Scaffold-Hopping Potential of Fragment-Based *De Novo* Design: The Chances and Limits of Variation

Bjoern A. Krueger[1], Axel Dietrich[2], Karl-Heinz Baringhaus[2] and Gisbert Schneider[*,1]

[1]*Johann Wolfgang Goethe-Universität, Institut für Organische Chemie und Chemische Biologie, Siesmayerstr. 70, D-60323 Frankfurt am Main, Germany*

[2]*Sanofi-Aventis Deutschland, Drug Design, Industriepark Höchst, Frankfurt am Main, Germany*

Abstract: The identification of new lead structures is a pivotal task in early drug discovery. Molecular *de novo* design of ligand structures has been successfully applied in various drug discovery projects. Still, the question of the scaffold hopping potential of drug design by adaptive evolutionary optimization has been left unanswered. It was unclear whether *de novo* design is actually able to leap away from given chemotypes ("activity islands"), allowing for rescaffolding of compounds. We have addressed these questions by scrutinizing different scoring functions of our *de novo* design software Flux for their ability to enable scaffold-hops for various target classes. We evaluated both the potential bioactivity and the scaffold diversity of *de novo* generated structures. For several target classes, known lead structures were reconstructed by the *de novo* algorithm ("lead-hopping"). We demonstrate that for one or multiple templates of a given chemotype, other chemotypes are reached during *de novo* compound generation, thus indicating successful scaffold-hops.

Keywords: Drug discovery, *de novo* design, machine learning, scaffold-hopping, similarity, building block.

INTRODUCTION

De novo design describes the concept of generating new molecular structures with desired pharmacological activities "from scratch", thus complementing two established methods in drug discovery, high-throughput screening (HTS) and virtual screening [1]. Both methods search through large chemical spaces mostly generated using real or virtual parallel synthesis to find suitable candidates for pharmaceutical development [2]. A differentiation can be made between receptor- and ligand-based design concepts. While the first approach focuses on a given structure of a receptor's binding site, strictly ligand-based design has its primary application domain where substantial information about a target's three-dimensional (3D) molecular structure is unavailable, *e.g.,* for many G-protein coupled receptors (GPCRs) [3]. The required information about protein-ligand interaction is extracted from available reference ligands and used as a guideline for the generation of new molecules [4]. In fragment-based *de novo* design, new molecules are assembled from small (sub-)molecular fragments or "building blocks" [5], often by applying pseudo-chemical rules in their recombination, *e.g.* RECAP ("Retrosynthetic Combinatorial Analysis Procedure") [6]. Although the use of fragments instead of atoms reduces the sampling rate of chemical space (and thus increasing the risk of "missing" novel compounds), the assembly of novel structures out of a given set of building blocks according to (pseudo-) chemical rules provides for a likely better synthetic accessibility rate of the new compounds [7].

The aim of the present study was to search for novel and potentially active compounds in different target areas. Using an evolutionary algorithm, we performed molecular mutation and recombination of building blocks, and evaluated the "fitness" of newly generated structures. Designs were optimized by rating their similarity to one or more given reference molecules, the "templates". Finally, the scaffold diversity of the generated *de novo* constructs was evaluated and the "scaffold hopping" potential was assessed [8].

MATERIALS & METHODS

Methodological Concept

New molecular structures were generated *de novo* using the software Flux [9] ("Fragment-based ligand building reaxions") version 0.44. Flux generates variations of molecular structures by exchanging a randomly determined molecular fragment or "building block" with another building block of compatible chemical type, drawn from a previously generated stock of building blocks. The "fitness" of the new structures is evaluated by similarity analysis to one or more given reference structures or "templates". An evolutionary algorithm is used to identify the "fittest" *de novo* constructs that become "parents" of the next cycle ("generation") of molecule mutation. This step is repeated until a defined number of generations is reached, or the fitness score converges.

We generated *de novo* constructs for four therapeutical targets: angiotensin converting enzyme (ACE) inhibitors, angiotensin-II receptor antagonists, aldose reductase inhibitors, and dopamine receptor antagonists. Three design approaches were followed:

1. A single reference structure was presented for evolutionary optimization, mimicking the situation in a so-called "me too" project, where a single molecular structure is often the only information at hand. We performed a literature search, choosing a "first-in-class" compound as a typical representative, *e.g.,* the

[*]Address correspondence to this author at the Johann Wolfgang Goethe-Universität, Institut für Organische Chemie und Chemische Biologie, Siesmayerstr. 70, D-60323 Frankfurt, Germany; Tel: +49 (0) 69 798 24873/4; Fax: +49 (0) 69 798 24880; E-mail: g.schneider@chemie.uni-frankfurt.de

1386-2073/09 $55.00+.00

first drug on the market in the corresponding target class.

2. Five most similar reference compounds with known activity on a target were used as design templates. Structures were compiled from the Derwent World Drug Index (WDI, 2006; Thomson Scientific, Philadelphia, USA). The templates were equally weighted for the *de novo* design run.

3. To sample structurally more diverse ligands of a given target, we used the five structurally most dissimilar compounds from the WDI, covering different chemotypes or scaffold classes.

In all three approaches, we started 100 parallel and independent *de novo* runs with the software Flux. A run lasted for 100 generations with 100 individual structures each. The design resulted in several hundred thousand unique virtual *de novo* constructs.

Generation of Drug-Like Molecular Fragments

We used the WDI as a starting point for a collection of drug-like molecules. After "washing" this set of substances by removing ions, radioactives, and non-organic compounds, further filters were applied subsequently: First, we removed unwanted mechanisms of action to discard substances not targeting human pharmacology, *e.g.*, insecticides. Then we applied a REOS ("rapid elimination of swill") [10] filter followed by a filter according to Hann [11]. In these steps, structures which have a potential negative effect on binding and affinities were removed, *e.g.*, anhydrides, sulphohalides, peroxides or alpha-beta-unsaturated nitriles. These unwanted substructures were provided as a list of SMIRKS (Daylight Chemical Information Systems Inc., Aliso Viejo, CA, USA). A search for these substructures was performed using the structure search feature implemented in Pipeline Pilot version 5.5 (SciTegic Inc., San Diego, CA, USA). Finally, we applied a modified version of Lipinski's rule of five [12]. In contrast to the original rules, we discarded a molecule if a single criterion was violated (molecular weight < 700 AND number of organic atoms ≥ 1 AND number of rotatable bond ≤ 10 AND number of hydrogen bond acceptors < 10 AND number of hydrogen bond donors ≤ 7), thereby rigorously reducing the number of WDI structures. Overall, a total number of 24,301 "drug-like" structures remained. This set of compounds was subsequently subjected to a dissection according to the RECAP rules by applying the software RetroFlux, a part of the Flux package, resulting in a set of 10,497 unique building blocks of different RECAP-type.

Similarity Searching

To establish a means for comparison between the *de novo* constructs and known reference structures, two different similarity metrics were used [13]: CATS2D [8] descriptors and functional connectivity fingerprints (FCFP-4; SciTegic Inc., San Diego, CA, USA).

CATS2D: Calculation of the topological CATS descriptor [8] in this study was done by applying the internal implementation within the *de novo* software Flux [9], and *via* the software SpeedCATS [14] (version 1.1; Schneider Consulting GbR, Oberursel, Germany). This two-dimensional descriptor encodes potential two-point pharmacophores in the form of an alignment-free correlation vector [14]. For each molecule, the occurrence frequencies of 15 potential pharmacophore points (PPP; hydrogen-bond donor, hydrogen-bond acceptor, lipophilic, positive ionizable, negative ionizable) pairs were determined for topological distances of 0 to 9 bonds. This resulted in a 150-dimensional descriptor vector. The similarity of CATS descriptor vectors **A** and **B** was computed using the Manhattan distance metric D (Eq. **1**). The distance D(**A**,**B**) corresponds to a value in the interval [0,+∞] where a value of 0 indicates identity between **A** and **B**. When the distance D(**A**,**B**) increases, so does the dissimilarity between **A** and **B**.

$$\mathrm{D}(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{150} \left| A_i - B_i \right| \qquad \text{Eq. (1)}$$

Functional Connectivity Fingerprints: The principle of connectivity fingerprints [15] is based on the occurrence of specific fragments in ring-like substructures. Each atom is represented as a string of "extended connectivity" values that are determined by a modified Morgan's algorithm [16]. In an initial step, each atom is assigned an atom type (hydrogen-bond donor, hydrogen-bond acceptor, positively ionizable, negatively ionizable, aromatic, halogen). This code in combination with information about the atoms' bonds and neighbors are projected into an address space of 2^{32} Bit. In this study, the FCFP-4 descriptor has been used considering topological distances of up to four bonds. The computation of the FCFP-4 descriptors was carried out using the software Pipeline Pilot 5.5 (SciTegic Inc., San Diego, CA, USA). Similarity searches were performed using the Pipeline Pilot function "similarity search", which provides an implementation of the Tanimoto coefficient [17,18] (Eq. **2**),

$$S_{A,B} = \frac{c}{a+b-c}, \qquad \text{Eq. (2)}$$

where *a* represents the number of bits set in bitstring *A*, *b* is the number of bits set in bitstring *B*, and *c* is the number of corresponding bits set in both molecules. The value of *T* ranges from zero (no similarity) to one (maximum similarity).

Selection of Target Reference Structures

Angiotensin converting enzyme inhibitors: We performed a database search for the mechanism of action keyword "ACE-Inhibitor" in the Derwent World Drug Index (version July 2006) using the software ISIS/base version 2.5 (MDL Information Systems). To ensure comparability to our previously collected fragment set, a filtering step using the above mentioned drug-like filter was subsequently performed resulting in a set of reference actives. As a template structure for evolutionary *de novo* design, we chose Lisinopril (**1**), which together with the smaller Captopril (**2**), is the only non-prodrug ACE inhibitor on the market to our knowledge.

For the second *de novo* approach with a set of focused reference structures of the same or similar chemotype, we conducted a similarity search in the reference actives for

Lisinopril (FCFP-4 descriptor, Tanimoto index). The four most similar substances (**3**-**6**) and Lisinopril (**1**), formed our set of five “focused” reference structures.

In a third approach, we wanted to present a set of most diverse target reference structures to the evolutionary *de novo* algorithm to span the whole range of known chemotypes in this target class. We performed a search for five maximally diverse ACE inhibitors using the Pipeline Pilot command “select diverse compounds” (**1**, **2**, **7**, **8**, **9**), which is based on a maximum dissimilarity analysis (FCFP-4 descriptor, Tanimoto index) [19-22]:

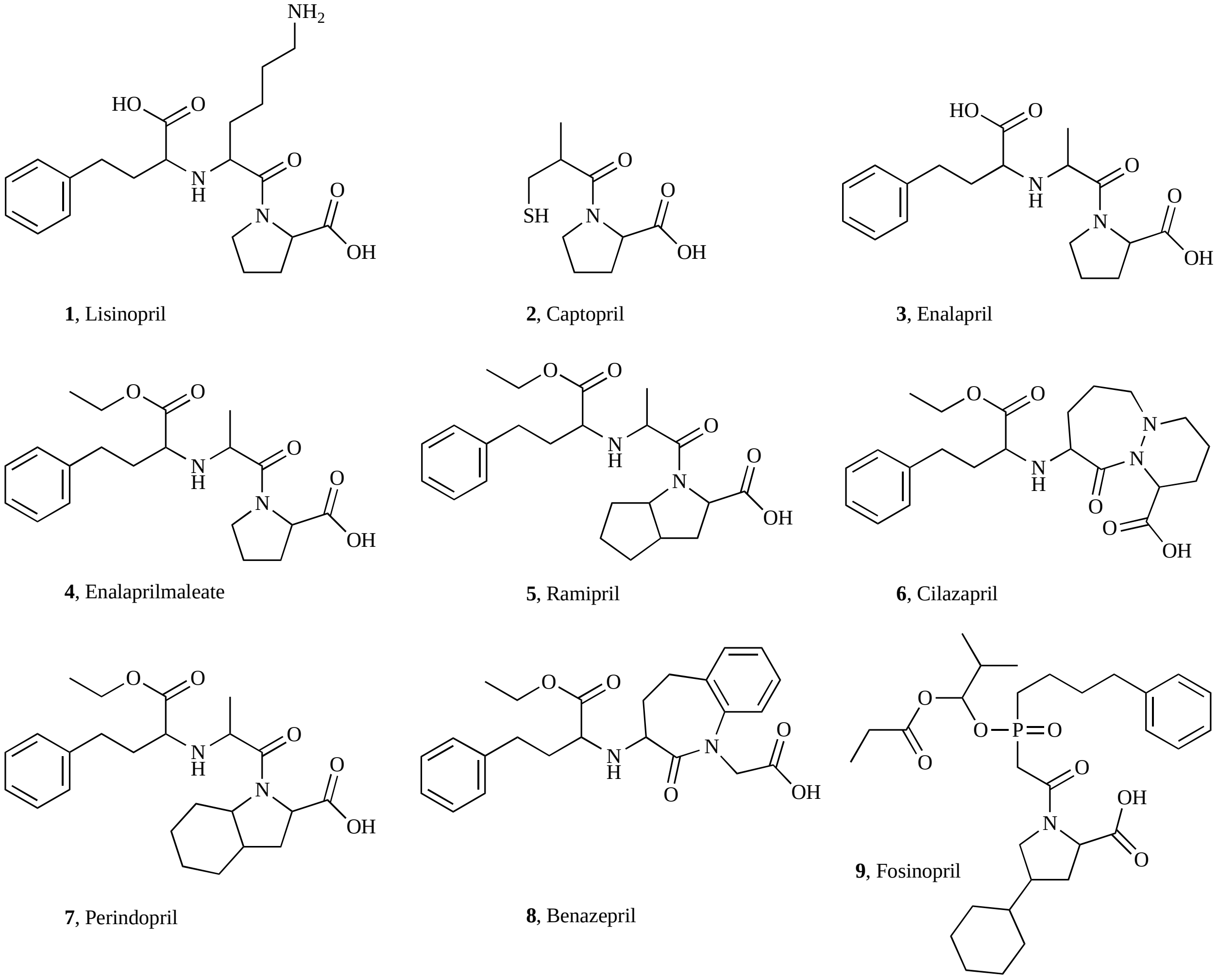
1. Initialize the subset by drawing one compound from the pool of molecules.
2. Calculate the (dis-)similarity between each remaining compound in the pool and the compounds in the subset.
3. Select the compound from the pool that is most dissimilar to the compounds in the subset. Add the newly selected compound to the subset.
4. Iterate from Step 1 until the desired number of compounds in the subset is reached.

Aldose reductase inhibitors: A WDI database search with the mechanism of action keyword “Aldose-reductase-inhibitor” (ARI) produced a set of actives for a second target class, which were subsequently filtered using our drug-likeness criteria. We chose Epalrestat (**10**) as template for *de novo* design with a single template structure. A similarity search (FCFP-4, Tanimoto) for Epalrestat in our actives set provided us with four similar compounds (**11-14**), which were used as templates for focused *de novo* design. Performing a maximum dissimilarity search, we found the five most dissimilar structures in the ARI actives set of actives (**15-19**).

Angiotensin-II receptor antagonists: To collect a set of actives for antagonists of the angiotensin-II receptor (AT2), we performed a database search in the WDI for the mechanism of action keyword “Angiotensin-2-inhibitor” once again followed by a filtering step with our custom drug-like filter. We chose the “first-in-class” drug Losartan (**20**) as reference structure for optimization with a single reference structure.

For parallel focused optimization with five similar reference structures, we found four substances (**21-24**) of our compiled actives set to be most similar (FCFP-4, Tanimoto) to Losartan.

Finally, a maximum dissimilarity search (FCFP-4. Tanimoto) resulted in the five diverse compounds (**24-29**) spanning the complete structural range of AT2 antagonists.



1, Lisinopril

2, Captopril

3, Enalapril

4, Enalaprilmaleate

5, Ramipril

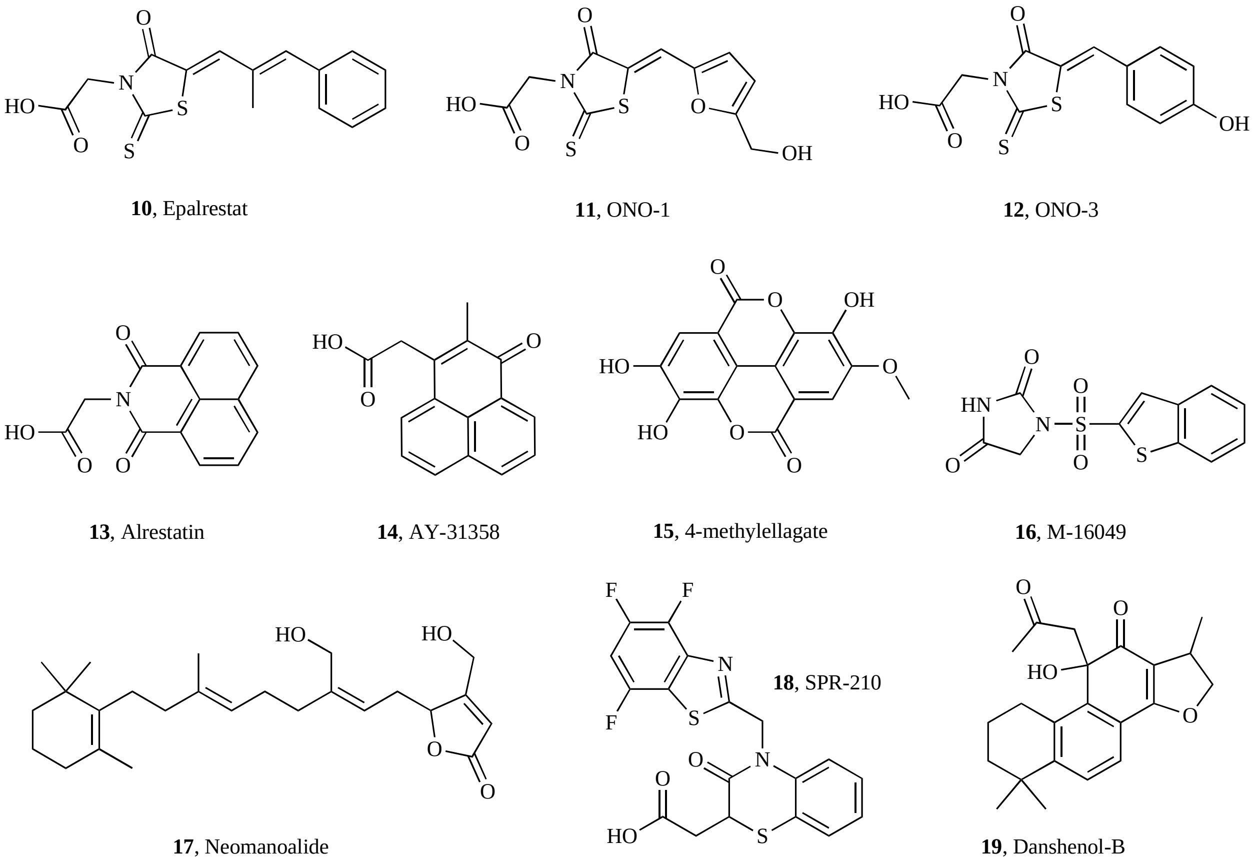
6, Cilazapril

7, Perindopril

8, Benazepril

9, Fosinopril

Scheme 1.



Scheme 2.

Similarity Anaysis of *De Novo* Generated Compounds

During evaluation of the *de novo* generated molecules we were challenged with the problem of large numbers. While a similarity analysis or even activity prediction is difficult with tens of thousands of molecules, it becomes a major computational task with several hundred thousand or even millions of structures. Thus, we needed a method to reduce the number of probable inactive *de novo* compounds while retaining a high proportion of potential actives.

We decided to calculate a target class-dependent "best-split" similarity value to establish a pre-screening similarity search with this best-split value as similarity radius or threshold around each of our known actives from the previously compiled actives set. If a *de novo* compound is found within a similarity radius smaller than the calculated threshold value for any of our known actives, the compound was classified as "potentially active" according to the "similar property principle" [23] or the principle of "neighborhood behavior" [24], respectively. *De novo* compounds thus labeled as potential actives were then subjected to further analysis. *De novo* compounds not showing a similarity of at least the defined threshold value to any known active were discarded due to high probability of inactivity. Again, we applied the Tanimoto index between two FCFP-4 fingerprints as similarity measurement.

As similarity threshold for a given target class, we determined the corresponding receiver operating characteristic (ROC). Enhancing the information content resulting e.g., from determining the Matthews correlation coefficient (MCC) [25,26], ROC provides a means of assuring the correctness of a binary classification [27,28] by calculating both sensitivity (*Se*, percentage of true positives found, Eq. **3**) and specificity (*Sp*, percentage of true negatives found, Eq. **4**):

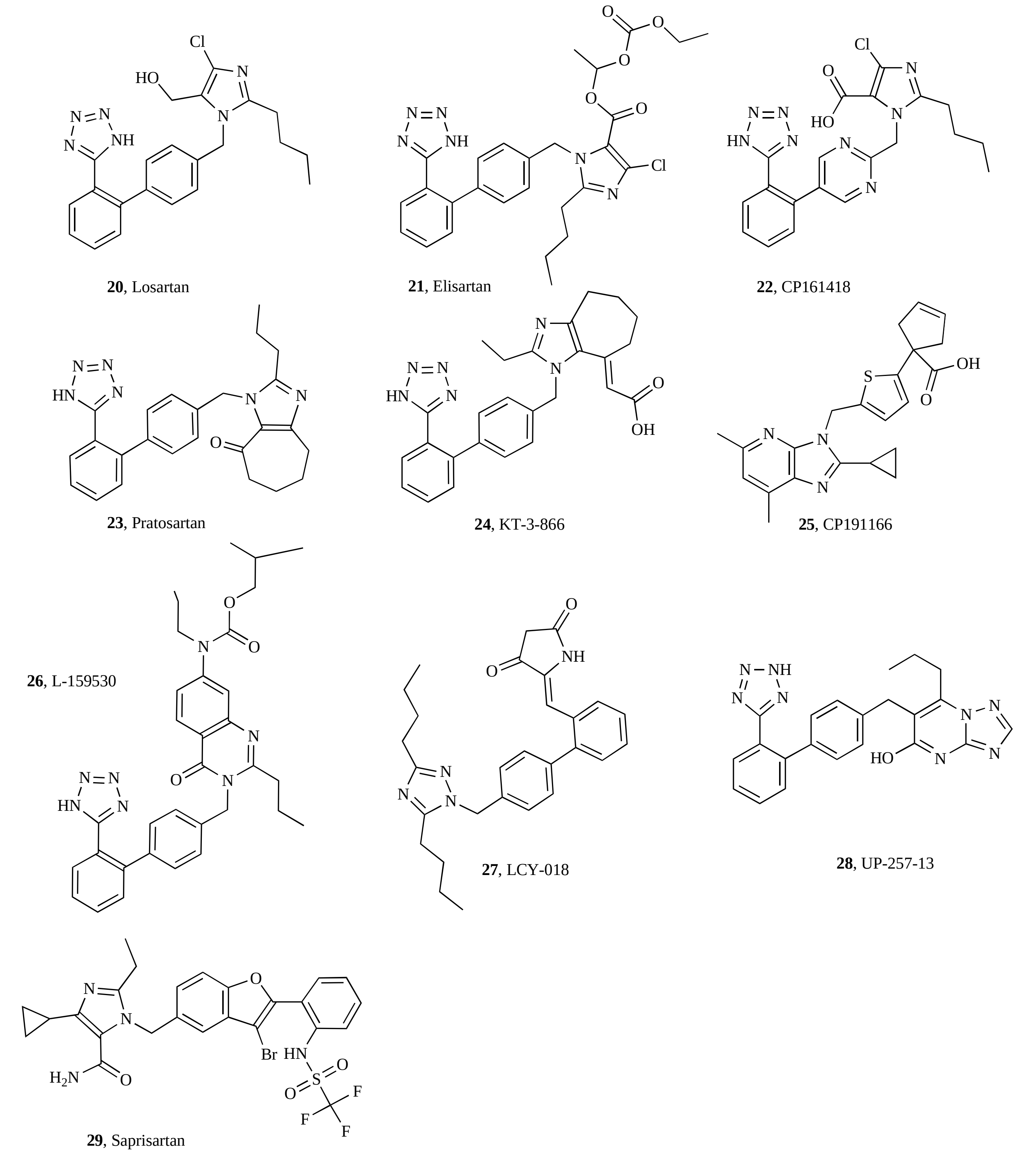
$$Se = \frac{TP}{TP + FN}, \qquad \text{(Eq. 3)}$$

$$Sp = \frac{TN}{TN + FP}, \qquad \text{(Eq. 4)}$$

where *TP* is the number of true positives, *FP* is the number of false positive, *TN* is the number of true negatives, and *FN* is the number of false negatives.

When sensitivity is plotted against (1 - specificity), the area under the ROC curve (AUROC) indicates the overall quality of the prediction. At the same time, the plot provides an optimum value for the "best-split" separation between the two classes [29]. While ROC plots and AUROC scores have been successfully applied as a means of evaluating prediction correctness in many research fields [30, 31], they have only recently been adopted as standard statistics in chemoinformatics [32].

To evaluate sensitivity and specificity, we calculated the similarity (FCFP-4, Tanimoto score) of each of the compounds from our set of known actives to all available known inactives. As inactives, we chose all compounds from the



20, Losartan

21, Elisartan

22, CP161418

23, Pratosartan

24, KT-3-866

25, CP191166

26, L-159530

27, LCY-018

28, UP-257-13

29, Saprisartan

Scheme 3.

WDI filtered with our previously mentioned filters, removing the target class's known actives beforehand. We obtained a ROC value specific for each known active compound of the target class. To get a global ROC value, we calculated the mean of all individual ROC values. Finally, we used the resulting global ROC value as similarity threshold for a similarity search with all compounds from our set of known actives in a given target class.

The ROC value can be interpreted as a value up to which a given descriptor (here: FCFP-4) is able to optimally discriminate between actives and inactives of a target class. Below this value is *terra incognita*, which cannot be classified using the corresponding descriptor. One result is shown as an example in Fig. (**1**).

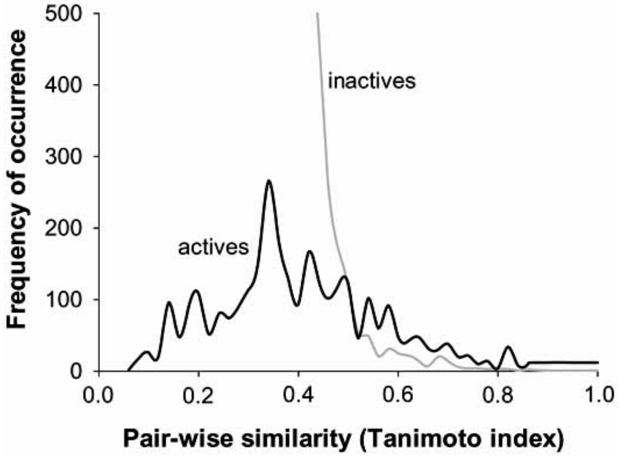


Fig. (1). Motivation of a "best-split" similarity threshold. The black line shows a similarity histogram, *e.g.,* the distribution of frequencies of occurrence for pair-wise similarity scores in our set of actives (here: ACE inhibitors). The grey line indicates the corresponding similarity histogram for the target's inactives. Since the frequency of occurrence for inactives with a pair-wise similarity of 0.0 is higher than 500, only the beginning of the inactives' histogram is visible. It can be seen that at a given point, the "similarity threshold", the number of inactives (false positives) starts to grow exponentially while the number of actives (true positives) is only mildly increased. The similarity threshold thus represents a "best-split" between the number of true positives and false positives for a specific target class. Please note that the similarity thresholds used in this study have been acquired by computational means (ROC scores).

Evaluation of Scaffold Diversity

To analyze the diversity of the newly generated *de novo* constructs, we generated the Murcko scaffolds [33] for each *de novo* compound and all members of the actives set using a custom MOE script [34] (Molecular Operating Environment, version 08.2006, Chemical Computing Group). Murcko scaffolds consist only of a molecule's cyclic substructures with their shortest linkers, that is, the shortest bond path between two rings.

In two different approaches, atom-based Murcko scaffolds (information about hetero atoms in rings and linkers is retained) and graph-based scaffolds (all information about hetero items in rings and linkers is discarded) were generated (Fig. **2**). For each scaffold, both MACCS keys [35] and FCFP-4 fingerprints were computed and clustered by maximum dissimilarity employing "complete linkage" (Eq. **5**): The distance of two clusters **A** and **B** is defined by the maximal distance between two members *a* and *b* of clusters **A** and **B**.

The 166 public MACCS keys were computed using MOE version 08.2006, the FCFP-4 fingerprints by Pipeline Pilot 5.5. All clustering was done using the clustering component of Pipeline Pilot. The resulting clusters gave a first impression of number, size and novelty of activity islands into which the *de novo* constructs aggregated.

$$D(\mathbf{A},\mathbf{B}) = \max_{a\in\mathbf{A},b\in\mathbf{B}}\{D(a,b)\} \qquad \text{(Eq. 5)}$$

RESULTS AND DISCUSSION

Generation of Drug-Like Molecular Fragments

After "washing" the Word Drug Index, 80,499 structures remained. Applying several filtering steps (REOS, Hann, drug-like) yielded 24,301 compounds, which formed the base of our drug-like molecule pool. Dissection with our implementation of RECAP in the software *Retroflux* resulted in 10,497 unique fragments with at least one cleaved (unvalenced) bond. These fragments were then used as building blocks during the *de novo* generation of molecules. As reference sets, we compiled 73 ACE inhibitors, 80 aldose reductase inhibitors and 59 AT2 antagonists, all with known activity towards their respective targets.

De Novo Design with a Single Reference Structure

ACE inhibitors: During our *de novo* design runs using Lisinopril (**1**) as template structure, 259,829 unique *de novo* constructs were generated. Lisinopril was not rebuilt during the evolutionary optimization, but the best *de novo* generated construct approached Lisinopril with a Manhattan distance of 8.8 (Manhattan on CATS2D Fingerprints).

In the next step, we performed a similarity search for all compounds from the ACE actives set, resulting in 12 known actives (**30-41**) being found in the *de novo* constructs, some of them with a considerable dissimilarity to the reference Lisinopril (annotated as Tanimoto score in brackets next to the corresponding structure).

AT2 antagonists: A number of 972,495 unique virtual constructs were computed during the *de novo* design with Losartan (**20**) as reference structure, and again the evolutionary algorithm was unable to reproduce the template structure. The best-ranked *de novo* constructs approached Losartan up to a Manhattan distance of 3.4 (CATS2D). A number of 5,792 virtual constructs had a similarity score ≥ 0.48 (target class-specific similarity threshold; FCFP-4, Tanimoto) to

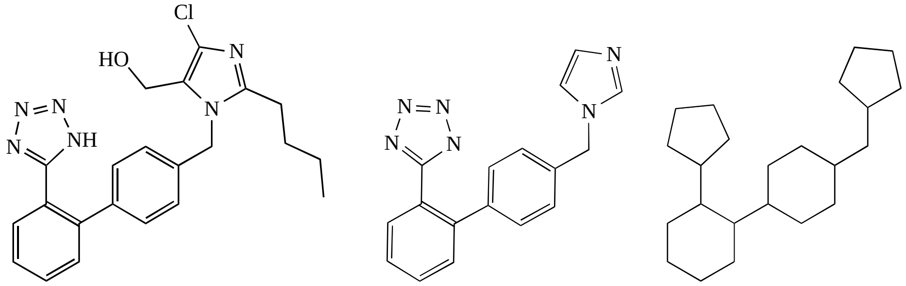
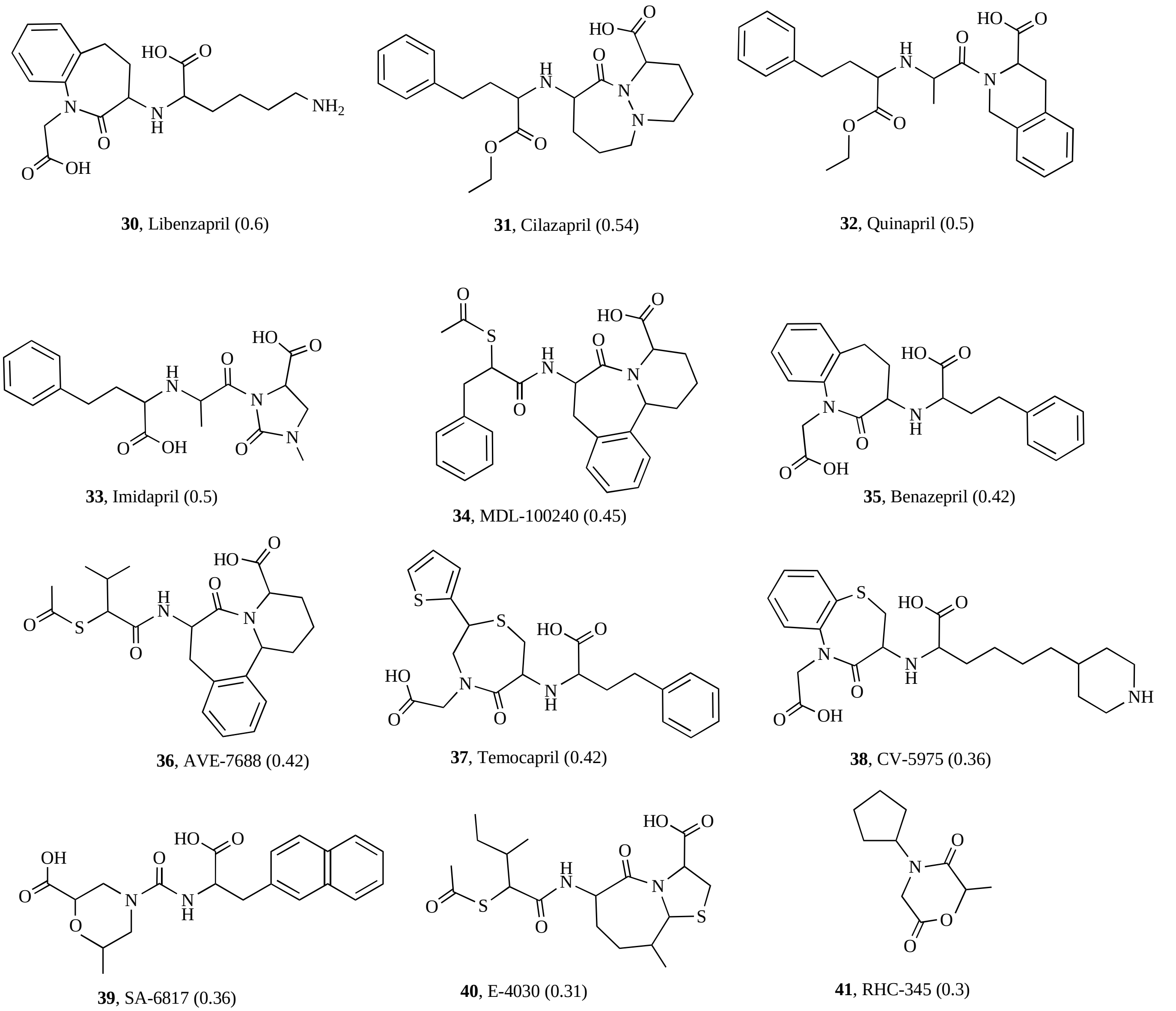


Fig. (2). Example for a generation of Murcko scaffolds from Losartan (left): atom-based (middle) and graph-based (right).



30, Libenzapril (0.6)

31, Cilazapril (0.54)

32, Quinapril (0.5)

33, Imidapril (0.5)

34, MDL-100240 (0.45)

35, Benazepril (0.42)

36, AVE-7688 (0.42)

37, Temocapril (0.42)

38, CV-5975 (0.36)

39, SA-6817 (0.36)

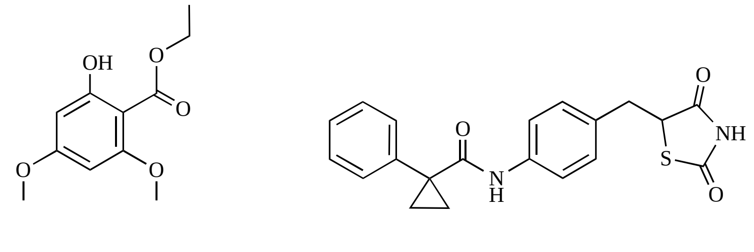
40, E-4030 (0.31)

41, RHC-345 (0.3)

Scheme 4.

any of the compounds from the known actives set and were classified as potential actives.

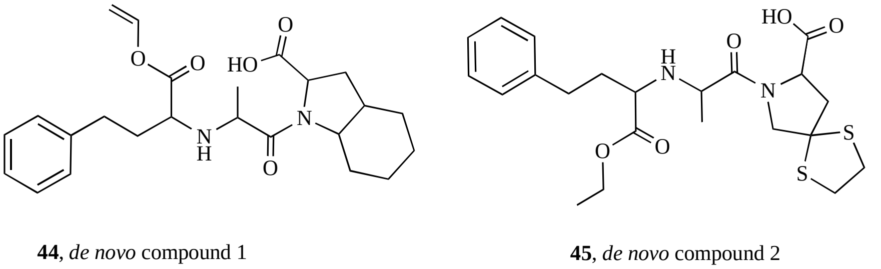
Aldose reductase inhibitors: A number of 250,753 virtual constructs were generated during *de novo* design with Epalrestat (**10**) as single template. The best-ranked *de novo* approached Epalrestat up to a Manhattan distance of 3.7 (CATS2D descriptor). Two known actives, (**42**) and (**43**), from WDI were rebuilt. A number of 1,030 virtual constructs had a similarity score ≥ 0.59 (target class-specific similarity threshold; Tanimoto, FCFP-4) to any compound from the actives set and thus were considered potentially active.



42, Xantoxylin-Carboxylate-Ethylester (0.21)

43, DN-108 (0.16)

Scheme 5.


44, *de novo* compound 1

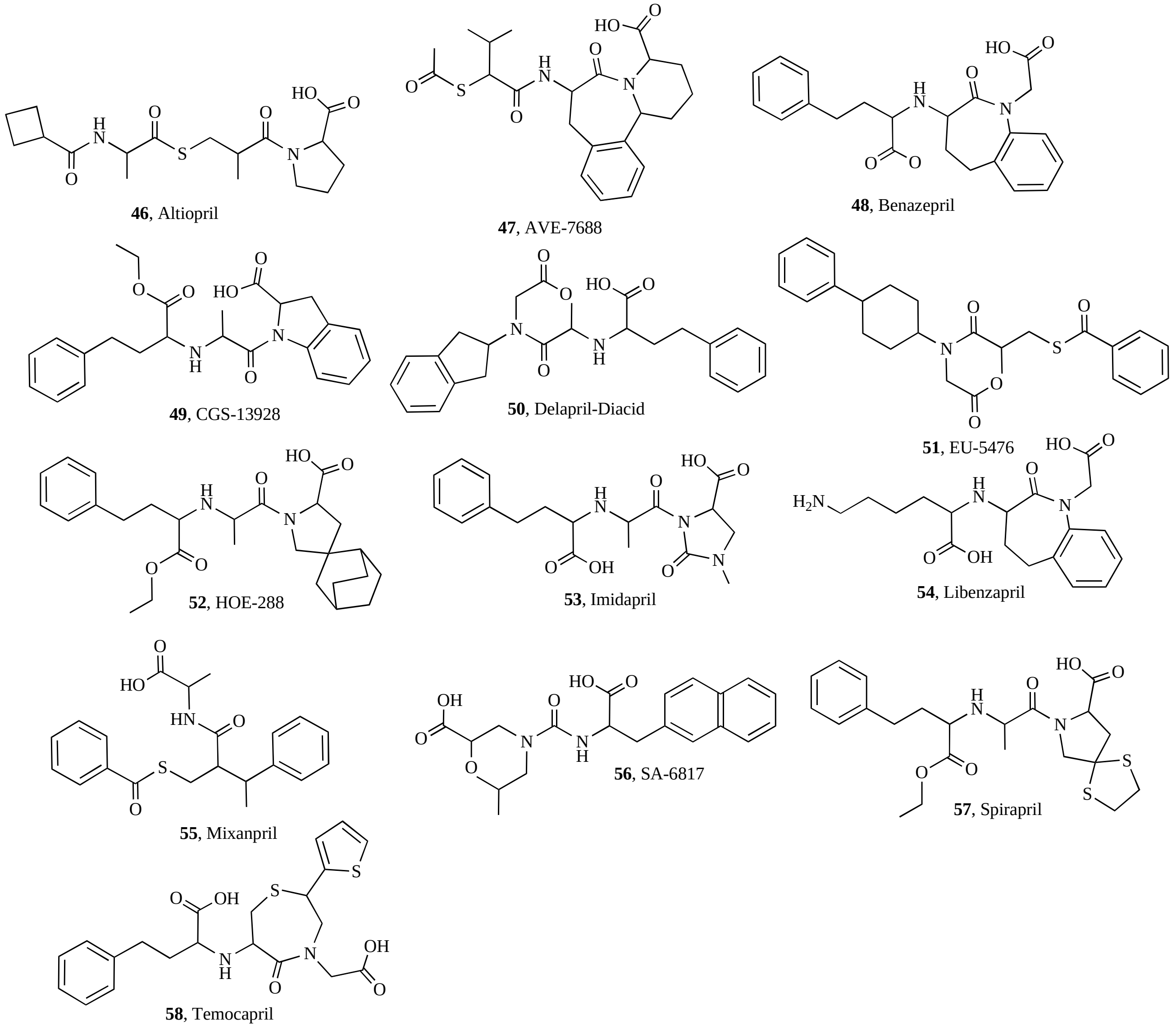
45, *de novo* compound 2

Scheme 6.

De Novo Design with a Focused Set of Five Similar Templates

ACE inhibitors: In this approach, a number of 329,244 unique virtual constructs were generated. The two best-ranked *de novo* constructs (**44**) and (**45**) each reached a distance of *D* = 5.2 (Manhattan, CATS2D).

Structure (**44**) was reconstructed three times during the *de novo* design, whereas molecule (**45**) was reconstructed two times. A number of 50 additional constructs scored best by Flux were found in the immediate chemical neighborhood of these two molecules (data not shown), implicating a convergence of the algorithm on a specific activity island. Although none of the given template structures could be found during *de novo* generation, 13 compounds (**45**-**58**) that were not presented as template were reconstructed from the actives reference set.


46, Altiopril

47, AVE-7688

48, Benazepril

49, CGS-13928

50, Delapril-Diacid

51, EU-5476

52, HOE-288

53, Imidapril

54, Libenzapril

55, Mixanpril

56, SA-6817

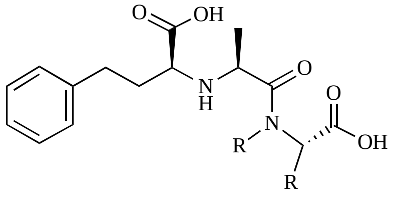
57, Spirapril

58, Temocapril

Scheme 7.

While visually inspecting the generated *de novo* constructs, we found a particular substructural motif (**59**) frequently among the best ranked *de novo* constructs. We assume this to be the Enalapril motif, the maximum common substructure of the five presented template compounds (**1**) and (**3**-**6**).

To test this hypothesis, we conducted a maximum common substructure search (MCSS) [36,37] using the Pipeline Pilot component "Substructure Search" for the Enalapril motif (**59**) in the 1,000 *de novo* constructs ranked highest according to the Flux fitness score. The maximum common substructure could be found in 219 of these virtual constructs.



59, Enalapril motif

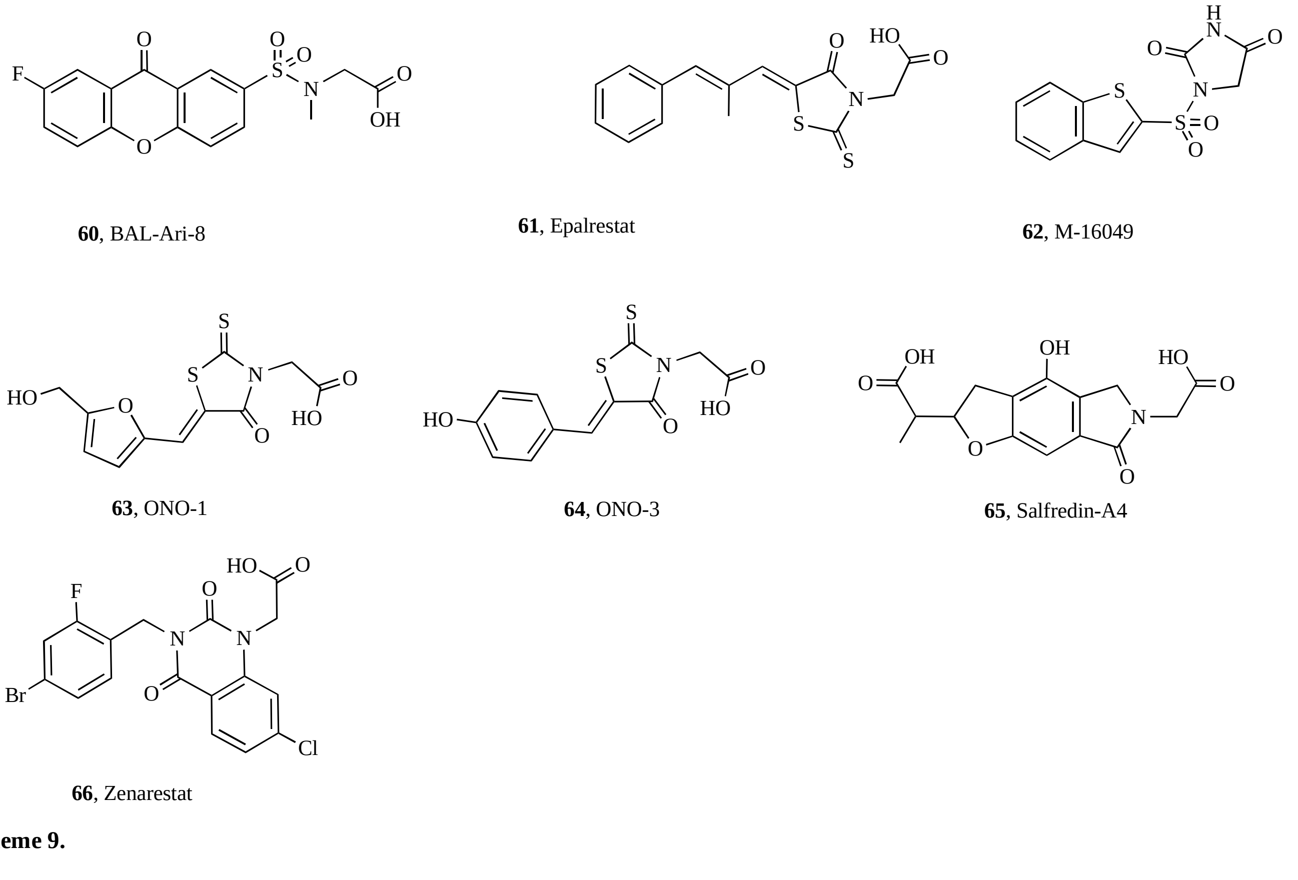
Scheme 8.

Based on this finding, we performed a MCSS in all 329,244 *de novo* generated constructs, retrieving the Enalapril motif 855 times (0.26%). A final MCSS was done within all constructs with a similarity ≥ 0.52 (target-specific similarity threshold for ACE inhibitors; Tanimoto, FCCFP-4) to any of our known actives. Among these 7,594 constructs the Enalapril motif was found 854 times (11.2%), which is almost identical to searching the 43-fold amount of *de novo* constructs. We conclude that selecting *de novo* constructs with a similarity score above a computed target-specific threshold value to any of our known actives not only greatly reduced the number of structures which need to be considered, but also considerably enriched the resulting set of *de novo* constructs with desirable structural motifs or chemotypes.

Angiotensin-II receptor antagonists: A number of 650,616 unique virtual constructs were generated. The best-scored *de novo* construct had a Manhattan distance of D = 6.5 (CATS). A number of 3,980 constructs showed a similarity score ≥ 0.58 (FCFP-4, Tanimoto) to at least one known active compound. Neither the given templates nor any other known actives were reconstructed.

Aldose reductase inhibitors: A number of 165,785 unique *de novo* constructs were generated. Three of the five presented reference structures were found again (Epalrestat (**61**), ONO-1 (**63**), ONO-3 (**64**)), four more known but not presented actives (**60**), (**62**), (**65**-**66**) were also found. Two known actives (Salfredin-A4 (**65**), Zenarestat (**66**)) were each rebuilt twice. A number of 990 *de novo* compounds showed a similarity score above the target class-specific similarity threshold of 0.59 (FCFP-4, Tanimoto).

We conclude that the retrieval of three presented reference structures (Epalrestat (**10**), ONO-1 (**11**) and ONO-3, (**12**)) is a further indicator for successful reconvergence of the evolutionary algorithm on an activity island defined by a desired chemotype. The generation of *de novo* constructs with known activity and considerable structural difference to the presented templates on the other hand could be further-



60, BAL-Ari-8

61, Epalrestat

62, M-16049

63, ONO-1

64, ONO-3

65, Salfredin-A4

66, Zenarestat

Scheme 9.

more considered as an indication for successful scaffold hopping by *de novo* design.

De Novo Design with a Diverse Set of Five Templates

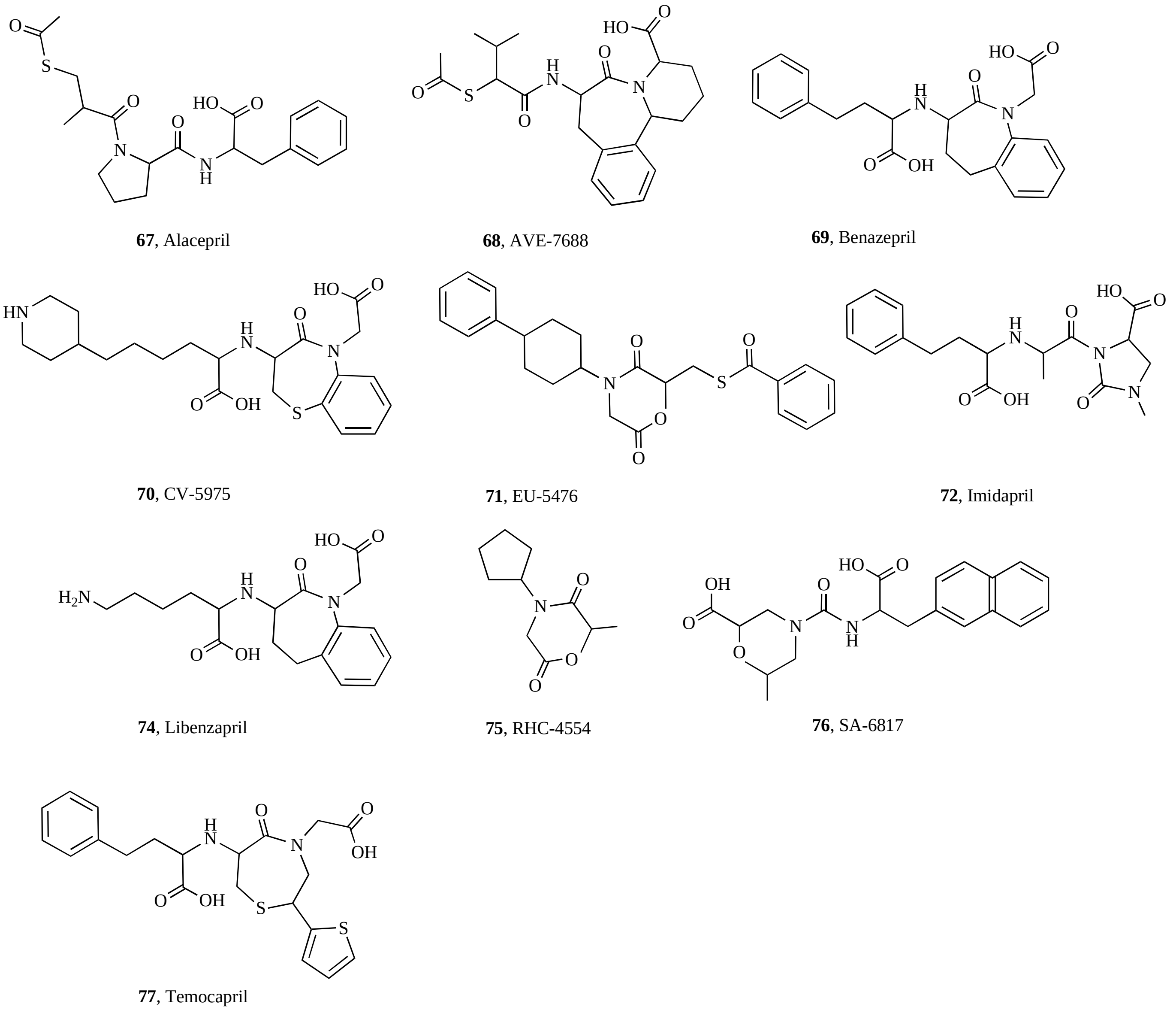
We analyzed whether the presentation of five maximally diverse reference structures spanning the whole range of structural diversity within ligands of a certain target class would increase the probability of scaffold hopping. Ideally, the newly generated *de novo* structures would show a high degree of structural diversity while still maintaining a reasonable probability of activity. For brevity, only results for ACE inhibitors and aldose reductase inhibitors are presented here. The outcome was similar for the other activity classes (not shown).

ACE inhibitors: A number of 333,299 unique virtual constructs were generated from five maximally diverse template structures (**1**, **2**, **7-9**). No given template structure was found again, the best-ranked *de novo* compound showed a fitness score of *D* = 7.8 (Manhattan, CATS2D), which is slightly better than for optimization with a single template (*D* = 8.8), but worse than optimization with a set of similar templates (*D* = 5.2). Still, ten of the known reference compounds (**67-77**) were reconstructed. A number of 4,999 *de novo* constructs showed a similarity score ≥ 0.52 (FCFP-4, Tanimoto) to any known active compound, which is approximately half the amount reached when optimizing with a single template.

Aldose reductase inhibitors: A number of 376,455 unique *de novo* constructs were created in this approach. 1,915 constructs were found with a similarity score ≥ 0.59 (target-specific similarity threshold), and seven known reference compounds (**78-84**) were rebuilt, two of which were presented as template structures (**81** and **84)**.

Evaluation of Scaffold Diversity

Although the retrospective retrieval of known reference compounds, which were not presented as template structures during the evolutionary *de novo* generation, is already a



67, Alacepril

68, AVE-7688

69, Benazepril

70, CV-5975

71, EU-5476

72, Imidapril

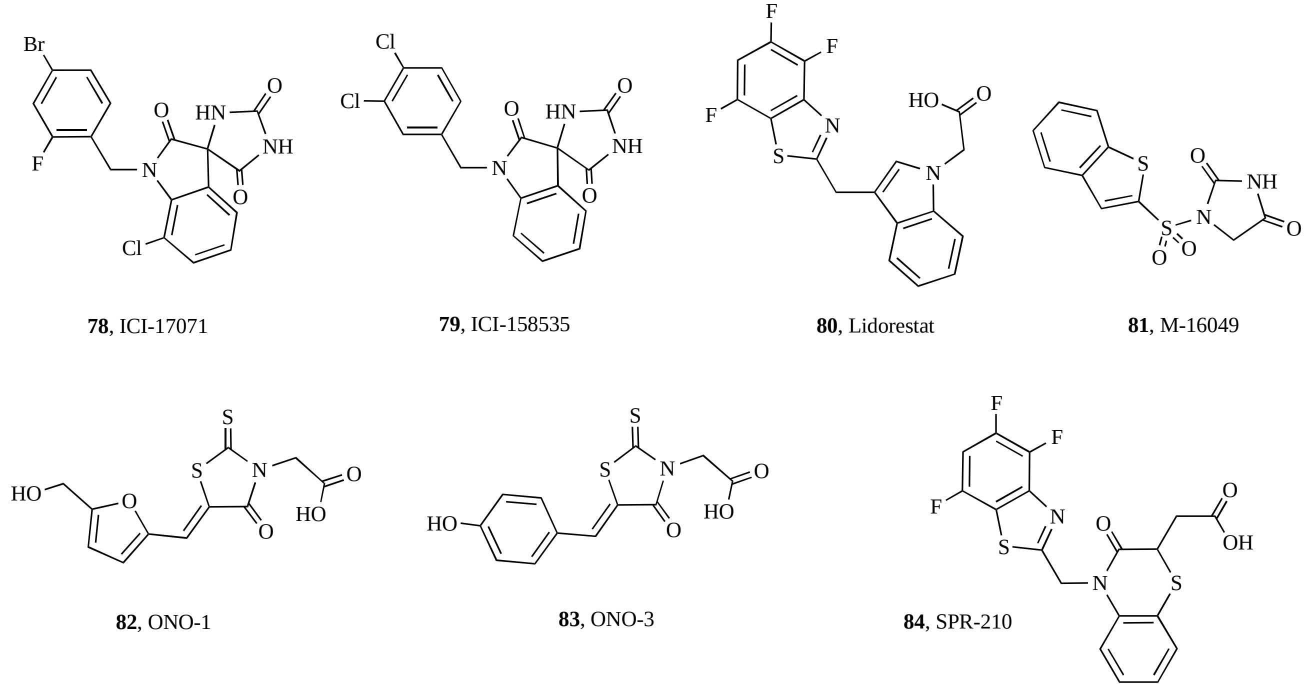
74, Libenzapril

75, RHC-4554

76, SA-6817

77, Temocapril

Scheme 10.



78, ICI-17071 **79**, ICI-158535 **80**, Lidorestat **81**, M-16049

82, ONO-1 **83**, ONO-3 **84**, SPR-210

Scheme 11.

strong indicator of successful scaffold-hopping, we wanted to analyze in more detail whether *de novo* design is also applicable for the generation of lead structures with completely novel scaffolds. The convergence of the *de novo* algorithm on new activity islands, which are not occupied by known actives could broaden the drug designer's perspective, thus optimally leading to new ideas and ultimately to novel drugs. At the same time, (partial) reconvergence of the algorithm into known activity islands proofs that *de novo* design is capable of producing "meaningful" results.

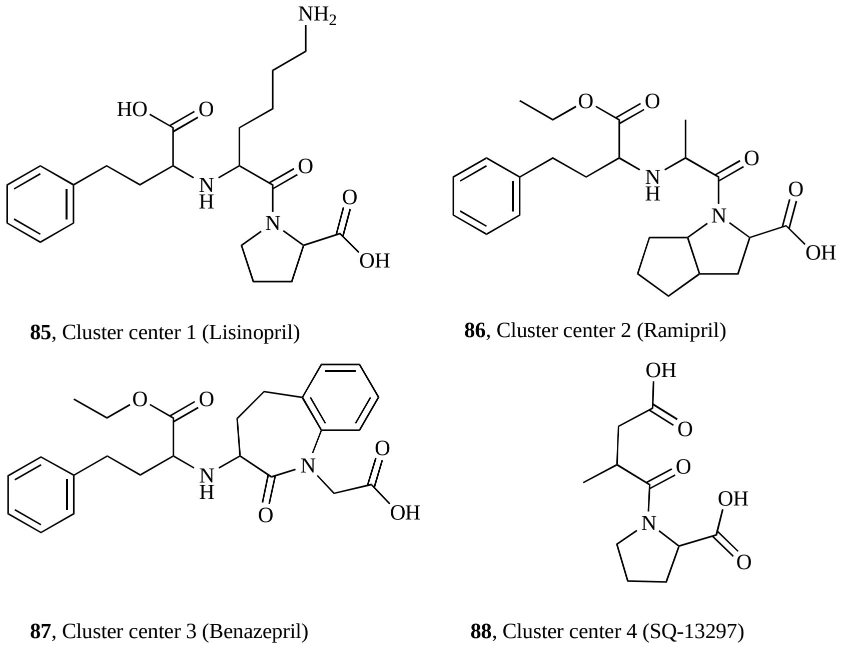
In order to evaluate existing activity islands we first clustered all known actives from the given target class using FCFP-4 fingerprints and the maximum dissimilarity clustering algorithm. The resulting major clusters were considered to be the activity islands of structures from the corresponding target area. The same procedure was applied to a unified set of known actives and *de novo* compounds. Then, we generated atom- and graph-based Murcko scaffolds ("frameworks") of both actives and *de novo* generated compounds. The reduction of information content by focusing only on the area of interest, that is, the molecular scaffold, might allow for a better overview of structural diversity. Finally, we evaluated if the activity islands determined by major clusters of known actives were re-populated during the *de novo* algorithm, and checked for newly emerged clusters. The analysis is restricted to ACE inhibitors for reasons of clarity and brevity.

ACE inhibitors – optimization with a single reference structure: The clustering of all 73 reference ACE inhibitors using the FCFP-4 fingerprint and the maximum dissimilarity algorithm yielded 16 different clusters. The four largest clusters cover 82% of all molecules, while the remaining 12 clusters each contain 5% or less of all actives. We thus considered the four major clusters (cluster centers: **85-88**) to represent the major ACE activity islands.

During the clustering of the unmodified structures of the 73 reference compounds and 9,416 potentially active (Note: we used a random forest model predicting "potential" activity; details not shown) *de novo* designed compounds, 82 clusters emerged. Of those 82 clusters, 30 clusters contained 7,662 structures (80.7%) that were considered further. Three of the seven largest clusters contained known actives; four major clusters (cluster centers: **89** – **82**) did not contain any known active indicating the creation of novel chemotypes.

Following this idea, we then generated 2,581 unique graph-based Murcko scaffolds from the same 73 known actives and 9,416 *de novo* compounds. The Murcko scaffolds were subsequently subjected to clustering by maximum dissimilarity. From the resulting 99 clusters, 18 clusters contained 7,856 scaffolds (83.4%), and the largest cluster had 1,693 members (18%).

While trying to cluster the graph-based scaffolds, we realized that the abstract representation of a molecule by only its ring systems and shortest linkers might indeed discard too much information (data not shown), and instead we decided to count the frequency of occurrence of different graphs. This still provided us with a good picture of how often the *de novo* algorithm converged into a specific basic molecular scaffold. We found that in the 2,581 unique graph-based scaffolds, 2,479 scaffolds (96%) were singletons, meaning they appeared just once. This shows that the *de novo* algorithm indeed sampled chemical space broadly, resulting in many different scaffolds. On the other hand, 55 scaffolds (2%) appeared frequently (> 100 times). We considered those as an indication of convergence into specific chemotypes. The ten most frequently designed scaffolds show characteristic motifs of the ACE inhibitor class while retaining a considerable diversity among their scaffolds (Fig. **3**).



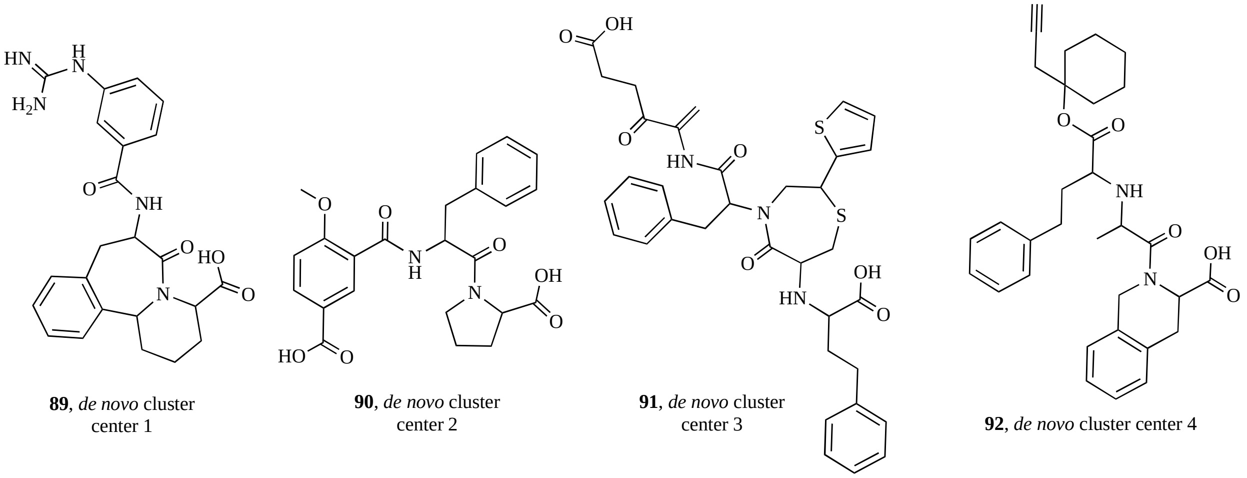
85, Cluster center 1 (Lisinopril)

86, Cluster center 2 (Ramipril)

87, Cluster center 3 (Benazepril)

88, Cluster center 4 (SQ-13297)

Scheme 12.



89, *de novo* cluster center 1

90, *de novo* cluster center 2

91, *de novo* cluster center 3

92, *de novo* cluster center 4

Scheme 13.

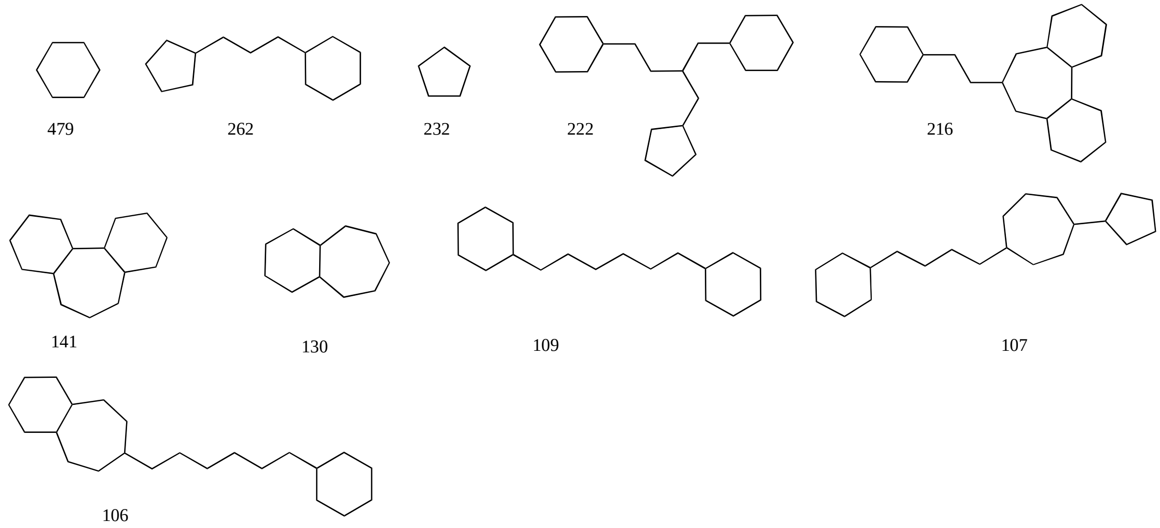
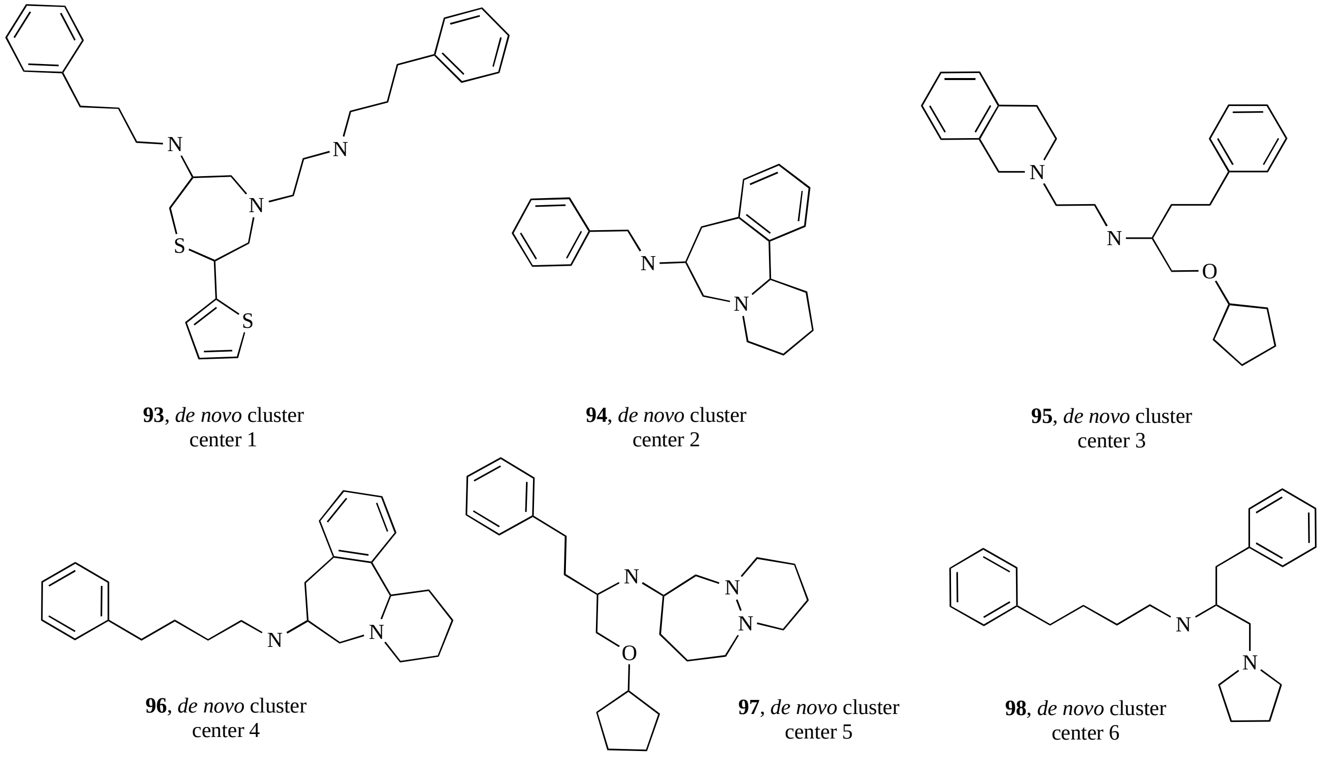


Fig. (3). Most frequently designed scaffolds of potential ACE inhibitors (absolute numbers).



93, *de novo* cluster center 1

94, *de novo* cluster center 2

95, *de novo* cluster center 3

96, *de novo* cluster center 4

97, *de novo* cluster center 5

98, *de novo* cluster center 6

Scheme 14.

In a final approach, we generated atom-based Murcko scaffolds from the same 73 known actives and 9,416 *de novo* compounds. A total of 4,795 different scaffolds were generated which were then clustered. A number of 369 different clusters emerged, with the largest 10 clusters covering 21.8% of all scaffolds. In four of the six largest clusters, (cluster centers: **93**-**98**) known actives were found. The remaining two clusters contained only *de novo* generated structures. Even though the molecular similarity of the atom-based scaffolds showed their common activity on the given target, there was enough structural diversity to indicate successful lead hopping.

CONCLUSIONS

We were able to present several examples of successfully performed scaffold-hopping during the *de novo* generation of drug-like compounds. Under different conditions, 27% of all compounds from a set of known ACE inhibitors and 17% of known aldose reductase inhibitors were reconstructed by an evolutionary *de novo* algorithm. It became clear that for some target classes like the angiotensin-II receptor antagonists, fragment-based ligand *de novo* design is still a challenge. One explanation of the evolutionary algorithm's poor reconvergence into these desired chemotypes might be the retrosynthetic inaccessibility of the predominantly ring-system-based AT2 ligands: the RECAP rules in our *de novo* design tool Flux do not allow for ring systems to be cleaved or new ring systems constructed. Reconvergence into desired activity islands thus is hindered. It remains a future task to address the problem of cyclic chemistry in *de novo* design.

ACKNOWLEDGEMENTS

Thorsten Naumann and Andreas Evers are thanked for valuable discussion and support. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, and the Deutsche Forschungsgemeinschaft (SFB 579, project A11).

REFERENCES

[1] Schneider, G.; Fechner, U. *Nat. Rev. Drug Discov.*, **2005**, *4*, 649-663.

[2] Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. *Nat. Rev. Drug Discov.*, **2003**, *2*, 369-378.

[3] Lipinski, C.; Hopkins, A. *Nature*, **2004**, *432*, 855-861.

[4] Wise, A.; Gearing, K.; Rees, S. *Drug Discov. Today*, **2002**, *7*, 235-246.

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

[6] Lewell, X.O.; Budd, D.B.; Watson, S.P.; Hann, M.M. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 511-522.

[7] Schneider, G.; Fechner, U. *Nat. Rev. Drug Discov.*, **2005**, *4*, 649-663.

[8] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew. Chem. Int. Ed.*, **1999**, *38*, 2894-2896.

[9] Fechner U.; Schneider G. *J. Chem. Inf. Model*., **2006**, *46*, 699-707.

[10] Walters, W.P.; Stahl, M.T.; Murcko, M.A. *Drug Discov. Today*, **1998**, *3*, 160-178.

[11] Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 897-902.

[12] Lipinski, C.A.; Lombardo, F.; Dominy B.W.; Feeney, P.J. *Adv. Drug Deliv. Rev.*, **1997**, *23*, 3-25.

[13] Willett, P. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.

[14] Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. *J. Comput.-Aided Mol. Design*, **2003**, *17*, 687-698.

[15] Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K; Jacoby, E. *Org. Biomol. Chem*., **2004**, *2*, 3256-3266.

[16] Morgan, H. L. *J. Chem. Doc.*, **1965**, *5*, 107-113.

[17] Jaccard, P. *Bull. Soc. Vaud. Sci. Nat*., **1901**, *37*, 241-272.

[18] Tanimoto, T.T. *IBM Internal Reports*, **1957**, *17th November*.

[19] Kennard, R.W.; Stone, L.A. *Technometrics*, **1969**, *11*, 137–148.

[20] Hansen, P.; Delattre, M. *J. Am. Stat. Assoc.*, **1978**, *73*, 397-403.

[21] Willett, P. *J. Comput. Biol.*, **1999**, *6*, 447-457.

[22] Snarey, M.; Terrett, N.K.; Willett, P.; Wilton, D.J. *J. Mol. Graphics Model.*, **1997**, *15*, 372-385.

[23] Johnson, M.A.; Maggiora, G.M. *Eds. Concepts and Applications of Molecular Similarity;* New York: Wiley, **1990**.

[24] Horvath, D.; Jeandenans, C. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 680-690.

[25] Matthews, B.W. *J. Mol. Biol.*, **1968**, *33*, 491-497.

[26] Matthews, B.W. *Annu. Rev. Phys. Chem.*, **1976**, *27*, 493–523.

[27] Centor, R.M. *Medical Decision Making*, **1991**, *11*, 102-106.

[28] Veropoulos, K.; Campbell, C.; Cristianini, N. Controlling the Sensitivity of Support Vector Machines. In *Proceedings of the International Joint Conference on Artificial Intelligence,* **1999**.

[29] Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C.A.F.; Nielsen, H. B. *Bioinformatics*, **2000**, *16*, 412-424

[30] Givehchi, A.; Schneider, G. *Mol. Div.*, **2004**, *9*, 371-383.

[31] Glick, M.; Jenkins, J.; Nettles, J.; Hitchings, H.; Davies, J. *J. Chem. Inf. Model.*, **2006**, *46*, 193-200.

[32] Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. *J. Med. Chem.*, **2005**, *48*, 2534-2547.

[33] Bemis, G.W.; Murcko, M.A. *J. Med. Chem*., **1996**, *39*, 2887-2893.

[34] Grabowksi, K.; Schneider, G. *Curr. Chem. Biol.*, **2007**, *1*, 115-127.

[35] Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 1273-1280.

[36] McGregor, J.; Willett, P. *J. Chem. Inf. Comput. Sci.*, **1981**, *21*, 137-140.

[37] Raymond, J.W.; Willett, P. *J. Comput.-Aided Mol. Des.*, **2002**, *16*, 521-533.

Received: June 30, 2007 **Revised: September 4, 2007** **Accepted: September 6, 2007**