Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design

Uli Fechner and Gisbert Schneider*

Johann Wolfgang Goethe-Universität, Institut für Organische Chemie und Chemische Biologie, Marie-Curie-Str. 11, D-60439 Frankfurt, Germany

Received August 31, 2005

It is demonstrated that the fragmentation of druglike molecules by applying simplistic pseudo-retrosynthesis results in a stock of chemically meaningful building blocks for de novo molecule generation. A stochastic search algorithm in conjunction with ligand-based similarity scoring (Flux: *f*ragment-based *l*igand b*u*ilder rea*x*ions) facilitated the generation of new molecules using a single known reference compound as a template. This molecule assembly method is applicable in the absence of receptor-structure information. In a case study, we used imantinib (Gleevec) and a Factor Xa inhibitor as the reference structures. The algorithm succeeded in redesigning the templates from scratch and suggested several alternative molecular structures. The resulting designed molecules were chemically reasonable and contained essential substructure motifs. A comparison of molecular descriptors suggests that holographic descriptors might be advantageous over binary fingerprints for ligand-based de novo design.

INTRODUCTION

The process of drug discovery commonly starts with the selection of a suitable biological target. Several methods can be deployed to the subsequent task, that is, the attainment of hits and ultimately a lead candidate.[1] This can be accomplished by, for example, high-throughput screening (HTS), virtual screening, or computational de novo design. The concept of both HTS and virtual screening is to find a "needle in the haystack" by searching for hits in large compound collections. These collections overlap to a certain extent between different pharmaceutical companies, thereby potentially leading to the detection of identical hits by companies that work on the same drug target. Application of de novo design minimizes the risk of identical hit retrieval as it does not directly rely on existing compound collections. Hence, de novo design may provide a competitive advantage.[2]

De novo design aims at generating novel molecular structures that exhibit desired pharmacological activity and match the binding pattern of a particular biological target. This binding pattern can be defined by either the three-dimensional (3D) target structure (structure-based) or at least one known reference ligand (ligand-based). The majority of current computer-based molecular design approaches are structure-based; that is, they rely on the availability of the three-dimensional target structure. Structure-based design cannot be exerted when a high-resolution structure of a biological target is unavailable. This applies to many membrane-bound receptors including the large group of G-protein coupled receptors.[3] Structure-based design concepts are also faced with the challenge of accurately predicting the binding energy of the generated molecules (scoring function). Apart from very few implementations, structure-based de novo design software does not usually regard the problem of protein flexibility.[4] Nevertheless, structure-based de novo design has been successfully employed for the generation of bioactive molecular structures.[5,6] Ligand-based de novo design complements structure-based approaches by working independently from the 3D target structure.[7] Instead, one or more ligands that bind to the biological target of interest are a prerequisite. The 3D structure of these known active molecules may or may not be taken into account. The design task is then guided by maximizing the similarity between the generated molecules and the known active(s). This calculation depends on a meaningful descriptor and a similarity coefficient.[8]

De novo design approaches can also be differentiated with regard to their building blocks. Molecules are assembled on either a per-atom or per-fragment basis. The atom-based method theoretically allows the generation of any molecule by enabling the addition, removal, and replacement of a single atom. Yet, this advantage comes with severe drawbacks. First, the chemical search space that includes all theoretically possible molecules with a druglike molecular weight is huge.[9] By changing the generated structures atom-wise, only small steps in this space are rendered possible at a time. Second, many of the structures that are created by an atom-based molecular generation algorithm are not amenable to chemical synthesis, a problem encountered by many de novo design procedures.[10] In contrast, fragment-based de novo design can only spawn molecules that cover a restricted region of the chemical space where the location and size of this region directly depend on the amount and type of employed fragments. A potential pitfall of fragment-wise alteration is the larger step size in chemical space compared to atoms as building blocks. This can be circumvented by compiling a fragment set that is diverse with respect to molecular size. Such a fragment compilation allows both small and larger steps to be made.

Here, we present a ligand-based concept of straightforward de novo molecule generation (Flux: *f*ragment-based *l*igand b*u*ilder rea*x*ions). It implements a stochastic search algorithm to navigate in chemical space. A restricted set of reaction

* Corresponding author tel:+49-69 798 24873; fax:+49-69 798 24880; e-mail: gisbert.schneider@modlab.de.

10.1021/ci0503560 CCC: $33.50
Published on Web 12/10/2005

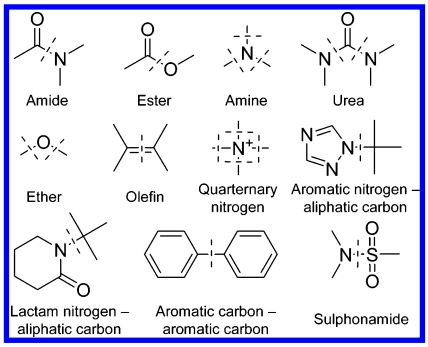


Figure 1. Eleven RECAP bond-cleavage types. This set of reaction schemes was employed for both the virtual retrosynthesis of the drug database to yield molecular fragments and the synthesis of these fragments during a de novo design run.

schemes and fragments is involved in the molecule assembly process to ease chemical synthesis of the ligands that are proposed by Flux. The combination of two descriptors and two similarity metrics resulted in four different ligand-based fitness functions to asses the quality of our designed structures. Additionally, we subjoined basic drug-likeness filters. We evaluated our molecule generator with two redesign studies using a kinase inhibitor and a Factor Xa inhibitor as templates.

METHODS

The Molecular Building Blocks. Building blocks for our molecule generator were yielded by virtual retrosynthesis of the COBRA data set with a limited set of reaction schemes. The COBRA data set is a collection of reference molecules for ligand-based library design compiled from recent scientific literature.[11] The data set is nonredundant and annotated by target receptor information and activity data. For this study, COBRA version 2.1 was used. It consists of 4705 compounds. Our virtual retrosynthesis approach very closely follows the concept of RECAP (retrosynthetic combinatorial analysis procedure) conceived by Lewell and co-workers.[12] We applied the same 11 bond-cleavage types derived from common chemical reactions (Figure 1). Cyclic structures were left intact; that is, ring bonds were not decomposed. In the original publication, reactions were avoided that give rise to small functional groups (including fragments up to four carbon atoms). We only excluded reactions where single hydrogen atoms emerge as products and explicitly included small fragments. Small building blocks allow for small steps in chemical space and, thus, facilitate "fine-tuning" during the molecular design process. Fragmentation of drug molecules, like those contained in the COBRA data set, is expected to yield building blocks that cover druglike areas of the chemical space. Hence, the search space of our molecule generator is limited in a meaningful manner as a result of its reduction to druglike regions of the chemical space.

Our virtual retrosynthesis algorithm iterates over the 11 bond-cleavage types. In each iteration cycle, all the molecules of the COBRA data set that are still uncleaved as well as all the fragments that were already dissected by the same or a previous bond-cleavage type are subject to virtual retrosynthesis. Thus, the resultant fragments may contain attachment sites that emerged from more than a single bond-cleavage type.

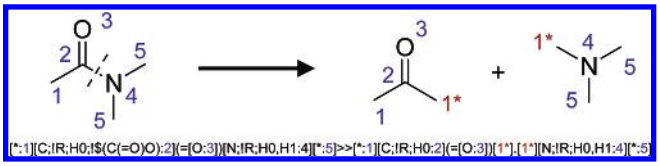


Figure 2. Depiction and SMIRKS for amide bond-cleavage type. Blue numbers denote the indices that ensure an unambiguous assignment of atoms in reactants and products. The red symbol "1*" is used to mark both the reaction type and the position where the reaction occurred.

The number of building blocks per bond-cleavage type considerably varied from type to type. These numbers could have been assimilated by applying artificial assignments or swaps of cleavage types to the existing cleaved fragments. We decided not to employ such operations because we wanted to retain the bond-cleavage type origin of each fragment. Intentionally, we accepted an overall smaller number of building blocks to increase the chance of synthetic feasibility. Flux does not lead to arbitrary combinations of *generic* building blocks; rather, a straightforward concept of simplistic virtual chemical reactions was implemented.

The 11 reaction schemes were defined by Daylight SMIRKS as implemented in the Daylight toolkit function (Daylight Chemical Information Systems Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA 92691).[13] "Virtual atoms" were joined to the resultant fragments to mark both the position where a reaction took place and the type of reaction. The "virtual atoms" of the 11 reaction schemes were specified as "1*" to "11*" in the SMARTS language (see Figure 2 for an example). Molecule dissection was performed with the program retroFlux, which is our slightly modified reimplementation of RECAP and is freely available from the Daylight contrib directory at http://www.daylight.com/download/contrib/.

The Fitness Function. Navigation of our molecule generator was guided by optimization of the similarity between the generated virtual molecules and a known active reference molecule ("template structure").[14] The calculation of pairwise similarity relies on a coordinate system that defines the chemical search space and a metric that defines distance in this space. We employed two descriptors to span a high-dimensional chemical space: the Daylight fingerprint (FP) and the Ghose and Crippen descriptor. The FP is a hashed fingerprint.[13] We calculated fingerprints with a length of 512 bits using the respective function of the Daylight toolkit. Ghose and Crippen defined 120 atom types to predict the hydrophobicity and molar refractivity of small organic molecules.[15] These atom types were also successfully employed as a 120-dimensional holographic fingerprint descriptor.[16,17] SMARTS definitions of the 120 atom types and SMARTS matching functions of the Daylight toolkit were used to compute the counts of different Ghose and Crippen atom types occurring in a molecule. This calculation gave rise to the nonbinary Ghose and Crippen fingerprint (GC).

Two indices were applied for distance (or similarity) calculations: the Tanimoto coefficient (also known as the Jaccard coefficient) and the Euclidian distance.[8] We chose these two similarity coefficients because of their widespread use in cheminformatics. Values of the Tanimoto coefficient are in [0;+1] for the binary version (eq 1) and [$-1/3$;+1]

for the nonbinary version (eq 2), where a value of 1 indicates identity. Depending on the descriptor, the binary (FP) or nonbinary (GC) equation of the respective index was applied. Results of the Euclidian distance (eq 3) are in [0;+∞], where zero refers to identity.

$$S_{A,B} = \frac{c}{a + b - c} \tag{1}$$

$$S_{A,B} = \frac{\sum_{j=1}^{j=n} x_{jA}x_{jB}}{\sum_{j=1}^{j=n}(x_{jA})^2 + \sum_{j=1}^{j=n}(x_{jB})^2 - \sum_{j=1}^{j=n} x_{jA}x_{jB}} \tag{2}$$

$$D_{A,B} = \sqrt{\sum_{j=1}^{j=n}(x_{jA} - x_{jB})^2} \tag{3}$$

In eqs 1−3, *A* and *B* are objects (here, molecules), *i* and *j* are attributes of these objects, *n* is the total number of attributes of an object, x_{jA} is the value of the *j*th attribute of object *A*, *a* is the number of bits set to "on" in object *A*, *b* is the number of bits set to "on" in object *B*, *c* is the number of bits set to "on" in objects *A* and *B*, $S_{A,B}$ denotes the similarity between objects *A* and *B*, and $D_{A,B}$ is the distance between objects *A* and *B*.

Optimization Algorithm. The evolutionary pressure of "selection" and the evolutionary operators "crossover" and "mutation" can be simulated in stochastic search algorithms to optimize solutions of a wide variety of problems.[18] These algorithms are based on the ideas described by Charles Darwin in 1859.[19] Whereas genetic algorithms act on a separately encoded transformation of the objective variables (the "chromosome"), evolution strategies usually operate directly on the values to be optimized.[20] Although evolutionary algorithms are available in a plethora of flavors, a common feature among all evolutionary algorithms is their cyclic process of variation and selection (Figure 3). For the present study, we implemented a simplistic (1, λ) evolution strategy (ES) without adaptive step-size control.[21] The 1 in (1, λ) ES states that only the fittest individual of one generation survives; λ states the number of children per generation. Selection of the best is performed among the offspring only; that is, the parent dies out. This sometimes facilitates escaping local optima in the fitness landscape.[22] More advanced ES algorithms will be compared in de novo design exercises, and the results will be published elsewhere (in preparation).

The algorithm starts at an arbitrary point in the search space by assembling a random molecular structure. This initial structure is the parent structure of the first generation. According to the *molecule mutation algorithm* (described in the next section), λ offspring structures are generated. The fitness of the offspring is determined, and the structure with the best fitness is selected as the parent of the next generation. If the termination criterion is satisfied, the algorithm terminates; otherwise, the parent breeds another generation of λ molecules. If identical offspring structures are spawned within the same generation, only one of them is kept. This can lead to a smaller number of offspring for a generation

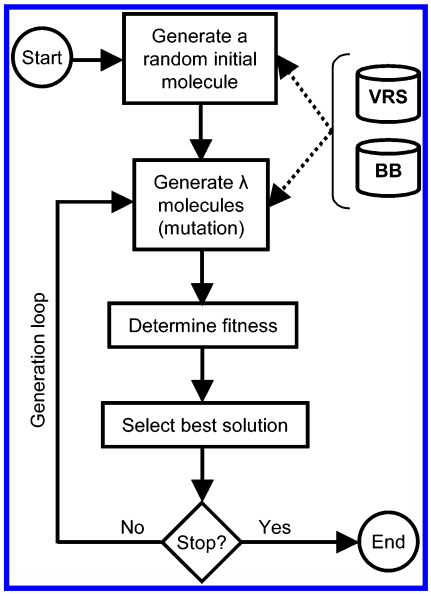


Figure 3. Schematic of the evolutionary algorithm employed in our de novo design software Flux. The virtual reaction schemes (VRS) and the building blocks (BB) provide the basis of the molecule mutation algorithm and the random structure assembly algorithm.

than λ. We did not iterate the mutation procedure as long as λ offspring were reached because the number of building blocks per bond-cleavage type considerably varied from type to type. Hence, the number of possible different child structures for certain cleavage types was limited.

The random structure that serves as a starting point is constructed according to the *random structure assembly algorithm*. The algorithm begins by randomly picking an initial fragment, $\text{fragment}_{\text{initial}}$, with *n* attachment sites, site_1 to site_n, from the stock of building blocks. It is guaranteed that $n \geq 1$ as we only included building blocks that contain at least one attachment site. Each attachment site emerged from pseudo retrosynthesis by retroFlux with a particular virtual reaction scheme and can be related to that by its labeling with the respective "virtual atom". A fragment, fragment_1, that is complementary to site_1 is then randomly chosen from the stock of building blocks and attached by virtual synthesis. Two fragments are complementary if (i) they share the same type of attachment site of the 11 types available and (ii) their "polarity", if there is any for the particular reaction type, is complementary. Five of the 11 bond-cleavage types give rise to two different fragment polarities ("amide", "ester", 'aromatic nitrogen−aliphatic carbon", "lactam nitrogen−aliphatic carbon", and "sulphonamide"). For example, the virtual retrosynthesis of an amide bond yields one fragment with an amine group and one fragment with a carboxy group: a virtual synthesis can only be carried out between a fragment with an amine group and a fragment with a carboxy group (complementary polarities), not between two fragments with an amine group and not between two fragments with a carboxy group (same polarity). Note that two of the 11 reaction schemes require more than two educts: "amine" has three and "quarternary nitrogen" has four educts. If the attachment site_1 of $\text{fragment}_{\text{initial}}$ corresponds to "amine" or "quarternary nitrogen", one or two additional fragments are randomly picked from the stock of building blocks prior to deployment of the virtual synthesis. The complete structure, structure_1, is yielded by

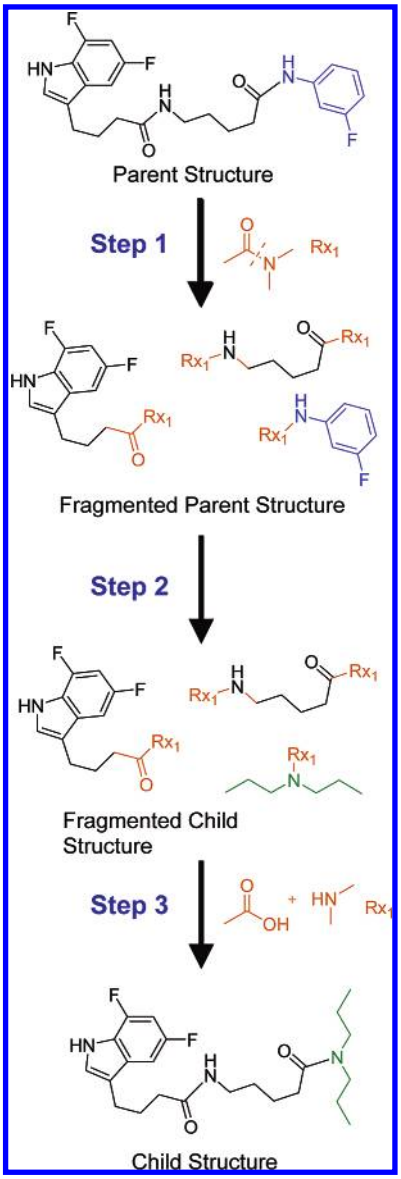


Figure 4. Exemplary run of the molecule mutation algorithm. **Step 1:** The parent structure is exhaustively dissected with the randomly picked amide bond reaction scheme. **Step 2:** One of the three parent structure fragments is then randomly selected (blue) and replaced by another one. **Step 3:** Finally, the child structure is obtained by virtual synthesis using the same reaction scheme that was applied for the dissection of the parent structure. Both the uncleaved and the cleaved amide bond motifs are highlighted.

iterating the selection-synthesis procedure n times where fragment_w is randomly picked from the stock of building blocks in the wth selection-synthesis cycle. In structure_1, all n attachment sites of $\text{fragment}_{\text{initial}}$ are saturated. If at least one of the fragments fragment_1 to fragment_n has more than one attachment site, structure_1 has at least one unsaturated attachment site. This is due to the fact that the virtual synthesis is only carried out between $\text{fragment}_{\text{initial}}$ and not-yet-connected fragments; that is, no intramolecular reactions are performed. If structure_1 contains one or more unsaturated attachment sites, $\text{fragment}_{\text{initial}}$ is set to structure_1 and the random structure assembly algorithm starts over again. Otherwise, the algorithm terminates.

The only two variables that have to be specified for our limited version (lacking adaptive step size control) of a (1, λ) ES are the number of generations (g) and the number of

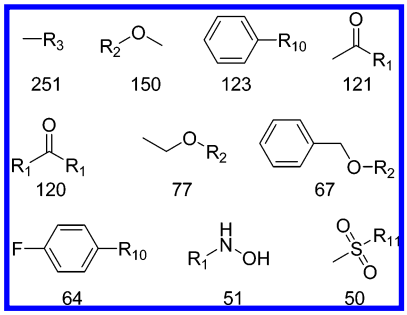


Figure 5. The 10 most frequent fragments that were yielded by a virtual retrosynthesis of the COBRA data set. Numbers denote the occurrences; R_1 to R_{11} refer to the respective bond-cleavage type as shown in Figure 1 that was virtually carried out to obtain the fragment.

children per generation (λ), which were set to 50 and 100, respectively. In preliminary experiments, we observed a convergence of the fitness within the specified number of generations, indicating that our selection of g and λ are sufficient for convergence on a local fitness optimum (not shown). Additional criteria for termination of the algorithm prior to the fixed number of generations were not employed. This particular population-based optimization algorithm does not guarantee that the parent structure of the current generation is superior to parent structures of previous generations in terms of the fitness function. A comprehensive analysis of the candidate compounds may, therefore, include the parent structures of all generations.

Molecule Mutation Algorithm. In each generation of the evolution strategy, the current parent structure is mutated to yield λ child structures. The molecule mutation algorithm starts with the random selection of one of the 11 virtual synthesis schemes reaction_1 to reaction_{11}. The chosen reaction, reaction_x, is then applied to exhaustively retrosynthesize the parent structure. If the parent structure is not amenable to virtual retrosynthesis with reaction_x, that is, the dissection with reaction_x does not lead to any fragments, another reaction is randomly chosen. Successful retrosynthesis of the parent structure gives rise to a set of fragments $F_{\text{parent}} = \{\text{fragment}_1, ..., \text{fragment}_n\}$ with $|F_{\text{parent}}| \geq 2$. Each of the n fragments contains at least one attachment site, site_x, where site_x is an attachment site that was obtained by retrosynthesis with the virtual reaction scheme reaction_x. One of the members of F_{parent} is then randomly picked ($\text{fragment}_{\text{original}}$), and a set of fragments L is compiled that contains all building blocks from the stock that are compatible with $\text{fragment}_{\text{original}}$. Fragments are compatible with $\text{fragment}_{\text{original}}$ if they share (i) the same type of attachment site, site_x, and (ii) the same kind of polarity, provided that retrosynthesis with reaction_x leads to fragments of different polarity. One fragment, $\text{fragment}_{\text{exchange}}$, is then randomly selected from the set of compatible fragments L. The set of fragments of the child structure are $F_{\text{child}} = (F_{\text{parent}}/\{\text{fragment}_{\text{original}}\}) \cup \text{fragment}_{\text{exchange}}$; that is, $\text{fragment}_{\text{original}}$ is replaced by $\text{fragment}_{\text{exchange}}$. Finally, the fragmented child structure is assembled with the same reaction, reaction_x, that was employed to dissect the parent structure.

If $\text{fragment}_{\text{original}}$ and $\text{fragment}_{\text{exchange}}$ have the same number of attachment sites, the child structure is complete and the molecule mutation algorithm terminates. If $\text{fragment}_{\text{original}}$ has less attachment sites than $\text{fragment}_{\text{exchange}}$, the unsaturated attachment site(s) of the child structure have to be saturated

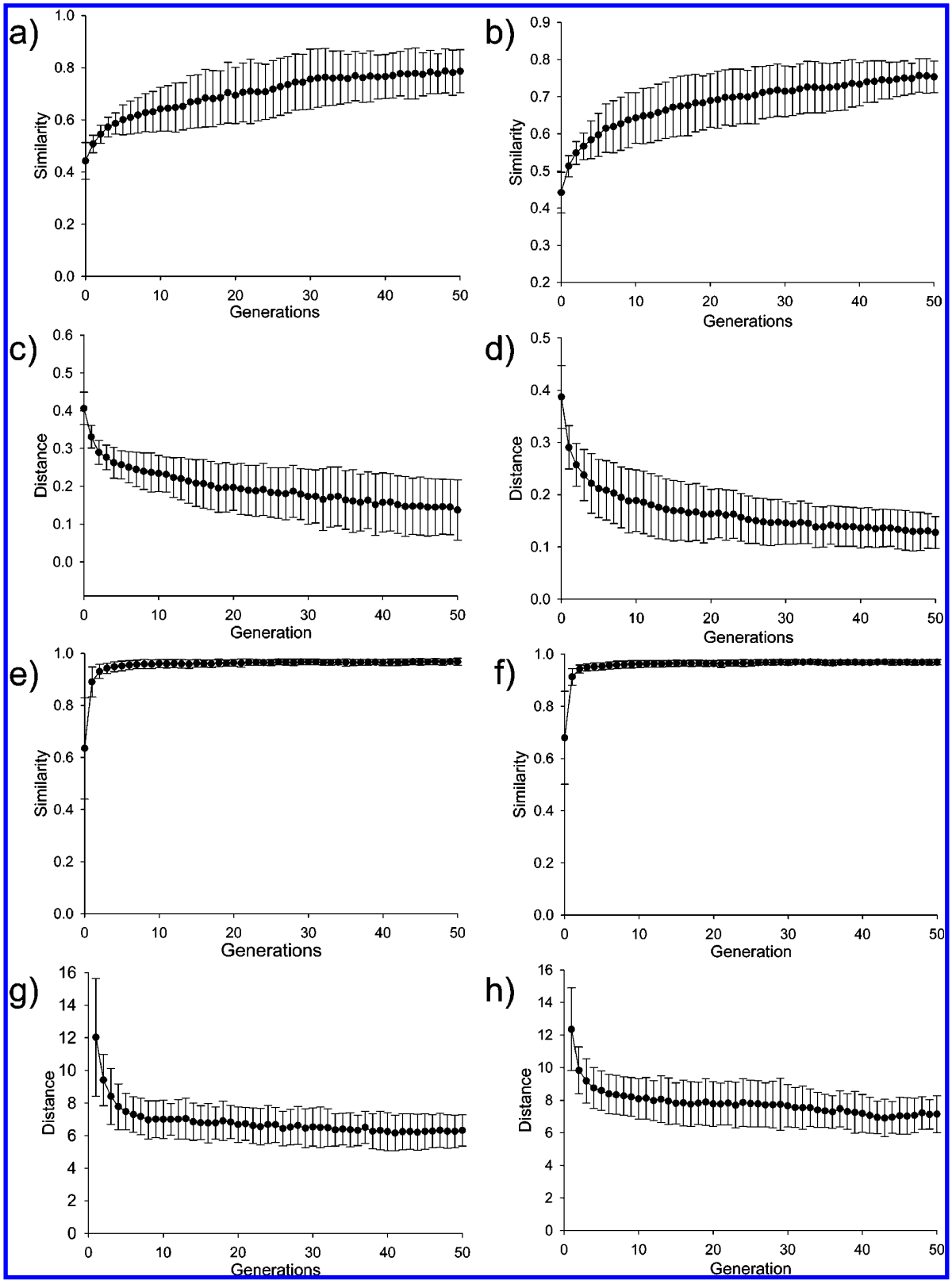


Figure 6. Average fitness values and standard deviations of 50 design runs in the FP descriptor space (a−d) and GC descriptor space (e−h). The left column of plots shows the results for template molecule **1**, the right column for template molecule **2**. If the ordinate is labeled "Similarity", the Tanimoto index was employed; if it is labeled "Distance", the Euclidian distance was used.

one after the other with compatible fragments that are randomly picked from the stock of building blocks. A child structure can, therefore, grow compared to the corresponding parent structure. We did not observe considerable growth because, in the present study, most of our building blocks (approximately 73%) were univalent, thereby limiting the chance of picking a fragment with more than one attachment site. If $\text{fragment}_{\text{original}}$ contains more attachment sites than $\text{fragment}_{\text{exchange}}$, the child structure can also be smaller than its parent. The individual steps of the virtual synthesis of the fragmented child structure are carried out in reverse order compared to the retrosynthesis of the parent structure. It is,

thus, clearly defined which fragments of F_{child} do not have a binding partner. Figure 4 illustrates an example of the molecule mutation algorithm.

Basic Drug-Likeness Filter. We implemented coarse-grained filters to prevent the generation of molecules that have a reduced chance of oral bioavailability and membrane permeation. These filters follow ideas of the "rule of five".[23] Structures with less than 15 or more than 50 non-hydrogen atoms were eliminated. We also excluded designed molecules where the sum of oxygen and nitrogen atoms was greater than 12 to roughly appraise an upper limit of the sum of hydrogen-bond acceptors and donors. The removal of these structures took place after the creation of λ child structures. A generation, therefore, may contain less than λ children not only because of the breeding of identical children structures but also because of the application of the drug-likeness filter. The calculated molecular weight (cMW) and octanol−water partition coefficient (clog *P*) were obtained from ChemDraw 7.03 (Cambridge Soft Corporation, 100 Cambridge Park Drive, Cambridge, MA 02140).

Implementation Details. Both retroFlux and Flux were written in ANSI compatible C and extensively employ functionality of the Daylight toolkit. Molecules and fragments are read and written in SMILES format. Uniqueness of structures within one generation is guaranteed by testing for string equality of unique SMILES. SMARTS matching is used to ensure that fragments are compatible or complementary in terms of their polarity. The SMARTS language was also employed to define the 120 GC atom types, and the SMARTS matching functionality of the Daylight toolkit served as a basis for the calculation of the GC fingerprint. The 11 bond-cleavage types were written in the SMIRKS language, and virtual synthesis and retrosynthesis were facilitated by the SMIRKS matching feature.

RESULTS AND DISCUSSION

We developed a ligand-based de novo design program ("Flux") to provide a means for virtual molecule generation. Fragments were used as molecular building blocks. These fragments were gained by virtual retrosynthesis of a drug database with a set of 11 bond-cleavage types. The same set of reaction schemes was then employed to link individual fragments, thereby yielding candidate compounds. Chemical similarity between a known ligand against the biological target of interest and the design candidates served as a scoring function. This chemical similarity was defined by a descriptor (Daylight fingerprints or GC descriptor) and a distance index (Euclidian metric or Tanimoto coefficient). We choose two biological targets to demonstrate the principal capabilities of our de novo design approach: molecule **1** (Gleevec, imantinib), an inhibitor of Abl protein−tyrosine kinase,[24] and molecule **2** (cMW = 572 Da, clog *P* = 4.02), a Factor Xa inhibitor picked from a large Ugi three-component combinatorial library.[25,26]

Virtual retrosynthesis of the 4705 structures in the COBRA version 2.1 data set yielded 3788 unique building blocks, that is, fragments with at least one attachment site. Amine and amide cleavage were most frequently performed, followed by aromatic nitrogen−aliphatic carbon and aromatic carbon−aromatic carbon dissections. Urea, ether, olefin, and quarternary nitrogen cleavage reactions contributed fewer

Table 1. Average Values (Standard Deviations in Brackets) of the Overall 100 Best Designs for the Different Combinations of Molecular Descriptors (FP or GC) and Fitness Measures (Tanimoto index *S* or Euclidian Distance *D*)

descriptor	fitness measure	template molecule **1**	template molecule **2**
FP	S	0.86 (0.10)	0.78 (0.05)
FP	D	0.10 (0.09)	0.11 (0.03)
GC	S	0.98 (0.01)	0.98 (0.01)
GC	D	5.34 (1.0)	5.91 (0.85)

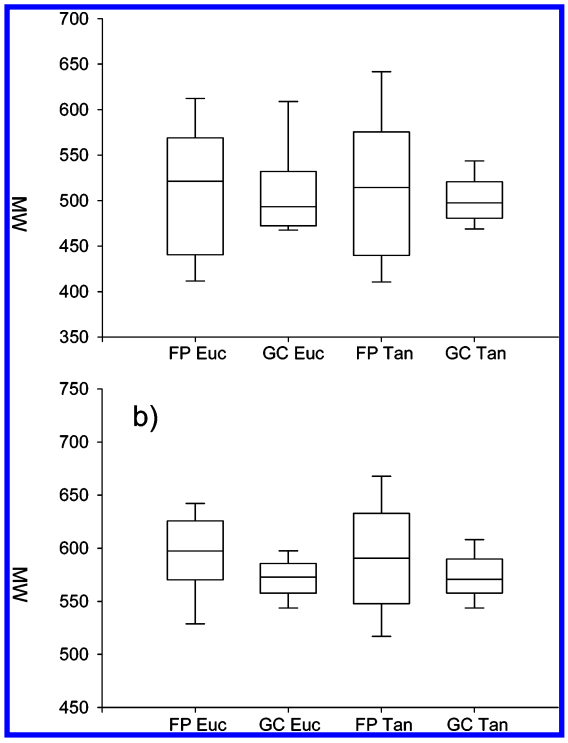
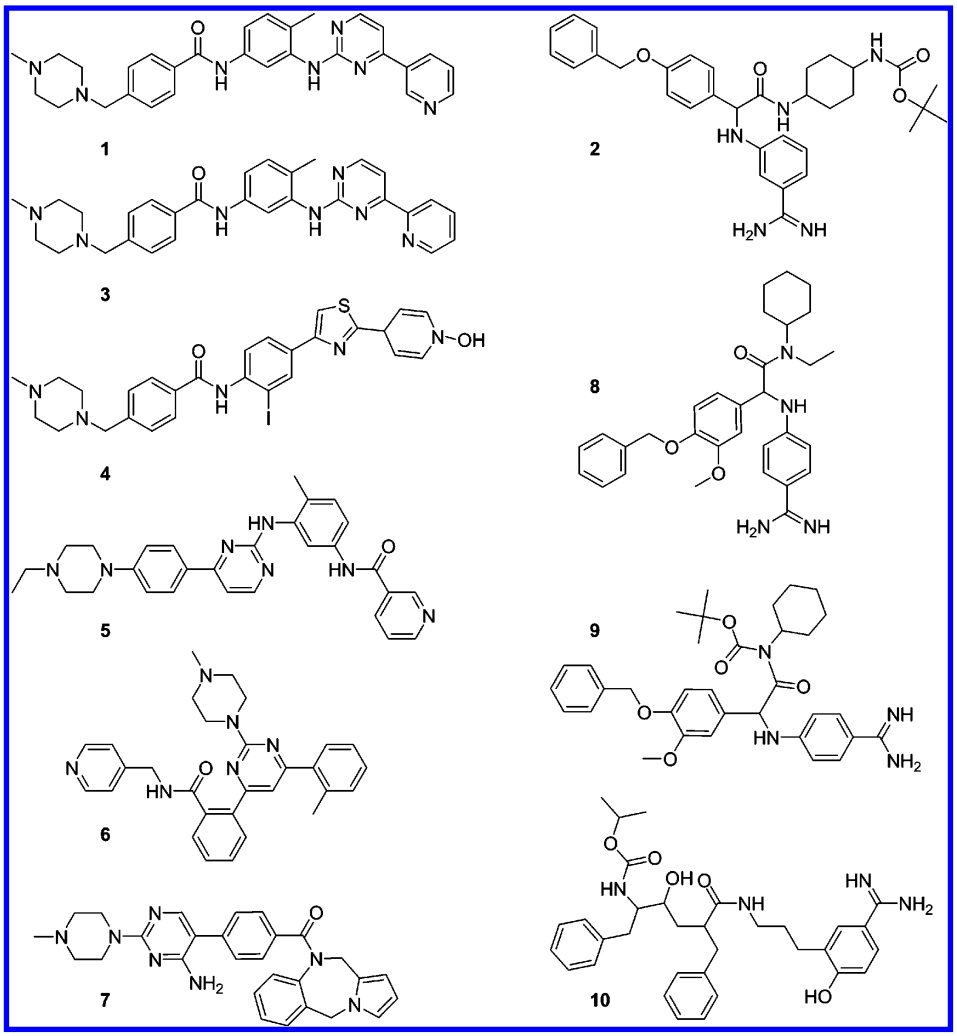


Figure 7. Box plots showing the molecular weight (MW) distribution of the 100 best designed structures. Reference molecules were Gleevec (a) and molecule **2** (b). The line within the box indicates the median, the lower boundary of the box marks the 25th percentile, and the upper boundary gives the 75th percentile. Whiskers mark the 90th and 10th percentiles.

building blocks, which is in agreement with the original RECAP application to fragmentation of the Derwent World Drug Index.[12] A total of 1531 structures of COBRA were not dissected at all by any of the 11 bond-cleavage types. Most of the fragments were univalent (2784), but we also gained bivalent building blocks (932), fragments with three attachment sites (64), and a few fragments with four attachment sites (8). This set of unique building blocks provided the stock of fragments for a virtual assembly of novel molecules. The 10 most frequent fragments obtained from COBRA are shown in Figure 5. Linker fragments and small side chains such as the methyl or methoxy group were formed most frequently, and the benzene ring occurred as the most frequent ring system after application of *retro*Flux, which is in agreement with previous studies of the most prominent building blocks of drug molecules.[27,28]

As a conservative estimation, the virtual search space contained approximately 2 million molecules, assuming only two fragments to be linked together. Such a small search space must be seen as a test case for our approach, although it may represent a realistic scenario as the number of readily available building blocks for parallel synthesis is also limited.

Scheme 1[a]



[a] Designed molecules (**3**−**7**, **8**−**10**). The respective descriptor (GC or FP) and the distance index (Euc or Tan) to the reference molecule (**1** Gleevec or **2** Ugi Factor Xa inhibitor) are **3** GC Euc 2.450, **4** GC Euc 3.603, **5** GC Tan 0.993, **6** GC Tan 0.987, **7** GC Tan 0.986, **8** FP Euc 0.098, **9** FP Tan 0.838, and **10** GC Euc 3.871.

It is evident that more elaborate optimization techniques, including the use of adaptive search parameters, minimize the risk of premature convergence and can help escape local optima. We wish to stress that, in the present study, we did not explicitly analyze the search space for the existence of local optima, so that our results could be an artifact resulting from insufficient sampling. It remains an open question how rugged such a virtual fitness landscape actually is, and to what extent local optima influence the outcome of a de novo design run.[29]

A first test of Flux was to see whether the algorithm is able to redesign the two template structures. Figure 6 shows the development of the fitness during optimization. The fixed numbers of generations and offspring were sufficient in some cases to obtain the original template. Although this result shows that simplistic stochastic optimization was sufficient to cope with the search space size, in most of the design trials, the optimization process terminated before optimal fitness values were yielded. It is noteworthy that the optimization runs in the FP descriptor space (Figure 6A−D) did not reach a marked plateau within the given number of generations, whereas the processes converged in GC descriptor space (Figure 6E−H). A particular observation was made for the combination of GC descriptors and the continuous Tanimoto index (eq 2): for both template molecules, the optimization processes rapidly yielded high similarity values ($S > 0.9$) with small standard deviations. One might assume that close analogues of the template structures were assembled early in the course of molecule optimization. Still, we must be cautious with such an assumption due to the fact that different descriptors will result in different absolute levels of similarity values. Furthermore, we used two versions of the Tanimoto index, binary (eq 1) and continuous (eq 2).

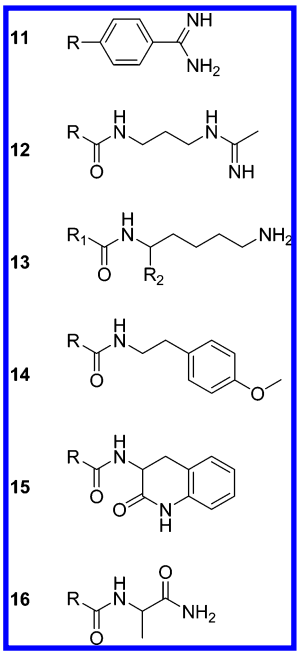
Table 1 summarizes the yielded average fitness values for the overall best 100 designed molecules. For both templates, the GC descriptor yielded better (i.e., greater) fitness values with the Tanimoto index S (eqs 1 and 2) than the FP descriptor but worse results with the Euclidian distance metric D (eq 3). The Tanimoto similarity was 0.98 on average for the GC data. Again, this result might suggest that this combination of descriptor and fitness index is principally suited for ligand-based de novo design. We wish to stress that such a general assumption might be misleading as a result of the problem of having different descriptors and distance indices that impede a straightforward comparison.

A further goal of our design exercise was to see whether our fragment approach does indeed result in meaningful variations of the template molecules, as we used only molecular building blocks that were obtained from dissecting known drugs and lead structures. An analysis of average values of MW of the 100 fittest designed structures reveals that the FP descriptor tends to lead to a higher molecular weight of the designed structures compared to the GC descriptor (Figure 7). No difference was observed between the two similarity indices used here. Again, one might speculate that the "quantitative" GC descriptor is more suitable for de novo design with regard to the molecular weight of the template structures (molecule **1**, MW = 494 Da; molecule **2**, MW = 572 Da): on average, the GC descriptor resulted in smaller differences of the MW between the templates and their respective design variants. This might be a consequence of the use of the GC substructure histogram ("holographic" or "quantitative" descriptor), whereas the FP descriptor was binary. Furthermore, as a result of the nature of Daylight FPs generation, namely, hashing and binary representation of substructure paths, information about molecular size (MW) and atom environments is not encoded explicitly. Therefore, molecules with a higher molecular weight were assembled. Binary fingerprints certainly are well-suited for molecule retrieval and proved their value in retrospective virtual screening exercises,[30,31] but their usefulness for prospective de novo design might be limited. We also investigated the clog *P* values of the designed structures and compared them to those of the template molecules but found no statistically significant difference for the GC and FP descriptors, nor for the two similarity indices (data not shown).

Molecules **3**−**7** (Scheme 1) represent selected high-scoring structures obtained in the various design runs using Gleevec as the template. While **3** is a very close analogue of **1**, the thiazole ring introduced in **4** represents a linker structure found with the GC descriptor and Euclidian distance metric on rank seven of the overall best designs. Such suggestions might be useful for further structure refinement. Structures **6** and **7** still contain several substructure elements that are also contained in the template **1** but are clearly less similar. Tricyclic moieties as in **7** were frequently observed in our designs, which might be attributed to the particular set of building blocks used for structure generation. With the exception of the iodine in molecule **4**, which bears the potential of unwanted reactivity, the designed structures appear reasonable and amenable to synthesis.

Molecules **8**−**10** represent selected high-scoring structures obtained with the Ugi structure **2** as the template. They contain variations of the peptide-like backbone of the template **2**, which can be desirable to obtain backup structures in a drug discovery project. Structures **8** and **9** contain well-known motifs of Factor Xa ligands or other members of trypsin-like serine protease inhibitors.[32] For example, they possess a benzamidine for binding to the S1 pocket and might be able to form additional interactions with the target protein, for example, hydrogen bonds to Gly216 (S3) found in the majority of inhibitor complexes with serine proteinases, or lipophilic interactions with residues forming the aryl-binding site (S4) of Factor Xa.[33] Automated docking studies could be applied subsequently to get an idea of potential binding modes.[34] Although candidate design **10** contains similar substructure elements, the compound very likely does not match the required pharmacophore for Factor Xa binding. Scheme 2 shows examples of molecular fragments that were found in top-ranking designs that were proposed by our algorithm as a substitution for the benzamidine in molecule **2**. Benzamidine **11** was observed most frequently as it is identical to the template, but we also observed arginine- and lysine-like fragments that might bear the potential to form hydrogen-bonding interactions with Asp189 at the bottom of the S1 pocket. Structures **14**−**16** represent some of the wilder designs that were proposed by Flux.

Scheme 2[a]



[a] Fragments of designed molecules that potentially bind to the S1 pocket of thrombin. All designs were obtained with the GC descriptor.Fragments **11**−**12**, **14**−**16**: Euclidian distance; molecule **13**: Tanimoto index.

It should be kept in mind that automatically designed structures cannot be expected to represent lead structures or even binding molecules. Rather, de novo design should be regarded as an "idea generator". Nevertheless, in related studies, it was demonstrated that fragment-based de novo design does actually yield potent novel ligands.[2,14,35]

A single de novo design run with Flux took approximately 150 min on a 2.4 GHz Intel Celeron with 2 GB of RAM. In our view, this time requirement does not render the current implementation applicable to high-throughput de novo design, for example, in conjunction with automated molecular docking for "in situ" molecule assembly. The most time-consuming part was shown to be the SMARTS/SMIRKS matching procedures. We are currently working on an improved version of the fragment assembly algorithm. Moreover, we are looking forward to generating building blocks from a data set that comprises a greater number of molecules so that we can apply our de novo design program to a larger search space. It is evident that a significantly enhanced search space requires robust search strategies for systematic navigation. Small search spaces are amenable to exhaustive enumeration.[36] But, in real-world scenarios, one has to cope with huge chemical spaces that demand more sophisticated searching.[2]

ACKNOWLEDGMENT

The authors are grateful to Norbert Dichter for setting up the LSF Linux cluster and to Dr. Petra Schneider for compiling the COBRA subsets. U.F. is thankful for a fellowship granted by Sanofi-Aventis Pharma Deutschland GmbH. This research was supported by Daylight Chemical Infromation Systems, Inc., the Deutsche Forschungsgemeinschaft (SFB 579, A11), and the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

REFERENCES AND NOTES

(1) Bleicher K. H.; Böhm H.-J.; Müller K.; Alanine, A. I. Hit and Lead Generation: Beyond High-throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.
(2) Schneider, G.; Fechner, U. Computer-based *De Novo* Design of Drug-like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.
(3) Lowrie, J. F.; Delisle, R. K.; Hobbs, D. W.; Diller, D. J. The Different Strategies for Designing GPCR and Kinase Targeted Libraries. *Comb. Chem. High Throughput Screening* **2004**, *7*, 495−510.
(4) Zhu, J.; Fan, H.; Liu, H.; Shi, Y. Structure-based Ligand Design for Flexible Proteins: Application of New F−DycoBlock. *J. Comput.-Aided Mol. Des*. **2001**, *15*, 979−996.
(5) Böhm, H.-J. Computational Tools for Structure-based Ligand Design. *Prog. Biophys. Mol. Biol.* **1996**, 197−220.
(6) Congreve, M.; Murray, C. W.; Blundell, T. L. Structural Biology and Drug Discovery. *Drug Discovery Today* **2005**, *10*, 895−907.
(7) Stahura, F. L.; Bajorath, J. New Methodologies for Ligand-based Virtual Screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.
(8) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 983−996.
(9) Lipinski, C.; Hopkins, A. Navigating Chemical Space for Biology and Medicine. *Nature* **2004**, *432*, 855−861.
(10) Böhm, H.-J.; Banner, D. W.; Weber, L. Combinatorial Docking and Combinatorial Chemistry: Design of Potent Non-peptide Thrombin Inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 51−56.
(11) Schneider, P.; Schneider, G. Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR Comb. Sci.* **2003**, *22*, 713−718.
(12) Lewell, X. O.; Budd, 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.
(13) *Daylight Theory Manual*. http://daylight.com/dayhtml/doc/theory/theory.toc.html.
(14) Schneider, G.; Clement-Chomienne, O.; Hilfiger, L.; Schneider, P.; Kirsch, S.; Böhm, H.-J.; Neidhart, W. Virtual Screening For Bioactive Molecules by Evolutionary *De Novo* Design. *Angew. Chem., Int. Ed.* **2000**, *39*, 4130−4133.
(15) Viswanadhan, V. N.; Ghose, A. K.; Reyankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for An Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(16) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating Between Drugs and Nondrugs. *J. Med. Chem*. **1998**, *41*, 3325−3329.
(17) Roche, O.; Schneider, P.; Zuegge, J.; Guba, W.; Kansy, M.; Alanine, A.; Bleicher, K.; Danel, F.; Gutknecht, E. M.; Rogers-Evans, M.; Neidhart, W.; Stalder, H.; Dillon, M.; Sjogren, E.; Fotouhi, N.; Gillespie, P.; Goodnow, R.; Harris, W.; Jones, P.; Taniguchi, M.; Tsujii, S.; von der Saal, W.; Zimmermann, G.; Schneider, G. Development of a Virtual Screening Method for Identification of "Frequent Hitters" in Compound Libraries. *J. Med. Chem.* **2002**, *45*, 137−142.
(18) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Redwood City, CA, 1989.
(19) Darwin, C. *On the Origin of Species a Facsimile of the First Edition*; Harvard University Press: Cambridge, Massachusetts, 1975.
(20) Saravanan, N.; Fogel, D. B.; Nelson, M. A Comparison of Methods for Self-Adaptation in Evolutionary Algorithms. *Biosystems* **1995**, *36*, 157−166.
(21) Rechenberg, I. *Evolutionsstrategie '94*; Frommann-Holzboog: Stuttgart, Germany, 1994.
(22) Schneider, G.; Wrede, P. Artificial Neural Networks for Computer-Based Molecular Design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175−222.
(23) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(24) Buchdunger, E.; Zimmermann, J.; Mett, H.; Meyer, T.; Müller, M.; Druker, B. J.; Lydon, N. B. Inhibition of the Abl Protein-Tyrosine Kinase In Vitro and In Vivo By a 2-Phenylaminopyrimidine Derivative. *Cancer Res.* **1996**, *56*, 100−104.
(25) (a) Ugi, I.; Meyr, R.; Fetzer, U.; Steinbrückner, C. Versuche mit Isonitrilen. *Angew. Chem*. **1959**, *71*, 386. (b) Ugi, I.; Steinbrückner, C. Über ein neues Kondensations-Prinzip. *Angew. Chem*. **1960**, *72*, 267−268.
(26) Weber, L. The Application of Multi-Component Reactions in Drug Discovery. *Curr. Med. Chem.* **2002**, *9*, 2085−2093.
(27) (a) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893. (b) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(28) (a) Wagener, M.; van Geerstein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 280−292. (b) Xu, Y. J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features That Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926. (c) Niwa, T. Prediction of Biological Targets Using Probabilistic Neural Networks and Atom-Type Descriptors. *J. Med. Chem.* **2004**, *47*, 2645−2650.
(29) (a) Kauffman, S. A. *The Origins of Order: Self-Organization and Selection in Evolution*; Oxford University Press: New York, 1993. (b) Schneider, G.; So, S.-S. *Adaptive Systems in Drug Design*; Landes Bioscience: Georgetown, TX, 2001.
(30) (a) Willett, P. Similarity-Based Approaches to Virtual Screening. *Biochem. Soc. Trans.* **2003**, *31* (3), 603−606. (b) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(31) Fechner, U.; Paetz, J.; Schneider, G. Comparison of Three Holographic Fingerprint Descriptors and Their Binary Counterparts. *QSAR Comb. Sci*. **2005**, *24*, 961−967.
(32) Banner, D. Principles of Enzyme−Inhibitor Design. In *Protein−Ligand Interactions: From Molecular Recognition to Drug Design*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, Germany, 2003; p 163.
(33) (a) Brandstetter, H.; Kühne A.; Bode, W.; Huber R.; von der Saal, W.; Wirthensohn, K.; Engh, R. A. X-ray Structure of Active Site-Inhibited Clotting Factor Xa−Implications for Drug Design and Substrate Recognition. *J. Biol. Chem.* **1996**, *271*, 29988−29992. (b) Krovat, E. M.; Frühwirth, K. H.; Langer, T. Pharmacophore Identification, In Silico Screening, and Virtual Library Design for Inhibitors of the Human Factor Xa. *J. Chem. Inf. Model*. **2005**, *45*, 146−159.
(34) (a) Alvarez, J. C.; High-Throughput Docking as a Source of Novel Drug Leads. *Curr. Opin. Chem. Biol.* **2004**, *8*, 365−370. (b) Schneider, G.; Böhm, H.-J. Virtual Screening Fast Automated Docking Methods. *Drug Discovery Today* **2002**, *7*, 64−70.
(35) Rogers-Evans, M.; Alanine, A. I.; Bleicher, K. H.; Kube, D.; Schneider, G. Identification of Novel Cannabinoid Receptor Ligands Via Evolutionary De Novo Design and Rapid Parallel Synthesis. *QSAR Comb. Sci*. **2004**, *23*, 426−430.
(36) Boda, K. SynSPROUT: Generating Synthetically Accessible Ligands by *De Novo* Design. Ph.D. Thesis, School of Chemistry, University of Leeds, Leeds, U. K., 2002.

CI0503560