Tagged Fragment Method for Evolutionary Structure-Based De Novo Lead Generation and Optimization

Qian Liu,*,[†] Brian Masek,[†] Karl Smith,[†] and Julian Smith[‡]

Tripos, Inc., 1699 South Hanley Road, St. Louis, Missouri 63144, and Tripos Discovery Research, Ltd.,
Bude-Stratton Business Park, Bude, Cornwall EX23 8LY, United Kingdom

Received June 25, 2007

Here we describe a computer-assisted de novo drug design method, EAISFD, which combines the de novo design engine EA-Inventor with a scoring function featuring the molecular docking program Surflex-Dock. This method employs tagged fragments, which are preserved substructures in EA-Inventor used for base fragment matching in Surflex-Dock, for constructing ligand structures under specific binding motifs. In addition, a target score mechanism is adopted that allows EAISFD to deliver a diverse set of desired structures. This method can be used to design novel ligand scaffolds (lead generation) or to optimize attachments on a fixed scaffold (lead optimization). EAISFD has successfully suggested many known inhibitor scaffolds as well as a number of new scaffold types when applied to p38 MAP kinase.

Introduction

Enrichment of the drug candidate pipeline is essential for continuing success of pharmaceutical companies. With the increase of therapeutic targets available from human genome sequencing, discovery of novel lead compounds as potential drug candidates looks promising but is also competitive and challenging. In silico virtual screening has been a major design strategy for identifying lead candidates. With continuing growth of available organic compounds from both internal drug discovery programs and external vendors, virtual screening has gradually become a time-consuming approach, not even to mention screening against virtual compounds generated from combinatorial libraries. Computer-assisted de novo drug design methods, such as LEGEND,[1] LUDI,[2] and LeapFrog,[3,4] attracted researchers' attention about a decade ago for obtaining structurally novel drug candidates. However, some major problems associated with these early de novo design methods have prevented most medicinal and computational chemists from making a serious commitment to the use of computational de novo drug design in routine discovery work. The problems include (1) producing chemically invalid structures or structures that do not have drug-like properties; (2) poor synthesizability of the suggested putative ligands; and (3) low structural diversity. Among these, (1) and (2) are the most frequently raised issues.

Often the ease of the synthesis runs counter to the novelty or desirability of the suggested ligands. There are two types of approaches for assessing synthesizability of computer-designed structures in de novo drug-design systems. The first approach combines rules from a reaction knowledgebase into the structure building engine to produce synthetically feasible drug candidates.[5−7] Drug design using such a system can be time-consuming and, therefore, is impractical for suggesting the large diversity of the candidate ligand structures available to the target. Designed products are also limited by the reactions provided in the knowledge base. The second approach, which is implemented by most existing de novo design methods, assesses synthesizability of feasible candidates in a postprocessing step. To minimize the amount of time and effort spent in this postprocessing step, the de novo design phase should strive to produce sensible drug-like compounds with structures that could be synthesized either as they are or with minor modification. If the set of structures resulting from this process is a reasonable number of broadly sampled structurally diverse compounds, there is a greater chance that a medicinal chemist inspecting these results will find structures that could be either interesting new lead scaffolds or templates leading to viable scaffolds.

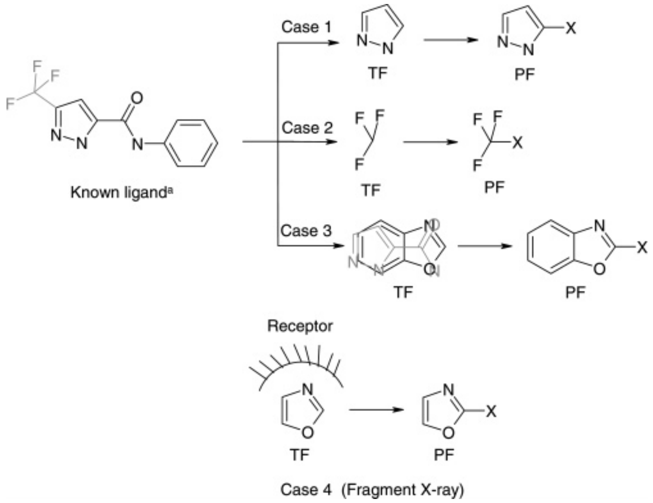

Figure 1. Illustration of the TF concept. [a]Known small molecule or peptide ligand cocrystallized with receptor protein. PF: EA-Inventor preserved fragment. Case 1: A fragment (pyrazole) of the ligand is chosen as a TF for scaffold hopping or partial structure design. Case 2: TF is an artificially attached fragment (trifluoromethyl) for full ligand structure design. Case 3: Scaffold (benzoxazole) from a proprietary lead series is superimposed onto the ligand (gray) for lead optimization. Case 4: An experimentally identified weak binder is treated as a TF for structural expansion.

In recent years, renewed effort has been seen in developing de novo design methods with emphasis on improving the drug-like characteristics of the designed ligands.[8−10] There are also recent reports of successful projects utilizing de novo design methods.[11−15] One such example came from Lloyd et al., who have made enhancements to their receptor structure-based de

* To whom correspondence should be addressed. Phone: 636-207-8971. Fax: 314-647-9241. E-mail: qliu@tripos.com.
[†] Tripos, Inc.
[‡] Tripos Discovery Research, Ltd.

10.1021/jm070750k CCC: $37.00
Published on Web 10/06/2007

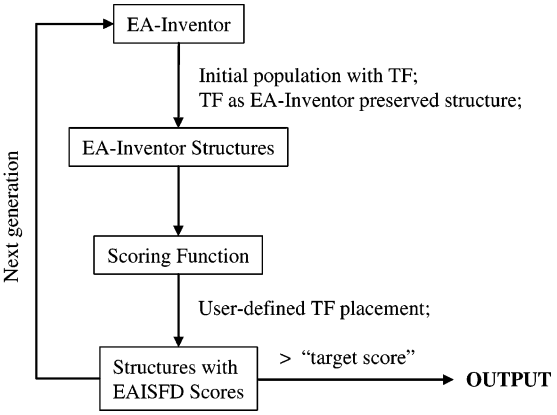
EA-Inventor
Initial population with TF;
TF as EA-Inventor preserved structure;
EA-Inventor Structures
Scoring Function
User-defined TF placement;
Structures with EAISFD Scores
> "target score"
OUTPUT
Next generation

Figure 2. EAISFD workflow.

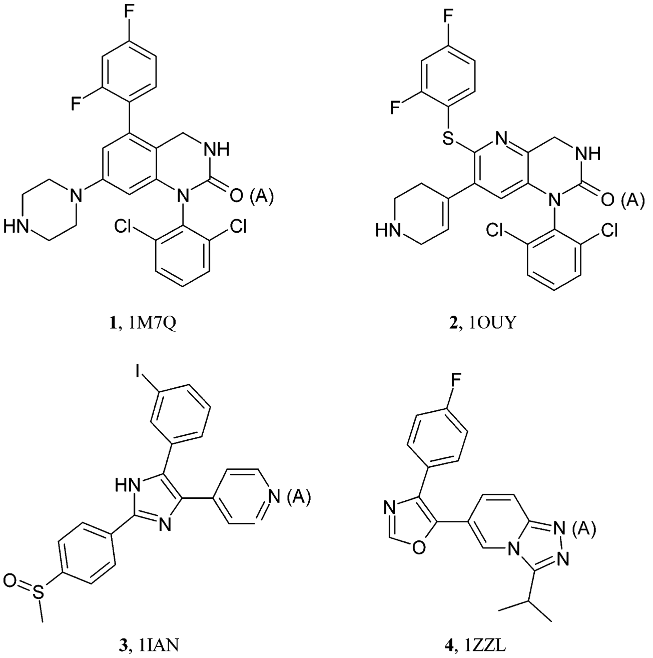

Figure 3. P38 MAP kinase ligands in binding mode 1. The atom marked with "(A)" forms H-bonding interaction with M109NH.

novo design program, Skelgen,[9] by also considering pharmacophore and pseudoreceptor information obtained from known ligands.[16] This approach was successfully used in scaffold hopping to novel ligands. Even though it requires the bound conformation of the ligands to produce reliable results, it is one of few encouraging de novo design studies in recent published work.

Computer-aided de novo drug design approaches, when applied in structure-based design, may not necessarily be effective in producing biologically meaningful structures such as ones that incorporate important receptor interactions if the information of the experimentally determined binding motifs of the known ligands is not taken into account in the design process. On the other hand, prioritization of the computer-generated structures is often based on binding affinities predicted by a generic scoring method, adapted from structure-based virtual screening. Despite intensive research, the relative binding affinities predicted by these methods still do not fully mirror experimentally observed ones.[17-19] However, the reliability of the methods for scoring ligands under restricted binding motifs is expected to be higher than scoring broadly and arbitrarily bound ones simply because the receptor side factors counted by the scoring functions become more consistent for the former.

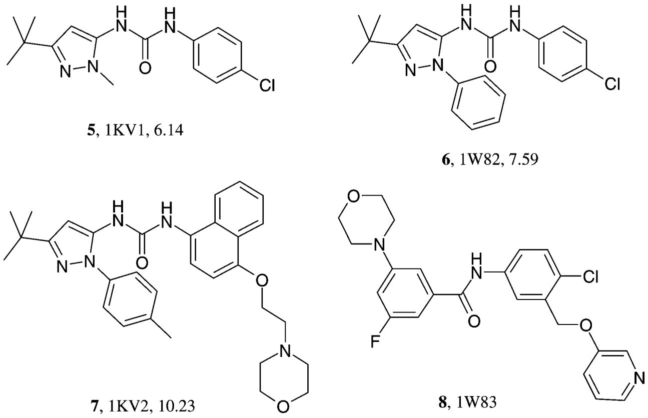

Figure 4. P38 MAP kinase ligands in binding mode 2. Surflex-Dock scores ($-\log(K_d)$) for **5**−**7** under the native binding poses are shown.

Therefore, a receptor-based de novo drug design system that is able to generate ligands, which share a specific binding motif, should be more reliable from the aspect of enforcing proven ligand binding mode as well as the aspect of prediction accuracy of relative binding affinities. Research experience has also shown that ligands known to bind to a receptor often satisfy certain predicted binding affinity score cut-offs, even if the correlation between the score values and the experimentally observed binding affinities cannot be strongly detected. Taking Surflex-Dock[20-22] as an example, experimentally recognized binders often have Surflex-Dock scores higher than 3.0 ($-\log(K_d)$). With this in mind, estimated scores should rather be treated somewhat loosely, for example, being associated with a score cutoff rather than being considered as a strict measure for structure prioritization when they are used to guide ligand design.

EA-Inventor[23] is an evolutionary algorithm based de novo design program that relies on an external scoring function for guiding the structure building process (see Methods). A major step toward receptor structure-based de novo ligand design with EA-Inventor is to develop a scoring function that can accurately estimate binding affinities to guide EA-Inventor's structure modification. In EA-Inventor, structures evolve across multiple generations. A scoring function that can effectively guide EA-Inventor's structure modification should consistently recognize and score the structural changes occurring across multiple generations. Standard docking protocols try to find the most probable binding poses for a given structure, and thus, diverse EA-Inventor structures are likely to be docked in different regions of the ligand binding site or exhibit a different binding mode in the same region of the ligand binding site. Therefore, the receptor side environment for scoring related EA-Inventor structures using such docking protocols lacks consistency throughout structural evolution. Such a scoring scheme provides less meaningful information on what structure modification is favorable. Surflex-Dock is a reasonable choice for scoring EA-Inventor structures in two aspects: (1) it allows the user to select a substructure from the ligands to be docked as a base fragment. The base fragment is consistently positioned in the ligand binding site in the docking process, and thus, scoring structures under specific binding motifs can be realized; and (2) Surflex-Dock's docking accuracy and virtual screening utility are proven by multiple experiments.[24-26] In this study, we have developed a new de novo design method, EAISFD, by coupling EA-

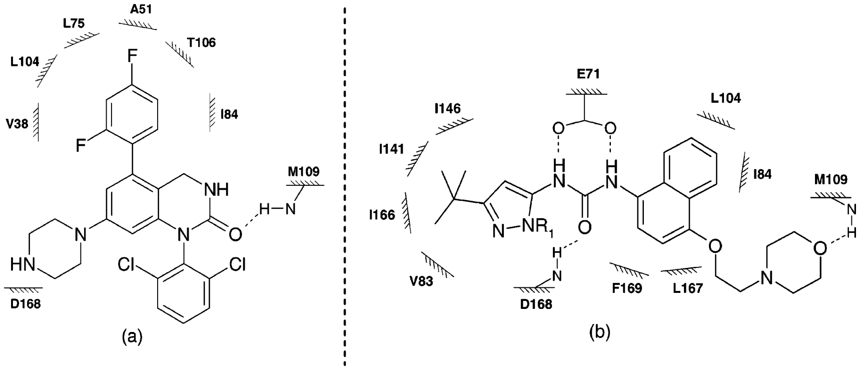

Figure 5. Schematic representation of **1** (a) and **5** (b) binding to p38 MAP kinase. H-bonding interactions are depicted as dotted lines. Amino acids are expressed in single letter code.

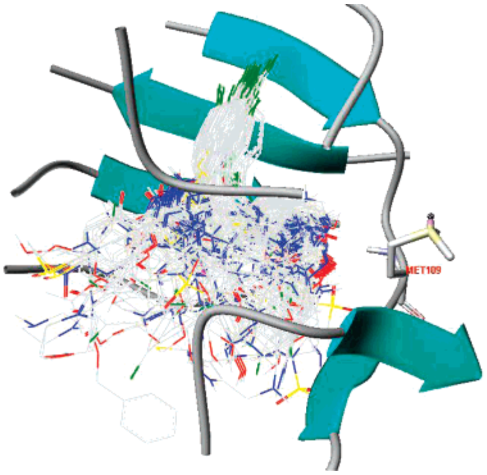

Figure 6. P38 MAP kinase inhibitors docked in 1M7Q.

Inventor with a "Tagged Fragment" (TF[a])-based scoring function that utilizes the Surflex-Dock program. This new method is suited for either suggesting novel scaffolds or optimizing a chemical series with a specific scaffold in common. To examine EAISFD's performance, we applied it to p38 MAP kinase and assessed its ability in reproducing the scaffolds of known p38 MAP kinase inhibitors as well as suggesting novel ones. Optimization of a ligand substructure was also attempted for a ligand of the same target.

Methods

EA-Inventor. EA-Inventor[23] is based on an Evolutionary Algorithm that operates on the connection tables of an initial population of structures to Invent new structures with improved "scores" related to properties that one wishes to optimize through multiple generations. De novo design with EA-Inventor utilizes two components, the EA-Inventor component that controls the evolutionary process that is responsible for all structure modification and a scoring function component that grades each invented structure. More specifically, a set of user provided structures (initial population) are modified by EA-Inventor in the first generation, and the new set of structures is passed to the scoring function for evaluation. Only the structures with good scores survive and remain in the second generation. This process is repeated until EA-Inventor accomplishes the number of generations specified by the user. Because EA-Inventor is a generic structure invention engine, it can be combined with any scoring function to form a unique EA-Inventor program. EA-Inventor has the following key features: (1) It contains a fragment library with over 1300 fragments extracted from MDL Drug Data Report.[27] These "drug-derived" fragment structures are structural sources of EA-Inventor's de novo design. (2) There are 32 chem-evolutionary operators that enable generation of any valid chemical structures obeying valence rules. (3) It is capable of generating large number of structures in a single run. (4) A substructure or substructures can be preserved during structure modification. This is useful for either designing new R-groups while the scaffold is preserved or designing new scaffolds while R-groups are preserved.

Surflex-Dock. Surflex-Dock is an automated method that docks ligands into a receptor's ligand binding site using a protomol based approach and an empirically derived scoring function.[20−22] The protomol is a computational representation of a putative ligand that binds to the intended binding site and is a unique and essential element of the docking algorithm. Surflex-Dock's scoring function, which contains hydrophobic, polar, repulsive, entropic, and solvation terms, was trained to estimate the dissociation constant (K_d) expressed in $-\log(K_d)$ unit. In addition to the automated docking procedure, the function of Surflex-Dock has recently been enhanced by incorporating a base fragment matching algorithm that allows prepositioning a fragment of the ligand being docked in the binding site. The fragment is allowed to shift from its original position in certain degree during pose optimization. This is important when the position of the base fragment is not completely fixed. Ligand docking with the base fragment matching feature is intended to yield docking and scoring of ligands constrained to match a specific binding motif.

EAISFD. EAISFD combines the de novo design engine EA-Inventor for structure evolution with Surflex-Dock for docking and scoring to yield a receptor structure-based de novo design method, referred to as EAISFD in the sequel. In EAISFD we have also linked the preserved substructure concept in EA-Inventor with the base fragment feature in Surflex-Dock. We refer to this preserved substructure/base fragment as a TF in EAISFD. By introducing the TF concept, the binding affinities of invented structures can be estimated under specific ligand binding motifs. Because the base fragment is not held completely rigid during docking, a hydrophobic fragment, which is not as directional as fragments rich in H-bonding interactions, can also be treated as a TF.

Figure 1 describes four scenarios for TF determination in a real drug discovery environment. A TF can be either a fragment of the ligand (case 1) or a new fragment attached to the ligand (case 2) in a crystal structure complex. Such TF strategies are useful for partial or full ligand structure design where the TF serves to anchor key binding interactions. For example, while case 1 is an obvious choice for lead optimization, case 2 is specifically suited for designing new ligand structures that occupy the whole ligand binding area of the receptor from scratch. Scaffold hopping can be achieved by adopting either scheme. As an extension to case 1, case 3 shows that lead optimization of a proprietary chemical series can be realized by choosing a proprietary scaffold structure as a TF and superimposing it onto the ligand in the crystal structure complex based on their pharmacophoric compatibility. Fragment

[a] Abbreviations: TF, tagged fragment; ADMET, absorption, discretion, metabolism, excretion, and toxicity; PDB, Protein Data Bank; M109NH, main chain NH moiety of Met 109; WDA, World Drug Alerts Plus; 2F4CLPHE, 2-fluoro-4-chlorophenyl.

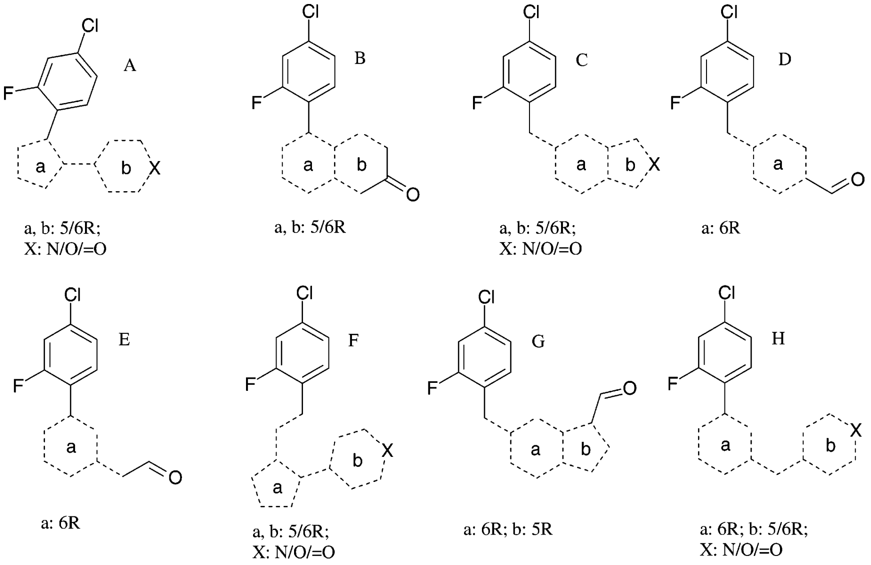

Figure 7. Generic representation of the eight core types (A-H) of binding mode 1 inhibitors. 5(6)R: five (six)-membered ring. 5/6R: five- or six-membered ring. Dashed lines represent any atom/bond types. Any atom type also covers the atom that is connected to both dashed and solid lines.

based drug discovery strategies have been gaining more and more popularity in recent years.[28] Low molecular weight fragments with weak binding affinities are discovered by applying experimental technologies such as NMR and X-ray crystallography. Fully enumerated ligand structures can be constructed by converting such weak binder fragments to EAISFD's TF's (case 4).

To achieve the balance of structural diversity and desired binding affinity, a "target score" parameter is implemented in EAISFD in the way that any EA-Inventor structures whose Surflex-Dock scores reached the target score are exported in the result set. Meanwhile, these structures are removed from EA-Inventor's ongoing generation so that a wider variety of structures can be sampled and reported. The value of the target score should be determined based on the objective of the project. Lead optimization requires higher target scores while lower scores are appropriate for scaffold hopping. Docking scores of known ligands often provide valuable references for defining the target score. EAISFD workflow is described in Figure 2.

EAISFD Assessment Protocol. We examined EAISFD's utility for both scaffold hopping and lead optimization. For scaffold hopping, we assessed the extent to which EAISFD was able to invent scaffolds that are similar to the scaffolds found in known ligands. EAISFD should be able to produce the same or similar ligand scaffolds if the scaffolds are predicted to be good binders to the target receptor protein by Surflex-Dock. We further examined and classified the scaffolds designed by EAISFD that did not exist in the known ligands as an extended evaluation. To make sure the new scaffolds are valid from the chemistry aspect, we intended to choose a rather simple approach than conducting a real chemical reaction based assessment because reaction design is more suited for individual chemical structures than for generic chemical scaffolds. Here, we examined the availability of the scaffolds by searching the ZINC[29] compound database with a generic 2D query built from each scaffold type.

In contrast to scaffold hopping, it is rather difficult to set up a standard protocol for assessing EAISFD's ability to optimize lead structures because the number of optimized leads reported is often limited, and the structures often reflect a mixture of multiple design objectives such as enhancing drugability involving ADMET properties while maintaining reasonable binding affinities at the target. Our strategy in assessing lead optimization ability of EAISFD is to examine the structural diversity, drug-likeness, as well as the existence of important pharmacophores of the partial structures designed by EAISFD.

Results and Discussion

1. Data Preparation. A well-defined drug target was required for EAISFD assessment. We looked for a drug target that (a) has experimentally determined three-dimensional structure of the target receptor protein in complex with a ligand molecule and (b) has a variety of confirmed active ligands which show similar binding motifs. P38 MAP kinase is an extensively studied protein kinase target, and initial investigation showed that it qualifies as a suitable drug target for EAISFD assessment. A number of crystal structures of p38 MAP kinase in complex with small molecule inhibitors were extracted from Protein Data Bank[30] (PDB), and the ligand−receptor interaction was examined. As a result, two ligand binding sites, the ATP binding pocket and an adjacent allosteric binding site, were recognized, which is consistent with a previous report.[31] We classified p38 MAP kinase inhibitors based on the two distinct ligand binding modes. Small molecule inhibitors **1**, **2**, **3**, and **4** occupy the ATP binding pocket (which we refer to as binding mode 1) as revealed by crystal structures 1M7Q, 1OUY, 1IAN, and 1ZZL (Figure 3). The halogen-substituted phenyl rings in the ligands occupy an inner hydrophobic pocket of the binding site. In addition, H-bonding interactions with the main chain NH moiety of Met 109 (M109NH) through the H-bond acceptor marked with "(A)" in all four ligands were observed (Figure 5a). M109NH was reported to form H-bonding interaction with N-1 atom of ATP adenine,[32] and thus, formation of the H-bonding interaction with M109NH is considered important for inhibitors occupying the ATP pocket.

The representative ligands associated with the non-ATP allosteric binding site (binding mode 2) were extracted from four crystal structures 1KV1, 1W82, 1KV2, and 1W83 (Figure 4). The Surflex-Dock scores shown for **5** and **7**, which were estimated based on their native binding poses, were used to determine the target scores for the related EAISFD runs. Analysis of the ligand binding mode suggested that the pyrazole ring in **5**−**7** fit in a relatively hydrophobic receptor cavity, while the urea in **5**−**7** as well as the amide in **8** form H-bonding interactions with Glu 71 (E71) and Asp 168 (D168) (Figure 5b). The phenyl rings next to the urea or the amide in all four ligands show hydrophobic interactions with four hydrophobic amino acid residues, I84, L104, L167, and F169, as illustrated in Figure 5b. Further observation revealed that the morpholino oxygen in **7** or the pyridinyl nitrogen in **8** is involved in M109NH interaction.

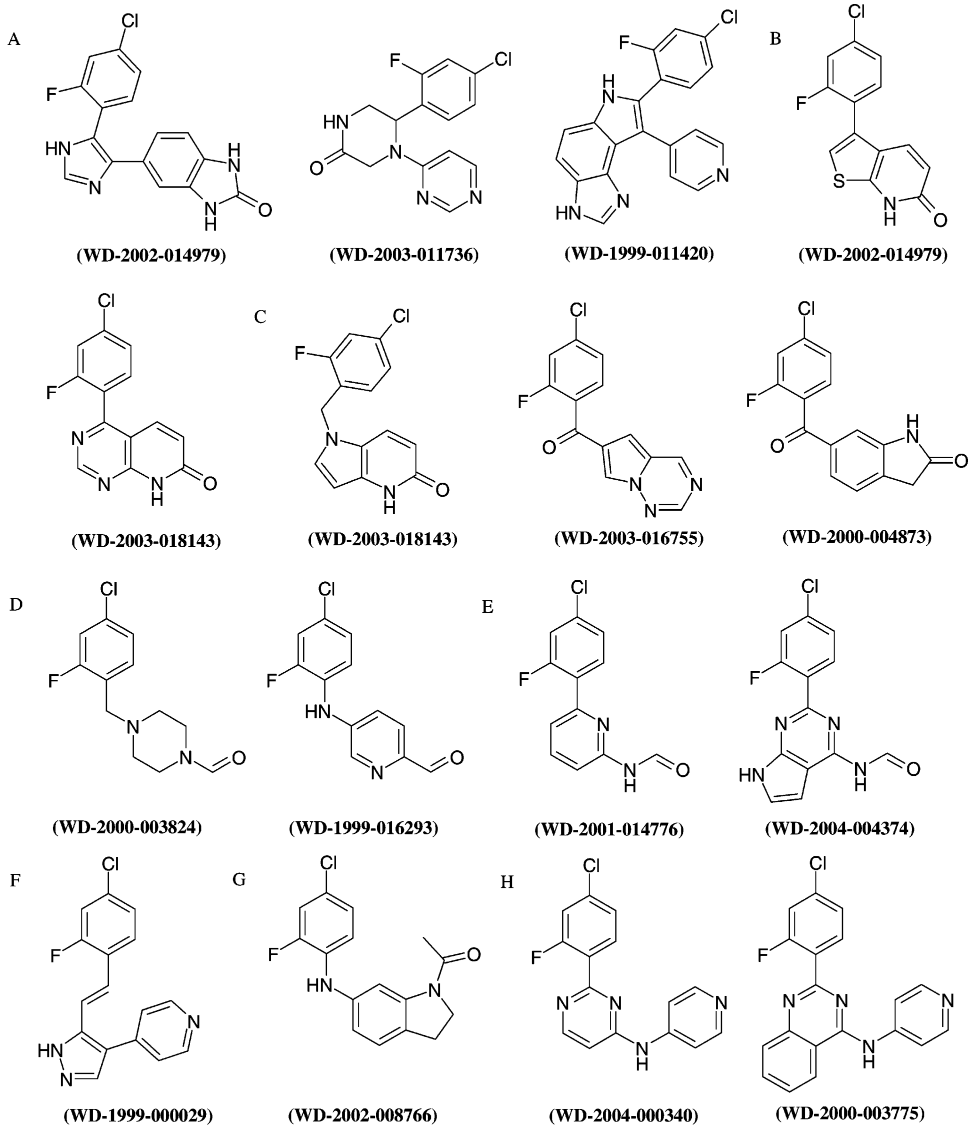

Figure 8. Examples of core structures in eight known core types of binding mode 1 inhibitors. Aromatic rings equivalent to the halogen-substituted phenyl in all structures are replaced by 2F4CLPHE.

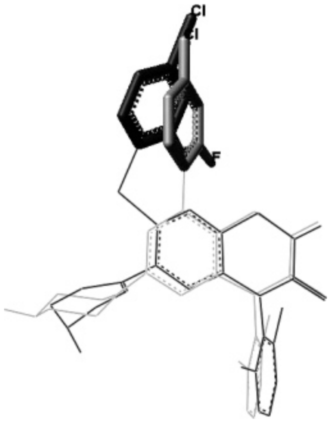

Figure 9. TF1 (black) and TF2 (gray).

We also searched Derwent World Drug Alerts Plus[33] (Derwent WDA) for registered p38 MAP kinase inhibitors, aimed at collecting all available scaffold types for EAISFD scaffold hopping ability assessment. A total of 209 compounds were identified with the indication of p38-kinase-inhibitor in the "mechanism of action" field. Only 98 structures that are likely binding mode 1 inhibitors as based on pharmacophoric similarity to known ATP binding site inhibitors were extracted for further confirmation using Surflex-Dock.

2. Scaffold Hopping. From the binding motifs of ligands **1**−**4**, a halogen-substituted phenyl ring fits into a conserved inner hydrophobic pocket and thus is considered an important structural element in the ligand scaffolds. It is connected to the M109NH binding fragment either directly or through a linker fragment. Considering the low structural diversity and limited mobility, this hydrophobic ring is a suitable candidate for use as the TF in EAISFD for designing the remaining scaffolds of the binding mode 1 ligands. To verify that the known ligands can be successfully docked and scored when related to their expected M109NH binding motif using EAISFD scoring function, we docked the 98 possible binding mode 1 inhibitors with Surflex-Dock based on several TF positions. This is a crucial step because EAISFD will not suggest structures that do not score well by its scoring function. The crystal structure of p38 MAP kinase from 1M7Q was used to define the ligand binding site environment for Surflex-Dock. Crystal structure 1OUY, 1IAN, and 1ZZL were superimposed onto 1M7Q based on protein sequence similarities. The halogen-substituted phenyl rings in the cocrystallized ligands **1**−**4** were extracted from the superimposed positions, and were treated as candidate positions

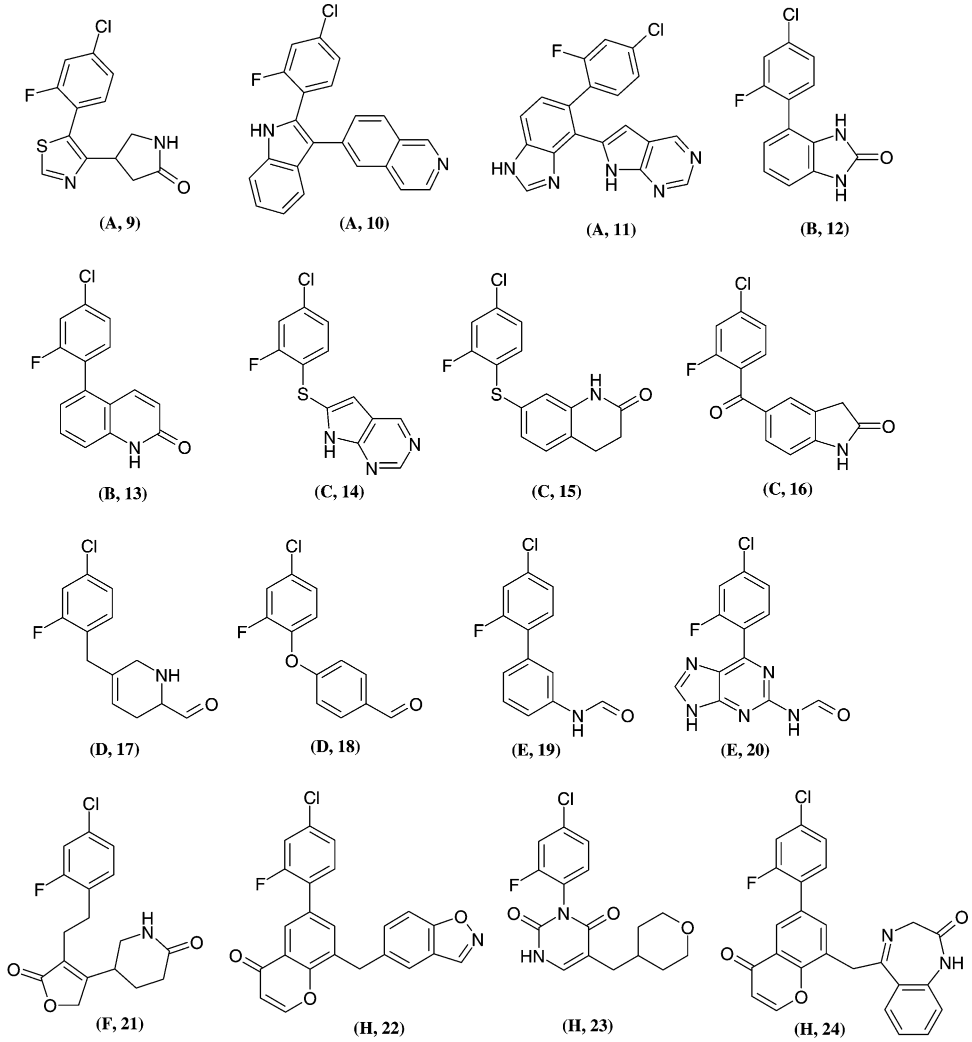

Figure 10. Structures invented by EAISFD in seven of the eight known core types of binding mode 1 inhibitors. Letter under each structure represents the core type.

Table 1. EAISFD Setup and Results for Scaffold Hopping[a]

	TF			
	TF1		TF2	
	RUN1	RUN2	RUN3	RUN4
target score (−log(K_d))	3.0	5.0	3.0	5.0
total EAISFD structures	9035	3654	6452	1906
structures with M109NH interactions	566	472	426	456
avg. mol. weight[b]	346.7	427.6	315.36	370.7
scaffold types[c]	A(14), B(5), C(18), D(1), E(1), F(1), H(3), A1(12), B1(1), C2(1), D1(5), D2(15), E1(1), E2(6), H1(6)	A(3), B(4), C(12), A1(3), B1(4), C1(2), D2(5), E2(2), H1(5)	A(10), B(7), C(16), D(1), E(6), A1(2), D1(1), D2(16), E2(5), F1(2), G1(1)	B(6), C(1), D(1), E(5), B1(2), C2(1), D1(3), D2(10), E1(1), E2(8)

[a] Generations, 100; population size, 200. [b] Calculated for structures with M109NH interaction. [c] From structures with M109NH interactions. Count of unique structures within the same scaffold type is shown in parenthesis.

of the TF. The halogen-substituted phenyl in each of **1**−**4** was replaced with 2-fluoro-4-chlorophenyl (2F4CLPHE) to form a unique base fragment matching for Surflex-Dock. Ligands **1**−**4** were docked with a 2F4CLPHE base fragment in each of the four candidate positions obtained above and the best docking pose of each ligand was compared to the native pose in the crystal structure. We observed that the native poses of ligands **1**−**4** were reproduced by the highest scoring poses when 2F4CLPHE was placed in the position of ligand **2**, except that **3** is shifted away from M109NH slightly and thus not in the optimal distance of the H-bonding interaction with M109NH. The structures of the 98 WDA inhibitors were docked under similar conditions after replacing the halogen-substituted phenyl or a corresponding substructure with 2F4CLPHE. Examination of the best scored docking poses suggested that 90 out of the 98 inhibitors demonstrated the important M109NH interactions. A two-dimensional projection of the 3D view of the binding motifs for the 90 inhibitors is shown in Figure 6, where 2F4CLPHE tightly clustered around the initial base fragment position.

Scaffolds of the 90 ligands that successfully docked and scored were collected for defining ligand core structure types.

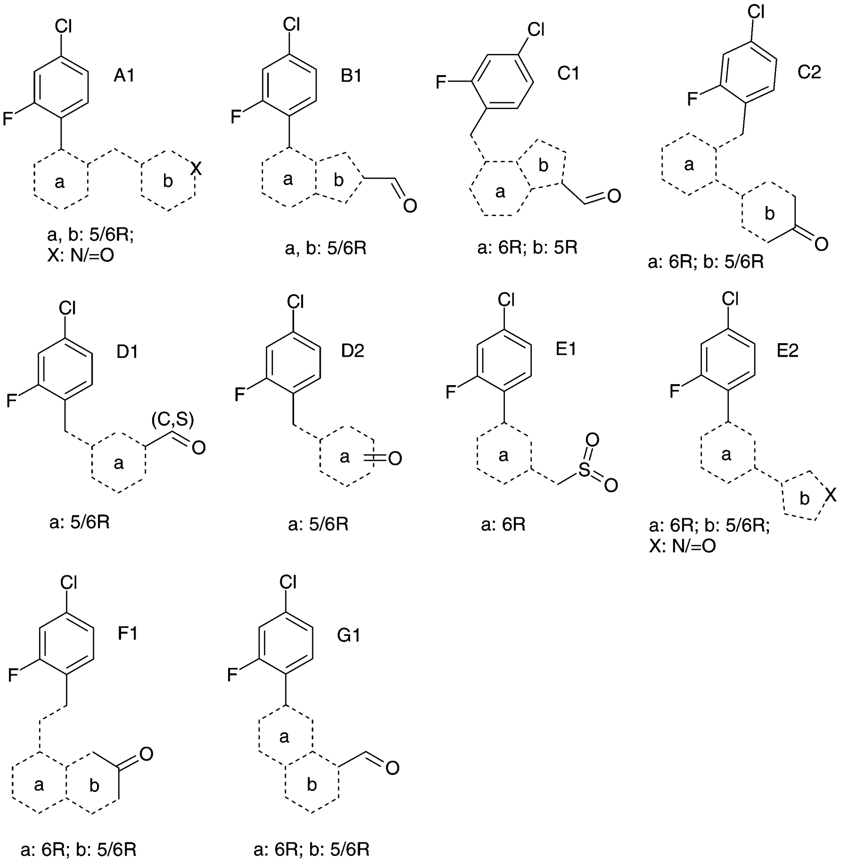

Figure 11. Generic representation of 10 new core types of binding mode 1 inhibitors invented by EAISFD. Core types are named based on structural similarity to known core types. 5(6)R: five (six)-membered ring. 5/6R: five- or six-membered ring. Dashed lines represent any atom/bond types. Any atom type covers the atom that is connected to both dashed and solid lines.

A total of 49 unique core structures were extracted from the 90 inhibitors by retaining 2F4CLPHE, the substructures with an H-bond acceptor interacting with M109NH, and any linker groups. These 49 core structures were redocked under the same condition as before, and the Surflex-Dock scores (range = 2.4−5.6; mean = 3.82) were recorded for reconfirmation. We visually inspected the 49 cores, plus the cores from **1**−**4**, and categorized them into eight core types based on structural characteristics of the linker groups (Figure 7). As shown in Figure 7, the linker group can be a single non-ring atom (type C), a ring moiety (type A), or combinations of non-ring atoms and rings. A single ring in a core type can also be extended to fused rings as long as the relative geometrical location of the important H-bond acceptor interacting with M109NH is not affected. For example, **WD-2002-014979** and **WD-1999-011420** are both defined in type A despite the fused rings in the linker group of **WD-2002-014979**. An exception to such linker group description is type B, which has no linker group. Also note that the classification is irrelevant to patentability of the structures. We intended to observe how many core types can be reproduced by the structures generated by EAISFD. Examples of the original core structures from p38 MAP kinase inhibitors in each core type are shown in Figure 8.

In an attempt to understand the influence of various EAISFD parameters on the results, we perfomed multiple EAISFD runs with different parameter sets. One important parameter is the target score which controls the maturity of EAISFD structures. Two target score values, 3.0 and 5.0, were selected by considering the Surflex-Dock score range obtained from docking the set of 49 core structures. We anticipated that higher target score values would likely encourage larger and more complex EAISFD structures and vice versa. Another parameter related to the design strategy is the selection of a TF. We chose 2F4CLPHE as the TF and placed it onto two of the four candidate positions obtained previously. The first position was from **2** (TF1 in Figure 9) considering its outstanding performance in reproducing the experimental binding motifs of the known ligands. The position from **1** (TF2 in Figure 9) was the next to be considered for investigating EAISFD's sensitivity to the orientation of the TF.

The results of four EAISFD runs with the combination of two target scores and two TF positions are summarized in Table 1. The number of EAISFD structures forming H-bonding interaction with M109NH is about the same in all four result sets despite the significant differences observed in the total number of EAISFD structures met the target scores. The formation of an H-bonding interaction was examined by checking for the existence of an H-bond acceptor within 2.9 Angstroms from M109NH. We should point out that the formation of this H-bonding interaction was not explicitly defined as a design criterion but was taken into account in the post filtering step. Of the eight known core types, RUN1 recovered seven demonstrating the best performance among the four EAISFD runs, followed by five in RUN3, four in RUN4, and three in RUN2. Core **F** and **H** are only found in RUN1, while **G** is missing in all four runs. The results implied that the looser target score, 3.0, was able to reproduce more known ligand cores and, as expected from our docking studies, TF1 is preferred by EAISFD for the current study. Examples of EA-Inventor structures invented by EAISFD in the seven known core types are shown in Figure 10. While most structures are noticeably drug-like and synthetically feasible, structural modi-

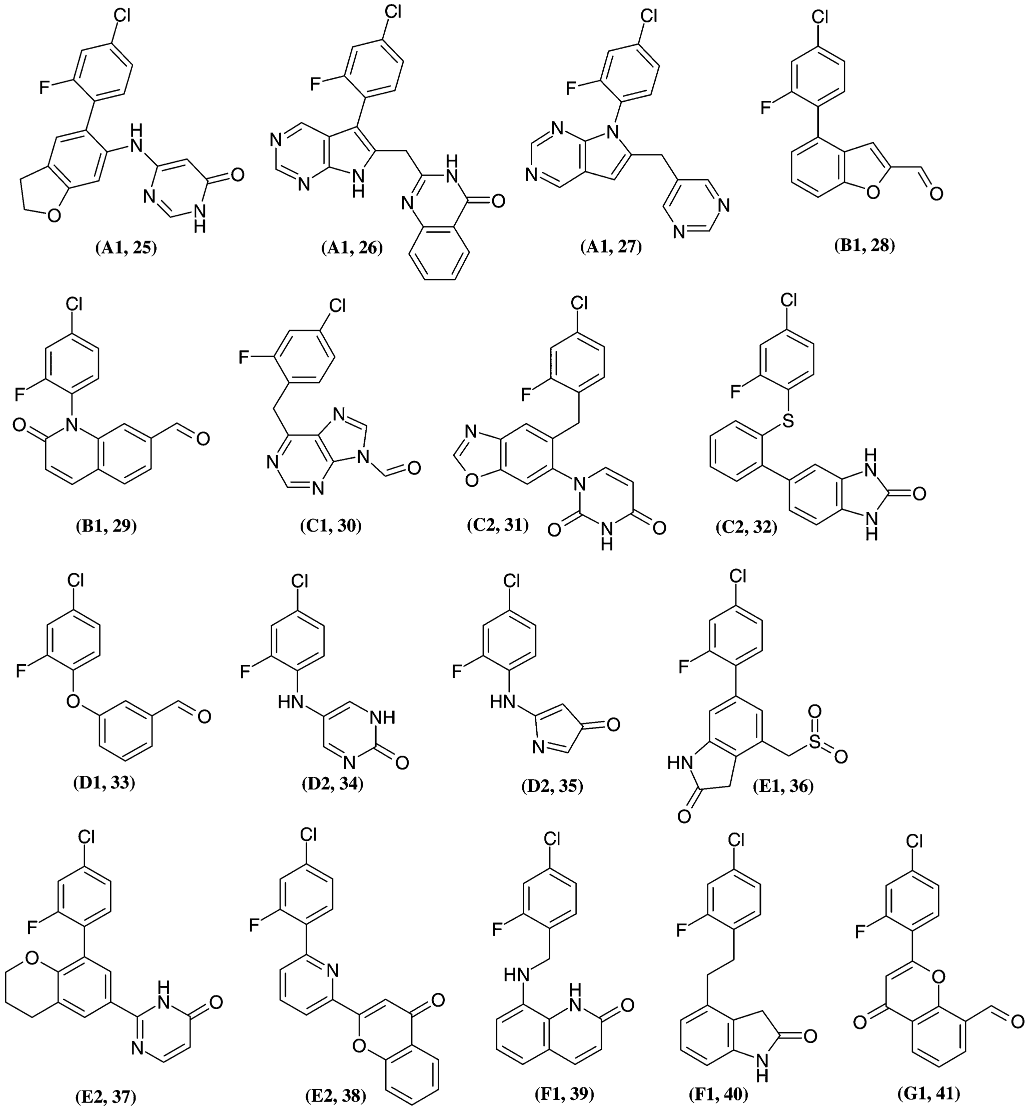

Figure 12. Structures invented by EAISFD in the 10 new core types of binding mode 1 inhibitors.

fication may be required for individual ones like **9**, **17**, and **21**, which have chiral carbons. A chiral carbon can be built in an EA-Inventor structure in the structure construction process by chance, and such a chiral center may or may not be essential for the specific receptor interaction. Thus, we strongly encourage the project scientists to explore the nonchiral structures that can be resulted through certain structure conversion and then reassess the new structures by EAISFD scoring function. Similar structural modification can be applied to any EA-Inventor structures that are interesting but cannot be synthesized as they are. As such, EA-Inventor structures provide valuable structural templates that may lead to new drug candidates even if they are not synthetically feasible in the first glance.

In addition to the known ligand cores, 10 new types of cores, all with reasonable M109NH binding motifs, were invented by EAISFD (Figure 11). Each new core was named after a known one with which it has the highest structural similarity among the known cores in our judgment. For example, **A1** was named because it can be formed by inserting a single atom between the linker ring and the M109NH binding moiety in core **A**. RUN1 alone gave eight of the ten new cores, which is the most productive run in delivering new cores (Table 1), while RUN2, RUN3, and RUN4 all produced six new cores. The lack of unique cores in RUN2 and RUN4 is likely due to low overall structural diversity of the structures which they generated caused by excessive structural optimization for achieving the higher target score, 5.0, as evidenced by the higher average molecular weight. Comparison between TF1 and TF2 showed that core **A1**, **C1**, and **H1** are only found in TF1 based EAISFD runs, while **F1** and **G1** are unique to TF2. Such different results can be due to either the geometrical difference between TF1 and TF2 or the fact that EA-Inventor is a stochastic approach that is subjected to probabilistic factors. To examine the existence of available chemical structures possessing these new types of core structures as an assessment of the validity of the core types produced by EAISFD in chemistry point of view, we built generic structural queries reflecting the core structure descriptions shown in Figure 11 and searched the ZINC[29] database. As a result, 9 out of the 10 core structures were found in the hits. No hit was returned for core **A1** from the search. Nonetheless, compounds with **A1** core type is theoretically synthesizable based on our chemistry assessment. Examples of EAISFD structures in the new ligand core types are shown in Figure 12.

EAISFD successfully reproduced scaffolds of known p38 MAP kinase inhibitors and also generated a number of novel cores predicted to bind in a similar mode. As such, these results illustrate how this tool provides a valuable starting point for medicinal chemists wishing to patent navigate, particularly in crowded IP space. The subtle yet distinct variation between the

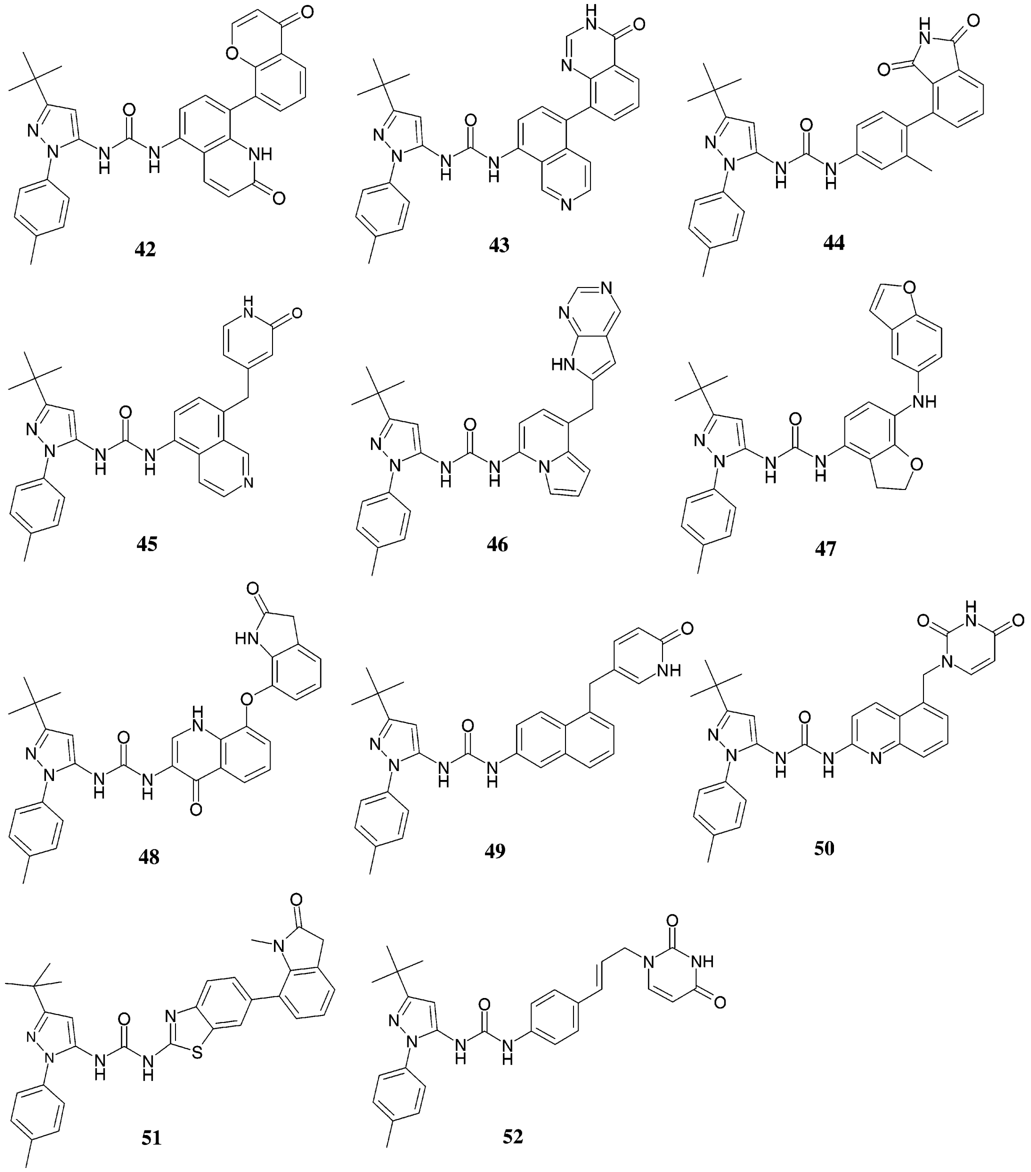

Figure 13. Analogue structures of **7** invented by EAISFD with the M109NH binding motif.

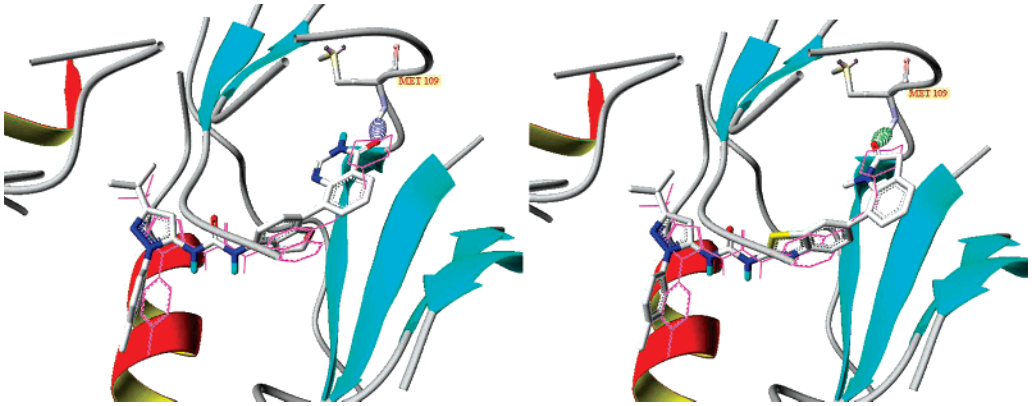

Figure 14. Graphical expression of the M109NH binding motif by **47** (left) and **51** (right). Ligand **7** is represented in pink line expression.

cores identified offers many new routes to evaluate, explore, and refine. The results showed that EAISFD, when it is used for scaffold hopping, is somewhat but not strictly dependent on TF positions at looser target scores. Moreover, experimenting with multiple TF positions can improve the odds of finding more ligand cores. Our study also indicate that a less stringent target score (3.0 in this study) will encourage broader sampling of the chemistry space by EAISFD and lead to diverse scaffolds.

By implementing the TF mechanism, the time required for scoring EA-Inventor structures using Surflex-Dock is reduced dramatically comparing to a standard Surflex-Dock docking based scoring protocol, provided that the actual scoring time is

Table 2. Timing for EAISFD Scoring Function[a]

	set 1		set 2	
	EAISFD	non-TF[c,d]	EAISFD	non-TF
time[b]	4 min	16 min	7 min	37 min

[a] Set 1: 100 structures; 200 < mol. weight < 300. Set 2: 100 structures; 400 < mol. weight < 500. All structures in sets 1 and 2 contain 1-methyl-3-*t*-butyl-pyrazole. [b] Time was measured on Linux system with an Intel Xeon 3.20 GHz CPU. [c] TF: 1-methyl-3-*t*-butyl-pyrazole. [d] Standard Surflex-Dock protocol based scoring.

Table 3. EAISFD Setup and Results for Substructure Optimization[a]

structures from EAISFD	2219
structures with M109NH interactions	63
avg. mol. weight[b]	521.8

[a] Generations, 50; population size, 300. Target score ($-\log(K_d)$): 9.0. [b] For structures with M109NH interactions.

influenced by the size and flexibility of the molecules. Table 2 shows that scoring 100 drug-size structures using EAISFD scoring function requires only about one-fifth of the time needed by the standard Surflex-Dock protocol.

3. Substructure Optimization. In contrast to scaffold hopping, which aims at identifying novel ligand core structures, the goal of lead optimization is to identify alternatives of the substructures relative to a fixed core for enhancing potency and improving ADMET profiles of the lead series. By treating the ligand core structure as the TF, EAISFD is capable of suggesting novel substructures attached to the ligand core as long as the TF position reflects the core structure's true binding motif within the receptor.

Among inhibitors under binding mode 2, **7** is reported to be highly potent and selective.[30] The distinctive properties of **7** over **5** and **6** is clearly attributed to the characteristic substructure that forms H-bonding interaction with M109NH through the morpholino oxygen. Designing substructures with such H-bonding abilities is a logical strategy for lead optimization. In this example, we selected 1-[3-*tert*-butyl-1-*p*-tolyl-1*H*-pyrazol-5-yl]urea in **7** as the preserved substructure of EA-Inventor, and then let EAISFD suggest alternatives to the remaining partial structures that have good estimated binding affinities through M109NH interaction. The 1KV2 protein−ligand complex was used for Surflex-Dock setup. The 3-*tert*-butyl-1*H*-pyrazole moiety with the initial placement from **7** within 1KV2 was used the TF. Based on the docking scores of ligand **6** (7.59) and **7** (10.23), a target score of 9.0 was chosen in a subsequent EAISFD run to achieve sufficient binding affinities of the structures designed by EAISFD.

The 2219 structures derived from a single EAISFD run were filtered for structures formed H-bonding interactions with M109NH using the same H-bonding criteria as described in the scaffold hopping section. A total of 63 structures met this criterion (Table 3). Surveying the structures whose binding motifs do not involve H-bonding with M109NH revealed that some structures reached the target score through hydrophobic and H-bonding interactions irrelevant to M109NH. Such structures may have good binding affinities to p38 MAP kinase but likely lack the specificity profile of **7** that was achieved by the M109NH binding motif. The 63 structures were visually inspected for suitability and drug-likeness. Examples of 11 EAISFD structures that represent five substructure types are depicted in Figure 13. Three of the five substructure types are exemplified by structures **42**−**44**, **45**−**47**, and **48**−**50**, respectively. Structures **51** and **52** each represent a unique substructure type. Structures **48**−**50** are the most similar ones to **7**. All structures except **46** are more rigid than **7**, which are usually characterized by target specific ligands. Figure 14 illustrates the M109NH binding motif by structures **49** and **50**.

As demonstrated by this experiment, EAISFD successfully constructed drug-like substructures as alternatives to the M109NH bonding fragment in **7**. We believe that these EAISFD structures provide valuable templates for the optimization of the lead series represented by **7**. For leads that do not have experimentally determined binding motifs, the binding pose of the scaffolds can be estimated by either docking- or pharmacophore-based overlay as illustrated by case 3 in Figure 1.

Conclusion

We developed a receptor structure-based de novo drug design method, EAISFD, which combines the de novo design engine EA-Inventor with the molecular docking program Surflex-Dock. This method uses a TA concept, which combines preserved substructures in EA-Inventor with a base fragment matching feature in Surflex-Dock. Instead of trying to fully optimize a lead, EAISFD generates a set of diverse structures predicted to exceed the designated threshold of the binding affinity. EAISFD has a number of benefits over typical de novo design approaches: (1) it can be used in both scaffold hopping and optimization of a lead series; (2) design can be focused on known receptor binding motifs; and (3) diverse drug-like structures are generated in a high-throughput manner through the use of a target score parameter. Designed structures that are interesting but cannot be synthesized as they are should be treated as templates that may lead to synthetically feasible drug candidates through synthesizability assessment and structural modification as required. A number of experiments showed that the results of EAISFD are influenced by the magnitude of the target score as well as the placement of the TF. In general, a low score threshold is suitable for scaffold hopping, while a higher threshold is usually required for designing accomplished ligand candidates. The TF placement not only ensures the ligands to be constructed under certain binding motifs within the receptor, but also influences the design strategies, scaffold hopping or lead optimization. Two examples are shown for the application of EAISFD in both tasks. Application of EAISFD on p38 MAP kinase successfully produced seven out of eight known ligand core structure types, and it is very encouraging to discover 10 novel core types in addition to the known ones. This implied that EAISFD is able to correctly capture the receptor binding requirement and construct diverse drug-like ligand structures effectively with the receptor information. In another application, EAISFD was used to suggest replacement of the partial structure of a known ligand within the scope of lead optimization. Five types of drug-like substructures all with the M109NH H-bonding interaction were successfully produced by EAISFD.

Acknowledgment. The authors are grateful to Bob Clark for suggestions on the experiment as well as comments on the manuscript.

References

(1) Nishibata, Y.; Itai, A. Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* **1991**, *47*, 8985−8990.

(2) Bohm, H. J. LUDI: The computer program LUDI: A new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.

(3) *SYBYL 7.3*, Molecular Modeling System; Tripos Inc.: St. Louis, MO, 63144−2913

(4) Makhija, M. T.; Kasliwal, R. T.; Kulkarni, V. M.; Neamati, N. *De novo* design and synthesis of HIV-1 integrase inhibitors. *Bioorg. Med. Chem.* **2004**, *12*, 2317−2333.

(5) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Aken, K. V.; Janssen, P. A. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765−2773.

(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) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and searching 10(20) synthetically accessible structures. *J. Comput.-Aided Mol. Des.* **2007**, *21* (6), 341−350.

(8) Boda, K.; Johnson, A. P. Molecular complexity analysis of *de novo* designed ligands. *J. Med. Chem.* **2006**, *49*, 5869−5879.

(9) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H. J.; Dean, P. M. A validation study on the practical use of automated *de novo* design. *J. Comput.-Aided Mol. Des.* **2002**, *16* (7), 459−478.

(10) Schneider, G.; Fechner, U. Computer-based *de novo* design of drug-like molecules. *Nature Rev.* **2005**, *4*, 649−663.

(11) Greer, J;, Erickson, J. W.; Baldwin, J. J.; Varney, M. D. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J. Med. Chem.* **1994**, *37*, 1035−1054.

(12) Iwata, Y.; Naito, S.; Itai, A.; Miyamoto, S. Protein structure-based *de novo* design and synthesis of aldose reductase inhibitors. *Drug Des. Discovery* **2001**, *17*, 349−359.

(13) Honma, T. Recent Advances in *de novo* design strategy for practical lead identification. *Med. Res. Rev.* **2003**, *23*, 606−632.

(14) Takano, Y.; Koizumi, M.; Takarada, R.; Takimoto, K.M.; Czerminski, R.; Koike, T. Computer-aided design of a factor Xa inhibtor by using MCSS functionality maps and a CAVEAT linker search. *J. Mol. Graph. Model.* 2003, *22* (*2*), 105−114

(15) Schmidt, J. M.; Mercure, J.; Tremblay, G. B.; Page, M.; Kalbakji, A.; Feher, M.; Dunn-Dufault, R.; Peter, M. G.; Redden, P. R. *De novo* design, synthesis, and evaluation of novel nonsteroidal phenanthrene ligands for the estrogen receptor. *J. Med. Chem.* **2003**, *46*, 1408−1418.

(16) Lloyd, D. G.; Buenemann, C. L.; Todorov, N. P.; Manallack, D. T.; Dean P. M. Scaffold hopping in *de novo* design. Ligand generation in the absence of receptor information. *J. Med. Chem.* **2004**, *47*, 494−496.

(17) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333−344.

(18) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, III, C. L. Assessing scoring functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(19) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(20) Pham, T. A.; Jain, A. J. Parameter estimation for scoring protein−ligand interactions using negative training data. *J. Med. Chem.*, **2006**, *49*, 5856−5868.

(21) Jain, A. N. Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 396−403.

(22) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.*, **2003**, *46*, 499−511.

(23) Unpublished.

(24) Polgar, T.; Keseru, G. M. Virtual screening for β-secretase (BACE1) inhibitors Reveals the importance of protonation states at asp32 and asp228. *J. Med. Chem.* **2005**, *48*, 3749−3755.

(25) Gruneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental conformation. *J. Med. Chem.* **2002**, *45*, 3588−3602.

(26) Lyne, P. D.; Kenny, P. W.; Cosgrove, D. A.; Deng, C.; Zabludoff, S.; Wendoloski, J. J.; Ashwell, S. Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *J. Med. Chem.* **2004**, *47*, 1962−1968.

(27) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(28) Erlanson, D. A. fragment-based lead discovery: A chemical update. *Curr. Opin. Biotechnol.* **2006**, *17* (6), 643−652.

(29) 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.

(30) 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.

(31) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S., Regan, J. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9* (4), 268−272.

(32) Tong, L.; Pav, S.; White, D.; Rogers, S.; Crane, K.; Cywin, C. A. Highly specific inhibitor of human p38 MAP kinase binds in the ATP pocket. *Nat. Struct. Biol.* **1997**, *4*, 311−316.

(33) Thomson Corporation (http://www.thomson.com/).

JM070750K