# Similarity Searching Using Fingerprints of Molecular Fragments Involved in Protein−Ligand Interactions

Lu Tan, Eugen Lounkine, and Jürgen Bajorath*

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

Received September 05, 2008

To incorporate protein−ligand interaction information into conventional two-dimensional (2D) fingerprint searching, interacting fragments of active compounds were extracted from X-ray structures of protein−ligand complexes and encoded as structural key-type fingerprints. Similarity search calculations with fingerprints derived from interacting fragments were compared to fingerprints of complete ligands and control fragments. In these calculations, fingerprints of interacting fragments produced significantly higher compound recall than other fingerprints. These results indicate that ligand fragments involved in protein−ligand interactions carry much activity-specific chemical information that can be exploited in similarity searching without explicitly accounting for interaction information.

## INTRODUCTION

Ligand-[1−3] and structure-based[3−5] computational screening approaches have long coexisted in pharmaceutical research. Thus far, only a limited number of efforts have been made to combine ligand- and structure-based methods. In several studies, similarity searching and docking have been applied in a serial manner[6−9] and it has also been attempted to merge these approaches through parallel compound selection schemes or data fusion.[10] In addition, extensive comparisons of various ligand- and structure-based methods have been performed.[11] Despite such serial applications and method comparisons, a fully integrated ligand- and structure-based screening approach is currently not available. Importantly, structural knowledge can also be included in the computational search for novel active compounds by making use of protein−ligand interaction, rather than target structure information. For example, the Structural Interaction Fingerprint (SIFt) approach[12,13] transforms protein−ligand interaction information into a one-dimensional bit string representation that can be used to compare protein−ligand complexes or filter three-dimensional (3D) databases for candidate compounds. An alternative approach, Fingerprints for Ligands And Proteins (FLAP)[14] describes protein binding sites and ligand interaction as four-point pharmacophore fingerprints and a binding site shape component. Similar to SIFt, FLAP is also versatile in its use. It can be applied, for example, to compare protein binding sites and perform structure- or pharmacophore-based search calculations. Interaction or pharmacophore fingerprints are conformation-dependent approaches. However, different from docking methods, they do not require the calculation of binding energies to account for protein−ligand interactions.

Departing from currently available computational approaches that take target structure or protein−ligand interaction information explicitly into account, we have considered ways of incorporating interaction information into conventional 2D fingerprint similarity searching[1−4] without the need to explicitly encode such interactions. We reasoned that the formation of specific interactions ultimately is dependent on the complementarity of protein binding site and ligand features. Therefore, we can formulate the hypothesis that interaction information might be implicitly captured by sets of ligand atoms that engage in interactions with the target protein. Then it should be possible to add interaction information to ligand-based screening by putting emphasis on the interacting parts of ligands during similarity evaluation.

In this study, we have tested this hypothesis by extracting fragments of ligands available in complex crystal structures that exclusively consisted of interacting atoms. We then calculated conventional structural key-type fingerprints for interacting fragments, complete ligands, and control fragments that were obtained by random deletion of ligand atoms or by deletion of interacting atoms. These fingerprints were compared in standard search calculations using multiple reference compounds. For three of the four compound classes that we have studied, fingerprints calculated for interacting fragments produced in part significantly higher recall of active compounds than fingerprints of complete ligands. In contrast, fingerprints of control fragments largely lost their ability to recognize active molecules. Taken together, these findings indicate that the interacting atoms of ligands capture activity-specific chemical information that can be easily utilized in ligand-based similarity searching.

## MATERIALS AND METHODS

**Protein−Ligand Complexes.** We have selected target proteins for which different protein−ligand complex structures were available in the Protein Data Bank (PDB),[15] based on the following criteria: (a) only noncovalent protein−ligand interactions, (b) multiple unique ligands (i.e., no close analogs), (c) no target protein mutants, and (d) high crystallographic resolution. Accordingly, four enzymes were se-

* To whom correspondence should be addressed. Tel.: +49-228-2699-306. Fax: +49-228-2699-341. E-mail address: bajorath@bit.uni-bonn.de.

10.1021/ci800322y CCC: $40.75 
Published on Web 11/13/2008

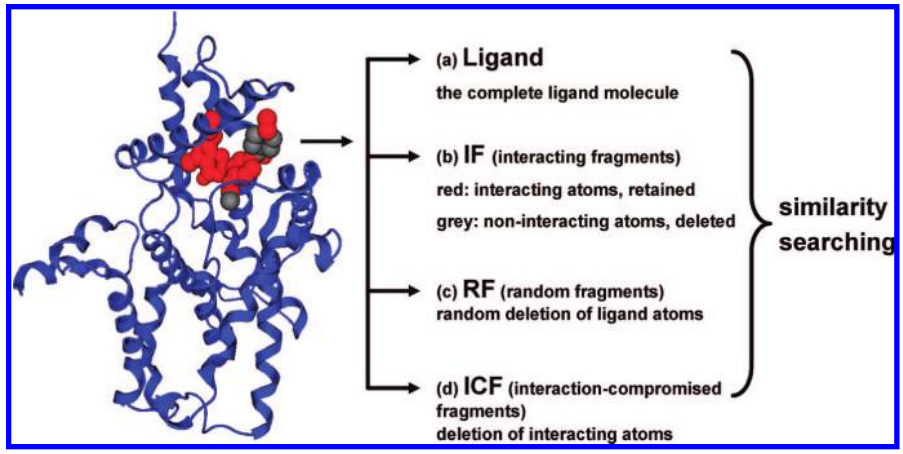


**Figure 1.** Method overview. The complex of PD with vardenafil is shown (PDB entry 1xot) with interacting ligand atoms colored red. Interacting, random, and interaction-compromised fragments are described. From these different types of fragments, fingerprints are calculated for similarity searching.

lected for our study: thrombin (TH), phosphodiesterase IV (PD), factor Xa (XA), and dihydrofolate reductase (DR). Ten complex structures were selected in each case, with the exception of XA, for which eight structures were selected. The selected PDB structures and their crystallographic resolution are reported in Table S1 in the Supporting Information and Figure S1 in the Supporting Information shows 2D structures of all of the crystallographic ligands.

**Ligand Fragments.** Three types of fragments were generated in this study: (1) interacting fragments (IFs), which only consisted of ligand atoms involved in protein−ligand interactions; (2) random fragments (RFs), which were generated by a random deletion of ligand atoms; and (3) interaction-compromised fragments (ICFs), which were generated via the deletion of interacting atoms. RFs and ICFs served as control fragments for IFs. For each IF consisting of $n$(ligand) − $k$(not interacting) atoms, five corresponding RFs and ICFs were calculated by random deletion of $k$(ligand) and $k$(interacting) atoms, respectively. Exemplary IF, RF, and ICF sets are shown in Figure S2 of the Supporting Information.

To determine the interacting atoms, hydrogen bonding, ionic, and van der Waals (vdW)/hydrophobic interactions between protein and ligand atoms were calculated using the Molecular Operating Environment (MOE).[16] The cutoff distance for hydrogen bonds was set to 3.8 Å, and for ionic and vdW interactions, a 4.5 Å cutoff distance was applied.

Fragment generation is summarized in Figure 1. The PD inhibitor shown, vardenafil, consists of a total of 34 heavy atoms, 26 of which are interacting atoms that form the IF. Five unique RF are generated by random deletion of eight interacting or noninteracting atoms and five unique ICF by random deletion of eight interacting atoms. Interacting and control fragments might be coherent or consist of multiple subfragments, depending on which ligand atoms are deleted.

**Fingerprints and Similarity Searching.** For each ligand and the corresponding fragments, MACCS structural keys[17] were calculated and used as a fingerprint for similarity searching. Thus, interacting and control fragments were not encoded in a fingerprint format; rather, for each fragment set, the corresponding MACCS keys were derived. The Tanimoto coefficient (Tc)[18] was calculated as a measure of fingerprint similarity. Each ligand and fragment was used as an individual template for similarity search calculations. For RFs and ICFs, compound recall rates were averaged over calculations using five unique control fragments. Furthermore, similarity searching using multiple reference compounds and fragments was performed by applying the $n$ Nearest Neighbor ($n$NN) search strategy,[18] which averages the Tc values for all templates to yield the final similarity value for a database compound. Here, contributions from all ligands (i.e., 10NN for TH, PD, and XA, and 8NN for DR) or fragments were considered equally. Recovery rates of active compounds were determined for the top scoring 200 database compounds and also monitored in cumulative recall curves over 1000 compounds for $n$NN search calculations.

**Compound Database.** As a background database for similarity searching, 10 000 compounds with a molecular weight of ≤600 Da and no activity against the four targets were randomly taken from the Molecular Drug Data Report (MDDR).[19] For our four target proteins, 59−640 known active compounds with pairwise Tc values of ≤0.80 (thereby avoiding the inclusion of closely related analogs) were also taken from the MDDR and added to the screening database as potential hits, as reported in Table 1.

**Table 1.** Active Compounds[a]

| protein | PDB complexes | reference compounds | average Tanimoto coefficient, Tc | active MDDR compounds |
|---|---|---|---|---|
| TH | 1nm6, 1qbv, 1w7g, 1bcu, 1sl3, 1vzq, 1c1u, 1c4v, 1mu6, 1a4w | 10 | 0.48 | 238 |
| PD | 1mkd, 1ro6, 1xlx, 1xlz, 1xm6, 1xmu, 1xot, 1zkn, 2fm0, 1y2c | 10 | 0.40 | 640 |
| XA | 1ezq, 1fax, 1fjs, 1g2l, 1kye, 1v3x, 1z6e, 2fzz | 8 | 0.49 | 393 |
| DR | 1boz, 1hfp, 1kms, 1kmv, 1mvs, 1ohj, 1s3u, 1s3w, 1yho, 2c2s | 10 | 0.55 | 59 |

[a] The PDB identifiers are given for each of the complexes, together with the total number of crystallographic reference compounds and their average pair-wise MACCS Tanimoto similarity. In addition, the number of active MDDR compounds available for each class is reported.

**Table 2.** Atom Numbers and Structural Keys[a]

| protein | heavy atom numbers | | structural keys | | | |
|---|---|---|---|---|---|---|
| | ligand | fragments (IF, RF, ICF) | ligand | IF | RF | ICF |
| TH | 31 | 19 | 60 | 38 (32) | 32 (25) | 32 (25) |
| PD | 24 | 19 | 50 | 38 (34) | 35 (27) | 35 (29) |
| XA | 37 | 23 | 65 | 41 (35) | 35 (28) | 43 (36) |
| DR | 27 | 15 | 50 | 27 (23) | 22 (17) | 24 (20) |
| average | 29 | 19 | 56 | 36 (31) | 31 (24) | 33 (27) |

[a] For every target enzyme, the average numbers of ligand and fragment atoms are reported and the average number of MACCS structural keys calculated from them. The number of structural keys that are shared by ligands and fragments is reported in parentheses.

## RESULTS AND DISCUSSION

**Interacting and Control Fragments.** The basic idea underlying the approach reported herein is to reduce active compounds to fragments that are involved in significant interactions with their biological targets and evaluate the activity-related chemical information content of these fragments in similarity searching. As reported in Table 2 and in Table S2 in the Supporting Information, all crystallographic ligands studied here contained a varying number of atoms that did not participate in measured protein−ligand interactions. On average, ligands consisted of 29 non-hydrogen atoms, 19 of which were involved in protein−ligand interactions, i.e. hydrogen bonds, ionic interactions, or vdW contacts. These atoms constituted interacting fragments.

To evaluate the information content of the IFs, the availability of carefully designed control fragments was considered to be of crucial importance. Therefore, we generated two types of control fragments equal in size to IF, randomly generated fragments (where we did not distinguish between interacting and noninteracting atoms (RF)), and interaction-compromised fragments (where we exclusively deleted interacting atoms, also in a random fashion (ICF)). RF provide a control for the general loss of chemical information that is associated with atom deletion and fragmentation, whereas ICF specifically evaluate the omission of atoms that are most relevant for the formation of specific protein−ligand interactions. Thus, IF and ICF represent "inverse" fragment designs.

**Structural Keys.** From ligands and corresponding fragments, MACCS structural keys were calculated to provide

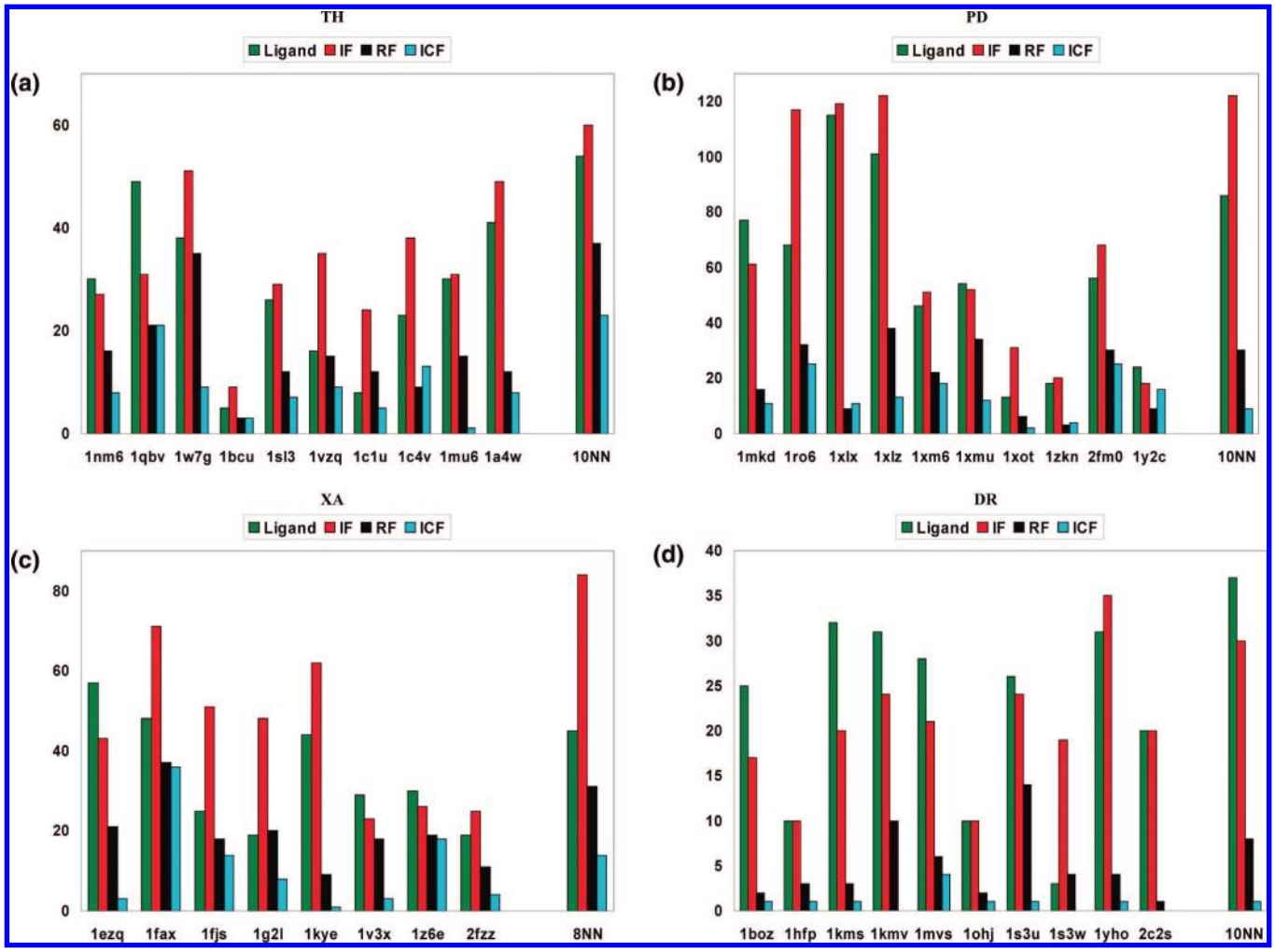


**Figure 2.** Recall of active compounds. For each target enzyme and individual ligand, recovery rates are reported for the 200 top-ranked database compounds in similarity search calculations, using the complete ligand (green), IF (red), RF (black), and ICF (cyan): (a) TH; (b) PD; (c) XA; (d) DR. In each case, the *y*-axis reports the total number of recovered active compounds. For RFs and ICFs, the results are averaged for five randomly generated control fragments. In addition to search results for individual ligands (labeled with the PDB identification of their respective complexes), results of *n*NN calculations that equally take contributions of all ligands or fragments into account are also reported. For DR, ICF search calculations failed to recover any active compounds for three ligands (1kmv, 1s3w, and 2c2s).

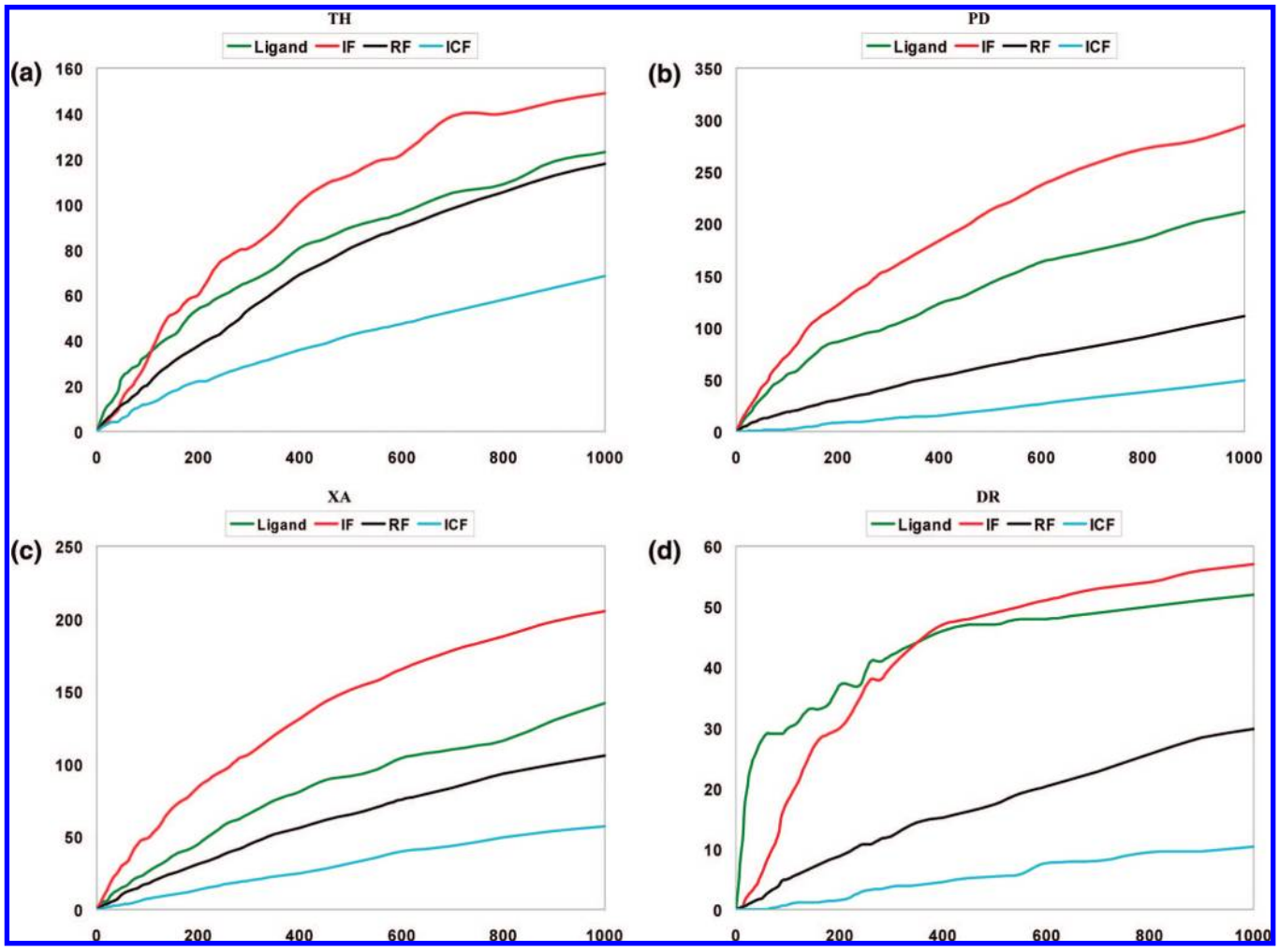


**Figure 3.** Cumulative recall curves. For *n*NN calculations, recall curves are reported for the top 1000 database compounds: (a) TH, (b) PD, (c) XA, and (d) DR. The *y*-axis reports the total number of recovered active compounds and the *x*-axis is the number of ranked database molecules. The color code is according to that shown in Figure 2.

a common reference frame for their comparison and enable similarity searching. Complete ligands produced, on average, 56 of 166 standard MACCS keys, compared to 36 keys for IF, 31 for RF, and 33 for ICF (see Table 2). Thus, key sets of our interacting and control fragments were of comparable size, but there was a considerable reduction in the structural information associated with fragmentation, compared to the original ligands. Structural keys calculated from fragments largely overlapped with those of ligands, but were not exact subsets, as also reported in Table 2. Although fragments yielded, on average, ~20 fewer keys than ligands, they also produced 4−8 structural keys that were not generated by ligands, because the bonding patterns and, thus, molecular topology changed, as a consequence of atom deletions.

**Similarity Searching.** The MACCS fingerprints of ligands and corresponding fragments were subjected to similarity searching using single and multiple templates. Figure 2 shows the recall of active molecules among the top ranked 200 database compounds. Strikingly, the recovery rates for interacting fragments are at least comparable to those achieved by individual ligands in essentially all cases. In fact, for three targets (TH, PD, and XA), the recovery rates of interacting fragments were generally higher. By contrast, recovery rates of random control fragments were always significantly lower than that observed for IF or complete ligands. Moreover, for the majority of ligands across all targets, ICF recovery rates were lower than RF rates. The results of *n*NN multiple reference compound similarity search calculations, which are also reported in Figure 2, further substantiated these trends. When contributions from all ligands and fragments were taken into account, IFs displayed the best search performance for three of four targets, except for DR, for which complete ligands were superior. For DR, crystallographic ligands, overall, were more similar than for other classes but produced fewer interacting fragments (see Table 1), which is likely to explain the advantage of the entire molecule over IF similarity searching in this case. Furthermore, RF showed significantly reduced recall for all four targets and recovery rates were always further reduced for ICF. The cumulative recall curves in Figure 3 also show that IF similarity searching produced overall highest recall of active compounds, whereas ICF produced lowest.

**Interaction Information.** These results of our similarity search calculations showed that fragments consisting of atoms involved in protein−ligand interactions were most effective in detecting active compounds, although they were smaller than entire ligands and produced fewer structural keys. In contrast, a random control fragment of equal size to IF displayed a reduced ability to detect active compounds. Moreover, deliberate deletion of interacting atoms in ICF further reduced the ability of control fragments to detect active compounds.

As discussed previously, fragmentation generally led to fingerprints containing fewer structural keys than those of original ligands. In principle, this corresponds to a loss in chemical information that might lead to reduced similarity search performance if this information is relevant for the detection of SARs. This is consistent with our observation that fingerprints of RF produced lower compound recall than ligands. Furthermore, if only compound class-specific information was lost, search performance should further decrease, which is what we observed for ICF fingerprints. The general loss in information content of fragment fingerprints should also apply to IF. However, in this case, search performance was high and often exceeded the performance of the original ligands. This is attributed to the fact that IF contain only the most significant parts of ligands, which increases their specificity in fingerprint similarity searching.

The observed effects were systematic and indicated that IF implicitly captured compound class-specific interaction information and formed 2D pharmacophore elements. Our observation that IF generally produced higher recall in *n*NN search calculations than ligands or control fragments further suggests that chemical features shared by interacting fragments of different ligands were decisive for the increased ability of IF to detect active compounds.

## CONCLUSIONS

In this study, we have systematically analyzed ligands that are available in complex crystal structures and the interacting fragments isolated from them. Calculations of standard structural key-type fingerprints provided a common reference frame for the evaluation of ligands and fragments in similarity search calculations. Fingerprints of interacting fragments implicitly capture protein−ligand interaction information without the need to encode interactions directly. Control fragments that are equal in size to interacting fragments that were generated by random deletion of ligand atoms or deletion of atoms involved in protein−ligand interactions largely lost their ability to detect active compounds in similarity searching. By contrast, interacting fragments recovered, in most cases, more active compounds than entire ligands. A general strength of the approach is that interacting fragment information is exclusively derived from experimental structures and, thus, is not vulnerable to modeling errors or false-positive activity assignments. We have shown that 8−10 diverse crystallographic ligands were sufficient to improve fingerprint search performance on the basis of IF information. Hence, for structurally similar ligands with highly conserved interaction patterns, fewer molecules will be sufficient. Taken together, our results suggest that interacting fragments capture much compound class-specific information and that the similarity search performance of conventional structural fingerprints can be further increased using interacting fragments as templates.

**Supporting Information Available:** Figure S1 shows 2D structures of all crystallographic ligands, and Figure S2 shows exemplary IF, RF, and ICF sets. Table S1 reports all PDB entries, and Table S2 gives an analysis of atom counts and structural keys for all of the crystallographic ligands. This information is available free of charge via the Internet at http://pubs.acs.org.

## ACKNOWLEDGMENT

L.T. is supported by a fellowship of the Graduiertenkolleg (GRK) 804 of the Deutsche Forschungsgemeinschaft. The authors thank Hanna Geppert and Martin Vogt for helpful discussions.

## REFERENCES AND NOTES

(1) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(2) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.
(3) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(4) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.
(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
(6) Wei, D. Q.; Zhang, R.; Du, Q. S.; Gao, W. N.; Li, Y.; Gao, H.; Wang, S. Q.; Zhang, X.; Li, A. X.; Sirois, S.; Chou, K. C. Anti-SARS drug screening by molecular docking. *Amino Acids* **2006**, *31*, 73–80.
(7) Barreiro, G.; Guimarães, C. R.; Tubert-Brohman, I.; Lyons, T. M.; Tirado-Rives, J.; Jorgensen, W. L. Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring. *J. Chem. Inf. Model.* **2007**, *47*, 2416–2428.
(8) Tikhonova, I. G.; Sum, C. S.; Neumann, S.; Engel, S.; Raaka, B. M.; Costanzi, S.; Gershengorn, M. C. Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *J. Med. Chem.* **2008**, *51*, 625–633.
(9) Lin, T. W.; Melgar, M. M.; Kurth, D.; Swamidass, S. J.; Purdon, J.; Tseng, T.; Gago, G.; Baldi, P.; Gramajo, H.; Tsai, S. C. Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of mycobacterium tuberculosis. *Proc. Natl. Acad. Sci., U.S.A.* **2006**, *103*, 3072–3077.
(10) Tan, L.; Geppert, H.; Sisay, M. T.; Gütschow, M.; Bajorath, J. Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **2008**, *3*, 1566–1571.
(11) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(12) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
(13) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-based design of target-focused libraries using protein−ligand interaction constraints. *J. Med. Chem.* **2006**, *49*, 490–500.
(14) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.
(15) 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.
(16) *Molecular Operating Environment (MOE), Version 2007.09*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2007.
(17) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.
(18) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
(19) *Molecular Drug Data Report (MDDR), Version 2005.2*; Symyx Software: San Ramon, CA, 2005.

CI800322Y