated differentiation element) is to devise adequate synthetic protocols for fragment diversification.<sup>55</sup> As shown in Figure 3 and Table S2, there are still ample opportunities to generate novel fragment matter even when considering limited structural complexity (to illustrate this, it is worth noting that although the data set of fragment hits presented here and the representative commercial fragment library we analyze for comparison display comparable distributions of molecular properties (Figure 2), there is not a single fragment structure shared between the two sets). Furthermore, 7-membered rings are under-represented in both sets, both as individual rings and as part of fused systems. Their peculiar conformational preferences<sup>56</sup> and projected substitution vectors represent interesting design features that are well suited to a fragment-based context. A number of additional fragment design considerations, directly derived from the analysis of the molecular interactions between the fragment hits and target proteins studied here, will be presented in the following sections.

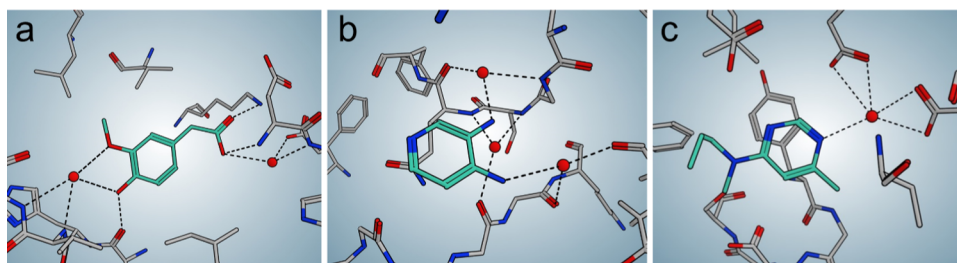
**Interactions of the Fragment Hits.** Fragments are smaller than lead compounds, and thus tend to have fewer productive interactions with target proteins (Figures S4 and S6). Although the fragment–protein interactions are typically referred to as “higher-quality” interactions, the net effect is a weaker binding affinity. The thermodynamics underlying fragment–protein binding is an area of active study.<sup>57–59</sup> Our current analysis of fragment hit–protein complexes serves to identify propensities in these interactions as approximated with surface-based and interaction-count descriptors. It is envisaged that these descriptors, together with the associated functional-group preference, could support triaging of fragment-screening results in virtual campaigns (e.g., molecular dynamics–based screening of a cocktail of different fragments) by, for example, reducing the number of false positives.

**Solvent Exposure and Protein Complementarity.** More than 70% of the fragment hits surveyed here, as validated by X-ray crystallography, reduce their total solvent exposure by >80% upon binding, a finding that is consistent with previously published comparisons of primary and secondary fragment-binding sites from proprietary databases.<sup>51</sup> Accordingly, we find that the polar fraction of the surfaces of fragments in this data set is almost entirely buried (>80%) upon binding. The

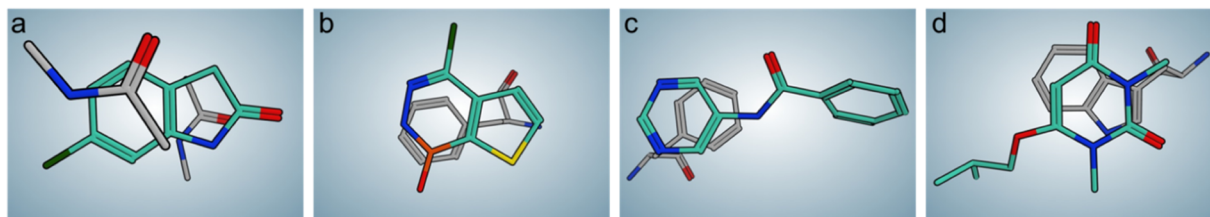
fragments consistently bury their polar SA regardless of the diverse physicochemical features of the observed protein pockets. Fragments tend to bury on average about twice as much apolar surface as polar surface (Figure 4), in line with the observed polar/apolar atomic composition of the fragments (Figure 2). Importantly, larger ligands of higher affinity reduce their total, polar, and apolar solvent exposure to a lesser degree than do fragments (Figure S5). Thermodynamic analyses have shown, however, that fragments cannot rely entirely on apolar desolvation as the main driver for binding.<sup>57,60</sup> Indeed, the large amounts of buried polar areas suggest a very effective use of fragment H-bond donor and acceptor functionalities. In most cases, the protein H-bond donors and acceptors are completely isolated from solvent (Figure 6) by virtue of significant apolar surface burial and are fully engaged in H-bonds with the fragment (see next section).

Taken together, these results are consistent with previous findings of enthalpically favored binding events at protein hotspots that are composed of polar sites buried in a lipophilic environment.<sup>57,60</sup> Given the observed fragment size and fraction of polar atoms, as well as typical H-bond chemical functionalities, a potential strategy to maximize molecular interaction diversity would be to present a minimum set of individual polar pharmacophoric elements, as opposed to distributing several pharmacophores on a given fragment. This strategy would provide fragments with greater freedom to satisfy the geometric constraints for optimal interactions. It would also result in better sampling of the reduced pharmacophoric space during fragment screening,<sup>16</sup> and additional pharmacophores could be built in and evaluated during the subsequent phase of fragment growing.

**H-Bonds to Protein and Water.** Stabilizing polar interactions are a recurring feature of the majority of the fragment–hit complexes (93%, Figure 6). These include strong attractive interactions, such as H-bonds to protein and to structural water molecules, as well as coordination bonds to catalytic metal ions. 87% and 58% of the complexes are stabilized by at least one or two H-bonds to the protein, respectively, most of which are completely isolated from solvent upon fragment binding (Figure 6). This result supports the previous finding of an average of two H-bonds per fragment from a minimally overlapping data set (the two overlapping entries are PDB IDs



**Figure 10.** Fragment hits H-bonded to structural water molecules. (a, b) Water molecules as part of extended nonbonded interaction networks (PDB ID: SMOH), (b, c) water molecules as the only H-bond partners for the fragment (PDB IDs: SNOW and 4Y3P, respectively). Fragment hits are depicted as bold sticks (cyan carbon atoms), water molecules as red spheres, and relevant H-bonds as dashed black lines.



**Figure 11.** Fragment hits engaged in arene-based interactions with the protein as the main attractive interaction in the absence of protein- and water-mediated H-bonds. (a) Arene interaction with backbone amide bonds (PDB ID: 4K2Y), (b, c) arene interaction with the side chains of phenylalanine (PDB IDs: 4Y37 and 5ISW), and (d) arene interaction with the side chain of tryptophan (PDB ID: 5JAN). Fragment hits are depicted as bold sticks (cyan carbon atoms).

3ESS and 3FGD).<sup>57</sup> Importantly, the substantial network of observed H-bonds provides an interaction context for the systematic and significant degree of fragment- and protein-polarity burial observed. The desolvation of polar groups on the fragment and the protein, as directed by the formed H-bonds, may result in an important enthalpic contribution to fragment binding, enhancing apolar desolvation, as previously reported.<sup>57,58,60</sup> It is noteworthy that the larger ligands cannot match the fragments' share of H-bonds per polar atom, and that the additional H-bonds formed tend to be more solvent-exposed than the ones established by fragments (Figure S6).

Fragment hits display a slight preference (58%) for establishing H-bonds to side-chain groups. Glycine is also highly represented as an H-bond target. These findings emphasize the importance of H-bond-site accessibility and geometric constraints, in addition to the need to populate diverse H-bond functionalities in fragment libraries, as summarized in Table 2. Functional groups with dual H-bond accepting and donating character (e.g., amide or alcohol groups) are particularly attractive for interaction-sampling purposes and can be further complemented by groups whose protomeric and tautomeric states can be influenced by the protein environment. Overall, the functional-group diversity observed in the fragments analyzed in the current study strengthens the conceptual appeal of fragment-based methods. An instructive example of fragment adaptability to molecular interactions is provided by selected fragments that demonstrate paired H-bonds to the side chain of asparagine residues across the present data set (Figure 9). In these fragments, nine individual atom types engaged the asparagine in paired H-bonds. The observed chemical and topological diversity is very inspiring for molecular design purposes, as it indicates opportunities for original bioisosteric replacements as well as for the optimization and diversification of pharmacophoric elements.

Water plays an important role in the binding of fragments to proteins.<sup>58,59</sup> When only high-resolution ( $\leq 1.5$  Å) structures

are considered, 46% of the fragment hits establish at least one H-bond to structural water molecules in the binding pocket. Water molecules could form, for example, extended nonbonded interaction networks by filling pocket cavities and offering interaction hotspots for fragments (Figure 10a,b). Importantly, in several cases, water-mediated H-bonds are the only polar interaction for the fragment hits (Figure 10b,c). This limits the ability of the energetic and solvation approximations used in current modeling software to adequately characterize and predict fragment binding using computational methods such as docking and molecular dynamics.<sup>59,61–63</sup>

**Beyond H-Bonds.** Although we are still far from a complete understanding of the energetics associated with H-bonds and metal-coordination bonds, these classes of polar interactions are relatively well studied, and medicinal chemists are accustomed to optimizing compounds based on them. Additional types of directional molecular interactions have only recently started to become more widely recognized, including arene-based contacts,<sup>27</sup> weak H-bonds, such as carbon H-bonds,<sup>50</sup> CH/ $\pi$  H-bonds,<sup>64</sup> halogen bonds,<sup>49</sup> and sulfur-mediated contacts.<sup>48</sup> More than half of the fragment hits display at least one such interaction, with arene contacts being the most frequent (42%). Although their occurrence is limited in comparison to canonical H-bonds, these additional directional interactions are likely to make important contributions to overall affinity in the context of fragment binding, where a reduced number of atoms is available for interactions. In a number of fragment hit–protein structures, such interactions stabilize the complex in the absence of more specific polar interactions (such as H-bonds), as shown by selected examples in Figure 11. Here, arene groups on the fragment hits are sandwiched against peptide bonds (Figure 11a) and stacked against the aromatic side chains of phenylalanine and tryptophan residues (Figure 11b–d).

The large variety of fragment heteroaromatic arrangements that are able to establish arene-type interactions is an

indication of future opportunities for the design of novel fragments and, more importantly, for the generation of intellectual property during fragment optimization. The thienodiazaborinine scaffold engaged in a face-to-face stacking interaction with phenylalanine 291 in the binding pocket of endothiapepsin (PDB ID: 4Y37, Figure 11) is an excellent example of under-represented and innovative heterocycles that could open up relevant pharmacophoric and chemical spaces for exploitation in fragment-based campaigns. The fine-tuning and optimization of such interactions at a fragment level still represent a significant challenge, given the conspicuous polarization and marked dispersive characteristics. To this end, the ability to query and mine existing structural data for nonbonded interactions, and to readily visualize<sup>65</sup> them in the context of a fragment-evolution effort, would greatly facilitate progress in this area.

## CONCLUSIONS

Analysis of the fragment–protein complexes curated here highlights salient features of validated fragment hits originating from fragment-based screening efforts. Despite limitations in sample size at both the fragment and the protein levels, we believe that this data set offers important insights relevant to hit-discovery activities, including the design and selection of fragments for fragment-screening libraries and the evaluation of the quality of fragment hits. The observed topological and functional-group diversity of fragments coupled with their polarity–lipophilicity balance could, for example, inform fragment-library selection and expansion schemes. Likewise, the observed surface and interaction-based propensities of the fragment–protein complexes could support the development of intuitive classification methods during *in silico* pocket and fragment-hit identification. As the number of deposited protein structures with bound fragment hits increases, the preliminary analysis presented here could be updated and used to refine empirical potentials for protein–fragment interactions and develop probabilistic models for molecular design applications. Our analysis emphasizes the essential role played by crystallographers and the importance of structural information in fragment-based thinking and methodologies.

The structural and chemical details revealed by publicly available protein–fragment hits have the potential to significantly impact molecular design and drug discovery. We believe that the ability of users of fragment-based approaches to distill this information for compound design can be a significant determinant of success during the fragment-hit evaluation and the fragment hit-to-lead phases. In these processes, interactive visualization of bound fragment hits across drug discovery projects could further enhance design and idea generation.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jmedchem.8b01855.

Figure S1: frequency distribution of interactions per fragment pocket; Figure S2: frequency distribution of total, polar, and apolar buried surface area (SA) of the fragments; Figure S3: characteristics of fragment hit protein targets; Figure S4: comparison between distributions of fragment hit and ligand features; Figure

S5: comparison between surface characteristics of the fragment– and ligand–pocket complexes; Figure S6: comparison between interactions of the fragment– and ligand–pocket complexes; Table S1: occurrence of multiple fragment hit binding sites across various protein targets; Table S2: occurrence of Bemis–Murcko frameworks across 462 unique fragment hits; Table S3: occurrence of explicit individual ring assemblies across 462 unique fragment hits; Table S4: descriptive statistics of the fragment buried SA upon binding; Table S5: fragment atoms involved in coordination bonds to structural metal ions in protein binding pockets; Table S6: fragment atoms involved in H-bonds with structural water molecules in protein binding pockets with crystallographic resolution  $\leq 1.5$  Å; Table S7: fragment atoms involved in sulfur contacts, halogen bonds, and carbon hydrogen bonds with the protein; Table S8: PDB codes for the 445 protein–ligand complexes; Table S9: PDB codes, SMILES strings, and citation information for the 489 pocket–fragment complexes (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

\*E-mail: Fabrizio.Giordanetto@DEShawResearch.com. Tel: (212) 478-0822. Fax: (212) 845-1822 (F.G.).

\*E-mail: David.Shaw@DEShawResearch.com. Tel: (212) 478-0260. Fax: (212) 845-1286. (D.E.S.).

### ORCID

Fabrizio Giordanetto: 0000-0001-9876-9552

David E. Shaw: 0000-0001-8265-5761

### Notes

The authors declare the following competing financial interest(s): This study was conducted and funded internally by D. E. Shaw Research, of which D.E.S. is the sole beneficial owner and Chief Scientist, and with which all authors are affiliated.

## ACKNOWLEDGMENTS

The authors thank Michael Eastwood, David Borhani, Yakov Pechersky, and Dina Sharon for helpful discussions and a critical reading of the manuscript and Rebecca Bish-Cornelissen and Berkman Frank for editorial assistance.

## ABBREVIATIONS

ECFP4, extended connectivity fingerprints; PDE10A2, cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A2; Hsp90, heat-shock protein 90; PAINS, pan-assay interference compounds; BAZ2B, bromodomain adjacent to zinc finger domain protein 2B

## REFERENCES

- (1) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A “Rule of Three” for Fragment-Based Lead Discovery? *Drug Discovery Today* **2003**, *8*, 876–877.
- (2) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274*, 1531–1534.
- (3) Scott, D. E.; Coyne, A. G.; Hudson, S. A.; Abell, C. Fragment-Based Approaches in Drug Discovery and Chemical Biology. *Biochemistry* **2012**, *51*, 4990–5003.
- (4) Erlanson, D. A. Introduction to Fragment-Based Drug Discovery. *Top. Curr. Chem.* **2012**, *317*, 1–32.

- (5) Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nat. Chem.* **2009**, *1*, 187–192.
- (6) Visegrádý, A.; Keserű, G. M. Fragment-Based Lead Discovery on G-Protein-Coupled Receptors. *Expert Opin. Drug Discovery* **2013**, *8*, 811–820.
- (7) Murray, C. W.; Verdonk, M. L.; Rees, D. C. Experiences in Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **2012**, *33*, 224–232.
- (8) Heikkilä, T. J.; Surade, S.; Silvestre, H. L.; Dias, M. V. B.; Ciulli, A.; Bromfield, K.; Scott, D.; Howard, N.; Wen, S.; Wei, A. H.; Osborne, D.; Abell, C.; Blundell, T. L. Fragment-Based Drug Discovery in Academia: Experiences From a Tuberculosis Programme. In *From Molecules to Medicines*; Sussman, J. L.; Spadon, P., Eds.; NATO Science for Peace and Security Series A: Chemistry and Biology; Springer: Netherlands, 2009; pp 21–36.
- (9) Zartler, E. R. Fragonomics: The -omics with Real Impact. *ACS Med. Chem. Lett.* **2014**, *5*, 952–953.
- (10) Romasanta, A. K. S.; van der Sijde, P.; Hellsten, I.; Hubbard, R. E.; Keseru, G. M.; van Muijlwijk-Koezen, J.; de Esch, I. J. P. When Fragments Link: A Bibliometric Perspective on the Development of Fragment-Based Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1596–1609.
- (11) Tsai, J.; Lee, J. T.; Wang, W.; Zhang, J.; Cho, H.; Mamo, S.; Bremer, R.; Gillette, S.; Kong, J.; Haass, N. K.; Sproesser, K.; Li, L.; Smalley, K. S.; Fong, D.; Zhu, Y. L.; Marimuthu, A.; Nguyen, H.; Lam, B.; Liu, J.; Cheung, L.; Rice, J.; Suzuki, Y.; Luu, C.; Settachatgul, C.; Shellooe, R.; Cantwell, J.; Kim, S. H.; Schlessinger, J.; Zhang, K. Y.; West, B. L.; Powell, B.; Habets, G.; Zhang, C.; Ibrahim, P. N.; Hirth, P.; Artis, D. R.; Herlyn, M.; Bollag, G. Discovery of a Selective Inhibitor of Oncogenic B-Raf Kinase with Potent Antimelanoma Activity. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3041–3046.
- (12) Bollag, G.; Tsai, J.; Zhang, J.; Zhang, C.; Ibrahim, P.; Nolop, K.; Hirth, P. Vemurafenib: The First Drug Approved for BRAF-Mutant Cancer. *Nat. Rev. Drug Discovery* **2012**, *11*, 873–886.
- (13) Souers, A. J.; Levenson, J. D.; Boghaert, E. R.; Ackler, S. L.; Catron, N. D.; Chen, J.; Dayton, B. D.; Ding, H.; Enschede, S. H.; Fairbrother, W. J.; Huang, D. C. S.; Hymowitz, S. G.; Jin, S.; Khaw, S. L.; Kovar, P. J.; Lam, L. T.; Lee, J.; Maecker, H. L.; Marsh, K. C.; Mason, K. D.; Mitten, M. J.; Nimmer, P. M.; Oleksijew, A.; Park, C. H.; Park, C.-M.; Phillips, D. C.; Roberts, A. W.; Sampath, D.; Seymour, J. F.; Smith, M. L.; Sullivan, G. M.; Tahir, S. K.; Tse, C.; Wendt, M. D.; Xiao, Y.; Xue, J. C.; Zhang, H.; Humerickhouse, R. A.; Rosenberg, S. H.; Elmore, S. W. ABT-199, a Potent and Selective BCL-2 Inhibitor, Achieves Antitumor Activity While Sparing Platelets. *Nat. Med.* **2013**, *19*, 202–208.
- (14) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhotti, H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discovery* **2016**, *15*, 605–619.
- (15) Jencks, W. P. On the Attribution and Additivity of Binding Energies. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 4046–4050.
- (16) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.
- (17) Raymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722–730.
- (18) Schuffenhauer, A.; Ruedisser, S.; Marzinzik, A. L.; Jahnke, W.; Blommers, M.; Selzer, P.; Jacoby, E. Library Design for Fragment Based Screening. *Curr. Top. Med. Chem.* **2005**, *5*, 751–762.
- (19) Schiebel, J.; Radeva, N.; Krimmer, S. G.; Wang, X.; Stieler, M.; Ehrmann, F. R.; Fu, K.; Metz, A.; Huschmann, F. U.; Weiss, M. S.; Mueller, U.; Heine, A.; Klebe, G. Six Biophysical Screening Methods Miss a Large Proportion of Crystallographically Discovered Fragment Hits: A Case Study. *ACS Chem. Biol.* **2016**, *11*, 1693–1701.
- (20) Murray, C. W.; Blundell, T. L. Structural Biology in Fragment-Based Drug Design. *Curr. Opin. Struct. Biol.* **2010**, *20*, 497–507.
- (21) Johnson, C. N.; Erlanson, D. A.; Murray, C. W.; Rees, D. C. Fragment-to-Lead Medicinal Chemistry Publications in 2015. *J. Med. Chem.* **2017**, *60*, 89–99.
- (22) Johnson, C. N.; Erlanson, D. A.; Jahnke, W.; Mortenson, P. N.; Rees, D. C. Fragment-to-Lead Medicinal Chemistry Publications in 2016. *J. Med. Chem.* **2018**, *61*, 1774–1784.
- (23) Mortenson, P. N.; Erlanson, D. A.; de Esch, I. J. P.; Jahnke, W.; Johnson, C. N. Fragment-to-Lead Medicinal Chemistry Publications in 2017. *J. Med. Chem.* **2018**, DOI: 10.1021/acs.jmedchem.8b01472.
- (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) Deller, M. C.; Rupp, B. Models of Protein-Ligand Crystal Structures: Trust, but Verify. *J. Comput.-Aided Mol. Des.* **2015**, *29*, 817–836.
- (26) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins* **2009**, *75*, 187–205.
- (27) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, S061–S084.
- (28) JChem, Version 16.4.25; ChemAxon, 2016. <http://www.chemaxon.com>.
- (29) Carugo, O.; Bordo, D. How Many Water Molecules Can Be Detected by Protein Crystallography? *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1999**, *55*, 479–483.
- (30) *Molecular Operating Environment (MOE)*, 2015.10; Chemical Computing Group, Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2015.
- (31) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- (32) Petrauskas, A. A.; Kolovanov, E. A. ACD/Log P Method Description. *Perspect. Drug Discovery Des.* **2000**, *19*, 99–116.
- (33) *ACD Percepta Batch, 2012 Release*; Advanced Chemistry Development, Inc.: 8 King Street East, Suite 107, Toronto, Ontario, M5C 1B5, Canada, 2016.
- (34) *Vortex*, 2016.08; Dotmatics Limited: The Old Monastery, Windhill, Bishops Stortford, Herts CM23 2ND UK, 2016.
- (35) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, No. 168.
- (36) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
- (37) Chen, Y. C.; Tolbert, R.; Aronov, A. M.; McGaughey, G.; Walters, W. P.; Meireles, L. Prediction of Protein Pairs Sharing Common Active Ligands Using Protein Sequence, Structure, and Ligand Similarity. *J. Chem. Inf. Model.* **2016**, *56*, 1734–1745.
- (38) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
- (39) Radeva, N.; Krimmer, S. G.; Stieler, M.; Fu, K.; Wang, X.; Ehrmann, F. R.; Metz, A.; Huschmann, F. U.; Weiss, M. S.; Mueller, U.; Schiebel, J.; Heine, A.; Klebe, G. Experimental Active-Site Mapping by Fragments: Hot Spots Remote from the Catalytic Center of Endothiapepsin. *J. Med. Chem.* **2016**, *59*, 7561–7575.
- (40) Köster, H.; Craan, T.; Brass, S.; Herhaus, C.; Zentgraf, M.; Neumann, L.; Heine, A.; Klebe, G. A Small Nonrule of 3 Compatible Fragment Library Provides High Hit Rate of Endothiapepsin Crystal Structures with Various Fragment Chemotypes. *J. Med. Chem.* **2011**, *54*, 7784–7796.
- (41) Recht, M. I.; Sridhar, V.; Badger, J.; Bounaud, P.-Y.; Logan, C.; Chie-Leon, B.; Nienaber, V.; Torres, F. E. Identification and Optimization of PDE10A Inhibitors Using Fragment-Based Screening by Nanocalorimetry and X-Ray Crystallography. *J. Biomol. Screening* **2014**, *19*, 497–507.
- (42) Shipe, W. D.; Sharik, S. S.; Barrow, J. C.; McGaughey, G. B.; Theberge, C. R.; Uslander, J. M.; Yan, Y.; Renger, J. J.; Smith, S. M.; Coleman, P. J.; Cox, C. D. Discovery and Optimization of a Series of Pyrimidine-Based Phosphodiesterase 10A (PDE10A) Inhibitors through Fragment Screening, Structure-Based Design, and Parallel Synthesis. *J. Med. Chem.* **2015**, *58*, 7888–7894.
- (43) Murray, C. W.; Carr, M. G.; Callaghan, O.; Chessari, G.; Congreve, M.; Cowan, S.; Coyle, J. E.; Downham, R.; Figueroa, E.;

Frederickson, M.; Graham, B.; McMenamin, R.; O'Brien, M. A.; Patel, S.; Phillips, T. R.; Williams, G.; Woodhead, A. J.; Woolford, A. J.-A. Fragment-Based Drug Discovery Applied to Hsp90. Discovery of Two Lead Series with High Ligand Efficiency. *J. Med. Chem.* **2010**, *53*, 5942–5955.

(44) Davies, N. G. M.; Browne, H.; Davis, B.; Drysdale, M. J.; Foloppe, N.; Geoffrey, S.; Gibbons, B.; Hart, T.; Hubbard, R.; Jensen, M. R.; Mansell, H.; Massey, A.; Matassova, N.; Moore, J. D.; Murray, J.; Pratt, R.; Ray, S.; Robertson, A.; Roughley, S. D.; Schoepfer, J.; Scriven, K.; Simmonite, H.; Stokes, S.; Surgenor, A.; Webb, P.; Wood, M.; Wright, L.; Brough, P. Targeting Conserved Water Molecules: Design of 4-Aryl-5-cyanopyrrolo[2,3-D]pyrimidine Hsp90 Inhibitors Using Fragment-Based Screening and Structure-Based Optimization. *Bioorg. Med. Chem.* **2012**, *20*, 6770–6789.

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

(46) Enamine: Golden Fragment Library (2016). [https://enamine.net/download/FL/Enamine\\_Golden\\_Fragment\\_Library.pdf](https://enamine.net/download/FL/Enamine_Golden_Fragment_Library.pdf) retrieved Dec 23, 2018.

(47) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. I. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(48) Zhang, X.; et al. Intermolecular Sulfur···Oxygen Interactions: Theoretical and Statistical Investigations. *J. Chem. Inf. Model.* **2015**, *55*, 2138–2153.

(49) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116*, 2478–2601.

(50) Horowitz, S.; Trievel, R. C. Carbon-Oxygen Hydrogen Bonding in Biological Structure and Function. *J. Biol. Chem.* **2012**, *287*, 41576–41582.

(51) Ludlow, R. F.; Verdonk, M. L.; Saini, H. K.; Tickle, I. J.; Jhoti, H. Detection of Secondary Binding Sites in Proteins Using Fragment Screening. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 15910–15915.

(52) Hall, D. R.; Kozakov, D.; Whitty, A.; Vajda, S. Lessons from Hot Spot Analysis for Fragment-Based Drug Discovery. *Trends Pharmacol. Sci.* **2015**, *36*, 724–736.

(53) Radoux, C. J.; Olsson, T. S. G.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Identifying Interactions That Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.* **2016**, *59*, 4314–4325.

(54) Devine, S. M.; Mulcair, M. D.; Debono, C. O.; Leung, E. W. W.; Nissink, J. W. M.; Lim, S. S.; Chandrashekar, I. R.; Vazirani, M.; Mohanty, B.; Simpson, J. S.; Baell, J. B.; Scammells, P. J.; Norton, R. S.; Scanlon, M. J. Promiscuous 2-Aminothiazoles (PrATs): A Frequent Hitting Scaffold. *J. Med. Chem.* **2015**, *58*, 1205–1214.

(55) Murray, C. W.; Rees, D. C. Opportunity Knocks: Organic Chemistry for Fragment-Based Drug Discovery (FBDD). *Angew. Chem., Int. Ed.* **2016**, *55*, 488–492.

(56) Brameld, K. A.; Kuhn, B.; Reuter, D. C.; Stahl, M. Small Molecule Conformational Preferences Derived from Crystal Structure Data. A Medicinal Chemistry Focused Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 1–24.

(57) Ferenczy, G. G.; Keserü, G. M. Thermodynamics of Fragment Binding. *J. Chem. Inf. Model.* **2012**, *52*, 1039–1045.

(58) Ferenczy, G. G.; Keserü, G. M. On the Enthalpic Preference of Fragment Binding. *Med. Chem. Commun.* **2016**, *7*, 332–337.

(59) Rühmann, E.; Betz, M.; Heine, A.; Klebe, G. Fragment Binding Can Be Either More Enthalpy-Driven or Entropy-Driven: Crystal Structures and Residual Hydration Patterns Suggest Why. *J. Med. Chem.* **2015**, *58*, 6960–6971.

(60) Olsson, T. S. G.; Williams, M. A.; Pitt, W. R.; Ladbury, J. E. The Thermodynamics of Protein–Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* **2008**, *384*, 1002–1017.

(61) Sándor, M.; Kiss, R.; Keserü, G. M. Virtual Fragment Docking by Glide: A Validation Study on 190 Protein–Fragment Complexes. *J. Chem. Inf. Model.* **2010**, *50*, 1165–1172.

(62) Verdonk, M. L.; Giangreco, L.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking Performance of Fragments and Druglike Compounds. *J. Med. Chem.* **2011**, *54*, 5422–5431.

(63) Klebe, G. Applying Thermodynamic Profiling in Lead Finding and Optimization. *Nat. Rev. Drug Discovery* **2015**, *14*, 95–110.

(64) Nishio, M. The CH/ $\pi$  hydrogen bond in chemistry. Conformation, supramolecules, optical resolution and interactions involving carbohydrates. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13873–13900.

(65) Korb, O.; Kuhn, B.; Hert, J.; Taylor, N.; Cole, J.; Groom, C.; Stahl, M. Interactive and Versatile Navigation of Structural Databases. *J. Med. Chem.* **2016**, *59*, 4257–4266.

(66) The UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43*, D204–D212.