Ligand-Target Interaction-Based Weighting of Substructures for Virtual Screening

Thomas J. Crisman,[†,§] Mihiret T. Sisay,[†,‡,§] 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 Pharmazeutisches Institut, Pharmazeutische Chemie I, Rheinische Friedrich-Wilhelms-Universität, An der Immenburg 4, D-53121 Bonn, Germany

Received July 8, 2008

A methodology is introduced to assign energy-based scores to two-dimensional (2D) structural features based on three-dimensional (3D) ligand-target interaction information and utilize interaction-annotated features in virtual screening. Database molecules containing such fragments are assigned cumulative scores that serve as a measure of similarity to active reference compounds. The Interaction Annotated Structural Features (IASF) method is applied to mine five high-throughput screening (HTS) data sets and often identifies more hits than conventional fragment-based similarity searching or ligand-protein docking.

INTRODUCTION

Molecular substructures or fragments are extensively utilized in computer-aided drug design and chemoinformatics and have a long history in chemical and pharmaceutical research.[1−4] Fragments are among the most popular molecular descriptors for compound clustering or database searching[1,2] and are frequently used for de novo compound design.[3] Furthermore, they often serve as building blocks for structure-based ligand design[4] and other fragment linking schemes[5] and are also utilized for synthesis planning.[6]

Structural fingerprints represent a particularly popular format of fragment descriptors and are applied in compound clustering[7] and similarity searching.[8,9] Structural fingerprints designed for similarity searching can essentially be divided into two categories: keyed fingerprints and feature collections. Keyed fingerprints typically have a fixed format where each bit position is associated with a particular fragment whose presence or absence in a test molecule is monitored. Pioneering developments of such fragment dictionary-based fingerprints include MACCS structural keys[10] or BCI fingerprints.[11,12] Recently, keyed fingerprints having a more variable format have also been introduced that encode varying numbers of compound class characteristic substructures.[13] The second class of structural fingerprints, termed here feature collections, essentially represents layered atom environments that are systematically calculated for test molecules and recorded as individual strings or features. Since the number of accessible atom environments (and thus strings) can become exceedingly large, environments cannot be assigned to predefined bit positions but are stored as individual sets of features. Thus, in contrast to keyed fingerprints, varying numbers of strings are generated for different test molecules. Similarity measures are then defined using set operators. For example, the intersection between two sets would correspond to the number of shared "1" bit positions in a keyed fingerprint. Pioneering designs of structural atom environment fingerprints include Molprint2D[14,15] or Scitegic's Extended Connectivity Fingerprints (ECFPs) that are implemented in the Pipeline Pilot software.[16] For the generation of ECFPs, a code is assigned to each nonhydrogen atom consisting of its mass, charge, element type, and the number of bonds to other atoms. The atom code is combined with bond information and codes of neighboring atoms through a hashing procedure; features are sampled until a predefined bond diameter (layer) is reached. These features represent substructures and are recorded as (large) integers for each molecule.

Similarity searching using fragment-type fingerprints is a typical ligand-based 2D mapping procedure. Fingerprints of reference molecule(s) are calculated and compared to corresponding fingerprints of database compounds; fingerprint overlap (corresponding to the number of shared fragments) is quantified via a similarity coefficient.

We have developed an alternative approach to conventional 2D fragment mapping that takes 3D interaction information into account. Interaction fingerprints have previously been reported that encode specific protein−ligand interactions.[17,18] However, different from such representations, we have aimed at designing a ligand-centric fragment approach that utilizes ensembles of 2D structural features and quantitatively annotates them with interaction information. For a set of active reference molecules, interactions in crystallographic ligand-target complexes are scored using an energy function, and atom-based scores are added to substructural features calculated for ligands using extended connectivity fingerprints. The so derived sets of annotated substructures are used for database searching. Herein we report the development and application of the Interaction Annotated Structural Features (IASF) method.

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

† Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry.

§ These authors contributed equally to this work.

‡ Pharmazeutisches Institut, Pharmazeutische Chemie I.

10.1021/ci800229q CCC: $40.75
Published on Web 09/27/2008

Table 1. Target Proteins and Screening Data[a]

	screening IDs	total actives	total inactives	inactive subset	PDB IDs of complexes	docking template
PKA	524, 548	62	64814	6456	1Q8T, 1RE8, 1REJ, 1REK, 1YDS, 1YDT, 2ERZ, 1SVE, 1SVG, 1SVH	1XH8
THR	1046, 1215	223	216693	21929	1WAY, 1WBG, 2C8W, 2C8X, 2C8Y, 2C90, 2C93	1A4W
HIV	565, 651	390	63969	6066	1C0T, 1C0U, 1KLM, 1RT1, 1RT2, 1RTH, 1RTI, 1TKT, 1TKX, 1TKZ, 1TL1, 1TL3	1RTH
HSP	595	300	66228	6548	1UY7, 1UY8, 1UYC, 2VCI, 2VCJ	2BYI
JNK	746	366	59422	5883	1JNK, 1PMN, 1PMQ, 1PMU, 1PMV, 2B1P, 2EXC, 2O0U, 2O2U, 2OK1	2B1P

[a] "Screening IDs" reports the PubChem bioassay identifiers and "total actives" and "total inactives" the number of hits and inactive compounds in each screening data set, respectively. "Inactive subset" gives the number of randomly selected inactive compounds used as background for virtual screening calculations.

MATERIALS AND METHODS

Structural Features. Structural feature ensembles of ligands in complex crystal structures were generated using extended connectivity fingerprints with a bond radius of four (ECFP4)[19] using the Scitegic Pipeline Pilot Student edition 6.1.5.0.[16]

Scoring Function. For scoring of features in bound ligands, individual components of the FlexX scoring function were used.[20,21] Prior to scoring, the active site region for each ligand was defined by including residues within a 6.5 Å radius around each ligand atom. This radius was chosen in order to ensure that longer range interactions were also taken into account. Selected score components included the FlexX energy (match) scores for neutral hydrogen bonds and ionic and aromatic interactions as well as the contact score accounting for lipophilic and van der Waals contacts. For example, the FlexX scoring function[21] assigns an energy score of −4.7 kJ/mol to an ideal hydrogen bond, −8.7 kJ/mol to a strong ionic interaction, and −0.17 kJ/mol to a nonpolar van der Waals contact (for fragment scoring, energy units are omitted). Energy score components were selected that could be separated into per atom contributions in a meaningful way.

Feature Annotation. For each ligand, calculated energy score components were divided into individual atom contributions. ECFP4 features are by design in part overlapping, and each feature must be independently annotated with interaction information. For each ECFP4 feature, contributions of participating atoms were summed to generate the feature score. Features only occurring in a single ligand within an activity class were not included in the analysis and not scored. For features generated by multiple ligands, individual feature scores were averaged to obtain an activity class feature score.

Target Proteins and Ligand Sets. For our analysis, we required complex crystal structures of proteins with multiple ligands. Five target proteins were chosen for which HTS data sets were publicly available including c-jun N-terminal kinase 3 (JNK), heat shock protein 90 (HSP), human immunodeficiency virus reverse transcriptase (HIV), protein kinase A (PKA), and thrombin (THR). For each protein, complexes with different inhibitors available in the Protein Data Bank (PDB)[22] were selected, as summarized in Table 1. The two-dimensional structures of all ligands are shown in Figure .

Screening Data Sets. Five screening data sets, one for each target protein, were obtained from PubChem.[23] These data sets ranged in size from ∼60,000 (JNK) to ∼217,000 (THR) tested compounds and contained between 62 (PKA) and 390 (HIV) hits (Table 1).

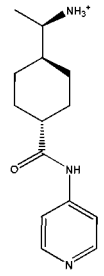
Reference Calculations. For comparison with IASF, standard fingerprint search calculations were carried out with ECFP4 and MACCS. Each HTS data set was searched using the crystallographic inhibitors as reference molecules, and the recall of active compounds among the top 500 screening set compounds (ranked by Tanimoto similarity[19]) was monitored. Enrichment factors over random selection were also calculated. As a fingerprint search strategy for multiple reference compounds, *1-NN* nearest neighbor searching[24] was applied for both fingerprints. This means that for each screening set compound, the Tanimoto similarity to each of the reference molecules is calculated, and the highest value is selected as the final similarity score. In addition to similarity searching, docking calculations with fully flexible ligands were also carried out using FlexX with default parameter settings.[25] In order to meet the computational expense of these calculations and enable a direct comparison with the ligand-based methods, approximately 10% of the total number of inactive compounds were randomly selected from each HTS set (Table 1), and all hits were added. These subsets were used for all (i.e., IASF, fingerprint, and docking) calculations. As a docking template, the target protein with the highest crystallographic resolution available in the PDB was used after removal of the bound ligand (Table 1). For fragment scoring and docking calculations, the same active site regions (as defined above) and the standard FlexX scoring function were used.

As an additional control for IASF calculations, substructure searching was also carried out. For this purpose, maximum common substructures were extracted from the crystallographic ligands sharing identical or similar scaffolds using Pipeline Pilot. These substructures were then used to search the entire screening sets and retrieve all molecules containing them.

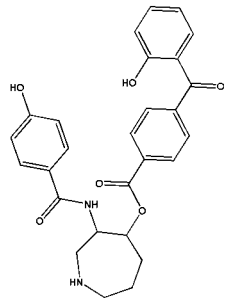
RESULTS AND DISCUSSION

Interaction Annotated Structural Features Method. Molecular fragments are typically used as binary ("present/absent") descriptors in molecular similarity research but have also been applied in weighted form,[26−28] for example, by calculating their frequency of occurrence in active compounds. The underlying idea of the IASF approach is to go beyond statistical analysis of fragment distributions and directly incorporate ligand-target interaction information in

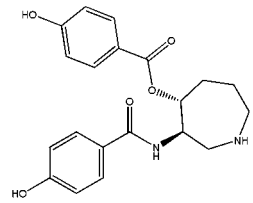
PKA



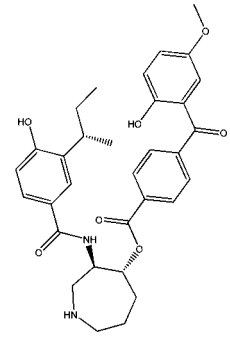
1Q8T



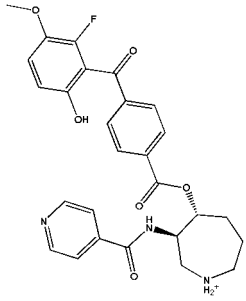
1RE8



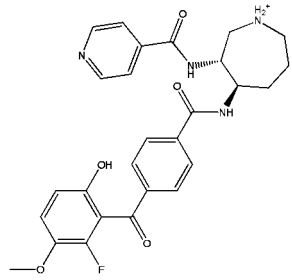
1REJ



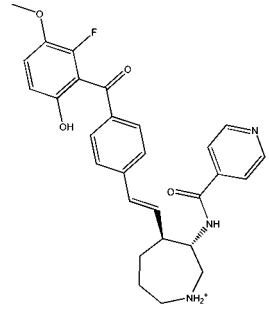
1REK



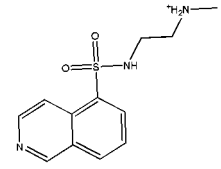
1SVE



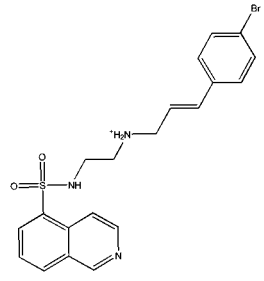
1SVG



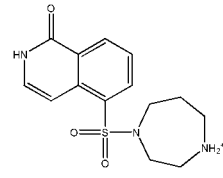
1SVH



1YDS

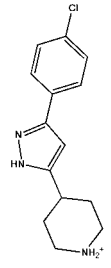


1YDT

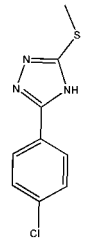


2ERZ

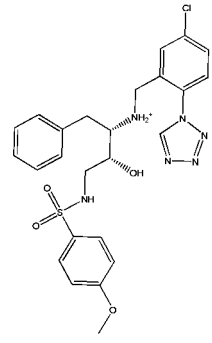
THR



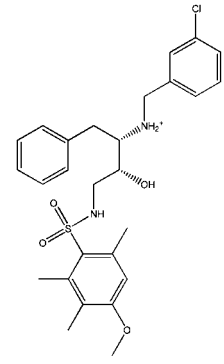
1WAY



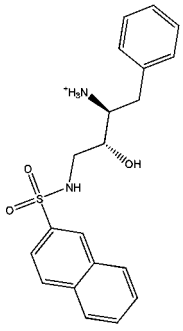
1WBG



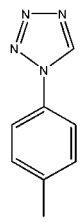
2C8W



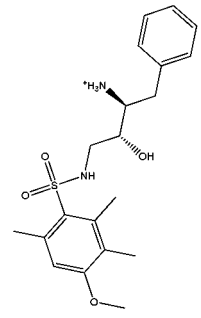
2C8X



2C8Y

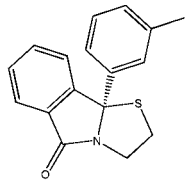


2C90

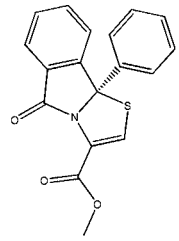


2C93

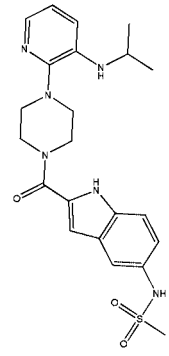
HIV



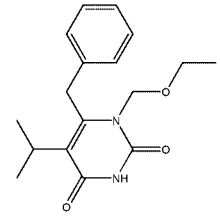
1C0T



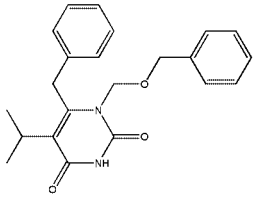
1C0U



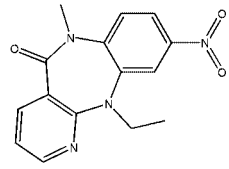
1KLM



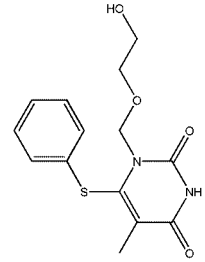
1RT1



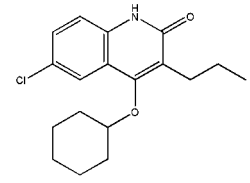
1RT2



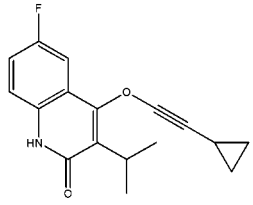
1RTH



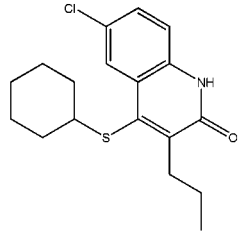
1RTI



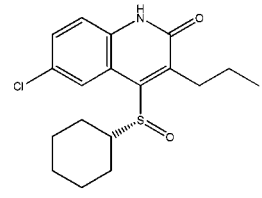
1TKT



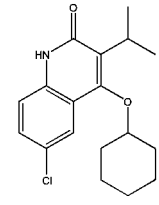
1TKX



1TKZ

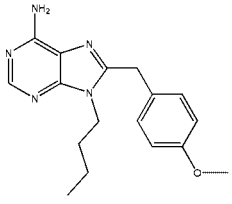


1TL1

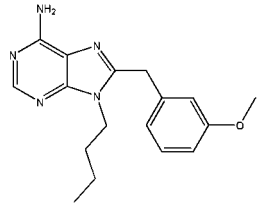


1TL3

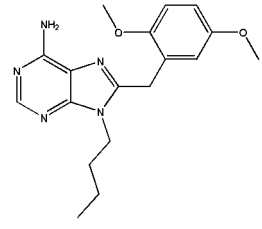
HSP



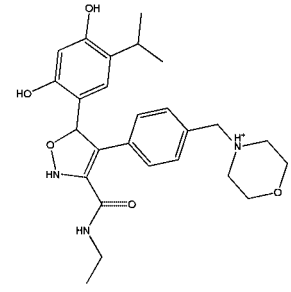
1UY7



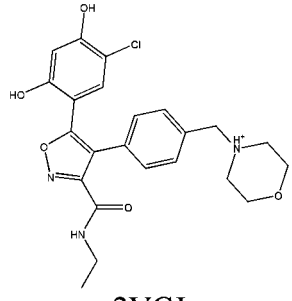
1UY8



1UYC



2VCI



2VCJ

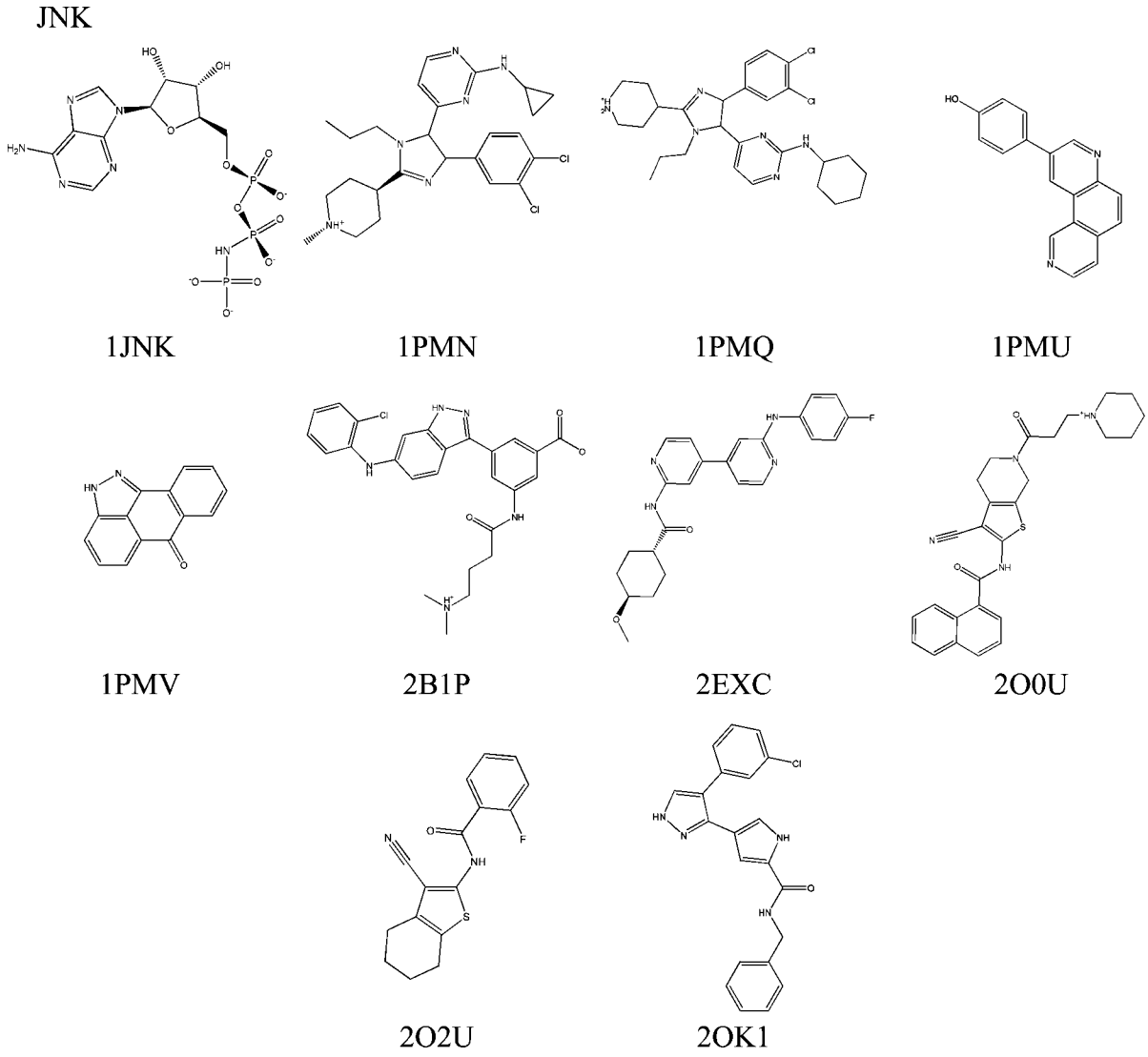


Figure 1. Ligands from X-ray structures. The structures of all ligands used for feature annotation are shown.

fragment-based virtual screening. The different steps involved in IASF calculations are summarized in Figure 2a, and its key aspects are highlighted in Figure 2b. Structural features are generated for ligands available in crystal structures of ligand-target complexes and annotated with energy scores reflecting atomic contributions to interaction energies. Database compounds are screened for corresponding features and obtain cumulative energy scores based on the features they contain. This produces a compound ranking by increasing cumulative scores that are utilized as a measure of similarity between active and database compounds. It would be considerably more difficult to determine cumulative score cutoff values as an indicator of activity, due to the compound set dependence of feature annotation and the relatively small size of the available crystallographic reference sets. Accordingly, as a basis for IASF compound selection, a ranked list is more reliable than predefined threshold values.

Incorporation of experimental ligand-target interaction information into fragment matching is facilitated through the application of an energy function. The accuracy of energy-based scoring functions in structure-based virtual screening is generally limited.[29,30] In IASF calculations, partly overlapping structural features are scored focusing on selected interactions without the need to calculate a global free energy minimum, which represents the major limitation of scoring functions. Conceptually, IASF is best rationalized as a hybrid approach that combines 2D fragment matching with 3D interaction-based fragment weighting. Cumulative scoring on the basis of atom-based interaction energy values balances these contributions.

IASF Analysis. As reported in Table 1, between five (HSP) and 12 (HIV) different crystallographic ligands were used as reference molecules for feature generation and scoring. Comparable or larger numbers of complex structures with different ligands are available for a variety of target proteins in the PDB that could be subjected to IASF analysis. However, our choice of targets was largely determined by the availability of HTS data because we intended to evaluate the approach on experimental screening data sets. Compared to artificially assembled compound benchmarking sets, screening data sets generally provide a more realistic and challenging basis for method comparisons because screening sets consist of experimentally confirmed active and inactive molecules. In addition, screening hits are usually structurally diverse and chemically less complex than highly optimized active compounds that are often used for benchmarking and easier to distinguish from database compounds than screening hits.

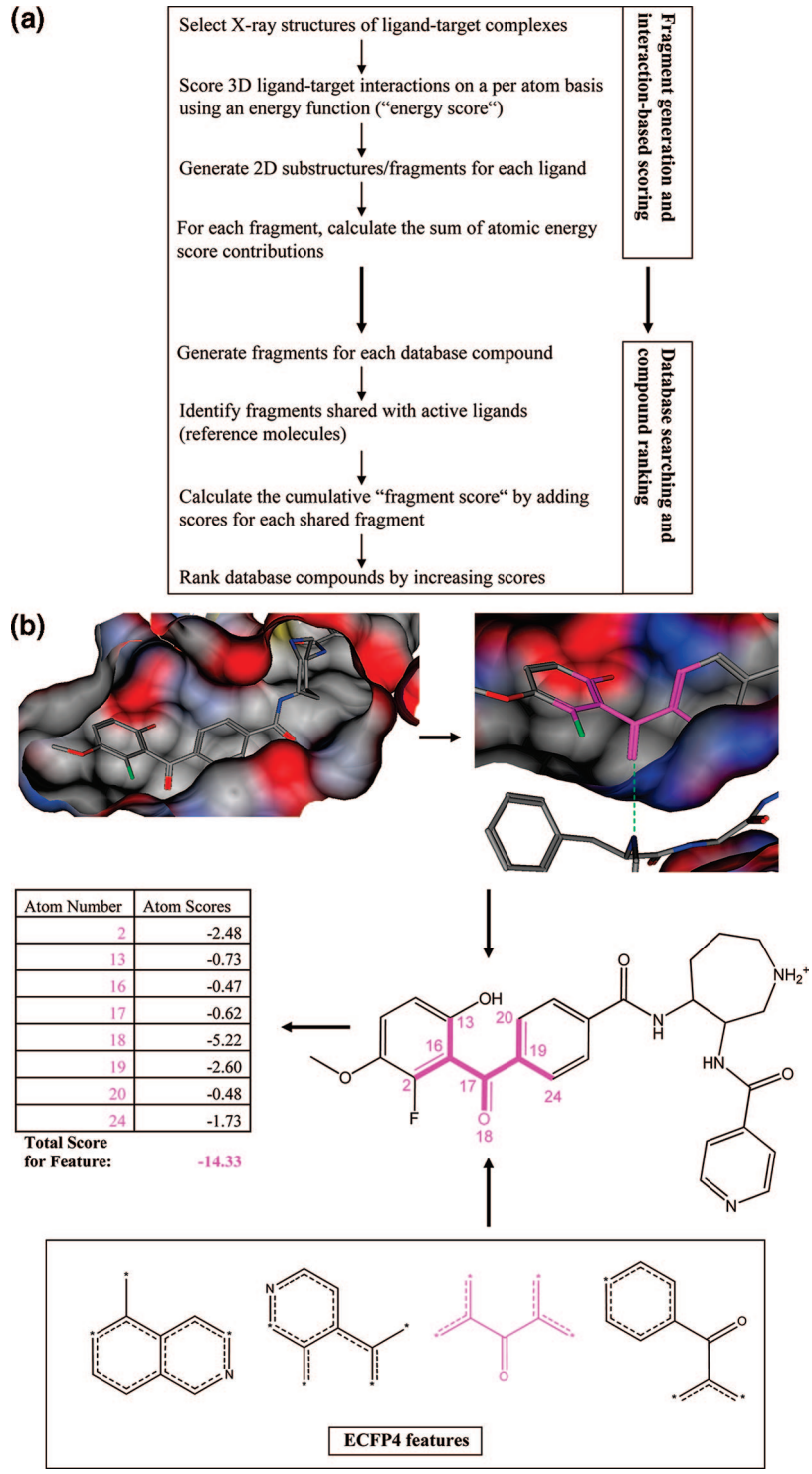


Figure 2. Outline of the IASF approach. In (**a**), a flowchart-like diagram is shown that summarizes the different steps involved in fragment generation, interaction-based scoring, and database searching. In (**b**), the process is further illustrated using a protein kinase A inhibitor as an example. In the upper left, the crystal structure of the inhibitor complex is shown (PDB ID 1SVG). The representation in the upper right focuses on a hydrogen bonding interaction involving carbonyl oxygen 18. The structure of the inhibitor is displayed in the center. Carbonyl oxygen 18 is part of an ECFP4 feature computed for the ligand (consistently shown in magenta). Examples of ECFP4 features are shown at the bottom. On the basis of the calculated atom-based energy scores, the total score for the highlighted feature is obtained. If a database compound is found to contain this feature, it is assigned a score of −14.33. If the compound would contain another annotated feature with a score of −11.25, its cumulative score would thus become −25.58.

In principle, IASF analysis can be carried out using any fragment ensembles and energy functions. Extended connectivity fingerprints were used here because they generate feature sets for individual molecules that are more specific than, for example, predefined fragment dictionaries. Furthermore, the FlexX scoring function was chosen for two reasons. FlexX energy component values can be easily parsed into atomic contributions, and, in addition, the FlexX scoring function emphasizes both uncharged and charged hydrogen bond interactions. The latter aspect was considered important for feature annotation because we intended to primarily capture molecular recognition and specificity determinants

Table 2. Feature and Score Distributions[a]

	annotated features	features per ligand	atoms min	atoms max	score min	score max	score av
PKA	110	10.0	1	10	−18.29	−0.32	−6.91
THR	79	11.2	1	11	−11.58	−0.13	−3.65
HIV	93	7.8	1	10	−15.33	−0.33	−5.71
HSP	74	14.8	1	9	−14.13	−0.14	−4.38
JNK	105	10.5	1	10	−23.97	0.00	−5.50

[a] "Annotated features" gives the total number of accepted ECFP4 features per compound reference set (i.e. features occurring in at least two reference molecules) and "features per ligand" the average number of generated features per crystallographic reference molecule. "Atoms min" and "max" give the minimum and maximum number of atoms per feature, respectively. "Score min", "max", and "av" report the minimum (best), maximum, and average scores per reference set, respectively.

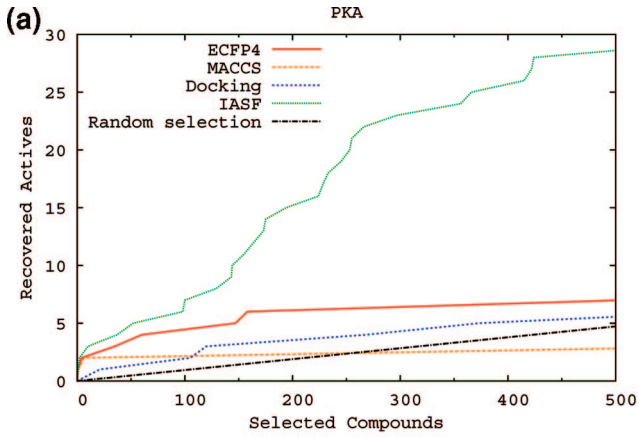
that can be directly assigned to ligand fragments. By contrast, scoring shape complementarity between ligand and target is difficult on the basis of 2D substructures and is here only partly and indirectly accounted for through the inclusion of nonpolar van der Waals contacts (and steric overlap penalties).

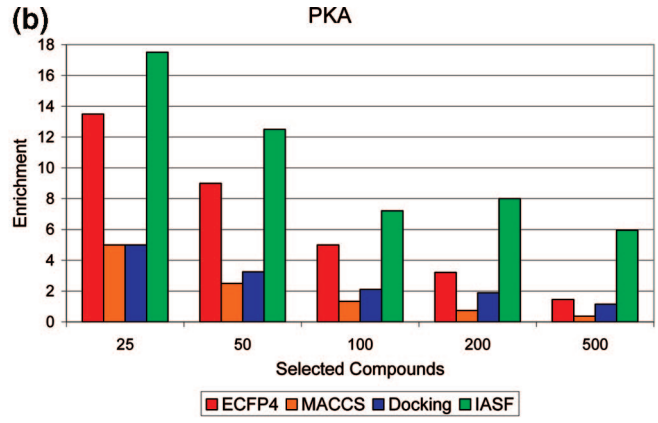
The feature and score distributions for our five compound classes are reported in Table 2. Between 74 (HSP) and 110 (PKA) annotated ECFP4 features were generated per ligand set. These features were used for IASF similarity searching, as discussed below. The average number of features per ligand ranged from 7.8 (HIV) to 14.8 (HSP), and individual features contained between one and 11 non-hydrogen atoms. Feature scores were of comparable magnitude for the different ligand sets, and, in each case, features were found with scores close or equal to zero (which means that they were essentially not involved in interactions accounted for by the score components).

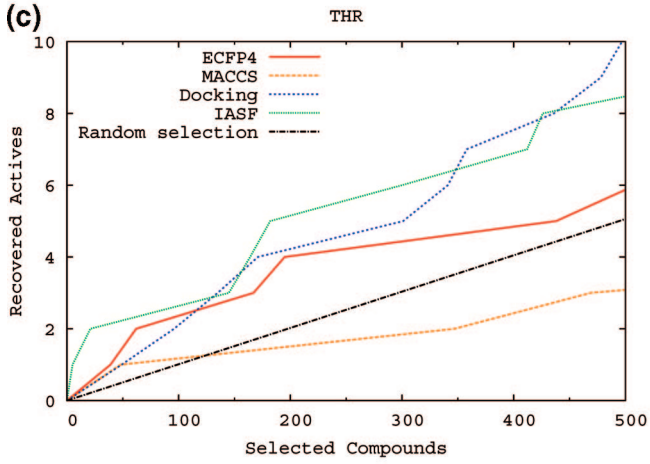
Virtual Screening and Substructure Searching. Following feature generation and annotation, IASF was applied to mine compounds from five HTS data sets in comparison with reference methods, ECFP4 and MACCS fingerprint similarity searching, and FlexX docking. The results are summarized in Figure . As expected, these screening sets provided challenging test cases for all methodologies. Active compounds were retrieved in essentially all calculations, but there frequently was only a 2−3-fold enrichment over random selection. IASF produced overall the highest compound recall and enrichment factors in three of the five cases, PKA, THR, and HIV. In the case of PKA (Figure a,b), IASF clearly dominated the calculations. On THR (Figure c,d), IASF outperformed the reference methods for small selection sets. For HIV (Figure e,f), only IASF and docking calculations produced meaningful recall and enrichment factors. IASF performance was lowest in the case of HSP (Figure g,h) where only five crystallographic ligands were utilized for fragment generation and annotation. However, IASF was consistently superior to ECFP4 in this case. Here docking produced the highest recall followed by MACCS keys. For JNK (Figure i,j), IASF produced the highest recall of hits and enrichment factors for the first 50 screening set compounds. For larger sets, ECFP4 searching produced a higher recall, but the enrichment factors were comparable. Overall IASF performed best on our test cases, in particular, for small compound selection set sizes. Importantly, the performance of IASF was generally superior to ECFP4 searching, although ECFP4 generated a total of approximately 300 to 500 features for each of the five reference sets (compared to between 74 and 110 annotated IASF features). In four of the five cases, IASF was also superior to docking calculations. These findings demonstrate the gain in target-specific information achieved by 3D interaction annotation of ECFP4 features. IASF calculations displayed a notable tendency to enrich active compounds in relatively small database selection sets. This behavior suggests that annotation with interaction information renders search calculation specific for subsets of active compounds that engage in similar interactions. Thus, as intended, IASF calculations focus search calculations on selected structural features (and thereby depart from general structural feature matching).

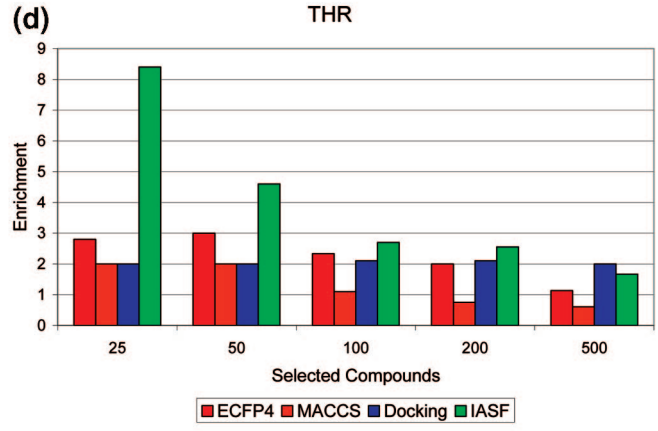
Substructure search calculations using the largest common substructures shared by subsets of crystallographic ligands containing similar scaffolds essentially failed to retrieve active compounds from the screening sets. The results are summarized in the Supporting Information, Table 1. In 12 of 14 substructure search calculations, no hits were identified. In two calculations on HIV and JNK, 56 and 107 active compounds were retrieved together with 13,730 and 15,203 inactive compounds, respectively, thus providing no meaningful search results. These findings demonstrate that IASF calculations are not related to substructure searching.

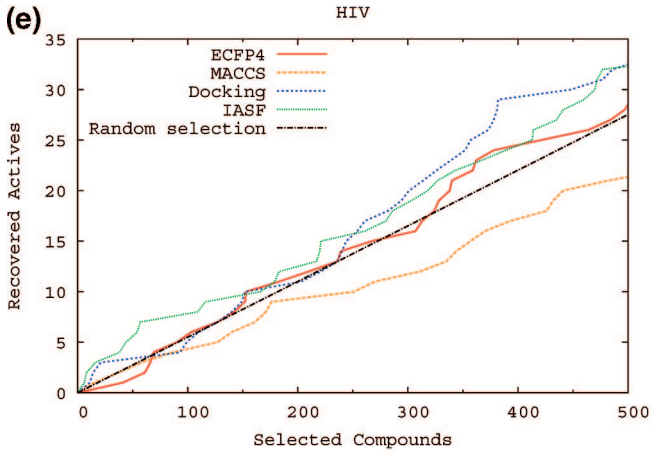
Analysis of Annotated Features. We have also studied selected features and the interactions they form. For example, the substructure highlighted in Figure 2b obtained the third best score among all PKA features. It is a part of the benzophenone motif that is a signature for this series of PKA inhibitors.[31] In the crystal structure, the carbonyl oxygen of the benzophenone moiety forms a hydrogen bond to the amide of PKA residue Phe54. Figure 4 shows the top 12 PKA features in the context of the inhibitor structures they were derived from. All of these features contain heteroatoms that function as hydrogen bond donors or acceptors, consistent with our choice of FlexX score components and our expectation. For example, the top two features contain major portions of the isoquinoline moiety including the nitrogen that forms a hydrogen bond to the adenine ring of ATP, the kinase cofactor. Feature annotation also discriminates between more and less conserved interactions. For example, if a heteroatom within a feature forms a hydrogen bond in only one of several inhibitors it occurs in, as also observed for PKA inhibitors of the benzophenone and azepane series depicted in Figure 2b, the average absolute value of the score of this feature is low for the reference

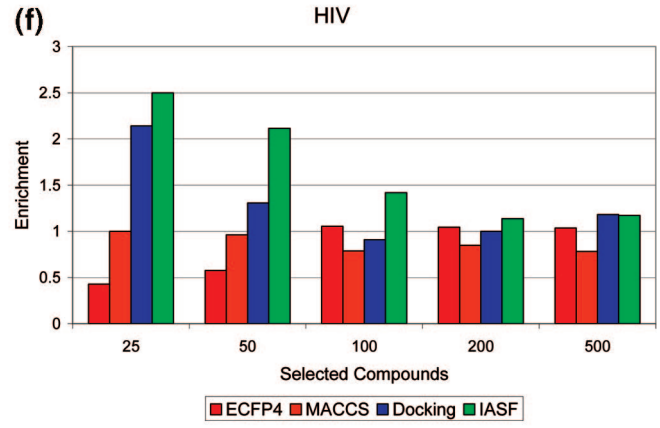

(a)
PKA
ECFP4
MACCS
Docking
IASF
Random selection
Recovered Actives
Selected Compounds

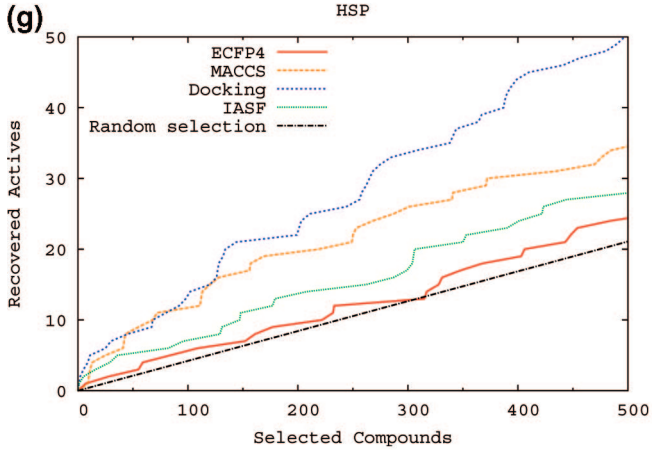

(b)
PKA
Enrichment
Selected Compounds
ECFP4 MACCS Docking IASF

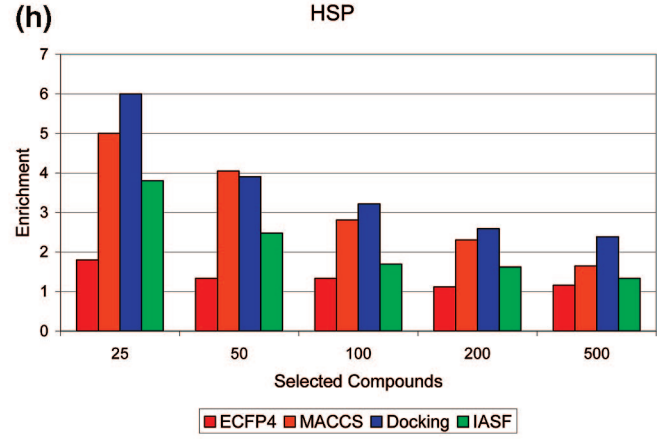

(c)
THR
ECFP4
MACCS
Docking
IASF
Random selection
Recovered Actives
Selected Compounds


(d)
THR
Enrichment
Selected Compounds
ECFP4 MACCS Docking IASF


(e)
HIV
ECFP4
MACCS
Docking
IASF
Random selection
Recovered Actives
Selected Compounds


(f)
HIV
Enrichment
Selected Compounds
ECFP4 MACCS Docking IASF


(g)
HSP
ECFP4
MACCS
Docking
IASF
Random selection
Recovered Actives
Selected Compounds


(h)
HSP
Enrichment
Selected Compounds
ECFP4 MACCS Docking IASF

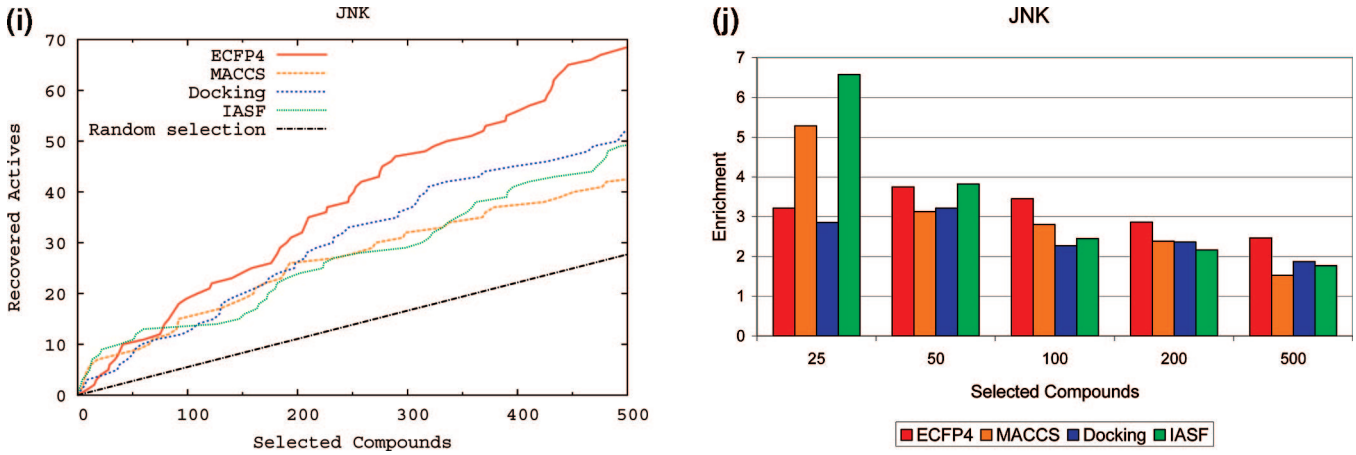


Figure 3. Virtual screenings trials. For each target, two graphs are shown that report recall curves and enrichment factors, respectively, for search calculations using IASF and reference methods. (**a**), (**b**): PKA; (**c**), (**d**): THR; (**e**), (**f**): HIV; (**g**), (**h**): HSP; (**i**), (**j**): JNK.

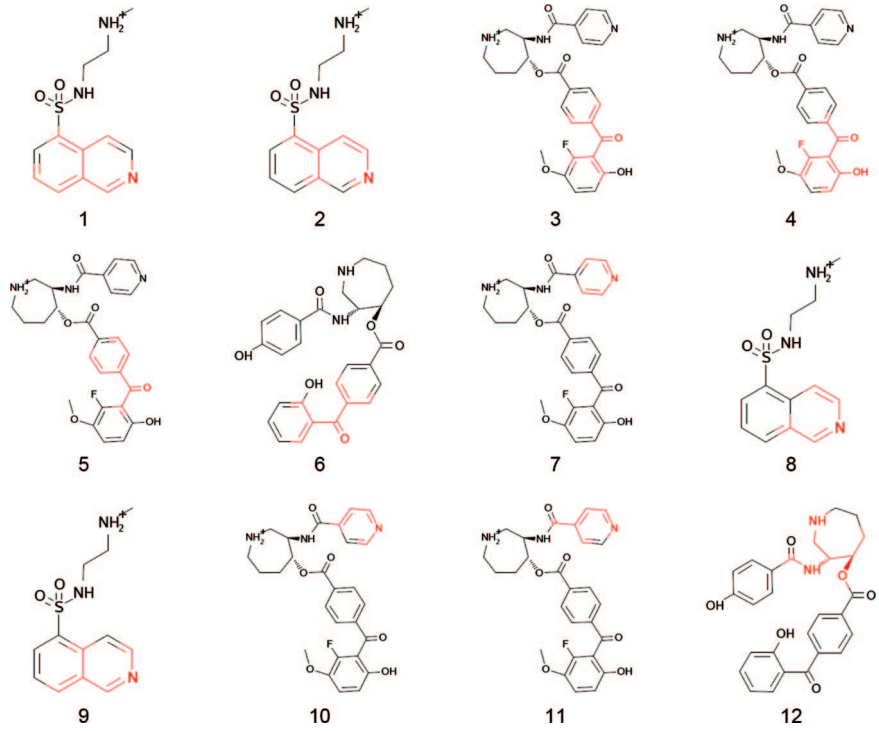


Figure 4. Exemplary features. Shown are the 12 top-scoring substructures (red) found in protein kinase A inhibitors.

set. Thus, substructures involved in interactions that are conserved among reference set inhibitors are most highly weighted.

CONCLUSIONS

A 2D/3D hybrid methodology is introduced that adds ligand-target interaction information to sets of substructures derived from active compounds. Interaction information is extracted from crystallographic data through the application of an energy function. Annotated substructures are used to search databases for active compounds. The methodology is ligand-centric because it relies on mapping of interaction-weighted substructures. Database compounds are assigned cumulative scores based on substructures they share with active reference compounds and the associated energy scores. In benchmark calculations on different HTS data sets, the hybrid approach mostly performed better than 2D (fingerprint) and 3D (docking) calculations. These findings suggest that substructure and interaction knowledge is highly complementary in nature and that there is considerable gain in structure−activity relationship information when these 2D and 3D components are combined.

ACKNOWLEDGMENT

We thank Hanna Geppert and Eugen Lounkine for their critical review of the manuscript. M.T.S. is supported by a fellowship from Graduiertenkolleg (GRK) 677 of the Deutsche Forschungsgemeinschaft (DFG).

Supporting Information Available: Derived substructures and the results of substructure search calculations (Table 1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594–602.

(2) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363–372.
(3) Mauser, H.; Stahl, M. Chemical fragment spaces for de novo design. *J. Chem. Inf. Model.* **2007**, *47*, 318–324.
(4) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons. *Nature Rev. Drug Discovery* **2007**, *6*, 211–219.
(5) Crisman, T. J.; Bender, A.; Milik, M.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Fejzo, J.; Hommel, U.; Davies, J. W.; Glick, M. "Virtual fragment linking": an approach to identify potent binders from low affinity fragment hits. *J. Med. Chem.* **2008**, *51*, 2481–2491.
(6) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
(7) Brown, R. D.; Martin, Y. C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ. Res.* **1998**, *8*, 23–39.
(8) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(9) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(10) *MACCS structural keys*; Symyx Software: San Ramon, CA, 2002.
(11) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.
(12) *BCI*; Digital Chemistry Ltd.: Leeds, U.K., 2008.
(13) Batista, J.; Bajorath, J. Similarity searching using compound class-specific combinations of substructures found in randomly generated fragment populations. *ChemMedChem* **2008**, *3*, 67–73.
(14) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
(15) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
(16) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2008.http://accelrys.com/products/scitegic/ (accessed June 2008).
(17) 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.
(18) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
(19) 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.
(20) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *3*, 470–489.
(21) Böhm, H. J. Development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G. T.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(23) PubChem BioAssays. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi (accessed June 2008).
(24) Hert, J.; Willet, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(25) *FlexX, version 2.2.0*; BioSolveIT GmbH: Sankt Augustin, Germany, 2007.http://www.biosolveit.de/FlexX/ (accessed June 2008).
(26) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165–179.
(27) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
(28) Jorgensen, A.; Langgard, M.; Gundertofte, K.; Pedersen, J. T. A fragment-weighted key-based similarity measure for use in structural clustering and virtual screening. *QSAR Comb. Sci.* **2006**, *3*, 221–234.
(29) 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.
(30) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
(31) Breitenlechner, C. B.; Wegge, T.; Berillon, L.; Graul, K.; Marzenell, K.; Friebe, W. G.; Thomas, U.; Schumacher, R.; Huber, R.; Engh, R. A.; Masjost, B. Structure-based optimization of novel azepane derivatives as PKB inhibitors. *J. Med. Chem.* **2004**, *47*, 1375–1390.

CI800229Q