

# Fragment Shuffling: An Automated Workflow for Three-Dimensional Fragment-Based Ligand Design

Britta Nisius<sup>§</sup> and Ulrich Rester\*

Bayer HealthCare AG, Global Drug Discovery, Lead Generation and Optimization, Aprather Weg 18a,  
D-42096 Elberfeld, Germany

Received December 16, 2008

Fragment-based approaches display a promising alternative in lead discovery. Herein, we present the automated fragment shuffling workflow for the identification of novel lead compounds combining central elements from fragment-based lead identification and structure-based *de novo* design. Our method is based on sets of aligned 3D ligand structures binding to the same target or target family. The implementation comprises three different ligand fragmentation methods, a scoring scheme assigning individual scores to each fragment, and the incremental construction of novel ligands based on a greedy search algorithm guided by the calculated fragment scores. The validation of our 3D ligand design workflow is presented on the basis of two pharmaceutically relevant drug targets. A retrospective study based on a selected protein kinase data set revealed that the fragment shuffling approach realizes extended results compared to the well-known BREED technique. Furthermore, we applied our approach in a prospective study for the design of novel non-peptidic thrombin inhibitors. The designed ligand structures in both studies demonstrate the potential of the fragment shuffling workflow.

## 1. INTRODUCTION

The early phases of commercial drug discovery programs are increasingly guided by information extracted from three-dimensional structures of target proteins. Therefore, the systematic exploration of the established and steadily growing knowledge on a target or target family and the corresponding ligands appears to be a promising way to speed up and further industrialize the target family oriented drug discovery.<sup>1</sup> Today, the Protein Data Bank (PDB)<sup>2</sup> is the major source for three-dimensional structures. Since the late 1990s X-ray structure analysis has emerged to a medium- to high-throughput technology.<sup>3</sup> Therefore, the number of structures is continuously growing: in 07/2008, more than 51,663 protein structures have been deposited in the PDB including 44,018 X-ray crystal structures and 7364 structures solved *via* NMR.

Fragment-based approaches have gained momentum and impact as a key discipline in the drug discovery process.<sup>4–6</sup> Fragment hits often exhibit good ligand efficiencies based on high binding affinities per heavy atom and thus are ideal starting points for lead optimization programs resulting in clinical candidates with good drug-like properties.<sup>7</sup> Over the last ten years, fragment-based research programs have provided more than 50 small molecule hits with favorable ligand efficiency values<sup>8</sup> that have been successfully advanced to lead structures.<sup>9</sup>

One simple fragment-based design concept for taking advantage of structural information is to use the known and

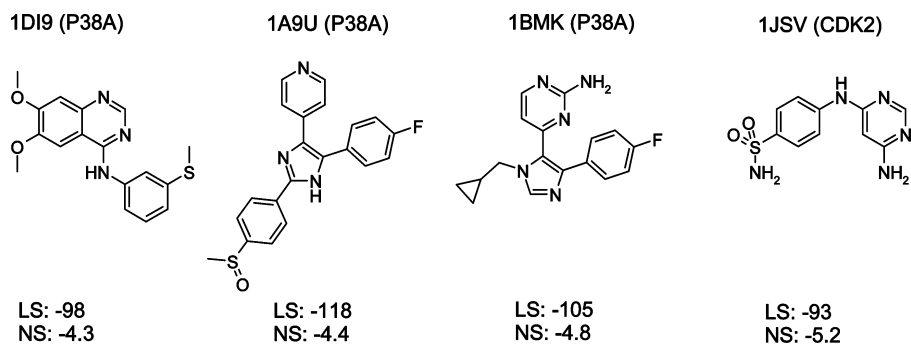
aligned 3D coordinates of two ligands and recombine fragments or substructures at overlapping bonds to generate novel molecules which are hybrids of two fragments. This concept reflects a common design strategy in medicinal chemistry. A computational implementation of this concept is offered by BREED.<sup>10</sup>

Replacing the central structural elements of a molecular scaffold by a new chemotype is the basic idea of scaffold hopping. Different computational methods have already been developed to help with this task, varying in the description of possible replacements, the query input, and the similarity measure used. In contrast to the “mix and match” program BREED, which combines only fragments derived from aligned 3D ligand structures, the tool RECORE<sup>11</sup> allows the scaffold replacement using 3D fragments generated by combinatorial enumeration of cuts of 3D compound libraries. These scaffold replacement approaches are related to fragment-based *de novo* design tools like MED-Hybridise<sup>12</sup> or FLUX.<sup>13,14</sup> FLUX is a 3D ligand-based concept for straightforward *de novo* molecule generation, combining molecular building blocks that match the 3D binding pattern of at least one known reference ligand, whereas the structure-based approach MED-Hybridise takes advantage of both chemical information from the PDB and from chemical supplier databases fitting the interaction surface of the target.

*In silico* fragment libraries are constructed by applying defined fragmentation schemes to known ligand data sets or large compound databases (e.g., the Cambridge Crystallographic Database<sup>11</sup> or the COBRA data set<sup>13,14</sup>). A widely used method for the computational generation of large fragment databases is RECAP,<sup>15</sup> which is based on defined chemistry-derived rules. These calculated libraries display the basis for scaffold replacement techniques as well as for fragment-based *de novo* design methods. Novel molecular

\* Corresponding author phone: +49 (0)214 30 65465; fax: +49 (0)214 30 50070; e-mail: ulrich.rester.ur@bayer-ag.de. Current address: Bayer AG, Corporate Development - Innovation, Kaiser-Wilhelm-Allee, W11, D-51368 Leverkusen, Germany.

<sup>§</sup> Current address: University of Bonn, B-IT Life Science Informatics, Dahlmannstrasse 2, D-53113 Bonn, Germany.



**Figure 1.** 2D structure representation of the four selected protein kinase ligands including corresponding PDB code, ligand score (LS), and normalized score (NS).

structures generated *via* these fragment-based *in silico* methods display excellent starting points for the development of lead structures.<sup>16</sup>

However, *de novo* design tools have to tackle the issue of combinatorial explosion. Therefore, different combinatorial search algorithms are applied to handle this problem: e.g. a stochastic search algorithm is implemented in the program FLUX,<sup>13,14</sup> the program LUDI<sup>17</sup> uses a breadth-first search while the program GROW<sup>18</sup> applies a greedy strategy. These search algorithms are based on a scoring function to assign fitness values to the sample space and to rank the generated molecules, thereby guiding the molecular design process by identifying the most promising compounds.<sup>16</sup>

Herein, we describe a novel and automated *in silico* method to create new compounds with good or even improved binding potential from a set of structurally diverse ligands binding to the same target or target-family. This fragment shuffling concept combines central elements from fragment-based drug discovery and structure-based *de novo* design. Our implemented workflow exploits calculated ligand fragment data sets derived from aligned and scored protein–ligand complex structures by recombining these fragments to novel molecules. The implemented scoring function guides the incremental construction of novel ligands toward ligands with optimal score values to avoid combinatorial explosion.

## 2. METHODS AND MATERIAL

**2.1. Data Sets.** Based on their relevance as potential drug targets in the pharmaceutical industry, we have selected the protein kinases P38A and CDK2 and the serine protease thrombin as model systems for the validation of our fragment shuffling workflow.

Protein kinases are responsible for the transfer of the gamma-phosphate of ATP to the hydroxyl side chains of substrate proteins. They play a key role in the regulation of transcription, in mediating immune responses, in regulating cell growth, and in many other functions.<sup>19</sup> A keyword search within the PDB (performed in 07/2008) results in 135 hits for “CDK2” (including 116 CDK2-inhibitor complex structures) and 77 PDB entries for “protein kinase p38” (comprising 59 protein kinase p38-inhibitor complexes). For an in-depth validation of our fragment shuffling workflow and for a direct comparison with the BREED approach we first focused on the BREED protein kinase reference data set comprising three P38A and one CDK2 protein–ligand X-ray structure. The 2D representation and the corresponding PDB codes of these four ligands are presented in Figure 1.

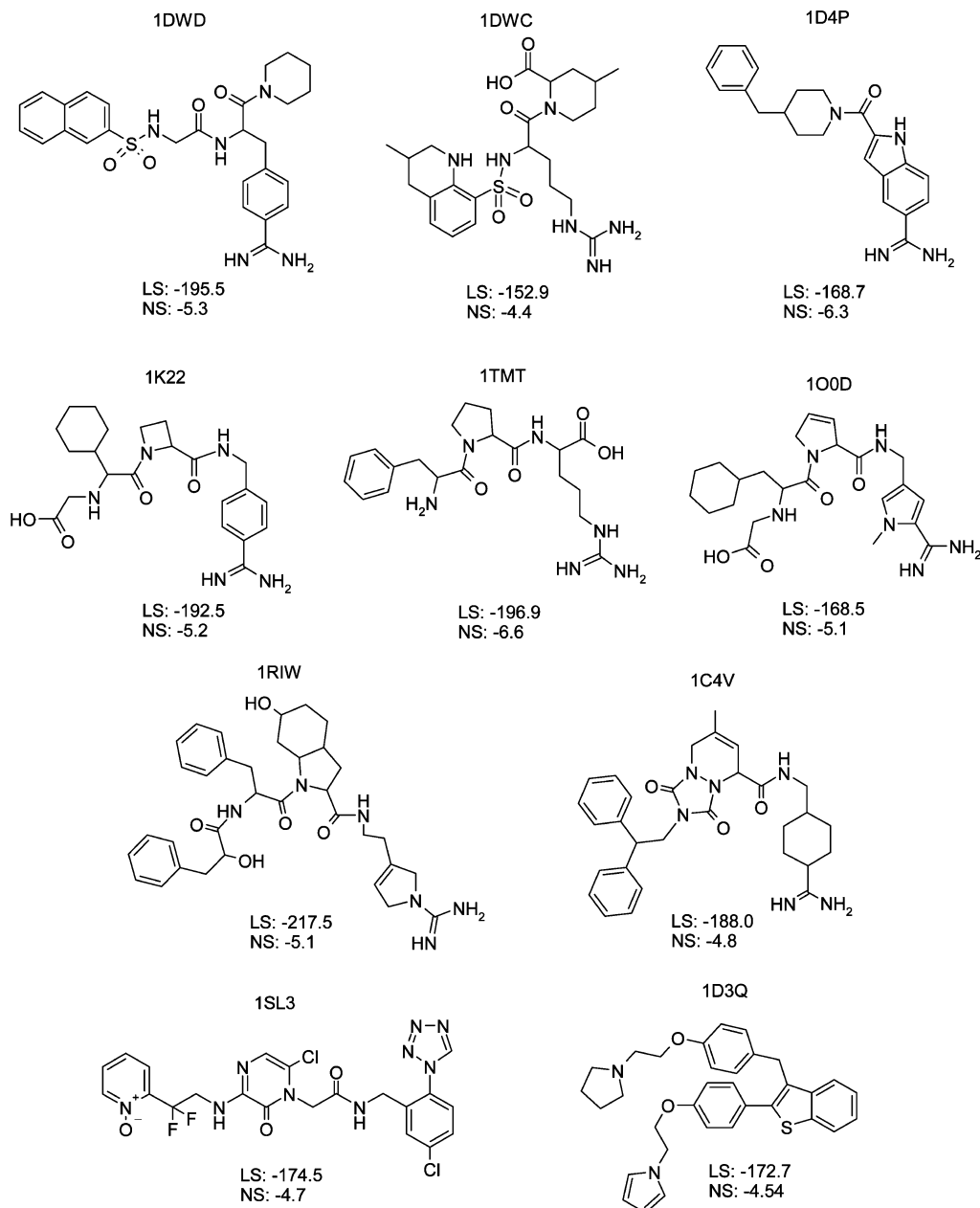
The serine protease thrombin is the central enzyme of the plasmatic blood coagulation cascade.<sup>20</sup> Blood coagulation is part of the physiological response to vascular injury, in which circulating zymogens of serine proteases are sequentially activated by limited proteolysis leading to the formation of a fibrin clot.<sup>21</sup> Within this network of reactions thrombin plays a pivotal role exhibiting procoagulant as well as anticoagulant and antifibrinolytic properties, very specifically interacting with a number of protein substrates, receptors, cofactors, carbohydrates, modulators, and inhibitors.<sup>22</sup> The high incidence of myocardial infarction and cardiovascular disease caused by thrombosis represents a leading cause of morbidity and mortality in the industrialized world.<sup>23</sup> Accordingly, the development of safe and selective thrombin inhibitors as potential antithrombotic drugs is an area of intense interest in the pharmaceutical industry.

An advanced PDB keyword search performed in 07/2008 results in 498 PDB entries for “thrombin”. A refined search considering only thrombin–ligand complexes solved via X-ray crystallography with a resolution below 2.5 Å yields 364 hits. For the prospective application of the fragment shuffling workflow ten structurally diverse thrombin inhibitors addressing different thrombin binding sites were selected. The corresponding 2D representations and the respective PDB codes are summarized in Figure 2.

**2.2. Fragment Shuffling Workflow.** Our overall “Fragment Shuffling” application flow for the identification of novel compounds with comparable or improved binding potential is based on a set of known ligands binding to the same target or target-family and can be subdivided into four steps (compare Figure 3).

**2.2.1. Step 1: 3D Structure Alignment and Protein–Ligand Complex Minimization.** The first step of the application flow comprises the alignment of the protein–ligand complexes to achieve comparable 3D coordinates. Our superimposition is based on the “protein structure alignment” and the “superposition” module of the Maestro modeling environment.<sup>24</sup>

The P38A complexes 1A9U<sup>25</sup> and 1BMK<sup>25</sup> are aligned on the reference structure 1DI9<sup>26</sup> using the default parameter settings of the protein structure alignment software, while the CDK2 protein–ligand complex 1JSV<sup>27</sup> is superimposed using the protein backbone atoms of the hinge regions. The aligned 3D structures of the four protein kinase ligands are shown in Figure 4. The ten thrombin–ligand X-ray structures were aligned based on the protein structure alignment module using default parameter settings.



**Figure 2.** 2D representation of the ten selected thrombin ligand structures including corresponding PDB code, ligand score (LS), and normalized score (NS).

The superimposed X-ray structures are subsequently prepared using Maestro's protein preparation wizard<sup>24</sup> including i) the correct assignment of bond orders, ii) the addition of hydrogens, iii) the optimization of H-bonds via sampling of water orientations, and iv) a forcefield-based minimization of the protein–ligand complex.

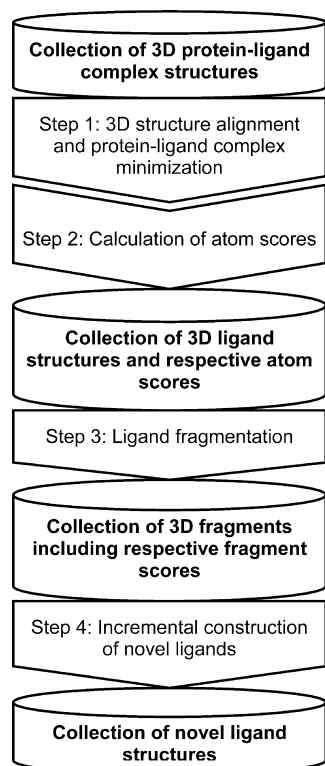
**2.2.2. Step 2: Calculation of Atom Scores.** The second step comprises the atom score calculation for each ligand atom. An atom score reflects the contribution of the specific ligand atom to the overall binding of the ligand to the corresponding target. The atom score calculation was carried out using an adapted version of FlexX<sup>28</sup> kindly provided by BiosolveIT.

The aligned and minimized 3D protein–ligand X-ray structures were exported from Maestro in PDB format<sup>29</sup> and prepared for the atom score calculation via the FlexX GUI.<sup>30</sup> Atom score calculation for all ligands was carried out taking all FlexX energy terms into account which contribute to the

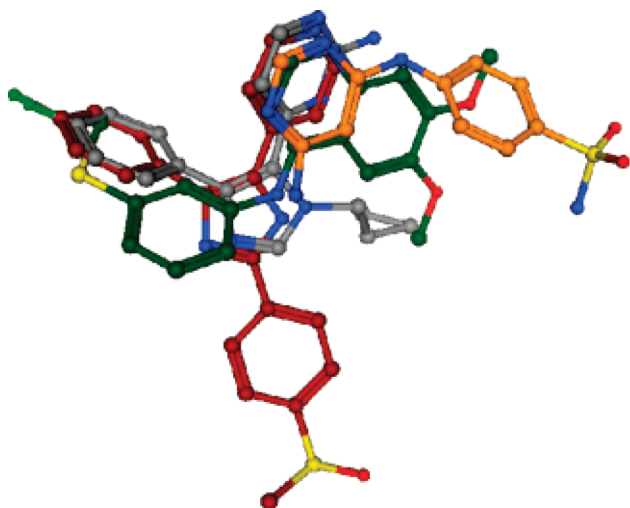
total energy: E\_MATCH, E\_LIPO, E\_ambiguous, E\_CLASH, E\_PLP, E\_SAS, E\_PMF. The results were exported and stored in SD-file format<sup>31</sup> containing the 3D structure information for each ligand atom and its corresponding atom scores. SD-files containing the aligned and scored ligands of the protein kinase and the thrombin data set are available in the Supporting Information.

**2.2.3. Step 3: Ligand Fragmentation and Fragment Score Calculation.** We have developed and implemented three different fragmentation schemes all comprising defined bond splitting rules.

The first and simplest fragmentation method can be summarized as follows: The ligand bond is cut if it is i) a single bond, ii) a nonterminal bond, and iii) an acyclic bond. Terminal bonds are excluded to avoid fragments containing just one heavy atom. Cutting only single bonds prevents loss of information about e.g. double bonds since during the incremental construction of novel molecules, fragments are



**Figure 3.** Overall fragment shuffling workflow for the three-dimensional fragment-based ligand design.



**Figure 4.** Superposition of the aligned and minimized 3D structures of the four protein kinase ligands with PDB codes: 1DI9 (green), 1A9U (red), 1BMK (gray), and 1JSV (orange).

always connected via a single bond. Cyclic bonds are not split to keep ring structures intact.

The second fragmentation method is an extended version of the ligand fragmentation scheme 1. Within this method each acyclic, nonterminal single bond complying with the following three rules is cut: i) it is not an amide bond, ii) it is not an ester bond, and iii) if it is a bond between two carbons, it is only cut if exactly one carbon is part of a ring structure.

The third fragmentation method was developed on the basis of the RECAP approach. RECAP includes 11 fragmentation rules based on retrosynthetic chemistry knowledge. Since the assembly of the fragments in our fragment shuffling workflow is geometrically guided we combined the 11

chemically guided RECAP rules into the four rules of scheme 3. The derivation of the four fragmentation rules from the RECAP rules is available in the Supporting Information. The resulting rules include the cleavage of every acyclic, non-terminal single bond between i) nitrogen and carbon, ii) oxygen and carbon, iii) two aromatic carbons, or iv) nitrogen and sulfur of sulfonamides.

The RECAP approach fragments a molecule by removing one of the 11 bonds, whereby the molecule splits up into non-overlapping fragments. For our purpose we have implemented a modified fragmentation method: A molecule is split into two fragments by assigning the cleaved bond to each of the fragments and therefore keeping the fragments overlapping. Therefore our fragmentation results in the formation of special anchor-atoms playing an important role during the reassembly of the fragments. The anchor-atom and its unique neighboring atom build up the bond that is cleaved during the fragmentation.

The fragments generated for the thrombin ligand with PDB code 1DWD<sup>32</sup> according to the three different fragmentation schemes are presented in Figure 5.

Our fragmentation approach additionally provides the calculation of fragment scores based on the atom scores calculated in step 2. These fragment scores reflect the potential binding affinity of the corresponding fragment and therefore can be used as a scoring scheme to create novel ligands with optimal overall molecule scores and to apply a search heuristic to avoid combinatorial explosion. Fragment scores are determined by summing up all atom scores of non-anchor atoms in a fragment. Anchor atoms are discarded during the calculation of atom scores since they are deleted during the reassembly of the fragments. To achieve comparable scores for different-sized fragments, the fragment scores are normalized according to the respective number of non-anchoring heavy atoms in the fragment.

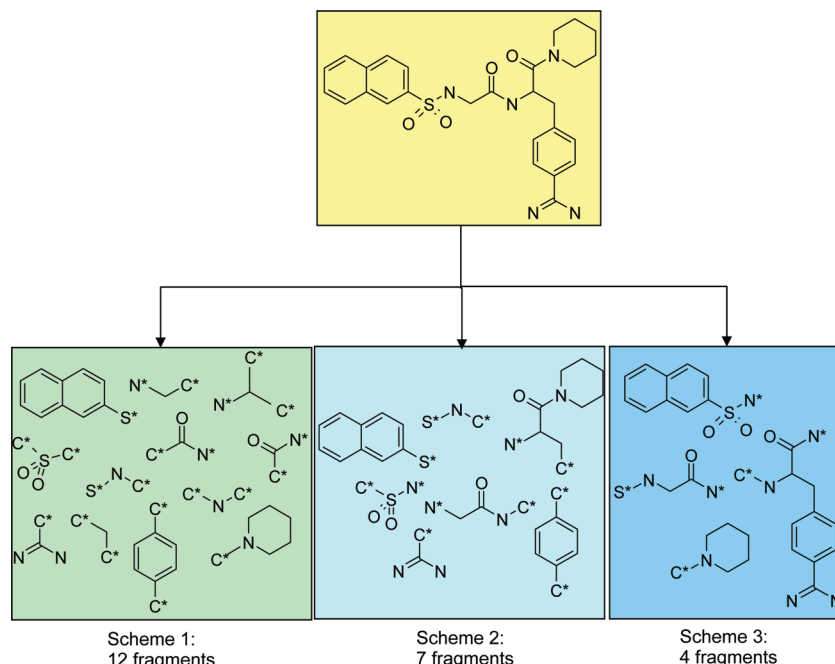
The fragmentation routines and the program for the incremental construction of novel ligands described below (compare chapter 2.2.4) were both implemented in Java (version 1.6.02).<sup>33</sup> The data input and output for both java programs are MDL SDF V2000-format files.<sup>31</sup> For the development of the fragmentation routine the Chemistry Development Kit,<sup>34,35</sup> an open source java library for structural chemo- and bioinformatics, was employed.

**2.2.4. Step 4: Incremental Construction of Novel Ligands.** The fragments derived from the fragmentation routines are used in step four to incrementally construct novel ligands via a tree search. The incremental construction is guided by the calculated fragment scores in order to construct ligands with optimal overall score values. A schematic overview of this iterative ligand construction is presented in Figure 6.

Incremental construction of ligands out of molecular fragments is widely used in computer-aided drug design applications, e.g. by *de novo* design tools like GROW<sup>18</sup> and LUDI<sup>17</sup> or docking methods like FlexX.<sup>36</sup>

In general, each incremental construction method defines linkage rules. Based on these rules the set of connectable fragments is determined. To decide which fragments are connectable, our approach only makes use of the 3D positions of the anchor atoms and their neighboring atoms: Two fragments can be connected if the distance between an anchor atom of fragment A and the neighboring atom of the anchor





**Figure 5.** Fragmentation of the thrombin ligand with PDB code 1DWD using three different fragmentation methods: scheme 1 (left), scheme 2 (middle), and scheme 3 (right). The resulting anchor atoms are labeled with a star.

atom of fragment B is lower than the user defined value of the parameter *max\_dist*.

The distance between the two atoms is calculated via their Euclidian distance

$$dist(x_1, y_1, z_1, x_2, y_2, z_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (1)$$

with  $x_1$ ,  $y_1$ , and  $z_1$  as the 3D coordinates of the anchor atom 1 of fragment A and  $x_2$ ,  $y_2$ , and  $z_2$  as the 3D coordinates of the neighboring atom 2 of fragment B. To link-up two fragments their anchor atoms are deleted, and a single bond between the corresponding neighboring atoms is established.

Our approach accomplishes the incremental construction of novel ligands using a tree search algorithm. Based on a user defined starting fragment, all connectable fragments are determined *via* eq 1. These fragments build up the first layer of the tree search. In this way, for each node in the tree all connectable fragments are calculated and appended as child nodes iteratively until no more fragments can be connected.

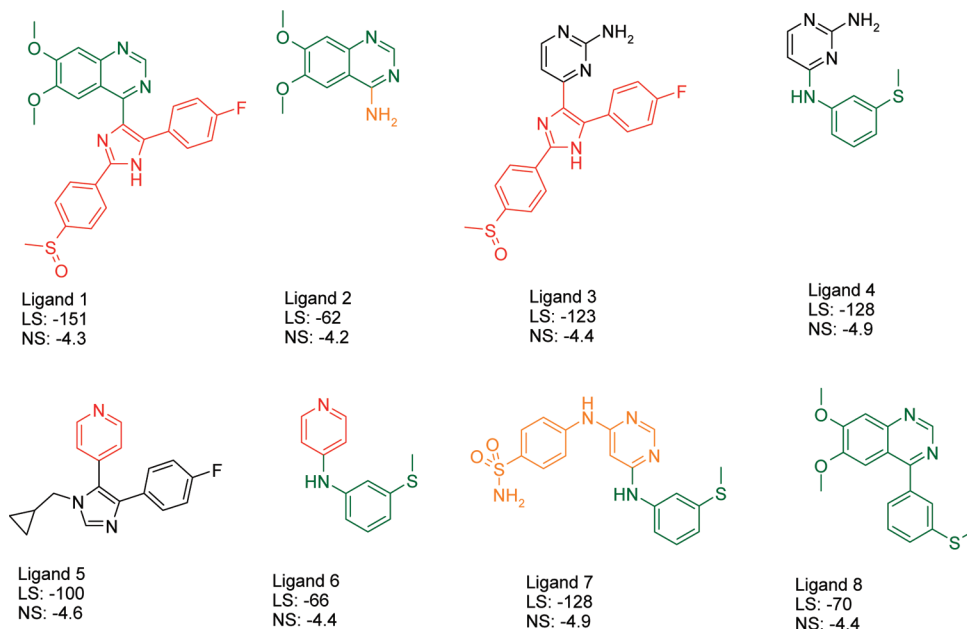
Since the size of the tree expands exponentially with the number of connectable fragment pairs, building up the whole tree structure for large data sets with more than 20 ligands can be resource intensive (i.e., long runtimes and exhaustive memory usage). Therefore, the incremental construction of the tree structure is based on a greedy strategy guided by the fragment scores calculated in step 2 of the workflow. The goal of the greedy strategy is to generate only those paths in the tree corresponding to novel ligands with favorable calculated binding energies as estimated by the summed-up fragment scores. Therefore, for each parent node the fragments with the best fragment scores are appended as child nodes. The maximum number of child nodes per parent node is specified by the user-defined parameter *N\_children*, thereby constraining the width of the tree. Additionally, it is reasonable to constrain the depth of the tree to avoid overly large ligand structures consisting of too many fragments. Therefore, our program provides the

parameter *N\_fragments* determining the maximum number of fragments, which are merged into a new ligand. The tree build-up based on the greedy strategy accounting the parameters *N\_children* and *N\_fragments* can be described as follows (compare Figure 6): Based on a user defined starting fragment all fragments which can be connected with the starting fragment are determined *via* eq 1. Among these fragments the *N\_children* fragments with the best fragment scores are extracted to form the child nodes of the starting fragment. In this way for each parent node in the tree structure, the child nodes are calculated recursively until no more fragments can be appended or until the user-defined depth of the tree given by the parameter *N\_fragments* is reached.

Additionally, the program includes a method to avoid intraligand overlaps: a further fragment may only be added to a path in the tree if it overlaps with none of the existing fragments. To determine whether two fragments overlap, the center of all nonanchor atoms is calculated for one fragment. Subsequently the maximum distance between this fragment center and the corresponding non-anchor atoms is determined. This maximum distance defines the radius of a sphere which is drawn around the center of this fragment. To prevent that two fragments overlap, none of the atoms of the connectable fragment may be located within this sphere.

Within our ligand-design workflow, the choice of the starting fragment has a strong impact on the resulting ligands. Therefore the program offers three different methods for the automatic determination of the starting fragment, i.e. i) the fragment with the highest fragment score (method 1), ii) the fragment with the highest number of connectable fragments (method 2), or iii) the fragment connectable with *N\_children* fragments with the best average fragment score (method 3) is chosen as the starting fragment. The application of these automatic methods to the protein kinase data set is presented below. Additionally, the program offers the option that the user defines the starting fragment manually.





**Figure 7.** Novel ligands designed by the fragment shuffling workflow based on the protein kinase data set. Fragments are color coded according to their respective input molecules with PDB codes 1DI9 (green), 1A9U (red), 1BMK (black), and 1JSV (orange).

molecules with a too high molecular weight. If the number of fragments in a designed molecule reaches the value of the parameter  $N\_fragments$ , the molecular design process is aborted. Therefore, the value for this parameters should not be chosen too small to prevent a pretermed abruption of the molecular design process. Furthermore, the runtime of the program is basically influenced by the parameter  $N\_children$ , whereas increasing the parameter  $N\_fragment$  leads to no exponential explosion of runtime.

As a rule of thumb, the parameter  $N\_fragments$  should never be chosen smaller than the average number of fragments per ligand obtained during the fragmentation of the respective ligand data set.

For our in silico experiments we applied  $N\_fragments = 15$  and  $N\_fragment = 10$  for the thrombin and protein kinase data set, respectively.

$N\_children$ . The parameter  $N\_children$  is crucial in terms of the runtime of the program and the quality of the novel constructed molecules. Therefore, this parameter should be chosen very carefully, especially for large data sets (i.e., more than 25 ligands). The runtime of the program increases exponentially with increasing values for  $N\_children$ . Additionally, the number of newly constructed ligands increases with the parameter  $N\_children$ . However, due to the greedy strategy the overall scores of these ligands do not increase in the same way: With increasing values for the parameter  $N\_children$  fragments with lower fragment scores are attached to the tree. Therefore, novel ligands constructed for large values of  $N\_children$  in general show overall scores inferior to the scores of novel ligands constructed for lower values of  $N\_children$ . In our experience a good choice for the parameter  $N\_children$ , for a data set with up to 20 ligands resulting in less than 150 fragments, is between 10 and 15. For larger data sets with more than 50 initial ligand structures we advise a smaller choice, e.g.  $N\_children = 5$ . If the initial choice of the parameter  $N\_children$  results in too few fragments, this parameter can be increased stepwise to avoid an explosion of the runtime.

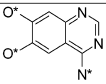
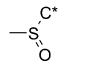
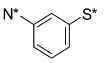

Using the advised values for the parameter  $N\_children$ , the runtime of the program for the incremental construction of novel ligands takes from several minutes up to a few hours.

$max\_dist$ . The parameter  $max\_dist$  defines the maximum distance between two connectable fragments and thereby determines the number of fragments that can be connected. The increase of  $max\_dist$  results in an increase of the number of possible fragment combinations amplifying the number of newly constructed ligands. Since two fragments are connected *via* a single bond, it is reasonable to choose the threshold  $max\_dist$  corresponding to the average length of a single bond. This fact complies with our experience that a maximum distance between two connectable fragments of about 1.2 Å is appropriate. If this choice of  $max\_dist$  results in too few connectable fragments, it is convenient to adjust the threshold up to 1.5 Å, followed by a ligand post-processing step.

**3.3. Retrospective Validation of the Fragment Shuffling Approach.** For an in-depth validation of our 3D fragment-based ligand design workflow and for a direct comparison with the BREED hybridization approach, we selected four kinase ligands (compare Figure 1) coming from the p38 MAP kinase (P38A) and cyclin-dependent kinase 2 (CDK2) crystal structures with respective PDB codes 1DI9, 1A9U, 1BMK, and 1JSV. Based on these four kinase inhibitors BREED generated eight compounds in a first pass ligand hybridization (compare molecules 18–25 in Figure 7 of reference 10).<sup>10</sup>

To validate our fragment shuffling approach we attempted to reconstruct these eight hybridized molecules. Therefore, we first generated three different fragment data sets according to the implemented fragmentation schemes 1–3, resulting in 26, 26, and 15 fragments, respectively (compare Table 1). Fragmentation schemes 1 and 2 generate identical fragment data sets. Due to the fact that ligand 1A9U cannot be split with fragmentation scheme 3, we focused on the 26

**Table 2.** Results of the Fragment Shuffling Approach Applied to the Protein Kinase Data Set Based on Different Starting Fragments<sup>a</sup>

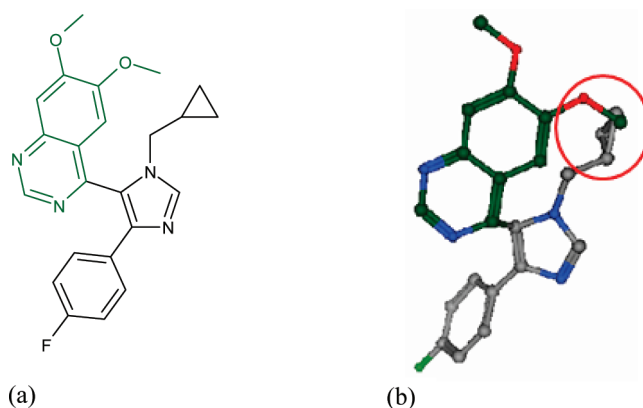
Starting fragment determination method	Fragment ID (PDB code_fragment number)	Fragment structure	Number of connectable fragments	Fragment score	Normalized fragment score	Designed ligands (compare Figure 7)
Method 1 Method 2	1DI9_F1		10	-49	-3.8	Ligand 1 Ligand 6
Method 3	1A9U_F2		2	-3.7	-1.9	Ligand 1 Ligand 8
Manual selection	1DI9_F7		4	-21	-2.1	Ligand 2 Ligand 4 Ligand 5
Manual selection	1A9U_F4		3	-30	-3.0	Ligand 2 Ligand 7

<sup>a</sup> 2D structures and IDs are presented as well as the number of connectable fragments, the fragment scores, the normalized scores, and the ID of the designed ligands.

fragments created via fragmentation scheme 1 for further *in silico* experiments. The fragment structures and their corresponding identifiers are available in the Supporting Information.

In a second step the starting fragments were automatically determined based on the three implemented identification routines. Method 1 (determine the fragment with the highest fragment score) and method 2 (determine fragment with the highest number of connectable fragments) both identified 1DI9\_F1 (compare Table 2), while method 3 (determine fragment connectable with the fragments with the best average fragment score) results in 1A9U\_F2 as the starting fragment. Based on these two root fragments, the incremental construction with  $N_{\text{fragments}} = 10$ ,  $N_{\text{children}} = 10$ , and  $\text{max\_dist} = 1.4$  assembles five molecules comprising three out of the eight hybridized BREED structures (ligands 1–3, Figure 7) as well as the respective 1DI9 and 1A9U seed molecules. While considering all 26 fragments as start structures, in total 12 ligands with unique 2D structures are built up. This set of ligands contains at first the four rebuilt kinase inhibitors. Additionally, seven out of the eight BREED processed hybrids are recreated (ligands 1–7, Figure 7). In addition to the seven recreated BREED structures, ligand 8 is generated only *via* our fragment shuffling approach. Although there is no published data for ligand 8 itself as a kinase inhibitor, it falls within the generic claims of a patent of CSF-1R receptor tyrosine kinase inhibitors<sup>38</sup> and a patent of tyrosine kinase inhibitors exhibiting selectivity for HER2.<sup>39</sup>

The only ligand we were not able to recreate is the 1DI9 (hinge) and 1BMK chimera (Figure 8 (a)), although it should be feasible using starting fragment 1DI9\_F1. However, due to the implemented filter to avoid intraligand overlaps during the incremental construction this chimera is discarded. The visual inspection of the 3D structure of the manually constructed chimera revealed that this ligand in fact shows internal overlaps (compare Figure 8 (b)); from that point of view it was reasonable that our fragment shuffling approach discarded this structure.

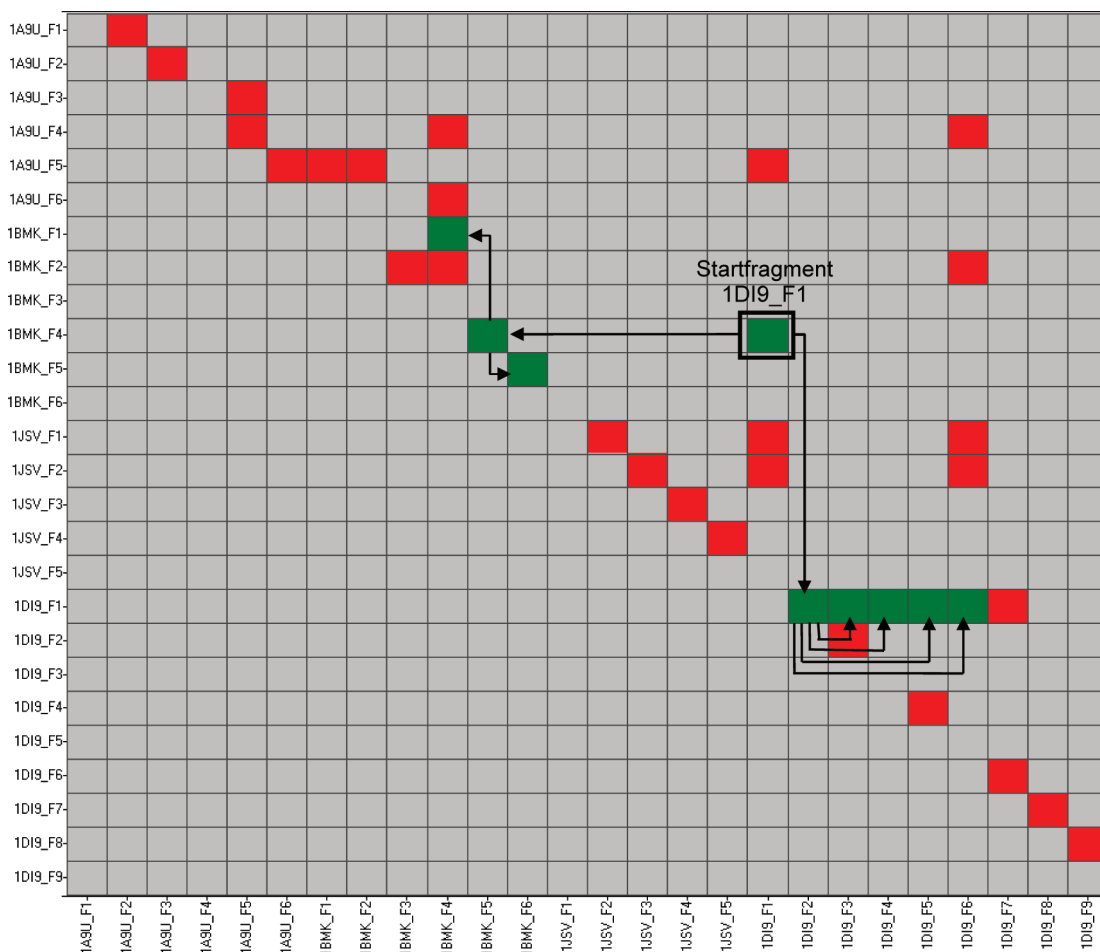


**Figure 8.** 2D representation (a) and 3D structure with overlapping fragments (b) of the 1DI9-1BMK chimera, which was not constructed due to the internal overlaps.

The two-dimensional connectivity matrix in Figure 9 highlights all possible connections derived from the 26 fragments calculated from the four protein kinase seed ligands using fragmentation method 1. The corresponding matrix elements for each pair of connectable fragments are labeled in red. The connections between the fragments of the overlapping 1DI9-1BMK chimera are highlighted in green. Therefore, this matrix encloses combined connectivity information for all potential ligands including the seven reconstructed hybrids as well as for the rejected molecule as indicated by the black arrows between the green matrix elements.

Besides the intraligand overlap filter, the fragment shuffling approach provides a scoring function to evaluate the binding potential of the designed ligands during the incremental construction. The fragment scores calculated during the fragmentation step are summed up for all fragments resulting in a ligand score for the designed molecules. To achieve comparable score values between differently sized ligands, the ligand scores can be normalized by the number of heavy atoms of the respective structure (normalized score).



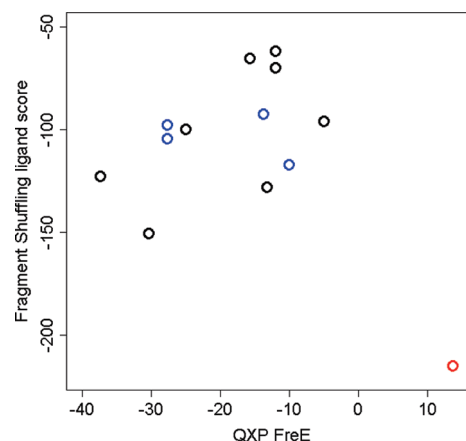


**Figure 9.** Connectivity matrix of the 26 fragments derived from the protein kinase data set. Connectable fragments are highlighted in red, while the connections of the 1DI9-1BMK chimera are shown in green with the potential corresponding molecular linkage indicated as black arrows.

The normalized score ranges from  $-4.3$  to  $-5.2$  for the four input structures and from  $-4.2$  to  $-4.9$  for the shuffled molecules.

However, it has to be kept in mind that a fragment may bind slightly different to the target protein, when it is recombined with other fragments. This fact is not considered in the current fragment-based ligand design setup. Therefore, we have implemented the ligand post-processing step to allow slight adjustments of the connected fragments of the novel ligands within the flexible binding side of the target using a forcefield-based minimization routine implemented in the QXP<sup>37</sup> modeling suite.

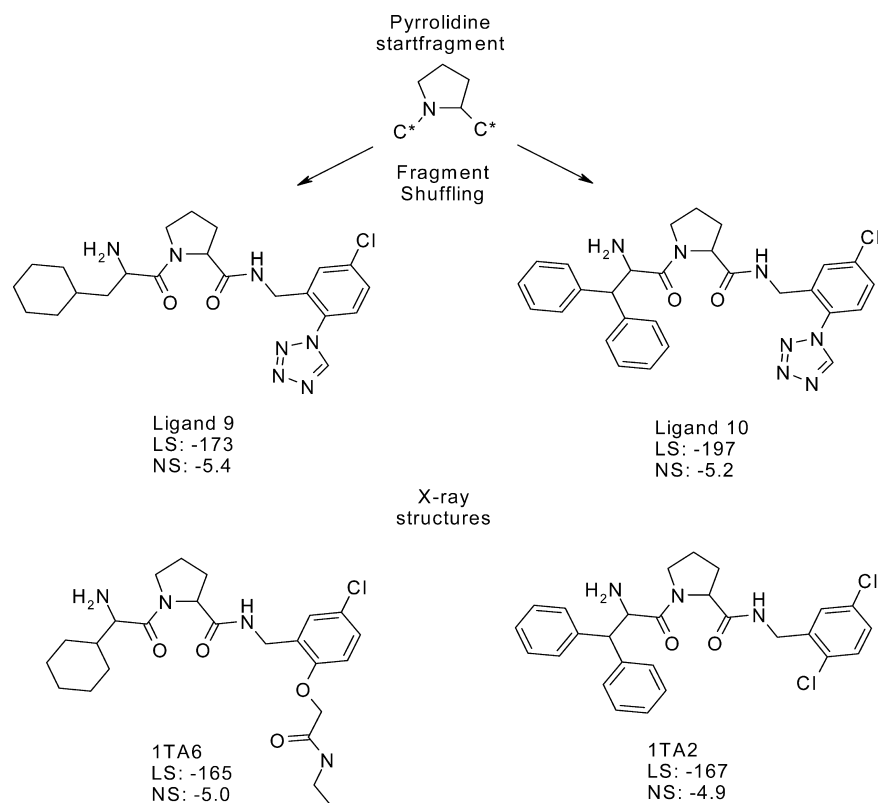
The comparison of these QXP free energy scores (Minimization routine: *dockmin*; Scoring term: FreE) and the fragment shuffling ligand scores can be used to identify outliers with distorted geometries as shown in Figure 10. Both scoring schemes are shown for the four input ligands and the shuffled kinase ligands (ligands 1–8) as well as for the manually built-up 1DI9-1BMK chimera. These two scoring schemes are in good agreement for the constructed ligands 1–8 (shown in black) and the initial protein kinase ligands (shown in blue). However, the manually constructed 1DI9-1BMK chimera, which was rejected due to internal overlaps, yields a high fragment shuffling ligand score but only a low QXP free energy score (Figure 10, highlighted in red). These findings support the assumption that the post-



**Figure 10.** Comparison of the fragment shuffling ligand scores with the QXP FreE scores. The four initial protein kinase ligands are shown as blue, and the eight shuffled ligands are highlighted as black dots. The red dot refers to the manually built-up 1DI9-1BMK chimera.

processing step identifies ligands with distorted geometries and therefore can be used to exclude potential false positives.

**3.4. From Peptides to Small Molecules: Identification of Thrombin Inhibitors.** During the past decades, thrombin has become the primary target for the development of anticoagulant agents for the prevention of thromboembolic disorders, and considerable efforts have been devoted to the



**Figure 11.** 2D representation of selected ligands designed by the fragment shuffling workflow based on the pyrrolidine start fragment and the thrombin fragment data set as well as thrombin inhibitors of two thrombin-ligand X-ray structures with PDB codes 1TA6 and 1TA2 and the calculated ligand scores (LS) and normalized scores (NS).

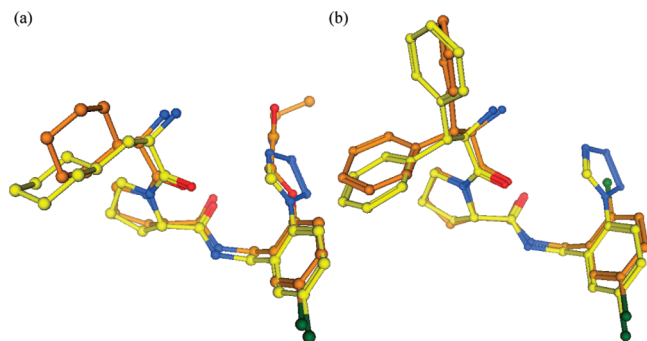
discovery of safe, orally active inhibitors of this enzyme.<sup>40</sup> In addition, more than 350 thrombin X-ray structures in complex with peptides, peptide-like ligands, and druglike molecules have been solved and deposited in the PDB database. Due to this large amount of available 3D structures and the pharmaceutical relevance of the target, we decided to use a set of thrombin inhibitors to further validate the fragment shuffling approach.

The starting point of our search for non-covalent, non-peptidic thrombin inhibitors was a set of 10 diverse thrombin ligands (compare Figure 2) addressing the S1 to S4 specificity pockets in the 3D complex with the serine protease. According to our fragment shuffling workflow, the aligned and scored thrombin ligands were fragmented via fragmentation scheme 1 resulting in 130 fragments. A comparison of the three different fragmentation schemes for the thrombin data set is shown in Table 1. The Supporting Information includes SD-files containing the fragments derived for the thrombin data set *via* the three different fragmentation schemes. Based on the central pyrrolidine core fragment of the peptidic Phe-Pro-Arg ligand (ligand PDB ID: 1TMT, Fragment ID: 1TMT\_5\_1) occupying the S2 site of the target protein and comprising 11 direct connectable fragments ( $max\_dist = 1.2$ ), the incremental construction module ( $N\_fragments = 15$  and  $N\_children = 10$ ) generated 408 unique ligands. A subsequent substructure analysis with regard to the meta-chlorobenzyl motif addressing the S1 pocket identified 65 potential non-basic thrombin ligands with ligand scores ranging from -126 to -228 (average ligand score: -198) compared to -182 as an average ligand score for the 10 thrombin inhibitor input structures. A visual inspection

of the post-processed ligands revealed two interesting chemotypes (compare ligand 9 and ligand 10 in Figure 11). These ligands are built up from fragments of the seed ligands with PDB codes 1O0D, 1TMT, 1RIW and 1SL3 and 1C4V, 1TMT, 1RIW, and 1SL3, respectively. Interestingly ligand 9 is described in a recent patent as a thrombin inhibitor.<sup>41</sup> Although there is no published data for ligand 10 itself, it falls within the claims of a patent of thrombin inhibitors.<sup>42</sup> Additionally, close analogues of ligand 9 and ligand 10 have been co-crystallized with human alpha-thrombin, and the X-ray structures have been deposited within the PDB with the respective PDB codes 1TA2<sup>43</sup> and 1TA6.<sup>43</sup> Besides this tetrazolyl substitution pattern of the P1 phenyl ring (versus chloro (1TA2) or 2-ethylcarbamoyl-ethoxy (1TA6)), which was derived from the fragment data set of the seed ligand 1SL3,<sup>44</sup> the shuffled and co-crystallized ligands are identical. The three-dimensional comparison of the shuffled ligands with the corresponding X-ray structures is presented in Figure 12, demonstrating the good overall fit of the *in silico* designed ligands and the experimental structures. These findings highlight the potential of our fragment shuffling workflow for the identification of novel lead structures.

#### 4. CONCLUSION AND OUTLOOK

We have implemented the fragment shuffling approach, a target-based ligand design workflow for scaffold hopping, comprising elements from both fragment-based drug discovery and structure-based *de novo* design. A set of 3D protein-ligand complexes belonging to a target or target-family displays the basis of our approach. In a first step these



**Figure 12.** 3D structure comparison of the shuffled thrombin ligands 9 and 10 (yellow) with the X-ray structures of two known thrombin inhibitors with respective PDB codes 1TA6 (a) and 1TA2 (b) highlighted in orange.

protein–ligand complexes are aligned and minimized. Secondly, the atom scores for each ligand atom are calculated followed by ligand fragmentation according to one of the three implemented fragmentation schemes and fragment score calculations. Finally, based on these fragments and their corresponding score values, novel ligands displaying hybrids of the input structures showing favorable binding potentials are constructed *via* a greedy strategy using the fragment scores as a scoring function.

Our fragment shuffling approach is related to other fragment-based scaffold replacement tools like BREED and RECORE as well as to the fragment-based *de novo* design applications MED-Hybridise and FLUX. Similar to our fragment shuffling workflow the structure-based approaches MED-Hybridise and BREED make use of 3D target information. However, we have to keep in mind that these structure-based approaches are not restricted to experimentally determined 3D structures (e.g., X-ray crystallography or NMR). They can easily be extended to any kind of computationally generated 3D protein–ligand structures (e.g., ligand docking), whereas the relevance of the novel designed ligands highly depends on the quality of the computationally determined 3D conformation. In contrast to structure-based approaches, the ligand-based approaches FLUX and RECORE are independent of the target protein and only require the 3D information of the query molecule(s) as well as non-target specific fragment databases.

Our fragment shuffling approach displays an extension of BREED since it enables the hybridization of multiple ligands within one iteration step. Furthermore, it includes a scoring function to guide the incremental construction toward ligands with optimal binding potential and offers a method to detect and exclude ligands with internal overlaps. For a direct comparison with the BREED approach, we applied our fragment shuffling approach onto a set of four protein kinase complexes. Based on this inhibitor data set our workflow shuffled seven out of eight BREED generated ligands, excluded one due to intramolecular overlaps and constructed a further protein kinase ligand falling within the generic claims of two patents.

To identify outliers with distorted geometries of the fragment shuffling approach, we compared the calculated scores of the designed protein kinase ligands with scores obtained during the QXP ligand refinement. This cross-scoring step identifies on the one hand true positives (good agreement of the scoring terms) and on the other hand detects

ligands with distorted geometries (poor agreement of the scoring terms) and thereby excludes these potential false positives.

Today, the combinatorial search heuristic within the incremental construction still has to tackle the issue of combinatorial explosion. As the number of input ligands in our workflow increases, the number of calculated fragments increases. For fragment sets comprising more than 200 fragments, the required computing time ascends from hours to several days, requiring additional memory space. Therefore, improvement of the search strategy becomes necessary to allow the application of the workflow to larger input fragment sets. Furthermore, an extension of the incremental construction rules accounting for synthesizability and chemical stability of the designed ligand structures<sup>15,45–47</sup> or the application of defined filter rules to screen out shuffled compounds containing undesirable groups<sup>48,49</sup> can be addressed.

However, the application of the current implementation of the fragment shuffling approach for the development of novel thrombin ligands has shown the potential of this workflow. Based on a central pyrrolidine-like starting fragment several novel non-covalent, non-peptide-like thrombin inhibitors were constructed including two interesting chemotypes falling within the claims of two recent patents of thrombin inhibitors. Additionally, crystal structures of close analogues of these novel ligands in complex with the target protein thrombin are available. A 3D alignment demonstrates the good overall fit of the *in silico* designed ligands and the experimental structures, highlighting the ability of the fragment shuffling workflow to identify novel lead structures with favorable binding affinities.

#### ACKNOWLEDGMENT

The authors would like to thank Holger Claussen and Christian Lemmen (BioSolveIT GmbH) for providing an updated version of the FlexX software package and Rolf Hilgenfeld (University of Lübeck), Hans Briem, Donald Bierer, and Alexander Hillisch for fruitful discussions and critical reading of the manuscript.

**Supporting Information Available:** Derivation of the four RECAP-analog fragmentation rules from the 11 original RECAP rules (Figure S1) and SD-Files containing the aligned and scored 3D structures of the initial protein kinase ligands: Protein\_Kinase\_Data set\_Initial\_Ligands.sdf, aligned and scored 3D structures of the initial thrombin ligands: Thrombin\_Data set\_Initial\_Ligands.sdf, protein kinase inhibitor fragments based on fragmentation scheme 1: Protein\_Kinase\_Data set\_Fragments\_Scheme 1.sdf, protein kinase inhibitor fragments based on fragmentation scheme 2: Protein\_Kinase\_Data set\_Fragments\_Scheme 2.sdf, protein kinase inhibitor fragments based on fragmentation scheme 3: Protein\_Kinase\_Data set\_Fragments\_Scheme 3.sdf, thrombin inhibitor fragment set based on fragmentation scheme 1: Thrombin\_Data set\_Fragments\_Scheme 1.sdf, thrombin inhibitor fragment set based on fragmentation scheme 2: Thrombin\_Data set\_Fragments\_Scheme 2.sdf, and thrombin inhibitor fragment set based on fragmentation scheme 3: Thrombin\_Data set\_Fragments\_Scheme 3.sdf. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- (1) *Chemogenomics in drug discovery*; Kubinyi, H., Müller, G., Eds.; Wiley-VCH: Weinheim, Germany, 2004.
- (2) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O. The Protein Data Bank: A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **1977**, *80*, 319–324.
- (3) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discovery* **2002**, *1*, 45–54.
- (4) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 600–672.
- (5) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J. Med. Chem.* **2004**, *47*, 3463–3482.
- (6) *Fragment-based approaches in drug discovery*; Jahnke, W., Erlanson, W. A., Eds.; Wiley-VCH: Weinheim, Germany, 2006.
- (7) *Fragment-based drug discovery*; Zartler, E. R., Shapiro, M. J. Eds.; Wiley: Chichester, UK, 2008.
- (8) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: A useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.
- (9) Alex, A.; Flocco, M. Fragment-based drug discovery: What has it achieved so far. *Curr. Top. Med. Chem.* **2007**, *7*, 1544–1567.
- (10) Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV Protease. *J. Med. Chem.* **2004**, *47*, 2768–2775.
- (11) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: A fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.* **2007**, *47*, 390–399.
- (12) *MED-Hybridise*; MEDIT SA, 2 rue du Belvédère, 911200 Palaiseau, France, 2008.
- (13) Fechner, U.; Schneider, G. Flux (1): A virtual synthesis scheme for fragment-based de novo design. *J. Chem. Inf. Model.* **2006**, *46*, 699–707.
- (14) Fechner, U.; Schneider, G. Flux (2): Comparison of molecular mutation and crossover operators for ligand-based de novo design. *J. Chem. Inf. Model.* **2007**, *47*, 656–667.
- (15) 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.
- (16) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.
- (17) Böhm, H. J. The computer program LUDI: a new simple method for the de-novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.
- (18) Moon, B. J.; Howe, W. J. Computer-aided design of bioactive molecules: A method for receptor-based de novo design. *Proteins* **1991**, *11*, 314–328.
- (19) Newton, A. C. Protein Kinase C: Structure, function, and regulation. *J. Biol. Chem.* **1995**, *270*, 28495–28498.
- (20) Hanessian, S.; Therrien, E.; Granberg, K.; Nilsson, I. Targeting thrombin and factor VIIa: Design, synthesis, and inhibitory activity of functionally relevant indolizidinones. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 2907–2911.
- (21) Huntington, J. A. Molecular recognition systems of thrombin. *J. Thromb. Haemostasis* **2005**, *3*, 1861–1872.
- (22) Bode, W. Structure and interaction modes of thrombin. *Blood Cells Mol. Dis.* **2006**, *36*, 122–130.
- (23) Semple, J. E. Design and construction of novel thrombin inhibitors featuring P3-P4 quaternary lactam dipeptide surrogates. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2501–2506.
- (24) *Maestro, version 8.0*; Schrödinger Inc.: 101 SW Main Street Suite, 1300 Portland, OR 97204, 2008. <http://www.schrodinger.com/> (accessed Feb 17, 2009).
- (25) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity in MAP kinases. *Structure* **1998**, *6*, 1117–1128.
- (26) Shewchuk, L.; Hassell, A.; Wisely, B.; Rocque, W.; Holmes, W.; Veal, J.; Kuyper, L. F. Binding mode of the 4-anilinoquinazoline class of protein kinase inhibitor: X-ray crystallographic studies of 4-anilinoquinazolines bound to cyclin-dependent kinase 2 and p38 kinase. *J. Med. Chem.* **2000**, *3*, 133–138.
- (27) Clare, P. M.; Poorman, R. A.; Kelley, L. C.; Watenpaugh, K. D.; Bannow, C. A.; Leach, K. L. The cyclin-dependent kinases cdk2 and cdk5 act by a random, anticooperative kinetic mechanism. *J. Biol. Chem.* **2001**, *276*, 48292–48299.
- (28) *FlexX, version 2.2.0 (pre)*; Biosolve IT GmbH, An der Ziegelei 75, 53757 Sankt Augustin, Germany, 2007.
- (29) Protein Data Bank, PDB File Format, 2000. <http://www.wwpdb.org/docs.html> (accessed Feb 17, 2009).
- (30) *FlexX GUI, version 1.0*; Biosolve IT GmbH, An der Ziegelei 75, 53757 Sankt Augustin, Germany, 2007.
- (31) *MDL SDF V2000 format*; Symyx Technologies, Inc., San Ramon, CA, U.S.A., 2005.
- (32) Banner, D. W.; Hadvary, P. Crystallographic analysis at 3.0 - A resolution of the binding to human thrombin of four active site-directed inhibitors. *J. Mol. Biol.* **1991**, *266*, 2085–2093.
- (33) *Java, version 1.6.02*; Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, CA 95054, 2007.
- (34) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source java library for chemo and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- (35) *Chemistry Development Kit CDK, version 1.0.1*; 2007. <http://sourceforge.net/projects/cdk/> (accessed Feb 17, 2009).
- (36) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
- (37) McMartin, C.; Bohacek, R. S. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.
- (38) Egeröd, L. Quinazoline compounds which inhibit CSF-1R receptor tyrosine kinase. EP 1488792, 2005 <http://www.freepatentsonline.com/EP1488792A3.html> (accessed Feb 17, 2009).
- (39) Myers, M. R.; Spade, A. P.; Maguire, M. P.; Persons, P. E. Protein tyrosine kinase aryl and heteroaryl quinazoline compounds having selective inhibition of HER-2 autophosphorylation properties. U.S. Patent 5721237, 1998. <http://www.patentstorm.us/patents/5721237.html> (accessed Feb 17, 2009).
- (40) Srivastava, S.; Goswami, L. N.; Dikshit, D. K. Progress in the design of low molecular weight thrombin inhibitors. *Med. Res. Rev.* **2005**, *25*, 66–92.
- (41) Selnick, H. G.; Young, M. B.; Nantermet, P. G.; Barrow, J. C.; Williams, P. D.; Lyle, T. A.; Staas, D. D.; Stauffer, K. J.; Sanderson, P. E. Preparation of amino acid benzylamide and 2-pyridylmethylamide derivatives as thrombin inhibitors. WO 02/064559, 2002. <http://www.wipo.int/pctdb/en/wo.jsp?wo=2002064559> (accessed Feb 17, 2009).
- (42) Lumma, W. C.; Tucker, T. J.; Witherup, K. M.; Brady, S. F.; Whitter, W. L.; Vacca, J. P.; Coburn, C. Thrombin inhibitors. WO 97/15190, 1997. <http://www.freepatentsonline.com/WO1997015190.html> (accessed Feb 17, 2009).
- (43) Tucker, T. J.; Brady, S. F.; Lumma, W. C.; Lewis, S. D.; Gardel, S. J.; Naylor-Olsen, A. M.; Yan, Y.; Sisko, J. T.; Stauffer, K. J.; Lucas, B. Y.; Lynch, J. J.; Cook, J. J.; Stranieri, M. T.; Holahan, M. A.; Lyle, E. A.; Baskin, E. P.; Chen, I.-W.; Dancheck, K. B.; Krueger, J. A.; Cooper, C. M.; Vacca, J. P. Design and synthesis of a series of potent and orally bioavailable noncovalent thrombin inhibitors that utilize nonbasic groups in the P1 position. *J. Med. Chem.* **1998**, *41*, 3210–3219.
- (44) Young, M. B.; Barrow, J. C.; Glass, K. L.; Lundell, G. F.; Newton, C. L.; Pellicore, J. M.; Rittle, K. E.; Selnick, H. G.; Stauffer, K. J.; Vacca, J. P.; Williams, P. D.; Bohn, D.; Clayton, F. C.; Cook, J. J.; Krueger, J. A.; Kuo, L. C.; Lewis, S. D.; Lucas, B. J.; McMasters, D. R.; Miller-Stein, C.; Pietrak, B. L. Discovery and evaluation of potent P1 aryl heterocycle-based thrombin inhibitors. *J. Med. Chem.* **2004**, *47*, 2995–3008.
- (45) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug. Discovery Des.* **1995**, *3*, 34–50.
- (46) Vinkers, H. M.; de Jonge, M. R.; Daevaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmermann, H.; van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPTimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765–2773.
- (47) Baber, J. C.; Feher, M. Predicting synthetic accessibility: application in drug discovery and development. *Mini-Rev. Med. Chem.* **2004**, *4*, 681–692.
- (48) Wunberg, T.; Hendrix, M.; Hillisch, A.; Lobell, M.; Meier, H.; Schmeck, C.; Wild, H.; Hinzen, B. Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discovery Today* **2006**, *11*, 175–180.
- (49) Lobell, M.; Hendrix, M.; Hinzen, B.; Keldenich, J.; Meier, H.; Schmeck, C.; Schohe-Loop, R.; Wunberg, H.; Hillisch, A. In silico ADMET traffic lights as a tool for the prioritization of HTS hits. *ChemMedChem* **2006**, *1*, 1229–1236.

CI8004572