Rationalization of the Performance and Target Dependence of Similarity Searching Incorporating Protein−Ligand Interaction Information

Lu Tan,[†] José Batista,[†,‡] 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 and In-Silico Center, JADO Technologies GmbH, Tatzberg 47-51, D-01307 Dresden, Germany

Received March 29, 2010

Interacting fragments (IFs) derived from protein−ligand complex crystal structures have previously been utilized to complement conventional two-dimensional (2D) similarity searching. In many instances, the (indirect) incorporation of three-dimensional (3D) interaction information through the use of IFs has further improved the search performance of fingerprints of unmodified ligands. However, for a number of targets, changes in the relative performance of conventional fingerprints and IF-based representations have also been observed, and reasons for these effects have thus far remained elusive. Herein, we analyze target−ligand systems that display different similarity search phenotypes. We study protein−ligand interactions and the resulting IF information at the molecular level of detail in order to better understand systematic differences in the search performance of unmodified ligands and IFs. The results show that the degree of conservation of IFs isolated from multiple crystallographic ligands is a major determinant of similarity search performance, regardless of whether IFs consist of coherent substructures or disjoint fragments. Conserved IFs focus similarity search calculations on 2D pharmacophore elements, as revealed by 3D interaction analysis. This leads to increased hit rates compared to unmodified reference ligand representations and to the recognition of smaller and less complex, yet structurally diverse hits. On the basis of these findings, one can predict for which targets the inclusion of IF information will likely result in improved 2D similarity search performance.

INTRODUCTION

Depending on the availability of ligand or protein structural information, ligand-[1,2] or structure-based[2,3] virtual screening methods are applicable to search for new active compounds. In cases where structures of ligand−target complexes are available, protein−ligand interaction information can be exploited in different ways to aid in virtual screening. Recently, the repertoire of methods that utilize protein−ligand interaction has been expanded with new approaches. A number of methods that directly or indirectly account for protein−ligand interaction have been developed and used for a wide range of applications.[4−8] For example, "structural interaction fingerprints"[4] (SIFt) and "expanded interaction fingerprints"[5] are bit string representations derived from protein−ligand interactions and applied to search for active compounds and to analyze predicted ligand binding modes. Other methodologies that capture protein−ligand interaction information include the "interaction-based accuracy classification"[6] (IBAC) that is primarily used to assess docking poses and the "fingerprint for ligands and proteins"[7] (FLAP) that is applied in pharmacophore-based search calculations. Furthermore, the "interaction annotated structural features"[8] (IASF) method is designed to combine fingerprint features with energy-based atom scores computed from protein−ligand complexes. Annotated feature sets are then utilized for similarity searching. Despite differences in their design, protein−ligand interaction fingerprints have in common that they directly encode interactions obtained from structural data.

Departing from this theme, we have recently introduced a conceptually different approach to indirectly utilize three-dimensional (3D) protein−ligand interaction information in 2D similarity searching that transforms interaction information into interacting fragments (IFs) of ligands.[9−11] IFs are generated from complex crystal structures by retaining ligand atoms that engage in well-defined interactions with the target protein and by omitting the remaining atoms. Hence, IFs can be rationalized as prioritized substructures isolated from ligands that represent their strongly interacting parts. IF information has originally been utilized in 2D similarity searching by calculating MACCS key[12] representations for interacting fragments. The resulting so-called IF-FPs have been explored in different ways including similarity searching instead of fingerprints of unmodified ligands,[9] transfer of IF information to ligands of related target proteins,[10] and scaling of IF-FPs based on multiple crystallographic reference molecules.[11] Feature frequency weights derived from IF-FPs were also utilized to scale MACCS fingerprints of active reference compounds for which no structural information was available. In many applications, IF-FPs further improved the performance of conventional fingerprint search calculations, indicating that IFs captured compound class-specific information.

Transforming IFs into MACCS key representations also had intrinsic limitations. For example, additional bits not present in MACCS fingerprints of unmodified ligands were frequently generated from disjoint IFs (i.e., IFs consisting

* Corresponding author. Telephone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.
[†] Rheinische Friedrich-Wilhelms-Universität.
[‡] JADO Technologies GmbH.

10.1021/ci1001197
Published on Web 05/14/2010

Table 1. Target Enzymes and Inhibitors[a]

no.	target	X-ray structures with inhibitors	actives
1	DP dipeptidyl peptidase IV	2QT9, 2QTB, 3C43, 3C45, 3D4L, 1N1M, 2BUB, 2FJP, 2OPH, 2P8S	135
2	XA factor Xa	1EZQ, 1F0S, 1FAX, 1FJS, 1KSN, 1KYE, 1NFY, 1X7A, 1Z6E, 2FZZ	393
3	HM HMG−CoA reductase	2Q6C, 2R4F, 3BGL, 3CCT, 3CCW, 3CCZ, 3CD0, 3CD7, 3CDA, 3CDB	246
4	P4 phosphodiesterase IV	1MKD, 1XLX, 1XLZ, 1XM4, 1XM6, 1XMU, 1XMY, 1XN0, 1XOQ, 2FM0	640
5	AR aldose reductase	1US0, 1Z3N, 2IKG, 2INE, 2INZ, 2IQ0, 2IQD, 2J8T, 2PDH	265
6	RT HIV-1 reverse transcriptase	1COT, 1DTT, 1EET, 1FK9, 1HNI, 1IKX, 1LW2, 1S6Q, 1S9E, 1SUQ, 1SV5, 1VRT	321
7	EL elastase	1B0F, 1BMA, 1BTU, 1E34, 1E36, 1EAS,1EAT, 1EAU, 1ELD, 1ELE, 1FZZ, 1H1B, 1HV7, 1INC, 2CV3, 2V35	179

[a] Target enzymes are listed (and abbreviated) and the PDB codes of enzyme−inhibitor complexes that were analyzed are provided. "actives" reports the number of inhibitors of each enzyme that were selected from the MDDR for similarity search applications.

of multiple substructures). Furthermore, in cases where MACCS search performance was intrinsically low, MACCS-based IF encoding improved the search performance but did not yield high hit rates. Therefore, we have also developed a flexible, rather than dictionary-based, representation of IFs on the basis of atom-centered fragmentation.[13] Following this approach, atom-centered fragments were calculated for unmodified ligands (termed AFs), and IFs derived from them (termed atom-centered interacting fragments or AIFs). The calculation of AFs and AIFs produces compound-specific feature sets of varying size that can also be utilized in 2D similarity searching. For this purpose, database compounds are screened for AF or AIF features and matching features are counted to produce a database ranking relative to a set of crystallographic reference compounds. We found that the AIF representations often performed better in similarity searching than AF sets of unmodified ligands.[13] Because AIF feature sets are smaller than corresponding AF sets, these findings suggested that encoding of interacting fragment information focused similarity searching on compound class-characteristic features. We also found that AIF representations usually achieved a higher compound recall than MACCS-based IF-FPs and other standard 2D fingerprints.[13] In addition, an early enrichment of active compounds in database selection sets has been a characteristic feature of AIF-based similarity searching.

The strong compound class dependence of fingerprints and other similarity search methods is a well-known, and largely unpredictable, phenomenon in ligand-based virtual screening.[1,2] Depending on the compound activity class, a given fingerprint might recognize many active compounds or fail. This compound class dependence essentially affects all similarity search approaches and also applies to AF and AIF representations. As we have applied the IF approach to an increasing number of targets, we have also observed other compound-class specific effects, i.e., notable differences in the relative search performance when using either unmodified ligands or IFs as references. In many instances, IF representations clearly outperformed unmodified reference compounds, whereas in others, essentially opposite observations were made. Reasons for these frequently observed differences in relative search performance were not apparent. Hence, we have set out to systematically study these effects by attempting to analyze similarity search performance for selected compound classes and targets in light of X-ray structural data and IF characteristics. The results of this analysis are presented herein. For all test cases studied here, it has been possible to provide a structural rationale as to why IF-based calculations might succeed or fail. These findings make it possible to predict on the basis of IF information if 3D interaction augmented similarity searching will be an attractive alternative to conventional 2D similarity search calculations.

MATERIALS AND METHODS

Protein−Ligand Complexes and Interacting Fragments. Multiple inhibitor complexes for seven enzymes were taken from the Protein Data Bank[14] (PDB), as summarized in Table 1. Target−ligand interactions were analyzed using the molecular operating environment[15] (MOE). Hydrogen-bonding, ionic, and van der Waals interactions with cutoff distances of 3.2, 4.0, and 3.8 Å, respectively, were calculated. IFs were generated by retaining ligand atoms involved in these intermolecular interactions and omitting the noninteracting atoms. As shown in Table 1, between nine and 16 crystallographic ligands were available per target and utilized as reference compounds for IF generation and similarity searching.

Fingerprint Similarity Searching. MACCS structural keys[12] (166 bit positions) were calculated for the original (unmodified) ligands and the corresponding IFs. As a similarity measure for fingerprint comparison, the Tanimoto coefficient[16] (Tc) was calculated. The k (k = number of crystallographic inhibitors) nearest-neighbor (kNN) similarity search strategy was applied that averages the Tc values of all k reference molecules to produce the final similarity value for a database compound. Hit rates were recorded for the 50 top-scoring database compounds.

Atom-Centered Fragmentation and Similarity Searching. As an independent encoding for similarity searching, atom-centered fragments[13] were calculated for both unmodified ligands and interacting fragments. On the basis of the molecular graph, each ligand atom was once considered as a central atom, and the surrounding atoms up to a bond distance of 2 (i.e., one or two bonds away from the central atom) were determined and combined, yielding an atom-centered fragment. In these fragments, atom- and bond-type

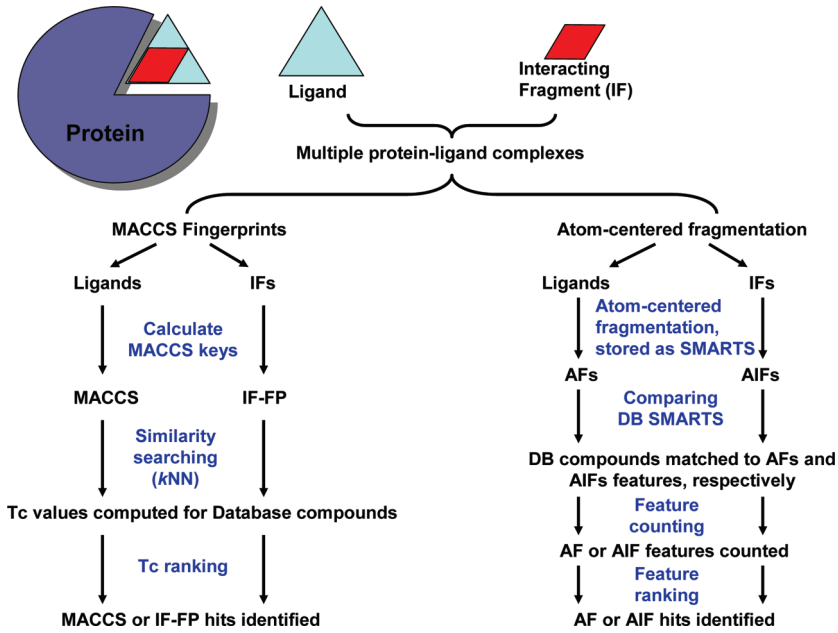

Figure 1. Method summary. An overview of the alternative similarity search strategies is provided. Complete ligands or their IFs are collected from multiple complex crystal structures and represented as MACCS keys or atom-centered fragments. For MACCS fingerprint similarity searching, Tanimoto similarity between database compounds and ligand reference sets is calculated. For AF and AIF representations, features common to a database compound and reference set are counted to generate a ranking.

information was retained. For all complete ligands active against a given target (i.e., the crystallographic reference set) and the corresponding interacting fragments, the union of all unique AFs or AIFs was generated, respectively, and stored as SMARTS[17] strings. For each screening database compound, its SMARTS string was calculated and compared to the AF and AIF ensembles, and matching features were counted. These feature counts were used to rank database compounds in the order of decreasing similarity to reference sets, and hit rates for the top-scoring 50 database compounds were determined. Figure 1 illustrates the MACCS key and atom-centered fragment similarity search strategies for unmodified ligands and their interacting fragments.

Database Compounds. As database compounds for similarity searching, 100 000 molecules were randomly selected from ZINC.[18] For each of our seven enzyme targets, known inhibitors with pairwise MACCS Tc values of ≤ 0.80 (i.e., to avoid the inclusion of very similar analogs) were selected from the MDL/Symyx Drug Data Report[19] (MDDR) and added to the screening database as potential hits. Depending on the target, between 135 and 640 inhibitors were obtained, as summarized in Table 1.

RESULTS AND DISCUSSION

Interacting Fragment-Based Similarity Searching. The IF approach was originally introduced to indirectly incorporate 3D protein−ligand interaction information into 2D similarity searching, i.e., without the need to directly encode such interactions. This made it possible to bridge between 2D and 3D search methods and to enrich 2D similarity searching with available interaction information. Proof-of-principle was established by using conventional MACCS key representations for interacting fragments and comparing their search performance with fingerprints of unmodified ligands. In these calculations, it was shown that the search performance of standard 2D molecular representations could indeed be increased when the search was focused on interacting parts of ligands, and others were omitted. Because IF information can be transferred to noncrystallographic ligands through frequency of occurrence based fingerprint scaling techniques, IF-based search calculations are not limited to the use of crystallographic reference compounds. With the AIF encoding, a more flexible and molecule-specific representation of interacting fragments was then introduced. Although AIF representations often further increased compound recall of MACCS, AF, and (MACCS-based) IF-FP search calculations, this was not always the case, dependent on the target. In some instances, the use of interacting fragment information reduced hit rates observed for unmodified reference molecules. Hence, what remained to be investigated were reasons for relative differences in search performance between unmodified ligands and IFs. Thus, we set out to relate search performance to structural details of protein−ligand interactions and the ensuing IF information. Although 2D similarity searching is applicable when only a single reference structure is available, we have deliberately focused on search calculations based on multiple reference compounds because these calculations are richer in ligand information than single reference searches, which typically results in a performance increase.

Targets. For our analysis, we selected seven target enzymes (Table 1) on the basis of three criteria: (i) availability of multiple well-refined enzyme−inhibitor complex structures, (ii) availability of more than 100 known inhibitors for similarity searching (Table 1), and (iii) apparent performance differences in conventional 2D fingerprint, atom-centered fragment, and interacting fragment-based similarity searching. Therefore, search calculations were initially carried out using unmodified crystallographic ligands or interacting fragments as reference compounds, and the results were compared.

Similarity Search Strategies. As illustrated in Figure 1, two different similarity search strategies were applied,

Table 2. Similarity Search Results[a]

		fingerprint		atom-centered fragments	
no.	target	MACCS	IF-FP	AF	AIF
1	DP	16	28	22	48
2	XA	18	32	38	42
3	HM	18	44	8	83
4	P4	46	12	90	77
5	AR	6	0	17	36
6	RT	4	2	20	18
7	EL	18	0	17	30

[a] For each target enzyme, hit rates are reported (in % for the 50 top-scoring database compounds) for representations of unmodified inhibitors (MACCS, AF) and interacting fragments (IF-FP, AIF).

reflecting the evolution of the IF approach. First, predefined MACCS structural keys were calculated for both reference molecules (ligands and IFs, respectively) and database compounds. Standard similarity searching was then carried out with MACCS and IF-FPs, applying the *k*NN search strategy for multiple reference compounds. Following this approach, Tc values for comparison with individual reference molecules are averaged to produce the final similarity score for a database compound, which was utilized for database ranking. Hence, contributions from all reference molecules are equally taken into account. Second, atom-centered fragments were calculated for reference molecules (ligands and IFs, respectively) and compared to database compounds. Atom-centered fragments matching the unique AF sets (unmodified ligands) and AIF sets (interacting fragments) were counted and provided the basis for database ranking. Systematic search calculations were carried out applying these alternative strategies, and the results were compared.

Similarity Search Performance. The search results are summarized in Table 2. In part, significant target-dependent variation of the performance of the applied molecular representations and search strategies were observed. For three

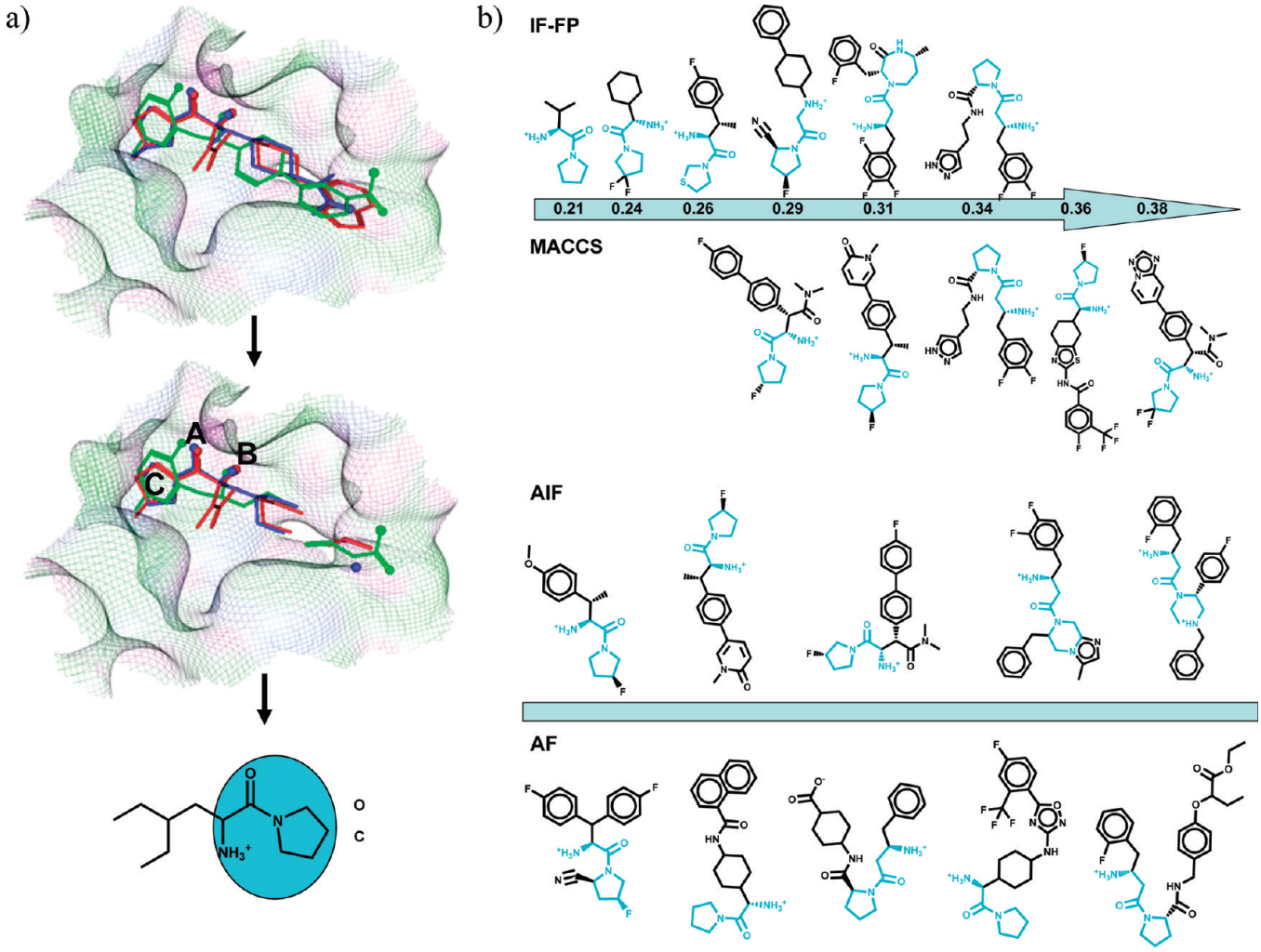

Figure 2. Dipeptidyl peptidase IV: Enzyme−inhibitor interactions, interacting fragments, and exemplary hits. (a) Active site with bound inhibitors (top), corresponding interacting fragments (middle), and a consensus fragment (bottom). Three representative inhibitors are colored red, green, and blue, respectively. Atoms involved in hydrogen-bond (H-bond) and ionic interactions are depicted as balls. Major interactions include: A, H-bonds; B, H-bonds and salt bridges; and C, van der Waals/hydrophobic interactions. The 2D structure of an exemplary IF is shown with a substructure conserved among inhibitors highlighted in cyan. The "O" and "C" atoms in the 2D depiction represent individual atoms resulting from IFs generation. Single atoms are occasionally also obtained for other targets but are usually not conserved in the IFs of different inhibitors. (b) Exemplary MDDR hits identified by similarity searching. The figure focuses on the structures of exemplary hits. The number of compounds shown depends on different molecular complexity levels or molecular weight ranges. Hits detected with MACCS-based ligand and IF fingerprints (IF-FP) are organized according to the fingerprint bit density they produce. In addition, AF and AIF hits are arranged according to increasing molecular weight. The core structures highlighted in cyan are identical or similar to the conserved IF moiety shown in (a). The same representation scheme as in this figure is utilized in Figures 3−8.

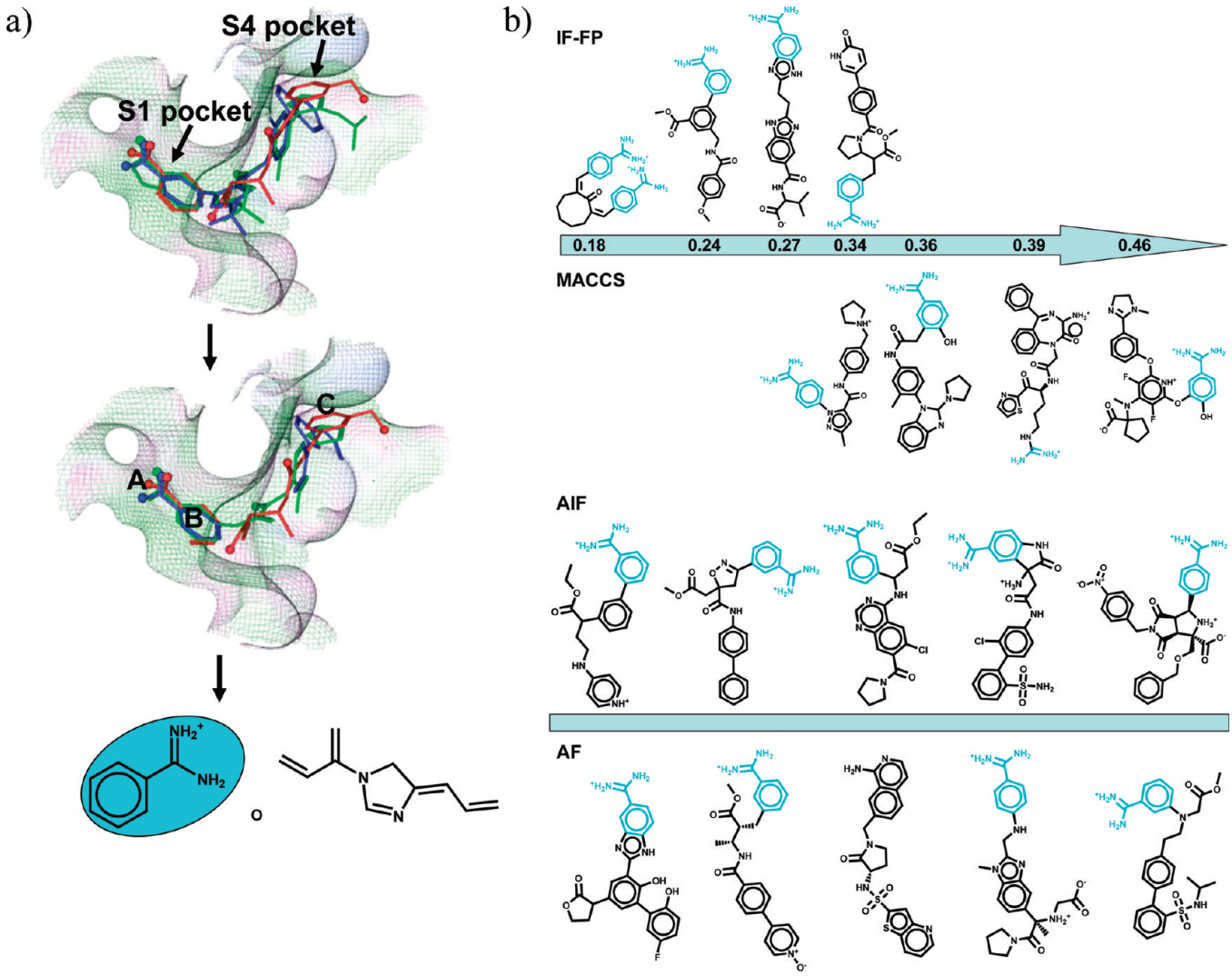

Figure 3. Factor Xa. (a) Dominant interactions include: A, H-bonds and salt bridges; B, hydrophobic; and C, highly conserved aromatic interactions. Ring structures isolated from region C are not conserved and thus not highlighted in (b) that shows representative hits with mapped IFs.

target enzymes, DP, XA and HM, IF-based search calculations produced much higher hit rates than whole-molecule-based calculations, for both MACCS and AF/AIF representations. IF-FPs increased the hit rates of MACCS keys by more than 10 or 20%, depending on the target, and AIF-based calculations further increased search performance by nearly doubling the IF-FP hit rates in two cases (DP and HM). Hit rates between 42 and 83% were observed for the 50 top-scoring database compounds. For HM, a striking increase in hit rates was observed from 8 for AF to 83% for AIF calculations.

For target P4, AF and AIF also produced very high hit rates of 90 and 77%, respectively, but in this case, search calculations utilizing unmodified ligands produced better results than IFs for both molecular representations, in contrast to the three targets discussed above. Similar observations were made for RT, where IFs produced slightly lower hit rates than ligands. In this case, however, MACCS representations essentially failed, with hits rates of only 4 and 2% for MACCS and IF-FPs, respectively. Only atom-centered fragments produced notable compound recall with hit rates of 20 and 18% for AF and AIF, respectively.

For the two remaining targets, AR and EL, similar observations were made. In these cases, MACCS hit rates were also only low (6 and 18%, respectively), but IF-FPs failed to detect any active compounds. However, the AF representation produced hit rates of 17% in both cases, and AIF further increased hit rates to 30 (EL) and 36% (AR). Thus, for these two target−ligand systems, only the AF encoding of IFs was successful, and AIF provided a clear advantage compared to MACCS and AF representations.

On the basis of these similarity search results, target enzyme−inhibitor systems could be divided into three categories:

(1) IF representations performed consistently better than ligands: DP, XA, and HM.

(2) Ligand representations performed better than IFs: P4 and RT.

(3) IFs performed better than ligands but only when atom-centered fragments were calculated (for MACCS, search performance was low or searches failed): AR and EL.

Thus, these targets provide examples for cases where interacting fragments either clearly improve standard ligand-based similarity search performance or fail to do so. Furthermore, they also illustrate differences in the search performance of alternative molecular representations and reveal an advantage of encoding IF information through atom-centered fragments, consistent with our earlier observa-

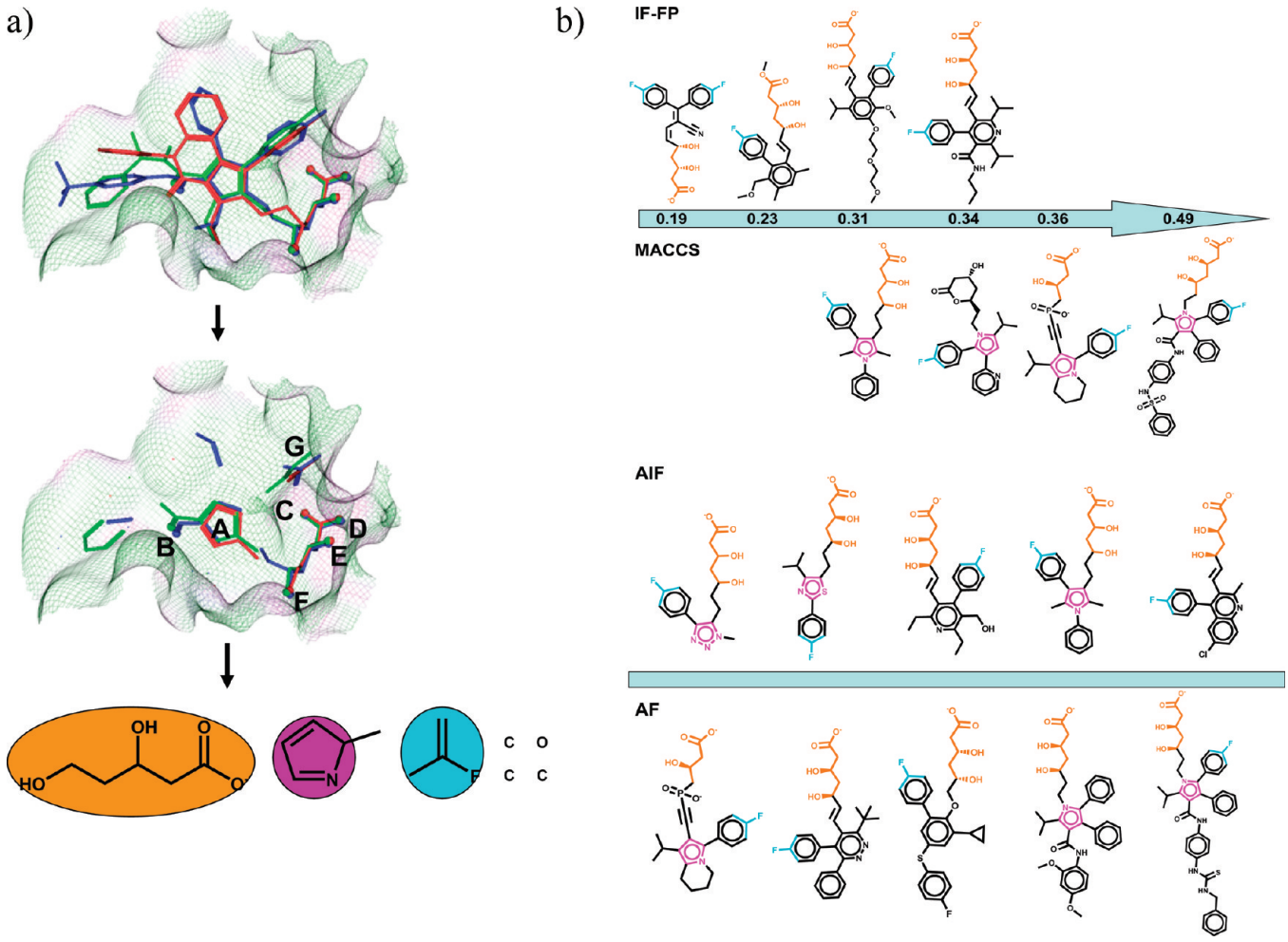

Figure 4. HMG-CoA reductase. (a) Important interactions include: A, hydrophobic interactions; B−F, an array of H-bonds and salt bridges; and G, hydrophobic interactions. Besides the highlighted fragments, four single atoms are shown that are part of the IF of the displayed inhibitor. In this case, the inhibitors produce a disjoint IF containing three conserved moieties that are color coded and mapped on similarity search hits in (b).

tions. We next analyzed structural details of the underlying enzyme−inhibitor interactions.

Details of Enzyme−Inhibitor Interactions and Similarity Search Hits. Figures 2a−8a show active site views of inhibitor complexes for each target and IFs derived from them, and Figures 2b−8b compare active compounds detected in similarity searching using different molecular representations and map interacting fragments on hits.

Dipeptidyl Peptidase IV. Figure 2a shows an overlay of three DP−inhibitor complexes. Strong DP−inhibitor interactions[20−22] included hydrogen bonds (H-bonds) with Asn710 and Arg125 (region A in Figure 2a), multiple ionic and H-bond interactions with Glu205, Glu206, and Tyr662 (B), and van der Waals/hydrophobic interactions involving ring structures (C). These interactions were conserved in different inhibitor complexes. The resulting IFs were coherent and contained a conserved core structure (Figure 2a, bottom). Hits identified in similarity search calculations also contained this core structure, as shown in Figure 2b, regardless of whether IFs or complete ligands were used as reference molecules and MACCS or atom-centered fragment representations. We also compared the fingerprint bit density of hits identified with MACCS and IF-FPs. Figures 2−8 show structures of exemplary hit molecules from the corresponding search calculations. The number of compounds that are shown in each case does not correspond to relative hit rates in Table 2. As shown in Figure 2b, MACCS fingerprints of unmodified ligands detected hits of higher bit density than IF-FPs, i.e., larger and more complex active compounds. By contrast, IF-FPs that produced much higher hit rates than MACCS preferentially detected smaller hits. Equivalent observations were made when hits identified with AFs and AIFs were compared. Hits consistently contained the conserved IF core but were otherwise structurally diverse. Taken together, these findings indicated that IF-based similarity searching focused the calculations on the IF consensus core of DP inhibitors, more so than search calculations based on complete ligands, leading to an increased search performance for both MACCS and atom-centered fragment representations.

Factor Xa. XA displayed the same similarity search phenotype as DP, i.e., encodings of interacting fragments performed consistently better than unmodified ligands. In Figure 3a, three inhibitor structures are superposed. Major enzyme−inhibitor interactions in the XA active site occurred in the S1 and S4 subsites.[23−25] In the S1 specificity pocket, inhibitors formed multiple salt bridges and H-bonds with residues Asp189 and Glu219 (region A in Figure 3a). In addition, strong van der Waals or aromatic interactions were observed in regions B and C involving three aromatic

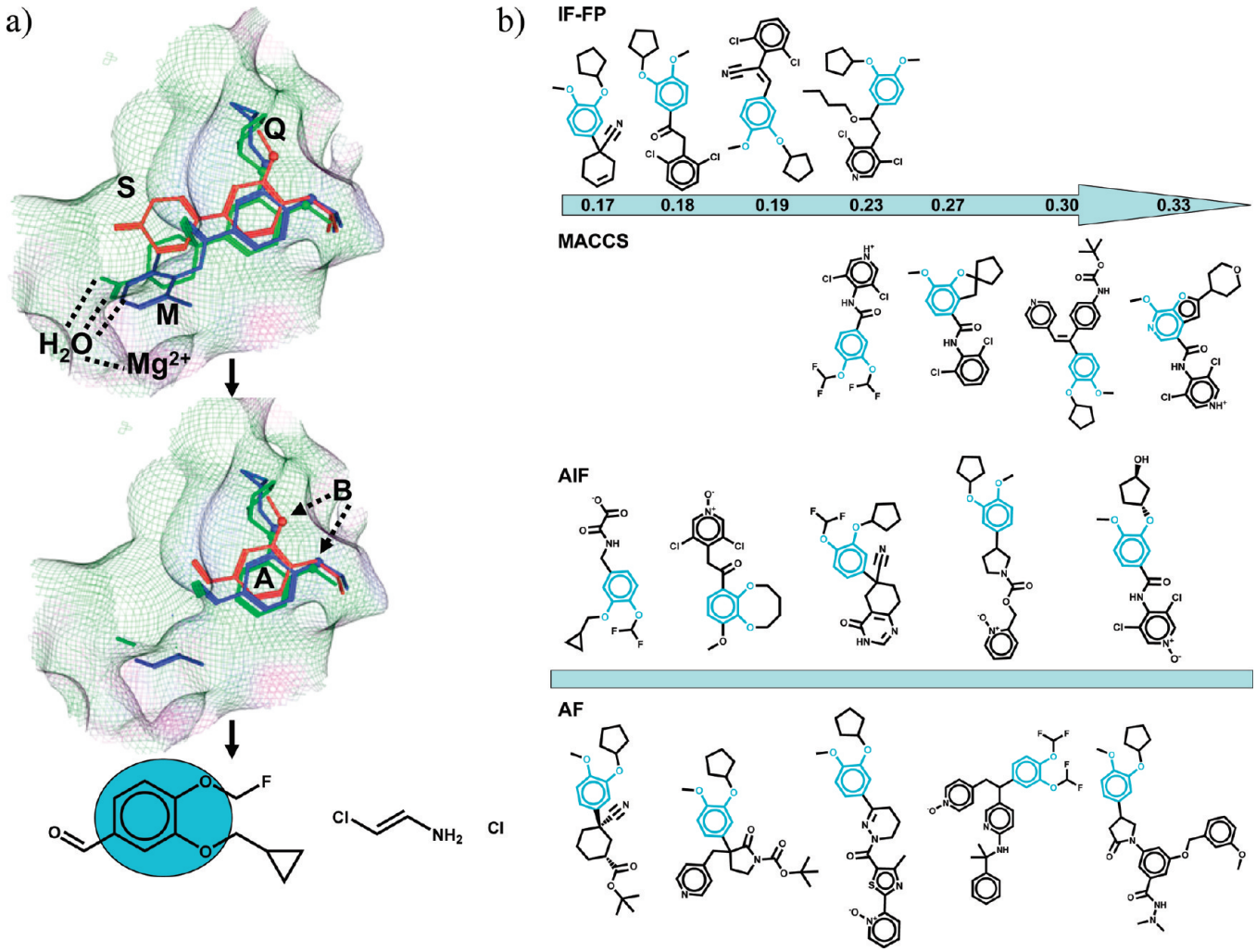

Figure 5. Phosphodiesterase IV. (a) The active site contains Q-, S-, and M-binding regions (top). Key interactions include: A, hydrophobic interactions and B, H-bonds (middle). Binding to the M site involves water-mediated complexation of a magnesium cation. Because only direct ligand−protein contacts are taken into account when determining interacting fragments, ligand atoms involved in complexing the cation are not contained in the IF (bottom). Only the conserved IF part (cyan) is highlighted in the hits in (b).

residues within the S4 pocket (Tyr99, Phe174, and Trp215). IFs isolated from bound inhibitors were primarily involved in conserved interactions in the S1 and S4 pockets. In this case, the IFs were disjoint and consisted of several parts, one of which, a benzamidine group occupying the S1 pocket, was conserved in all inhibitors and, hence, represented a consensus fragment (Figure 3a, bottom). By contrast, groups occupying the S4 pockets were not rigorously conserved. As shown in Figure 3b, the benzamidine consensus fragment also occurred in almost all active compounds detected in similarity search calculations using MACCS and atom-centered fragment representations. As observed for DP, the IF-FP also detected active compounds with lower bit density than MACCS in this case and achieved much higher hit rates (Table 2).

HMG−CoA Reductase. HM has been the target for which IF encodings led to the largest improvements in similarity search performance compared to ligands, with a MACCS hit rate of 18% increasing to an IF-FP rate of 44% and an AF hit rate of 8% increasing to 83% for AIF. Figure 4a shows an HM active site view with superposed inhibitors. Here enzyme−inhibitor interactions were highly conserved, including van der Waals/hydrophobic interactions in regions A and G and an array of H-bond and ionic interactions in subsites B to F.[26−28] HM inhibitors were found to place a conserved pyrrole group into subsite A and to display conserved H-bond interactions throughout the C−F sites. The resulting HM IFs were disjoint, similar to XA, but contained three consensus fragments, highlighted at the bottom of Figure 4a. The consensus fragments were found in most of the hits detected in search calculations using both unmodified ligands and IFs, as shown in Figure 4b. In HM search calculations, IF-FP and AIF produced much higher hit rates than their ligand-based counterparts and detected smaller and less complex active compounds, similar to the observations made for DP. Nevertheless, hits containing consensus fragments were also structurally diverse (Figure 4b). Thus, in the case of HM, IF representations clearly focused similarity searching on consensus fragments contained in otherwise structurally diverse inhibitors, which led to very significant increases in hit rates.

Phosphodiesterase IV. Inhibitor binding to P4 also involved largely conserved interactions.[29,30] Figure 5a reveals very similar binding modes of different inhibitors. The P4 active site can be divided into three subsites[30] including a glutamine-containing Q site, a largely hydrophilic and the solvent-exposed S site, and a magnesium cation-containing M site (Figure 5a, top). Strong interactions were formed in

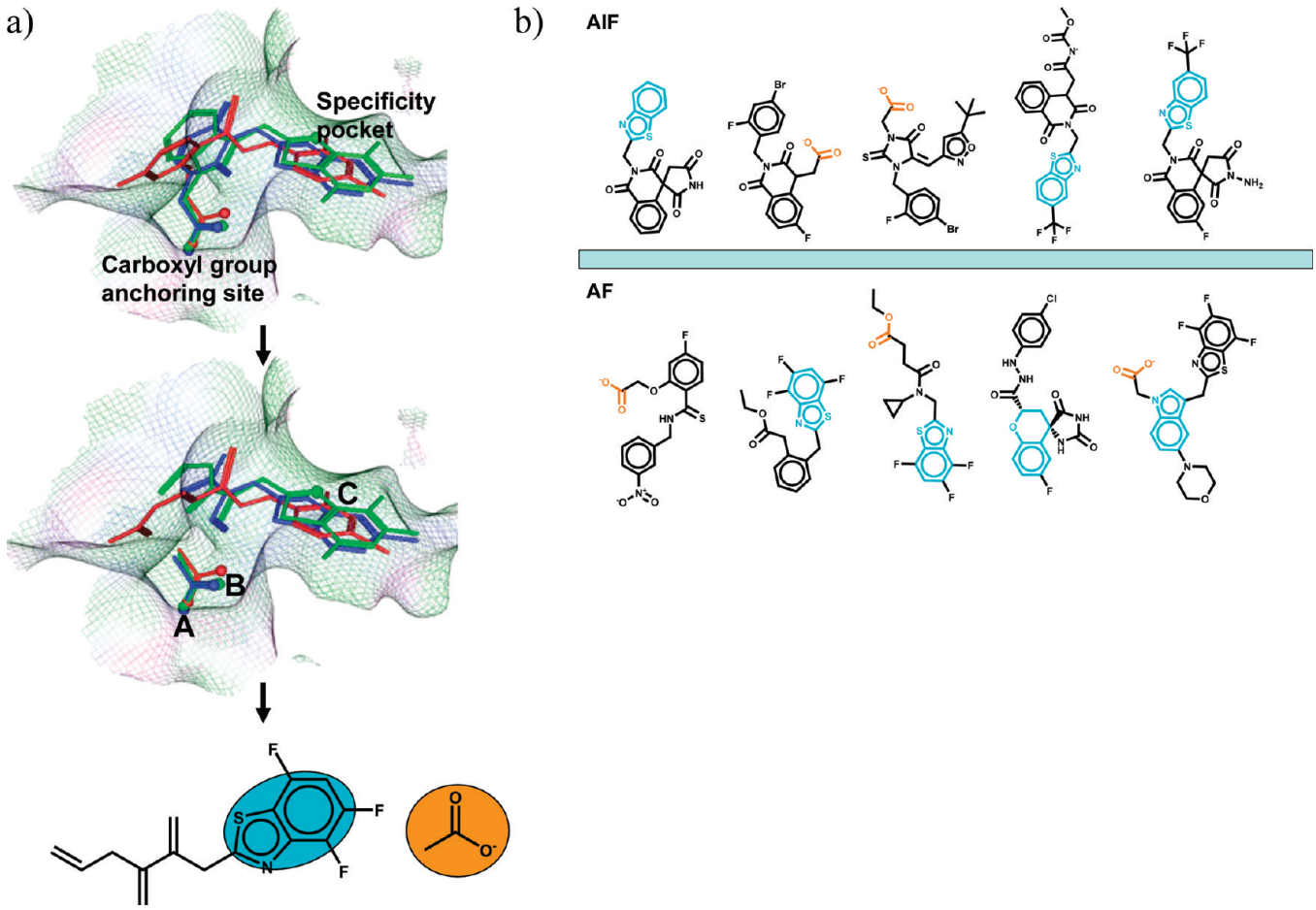

Figure 6. Aldose reductase. (a) The active site is divided into a carboxyl group anchoring pocket and a specificity pocket. Dominant interactions include: A and B, H-bonds and C, hydrophobic interactions. In (b), no hits identified with MACCS or IF-FP are shown because search calculations using MACCS-based representations produced only very low hit rates or no hits.

the Q site that is further divided into two regions A and B (Figure 5a, middle). In the A site, residues Phe446 and Ile410 formed a hydrophobic clamp harboring an aromatic ring of bound inhibitors. In B, conserved H-bonds were formed between ring substituents and the invariant residue Gln443. These conserved hydrophobic and H-bond interactions in the Q site resulted in a conserved IF substructure, a 1,2-dihydroxybenzene moiety, highlighted at the bottom of Figure 5a. The 1,2-dihydroxybenzene was the only IF consensus portion and was consistently found in similarity search hits, as shown in Figure 5b. However, in contrast to DP, XA, and HM, hit rates for P4 were higher for complete ligands than for IFs. How can one rationalize these findings? As shown in Figure 5a (top), inhibitor binding to P4 also involved conserved interactions in the M site, i.e., water-mediated H-bonds to an active site magnesium cation. Complexation of this cation represented another largely conserved interaction that is known to be a hallmark of potent P4 inhibitors.[30] However, this Mg^{2+} interaction was not accounted for by IFs (Figure 5a, middle) because IFs were derived only on the basis of direct protein−ligand contacts, not water- and/or ion-mediated interactions. Consequently, in the case of P4, the IF only provided an incomplete ensemble of binding determinants, which clearly rationalizes the observed reduced search performance of IF representations. Therefore, inhibitors forming ion-mediated interactions should best not be subjected to IF calculations (which can be determined by inspecting the crystallographic data).

Aldose Reductase. Similarity searching for AR inhibitors was compromised or failed when MACCS representations were used but yielded an acceptable hit rate for AF (17%, Table 2). These representation-dependent differences in search performance were consistent with our earlier observations that the flexible and molecular-specific atom-centered fragment encoding was generally superior to MACCS-based molecular representations.[13] Importantly, in the case of AR, a clear increase in hit rate compared to AF was also observed for AIF (17 versus 36%). Figure 6a (top) shows the active site view of AR−inhibitor complexes. Major interactions were distributed over a so-called carboxyl group anchoring site and a specificity pocket.[31−33] In the A and B regions (Figure 6a, bottom) H-bonds were formed between inhibitor atoms and aromatic residues Tyr48, His110, and Trp111 (that are invariant in AR across different species). The specificity pocket (C region) was occupied by a substituted benzothiazole moiety of the inhibitors. These conserved interactions were reflected at the IF level by two consensus substructures (Figure 6a, bottom) that were typically also found in an active compound identified in AF and, in particular, AIF searches (Figure 6b).

HIV-1 Reverse Transcriptase. The RT active site is known to exhibit a high degree of plasticity and permits a variety of compound binding modes.[34−36] Figure 7a (top, I) displays two distinct binding modes of non-nucleoside inhibitors termed "butterfly" and "horseshoe"[35] that resulted in the formation of different IFs. Moreover, structurally diverse

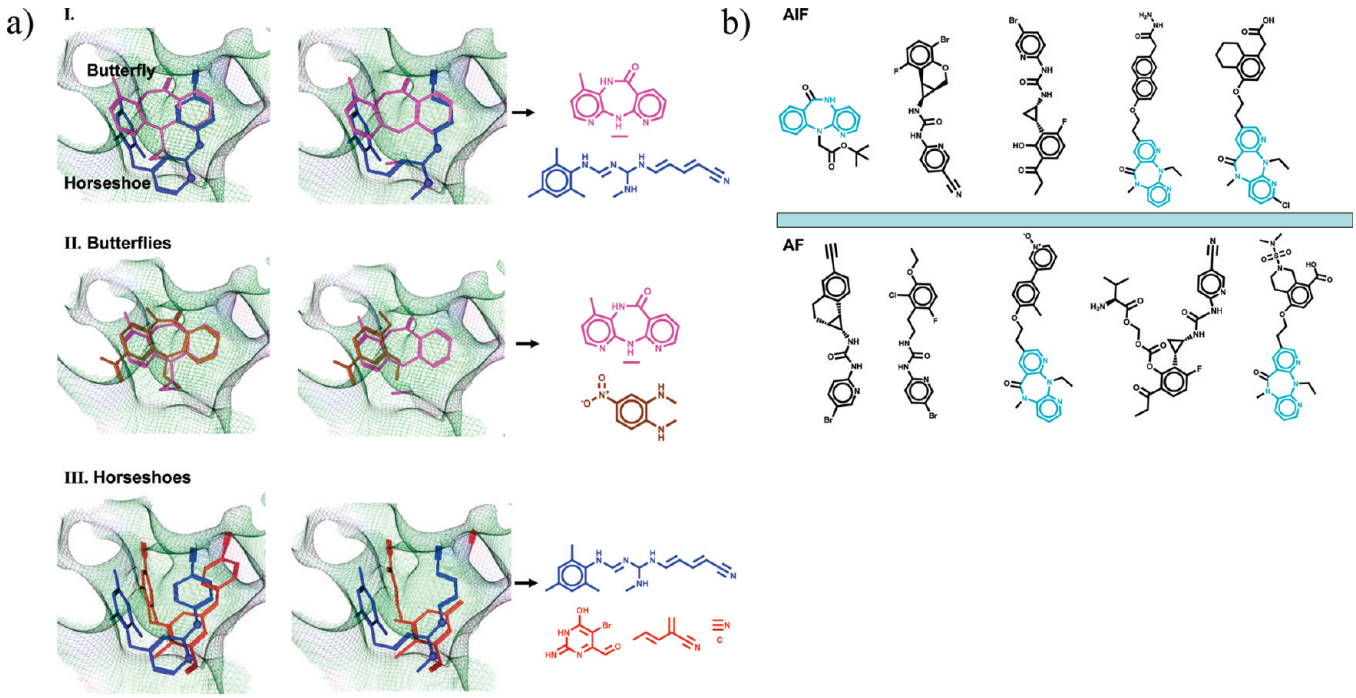

Figure 7. HIV-1 reverse transcriptase. (a) I. Two different compound binding modes termed "butterfly"(magenta) and "horseshoe" (blue) are shown on the left, yielding distinct IFs (middle, right). II. Diverse compounds adopting the butterfly binding mode also produce different IFs. III. Similar observations are made for inhibitors adopting the horseshoe binding mode. No conserved core structure is identified in RT inhibitors adopting these alternative binding modes. (b) No hits identified with MACCS or IF-FP are shown because these search calculations essentially failed. Different from Figures 2−6 where conserved substructures are color coded in both (a) and (b), in this figure (and in Figure 8), the coloring scheme of the 2D structures in (a) is applied to distinguish superimposed ligands and does not correspond to the substructure color code in (b).

active compounds could either adopt the butterfly or horseshoe binding mode, as shown in Figure 7a (middle, II and bottom, III, respectively), which again resulted in different IFs for these compounds. Hence, in general, different IFs were produced from compounds adopting the horseshoe and butterfly binding modes. Due to the plasticity of the RT active site, different interactions were found to stabilize each binding mode,[35,36] which also enabled structurally diverse compounds to bind in similar ways. As a consequence, IFs derived from different RT inhibitors displayed only little conservation.

In RT search calculations, only atom-centered fragment representations produced meaningful results (Table 2), although hit rates were generally low in this case. However, AIF performed slightly worse here than AFs, consistent with the absence of conserved IF features. Figure 7b shows hits identified with atom-centered fragment representations. Besides the occasional appearance of moieties reminiscent of "butterfly rings", these active compounds displayed no conserved substructures.

Elastase. EL is another enzyme for which inhibitors are known to adopt different binding models, regardless of their degree of structural similarity.[37−40] Figure 8a shows different compound binding modes in the EL active site. In Figure 8a (left, I), two inhibitors are shown that bound differently yet yielded, in part, very similar (but not identical) IF substructures. Furthermore, in Figure 8a (right, II), two inhibitors are displayed that bound similarly and shared many interactions, which again resulted in largely corresponding IFs. Hence, although EL inhibitors often adopt different binding modes, similar to RT, their IFs were usually rather similar (but not fully conserved), in contrast to RT inhibitors. Peptide-like IF components and IF functional groups were frequently found in EL similarity search hits (Figure 8b), although the inhibitors lacked well-defined common core structures (except peptide backbone-like components of varying size). The similarity search results corresponded to a phenotype similar to AR, i.e., MACCS and AF produced rather low hit rates (of 18 and 17%, respectively; Table 2), IF-FP calculations failed, but AIF further increased the AF hit rate to 30%. Hence, in this case, partly conserved IF information also provided a moderate advantage in similarity search calculations. However, hit rates were much lower than those observed for targets DP, XA, or HM.

Interaction Anatomy, Interacting Fragments, And Similarity Search Performance. Taken together, the results discussed above revealed a clear correlation between the similarity search performance and the conservation of IFs derived from multiple crystallographic ligands. We have found that encoding of conserved IFs consistently and significantly increased search performance of atom-centered fragments and also MACCS key representations. For targets, such as DP, XA and HM, significant increases in hit rates were observed when IF representations were utilized. Here IFs or IF subsets were strongly conserved. Moreover, in these cases, it made no difference whether IFs consisted of coherent (DP and XA) or multiple disjoint (HM) substructures. Compared to unmodified reference ligands, search calculations for these targets yielded many smaller and less complex (albeit structurally diverse) hits, which provided a clear indication that IF information focused similarity search calculations on critical parts of active compounds and thereby reduced search noise. In the presence of conserved IFs, or even only partly conserved IFs (AR, EL), the AIF representation was the method of choice for similarity searching and produced consistently the best results. By contrast, if IFs did not display compound class signature character, they

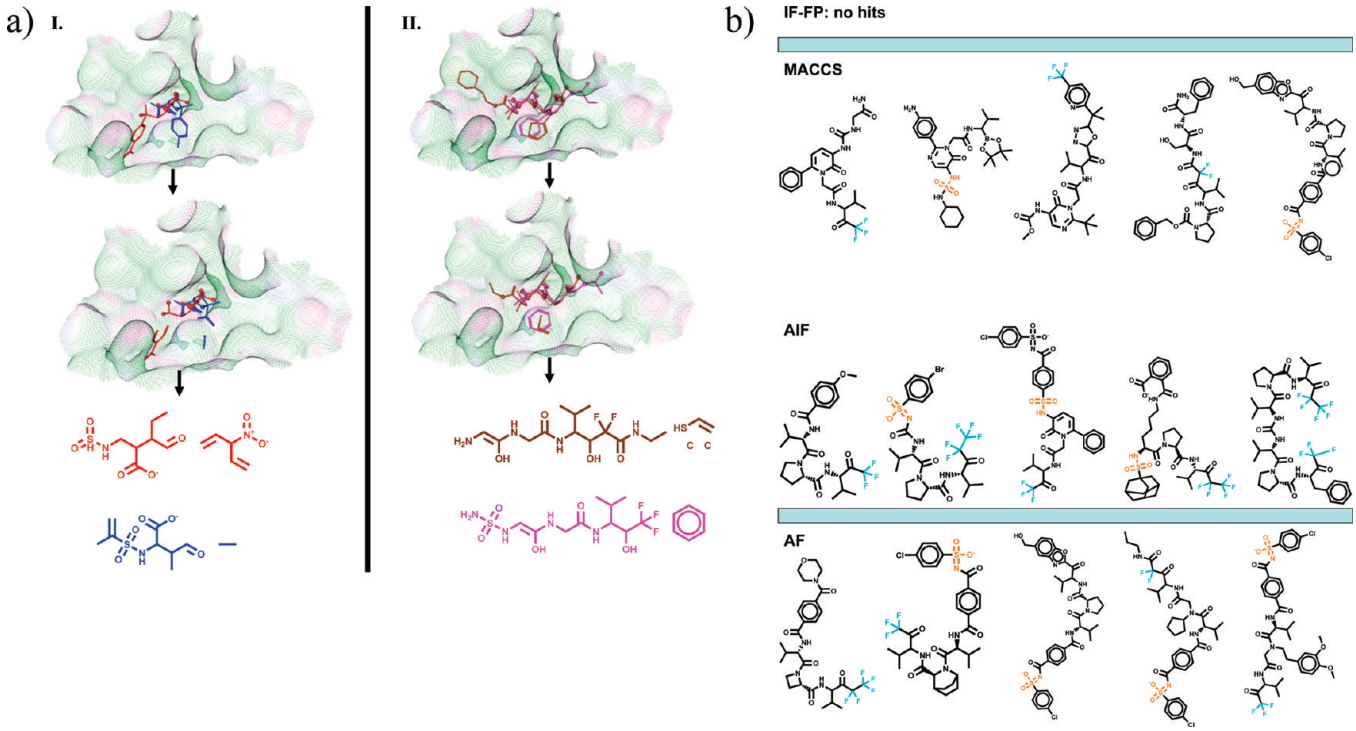

Figure 8. Elastase. (a) In I and II, inhibitors with distinct binding modes are shown. Compounds adopting each binding mode form in part variable interactions, yielding overlapping yet distinct IFs. In (b), hits obtained with AFs, AIFs, and MACCS are shown. In this case, IF-FP searching failed to retrieve compounds. The coloring scheme of the 2D structures in (a) is applied to distinguish superimposed ligands and does not correspond to the substructure color code in (b).

offered no advantage compared to intact ligands and usually resulted in reduced search performance. This was the case for P4, where IFs only incompletely accounted for critical enzyme−inhibitor interactions and also for RT, where different compound binding modes resulted in nonconserved IFs. From this point of view, the comparison of RT and EL has also been of interest. Similar to RT, EL inhibitors also adopted different binding modes but nevertheless produced similar IFs, in contrast to RT. Using these IFs also increased the search performance of unmodified EL inhibitor queries, consistent with the critical role of IF conservation observed for other targets (AR, DP, XA, and HM).

CONCLUSIONS

The Interacting fragment (IF) approach has originally been introduced to augment conventional 2D similarity searching with 3D protein−ligand interaction information, without the need to encode such interaction directly. Herein we have analyzed relative differences in the search performance of IF representations and unmodified ligands in similarity searching using multiple reference compounds. Exemplary enzyme−inhibitor systems were analyzed in detail that displayed different phenotypes in search calculations using different molecular representations. Interaction analysis was carried out in order to rationalize the derivation of IFs, and IFs obtained from different inhibitors were compared. On the basis of this analysis, we were able to establish a clear correlation between the degrees of IF conservation and similarity search performance. High IF-based search performance was consistently observed when IFs were conserved across different crystallographic ligands, leading to increasingly focused similarity search calculations. In light of these findings, conserved IFs can also be rationalized as a representation of 2D pharmacophore elements that emphasize compound class-specific activity determinants during similarity searching. This makes it also possible to predict in which search situations the application of the IF approach might be promising. IFs isolated from different inhibitor complexes can be compared, and the degree of their conservation determined, as described herein. If conserved IF elements are identified, their atom-centered interacting fragments (AIF) encoding is likely to provide a promising similarity search tool. Through feature frequency of occurrence analysis, AIF information can also be transferred from crystallographic reference sets to noncrystallographic reference compounds, hence providing different ways to exploit IFs in similarity searching. Focusing search calculations on strongly interacting and conserved parts of ligands displayed the general tendency to increase hit rates and enrich database selection sets with active compounds that were smaller and chemically less complex than many compounds, both hits and false-positives, identified using whole-molecule queries. These findings also indicate that IF-based similarity searching generally reduces the noise of search calculations.

ACKNOWLEDGMENT

L.T. is supported by a fellowship of the Graduiertenkolleg (GRK) 804 of the Deutsche Forschungsgemeinschaft.

REFERENCES AND NOTES

(1) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189–1202.
(2) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(3) 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.
(4) 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.

(5) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–1951.
(6) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Glordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871–881.
(7) 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.
(8) Crisman, T. J.; Sisay, M. T.; Bajorath, J. Ligand-target interaction-based weighing of substructures for virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964.
(9) Tan, L.; Lounkine, E.; Bajorath, J. Similarity searching using fingerprints of molecular fragments involved in protein-ligand interactions. *J. Chem. Inf. Model.* **2008**, *48*, 2308–2312.
(10) Tan, L.; Bajorath, J. Utilizing target-ligand interaction information in fingerprint searching for ligands of related targets. *Chem. Biol. Drug Des.* **2009**, *74*, 25–32.
(11) Tan, L.; Vogt, M.; Bajorath, J. Three-dimensional protein-ligand interaction scaling of two-dimensional fingerprints. *Chem. Biol. Drug Des.* **2009**, *74*, 449–456.
(12) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(13) Batista, J.; Tan, L.; Bajorath, J. Atom-centered interacting fragments and similarity search applications. *J. Chem. Inf. Model.* **2010**, *50*, 79–86.
(14) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(15) *Molecular Operating Environment (MOE)*, version 2008.10; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2008.
(16) 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.
(17) *SMARTS*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008.
(18) Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(19) *MDL Drug Data Report (MDDR)*, version 2005.2; Symyx Technologies, Inc.: Sunnyvale, CA, 2005.
(20) Kaelin, D. E.; Smenton, A. L.; Eiermann, G. J.; He, H.; Leiting, B.; Lyons, K. A.; Patel, R. A.; Patel, S. B.; Petrov, A.; Scapin, G.; Wu, J. K.; Thornberry, N. A.; Weber, A. E.; Duffy, J. L. 4-arylcyclohexylalanine analogs as potent, selective, and orally active inhibitors of dipeptidyl peptidase IV. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 5806–5811.
(21) Liang, G. B.; Qian, X.; Biftu, T.; Singh, S.; Gao, Y. D.; Scapin, G.; Patel, S.; Leiting, B.; Patel, R.; Wu, J.; Zhang, X.; Thornberry, N. A.; Weber, A. E. Discovery of new binding elements in DPP-4 inhibition and their applications in novel DPP-4 inhibitor design. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3706–3710.
(22) Duffy, J. L.; Kirk, B. A.; Wang, L.; Eiermann, G. J.; He, H.; Leiting, B.; Lyons, K. A.; Patel, R. A.; Patel, S. B.; Petrov, A.; Scapin, G.; Wu, J. K.; Thornberry, N. A.; Weber, A. E. 4-aminophenylalanine and 4-aminocyclohexylalanine derivatives as potent, selective, and orally bioavailable inhibitors of dipeptidyl peptidase IV. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2879–2885.
(23) Quan, M. L.; Lam, P. Y. S.; Han, Q.; Pinto, D. J. P.; He, M. Y.; Li, R.; Ellis, C. D.; Clark, C. G.; Teleha, C. A.; Sun, J.-H.; Alexander, R. S.; Bai, S.; Luettgen, J. M.; Knabb, R. M.; Wong, P. C.; Wexler, R. R. Discovery of 1-(3′-aminobenzisoxazol-5′-yl)-3-trifluoromethyl-N-[2-fluoro-4-[(2′-dimethylaminomethyl)imidazol-1-yl]phenyl]-1H-pyrazole-5-carboxyamide hydrochloride (razaxaban), a highly potent, selective, and orally bioavailable factor Xa inhibitor. *J. Med. Chem.* **2005**, *48*, 1729–1744.
(24) Maignan, S.; Guilloteau, J. P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal structures of human factor Xa complexed with potent inhibitors. *J. Med. Chem.* **2000**, *43*, 3226–3232.
(25) Smallheer, J. M.; Alexander, R. S.; Wang, J.; Wang, S.; Nakajima, S.; Rossi, K. A.; Smallwood, A.; Barbera, F.; Burdick, D.; Luettgen, J. M.; Knabb, R. M.; Wexler, R. R.; Jadhav, P. K. SAR and factor IXa crystal structure of a dual inhibitor of factors IXa and Xa. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 5263–5267.
(26) Pfefferkorn, J. A.; Choi, C.; Song, Y.; Trivedi, B. K.; Larsen, S. D.; Askew, V.; Dillon, L.; Hanselman, J. C.; Lin, Z.; Lu, G.; Robertson, A.; Sekerke, C.; Auerbach, B.; Pavlovsky, A.; Harris, M. S.; Bainbridge, G.; Caspers, N. Design and synthesis of novel, conformationally restricted HMG-CoA reductase inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4531–4537.
(27) Pfefferkorn, J. A.; Choi, C.; Larsen, S. D.; Auerbach, B.; Hutchings, R.; Park, W.; Askew, V.; Dillon, L.; Hanselman, J. C.; Lin, Z.; Lu, G. H.; Robertson, A.; Sekerke, C.; Harris, M. S.; Pavlovsky, A.; Bainbridge, G.; Caspers, N.; Kowala, M.; Tait, B. D. Substituted pyrazoles as hepatoselective HMG-CoA reductase inhibitors: discovery of (3R,5R)-7-[2-(4-fluoro-phenyl)-4-isopropyl-5-(4-methyl-benzylcarbamoyl)-2H-pyrazol-3-yl]-3,5-dihydroxyheptanoic acid (PF-3052334) as a candidate for the treatment of hypercholesterolemia. *J. Med. Chem.* **2008**, *51*, 31–45.
(28) Sarver, R. W.; Bills, E.; Bolton, G.; Bratton, L. D.; Caspers, N. L.; Dunbar, J. B.; Harris, M. S.; Hutchings, R. H.; Kennedy, R. M.; Larsen, S. D.; Pavlovsky, A.; Pfefferkorn, J. A.; Bainbridge, G. Thermodynamic and structure guided design of statin based inhibitors of 3-hydroxy-3-methylglutaryl coenzyme A reductase. *J. Med. Chem.* **2008**, *51*, 3804–3813.
(29) Lee, M. E.; Markowitz, J.; Lee, J.-O.; Lee, H. Crystal structure of phosphodiesterase 4D and inhibitor complex(1). *FEBS Lett.* **2002**, *530*, 53–58.
(30) Card, G. L.; England, B. P.; Suzuki, Y.; Fong, D.; Powell, B.; Lee, B.; Lu, C.; Tabrizizad, M.; Gillette, S.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Kim, S.-H.; Schlessinger, J.; Zhang, K. Y. J. Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure* **2004**, *12*, 2233–2247.
(31) Van Zandt, M. C.; Jones, M. L.; Gunn, D. E.; Geraci, L. S.; Jones, J. H.; Sawicki, D. R.; Sredy, J.; Jacot, J. L.; Dicioccio, A. T.; Petrova, T.; Mitschler, A.; Podjarny, A. D. Discovery of 3-[(4,5,7-trifluorobenzothiazol-2-yl)methyl]indole-N-acetic acid (lidorestat) and congeners as highly potent and selective inhibitors of aldose reductase for treatment of chronic diabetic complications. *J. Med. Chem.* **2005**, *48*, 3141–3152.
(32) Steuber, H.; Heine, A.; Podjarny, A.; Klebe, G. Merging the binding sites of aldose and aldehyde reductase for detection of inhibitor selectivity-determining features. *J. Mol. Biol.* **2008**, *379*, 991–1016.
(33) Howard, E. I.; Sanishvili, R.; Cachau, R E.; Mitschler, A.; Chevrier, B.; Barth, P.; Lamour, V.; Van Zandt, M.; Sibley, E.; Bon, C.; Moras, D.; Schneider, T. R.; Joachimiak, A.; Podjarny, A. Ultra-high resolution drug design I: details of interactions in human aldose reductase-inhibitor complex at 0.66 A. *Proteins: Struct., Funct., Genet.* **2004**, *55*, 792–804.
(34) Chamberlain, P. P.; Ren, J.; Nichols, C. E.; Douglas, L.; Lennerstrand, J.; Larder, B. A.; Stuart, D. I.; Stammers, D. K. Crystal structures of Zidovudine- or Lamivudine-resistant human immunodeficiency virus type 1 reverse transcriptases containing mutations at codons 41, 184, and 215. *J. Virol.* **2002**, *76*, 10015–10019.
(35) Das, K.; Clark, A. D.; Lewi, P. J.; Heeres, J.; De Jonge, M. R.; Koymans, L. M.; Vinkers, H. M.; Daeyaert, F.; Ludovici, D. W.; Kukla, M. J.; De Corte, B.; Kavash, R. W.; Ho, C. Y.; Ye, H.; Lichtenstein, M. A.; Andries, K.; Pauwels, R.; Boyer, P. L.; Clark, P.; Hughes, S. H.; Janssen, P. A.; Arnold, E. Roles of conformational and positional adaptability in structure-based design of TMC125-R165335 (etravirine) and related non-nucleoside reverse transcriptase inhibitors that are highly potent and effective against wild-type and drug-resistant HIV-1 variants. *J. Med. Chem.* **2004**, *47*, 2550–2560.
(36) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat. Struct. Biol.* **1995**, *2*, 293–302.
(37) Wright, P. A.; Wilmouth, R. C.; Clifton, I. J.; Schofield, C. J. 'pH-jump' crystallographic analyses of gamma-lactam-porcine pancreatic elastase complexes. *Biochem. J.* **2000**, *351*, 335–340.
(38) Wilmouth, R. C.; Westwood, N. J.; Anderson, K.; Brownlee, W.; Claridge, T. D.; Clifton, I. J.; Pritchard, G. J.; Aplin, R. T.; Schofield, C. J. Inhibition of elastase by N-sulfonylaryl beta-lactams: anatomy of a stable acyl-enzyme complex. *Biochemistry* **1998**, *37*, 17506–17513.
(39) Bernstein, P. R.; Andisik, D.; Bradley, P. K.; Bryant, C. B.; Ceccarelli, C.;, Jr.; Earley, R.; Edwards, P. D.; Feeney, S.; Gomes, B. C.; Kosmider, B. J.; Steelman, G. B.; Thomas, R. M.; Vacek, E. P.; Veale, C. A.; Williams, J. C.; Wolanin, D. J.; Woolson, S. A. Nonpeptidic inhibitors of human leukocyte elastase. 3. Design, synthesis, X-ray crystallographic analysis, and structure-activity relationships for a series of orally active 3-amino-6-phenylpyridin-2-one trifluoromethyl ketones. *J. Med. Chem.* **1994**, *37*, 3313–3326.
(40) Bernstein, P. R.; Gomes, B. C.; Kosmider, B. J.; Vacek, E. P.; Williams, J. C. Nonpeptidic inhibitors of human leukocyte elastase. 6. Design of a potent, intratracheally active, pyridone-based trifluoromethyl ketone. *J. Med. Chem.* **1995**, *38*, 212–215.

CI1001197