DOI: 10.1002/cmdc.200500102

FlexNovo: Structure-Based Searching in Large Fragment Spaces

Jörg Degen and Matthias Rarey*[a]

We present a new molecular design program, FlexNovo, for structure-based searching within large fragment spaces following a sequential growth strategy. The fragment spaces consist of several thousands of chemical fragments and a corresponding set of rules that specify how the fragments can be connected. FlexNovo is based on the FlexX molecular docking software and makes use of its incremental construction algorithm and the underlying chemical models. Interaction energies are calculated by using standard scoring functions. Several placement geometry, physicochemical property (drug-likeness), and diversity filter criteria are directly integrated into the "build-up" process. FlexNovo has been used to design potential inhibitors for four targets of pharmaceutical interest (dihydrofolate reductase, cyclin-dependant kinase 2, cyclooxygenase-2, and the estrogen receptor). We have carried out calculations using different diversity parameters for each of these targets and generated solution sets containing up to 50 molecules. The compounds obtained show that FlexNovo is able to generate a diverse set of reasonable molecules with drug-like properties. The results, including an automated similarity analysis with the Feature Tree program, indicate that FlexNovo often reproduces structural motifs as well as the corresponding binding modes seen in known active structures.

Introduction

The examination of a reasonable amount of chemical space is crucial for the identification of possible lead structures for the target of interest at the beginning of a drug discovery project. Until now, increasingly large experimental and in silico screening of compounds has been used to address this problem. However, the chemical space that these methods cover is small compared to the available compounds, estimated to be between 10^{60}[1,2] and 10^{100} molecules.[3] Evidently the compounds in this chemical space cannot be enumerated by any experimental or computational technique, therefore different approaches are needed to address this challenge.

An alternative to compound libraries is the use of "fragment spaces", which inherently cover a larger part of the chemical universe.[4–6] For reasons of efficiency these too cannot be enumerated, so programs are needed that exploit their combinatorial nature. Therefore, we have developed a new structure-based molecular design program, FlexNovo, to address this need.

The concept of fragment-based drug or lead discovery is not new. Several publications have dealt with this subject recently; though, within the experimental methods described, the term "fragment" is used slightly differently, often representing a small chemical molecular entity.[7–11] By using experimental techniques such as high-resolution X-ray crystallography or highly sensitive NMR screening, small weak-binding fragments can be identified. These fragments are then, if possible, developed into lead candidate structures.

There are many computational approaches available that use fragment spaces, or similar concepts, to generate either analogues to a ligand structure or molecules complementary to a given protein's active site. The first programs to perform such tasks were published in the late 1980s,[12] since when a multitude of de novo design methods have evolved (recently reviewed by Schneider et al.[13]). Here a few are briefly described in order to point out the challenges still to be addressed and to place our approach in the context of existing ones.

Many of the early programs for de novo design use generalized molecular structures, so called "template structures". These are combined to create molecular skeletons that satisfy the steric constraints of a given receptor. Atom-type assignment is then carried out on these molecular skeletons to generate molecules that have chemical properties complementary to the active site. Programs adopting this approach differ in the way they combine templates and the type of method they use for the treatment of conformational flexibility. Two programs that work in a stepwise manner are, for example, SPROUT,[14–16] which is based on graph-searching algorithms, and Skelgen,[17,18] which uses a Monte Carlo-type method.

Other approaches fit molecular fragments, functional groups, or atoms onto previously determined interaction sites within the protein cavity. One possibility is then to connect these by using "bridge fragments". One of the first approaches to make use of this technique was LUDI,[19,20] but there are a number of similar programs.[21–25] Alternatively, starting fragments, building blocks, or atoms can be extended sequentially. Such incremental methods were originally invented for de novo design, and are used, for example, in the programs GROW[26] and GroupBuild.[27]

[a] Prof. Dr. J. Degen, Dr. M. Rarey
Center for Bioinformatics (ZBH), University of Hamburg
Bundesstrasse 43, 20146 Hamburg (Germany)
Fax: (+49) 40-42838-7352
E-mail: rarey@zbh.uni-hamburg.de

Synthetic accessibility has always been, and remains, one of the key issues in computational de novo design.[28] However, a proposed lead or drug candidate must also satisfy a number of equally important prerequisites, such as pharmacological activity and ADMET properties,[29,30] which correlate, to some extent, to certain physicochemical properties.[31–35]

One way to incorporate synthetic accessibility in a de novo design approach is to encode it directly by choosing appropriate building blocks and synthesis rules. The (valid) assumption is that compounds created from these building blocks will have a high probability of being synthetically tractable.[36–44]

Lewell et al. first generated a fragment space by applying a "retrosynthetic combinatorial analysis procedure" (RECAP) to a collection of biologically active molecules from the Derwent World Drug Index (WDI) and thus obtained a set of approximately 30000 corresponding fragments.[4] By adapting these, Schneider et al. developed the ligand-based structure-generating program TOPAS,[5] and Rarey et al. used the "Feature Tree" descriptor[45] for similarity searching in large combinatorial chemistry spaces.[6] Stahl et al. used the structure-based de novo design program Skelgen with a set of approximately 600 hand-selected ring and linker fragments in a study to validate the automated use of structure-based de novo design.[46]

By contrast, the FlexNovo approach presented here treats fragments in a more comprehensive way. It works directly on the fragments themselves, rather than using generalized molecular structures such as templates or skeletons, and uses large fragment spaces (up to several thousand fragments). FlexNovo has a flexible fragment-placement engine, employing the underlying incremental construction algorithm of FlexX.[48] The underlying chemical model allows for an accurate treatment of fragments in terms of the description of their molecular interactions and conformational flexibility. The deterministic approach of FlexNovo, together with the use of well-defined connection rules, facilitates a comprehensive and user-customizable virtual synthesis of potentially bioactive molecules. Additionally, a multitude of physicochemical property ("drug-likeness"), placement geometry, and diversity filter criteria have been integrated into the fragment build-up process. This allows the algorithm to generate solutions fulfilling drug-likeness criteria in addition to offering prediction of reasonable binding orientations.

First we summarize the underlying chemical models and algorithms of FlexNovo (including the relevant parts of FlexX), and describe FlexNovo itself: the build-up algorithm and the physicochemical property, diversity and geometry filter implementations. Then we describe and discuss the results of the calculations we performed for four different targets. Finally we discuss key aspects of the calculations and the main technical issues.

Methods

This section describes the software kernel and the chemical models FlexNovo uses, followed by a conceptual and technical description of the type of fragment spaces the program deals with. Subsequently, we describe FlexNovo itself by giving a detailed outline of the main technical aspects, the underlying search algorithm, and the physicochemical property, diversity, and pose geometry filter functionality available.

The chemical models, scoring functions, algorithms, and pharmacophore models used in the FlexX docking program (and its extension, FlexX-Pharm) have been detailed previously.[47–50] Here an outline of those parts most relevant to FlexNovo is given.

FlexX is a program for semi-flexible protein–ligand docking in which multiple ligand conformations are examined while the protein structure is kept rigid. FlexX assigns interaction types and geometries to the atoms of the active site and the ligand. Each interaction group has an interaction type and an interaction center. The interaction types used in FlexNovo are listed in Table 1. The interaction center is the center of a sphere on which the interaction surface lies. FlexX forms interactions by matching interaction surfaces of compatible ligand and receptor interaction groups. If the interaction center of one group lies close to the interaction surface of the other, this interaction is considered to have an ideal geometry. FlexX tolerates some deviations from the ideal geometry and penalizes these in the scoring function.

Table 1. Interaction types used in FlexNovo.

Name	Interaction Group	Counter Groups
Directed interactions (level 3 interactions—strong)		
h_don	hydrogen bond donor	h_acc
h_acc	hydrogen bond acceptor	h_don
metal	metal atom	metal_acc
metal_acc	metal ligand	metal
"Hydrophobic directed interactions" (level 2 interactions)		
phenyl_center	phenyl ring (centroid)	phenyl_center, ch3_phe, amide
phenyl_ring	atom in phenyl ring	phenyl_center
ch3_phe	methyl group	phenyl_center
amide	amide bond	phenyl_center
"Hydrophobic undirected interactions" (level 1 interactions—weak)		
ch	CH group	ch, ch2, ch3, sulfur, aro
ch2	CH_2 group	ch, ch2, ch3, sulfur, aro
ch3	methyl group	ch, ch2, ch3, sulfur, aro
sulfur	sulfur atom	ch, ch2, ch3, sulfur, aro
aro	aromatic C atom	ch, ch2, ch3, sulfur, aro

FlexX divides the ligand into different components at each acyclic rotatable bond. For each component, a set of energetically favorable conformations is computed that are derived from a torsion-angle database. Combinations of different numbers of adjacent components form "base fragments". These are placed independently into the active site by considering multiple orientations and conformations. The ligand structure is built up by iterative addition of the remaining components. In each of these incremental construction steps, FlexX searches for new interactions and uses them to optimize the ligand position and conformation in the active site. Each placement is then assessed by an empirical scoring function, for example SCORE,[51] ScreenScore,[52] or PLP.[53] A greedy heuristic is used to retain the *k* best placements for the next incremental construction step.

FlexNovo also uses the Pharm extension[50] of FlexX that allows for pharmacophore-type constraints, such as interaction constraints and spatial constraints, during the docking calculation. For the first of these, the user can specify an interaction surface for an active-site atom that has to form an interaction with the ligand. For the second, the user can specify a volume in the active site and a SMARTS (www.daylight.com) expression for that volume that has to be satisfied by each ligand placement. The user can combine constraints by using logical expressions, which are evaluated for every placement of the ligand at each incremental construction step. Placements that fail to satisfy the pharmacophore constraints are discarded. Since these constraints can at the moment only be used for the initial fragment placement, they do not represent a complete pharmacophore definition, and are thus termed "anchor pharmacophores" below.

From molecules to fragment spaces

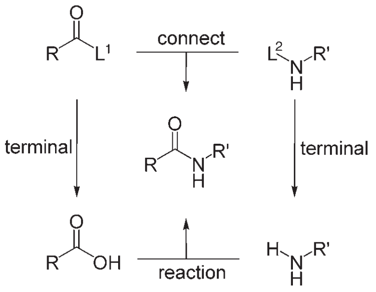
Unlike FlexX, which deals with one molecule at a time, FlexNovo uses fragment spaces. Essentially, the "atomic" elements of a fragment space are chemical fragments, each of which has one or more open valences, represented by "link" (dummy) atoms. Each link atom has a "type" representing one of the different chemical environments. The fragments span the fragment space according to the definition of the compatibility of the link types. For each pair of link types, an entry in a (symmetrical) compatibility matrix dictates whether or not this pair can be connected. Additionally, properties of the connecting bond (bond type, length, and a preferred initial torsion angle) and the connected atoms (hybridization) can be defined. Finally, for each link type, a "terminal group" has to be specified, which is a substitute for a link atom that has not been involved in fragment connection. Scheme 1 illustrates the two basic operations of a fragment space: fragment connection and link substitution.

In the context of FlexNovo, we distinguish between two types of structures resulting from the connection of two or more fragments; if the result of such a connection contains additional link atoms it is still termed a "fragment" for consistency; otherwise, it will be called a "molecule". In general, the link atoms involved in the connection of two fragments are removed upon their connection. If the difference between a fragment resulting from fragment connection and an initial ("atomic") fragment is relevant, it will be mentioned explicitly.

Fragments have been generated here in the same way as was done for the two approaches described above. Therefore, the number and types of the corresponding links and the basic properties of the space are the same.[6] Additionally, a couple of filtering steps are performed to generate a more application-tailored version of the fragment space (see below).

The FlexNovo program

FlexNovo is a structure-based molecular-design software package that operates on large fragment spaces. The input to the program consists of the definitions of the protein active site, the anchor-pharmacophore constraints (see above), and the fragment space. Furthermore, three sets of filter criteria may be specified to define certain compound property, pose geometry, and diversity criteria. These filters are mentioned during the description of the algorithm and are explained in detail in a separate paragraph. The result of a FlexNovo calculation is a collection of compounds that meet certain physicochemical property and diversity criteria. For each molecule contained in this collection, a number of different orientations (and conformations) in the active site, "placements", are generated and a score is calculated. Both the compounds and the placements can be saved to disk in standard data formats.



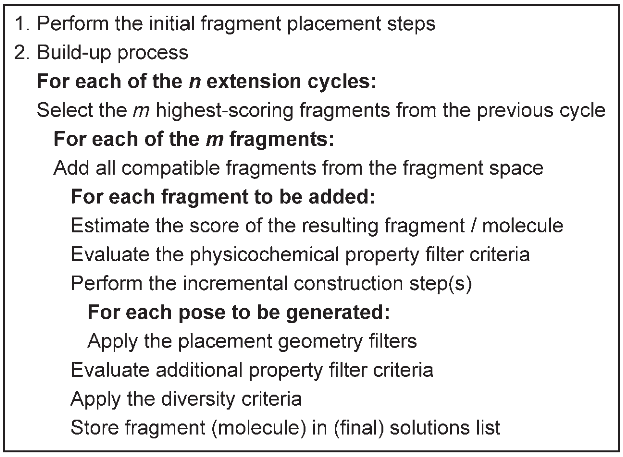
Scheme 1. Schematic illustration of two hypothetical fragments containing link atoms (L^1, L^2) and the results of connecting these fragments or of substituting the link atoms with the corresponding terminal group. For the latter, the chemical reaction modeled with the two link atoms above is shown.

The FlexNovo algorithm

A FlexNovo calculation consists of a preprocessing phase and a "build-up" phase. The preprocessing phase consists of two different docking calculations for all fragments. The first is performed without constraints and serves to estimate the highest possible score for each fragment according to the scoring function used for a particular receptor. (This information is stored on disk and used later in the build-up process.) The second calculation is done by using pharmacophore-type constraints for generating docking solutions for all fragments that are able to fulfill these constraints. The "placements" are used as the starting positions in the build-up process.

In the build-up phase, a fixed number of "extension cycles" is carried out. In each extension cycle, the fragments with the best scores from the previous cycle are identified, except for the first cycle, here the highest scoring initial (atomic) fragments from the docking calculation of the preprocessing step are selected. To every set of placements of each of these fragments all initial fragments from the fragment space are added exhaustively; that is, all possible combinations are generated as determined by the link compatibility definitions. In each fragment connection step, the algorithm considers multiple conformations for each fragment. During fragment connection, all possible stereoisomers (if any) for the resulting fragment are generated. Additionally, the algorithms account for fragment symmetry and structurally equal fragments to eliminate duplicate fragments or molecules.

Before each potential fragment is generated, its best achievable score is estimated. This is done by adding the value of the highest scoring placement of the fragment to be extended and the maximum score of the initial fragment to be added (derived in the preprocessing step or estimated on the fly). If a minimum number of fragments has already been generated (equal or more than the size of the partial solution queue) and the estimated maximum score is less than that of the worst scoring fragment, the new potential fragment is discarded. Additionally, a couple of property filter criteria are checked. If the potential fragment fulfills all these criteria, the initial fragment is added by performing an incremental construction step, applying the FlexX algorithm to all stored placements. During this step, the algorithm checks for certain pose-geometry properties, optimizes the orientation and conformation, and creates a set of corresponding placements for the fragment. This is followed by the evaluation of a second set of filter criteria, which accounts for additional fragment properties and provides diversity in the solution set for each extension cycle and in the final results list. If the result of the connection does not contain any additional link atoms, it cannot be extended further and is therefore put into the solution list; otherwise it is put into the partial solution queue for the next extension cycle. The size of each queue is variable, but limited to minimize memory usage. Upon completion of the final extension cycle, all results generated are placed into the final solution list. All open links are replaced by the corresponding terminal group at this point (if so specified by the user). This procedure involves another incremental construction step in which the algorithm searches for additional interactions of the terminal groups, optimizes the whole molecule orientation, conformation and score, and performs a clustering of placements that are very similar with respect to their RMSD. By default, FlexNovo does not perform any further optimization of the molecule; in particular, no force field minimization or molecular-dynamics simulation is carried out. Therefore, the predicted binding orientations of the results are just those generated by the build-up algorithm. Figure 1 contains a simplified illustration of the FlexNovo build-up algorithm described above.



```
1. Perform the initial fragment placement steps
2. Build-up process
   For each of the n extension cycles:
   Select the m highest-scoring fragments from the previous cycle
     For each of the m fragments:
     Add all compatible fragments from the fragment space
       For each fragment to be added:
       Estimate the score of the resulting fragment / molecule
       Evaluate the physicochemical property filter criteria
       Perform the incremental construction step(s)
         For each pose to be generated:
         Apply the placement geometry filters
       Evaluate additional property filter criteria
       Apply the diversity criteria
       Store fragment (molecule) in (final) solutions list
```

Figure 1. Main steps of the FlexNovo build-up algorithm.

FlexNovo **property filters**

In order to obtain more chemically reasonable results, several filter criteria are available. These can be used for an initial fragment prefiltering step and during the build-up calculations. The user can specify property ranges for molecular weight, the number of hydrogen-bond donors and acceptors, rings, nonterminal single bonds, tetrahedral stereocenters, molecular log *P*, refractivity values, and polar solvent-accessible surface (PSAS). For efficiency reasons, the calculation of these values is done directly within the program. The number of hydrogen bonding atoms is calculated as described by Lipinski et al.[31] The log *P* and refractivity values are calculated by using the atom substructure descriptors described by Wildman and Crippen.[54] The "number of rings" is the number of ring closures (#bonds−#atoms+1), and the PSAS is derived as fraction of the solvent-accessible surface (SAS) of a fragment or molecule. The user can also specify sets of inclusion and exclusion SMARTS expressions. This is probably most useful for prefiltering in order to exclude certain reactive or toxic molecular substructures, but can also be used during the build-up algorithm, for every possible intermediate solution as well as for the final results. Thus, the user can, for example, prevent the formation of unwanted chemical motifs or, conversely, request specific chemical groups.

The assumption when using the filter criteria is that all these values can be approximated, for any fragment, by summing the respective values of the initial (atomic) fragments of which it consists. Obviously, this is the case for molecular weight, number of rings, nonterminal single bonds, hydrogen-bonding atoms (according to the Lipinski approach), and an approximated PSAS calculation. Consequently these filters can be checked before actually constructing a fragment. In contrast, the log *P* and molar refractivity values and the number of tetrahedral stereocenters do not show a strictly additive behavior. Therefore, this part of a filter definition is only applied to molecules in the final solution list.

FlexNovo **diversity filters**

The fragments and molecules assembled by FlexNovo tend to be very similar. As the build-up procedure is based on a *k*-greedy approach, the number of fragments that the algorithm retains at the end of each extension cycle is limited compared to the number of fragments available in the fragment space. Therefore, FlexNovo contains two types of diversity measurements based on the number of initial (atomic) fragments they have in common.

The user can specify the maximum number of common fragments, or alternatively a minimum number of different fragments, for each pair of molecules in the final solution list. A similar approach has been used for Feature Tree fragment spaces.[6] Thus the occurrence of few high-scoring initial (atomic) fragments can easily be limited. In addition, the user can limit the number of total occurrences of each of these fragments for the partial-solution queue of an extension cycle. This applies a diversity criterion right from the beginning of the calculations.

Before a newly placed fragment is added to the solution list, it is checked against all other entries in the solution set in conjunction with the diversity filter criteria. Where there are conflicting entries, and the score of the new fragment is better than the conflicting ones, this is added to the solution set and the conflicting entries are removed. Otherwise the new fragment or molecule is discarded.

FlexNovo pose-geometry filters

FlexNovo also incorporates a set of pose-geometry filters, since the orientation in the active site of the receptor is of at least equal importance and should fulfill a number of criteria. The user is able to switch on a number of filter functions that account for improved pose geometries. The first two (described below) are part of the FlexX software package, so are described here briefly for completeness.

Polarity filter: A realistic fragment orientation should avoid close contact of charged polar atoms. If the scalar distance of two atoms of the fragment (or molecule) and the receptor is below a certain limit, and these atoms are not involved in hydrogen-bonding interactions, this placement will be removed from the pose set of this molecule.[55]

Repulsion filter: Another aspect is steric repulsion of fragment atoms beyond standard atom distance criteria. This can occur, for example, if atoms are close to each other due to a series of *cis*-oriented single bonds. A filter function checks for these "1–5 repulsions" and removes such placements.

Buriedness filter: Hydrophilic atoms of a fragment or molecule that are not involved in hydrogen-bonding interactions are checked against proximal receptor atoms. If a fragment atom is placed within a hydrophobic pocket of the receptor, but does not form polar interactions, the corresponding pose is discarded.

Saturation filter: This filter is included to avoid the generation of fragments or molecules containing several polar atoms that do not form interactions. If the ratio of polar atoms forming interactions to polar atoms that do not form interactions drops below a threshold specified by the user, the corresponding placement is discarded. It is important to note that a single interaction per atom is sufficient for this estimation, even if an atom is capable of forming more than one interaction at a time.

Clash filter: This filter is similar to the polarity filter but is used only for hydrogen atoms. If the scalar distance between two hydrogen atoms of a fragment or molecule and the receptor is below a certain limit, the corresponding pose will be deleted from the solution set.

Data preprocessing

The preprocessing consists of two docking calculations for all the fragments of the fragment space. One calculation is carried out by using the anchor-pharmacophore definition in order to generate fragment start placements for the build-up procedure. (These are stored in files for later use.) The second docking run is performed without any additional constraints and provides a score estimate for every fragment (which can also be saved to disk) and used during the build-up procedure. This issue is addressed in more detail in the discussion of the results and in technical sections.

Two fragment spaces were used for the calculations. The first was derived from a collection of molecules from the WDI by applying a modified version of the RECAP program.[4] (It has already been used for similarity searching with the Feature Tree program.[6]) This contains approximately 17 000 fragments and is called "WDI space" in the following paragraphs. By performing substructure filtering on the WDI space to exclude specific toxic and reactive substructures, we generated a second fragment space called "WDIsub" space. The substructures used for the filtering were deduced from the publications of Baurin et al.[10,56] and Kazius et al.,[57] though not all of these were used. Additionally, we limited the size of fragments to 300 Da, the number of hydrogen-bond donors and acceptors to six each, and the number of tetrahedral stereocenters to one. In this second fragment space approximately 4500 fragments remained after the filtering process.

FlexNovo parameterization

The same run parameters for the FlexNovo calculation were used for all targets. In each case, we performed two extension cycles so that the resulting molecules consist of up to three fragments. Due to the large variability in fragment size this is not, per se, a measure for the size of the resulting molecules. We kept up to 100 fragments for each extension cycle for further build-up, and considered up to 50 different placements for each fragment. Additionally, we defined several physicochemical property filters for each target in order to obtain solutions that satisfy drug-likeness criteria to a certain extent. Table 2 lists the corresponding property ranges, derived from target-specific collections of known active products.

Using different diversity parameters, we carried out multiple runs of the program for each target. The parameters were deliberately chosen such that the algorithm could, in principle, use the same start fragment for a whole solution set, in order explicitly to generate close analogues to known active structures. This is explained in more detail in the discussion of the individual results. For the estimation of binding energies, we used a modified version of the SCORE function[51] (as used in "standard" FlexX calculations).

Results and Discussion

This section starts with a brief description of the four targets and the output analysis procedure we used for the FlexNovo calculations. We then describe the calculations for the individual targets in detail and analyze the results. Finally, we summarize the characteristics of the calculations and provide some technical and statistical information.

Table 2. Physicochemical property ranges for the solutions of the FlexNovo calculations.

Property ranges	DHFR	CDK2	COX-2	ER
molecular weight	200–480	200–480	200–460	200–480
rings	0–4	0–6	0–4	0–6
non-terminal rot. bonds	0–12	0–12	0–10	0–12
hydrogen bond acceptors	0–13	0–9	0–8	0–8
hydrogen bond donors	0–7	0–6	0–3	0–5
calculated log*P*	−2.5–2.5	−2.5–6.0	1.0–7.0	1.0–9.0

The validation test set

We selected four targets that represent different design objectives in typical lead discovery projects. The binding sites of both dihydrofolate reductase (DHFR) and cyclin-dependant kinase 2 (CDK2) have rather solvent-exposed cavities and require specific hydrogen-bonding interactions. In contrast, cyclooxygenase-2 (COX-2) and the estrogen receptor (ER) have primarily lipophilic, buried cavities, and thus limited possibilities for forming directed interactions. The focus in the analysis of the candidate structures for each target depends therefore on the different structural aspects. The search constraints (described below) for the FlexNovo calculations reflect this.

Analysis of the FlexNovo results

A maximum of 50 molecules was generated for each target. All proposed molecules and their corresponding binding modes were visually inspected and compared to the predicted (docked) binding modes of known inhibitors. For the validation of FlexNovo, we were interested in finding structures that show high complementarity to a given protein active site, are chemically reasonable, satisfy drug-like properties to a certain extent, and have good predicted binding orientations. Given the size of the fragment spaces, it is unrealistic to expect to find close analogues for most known inhibitors. Therefore, we deliberately generated a limited set of solutions so that we could perform the analysis in a reasonable time and assess the performance of our current approach. From these results we picked one representative example of each combination of fragment space and target. Figure 2 shows the corresponding images, including the best predicted binding orientation of the solution together with the Connolly surface of the active site of the protein. The surface is color-coded such that surface patches resulting from heteroatoms are shown in the corresponding colors. This helps in distinguishing hydrophilic from lipophilic areas.

In addition we carried out similarity calculations on a set of known active molecules using the Feature Tree program.[6] For each of the known active structures in our set, only the molecule having the highest Feature Tree similarity from the solution set was considered. Here we focused on finding molecules with similar topology that are located in the same region of the active site according to the predicted binding mode. From these calculations, we picked pairs of known active structures and molecules from the solution list for each target. Figure 3 contains the corresponding images, which show the superimposed predicted binding orientations. Non-heteroatom-bound hydrogen atoms are not shown in the figure for clarity, but were included in the similarity analyses.

To complement the structural and topological analysis, we calculated distributions for the same set of physicochemical properties as was used for the filtering. We compared the distributions resulting from the solutions with the corresponding distributions of the collections of known active molecules. Finally, we performed a maximum common substructure (MCS)-based similarity analysis, from which we obtained an additional measure for the structural diversity of the solutions (Table 3, Figure 4).

Results for the four targets

Dihydrofolate reductase is a target in cancer and malaria research, frequently used for validating structure-based design approaches. The published structure (PDB ID: 4DFR) complexed with the Methotrexate was used for the calculation. The anchor-pharmacophore definition for the generation of the start placements was the formation of a hydrogen bond to Asp27. Scheme2 shows a 2D sketch of the key residues and the binding mode of Methotrexate.

Approximately 40% of the fragments in the fragment space could be successfully placed into the active site and satisfied the anchor-pharmacophore constraints. The predicted scores ranged from small positive energies to −32 (kJ mol^{-1}), which is partly due to the variability in fragment size but mainly reflects the interaction patterns.

For most molecules in the solution set, FlexNovo chose the same pteridine ring system for the initial fragment as that in Methotrexate, due to the high score for the specific interaction pattern it forms in the pocket. The adjacent aromatic ring system, which is present in many known active molecules, is well reproduced in all the solutions. Within the different sets, there are slight structural variations—bridging carbon or amine groups and fused or single aromatic rings—but all have the same predicted binding orientation. The adjacent amide bond, known mainly from Methotrexate, was often replaced in the solutions by either a secondary or tertiary amine. In some solutions one of these amine groups formed part of a fused ring system. This was followed by one or two carboxylic acid groups at varying distances. In general, the solutions mimic the binding mode of Methotrexate and other known active structures well, and have reasonable chemical and geometrical properties. In particular, FlexNovo suggested alternative bridging fragments and additional ring closures that are energetically beneficial in terms of entropy loss upon binding. Figure 2 shows two representative solutions, one for each fragment space. Despite employing the same starting fragment, they have different scaffolds that frequently occur in the solution sets.

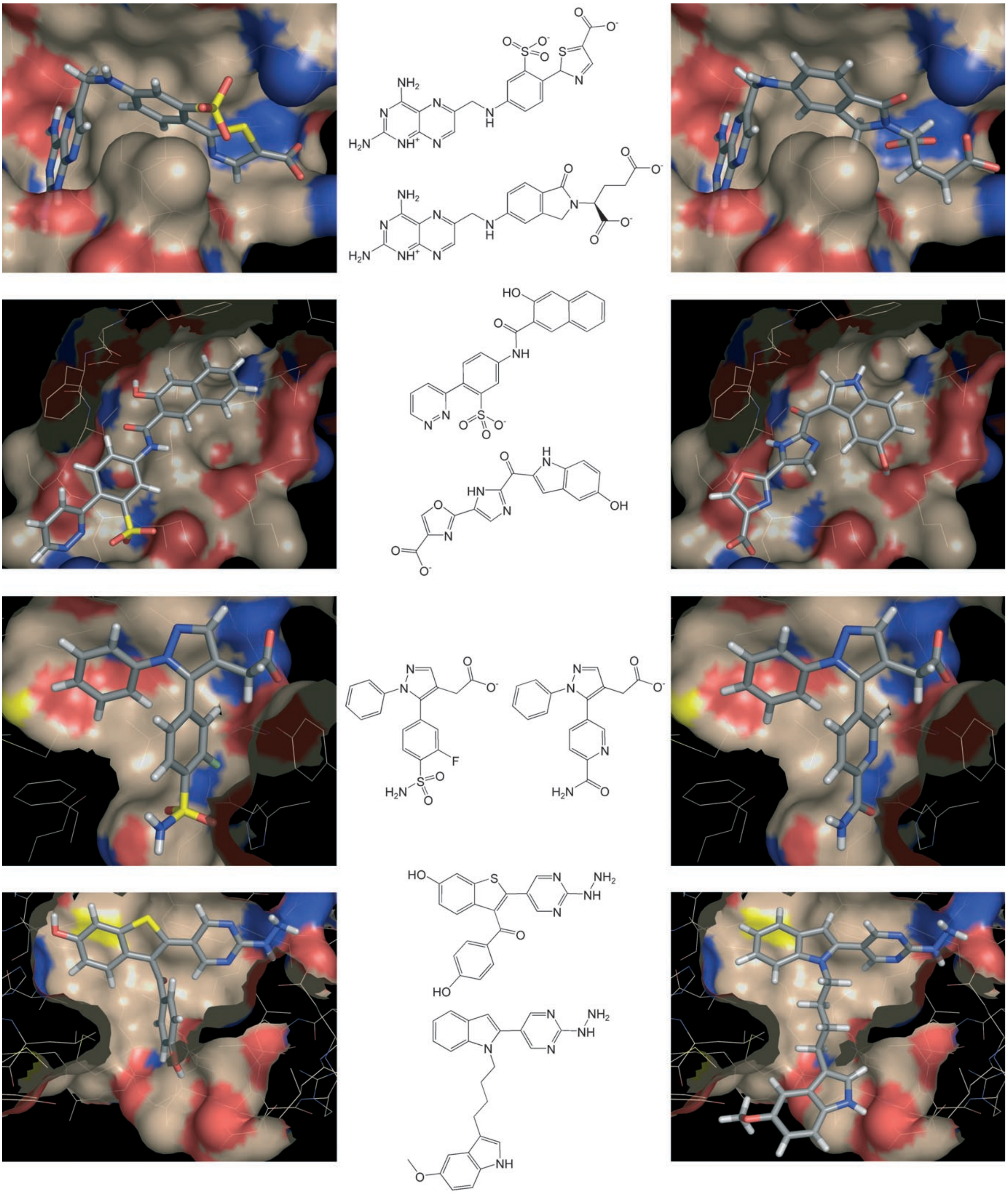


Figure 2. FlexNovo solution list molecules for the individual targets DHFR, CDK2, COX-2, and ER (from top to bottom) drawn according to their predicted binding orientation. The active site of the protein is shown with its Connolly surface. The top half of the cavity is removed. Surface patches resulting from heteroatoms are shown in the corresponding colors. The left column contains results from the WDIsub fragment space, the right from the WDI space. For COX-2, both results were generated by using the WDI fragment space.

Figure 3 shows some additional structurally diverse solutions that were identified by automated similarity analysis. The active set consisted of 68 structures (derived from Selassie et al.[58]) and Methotrexate. The solutions contain not only the

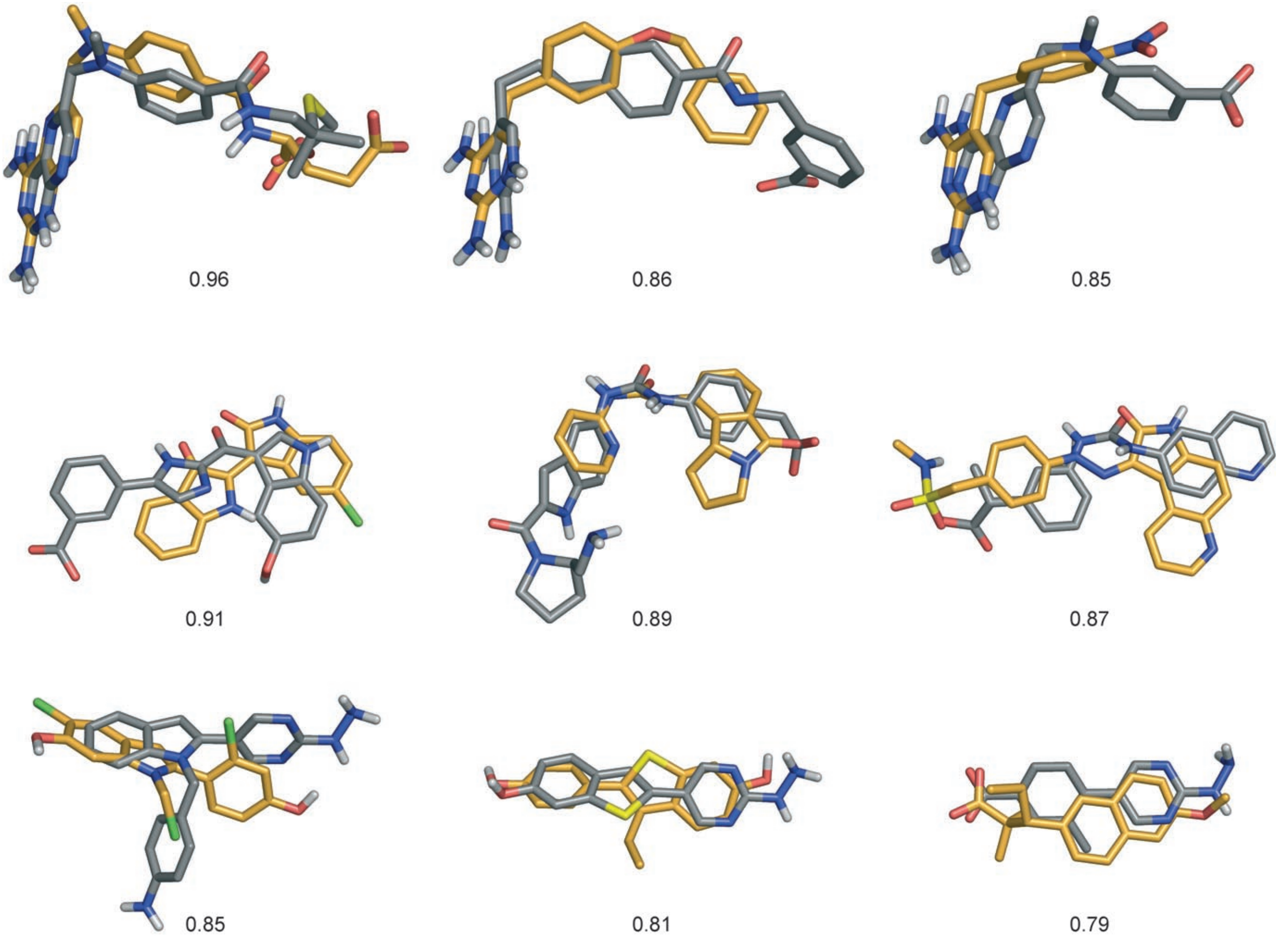
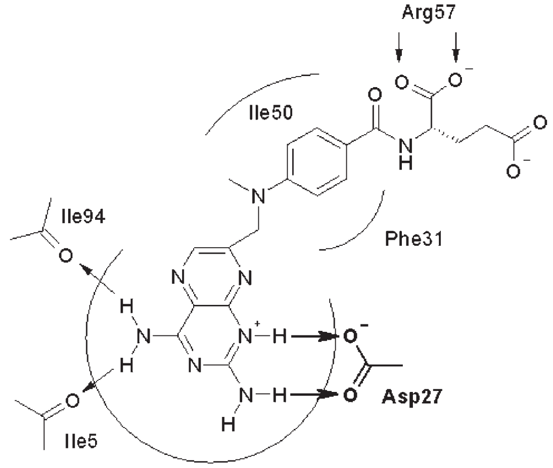


Figure 3. Superimposed pairs of FLEXNOVO solution list molecules (gray) for the targets DHFR, CDK2, and ER (from top to bottom), and known inhibitors (orange) with a high Feature Tree similarity value. Both structures are drawn according to their predicted/modeled binding orientation. Non-heteroatom-bound hydrogen atoms have been removed for clarity. Similarity values are given below each compound pair. FLEXNOVO structures are the closest analogues to the active structures with regard to Feature Tree similarity.

Table 3. Complete-linkage cluster analysis for four representative solution sets.

	DHFR	CDK2	COX-2	ER
clusters with MCS ≥ 10	8	6	7	7
clusters with MCS ≥ 15	17	14	18	16
clusters with MCS ≥ 20	29	31	46	31

The values indicate the amount of different clusters the corresponding solution set contains when requesting a maximum common substructure (MCS) of ≥ 10, ≥ 15, or ≥ 20 heavy atoms between all members in each cluster. The higher the amount of clusters for a given MCS value, the higher the diversity within the corresponding set. Values were derived from a combined solution set containing 100 molecules resulting from both fragment spaces.



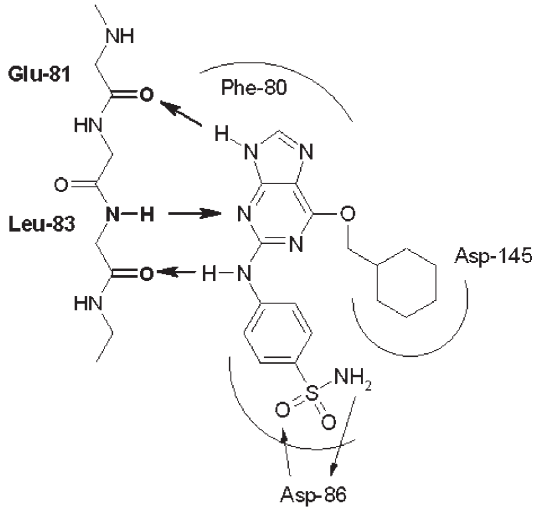
Scheme 2. 2D sketch of the active site of DHFR (PDB ID: 4DFR) complexed with Methotrexate (oriented according to the binding mode). The anchor-pharmacophore constraint used for the initial fragment placement was a hydrogen bond to Asp27.

substituted pteridine ring system but also the smaller, equally substituted pyrimidine ring system, often found in the active structures. The Feature Tree similarity values indicated that the proposed molecules have reasonable to high similarity to the corresponding active structures. Additionally, the predicted binding modes show that the structures have a reasonable orientation in the active site of the protein.

Cyclin-dependent kinase 2 plays a dominant role (among other kinases) in the modulation of diseases like cancer and is a well-established target in pharmaceutical research. ATP-competitive kinase inhibitors all form one or more hydrogen bonds to the "hinge region" of the active site. Therefore, an essential anchor-pharmacophore constraint for the generation of the start placements was the formation of a hydrogen bond to the central backbone NH group of Leu83. Most CDK2 inhibitors form additional hydrogen bonds to the nearby carbonyl groups of Glu81 and Leu83. These were also used as a constraint such that at least one interaction with these residues had to be formed. A 2D sketch (Scheme 3) of the key residues of the protein structure used for the calculation (PDB ID: 1DI8) indicates the binding mode of a substituted dimethoxyquinazoline ligand (taken from PDB ID: 1H1S).

The initial docking calculation yielded solutions for approximately 35% of the fragments that satisfy the specified anchor-pharmacophore constraints. The predicted scores were between small positive values and about −30 (kJ mol^{-1}). As with the DHFR example, a few high-scoring scaffolds dominate the final solution list. Figures 2 and 3 show the most frequently occurring examples.

In general, the resulting molecules were relatively large when compared to most known active molecules in our collection. This is due to the fact that the score of a molecule depends to some extent on its size. This is already evident from docking calculations, and becomes even more apparent in the case of the fragment spaces. Thus FlexNovo has a slight preference for larger compounds, though the user is able to control this behavior by defining appropriate property filter ranges.



Scheme 3. 2D sketch of the ATP binding site of CDK2 (PDB ID: 1DI8) complexed with a ligand (taken from PDB ID: 1H1S, oriented according to the binding mode). The anchor-pharmacophore constraints for the initial fragment placement were the formation of a hydrogen bond with the NH group of Leu83, and a hydrogen bond to the carbonyl group of either Glu81 or Leu83.

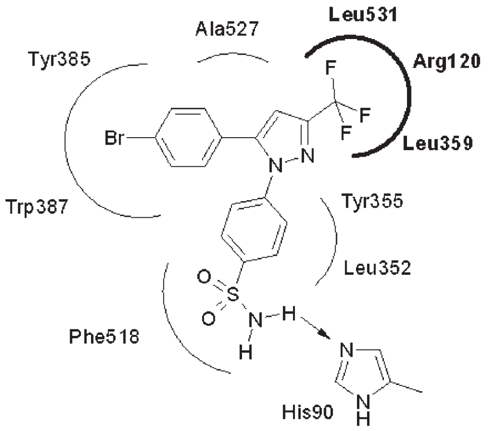
According to Traxler's binding model for ATP-competitive kinase inhibitors,[59] most of the resulting molecules occupy binding regions I and II, in addition to the adenine-binding region. The phosphate- and ribose-binding regions are omitted in most cases; this is due to the fact that FlexNovo extended the initial starting fragment preferentially along the hinge region of the active site rather than perpendicular to it. Overall, the solutions mimic the interaction pattern at the hinge region of the active site well, although not all have the perfect planar shape observed for the binding modes of most of the known active structures. Many of the generated molecules contain well-known structural motifs from known active structures that are positioned at similar regions of the active site. These include the indole (or indole-like) ring systems forming interactions with the backbone, bridging urea- or amide-like fragments, and oxazole or thiazole ring systems.

Figure 2 shows two solutions representing different classes (the central scaffolds of one are commercially available as Naphthol AS). Figure 3 shows a selection of predicted binding orientations of generated structures superimposed on modeled binding modes of known active structures taken from a set of 72 active structures[60–62] plus the dimethoxyquinazoline derivate. These show some more common structural motifs that were reproduced and frequently observed throughout the different solution sets. These figures demonstrate that FlexNovo is able to generate solutions that are both structurally similar to known ligands and have reasonable predicted binding orientations.

Cyclooxygenase-2 is a major therapeutic target for inflammatory diseases and the subject of pharmaceutical research in the areas of cancer and neurological diseases. For the COX-2 receptor, the structure complexed with the inhibitor SC558 was used for the calculations (PDB ID: 6COX). The structural element common to known active structures is an *ortho*-substituted ring system in one of the three lipophilic subpockets of the receptor. Consequently, the anchor-pharmacophore constraints for the initial fragment placement step consisted of the presence of such an *ortho*-substituted ring system in the subpocket formed by residues Leu531, Arg120, and Leu359 (among others) as shown in the 2D sketch (Scheme 4).

Solutions fulfilling the anchor-pharmacophore could only be found for about 13% of the fragments from the WDI space and for 20% of the fragments of the WDIsub space. The scores reached values of up to −25 (kJ mol^{-1}) in both cases. As in the previous examples there is a single starting fragment, a substituted phenyl-pyrazole system, that dominates all solution sets, due mostly to the high number of strong hydrogen-bonding interactions formed between the carboxylic acid group attached to the pyrazole ring and Arg120.

In general, the solutions generated by FlexNovo are more hydrophilic than most known active structures are. This is due to the fact that noninteracting polar groups are not penalized in the scoring function, an issue that is known from standard docking calculations. With fragment spaces this becomes more problematic due to the sheer quantity of fragments available. We compensated for this to a large extent with the saturation filter described above.



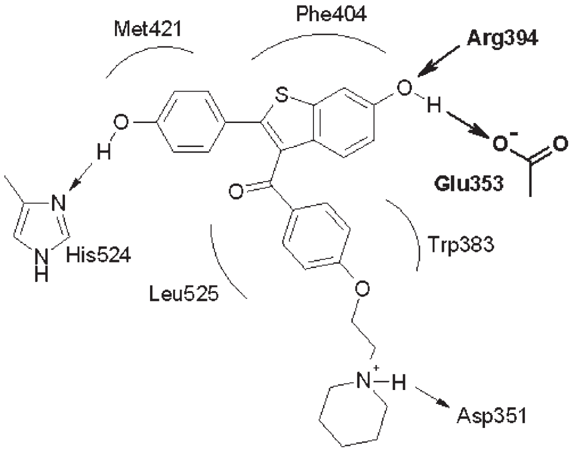
Scheme 4. 2D sketch of the active site of COX-2 (PDB ID: 6COX) complexed with the ligand SC558 (oriented according to the binding mode). The anchor-pharmacophore constraint for the initial fragment placement was an *ortho*-substituted aromatic ring positioned within a sphere of 1.5 Å radius in the subpocket formed by residues Leu531, Arg120, and Leu359.

Another characteristic of this particular example is that the active site of the COX-2 receptor is highly lipophilic. This make it difficult for the algorithm to find good solutions, since it depends on finding a few well-defined directed interactions in order to position the fragment (or molecule) properly in the pocket. Figure 2 shows two solutions that mimic the binding mode of most inhibitors almost exactly. Both are structurally very similar to many known active structures in the collection we used for comparison. Although this is a relatively demanding design case for the algorithm FLEXNOVO was able to find solutions that contain known structural motifs and occupy all three lipophilic subpockets of the active site by using only one initial pharmacophore-type constraint.

Two further characteristics are of interest. Firstly, halogen atoms are not contained in the underlying chemical model, so the algorithm cannot account for any halogen-bonding effects—basically they are treated as "steric bulk". Therefore, virtually no halogen-containing molecules are part of the solution set. Secondly, the carboxylic acid group attached to the pyrazole ring in the start fragment is not favorable for a possible COX-2 inhibitor. Nevertheless, the results are presented here in order to highlight the preference for directed polar interactions and the effect this imposes on the build-up algorithm and the consequent results. (The user can easily exclude such fragments by specifying an appropriate SMARTS expression, either in the anchor-pharmacophore definition that is used for the initial fragment placement step or during the build-up process itself.)

In contrast to the other targets, for which both fragment spaces performed roughly equally, the WDIsub fragment space performed worse than the WDI space in this case. Additionally, the automated similarity analysis did not yield any results that were sufficiently distinct from the ones already selected for Figure 2. Therefore, no molecules were found that had high similarity to the other 32 known active structures contained in our comparison set.[63–65] (Figure 3 shows no results for the COX-2 receptor.) This might be due to the very limited number of results we generated with FLEXNOVO and the relatively small set of highly similar active structures we used for the analysis.

Estrogen receptor agonists or antagonists are applied in the treatment of cancer and cardiovascular diseases, and the development of selective ER modulators is well established in pharmaceutical research. The protein structure complexed with Raloxifene was used for the calculations for ER (PDB ID: 1ERR). Most of the known agonists and antagonists for this receptor form hydrogen bonds to both Arg394 and Glu353. The anchor-pharmacophore definition we used for the initial fragment-placement step was less strict: only one hydrogen bond, to either of the two residues, was mandated for the start fragment. Scheme 5 shows the key residues of the active site of the receptor and indicates the binding mode of Raloxifene.



Scheme 5. 2D sketch of the active site of ER (PDB ID: 1ERR) complexed with Raloxifene (oriented according to the binding mode). The anchor-pharmacophore constraint for the initial fragment placement was the formation of a hydrogen bond to either Arg394 or Glu353.

Because of the unspecific anchor-pharmacophore definition, the algorithm was able to find solutions for about 63% of the fragments of the fragment space. As with the other targets, the scores range from low positive values to nearly −40 (kJ mol^{-1}) for the largest fragments. A couple of fragments dominate the final solution list, a phenomenon we observed for all the targets. In this case it is a pyrimidyl–hydrazine fragment that is common to all solutions shown in Figures 2 and 3. The score for this fragment is much better than those of other possible start fragments, which contain a single phenolic hydroxo group, for example, due to the higher number of hydrogen bonds that the hydrazyl-group is able to form with Arg394 and Glu353.

Steroid-like starting fragments were deliberately excluded from the calculations, so FLEXNOVO was not able to generate structures similar to the steroid structures in our set of known active structures.[66–69] Moreover, there is a conceptual limitation in the present approach for generating these kinds of structures; for algorithmic reasons, ring-closing reactions cannot be considered in the build-up process, so these types of reactions are neither modeled nor contained in the fragment spaces FlexNovo uses.

In general, FlexNovo reproduced the structural features of the non-steroidal estrogen agonists and antagonists very well. All results include two aromatic moieties that are connected by a single bond, ranging from Arg394 and Glu353 to His524, and thereby fill the pocket quite well. As mentioned above, FlexNovo used a hydrazyl group to satisfy the anchor-pharmacophore constraint instead of, for example, a phenolic hydroxo group (somehow similar to the case of COX-2, above). It has the same (scoring) reasons and can easily be controlled by the user by defining more specific filter criteria. However, the attached (second) ring system is, in most cases, either a substituted benzothiophene or indole-like fragment followed by an alkyl chain or a third aromatic ring.

Figure 2 shows two representative examples that contain all the features described above, and shows that FlexNovo reproduced the binding motif known from Raloxifene or other active structures well. In particular, FlexNovo positioned an indole ring NH group in exactly the same position as the corresponding NH group in the Raloxifene piperidine ring. Although the two ring systems have different electronic and chemical properties, FlexNovo was able to reproduce the main structural, geometrical, and topological aspects of different active structures by using just one initial pharmacophore-type constraint. Figure 3 contains additional results that have similar structures and binding orientation to those of the corresponding active structures.

Feature Tree similarity analysis: In order to increase the objectivity of the results we performed an automated similarity analysis using the Feature Tree program.[6] Different solution sets from FlexNovo were compared to a collection of known active structures. For each known structure, only the solution set molecule with the highest Feature Tree similarity was taken. As mentioned earlier, we focused on finding molecules that have topologically similar structural elements, located in the same region of the active site, according to the predicted binding mode. From these calculations, we picked different pairs of known active structures and solution list molecules for each target. Figure 3 presents the corresponding images, which show that FlexNovo reproduced common structural motifs and binding modes for a number of different known actives for three different target proteins.

Cluster analysis: As the Feature Tree descriptor is a measure of topological similarity, it is not well suited to the identification of common structural scaffolds within a set of molecules. Therefore we chose the number of heavy atoms contained in the maximum common (connected) substructure (MCS) as similarity measure. For each target, we combined the two solution sets to create a set of 100 results from both fragment spaces and carried out pairwise MCS calculations for all molecules within the set. With these results, we performed a complete-linkage hierarchical cluster analysis. We requested a minimum MCS of ten shared heavy atoms in order to assign them to the same structural class. Table 3 shows the number of different clusters obtained when using MCS values of ≥ 10, ≥ 15, and ≥ 20 heavy atoms as a cutoff. With this definition, at least six different structural classes were identified for each solution set. The higher the number of clusters, the higher the diversity was compared to the solution set as a whole. Therefore, the number of clusters increases with MCS value.

Physicochemical property distributions: Having evaluated the main structural and topological characteristics of the solutions, we calculated distributions for the same set of physicochemical properties that was used for filtering. The unified solution set described for the scaffold analysis above was also used for this analysis. We compared the distributions for the solutions generated by FlexNovo with the corresponding distributions for the collections of known active structures. Figure 4 shows a smoothed line representation of the normalized histograms obtained. Although the distributions are quite similar for some properties, they differ in one specific aspect (mentioned above): FlexNovo tends to generate solutions that are slightly larger and more hydrophilic than reference values. Nevertheless, Figure 4 shows that FlexNovo was able to generate solution sets that have reasonable property distributions, and the user can easily compensate for specific aspects by using stricter filter criteria.

General observations

From the results discussed above, we can make some general observations. FlexNovo generates solutions that have quite reasonable hydrogen-bonding patterns, particularly with respect to the predicted geometry, a characteristic that is known from FlexX docking calculations. The chemical-property filter criteria implemented can, to a large extent, compensate for difficulties resulting from the use of standard scoring functions. This is relevant since these effects are even more pronounced in the case of fragment spaces. Placement-geometry filters further compensate for undesirable attributes of the scoring function. Besides the fact that these filter criteria guide the algorithm to the construction of more sensible molecules, the overall performance of the program is enhanced significantly by speeding up the calculation, as shown below.

Typically, a couple of high-scoring scaffolds dominated the solution sets for each target. This is the case not only for the final solution list but also for the intermediate fragment queues. The diversity filters are useful for limiting the amount of identical (atomic) fragments that two intermediate fragments, or resulting molecules, have in common. Additionally, the number of occurrences of each initial (atomic) fragment can be limited for each extension cycle, thus introducing a degree of diversity from the beginning of the calculation.

FlexNovo benefits (even more than FlexX) from a good definition of the binding site of a protein. The stricter the search constraints, particularly concerning the size of the active site, the better the results will be. This also holds for the definition of the anchor-pharmacophore-type constraints for the initial fragment-placement step. Therefore, the selection of the start fragments is an important issue, although very strict substructure or property filtering should be handled with care. Figure 5 shows the predicted binding orientation of a solution obtained from a calculation for CDK2 with an earlier version of

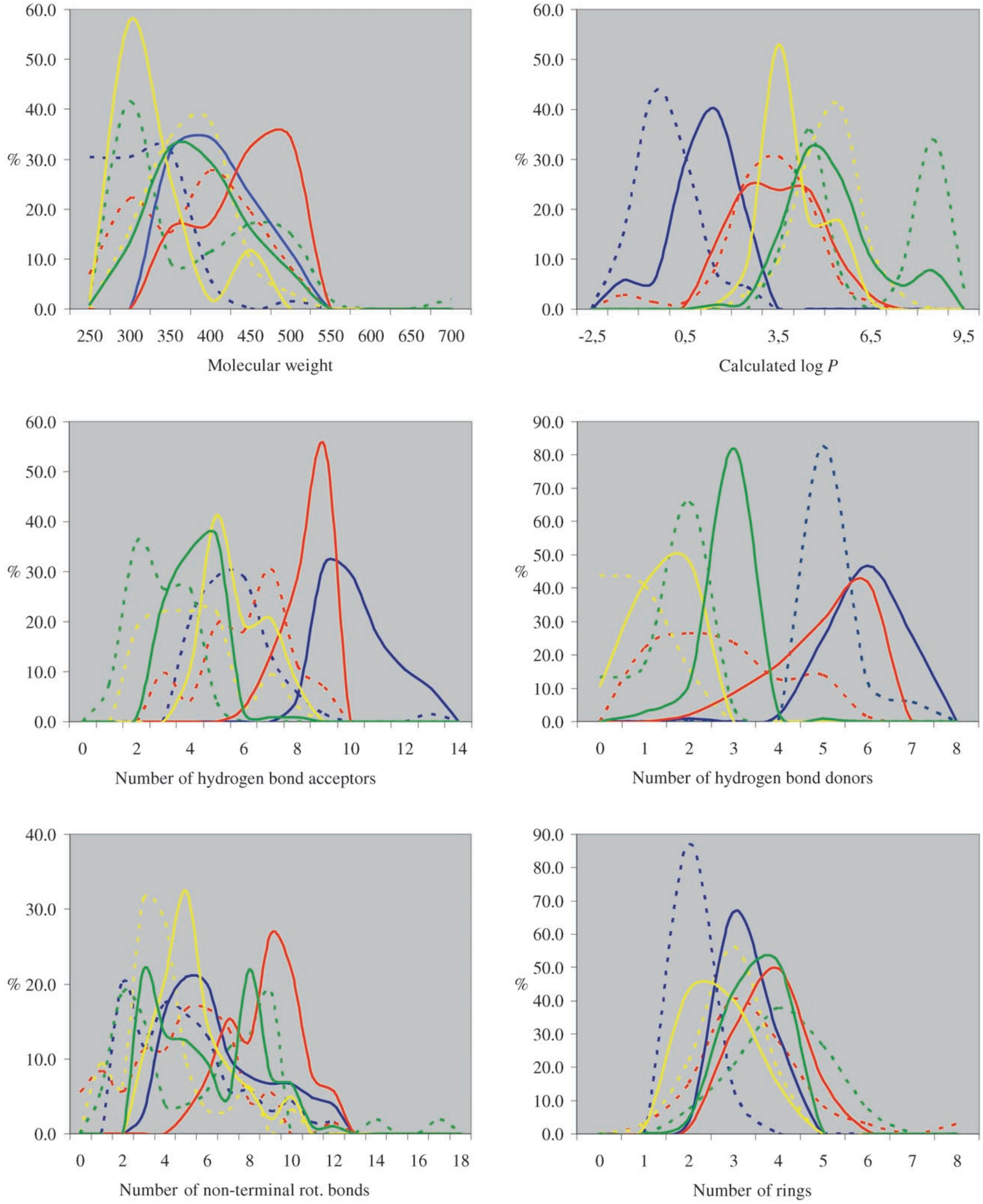


Figure 4. Physicochemical property distributions for solution sets generated by FlexNovo (——) and known actives (- - - -). Histograms are normalized and represented as smoothed line graphs. Each individual image contains results for the four different targets: DHFR (dark blue), CDK2 (red), COX-2 (yellow), and ER (green).

FlexNovo, in which no filter functionality was used, although the run parameters were identical. The result violates most of the physicochemical property filter criteria used for calculations and preprocessing described above. However, it mimics some of the main structural properties and the binding mode of Staurosporine nearly perfectly; this is quite astonishing bearing in mind the size of the fragment space and the number of intermediate solutions evaluated (see below). Although the so-

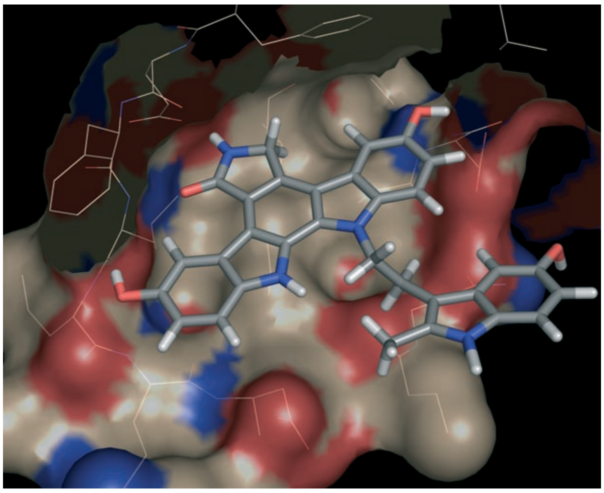


Figure 5. Predicted binding orientation of a molecule obtained from an earlier version of FlexNovo in which no filter criteria were used. The active site of CDK2 is shown with its Connolly surface. The top half of the cavity is removed. Surface patches resulting from heteroatoms are shown in the corresponding colors.

lution shown might not look like an ideal lead candidate structure, it does share the main structural features of CEP-1347 (Cephalon and Lundbeck[70]), which was targeted for the treatment of Parkinson's disease. (Phase II/III clinical trials were abandoned due to poor performance.)

Although we used FlexNovo in a fully automated mode for this evaluation, this is not a requirement. The user can perform all calculations in a stepwise manner by carrying out every extension cycle separately. Thereby, not only the start fragments but also the intermediate solutions obtained from each step can be hand-selected, rescored, or optimized with other methods and then used for the next extension cycle. Additionally, each filter property and program parameter can be adjusted separately for each build-up step.

Whereas the initial docking calculation (under anchor-pharmacophore type constraints) is essential for a FlexNovo calculation, this is not the case for the second docking run for determining the best fragment scores. FlexNovo can also estimate the maximum (protein-independent) score of a fragment within the build-up algorithm itself, albeit with less accuracy. In general, the score estimation should be as accurate as possible since this speeds up the calculation significantly.

Computing times and memory usage

All the FlexNovo calculations were carried out on a standard, single CPU workstation (Intel Xeon 2.8 GHz, 2 GB memory). Except for the diversity parameters, we used the same set of parameters for each individual calculation. Table 4 summarizes the main statistical information from these calculations, averaged over all calculations performed (therefore, the margin of the individual values is rather high). In general, the calculation becomes more expensive the lower the allowed maximum number of common initial (atomic) fragments between two solutions is. The number of fragments retained during each extension cycle, and the number of placements stored for each fragment also have a major influence on the runtime. The time needed for the preprocessing steps is not included in this runtime information since it is constant. On average, the preprocessing takes a couple of seconds per fragment resulting in, for example, one day of CPU time for the WDI fragment space.

Table 4. Computation times and memory usage for a FlexNovo calculation with standard parameters on a single CPU workstation for two different fragment spaces.

Fragment space	WDI	WDIsub
fragments contained	17000	4500
results generated	50	50
compounds evaluated	1.2 M–1.5 M	350 k–478 k
compounds placed	28 k–62 k	24 k–62 k
total process time	7.0 h–15.5 h	5.0 h–10.0 h
total memory usage	750 MB	200 MB

From these experiments, we can deduce the following principle characteristics: The amount of memory required and the number of compounds evaluated are basically dependent on the size of the underlying fragment space. For a typical scenario, the memory usage is between a couple of hundred megabytes (MB) and one gigabyte (GB). The number of compounds FlexNovo examined ranges from a couple of hundred thousand to several million. In contrast, the total CPU time of the calculation and the number of compounds that were actually placed in the active site depend principally on the quantity of results requested. The nature of the target protein also plays a crucial role. If, for example, the active site is relatively closed, the computation time is much smaller than if it is highly solvent-exposed. The tighter the search constraints for the algorithm, the less time it will take. For a typical scenario, a calculation takes a couple of hours, but a few days are necessary for a very broad search space.

There are two reasons for the discrepancy between the number of compounds FlexNovo evaluates and the number actually placed. The first lies in the initial docking calculation for all initial (atomic) fragments of the fragment space without any additional constraints. (FlexNovo uses this to estimate a score of fragments for each particular target.) The better this estimation is, the more fragments can be discarded in the build-up phase due to their low potential score. The second reason concerns the definition of the property ranges. The more precisely these are defined, the greater the number of fragments that can be discarded prior to construction. Table 4 shows the significant reduction in computing time resulting from these measures (particularly the score estimation).

The parameterization we chose, keeping up to 100 fragments per extension cycle for further build-up and up to 50 placements for each intermediate solution, turned out to be the best compromise between execution time and results obtained for all test cases. Lower values for the number of fragments resulted in less satisfactory results due to reduced coverage of the search space. With a smaller set of placements per fragment, often no solutions were found at later stages of the incremental construction procedure. On the other hand, increasing the number of placements (and considering a couple of hundred intermediate solutions) did not greatly improve the results. This is in agreement with findings related to the *k*-greedy approach in molecular-docking calculations: covering a broader search space does not necessarily improve the results significantly.[71]

Conclusions

The new structure-based molecular-design software package, FlexNovo, has been used to design potential inhibitors for four targets of pharmaceutical interest (DHFR, CDK2, COX-2, and ER). FlexNovo is based on the molecular-docking software FlexX, employs its incremental construction algorithm, and uses the same chemical models and scoring functions. Additionally, several placement geometry, physicochemical property (drug-likeness), and diversity filter criteria are directly integrated into the build-up process. The results obtained show that FlexNovo is able to generate diverse sets of molecules that are highly complementary to different target proteins. These molecules exhibit drug-like properties and have reasonable predicted binding orientations. The results, including those for an automated similarity analysis with the Feature Tree program, indicate that FlexNovo often reproduces structural motifs as well as the corresponding binding modes of known active structures.

The diversity of the results is a critical issue, as oftentimes a couple of high-scoring scaffolds tend to dominate the final solution list. FlexNovo has a powerful mechanism for limiting the number of initial (atomic) fragments two molecules have in common; this permits the generation of a more diverse set of results. Nevertheless, a more sophisticated measure of diversity might be desirable. This is an area we are aiming to improve. Despite the limitations of the current diversity estimation, the MCS-based scaffold analyses we carried out indicate that, for the combined set of 100 solutions, at least six different structural classes were generated for each target.

FlexNovo treats fragments in a more comprehensive manner than first-generation de novo design software tools. It works directly with the fragments themselves, rather than a generalized molecular (template) structure. Furthermore, FlexNovo handles very large fragment spaces with up to several thousand fragments and there are no fragment restrictions, such as their size. The underlying chemical model allows accurate modeling of molecular interactions and treatment of conformational flexibility. Due to this, the use of a deterministic algorithm, well-defined connection rules, and the many filter criteria available, FlexNovo provides a comprehensive and user-customizable virtual synthesis of molecules. The physicochemical property distributions we calculated reflect this.

However, there are some conceptual limitations. The first is related to the treatment of ring-closing reactions: for algorithmic reasons FlexNovo is not able to close rings during the structure-generation process, so there are no corresponding fragments or connection rules contained in the fragment space. Furthermore, not all energetic effects are considered in the evaluation of the binding energies due to the use of a standard scoring function; halogen atoms are not contained in the chemical model, so halogen-bonding effects are not considered during the build-up process. In general, a fragment space is used “as is”; this means that there are no additional data or fragment-refinement steps available beyond those included in the FlexX software package. Therefore, the user has to account for diversity and chemical property issues beforehand.

Another principal limitation is that pharmacophore-type constraints can only be used for the initial fragment-placement step. From our results we can imagine cases in which such an initial constraint might not be sufficient. Additionally, the user may want to define a more complex pharmacophore pattern that all the solutions FlexNovo generates should fulfill. Despite the current improvements, we envision this as an area that we wish to address in a future version of the program. Regardless of the present limitations, FlexNovo is a modular and user-customizable system, so that other fragment spaces, chemical models, scoring functions or build-up algorithms can be incorporated.

In view of the scope of the calculations performed here, we consider this to be a comprehensive first feasibility study of the software. The next step will be a substantial retrosynthetic analysis with synthesis and experimental testing of the results. FlexNovo can be considered as an alternative to a screening approach, with the advantage of covering a larger part of the underlying chemical space, in particular because it can easily be used in a parallel computing environment. Although the scale-up for the build-up calculation is probably not linear, an improvement of one or two orders of magnitude can realistically be achieved by using a typical computer cluster of, for example, 100 CPUs. Therefore, we think that the results obtained are encouraging and show that FlexNovo is a valuable tool for a structure-based design task. Additionally, FlexNovo might help in suggesting novel structural motifs, for designing a first library for a given target protein, and assisting in understanding the interactions and related issues of protein–ligand binding. The algorithmic and conceptual improvements we plan might further increase the capability of the current approach.

Acknowledgements

The authors would like to thank Andrea Zaliani (Lilly Research Laboratories, Hamburg, Germany) for numerous stimulating discussions and for careful reading of the manuscript, and one of the anonymous reviewers for helpful comments on the analysis of the FlexNovo results.

Keywords: chemical fragment spaces · computer chemistry · de novo design · drug design · protein–ligand interactions

[1] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50.
[2] C. M. Dobson, *Nature* **2004**, *432*, 824–828.
[3] W. P. Walters, M. T. Stahl, M. A. Murcko, *Drug Discovery Today* **1998**, *3*, 160–178.

[4] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511 – 522.

[5] G. Schneider, M.-L. Lee, M. Stahl, P. Schneider, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487 – 494.

[6] M. Rarey, M. Stahl, *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497 – 520.

[7] D. Fattori, *Drug Discovery Today* **2004**, *9*, 229 – 238.

[8] D. A. Erlanson, R. S. McDowell, T. O'Brien, *J. Med. Chem.* **2004**, *47*, 3463 – 3482.

[9] D. C. Rees, M. Congreve, C. W. Murray, R. Carr, *Nat. Rev. Drug Discovery* **2004**, *3*, 660 – 672.

[10] N. Baurin, F. Aboul-Ela, X. Barril, B. Davis, M. Drysdale, B. Dymock, H. Finch, C. Fromont, C. Richardson, H. Simmonite, R. E. Hubbard, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2157 – 2166.

[11] M. J. Hartshorn, C. W. Murray, A. Cleasby, M. Frederickson, I. J. Tickle, H. Jhoti, *J. Med. Chem.* **2005**, *48*, 403 – 413.

[12] R. A. Lewis, P. M. Dean, *Proc. R. Soc. London Ser. B* **1989**, *236*, 125 – 140.

[13] G. Schneider, U. Fechner, *Nat. Rev. Drug Discovery* **2005**, *4*, 649 – 663.

[14] V. J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos, A. P. Johnson, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207 – 217.

[15] V. J. Gillet, G. Myatt, Z. Zsoldos, A. P. Johnson, *Perspect. Drug Discovery Des.* **1995**, *3*, 34 – 50.

[16] A. P. Johnson, K. Boda, T. Lengyel, S. Weaver, *Proceedings of the 228th ACS National Meeting*, American Chemical Society, Philadelphia, PA, United States, **2004**.

[17] N. P. Todorov, P. M. Dean, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 175 – 192.

[18] N. P. Todorov, P. M. Dean, *J. Comput.-Aided Mol. Des.* **1998**, *12*, 335 – 349.

[19] H.-J. Böhm, *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61 – 78.

[20] H.-J. Böhm, Perspect. *Drug Discov. Design* **1995**, *3*, 21 – 33.

[21] M. C. Lawrence, P. C. Davis, *Proteins Struct. Funct. Genet.* **1992**, *12*, 31 – 41.

[22] C. M. W. Ho, G. R. Marshall, *J. Comput.-Aided Mol. Des.* **1993**, *7*, 623 – 647.

[23] D. E. Clark, D. Frenkel, S. A. Levy, J. Li, C. W. Murray, B. Robson, B. Waszkowycz, D. R. Westhead, *J. Comput.-Aided Mol. Des.* **1995**, *9*, 13 – 32.

[24] M. B. Eisen, D. C. Wilez, M. Karplus, R. E. Hubbard, *Proteins Struct. Funct. Genet.* **1994**, *19*, 199 – 221.

[25] A. R. Leach, S. R. Kilvington, *J. Comput.-Aided Mol. Des.* **1994**, *8*, 283 – 298.

[26] J. B. Moon, W. J. Howe, *Proteins Struct. Funct. Genet.* **1991**, *11*, 314 – 328.

[27] S. H. Rotstein, M. A. Murcko, *J. Med. Chem.* **1993**, *36*, 1700 – 1710.

[28] J. C. Baber, M. Feher, *Mini-Rev. Med. Chem.* **2004**, *4*, 681 – 692.

[29] J. F. Pritchard, M. Jurima-Romer, M. L. J. Reimer, E. Mortimer, B. Rolfe, M. N. Cayen, *Nat. Rev. Drug Discovery* **2003**, *2*, 542 – 553.

[30] P. D. Leeson, A. M. Davis, J. Steele, *Drug Discovery Today Technol.* **2004**, *1*, 189 – 195.

[31] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Delivery Rev.* **1997**, *23*, 3 – 25.

[32] T. I. Oprea, A. M. Davis, S. J. Teague, P. D. Leeson, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308 – 1315.

[33] M. M. Hann, A. R. Leach, G. Harper, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856 – 864.

[34] C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855 – 861.

[35] M. Vieth, M. G. Siegel, R. E. Higgs, I. A. Watson, D. H. Robertson, K. A. Savin, G. L. Durst, P. A. Hipskind, *J. Med. Chem.* **2004**, *47*, 224 – 232.

[36] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887 – 2893.

[37] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1999**, *42*, 5095 – 5099.

[38] R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103 – 108.

[39] R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037 – 1050.

[40] K. P. Cross, G. Myatt, C. Yang, M. A. Fligner, J. S. Verducci, P. E. J. Blower, *J. Med. Chem.* **2003**, *46*, 4770 – 4775.

[41] H. B. Broughton, I. A. Watson, *J. Mol. Graphics Modell.* **2004**, *23*, 51 – 58.

[42] C. Lazar, A. Kluczyk, T. Kiyota, Y. Konishi, *J. Med. Chem.* **2004**, *47*, 6973 – 6982.

[43] A. C. Pierce, G. Rao, G. W. Bemis, *J. Med. Chem.* **2004**, *47*, 2768 – 2775.

[44] K. Boda, A. P. Johnson, *J. Med. Chem.* **2006**, ASAP article.

[45] M. Rarey, J. S. Dixon, *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471 – 490.

[46] M. Stahl, N. P. Todorov, T. James, H. Mauser, H.-J. Böhm, P. M. Dean, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 459 – 478.

[47] M. Rarey, S. Wefing, T. Lengauer, *J. Comput.-Aided Mol. Des.* **1996**, *10*, 41 – 54.

[48] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470 – 489.

[49] M. Rarey, B. Kramer, T. Lengauer, *J. Comput.-Aided Mol. Des.* **1997**, *11*, 369 – 384.

[50] S. A. Hindle, M. Rarey, C. Bunning, T. Lengauer, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 129 – 149.

[51] H.-J. Böhm, *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243 – 256.

[52] M. Stahl, M. Rarey, *J. Med. Chem.* **2001**, *44*, 1035 – 1042.

[53] D. K. Gelhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, S. T. Freer, *Chem. Biol.* **1995**, *2*, 317 – 324.

[54] S. A. Wildman, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868 – 873.

[55] M. Stahl, H.-J. Böhm, *J. Mol. Graphics Modell.* **1998**, *16*, 121 – 132.

[56] N. Baurin, R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parrat, P. Greaney, D. Morley, R. E. Hubbard, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643 – 651.

[57] J. Kazius, R. McGuire, R. Bursi, *J. Med. Chem.* **2005**, *48*, 312 – 320.

[58] C. D. Selassie, Z.-X. Fang, R.-L. Li, C. Hansch, G. Debnath, T. E. Klein, R. Langridge, B. T. Kaufmann, *J. Med. Chem.* **1989**, *32*, 1895 – 1905.

[59] P. Traxler, P. Furet, *Pharmacol. Ther.* **1999**, *82*, 195 – 206.

[60] G. R. Rosania, Y.-T. Chang, *Expert Opin. Ther. Pat.* **2000**, *10*, 215 – 230.

[61] T. M. Sielecki, J. F. Boylan, P. A. Benfield, G. L. Trainor, *J. Med. Chem.* **2000**, *43*, 1 – 18.

[62] J. Dumas, *Expert Opin. Ther. Pat.* **2001**, *11*, 405 – 429.

[63] J. S. Carter, *Expert Opin. Ther. Pat.* **1998**, *8*, 21 – 29.

[64] R. W. Friesen, C. Brideau, C. C. Chan, S. Charleson, D. Deschenes, D. Dube, D. Ethier, R. Fortin, J. Y. Gauthier, Y. Girard, R. Gordon, G. M. Greig, D. Riendeau, C. Savoie, Z. Wang, E. Wong, D. Visco, L. J. Xu, R. N. Young, *Bioorg. Med. Chem. Lett.* **1998**, *8*, 2777 – 2782.

[65] A. S. Kalgutkar, *Expert Opin. Ther. Pat.* **1999**, *9*, 831 – 849.

[66] R. A. Magarian, L. B. Overacre, S. Singh, K. L. Meyer, *Curr. Med. Chem.* **1994**, *1*, 61 – 104.

[67] M. J. Coghlan, M. E. Kort, *Expert Opin. Ther. Pat.* **1999**, *9*, 1523 – 1536.

[68] L. D. Gegnas, *Expert Opin. Ther. Pat.* **2000**, *10*, 833 – 846.

[69] W. Sippl, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 559 – 572.

[70] M. S. Saporito, R. L. Hudkins, A. C. Maroney, *Prog. Med. Chem.* **2002**, *40*, 23 – 62.

[71] A. Griewel, M. Rarey in *Proceedings of the German Conference on Bioinformatics GCB 2005*, Gesellschaft für Informatik (GI), Hamburg, **2005**, pp. 119–130.

Received: December 22, 2005
Published online on June 14, 2006