

**Development of a computational method for
reaction-driven *de novo* design of druglike compounds**

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich 15

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Markus Hartenfeller

aus Friedberg

Frankfurt 2010

(D 30)

vom Fachbereich 15 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Anna Starzinski-Powitz

Gutachter: Prof. Dr. Gisbert Schneider
Prof. Dr. Holger Stark

Datum der Disputation:

Index of Contents

Index of Contents	3
Abbreviations	5
1 Introduction	6
1.1 Computer-assisted <i>De Novo</i> Design of Drug Candidate Structures	7
1.1.1 Scoring Strategies	8
1.1.2 Assembly Strategies	11
1.1.3 Search Strategies	16
1.1.4 Multi-objective Optimization and Feasibility by Chemical Synthesis	20
1.2 Examples of Software Tools for Molecular <i>De Novo</i> Design	22
1.2.1 LUDI	22
1.2.2 Skelgen.....	23
1.2.3 TOPAS/FLUX	24
1.2.4 BOMB.....	25
1.3 Outline	26
2 Material and Methods	28
2.1 Library of Chemical Reactions	28
2.2 Library of Synthesis Building Blocks	30
2.3 Design Algorithm	32
2.4 Scoring Function	36
2.4.1 Graph Kernel Method	36
2.4.2 Modification of the Graph Kernel Method	38
2.4.3 Pharmacophore Typing.....	39
2.4.4 Graph Reduction	41
2.5 Assessment of Scaffold Similarity	44
2.6 Implementation	45
3 Results and Discussion	46
3.1 Influence of Parameters on General Characteristics and Scaffold Diversity	46
3.1.1 General characteristics	47
3.1.2 Scaffold Diversity	49
3.2 Property Analysis of Designed Compounds	53
3.3 Exemplary DOGS Designs	57
3.3.1 Trypsin	57
3.3.2 Transforming Growth Factor β 1 Receptor.....	61
3.3.3 Estrogen Receptor	62
3.4 Practical Evaluation of the Software	64
3.4.1 Human γ -secretase	64
3.4.2 Human Histamine H ₄ -Receptor	66
4 Conclusions and Outlook	71
Summary	76
Zusammenfassung	78
References	84
Supplement	95
Coupling Reactions	95
Preprocessing Reactions	110
Analytical Spectra	112

Acknowledgement	118
Curriculum Vitae	119

Abbreviations

2D:	two-dimensional
3D:	three-dimensional
ADME:	Absorption, distribution, metabolism, excretion
AP:	Attachment point
DOGS:	Design of genuine structures
EA:	Evolutionary algorithm
ER:	Estrogen receptor
FEP:	Free energy perturbation
FGA:	Functional group addition
FGI:	Functional group interconversion
GPCR:	G-protein coupled receptor
<i>h</i> H ₄ R	Human histamine H ₄ -receptor
HIV-RT:	Human immunodeficiency virus reverse transcriptase
HTS:	High throughput screening
mg:	Molecular graph
NCE:	New/novel chemical entity
PSO:	Particle swarm optimization
QSAR:	Quantitative structure activity relationship
SSSR:	Smallest set of smallest rings
TGF:	Transforming growth factor
rg:	Reduced graph
VS:	Virtual screening

1 Introduction

One of the early and pivotal steps in drug development is the identification of structurally novel chemical entities (NCE) exhibiting a desired effect on a biological target molecule. Identification of NCEs may be approached by two complementary strategies: One can either search for NCEs in libraries of *already existing* small organic molecules (high throughput screening, HTS) or synthesize *new* molecules ‘from scratch’ that are tailored for a particular project (*de novo* design). Both strategies have their advantages and caveats. While the costs per tested molecule are typically of magnitudes lower for HTS than for *de novo* designed compounds,¹ HTS is limited to known regions of the chemical space. This can be a problem in case a HTS library does not contain appropriate molecules for the project at hand. In contrast, *de novo* design holds the appealing advantage to be theoretically unlimited and intrinsically innovative. On the other hand, custom synthesis of small organic molecules is comparably slow and more expensive. The two strategies can therefore be seen as complementary and can be employed in parallel in drug development campaigns.

Since the 1950s computer-assisted methods have found entrance to the drug development process.² For both strategies (HTS and *de novo* synthesis) *in silico* counterparts have been introduced to complement and support the traditional drug development methods. Like HTS, software for *virtual screening* (VS) evaluates large collections of available compounds with respect to their potential biological activity. A plethora of different approaches has been proposed for this purpose.^{2,3} Programs for computer-assisted *de novo* design suggest novel compounds supposed to possess desired pharmacological properties (Figure 1).

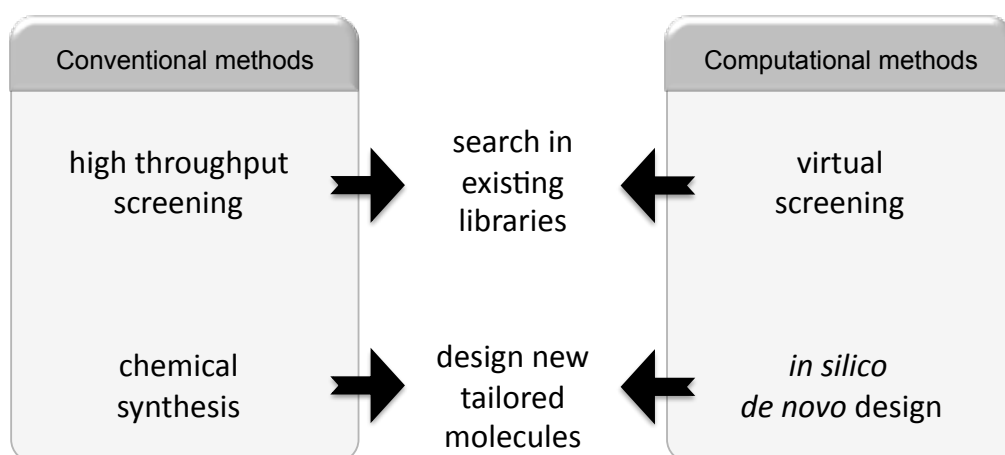


Figure 1. Computational counterparts for conventional drug development methods have been introduced throughout the last decades.

The goal of this work is the development of a new program for computer-assisted *de novo* design of drug candidate compounds.

1.1 Computer-assisted *De Novo* Design of Drug Candidate Structures

The first programs for computer-assisted *de novo* design (termed ‘*de novo* design’ in the following) were published about 20 years ago in the late 1980s.⁴ Table 1 presents an overview of existing computer programs for molecular *de novo* design. Software tools can be categorized by the strategies applied to address three pivotal elements of molecule design:

1. How is the quality of proposed molecules assessed (*scoring strategy*)?
2. How are molecules constructed (*assembly strategy*)?
3. How does the optimization progress based on the current state of knowledge (*search strategy*)?

Regardless of the way different approaches try to solve these challenges, almost all of them follow the fundamental concept to mimic the iterative process of drug discovery research in a real laboratory: molecules are generated, subsequently tested, and the results form the basis of the next round of synthesis. Search and assembly strategies correspond to the intellectual and technical work of a chemist, whereas scoring complies with testing the compounds for activity in a biological assay.

Table 1. Chronological overview of *de novo* design software and the applied type of scoring strategy. If available, a software name is given. Otherwise, the name of the first author is used (table continues on the next page).

Method/Name	Year of publication	Scoring	
		Ligand-based	Receptor-based
HSITE/2D Skeletons ⁵⁻⁷	1989		X
3D Skeletons ⁸	1990		X
Builder v1 ⁹	1992		X
LUDI ¹⁰⁻¹⁴	1992		X
NEWLEAD ¹⁵	1993		X
SPLICE ¹⁶	1993		X
GroupBuild ¹⁷	1993		X
CONCEPTS ¹⁸	1993		X
SPROUT ¹⁹⁻²²	1993		X
MCSS & HOOK ^{23,24}	1994		X
GrowMol ²⁵	1994		X
Chemical Genesis ²⁶	1995	X	X

PRO_LIGAND ²⁷⁻³²	1995	X	X
SMoG ³³⁻³⁵	1996		X
CONCERTS ³⁶	1996		X
PRO_SELECT ^{37,38}	1997		X
Skelgen ³⁹⁻⁴⁴	1997	X	X
Nachbar ^{45,46}	1998	X	
Globus ⁴⁷	1999	X	
DycoBlock ⁴⁸⁻⁴⁹	1999		X
LEA ⁵⁰	2000	X	
LigBuilder ⁵¹	2000		X
TOPAS ^{52,53}	2000	X	
F-DycoBlock ⁵⁴	2001		X
ADAPT ⁵⁵	2001		X
Pellegrini & Field ⁵⁶	2003	X	X
SYNOPSIS ⁵⁷	2003		X
CoG ⁵⁸	2004	X	
BREED ⁵⁹	2004	X	
Nikitin ⁶⁰	2005		X
LEA3D ⁶¹	2005		X
Flux ^{62,63}	2006	X	
FlexNovo ⁶⁴	2006		X
BOMB ⁶⁵	2006		X
Feher ⁶⁶	2008	X	
GANDI ⁶⁷	2008	X	X
COLIBREE ⁶⁸	2008	X	
SQUIRREL ^{nov} ^{69,70}	2009	X	
Hecht&Fogel ⁷¹	2009	X	X
FOG ⁷²	2009	X	
MED-Hybridise ⁷³	2009		X
MEGA ⁷⁴	2009	X	X
Fragment Shuffling ⁷⁵	2009	X	X
AutoGrow ⁷⁶	2009		X
BI CLAIM ⁷⁷	2009	X	
NovoFLAP ⁷⁸	2010	X	
PhDD ⁷⁹	2010	X	
GARLig ⁸⁰	2010		X
DOGS ⁸¹	2011	X	

1.1.1 Scoring Strategies

Early *de novo* design programs were exclusively based on *receptor-based* scoring schemes, *i.e.* the quality of proposed molecules is assessed by evaluating their potential to interact with a binding site on the receptor surface. This approach is limited to target proteins for which data about their three-dimensional (3D) structure is available, which is not the case for all targets of pharmaceutical relevance. For example, G-protein coupled receptors (GPCR) represent a target class of high interest for the pharmaceutical industry⁸² for which only little experimental data about 3D structures of its members could be collected so far.⁸³ Receptor-based tools were therefore soon augmented by the development of *ligand-based* scoring schemes to circumvent this shortcoming (Table 1). While receptor-based scoring relies on the

concept of complementarity to the binding pocket, ligand-based scoring schemes assess similarity (or distance) to known reference ligands exhibiting the desired biological activity. Following the ‘similarity principle’ stated by Johnson and Maggiora⁸⁴ compounds designed under the objective to show high structural similarity to the reference should have an increased probability to exhibit similar pharmacological properties.

Receptor-based scoring

Receptor-based approaches are closely related to computational strategies for molecular docking. While docking tries to place complete ligands into a binding pocket, *de novo* strategies construct the compound directly within the cavity (*in situ* construction). Both techniques share the objective to maximize the complementarity of the ligand to the binding site regarding shape and properties. Common approaches to estimate the quality of binding during the design process are therefore the same as for molecular docking, where three main strategies have emerged: (i) molecular force fields, (ii) empirical scoring functions, and (iii) knowledge-based scoring functions.^{85,86}

Force fields treat molecules as ensembles of balls (atoms) connected by springs (bonds). Each spring has optimal values for length, torsion angles and angles to other springs. Deviation from these optimal values result in strain. Accordingly, low strain energies correspond to favorable ligand conformations. Interaction with the receptor molecule is estimated by two terms for non-bonded interactions (Coulomb and van-der-Waals potentials, sometimes augmented by an explicit term for contributions of hydrogen bonds). A generalized force field term for non-covalent interactions is given in equation (1). It computes their contribution E to the binding energy between a ligand and a receptor for a given binding mode as

$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left[\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^6} + \frac{q_i q_j}{D r_{ij}} \right], \quad (1)$$

where A_{ij} and B_{ij} are parameters expressing repulsion and attraction of van-der-Waals interactions of atoms i and j at a distance r_{ij} , q_i is the point charge of atom i and D is the dielectric constant of the solvent.⁸⁶ For example, the docking software GOLD uses a molecular mechanics scoring function in its original implementation.^{87,88}

Empirical scoring functions are weighted sums of several separate components, where weights are determined by regression analysis. Weights are optimized in order to reproduce experimentally measured activity values of known ligand-receptor complexes. Individual

components represent different ligand-receptor interactions, which can be determined from a given binding pose. An example of a docking software implementing an empirical scoring function is FlexX.⁸⁹ The free energy of binding is calculated as presented in equation (2) (generalized from an example given in reference 86).

$$\Delta G = \Delta G_0 \sum_{i=1}^{\#it} [\Delta G_i * count_i * pen_i], \quad (2)$$

where ΔG_i represents the contribution (adjusted weight) of interaction type i , $count_i$ is the number of times this interaction type (it) is observed in the given receptor-ligand complex and pen_i is a penalty function accounting for deviations from the ideal interaction geometries for some interaction types like *e.g.* hydrogen bonds, salt bridges or aromatic interactions. The penalty must be evaluated for each observed interaction of such a type and is summarized in pen_i for all instances of an interaction type. ΔG_0 is a fixed ground term that is also adjusted during the fitting process.

Knowledge-based scoring functions rely on discrepancies between observable and expected distributions of atom pair occurrences. Based on the frequencies of atoms one can calculate a background probability of the chance that two atoms (one from the receptor and one from the ligand) are placed in a certain distance in a random ligand-receptor complex, given that they do not interact. This is compared to the counts of atom pairs observed in experimentally explored ligand-receptor complexes (training set) and finally transformed into interaction scores by an inverse formulation of the Boltzmann law.⁸⁵ Atom pairs that occur in higher frequencies than expected by chance result in negative interaction energies (attraction) while less frequently observed pairs score positive (repulsion). Ligand affinity in a given complex with a receptor is estimated by summing up individual scores of observed atom pairs derived from the training set. DrugScore^{90,91} is an example for a knowledge-based scoring function for molecular docking. Equation (3) calculates the contribution of atom pairs between atom types i and j at distance r to the interaction energy of the ligand-receptor complex.⁸⁶

$$E(i, j) = -k_B T \ln g_{ij}(r), \quad (3)$$

where k_B is the Boltzmann constant, T is the absolute temperature and function g_{ij} is a quotient of observed and background frequencies of atom pairs of type i and j at distance r . The total energy of binding is calculated as a sum of these terms for all pairs of atom types and a range of different distances.

Ligand-based scoring

In contrast to computing the complementarity of ligands with the binding site, ligand-based scoring schemes compare ligand candidates to a reference compound exhibiting desired properties and compute a similarity index (or the distance) between them in a descriptor space. For this purpose, the compounds have to be encoded by a mathematical representation allowing for efficient comparison. The concept of similarity also forms the basis of ligand-based virtual screening methods. As a consequence, almost every type of technique developed for VS also finds application in *de novo* design. For ligand comparison, a model representing the molecules and a metric measuring distances in the space of the model need to be selected. While receptor-based scoring inevitably requires accounting for 3D conformations of designed compounds, ligand-based approaches can also work on models based on topological 2D structures (an example for a *de novo* design software working on 2D representations is TOPAS^{52,53}). This can be of particular interest if no sound hypotheses about the binding modes of reference ligands exist or computational power and run time need to be saved.

Several ligand based *de novo* design programs use *pharmacophore* models for quality assessment. These methods compare molecules by the topological (2D^{62,63}) or spatial (3D^{69,70,78}) arrangement of potential interaction centers. Even straightforward substructure fingerprints accounting for the presence and absence of certain structural motifs have found application in *de novo* scoring strategies.^{52,53} Some tools also employ *pseudoreceptor* techniques⁹² and related methods like *molecular field analysis* (MFA²⁸) for scoring. These approaches calculate pharmacophoric and steric constraints of a hypothetical receptor pocket based on a 3D conformation of an active ligand and assess the score of a new compound by evaluating its complementarity to this cavity model, forming a bridge between receptor- and ligand-based methods.⁹³ Ligand-based scoring strategies can either be based on a single reference or an ensemble of known ligands. For example, a consensus pharmacophore model can be built from a multiple alignment of reference ligands. Some scoring techniques even require a whole set of known actives: *QSAR* (quantitative structure activity relationship) methods correlate biological activities of training set compounds with calculated descriptors to yield a predictive model for activity.⁵⁶

1.1.2 Assembly Strategies

Compound assembly strategies can be subdivided into *atom-based* and *fragment-based* approaches. Atom-based techniques build up new molecules atom by atom, whereas

fragment-based design relies on molecular fragments as building blocks. A fragment can be anything from a single atom to a polycyclic ring system. Most of the early *de novo* design tools were strictly atom-based. Modern approaches often provide a diverse selection of large and small virtual molecular entities for compound construction including a few single-atom fragments. Atom-based approaches have the advantages that fine-grained molecule sculpting can be performed and – though only theoretically – the complete universe of chemical structures can be constructed. These advantages come at a price: the huge number of potential solutions complicates a systematic search for actually useful, chemically stable and druglike compounds. The strictly atom-based approach is prone to produce a large fraction of chemically unstable and unreasonable compounds. Fragment-based approaches offer a shortcut to generating new ligands in a more meaningful way and significantly reduce the size of the search space. If fragments commonly occurring in drugs are used for molecule assembly the designed compounds have a high chance of being druglike themselves.^{52,53} In addition, the fragment-based approach improves the chance to produce chemically stable and synthetically feasible compounds. The reason is that fragment-based construction uses larger building blocks, which reduces the number of connection steps needed to assemble a new compound. It should be pointed out that all bonds formed by the software are artificial and therefore the chemical stability and accessibility of the virtual product cannot be guaranteed. The main advantage of fragment-based assembly over atom-based approaches is that many bonds of the designed structures are already predefined in a meaningful way by the fragments. It can be argued that this might be the major reason why the last purely atom-based *de novo* design program RASSE⁹⁴ was published over a decade ago. Nevertheless, using molecular fragments instead of atoms as building blocks alone does not guarantee to construct virtual compounds actually amenable to synthesis (this major objective of *de novo* design will be covered in more detail later).

Several techniques have been developed for automated assembly of molecules. *Alignment-based* methods like BREED⁵⁹ and the fragment shuffling approach⁷⁵ first align different ligands bound to the same protein (or a homologue protein with high sequence similarity) by a backbone overlay of 3D protein structures. Strategic bonds from different ligands brought to close proximity are detected, broken and the four resulting fragments are swapped to yield two new compounds representing hybrids of original ligands (Figure 2).

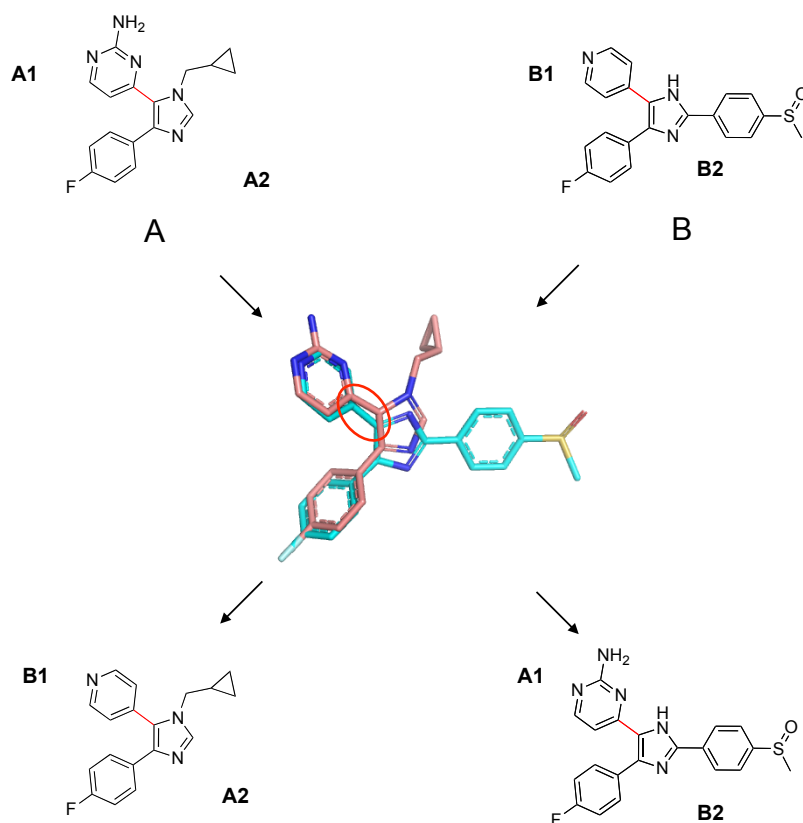


Figure 2. Original ligands A and B (*top*) are aligned in a first step (*center*). BREED⁵⁹ searches for strategic bonds (highlighted in red) and swaps fragments (A1, A2, B1 and B2) at this position in order to get hybrid structures (*bottom*) of original ligands.

Other approaches for structure assembly rely on *molecular force fields* and *docking* techniques. The basic idea is to independently place molecular fragments inside a binding cavity and connect them in successive steps. The software CONCERTS³⁶ is an early example of using molecular dynamics simulations for fragment placing. Fragments are moved according to a molecular force field to obtain low-energy orientations with respect to interactions with the binding site but without witnessing each other. Bonds can be formed between fragments that are brought to close proximity, but can also be broken in later steps. The constant rearrangement of bonds between fragments is supposed to result in compounds exhibiting interaction energies that are favorable to those of the unconnected fragments. Other *de novo* design programs employ docking tools in order to initially place fragments into a binding site. In general, two different strategies exist for this approach: *growing* and *linking*. Growing approaches^{8,17,19-22,25,33-35,64,94} start with one fragment that already satisfies key interactions with the receptor and add more fragments step by step in order to improve the affinity of the constructed compound, guided by the scoring function of the underlying

docking program (Figure 3A). The linking strategy^{10-14,15,67,79} first places several fragments at distinct parts of the pocket, which are then connected by linker fragments (Figure 3B).

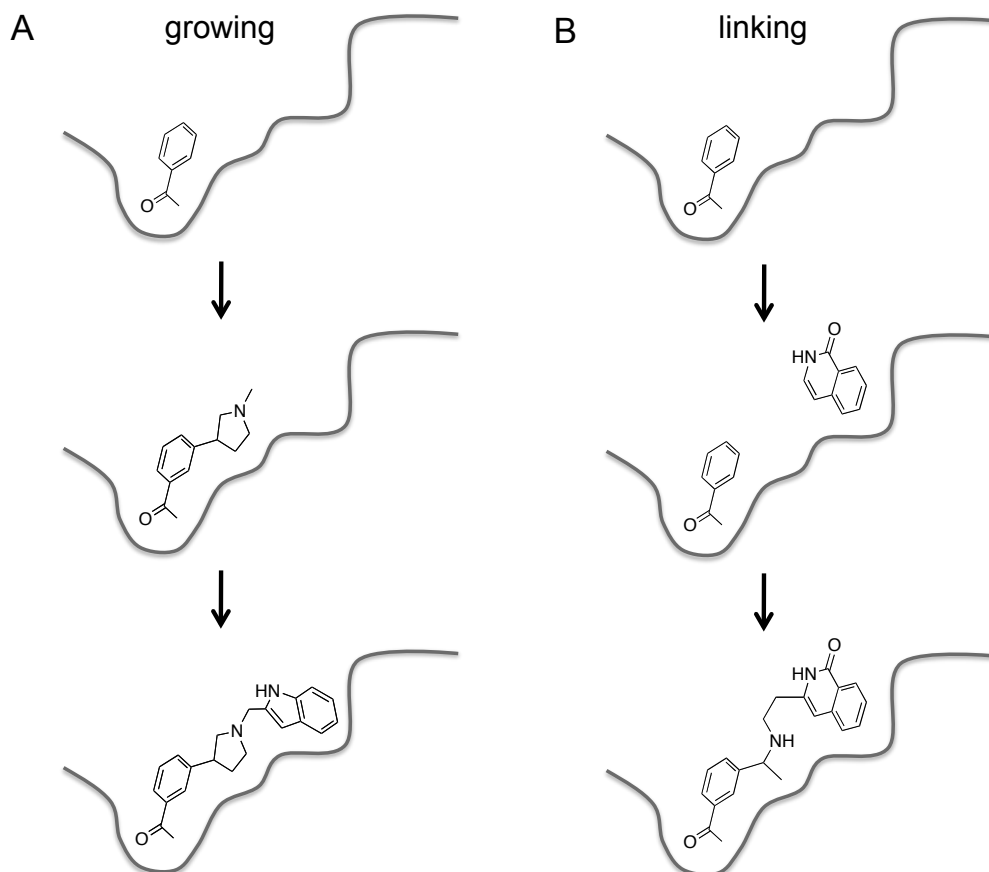


Figure 3. The growing strategy starts with a single building block and sequentially extends it by adding new fragments (A). The linking approach first saturates key ‘interaction hotspots’ of the cavity with building blocks and subsequently links them by special linker fragments (B).

Examples for assembly techniques mentioned so far incorporate knowledge about the receptor structure. In a recent publication Kutchukian *et al.* describe an algorithm for ligand design that is independent of receptor information.⁷² Their ligand-based *de novo* design tool uses connection statistics to assemble new compounds. The algorithm extracts connection frequencies of predefined molecular fragments from a training set of reference compounds. These counts are then converted to probabilities termed *transition probabilities* forming the basis of a growth strategy implemented as a Markov chain of first order. Following the idea of a Markov chain,⁹⁵ the process of growing a molecule can be seen as a walk on a graph, where each fragment is represented by a node. Edges between nodes are labeled with obtained transition probabilities. These labels represent probabilities to pass between nodes connected

by the edge. To grow a molecule, the algorithm starts with a randomly selected or given fragment (a node) and walks across the graph according to transition probabilities. Each time a node is visited the according fragment is added to the molecule (Figure 4). Since each node represents exactly one fragment, transition probabilities only depend on the fragment to be extended in the next step (first order property of the Markov chain). The process stops when all potential growth sites are saturated, a user-defined number of fragments or a given molecular mass is exceeded. The Markov chain is supposed to generate molecules that reproduce connection statistics of the training set, therefore exhibiting increased probability to show desired molecular properties.

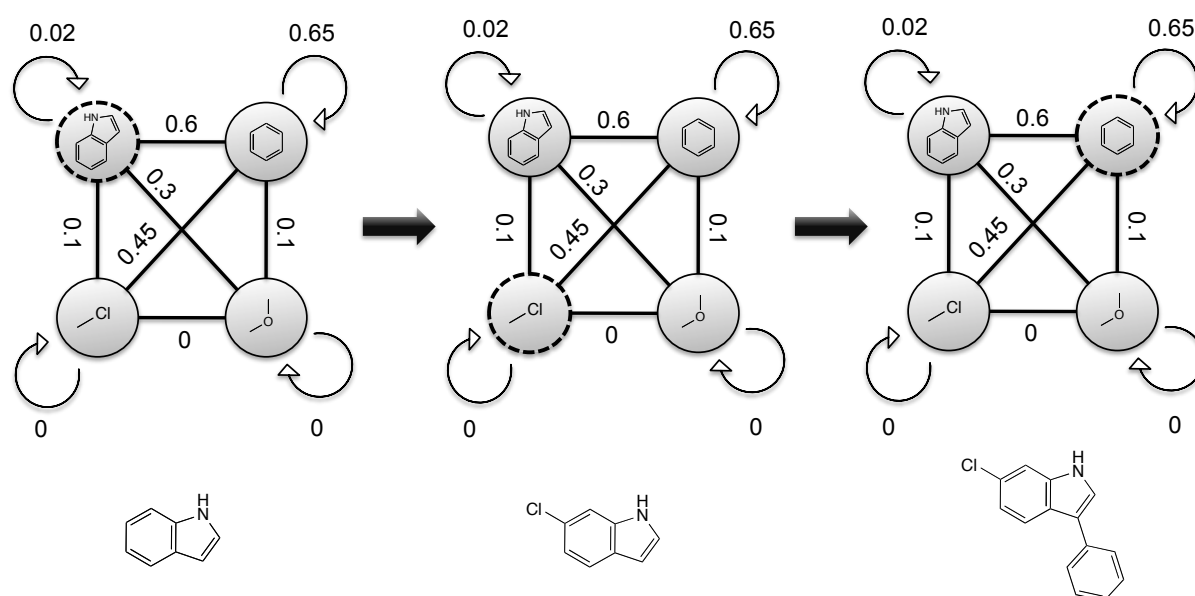


Figure 4. A Markov chain model for compound assembly. A Markov chain⁹⁵ represents a sequential graph traversal: Every time a node is visited (indicated by a dashed line) the corresponding building block is added to the growing molecule (*bottom*). Edge labels (weights) correspond to the probability to walk along an edge in order to get to the next node. Weights are determined by connection statistics of fragments observed in a training set of molecules. Please note that this graph represents a simplification. Typically, weights will not be symmetric.

Retrosynthesis rules form the basis of another category of ligand assembly strategies. Such rules define a set of substructures, each one built around a central bond that is deemed to be cleavable. Collections of compounds can be disassembled at these strategic positions to yield a set of molecular fragments. The same rules find application during the assembly process to construct new molecules by recombining the fragments. The most prominent representative of retrosynthetic rules is the *Retrosynthetic Combinatorial Analysis Procedure (RECAP)*⁹⁶. RECAP derives eleven cleavable bond types from common chemical reactions and defines

them by their structural environment (Figure 5). Examples of programs using the RECAP for mining and recombining molecular fragments to breed new druglike compounds are TOPAS^{52,53} and its direct successor FLUX^{62,63}. Reconnection is restricted to attachment sides originating from the same disassembly rule in order to enhance the probability to form chemically meaningful and stable bonds.

The most sophisticated assembly technique in the sense of incorporating chemical knowledge is the simulation of established reaction protocols for fragment connection. *Reaction-based approaches* use formalized reaction schemes to mirror the bond rearrangements of real synthesis steps in order to connect molecular building blocks. Established data formats for formalization of chemical reactions are the SMIRKS language⁹⁷ and the rxn file format⁹⁸. In case the building blocks are readily available (*e.g.* purchasable from a commercial vendor) this strategy not only enhances the chance to produce chemically reasonable compounds but also delivers direct blueprints for possible synthesis routes. The software SYNOPSIS⁵⁷ is such an example.

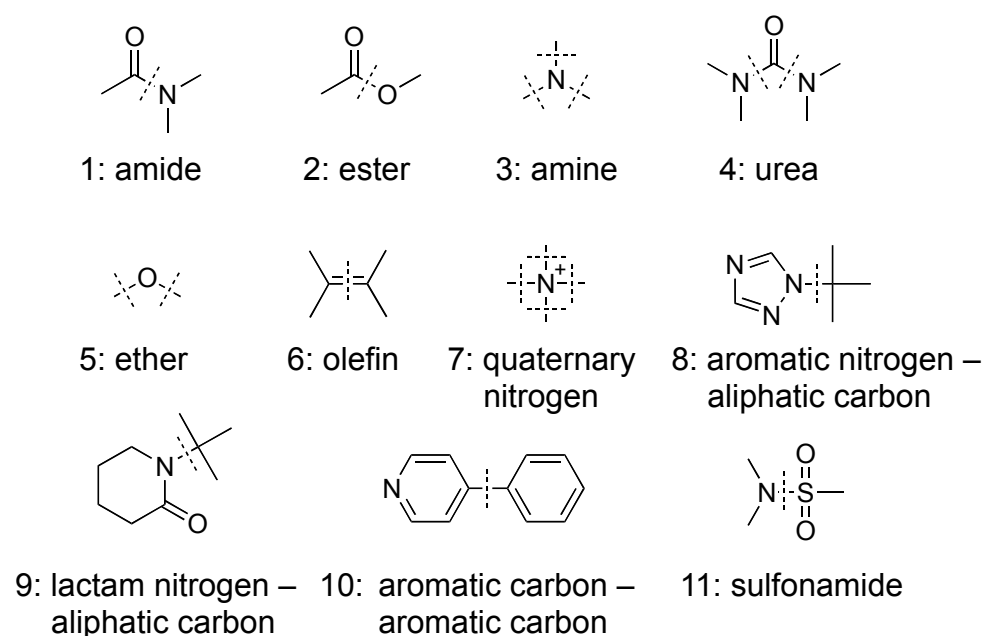


Figure 5. Eleven cleavable bond types defined by RECAP⁹⁶.

1.1.3 Search Strategies

De novo design is faced with an almost infinite search space of small organic molecules. An attempt to assess the actual size of this space resulted in an estimate of 10^{60} compounds.⁹⁹ This figure might slightly overestimate the real size of the space relevant for drug discovery

because it does not account for chemical stability and druglikeness. Nevertheless, it still prohibits approaches that try to enumerate all possible structures. For this reason, *de novo* design programs have to truncate the number of molecules they consider during a design run. Most of the programs apply *stochastic optimization* techniques to cope with such large search spaces.⁴ In order to understand how these algorithms work it helps to imagine the chemical space as a wavy surface. Each point on the surface represents a molecule. Similarity (to whatever quality) between molecules is expressed by distance in this space, so that similar molecules are close to each other. The quality of a compound (score, biological activity) is expressed by the height of the respective search point where better quality is expressed by a higher level. The basic idea of stochastic search algorithms is to explore the neighborhood of the current search point by sampling a few surrounding search points in close proximity (also termed ‘local search’). Information gained by this process is used to extrapolate about the actual structure of this subspace and move along the most promising direction. Successful application of local search strategies requires a ‘smooth’ response characteristic, *i.e.* small structural changes (movement in the space) result in small changes of the score, while large steps cause large differences of scores.¹⁰⁰ Although it has been shown that this is not generally the case in arbitrary chemical spaces and associated activity landscapes,^{101,102} search algorithms relying on local optimization have proven to deliver results of practical relevance and sufficient quality for many complicated optimization problems including molecular *de novo* design.^{103,104} However, appliance of stochastic optimization does not come without a drawback: Many optimization algorithms employ heuristics, *i.e.* they cannot guarantee to find the absolute optimal solution for a given problem. Their stochastic component renders it most likely that two runs of same algorithm applied to the same problem deliver different results. Typically, multiple runs of heuristic approaches have to be performed to yield statistically sound results of retrospective evaluations and enhance the probability to attain useful outcome in prospective studies. The reason for this is that even if the underlying scoring function responds smoothly to movements in the search space, *local optima* can still occur and trap local search strategies. A local optimum is a point in the search space that is assigned with a better score than all points in a certain neighborhood around it, while the search space still might offer better search points in regions beyond this neighborhood. The reliance on the local behavior around the current search point can trap a search algorithm. Although search techniques that support the ability to escape local optima have been developed (*vide infra*), there is still no guarantee to find the global optimum. Different results of optimization runs on

the same problem originate from the fact that the algorithm converges on different local optima due to its stochastic component.

Markov chains (*vide supra*) represent an example of a stochastic search strategy.^{72,95} Their random sampling procedure is often coupled to a *Metropolis criterion* in order to facilitate escaping from local optima.^{18,105,106} According to the Metropolis criterion, structural changes improving the score of a molecule are accepted in any case, whereas steps degrading the score might be rejected: the more a modification degrades the score, the higher the probability to reject it. According to the Metropolis criterion the probability to accept search point j coming from i is

$$P_{i \rightarrow j} = \min(1, e^{-(f(i)-f(j))/T}), \quad (4)$$

where $f(i)$ denotes the quality of search point i (better solutions receive higher values) and T is a constant scaling factor.

Simulated annealing techniques pick up this idea but dynamically change the calculation of rejection probabilities for a degrading step.^{57,107} At the beginning of an optimization run, degrading modifications have a higher chance to be accepted in order to prevent early convergence on a (most likely globally unfavorable) local optimum. In later steps, when the search space has already been explored more intensively and found optima are more likely to be of practical interest, the dynamic calculation of acceptance probabilities will tune the algorithm to preferably stay in the current search region. This is achieved by a more rigorous calculation of rejection probabilities for score-degrading movements in the search space. Computationally, simulated annealing is realized by constantly reducing the scaling factor T of equation (4) during the optimization.

Several stochastic search algorithms have been derived from optimization strategies observable in nature, of which *evolutionary algorithms* (EA) and *particle swarm optimization* (PSO) are prominent examples.¹⁰⁸ EA is an umbrella term for several optimization techniques inspired by the idea of biological evolution. A population of search agents (representing molecules in the context of *de novo* design) is iteratively exposed to random variation and selection. Variation is introduced by genetic operators like mutation and genetic crossover. Selection is performed according to the score of individuals assigned by a scoring function (also termed *fitness function* in this context). Better search agents are more likely to survive and continue to influence the search process, while less fit individuals die out. Evolutionary algorithms mainly differ in the way they encode individuals and how selection and variation

are implemented. Examples of *de novo* design software applying evolutionary algorithms are FLUX^{62,63} and LigBuilder⁵¹. PSO algorithms mimic the behavior of real swarms of animals searching, *e.g.*, for food resources.¹⁰⁹ A set of virtual search agents (termed *particles*) moves in the search space. A position in the search space equals a solution to the optimization problem. The direction of movement is influenced by communication and information exchange between particles about their individual search success. Communication is implemented as a *social memory*, which is accessible by every particle. The social memory stores the best search point found so far by the swarm as a collective. In addition, each particle also stores the best search point it has explored so far in its *personal memory*. Search points stored in the social and personal memory attract the particles during their search. As the search proceeds, promising regions will be explored thoroughly by many particles, while areas found to be less attractive are widely ignored. PSO has been introduced to the field of *de novo* drug design by the program COLIBREE.⁶⁸

Stochastic optimization is not the only way software tools for molecular *de novo* design try to manage the large search space they are confronted with. There are examples of programs applying deterministic search algorithms: FlexNovo⁶⁴ uses a grow strategy to connect molecular fragments. In a preprocessing step, each fragment is docked into the receptor binding site to obtain a single score which serves as a filtering criterion during the design process. Prior to the extension of the growing molecule by adding the next fragment, an estimation of the maximal score achievable by the extended molecule and a set of additional filters are applied to limit the number of potential fragments. Only fragments that have a good chance to improve the score of the molecule are considered. In addition, only the best *k* molecules emerging from an extension cycle are considered in subsequent steps. Thus, FlexNovo employs a ‘greedy’ strategy¹¹⁰ and a set of filtering criteria to reduce the search space to promising sub-regions.

A further strategy to cope with large numbers of potential search points is to employ scoring functions designed to feature *fragment additivity*, *i.e.* the score of a complete molecule is computed as the sum of scores its fragments. This offers the advantage to avoid scoring every possible fragment combination. Instead, each entry of fragment library can be scored alone, and optimal combinations of fragments can be computed without the need to explicitly assemble them. To illustrate the advantage of additive scoring schemes let us consider two fragment libraries, each containing 1,000 entries. A possible product is a combination of two fragments, one fragment from each library. Full enumeration of all possible products would

result in 1,000,000 molecules ($1,000^2$) and the same number of score calculations. An additive scoring function would only have to score each fragment once, which means that only 2000 ($1000 + 1000$) score calculations would be required. This simple example demonstrates that the search space grows exponentially with the number of fragments. One can expect the advantage of additive scoring schemes regarding computational cost to be more serious in practically relevant *de novo* design scenarios. Two examples of software tools that make use of additive scoring schemes are BI CLAIM⁷⁷ and a computer program proposed by Nikitin *et al.*⁶⁰. However, regardless of the advantages, it must be stated that additivity of ligand scores is a feature that is artificially introduced as it represents a simplifying assumption of the scoring scheme. Binding energies of ligand–receptor interactions cannot be expected to be additive in general.¹¹¹

1.1.4 Multi-objective Optimization and Feasibility by Chemical Synthesis

The primary task of *de novo* design is to propose novel compounds with a desired biological effect, *i.e.* affinity to a target macromolecule. Although scoring functions considerably differ in their approach to estimate biological activity, every *de novo* design algorithm takes this objective into account. For this reason it can be referred to as the *primary constraint* of *de novo* design. However, biological activity is not the only requirement for a compound to be a promising candidate for further investigation. Druglikeness, pharmacokinetic properties like absorption, distribution, metabolism and excretion (ADME), toxicity, off-target activity (selectivity), and accessibility by chemical synthesis are examples of *secondary target constraints*.⁴ Such objectives can either be directly addressed by an explicit scoring term or implicitly accounted for by the design strategy. For example, a fragment-based design approach based on fragments derived from known drugs implicitly considers druglikeness. In case additional scoring terms explicitly consider secondary constraints, *de novo* design becomes a multi-objective optimization task. One possibility to make multi-objective optimization compatible to one-dimensional optimization techniques is to calculate a combined score as a weighted sum of single scoring terms. This requires careful weighting of the different design objectives and is prone to lead to unfavorable results, especially in the case of conflicting design objectives.⁴ The reason is that in this case ‘average’ structures fulfilling all objectives on a comparable but overall weak level are likely to emerge. In contrast, Pareto optimization¹¹² delivers a collection of results that contains solutions focusing on different subsets of objectives (the so-called ‘Pareto front’). The Pareto front is formed by non-dominated solutions: a solution is dominated if there is at least one solution in the population featuring a better score in every optimization objective (Figure 6). Non-dominated

solutions therefore represent trade-offs between competing constraints. Pareto optimization does not need any weighting of objectives prior to scoring. It provides the user with a list of candidate solutions for every objective, leaving the decision to the user which of the objectives should be emphasized. Pareto-optimization has been introduced to *de novo* drug design in 2004 by Brown and coworkers.¹¹³ Two years before, the program MoSELECT¹¹² implemented Pareto optimization for the design of combinatorial libraries.

Secondary constraints do not necessarily have to be employed during the design process. Another way is to use them as filtering criteria after the actual design run to narrow down the number of structures of potential interest ('post-generation' scoring).^{114,115}

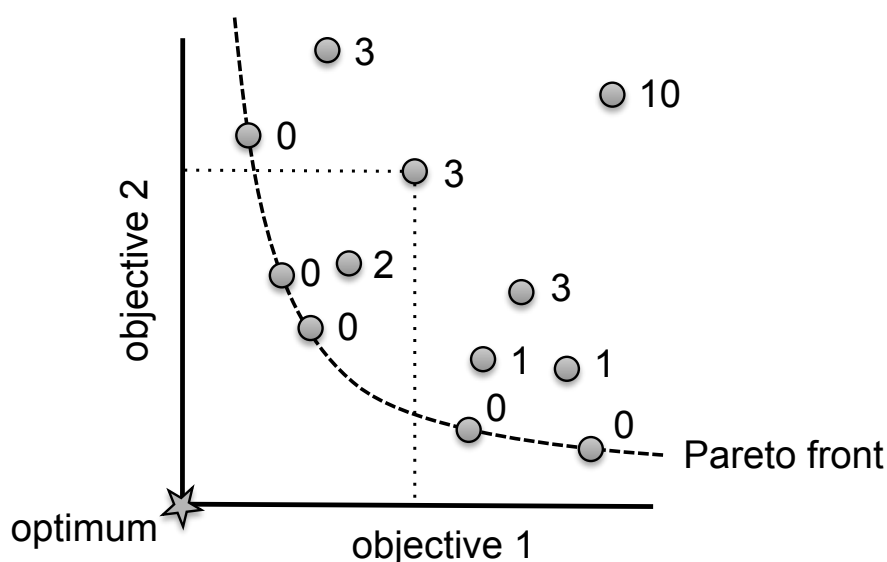


Figure 6. A set of solutions (grey circles) for a two-dimensional optimization problem. The figure next to each circle denotes the number of dominating solutions. For one solution, exemplary determination of the number of dominating solutions is presented (dotted lines). Three solutions are dominating, since they are better in all objectives. Non-dominated solutions form the Pareto front (dashed line; figure adapted from reference 4).

Among the aforementioned secondary design constraints, synthetic feasibility of proposed structures is of crucial importance for molecular *de novo* design.^{81,114} The actual synthesis of designed compounds is key to both practical evaluation of the software as well as drug design projects. The assembly process represents the part where synthetic feasibility can be incorporated directly during the design. Over the years of development in the field, steadily increasing effort has been put on this issue: from atom-based molecule build-up to rule-based assembly of fragments and, finally, virtual synthesis by established reaction protocols and available building blocks. Among all strategies mentioned only the latter approach is able to

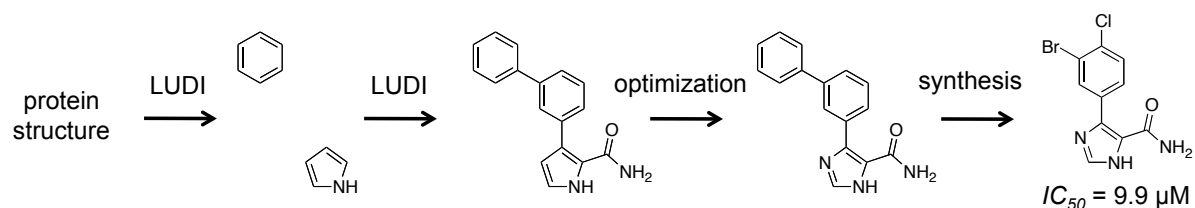
additionally propose synthesis routes for each designed compound, which can be an advantage of practical value.

Instead of implicitly accounting for chemical feasibility by an advanced assembly method, another strategy is to use a suitable scoring function (thereby making synthesizability an explicit design objective). For example, the software SYLVIA¹¹⁶ can be used to score designed molecules by their chemical feasibility after they have been assembled. However, this approach does not suggest synthesis routes. Additional software especially designed for this purpose can be employed for synthesis planning. For example the computer programs CAESA¹¹⁷ or Route Designer¹¹⁸ can be applied to suggest synthesis plans for designed compounds post-hoc.

1.2 Examples of *De Novo* Design Software Tools

1.2.1 LUDI

The program LUDI¹⁰⁻¹⁴ is an example of a software solution from the early days of computer-assisted *de novo* design. Despite being a pioneer in the field, LUDI still represents a sophisticated approach to receptor-based design and can be deemed one of the most successful *de novo* design tools.¹¹⁹ The first step of a LUDI construction run comprises the placement of molecular fragments within the receptor binding cavity. The fragment library can be defined by the user. Fragments are placed so that they satisfy potential *interaction centers* within the protein pocket. The algorithm accounts for directed characteristics of interactions (in particular hydrogen bonds) by a vector representation of interaction centers and complementary interaction sites of the fragments. Fragment positions are optimized by minimizing deviations from optimal orientations of interaction partners. Steric clashes with the protein are penalized. A second, empirical scoring function is employed in a subsequent step in order to rank all fragment poses. The most promising fragments placed within the binding site are then connected using linker fragments (linking approach) to yield complete ligand candidate structures. For example, LUDI has been used to design inhibitors of the human immunodeficiency virus reverse transcriptase (HIV-RT).¹²⁰ The scaffold identified by the software was slightly modified to simplify the synthesis. A series of structural variations and sidechain replacements resulted in a set of new structures inhibiting different enzymatic activities of HIV-RT (Scheme 1).



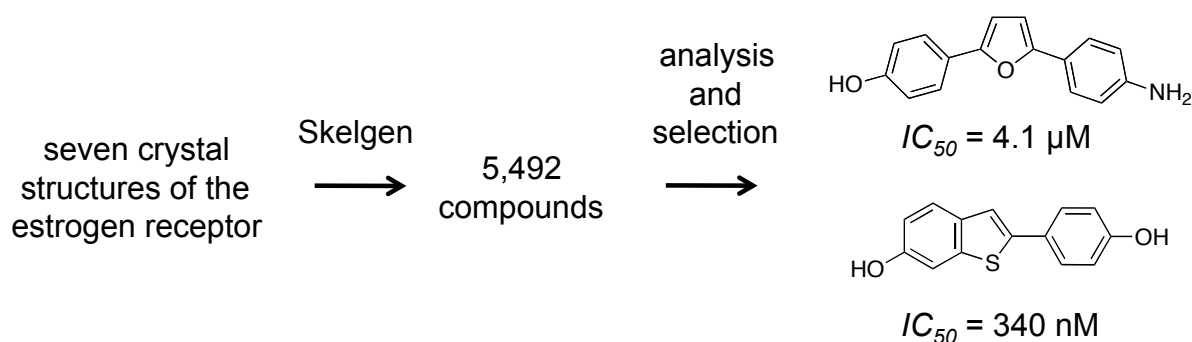
Scheme 1. LUDI¹⁰⁻¹⁴ has been successfully applied to design an inhibitor of the HIV-RT. Starting with the structure of the binding site, LUDI first placed fragments into sub-regions of the pocket and linked them by a phenyl moiety in a second step. The resulting structure was finally decorated with an amide sidechain by LUDI. Subsequent manual optimization exchanged the pyrrole ring with an imidazole to simplify the chemical synthesis. A series of compounds based on this scaffold has been synthesized and led to different active molecules for several enzymatic activities of the target (the given IC_{50} has been determined with respect to DNA polymerase activity of HIV-RT).

1.2.2 Skelgen

In the late 1990s, Todorov and Dean proposed two algorithms for computer-assisted molecule construction forming the backbone of the Skelgen software.^{39,40} Skelgen features a two-step process to generate new ligand candidates. The approach is similar to the idea of SPROUT¹⁹⁻²², which was published five years earlier in 1993. While the first step constructs bare molecular *skeletons*, the second step implements an atom type assignment in order to turn skeletons into complete virtual molecules. Molecular skeletons are constructed by the stochastic assembly of so called *template fragments*. For this purpose, Skelgen has access to a library of special fragments manually grouped into different template sets by the user. A template only consists of carbon and hydrogen atoms. After an initial skeleton has been generated, it is optimized to sterically fit the binding cavity of the receptor. During this process, the skeleton can be rotated and translated as a whole, and single bonds are rotated to sample the conformational space of the skeleton. In addition, fragments may be added, removed or exchanged. In the latter case, fragments are only replaced by other fragments belonging to the same template set. The whole optimization procedure is implemented as a simulated annealing process. Scores of skeletons are assessed by a scoring function that takes both intermolecular and intramolecular steric interactions into account as well as torsion energies. After the skeleton has been sterically optimized to fit the receptor binding site, element types are assigned to skeleton vertices in the second step. The aim is to maximize the complementarity of the emerging molecule to the pocket in terms of electrostatic and hydrogen bonding interactions. For this purpose an empirical scoring function is used. A branch-and-bound algorithm in combination with a depth-first search is employed to exclude unfavorable element type assignments and efficiently find good solutions for this combinatorial problem. Although originally implemented as a receptor-based method, a later

version of Skelgen also features a ligand-based design mode based on three-dimensional steric and pharmacophoric constraints derived from a reference ligand.⁴²

In a large study Firth-Clark and coworkers employed Skelgen to generate new ligand candidates for the human estrogen receptor (ER) α .¹²¹ Skelgen generated a total of 5,492 unique designs based on seven different crystal structures of the target protein. For each receptor structure the 50 top scoring molecules were selected for a subsequent clustering analysis. These 350 compounds could be clustered into 22 distinct sets based on common substructures. Out of 17 compounds picked for synthesis and testing (selection was performed to cover a broad range of clusters), five (30%) showed >40% inhibition at a concentration of 10 μM . Five compounds (of which four are structurally novel) have an $IC_{50} \leq 25 \mu\text{M}$. The most potent compound exhibits an IC_{50} of 340 nM (Scheme 2).



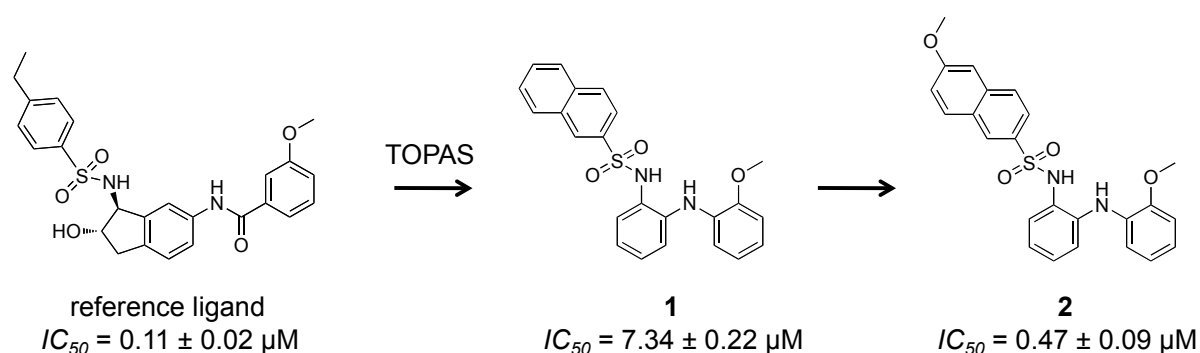
Scheme 2. The two most potent inhibitors designed by the Skelgen software³⁹⁻⁴⁴ for human estrogen receptor α .

1.2.3 TOPAS/FLUX

TOPAS^{52,53} and its successor FLUX^{62,63} are examples of a purely ligand-based *de novo* design paradigm. Designed molecules are evaluated by their similarity to a reference compound. For this purpose, the relative topological distributions of potential pharmacophore points on the two-dimensional molecule structure are calculated (CATS descriptor¹²²). Molecular fragments for construction are derived from the disassembly of known bioactive compounds by applying the RECAP⁹⁶ rules retrosynthetically. The same rules also guide the forward design process: during construction, only fragment attachment sites derived from the same cleavage rule can be reconnected (*i.e.* a carbonyl and a nitrogen attachment site are only allowed to form an amide bond if both have been part of an amide bond prior to disassembly). An evolutionary algorithm directs the search process: a strict selection criterion allows only the fittest compound of a ‘population’ to survive and produce offspring by random fragment exchange (mutation operator). In contrast to TOPAS, FLUX also features a crossover operator

recombining parts of ‘parent’ structures to generate new fragment combination in the next generation. Starting with a randomly assembled compound, the optimization process is supposed to breed structures with increasing fitness over time.

An example for successful ligand-based *de novo* design has been published by researchers at Roche in 2000.⁵³ TOPAS suggested structures supposed to block the human K⁺-channel Kv1.5. The top-scored molecule **1** was synthesized and showed the desired effect on the target (Scheme 3). Minor modification led to compound **2**, which is equal to the reference ligand with respect to potency.



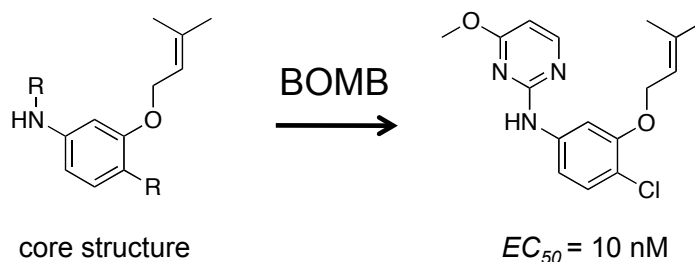
Scheme 3. Compound **1** designed by TOPAS and its close structural analog **2** were synthesized and block the human K⁺-channel Kv1.5.

1.2.4 BOMB

The software BOMB⁶⁵ features free energy perturbation (FEP) as a post-run scoring scheme to compute relative binding energies for the most promising designs. FEP makes use of the thermodynamic cycle in order to estimate differences of binding energies between (preferably close) structural analogs (relative binding energies). For this purpose, one ligand has to be ‘morphed’ into another by incremental small steps of structural changes. Each intermediate step needs to be evaluated in terms of binding energies to the receptor pocket. This time-consuming and computationally demanding process is one of the most sophisticated methods available to estimate relative differences in binding energies taking solvation effects into account. Since the morphing process works best on structurally similar compounds, BOMB generates series of ligands by decorating a fixed core fragment with various sidechains. The selection of the core fragment and its placement within the binding site is accordingly the first step of a design run. Several layers of fragments from a fixed set of building blocks that are clustered into multiple groups can be added to grow the seed structure to final ligand candidates. Each grown molecule is geometrically optimized within the pocket by a force field minimization and evaluated according to a QSAR-like scoring function that was trained

to reproduce experimentally determined activity values. Finally, FEP is employed to re-score and rank the most promising compounds after a design run.

BOMB has been successfully employed in a *de novo* design study to identify a series of new inhibitors of HIV-RT.⁶⁵ An example of a designed compound from this study is presented in Scheme 4.



Scheme 4. Example of a potent inhibitor of HIV-RT designed with the help of BOMB⁶⁵ based on a fixed core structure.

Similar to BOMB, a few *de novo* design programs start off with a user-defined fragment. Although this breaks with the concept of ‘pure’ *de novo* design to invent new molecules from scratch, it is a worthwhile strategy to incorporate knowledge about privileged fragments into the design process. Depending on the focus of the software, the seed fragment can be anything from molecular scaffold (*e.g.* BOMB, COLIBREE⁶⁸) to a set of sidechains (*e.g.* Recore¹²³).

1.3 Outline

This work presents a new approach to computer-assisted *de novo* design of ligand candidate structures. Special focus was put on the practical evaluation of the software. Only a small number of *de novo* design programs have been tested for their ability to propose synthetically feasible compounds by practical synthesis. This represents a problem computer-aided *de novo* design suffers from since the beginnings of the research field. The main reason might be the extensive costs and effort associated with chemical synthesis of candidate molecules. The decision to synthesize a compound depends on the estimated tradeoff between the ease of synthesis and its presumed chance to exhibit the desired biological activity. Enhancing the ease of synthesis of candidate molecules therefore raises the probability that some compounds will actually be selected for synthesis and practical testing. For this reason, the proposed

software DOGS (Design of Genuine Structures) was developed to maximize the chance of designed structures to be synthesizable with little effort. In fact, DOGS not only suggests new compounds but also provides the user with at least one possible synthesis pathway for each compound. DOGS features an assembly process based on available molecular building blocks and a set of established reaction schemes, which forces the software to follow up construction pathways representing direct blueprints for possible synthesis routes. Only a small number existing software tools (*e.g.* SYNOPSIS⁵⁷ and BI CLAIM⁷⁷) provide the user with synthesis routes for designed compounds.

Despite the suggestion of synthesis pathways, the reaction-driven construction of candidate molecules can be exploited in an additional way: Restrictions dictated by chemical reactions limit the number of constructible molecules in a well-motivated way. This can be exploited by the applied search algorithm, as the size of the search space is significantly narrowed down compared to an unconstrained combination of fragments. DOGS features a deterministic search algorithm implemented as a greedy strategy. Molecules are grown in a stepwise process, in which for each extension cycle not more than the best k of all generated molecules will be followed up in the next round.

Quality of designed products is assessed using a ligand-based scoring scheme. Similarity to the reference ligand is computed by a graph kernel method especially suited for the stepwise growing process. Two graph representations of molecules (*molecular graph* and *reduced graph*) have been implemented to allow for different levels of abstraction from the two-dimensional molecular structure.

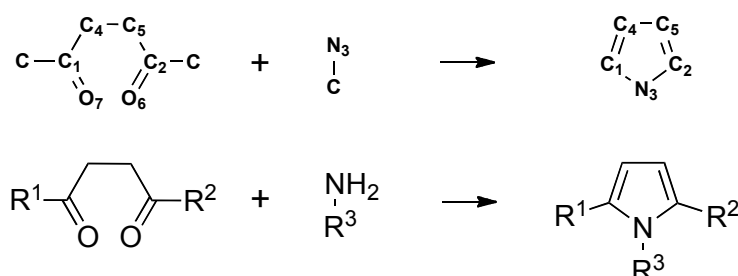
Theoretical evaluation of the software with respect to general properties of designed compounds was performed as well as analyses of generated scaffolds. Finally, DOGS was tested for its ability to contribute to a realistic drug design project in two practical case studies on ligand design for human γ -secretase and human histamine H₄-receptor.

2 Material and Methods

2.1 Library of Chemical Reactions

The way the DOGS algorithm builds up new candidate structures mimics a stepwise synthesis pathway as applied in a laboratory. This strategy is supposed to deliver a direct blueprint for the actual synthesis of proposed candidate structures. For this approach, established reaction protocols need to be formalized in order to make them processable by a computer. The reactions applied within DOGS were encoded using the formal language Reaction-MQL¹²⁴. Reaction-MQL is a line notation language that can be used to describe functional transformations of molecules. The specification of a reaction as a Reaction-MQL expression consists of an educt side on the left and a product side on the right. Educts are specified only by substructures that are directly involved or essential for the reaction (*reaction center*) in order to make the description applicable to wide spectrum of educts with variable substituent groups (R-groups). The product is described as bond rearrangements caused by the reaction (Scheme 5). All Reaction-MQL representations used here feature educts with variable R-groups in order to make them as generic as possible and broaden the spectrum of possible products.

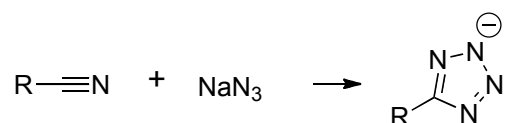
C-C1[!ring](=O7)-C4[!aromatic & bound(-H)]-C5[!aromatic & bound(-H)]-C2(=O6)-C ++
C[!bound(=O)]-N3[allHydrogens=2 & charge=0] >> Paal-Knorr pyrrole >> C1\$1-N3-C2=C5-
C4\$=1



Scheme 5. Example of a Paal-Knorr pyrrole reaction encoded as Reaction-MQL expression (*top*). Educt substructure descriptions (*left part*) are separated by ‘++’. Educt side and product side (*right part*) are separated by ‘>> ID >>’ where *ID* is an arbitrary identifier for the reaction. A direct structural representation of the line notation description including atom identifiers is shown in the middle. The conventional structural representation of the reaction (*bottom*) denotes variable parts of molecules by R-groups (R^x).

Catalysts or invariant educts are not denominated in the reaction string. For example, the reaction expression presented in Scheme 6 does not explicitly list sodium azide as an educt because it will not introduce variable sidechains on the product side. Invariant educts are implicitly included on the product side by adding relevant atoms. Only educts explicitly mentioned in a Reaction-MQL expression are considered as reaction components in DOGS. This means that the reaction of Scheme 6 is referred to as a one-component reaction in this work, although its real-life counterpart involves more than one educt.

C-C1#N2 >> Tetrazole >> C1\$1=N2-N[charge=-1]-N=N\$-1



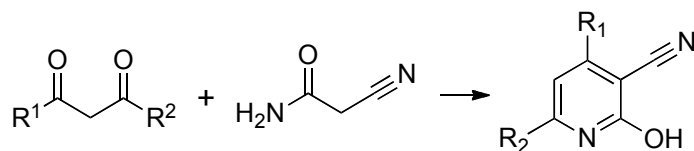
Scheme 6. Reaction-MQL representations (*top*) only list educts with variable sidechains. Atoms from invariant educts are automatically added when processing the reaction. For this reason the sodium azide is not explicitly present in the reaction expression, although it is part of the reaction (*bottom*).

A set of established reaction protocols was collected from the literature and encoded as Reaction-MQL expressions. Special focus was drawn on ring closure reactions forming substructures of pharmacological interest. Other selection criteria comprised high product yields, simple application, broad diversity with respect to educt R-groups and minimal exertion of toxic catalysts. Although preferred, a reaction did not necessarily have to fulfill all requirements to be considered.

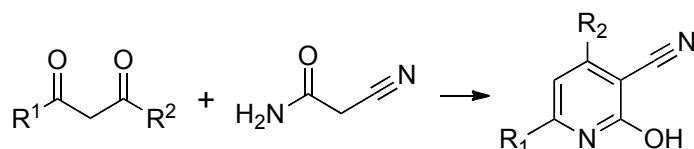
The collected set comprises 83 reactions, of which 58 are unique and 25 are charge or symmetry variations (a complete list can be found in section *Coupling Reactions* in the supplement). Out of the 58 unique reactions 34 describe ring formations. A reaction is classified as a ring closing reaction if the product contains a cyclic substructure that is not present in one of the educts. All reactions require one or two educts (one- and two-component reactions) and result in a single product ($A \rightarrow B$ or $A+B \rightarrow C$). The fact that each specification only describes one product guarantees a one-to-one assignment of a reaction and a product. While this simplifies the application of virtual reactions during the design process it raises a problem when a reaction involves a symmetric educt substructure and is not characterized to be regioselective. In this case, the reaction is described by two distinct Reaction-MQL

specifications, each forming one regioisomer. An example of such a reaction is presented in Scheme 7.

C1-C2[!ring](=O10)-C3[allHydrogens=2]-C4(=O11)-C5 >> 3-nitrile pyridine (symmetry 1) >>
 N\$1=C(-O)-C(-C#N)=C2(-C1)-C3=C4\$-1(-C5)



C1-C2[!ring](=O10)-C3[allHydrogens=2]-C4(=O11)-C5 >> 3-nitrile pyridine (symmetry 2) >>
 N\$1=C(-O)-C(-C#N)=C2(-C5)-C3=C4\$-1(-C1)



Scheme 7. Example of a reaction forming regioisomer products due to a symmetric educt substructure. The reaction is split in two separate reaction expressions in DOGS. Corresponding Reaction-MQL expressions are presented above each reaction scheme.

2.2 Library of Synthesis Building Blocks

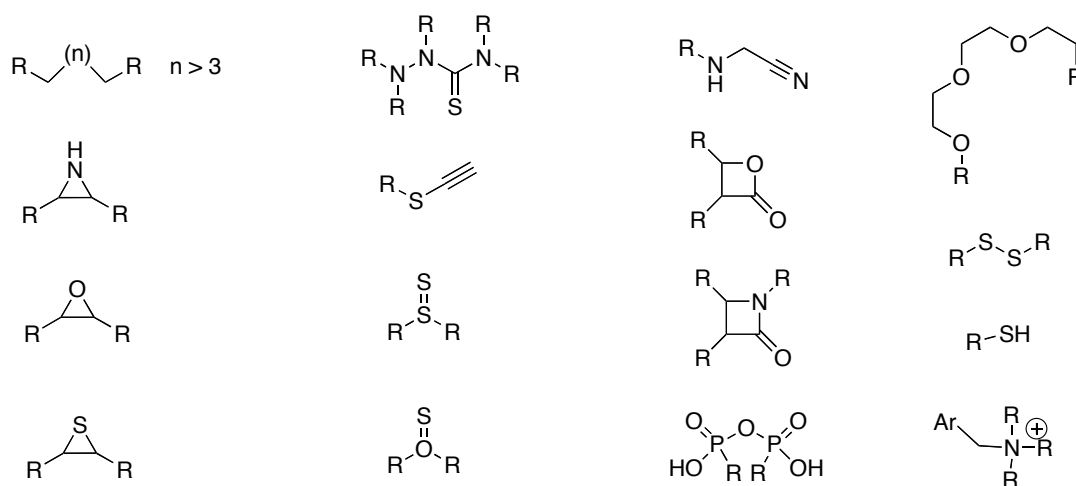
DOGS uses commercially available synthesis building blocks for the construction of new molecules. A subset of the Sigma-Aldrich¹²⁵ catalog containing about 56,878 chemical building blocks was downloaded as SDF file from the ZINC database.^{126,127} These compounds served as a basis to extract the final set of building blocks available to automated design by a three-step preparation protocol.

1) In the first step, building blocks were standardized and unsuitable entries were filtered out. For this purpose, a preprocessing routine was developed and implemented in the programming language SVL using the software MOE¹²⁸ (version 2009.10). This routine comprises multiple filtering criteria:

- Compounds with a molecular mass of less than 30 Da or more than 300 Da were removed.
- Compounds having more than 4 rings were removed.

- Compounds exhibiting any element type other than C, N, O, S, P, F, Cl, Br, I, B, Si and Se were removed.
- Compounds containing more than three fluoride atoms were removed.
- Compounds exhibiting atoms with incorrect valences were removed.
- Compounds exhibiting unwanted substructures (Scheme 8) were removed.
- Protonation states and formal charges were set according to MOE's washing routine (e.g. carboxylic acids were deprotonated, most primary, secondary and tertiary amines were protonated).
- Duplicate entries were removed.

Definitions of unwanted substructures (Scheme 8) were compiled on the basis of rules published by Hann and coworkers.¹²⁹



Scheme 8: Unwanted substructures according to Hann *et al.*¹²⁹ Building blocks containing one of these substructures are removed from the stock of building blocks for DOGS. (Ar: aromatic)

2) In the second step, the filtered compound set was processed by a collection of preprocessing reactions. A set of 15 functional group addition (FGA) and functional group interconversion (FGI) reactions was compiled from the literature and encoded as Reaction-MQL expressions (for a complete list of preprocessing reactions see section *Preprocessing Reactions* in the supplement). FGA/FGI reactions are supposed to introduce reactive functional groups to building blocks in order to make them applicable to coupling reactions in the design process. Every time a building block was converted by any of the 15 reactions, the original version was kept and the converted building block added to the library.

3) The final step of the preparation process comprises the annotation of reactive substructures present at each molecular fragment. Structural information and substructure annotations were then stored in a MySQL¹³⁰ database. For annotation, every building block was checked for the presence of any reactive substructure defined in the 83 Reaction-MQL expressions of coupling reactions. A bit vector storing this information was built for every synthesis fragment. The bit vector holds a '1' at a certain position if the respective substructure is present (*i.e.* the building block can serve as an educt for a certain reaction). Accordingly, the length of this bit vector is exactly the same as the number of reactive substructures defined by the coupling reactions. Storing this information together with each building block is supposed to speed up the selection of suitable reaction partners during the design process. In case a building block does not contain any of the defined reactive substructures (*i.e.*, all bits have zero values) the building block is neglected and not stored in the database because it will not be able to act as reaction partner during molecule construction.

Figure 7 summarizes the stepwise process of preparing the building block library. Starting with 56,878 synthesis fragments, the final library contains 25,144 entries.

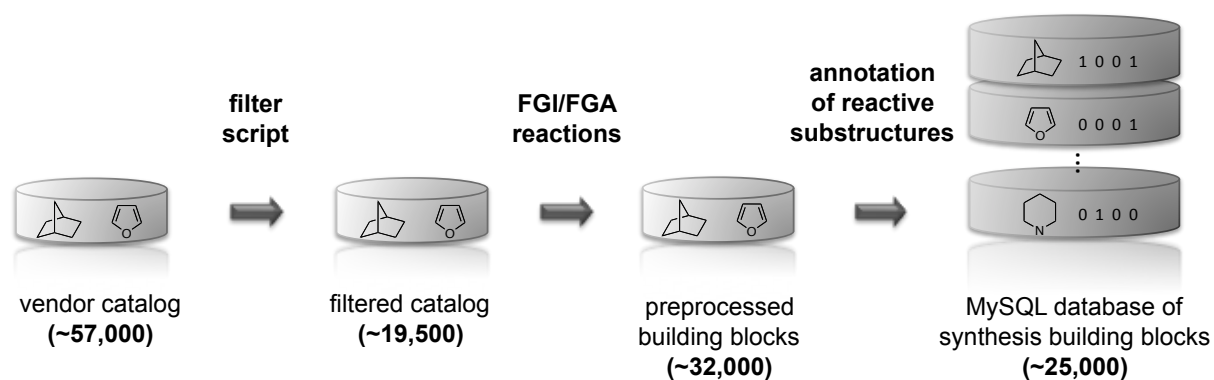


Figure 7. Preprocessing protocol setting up the DOGS building block library. Figures in parentheses give the number of building blocks involved at the respective stage.

2.3 Design Algorithm

DOGS generates new molecules by iterating through the design cycle. One design cycle comprises the modification of a current intermediate product by applying one of the chemical reactions from the library, *i.e.* the extension of the intermediate product. The product of one design cycle represents an intermediate product, which is modified in the subsequent iteration. A design cycle has two steps:

1. *Selection of applied reaction*

An intermediate product *Z* will typically exhibit more than one functional group that can be addressed by reactions from the reaction library. Each of these groups can potentially serve as an *attachment point* (AP) to connect another building block. In order to identify the most promising AP of *Z* and the reaction to apply, DOGS introduces the concept of *minimal dummy fragments*. A minimal dummy fragment is a virtual molecule that exclusively features the minimal structural demands that must be fulfilled to participate in a certain reaction. The application of this concept is supposed to estimate the minimum structural changes a reaction will introduce (Figure 8). The definition of a reaction therefore determines corresponding minimal dummy fragments, as they depend on the way a reaction defines reactive substructures involved. A one-component reaction does not define any minimal dummy fragment. It can directly be applied to a molecule without the involvement of a second reactant. Thus, structural changes to *Z* do not need to be estimated but are determined by simply applying the reaction. In contrast, a two-component reaction defines two minimal dummy fragments.

In order to extend *Z*, the algorithm first detects which of the implemented reactions can be applied to the attachment points offered by *Z*. Each of these reactions is applied to *Z* with a complementary minimal dummy fragment, leading to a list of *dummy products*. Here, one dummy product corresponds to exactly one reaction. By subsequently scoring the dummy products DOGS implicitly scores the corresponding reactions. The reaction breeding the top scoring dummy product is selected to be pursued in the next step. In case more than one top scoring reaction is identified all of them are considered in step 2.

2. *Selection of synthesis building block*

In case step 1 selected a one-component reaction it is directly applied, and *Z* is modified accordingly. Otherwise (two-component reaction), the reaction is performed using all building blocks from the library holding the respective reactive substructure (Figure 8). Every generated product is scored according to the scoring function. The top-scored compound is selected and represents the extended intermediate product for the next design cycle. In case more than one intermediate product scores favorable, all of them are considered for the next round. In order to truncate the number of molecules generated during each step and to prevent combinatorial explosion, the

maximal number of intermediate products proceeding to the next extension round is limited to 10.

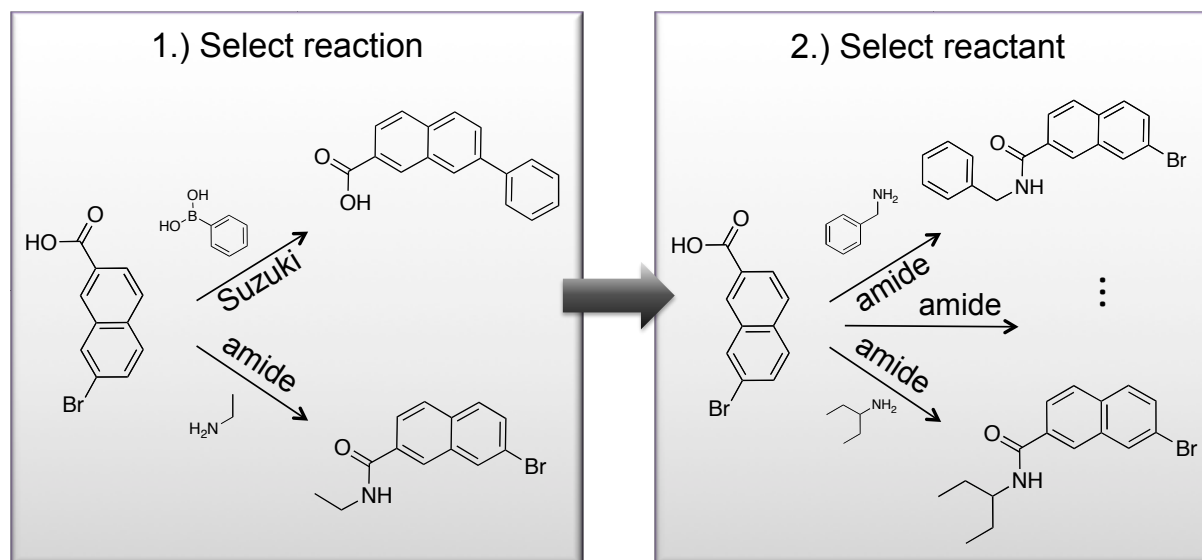


Figure 8. Two-step procedure of an extension cycle. Step 1 selects the reaction by scoring generated dummy products. In the example, only two reactions can be applied (Suzuki coupling and amide coupling), and the amide dummy product scores favorable. In step 2, all educts from the building block library exhibiting a suitable amine are added to the growing molecule. The top-scoring product represents the extended intermediate product and is selected for the next design cycle.

The building block a synthesis path starts with is selected among all entries of the library. For this purpose the algorithm evaluates every building block processed by the dummy reaction steps according to the scoring function. Each of the n top scoring building blocks are considered as a starting point for a distinct synthesis path. The value of n is defined by the user to control the number of compounds proposed during a design run.

Once the design of a new compound based on a selected starting building block is initiated it will be continued until one of two stop criteria is fulfilled.

The first stop criterion controls the molecular mass of designed compounds. The reference compound's mass (100%) defines a relative lower (70%) and upper (130%) bound. A constructed molecule has to exhibit a molecular mass lying within these boundaries to be accepted as a valid final product. During the design of a new molecule the algorithm continuously adds building blocks until the resulting intermediate product exceeds the lower mass boundary. Up to this step the extension of the intermediate product is accepted even if the score degrades from intermediate product i to $i+1$. Once the molecular mass of an intermediate product lies within the defined range, the algorithm will only accept a

subsequent extension step if it results in an improvement of the score. In case the addition of a building block leads to a lower score or causes the molecular mass to exceed the upper weight constraint, the last reaction step is neglected and the previous intermediate product is added to the list of final products.

The second stop criterion is supposed to truncate the number of synthesis steps in order to keep proposed synthesis pathways short. A pathway is interrupted regardless of any other condition when a certain number of synthesis steps (here: 4) is exceeded. In this case, the intermediate product formed by the last valid reaction step is added to the list of final products and a new synthesis pathway is launched based on another starting building block. Figure 9 presents the core of the design algorithm.

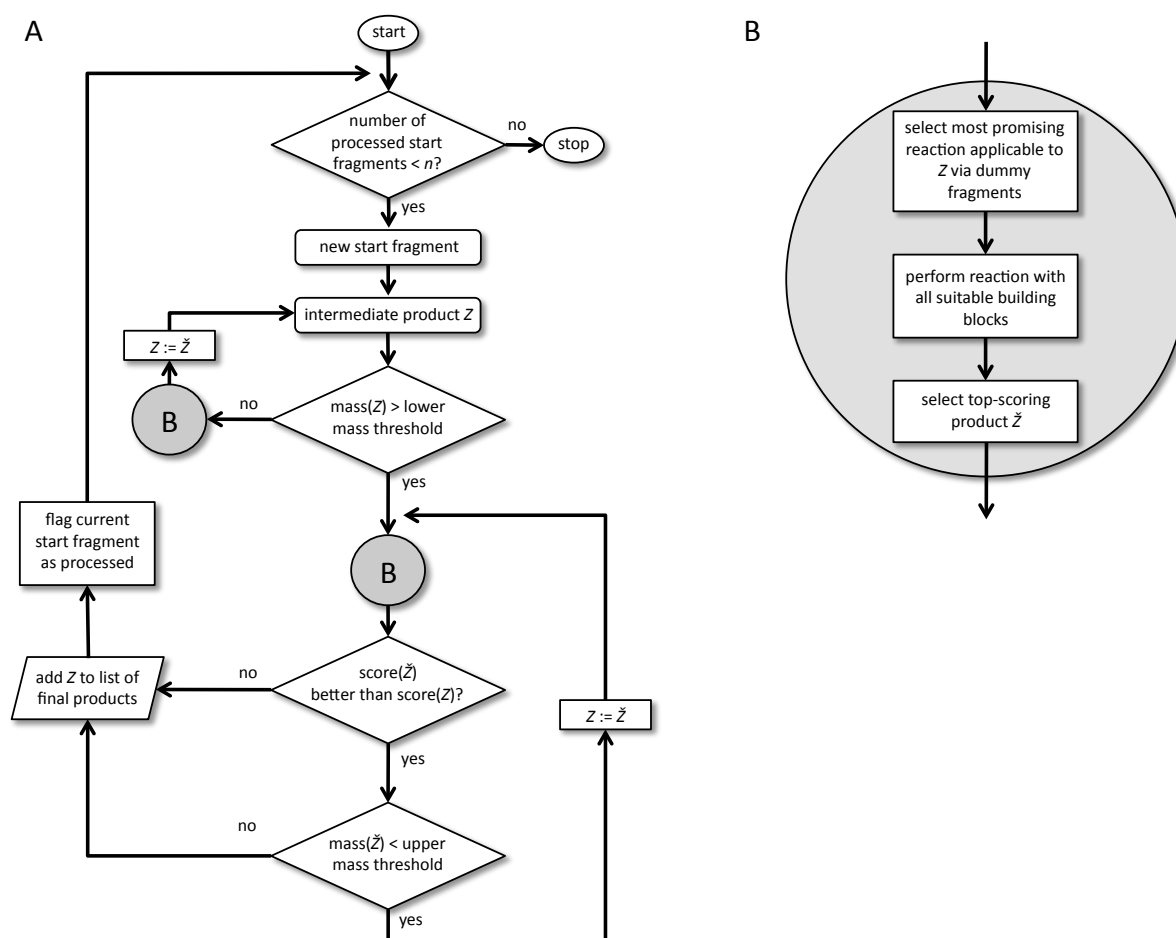


Figure 9. A: Flowchart of the DOGS design algorithm. The stop criterion for maximum number of reaction steps is not included. B: Detailed description of flowchart element B (grey circle). It comprises the key steps to modify intermediate product Z in order to yield \tilde{Z} by applying *in silico* reactions.

DOGS tries to construct at least one compound starting from each of the n building blocks considered as most promising starting points. It is possible that an initiated synthesis path does not produce a final product. This is the case if the growing intermediate product does not offer an attachment point to add another building block before it exceeds the minimal mass limit. DOGS automatically skips this particular synthesis and increments n by 1 to guarantee that at least n final products are generated. Typically, a run will result in more than n final products because synthesis pathways can split if different top scoring intermediate products are generated. In this case, more than one final product will be designed on the basis of a starting building block. All steps of the design algorithm are deterministic, *i.e.* two runs of DOGS with identical parameters will deliver identical results.

2.4 Scoring Function

2.4.1 Graph Kernel Method

The scoring function assesses the quality of a molecule with respect to the design objective. Products of each stage of a virtual synthesis pathway (dummy products, intermediate products, final products) are evaluated by the same scoring function. DOGS uses a 2D graph kernel method (ISOAK¹³¹) for scoring the designed molecules. The graph kernel was originally developed for similarity searching in virtual screening, where it has been successfully applied¹³². ISOAK can be readily employed as a scoring function for ligand-based *de novo* design, where, like in virtual screening, similarity to a given reference ligand forms the key objective for the design process.

ISOAK computes the similarity of two molecules A and B based on their two-dimensional topological structure. Molecules are interpreted as graphs where atoms are represented as vertices and covalent bonds as edges between vertices (*molecular graph*). Hydrogen atoms and corresponding bonds are removed from the graph.

In the first step, ISOAK computes a similarity value for each pair of vertices between A and B (Figure 10, step 1).

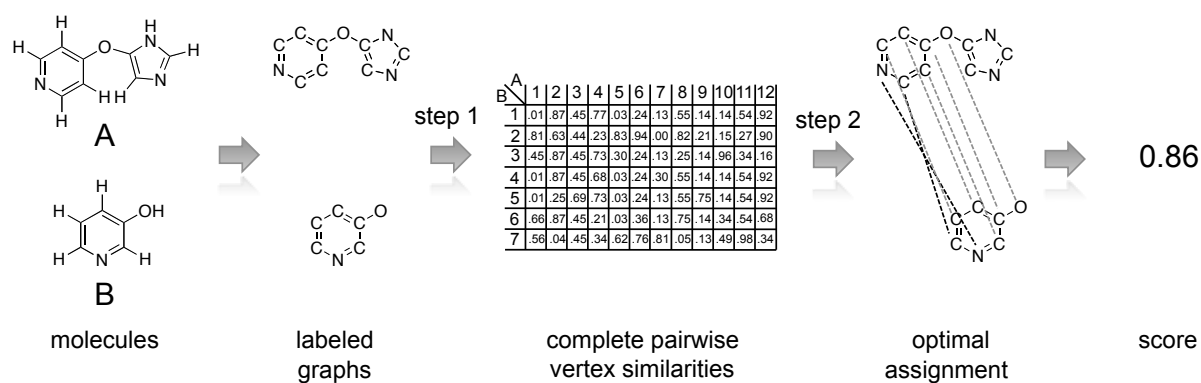


Figure 10. A similarity value for each pair of vertices between two labeled graphs is computed and stored in a matrix (step 1). An optimal assignment (step 2, dashed lines) of these vertex pairs maximizing the sum of similarities gives the final score. Some assignments are highlighted in black for better orientation.

The similarity of two vertices is influenced by two terms. The first term compares the isolated vertices themselves based on their *labels*. In the context of molecular graphs, examples for meaningful vertex labels are atom types, element types, pharmacophoric features (discrete labels) or partial charges and electronegativity (continuous labels). For the comparison of vertex labels a function $f_{vc}(v_1, v_2)$ is needed to compute a numerical value for a pair of labels expressing their similarity. For discrete labels the Dirac kernel can be used. The Dirac kernel is a simple function returning ‘1’ if the two labels are identical and ‘0’ otherwise. The second term of vertex similarity takes the local graph environment (surrounding vertices) into account. The basic idea behind the second term is that two vertices are similar if their topological neighbors are similar. This recursive measurement incorporates vertex similarities of neighbored vertices as well as a comparison of connecting edges. For edge comparisons, for example, the Dirac kernel based on bond order labels (single, double, triple) is applied. The recursive nature of this vertex similarity definition is expressed by an iterative computation, where vertex similarities of pairs of neighbored atoms used in the i -th iteration are taken from results of the previous iteration $i-1$. Similarities of iteration 0 are initiated with a standard value, e.g. 1. In each iteration, the final similarity of two vertices is computed as a weighted sum of the two components, where the influence of each component is controlled by a parameter α ($0 < \alpha < 1$). Component 1 (direct label comparison of vertices) is weighted by $1-\alpha$, while component 2 (recursive neighborhood comparison) is weighted by α . Higher values of α therefore increase the influence of the topological graph neighborhood on vertex comparison (Figure 11).

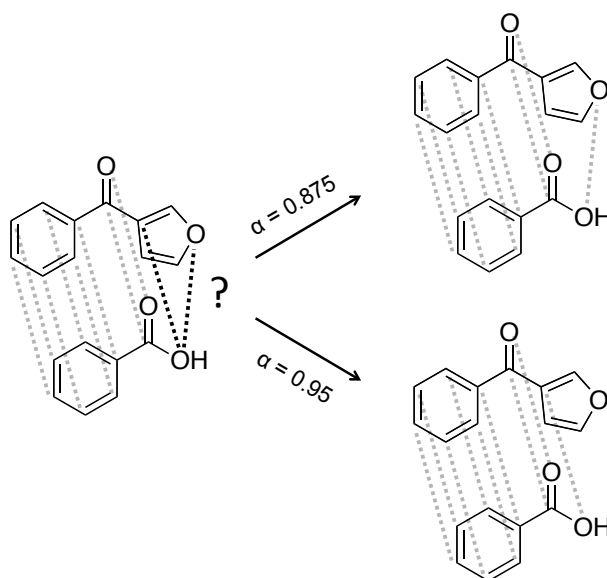


Figure 11. The assignment of the left part (grey dashed lines) is intuitive since the respective substructures of the two molecules are identical. The assignment of the oxygen atom of the smaller molecule (black dashes lines, *left*) depends on the ISOAK parameter α . Higher values of α emphasize on the local neighborhood and shift the assignment accordingly (*right*).

As a second step (Figure 10, step 2), calculated vertex similarities are used to compute an optimal assignment: Each vertex of the smaller graph is assigned to exactly one vertex of the larger graph. The assignment is optimal in the sense that it maximizes the sum of similarities for the assigned vertex pairs. In other words, for each vertex of the smaller graph ISOAK finds exactly one corresponding vertex in the larger graph. Note that it is not possible for a vertex to appear in more than exactly one pair, *i.e.* a vertex of the smaller graph cannot be assigned to a vertex of the larger graph that has already been assigned to another vertex and vice versa. The total similarity of two graphs is finally computed as the sum of all similarities between the assigned vertices.

2.4.2 Modification of the Graph Kernel Method

The ISOAK kernel as published¹³¹ and described above was slightly modified to adapt it to the requirements of DOGS. The following changes were introduced:

Edge labels: An additional label ‘aromatic bond’ has been introduced to complement the existing labels ‘single bond’, ‘double bond’ and ‘triple bond’. The obvious advantage is that now all bonds of aromatic systems are treated equally, which better reflects their actual

physical properties. The original implementation distinguishes between single and double bonds of aromatic systems, which can lead to artificial dissimilarity between identical substituted aromatic systems represented as different mesomeric resonance structures. The Dirac kernel based on these four discrete labels is applied to comparing edge labels.

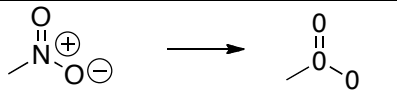
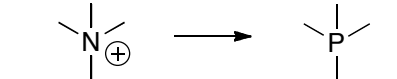
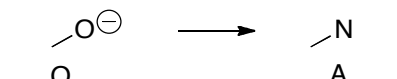
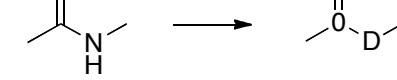
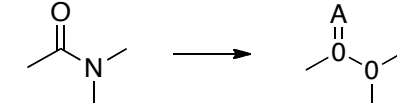
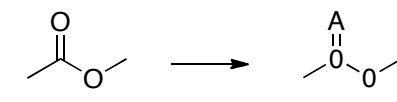
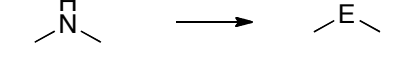
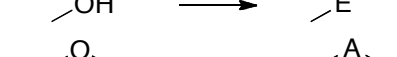
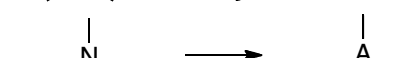

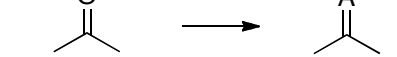
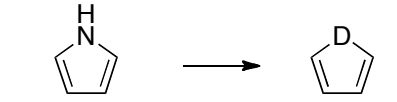
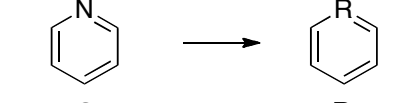
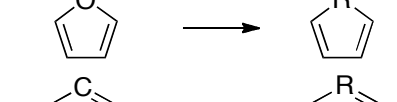
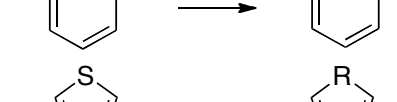
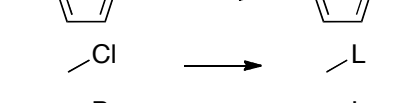
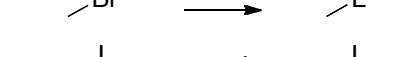
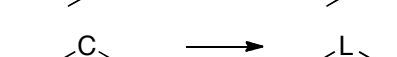
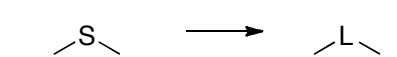
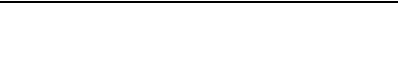

Vertex labels: Instead of labeling vertices with element types of corresponding atoms, the vertices are labeled with pharmacophoric features (pharmacophore typing is described in detail in section *Pharmacophore typing* below). Pharmacophore types describe atoms by their potential molecular interaction with a receptor molecule. Depending on their molecular environment, atoms of different element types can have the same pharmacophore type, which leads to an abstraction from the mere chemical nature of an atom, focusing on its potential to interact with a biological target. Since the goal of *de novo* design is to find novel chemical structures while keeping the desired biological effect of the reference ligand, a pharmacophore description of molecules ought to be beneficial in this context. For the comparison of pharmacophore vertex labels the Dirac kernel is used.

Graph reduction: Although designed for virtual screening of molecules, ISOAK is a general method to compare labeled graphs of any kind. The *reduced graph* representation of molecules was implemented as an alternative to the molecular graph introduced above. A reduced graph represents certain substructures of a molecule comprising more than one atom as single nodes: circular substructures as well as neighbored atoms sharing the same type ‘lipophilic’ or ‘no type’ are condensed to one vertex. Bit vectors are used to label vertices of reduced graphs (labeling and vertex comparison are described in detail in section *Graph reduction* below). A reduced graph represents a more abstract molecule description and is supposed to complement scoring based on the more detailed molecular graph. The user chooses which of the two graph representations will be applied in a design run of DOGS.

2.4.3 Pharmacophore Typing

Each vertex of a molecular graph is labeled by one of seven pharmacophoric features (A: hydrogen bond acceptor, D: hydrogen bond donor, E: hydrogen bond donor & acceptor, P: positive charge, N: negative charge, R: aromatic, 0: no other type) depending on the corresponding atom of the molecule. Typing is performed by applying a set of substructure definitions expressed as MQL¹³³ strings (Table 2). All atoms not explicitly typed by one of these rules are assigned to have no type (‘0’). Table 2 presents the typing rules in the order they are applied to a molecule. The order is important because an atom that has already been typed by one rule will not be typed again.

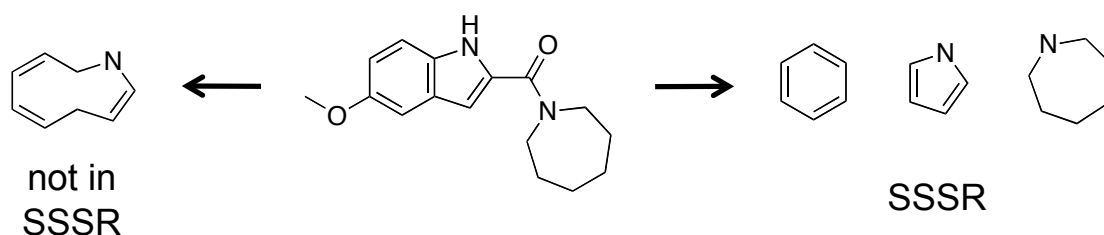
Table 2. Substructure definitions and corresponding typings.

MQL substructure definition	Type	Example
O[charge=-1]-N[charge=1]=O	0;0;0	
*[charge>0]	P	
*[charge<0]	N	
O=C-N[allHydrogens>0]	A;0;D	
O=C-N[allHydrogens=0]	A;0;0	
O=C-O[allHydrogens=0]	A;0;0	
N[allHydrogens>0 & !aromatic & !bound(-C=N) & !bound(-S=O)]	E	
O-H'	E	
Heavy'[!aromatic]-O-Heavy'[!aromatic]	A	
N[allHydrogens=0 & !{aromatic&totalConnections=3} & !{bound(-C=N) & !bound(=C)} & !{bound(-C:N) & !bound(=C)} & !bound(-S=O)]	A	
O=*[C P S N]	A	
N[{allHydrogens=1 & aromatic} {allHydrogens>0 & bound(-C=N)} {allHydrogens>0 & bound(-S=O)}]	D	
N[aromatic]	R	
O[aromatic]	R	
C[aromatic]	R	
S[aromatic]	R	
Cl	L	
Br	L	
I	L	
C[!bound(~N)&!bound(~O)]~*[C F Cl Br I S]	L	
S[!bound(~N)&!bound(~O)]~*[C H]	L	

2.4.4 Graph Reduction

The graph reduction process is supposed to convert the molecule into an acyclic graph in order to represent it on a higher level of abstraction from its atomic structure. This is achieved by condensing certain substructures consisting of more than one atom to single graph vertices. The reduced graph often contains fewer vertices than the number of heavy atoms in the molecule (in contrast to the molecular graph, which always exhibits as many vertices as there are heavy atoms in the molecule). There are three cases in which atoms neighbored in a molecule are condensed and represented by one vertex, *i.e.* (i) cyclic substructures, (ii) clusters of atoms typed as ‘lipophilic’, and (iii) clusters of atoms typed as ‘no type’. The pharmacophore typing is identical to the one described in section *Pharmacophore Typing* above. A cluster is defined as a set of atoms of the same type that form topological neighbors. Not all atoms of a cluster have to be directly connected via one bond but can also be linked via other atoms belonging to the same cluster.

Cyclic substructures can consist of more than one ring. In the following, the term ‘ring’ will mean a cyclic substructure that is part of the *smallest set of smallest rings* (SSSR). Practically speaking, the SSSR is the set of all cyclic substructures with minimal numbers of atoms. All ring atoms must be covered by this set. A ring that only consists of ring atoms that can be completely covered by a combination of smaller rings will not be part of the SSSR (Scheme 9). A more formal definition can be found in reference 134.



Scheme 9. The ring on the left side is not part of the smallest set of smallest rings (SSSR) of the molecule in the center because the corresponding ring atoms can be covered by a combination of two smaller rings. The SSSR of the molecule is given on the right.

The graph reduction algorithm represents each ring of the SSSR as one vertex. In case a ring system contains only atoms that do not belong to more than two rings (*e.g.* naphthalene) it is possible to represent each ring by a single vertex and connect them in such a way that their topological order in the molecule is preserved in the reduced graph (Figure 12A). There are, however, cases where this is impossible in a straightforward way. For example, it is not

possible to find an acyclic graph layout of all rings that are part of the SSSR for phenalene or adamantane which preserves their topological order in the molecule (Figure 12B). In order to solve this problem, the algorithm searches for atoms being part of more than two rings of the SSSR and combines these rings to one vertex in the reduced graph. Please note that this breaks the usual ‘one-ring-one-vertex’ relation between the molecule and the respective reduced graph. Ring systems represented by a single vertex will be termed *amalgamated* in the following.

In order to distinguish the reduced graph representation of two adjacent rings that are connected by a bond and two rings that share atoms, the corresponding vertices of reduced graphs representing the rings are connected by an edge of order one (‘single bond’) in the former case and two (‘double bond’) in the latter case (Figure 12C).

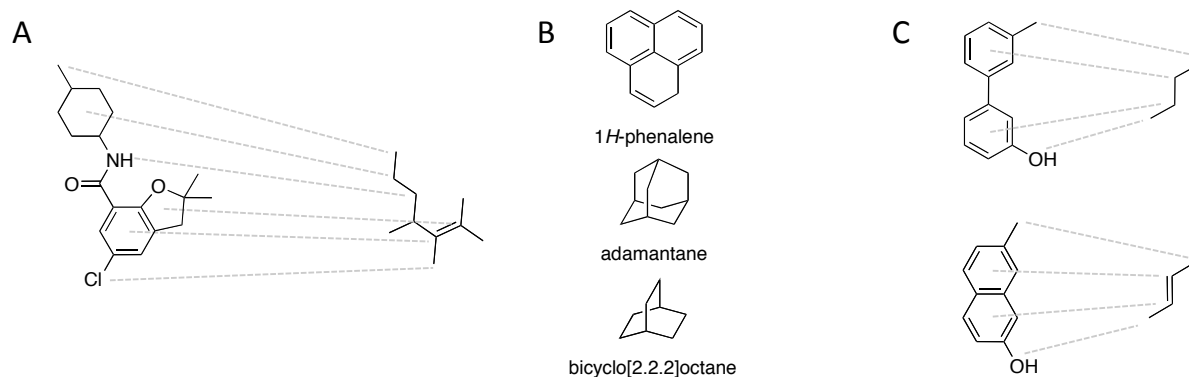


Figure 12. A: An example of a reduced graph representation. Dashed lines connect atoms or rings of the molecule (*left*) with their corresponding vertex of the reduced graph (*right*). For clarity only some lines are shown. B: Examples of polycyclic substructures (‘amalgamated’) represented by only one vertex in the reduced graph. C: Edges of order two are used to connect fused rings (*bottom*) in order to distinguish the shown cases of neighbored rings in reduced graph representation.

Labels of reduced graph vertices keep information about the atom(s) they represent. A bit vector of length nine stores which of the pharmacophore types are present in the respective substructure. Each of the seven pharmacophore types is represented by one bit. Two additional bits stand for ‘cyclic substructure’ and ‘amalgamated ring system’. A bit is set to ‘1’ if the corresponding feature is present in the substructure, ‘0’ otherwise. In addition, a vertex also stores the number of atoms it represents (atom count). Accordingly, a benzene substructure would be converted to a single vertex labeled by a bit vector with bits set for ‘ring’ and ‘aromatic’ and an atom count of six.

Bit vectors (bv) and atom counts (ac) are used to compute the similarity of two vertices A and B of reduced graphs. The similarity is computed by multiplying two components (equation 5).

$$f_{vc}(ac_A, ac_B, bv_A, bv_B) = sdFactor(ac_A, ac_B) * Ti(bv_A, bv_B) \quad (5)$$

Term 1 ($sdFactor$) returns a value between 0 and 1 depending on the difference between the atom count values of compared vertices (Equation 6), computed as

$$sdFactor(ac_A, ac_B) = \begin{cases} 1 & \text{if } |ac_A - ac_B| = 0 \\ 0.98 & \text{if } |ac_A - ac_B| = 1 \\ 0.9 & \text{if } |ac_A - ac_B| = 2 \\ 0.8 & \text{if } |ac_A - ac_B| = 3 \\ 0.5 & \text{if } |ac_A - ac_B| = 4 \\ 0.3 & \text{if } |ac_A - ac_B| = 5 \\ 0 & \text{if } |ac_A - ac_B| > 5 \end{cases} \quad (6)$$

Term 2 is the Tanimoto index (Ti) for bit vector comparison (Equation 7), calculated as

$$Ti(bv_A, bv_B) = \frac{c}{a + b - c} \quad (7)$$

where c is the number of bits commonly set to 1 in both vectors, a is the number of bits set to 1 in bv_A and b is the number of bits set to 1 in bv_B . Two identical bit vectors result in $Ti = 1$, while bit vectors with no set bits in common score with 0. Component $sdFactor$ can be seen as a penalty function for atom count differences modulating the Tanimoto index. In case the atom count of compared vertices is equal (e.g. two six-membered rings are compared), f_{vc} reduces to the Tanimoto index. If the difference between the atom counts exceeds five, f_{vc} will return 0 regardless of the calculated Ti for the bit vectors.

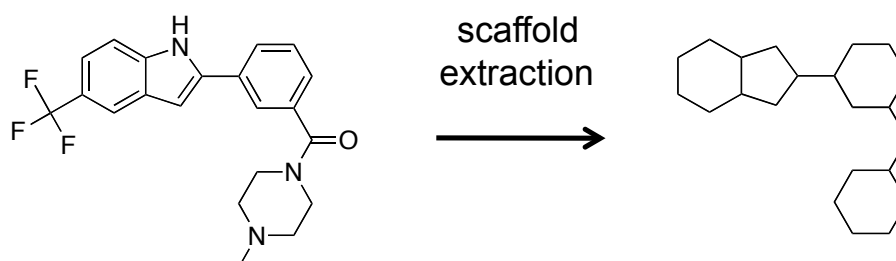
All other components of ISOAK including the edge comparison are identical to the molecular graph comparison. ISOAK can only process graphs with a maximum vertex connectivity of six, i.e. a vertex of a graph processed by ISOAK must not have more than six directly connected neighbors. While this will not happen in molecular graphs (typically, no element that is present in druglike molecules will form more than six covalent bonds), such cases can occur in reduced graphs. For example, naphthalene is represented as a single vertex and offers

up to eight positions for substitution. Molecules containing vertices with more than six neighbors in their reduced graph representation are excluded from subsequent steps and will be discarded.

2.5 Assessment of Scaffold Similarity

In order to assess the outcome of *de novo* design runs in terms of scaffold diversity and the software's potential to perform scaffold hops, a method to measure distances between scaffolds has been developed.

As there is no general definition of the term 'scaffold',¹³⁵ the definition of *graph frameworks* according to Bemis and Murcko¹³⁶ has been selected to describe the molecular scaffold in this work. Following their definition, a molecule's scaffold is extracted by keeping all cyclic substructures and the linker chains directly connecting them. Sidechains only connected to one or to no cyclic substructure are deleted. All retaining atoms are converted to carbon atoms, and all bonds are modified to single bonds (Scheme 10).



Scheme 10. Example of scaffold extraction according to the definition of Bemis and Murcko. The scaffold (*right*) only consists of the ring systems and the linker chains connecting them. All bonds have order one, and all atoms are converted to carbon atoms.

Scaffold similarities are computed as Euclidian distances in a descriptor space spanned by three descriptors ('number of rings', 'Petitjean', 'Kier1') from MOE¹²⁸ (v2009.10). These descriptors show comparably low to moderate cross correlations on an external test set (Table 3) and describe properties of the two-dimensional molecule framework. 'Number of rings' simply counts the number of all rings of the SSSR. 'Kier1' is the first of three kappa shape indices proposed by Hall and Kier.¹³⁷ It calculates a ratio between the number of atoms and bonds of a molecule $[(\#atoms-1)^2 / \#bonds^2]$. A slightly more elaborate measure for the 2D

shape of the molecule is computed by the ‘Petitjean’ descriptor¹³⁸: At first, an eccentricity value is determined for every atom of a molecule. It is defined as the longest of all shortest paths to every other atom in the molecule. The graph *radius* is the smallest atom eccentricity in the molecule and the graph *diameter* is the largest eccentricity value of the whole molecule. The descriptor is defined as (diameter-radius)/diameter.

Table 3. Descriptor correlations are expressed as the Pearson correlation coefficient (x100). Calculations were performed on a set of scaffolds derived from ~11,000 bioactive molecules.

	#Rings	Petitjean	Kier1
#Rings	100	-	-
Petitjean	21	100	-
Kier1	53	27	100

The external test set comprises about ~11,000 bioactive molecules with a molecular mass of <1000Da.¹³⁹ Scaffolds computed for this test set also served as a reference framework to establish a scaling procedure for descriptor values. Auto-scaling parameters (mean, standard deviation) extracted from test set descriptor values were applied to scale descriptor values of new compounds before computing Euclidian distances. This procedure adjusts the influence of each descriptor on the distance. The final distance between two scaffolds A and B was computed as given in equation (8).

$$d(A,B) = \sqrt{\left(\frac{rings(A) - 4.30}{1.12} - \frac{rings(B) - 4.30}{1.12}\right)^2 + \left(\frac{pj(A) - 0.47}{0.05} - \frac{pj(B) - 0.47}{0.05}\right)^2 + \left(\frac{k1(A) - 19.28}{5.64} - \frac{k1(B) - 19.28}{5.64}\right)^2} \quad (8)$$

2.6 Implementation

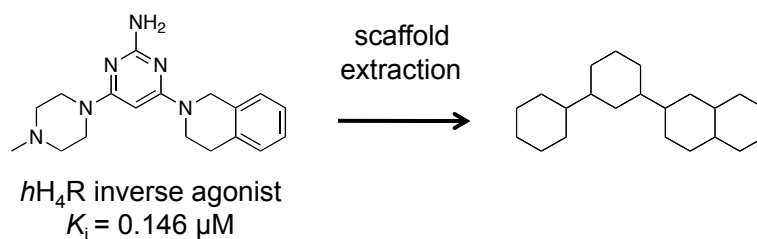
DOGS was implemented in Java¹³⁰ version 1.6 and uses the Chemistry Development Kit^{140,141} (CDK, version 1.0.2). Calculations were performed on an Apple Mac Pro with eight CPU cores (2 x 2.26GHz Quad-Core Intel Xeon) and 16GB RAM.

3 Results and Discussion

DOGS was evaluated theoretically with respect to general characteristics of the program like runtime, number of generated molecules and scaffolds. Designed scaffolds were assessed for their similarity to the reference scaffold in order to analyze the program's ability to propose ideas for novel scaffolds. DOGS designs were also investigated for general properties of interest for lead candidates, in particular druglikeness, synthesizability and calculated $\log P(o/w)$. In order to be of practical relevance, a *de novo* design software tool must be able to come up with molecules that can be synthesized and already show druglike properties. Exemplary compounds generated by the software for different target molecules (trypsin, TGF- β 1 receptor, estrogen receptor) were picked out and discussed with regard to their pharmacophoric features in comparison to the respective reference ligand. Here, DOGS was tested for its ability to generate compounds that capture the main features of the seed molecule while being structurally distinct. Finally, the software was analyzed in two practical *de novo* design case studies (H_4 receptor and γ -secretase). In both cases, proposed molecules were selected, synthesized and tested for their biological activity.

3.1 Influence of Parameters on General Characteristics and Scaffold Diversity

A goal of *de novo* design is to generate ideas for new scaffolds. In order to test DOGS for its ability to design compounds with scaffolds different from the scaffold of the reference ligand, result lists of runs started with different parameters were analyzed. The scaffold definition used in this investigation follows the one of Bemis and Murcko¹³⁶ (for details see section *Assessment of Scaffold similarity in Materials and Methods*). An inverse agonist of the human histamine H_4 -receptor served as reference ligand for all runs of this investigation (Scheme 11).¹⁴² The number of investigated start fragments was set to 200 in each case. For both graph representations, parameter α was varied from 0.1 to 0.9 in increments of 0.1, producing a total of 18 result lists.



Scheme 11. An inverse agonist of the histamine receptor served as reference ligand.¹⁴² The extracted scaffold on the right was used to analyze the similarity of scaffolds designed by DOGS.

3.1.1 General characteristics

Quantitative characteristics of each run are summarized in Figure 13. DOGS needed between 8 and 15 hours to finish a complete run. Runtimes have the tendency to rise with higher α values, which can be observed for both graph representations (Figure 13, *bottom charts*). The reason for this correlation is that higher α values increase the influence of the graph neighborhood on vertex comparison, which leads to more computational iterations until the comparison process converges. An exception to the general trend is observed for $\alpha = 0.4$ in reduced graph design mode. This is caused by the fact that all other runs were performed in parallel with other jobs on the same machine, while this run was performed on an idle machine.

In general, runtimes are comparable between the two molecule representations, giving rise to the assumption that additional computational costs caused by graph reduction are compensated by the faster comparison of less complex reduced graphs. Since the overall number of molecules designed during a run in reduced graph design mode is higher (Figure 13, *top charts*), the overall time needed to score a single molecule is lower for reduced graphs. Hence, the faster comparison even overcompensates the costs for graph reduction. The reason for an overall higher number of designed molecules in reduced graph mode may be addressed to the fact that the higher level of abstraction from the molecule increases the chance for different intermediate products receive the same score during design. In case more than one top scoring intermediate product occurs during construction, the process is split and more than one final product may be generated from the same start fragment.

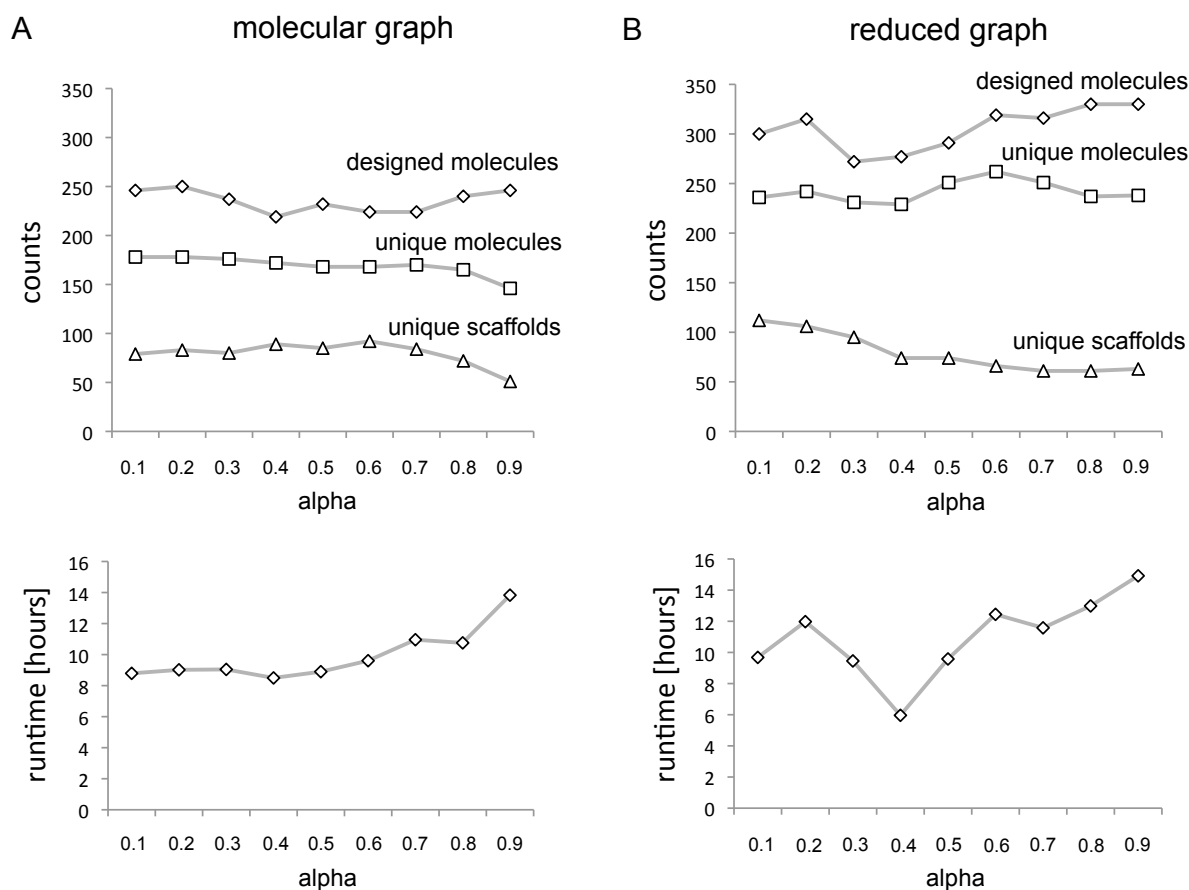


Figure 13. Influence of parameter α on different performance characteristics of DOGS for molecular graph (A) and reduced graph (B) representation. Lines connecting single measurements are given for better overview and do not represent interpolations.

The number of duplicate molecules produced during a run is widely independent of the molecular representation, and varies only slightly with changes of parameter α . It is impossible that duplicate molecules also have identical synthesis pathways, since the list of synthesis building blocks does not contain duplicates. However, it can be expected that a large fraction of duplicates result from only slightly differing synthesis routes. For example, two synthesis pathways can be identical except for the initial building blocks: In case the starting fragments of two synthesis pathways only differ in a halogen atom (bromide exchanged against iodide), which is substituted by an azide in the first step in both cases, they do not represent alternative synthesis strategies. However, duplicates may also be produced by significantly differing synthesis pathways. In this case, they add valuable information to the result list because they point to alternative synthesis strategies.

The number of unique scaffolds has the tendency to drop with elevating α values, although this effect is observable most distinctly at different parameter ranges for the two graph representations (0.6-0.9 for molecular graph representation and 0.1-0.4 for reduced graph).

3.1.2 Scaffold Diversity

In order to assess the influence of parameter α and the two different graph representations on the quality of outcome with respect to scaffold diversity, a numerical representation of scaffolds was used (for details see section *Assessment of Scaffold Similarity in Materials and Methods*). The descriptor representation allows for distance calculation between scaffolds of designed molecules and the reference scaffold. Figure 14 presents statistical parameters (median, average and standard deviation) of scaffold distance distributions as well as additional characteristics related with scaffold generation derived from analyzes of the 18 DOGS runs investigated in the former section.

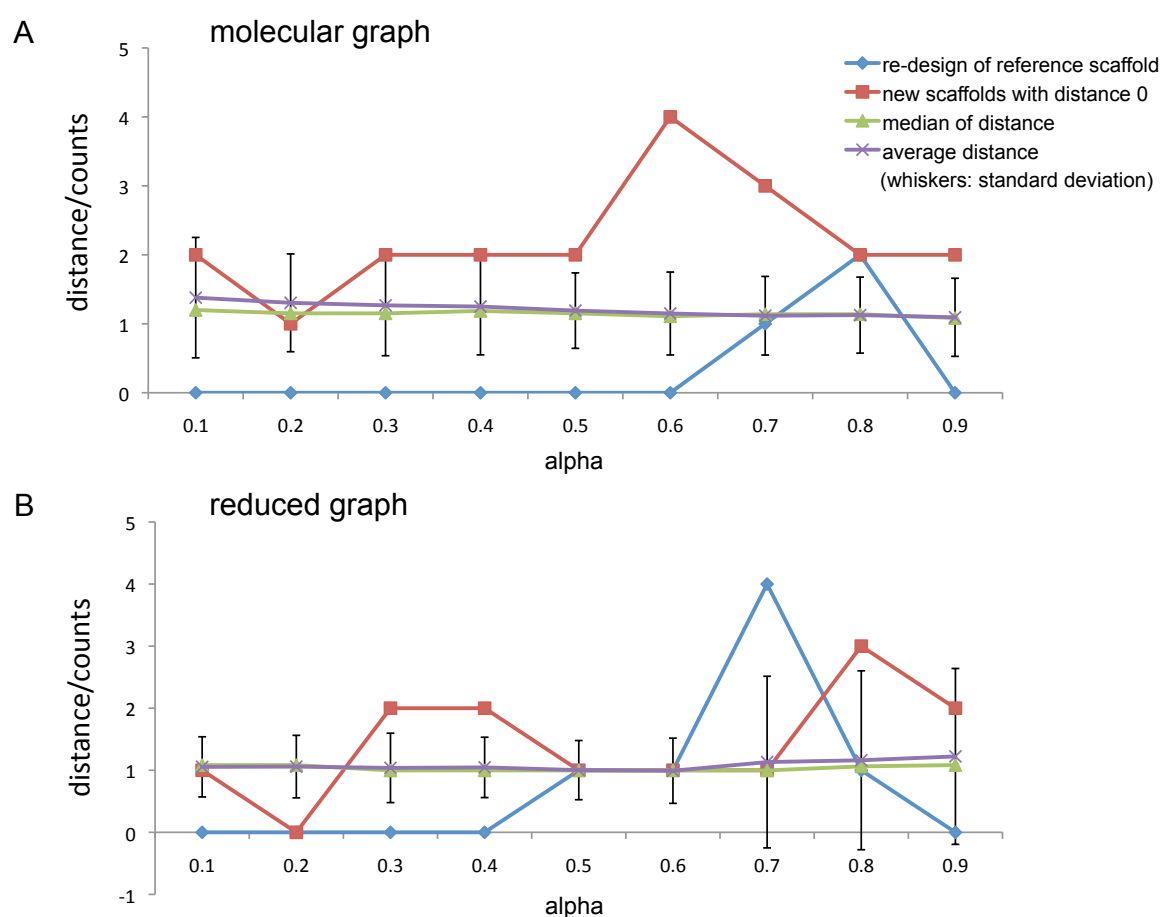


Figure 14. Influence of parameter α on characteristics and distribution of distances between designed scaffolds and the reference scaffold (A: molecular graph representation; B: reduced graph representation). Lines connecting single measurements are given for better overview and do not represent interpolations.

Median and average values of scaffold distance distributions only marginally differ between molecular graph and reduced graph representations at the same α level. The same holds true

when comparing different α levels of the same graph representation. Standard deviations slightly decrease with increasing α for molecular graph scoring. High α values in combination with the reduced graph design mode (0.7-0.9) produce scaffold distributions with increased standard deviations.

The number of molecules exhibiting the same scaffold as the reference molecule can serve as an evidence whether parameter combinations are able to generate close analogs of the reference molecule. In general, this is not the intention of *de novo* design, as the scope is to come up with innovative scaffolds. On the other hand, a low number of scaffold re-designs can be seen as an indicator that the algorithm designs ‘around’ the seed scaffold in scaffold space. In the given example, re-design of the reference scaffold is enhanced at higher levels of α for both molecule representations (0.6-0.8), while it is not observable at $\alpha = 0.9$. A possible explanation for this observation might be that such high levels of α influence the selection of initial building blocks in a way that they do not offer the potential to be transformed by suitable reactions to exhibit the reference scaffold.

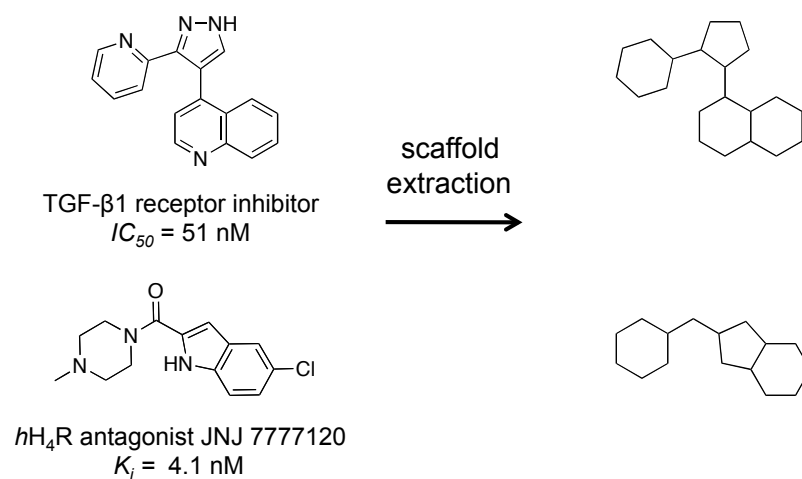
It is of greater interest for *de novo* design programs to be able to produce new scaffolds that feature pharmacophore and shape similarity compared to the seed. In order to test DOGS for the ability to propose new scaffolds similar to the reference, the number of scaffolds that structurally differ from the reference scaffold but exhibit minimal distance in the descriptor space (distance = 0) was counted in each result list (Figure 14). DOGS was able to at least design one new scaffold with distance 0 in every run except for one ($\alpha = 0.2$, reduced graph mode). The molecular graph representation always produced at least as many new ‘0-distance’ scaffolds on equal α levels as the reduced graph mode, in 66% of the cases even more.

Summarizing, these analyses do not give clear evidence on preferable parameters for DOGS runs. Visual inspection of several result lists based on different reference ligands revealed that molecular graph representation is preferably combined with α values in the high range (0.7-0.9). This is supported by the fact that the default setting for α is 0.875 in the original version of ISOAK for virtual screening.^{131,132} This method is almost identical to the molecular graph mode used here, as it also operates on topological molecule graphs. In contrast, reduced graph design works better on α values in the low to mid range (around 0.4). High α values tend to produce molecules exhibiting little similarity to the reference. This subjective finding is supported by the fact that exceptionally high standard deviations of scaffold distance distributions occur at these combinations of parameters (Figure 14B), giving evidence that

more scaffolds of larger distance to the reference are designed in these cases. The fact that result structures of the reduced graph representation were deemed to be more reasonable at lower α values than for the molecular graph may be addressed to the fact that ring systems commonly occur in druglike molecules and therefore play an important role in objective similarity measurements as well as in subjective human cognition. Reduced graphs represent rings or even ring systems by single vertices. By reducing the influence of the graph environment on vertex comparison *via* lower α values, ISOAK emphasized a direct vertex comparison, which means direct matching of rings in the case of reduced graphs. In contrast, matching of complete rings is enhanced by high α values in molecular graph design mode, since this forces ISOAK to put focus on the environment of compared atoms and incorporate connected ring atoms into vertex comparison.

In general, one might expect a more abstract molecule representation (reduced graph) to lead to more distant and diverse scaffold designs compared to a more detailed description of a molecule (molecular graph). Therefore, it might seem counterintuitive that design runs based on reduced graph representation do not breed scaffolds with a higher average distance to the reference than the molecular graph (Figure 14). It should be stated that the results of this investigation and drawn conclusions only hold for the selected descriptor space encoding the scaffolds. Future work needs to address whether similar observations can be made for different scaffold representations.

For further analysis of the influence of molecule representation on scaffold generation, another study of scaffold comparison based directly on the scaffold structures (instead of an abstraction by molecular descriptors) was performed. Four DOGS runs (reduced graph and molecular graph, each on $\alpha=0.875$ and $\alpha=0.4$) were carried out based on two different reference ligands: an inhibitor of the human transforming growth factor (TGF) $\beta 1$ receptor¹⁴³ and the *h*H₄R antagonist JNJ7777120¹⁴⁴ (Scheme 12). For these eight runs, distributions of scaffold distances to the reference scaffolds exhibit the same behavior as described in the former analysis for the *h*H₄R inverse agonist: no significant difference between distributions of reduced graph design and molecular graph design is observable (Figure 15).



Scheme 12. Reference ligands (a TGF β1 receptor inhibitor¹⁴³ and a *h*H₄-receptor antagonist¹⁴⁴) and extracted scaffolds.

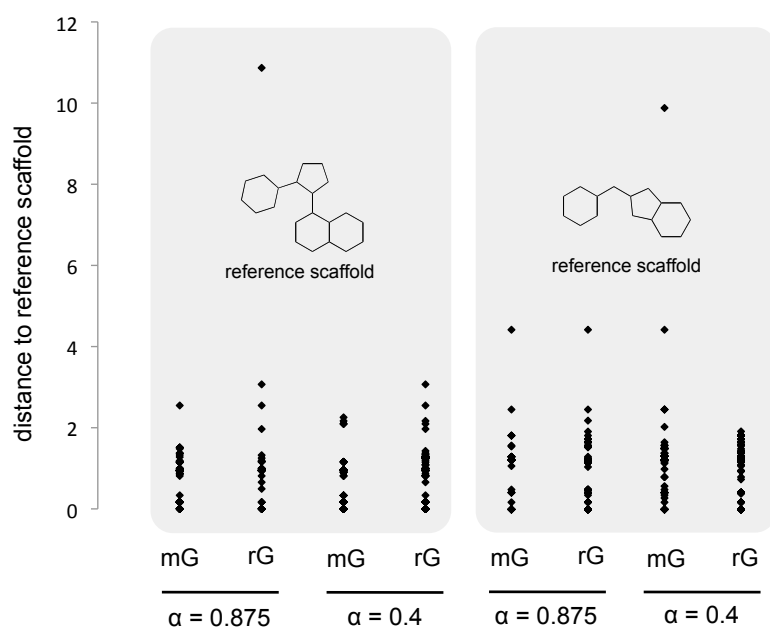


Figure 15. Distributions of distances between designed scaffolds and the respective reference scaffold for eight DOGS runs (*left*: TGF-β1 receptor inhibitor; *right*: JNJ 7777120; mG = molecular graph; rG = reduced graph).

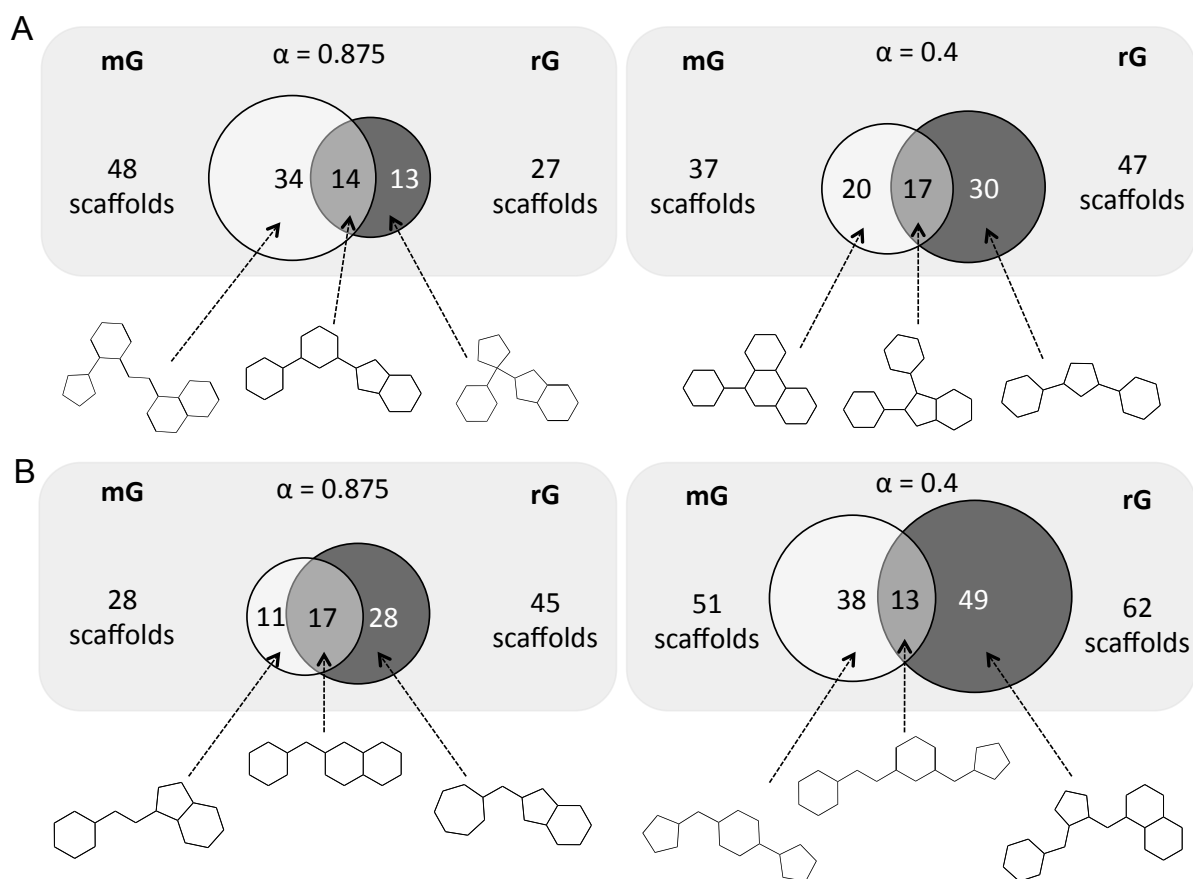


Figure 16. Overlaps of scaffold lists from runs based on molecular graph (mG) and reduced graph (rG) scoring (A: TGF- β 1 receptor inhibitor; B: JNJ 777120). The total number of scaffolds found in a run is given next to the corresponding circle. Figures inside of circle fractions describe numbers of scaffolds. Overlaps represent scaffolds constructed by both design modes. Examples of scaffolds for each fraction are given below.

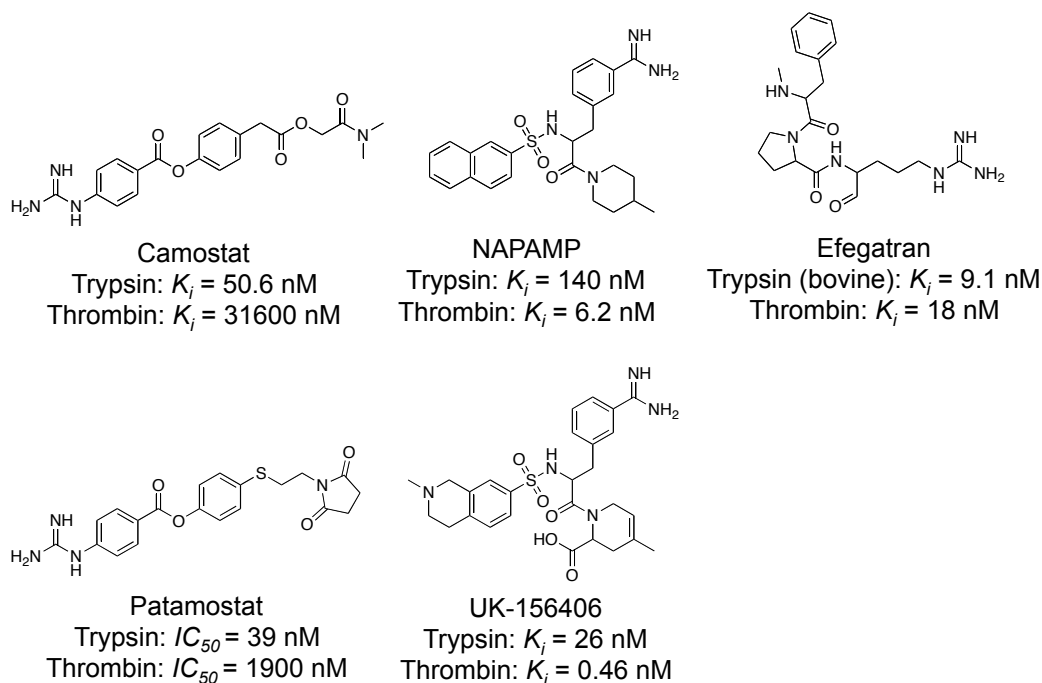
Nevertheless, a comparison of the scaffold lists produced with the two different abstraction levels on the same α value revealed that – albeit showing comparable distance distributions around the scaffold – they exhibit only small to moderate overlaps (Figure 16). This finding suggests that while reduced graph scoring did not jump farther away in the spanned descriptor scaffold space, it jumped into directions in structural scaffold space that considerably differ from those followed by molecular graph scoring. This leads to the conclusion that results produced on different abstraction levels of molecule representation are complementary. It is therefore worthwhile to apply both design modes to yield a richer pool of ideas for new scaffolds.

3.2 Property Analysis of Designed Compounds

De novo design programs are supposed to suggest compounds exhibiting druglike properties. Although successful *de novo* design campaigns will likely be followed by a process of

structural optimization in order to improve pharmacokinetic properties of designed compounds, it is evident that designed compounds should already show druglike properties themselves. Lipinski and coworkers have proposed four simple rules ('rule of 5'),¹⁴⁵ which have found wide acceptance as crude criteria for an estimation of oral bioavailability. These rules define negative guiding principles: with each additional failed criterion, the probability of showing poor absorption or permeation rises, which might lead to attrition in later steps of the drug development process. It is important to be aware of the fact that the 'rule of 5' present soft filters. Failing one of the filters does not necessarily mean that a molecule is not druglike and has no chance to become a drug. In fact, not all marketed drugs and drug candidates pass each of Lipinski's rules.¹¹⁹

In order to assess the druglikeness of DOGS designs, 'rule of 5' violations of 1,767 molecules originating from ten DOGS runs were computed using the descriptor implemented in the software MOE. Five trypsin inhibitors served as reference ligands for these runs (Scheme 13).



Scheme 13. Five trypsin inhibitors serving as reference compounds for DOGS design runs (Camostat¹⁴⁶, NAPAMP¹⁴⁷, Efegatran¹⁴⁸, Patamostat^{149,150}, UK-156406¹⁵¹).

For each reference, one run based on the molecular graph ($\alpha=0.875$) and a second run applying the reduced graph representation ($\alpha=0.4$) was performed. Strikingly, an analysis of 'rule of 5' violations shows that most of the compounds constructed by DOGS (78.5%)

violate less than two rules (Figure 17). Only 52 proposed molecules (3%) show three violations. The distribution of designed compounds mirrors the one of the reference ligands. A second analysis of druglikeness of DOGS designs was carried out for the same set of designs using an artificial neural network.¹⁵² This classifier had been trained on a set of drugs and non-drugs to score molecules between 0 (low druglikeness) and 1 (high druglikeness). Out of the 1,767 molecules designed by DOGS 904 (51%) receive a score >0.8 (Figure 18).

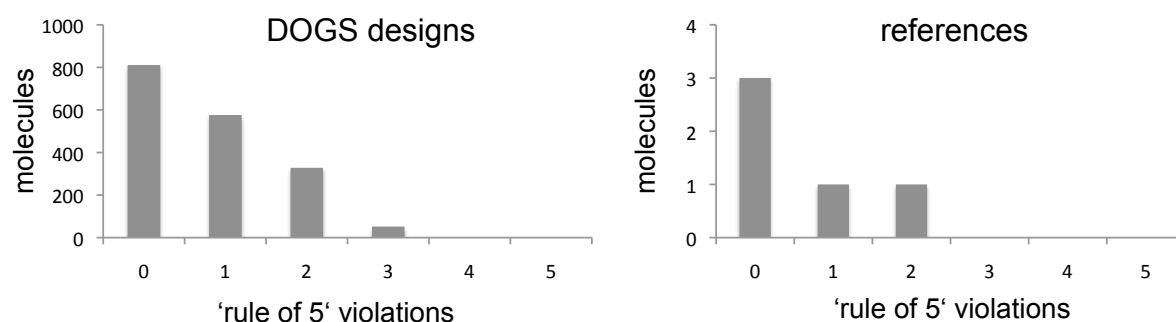


Figure 17. Distribution of 'rule of 5' violations of compounds designed by DOGS (*left*) and of respective reference compounds (*right*).

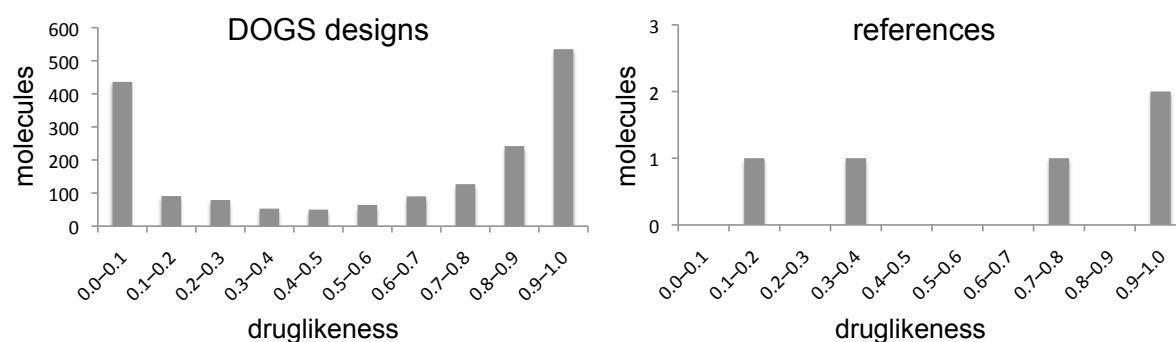


Figure 18. Distribution of druglikeness scores of compounds designed by DOGS (*left*) and the reference compounds of respective runs (*right*). Scores have been computed by a trained classifier (1 = high druglikeness).

Besides this result it is eye-catching that a considerable number of molecules (436) receive a poor druglikeness score below 0.1. This fact is less surprising if one considers that the set of reference compounds also contains a molecule deemed to be not druglike (Patamostat, score = 0.11). Compounds designed to maximize similarity to this reference can be expected to receive poor druglikeness scores as well.

Another relevant property for drug candidate molecules is lipophilicity.¹⁵³ A common parameter closely related to this property is the octanol-water partition coefficient ($\log P(o/w)$).¹⁵⁴ One of the Lipinski rules states that $\log P(o/w)$ values greater than 5 enhance the chance that a molecule will be poorly absorbed.¹⁴⁵ The $\log P(o/w)$ was calculated for the five trypsin reference ligands and the molecules designed by DOGS using the ' $\log P(o/w)$ ' descriptor implemented in MOE¹²⁸ (Figure 19).

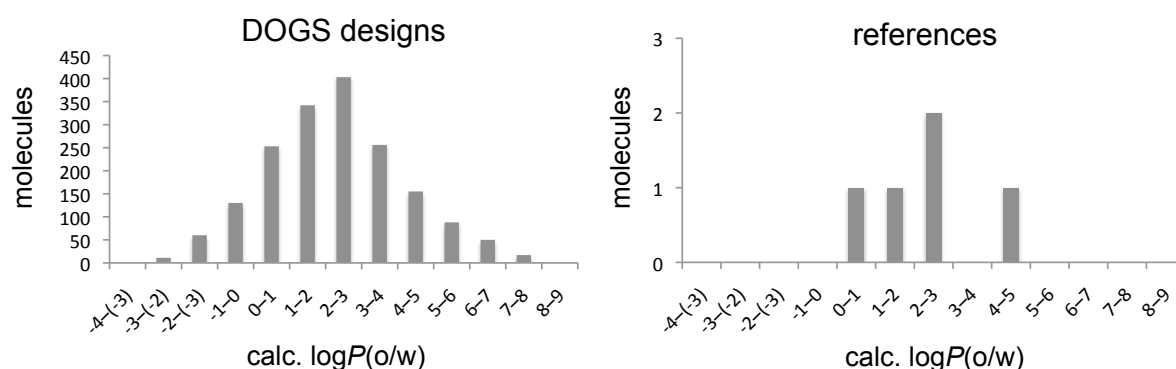


Figure 19. Distribution of calculated $\log P(o/w)$ scores of compounds designed by DOGS (*left*) and the reference compounds of respective runs (*right*).

The distribution of calculated $\log P(o/w)$ values of DOGS designs approximates a unimodal distribution centered around values between 2 and 3. This is in agreement with the distribution of values calculated for the reference ligands. DOGS was able to mimic this property of the references in the designed compounds, although it is not explicitly considered during the design.

It is of critical importance that molecules designed *in silico* not only exhibit desired properties but are also amenable to chemical synthesis in order to be of practical value for drug discovery projects. A molecular descriptor ('rsynth') implemented in the software package MOE¹²⁸ estimates synthesizability of molecules by the fraction of heavy atoms that can be traced back to starting material fragments resulting from retrosynthesis disconnection rules. A score of 1 means full coverage of atoms and expected high synthesizability. The *rsynth* descriptor was calculated for both the reference set and the set of *de novo* designed molecules (Figure 20). The majority of DOGS designs is deemed synthesizable (77% of compounds receive a score of >0.9).

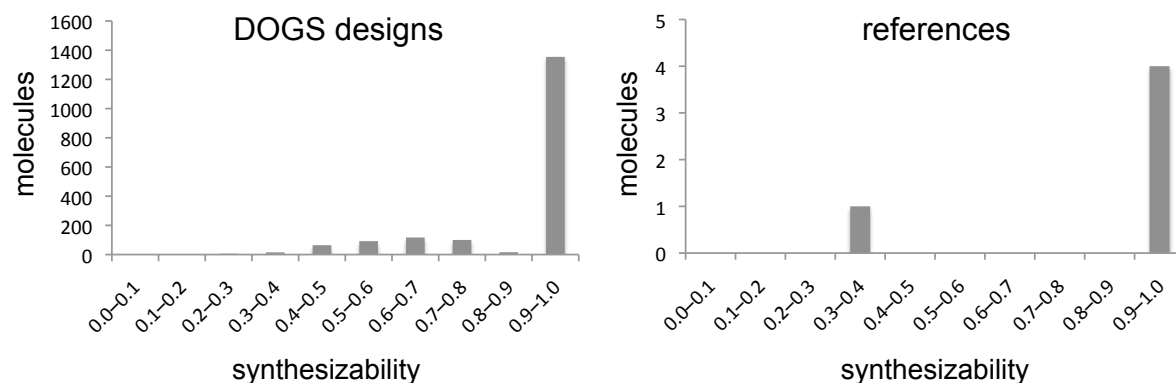


Figure 20. Distribution of estimated synthesizability scores of compounds designed by DOGS (*left*) and the reference compounds of respective runs (*right*). A score of 1 means perfect expected synthesizability.

Most of the remaining designs receive scores between 0.4 and 0.8. Reference compound UK-156406 is scored comparably low ($r_{\text{synth}} = 0.37$). A total of 35.5% (141 of 397) of all DOGS designs scoring below 0.8 originate from this reference ligand, which exceeds an expected fraction of 20% assuming that low-scoring designs come from all five references in equal parts. That means low synthesizability scores are enriched for molecules originating from a reference compound that is scored unfavorable as well.

In conclusion, this result may be considered a success of the DOGS approach to obtain synthesizability of *de novo* designed compounds.

Summarizing, DOGS is able to design overall druglike and chemically plausible molecules with a chance of being amenable to chemical synthesis. The proposed molecules resemble the reference compounds in properties that are not explicitly considered by the scoring function.

3.3 Exemplary DOGS Designs

3.3.1 Trypsin

Trypsin is a serine protease found in the digestive system of vertebrates. Its enzymatic activity comprises the cleavage of amide bonds in the protein backbone. Cleavage sites of trypsin are characterized by lying next to basic amino acids (arginine, lysine) in C-terminal direction.¹⁵⁵ The main reason for cleavage site specificity of trypsin is the S1 binding pocket (Figure 21), which is selectively filled by basic amino acid sidechains to interact with an aspartate residue at its bottom.¹⁵⁵ Inhibition of trypsin itself is of little pharmaceutical interest, but can be used as an example for case studies on serine proteases. From a pharmacological perspective,

trypsin represents an off-target for drug discovery projects directed to therapeutically relevant serine proteases like thrombin and factor Xa.¹⁵⁶

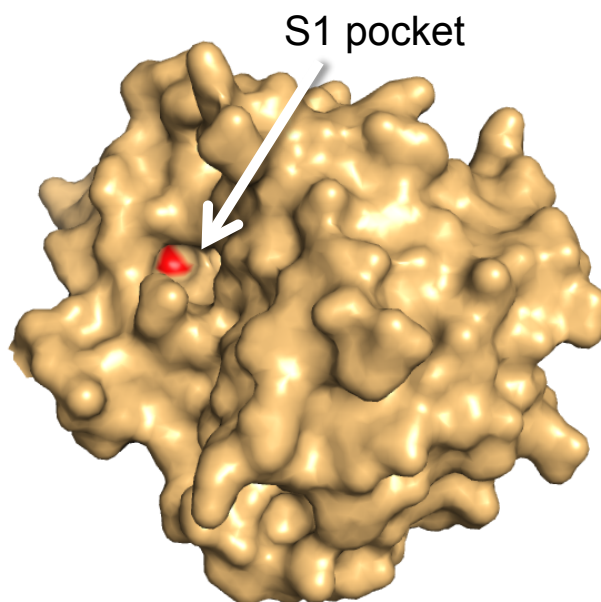
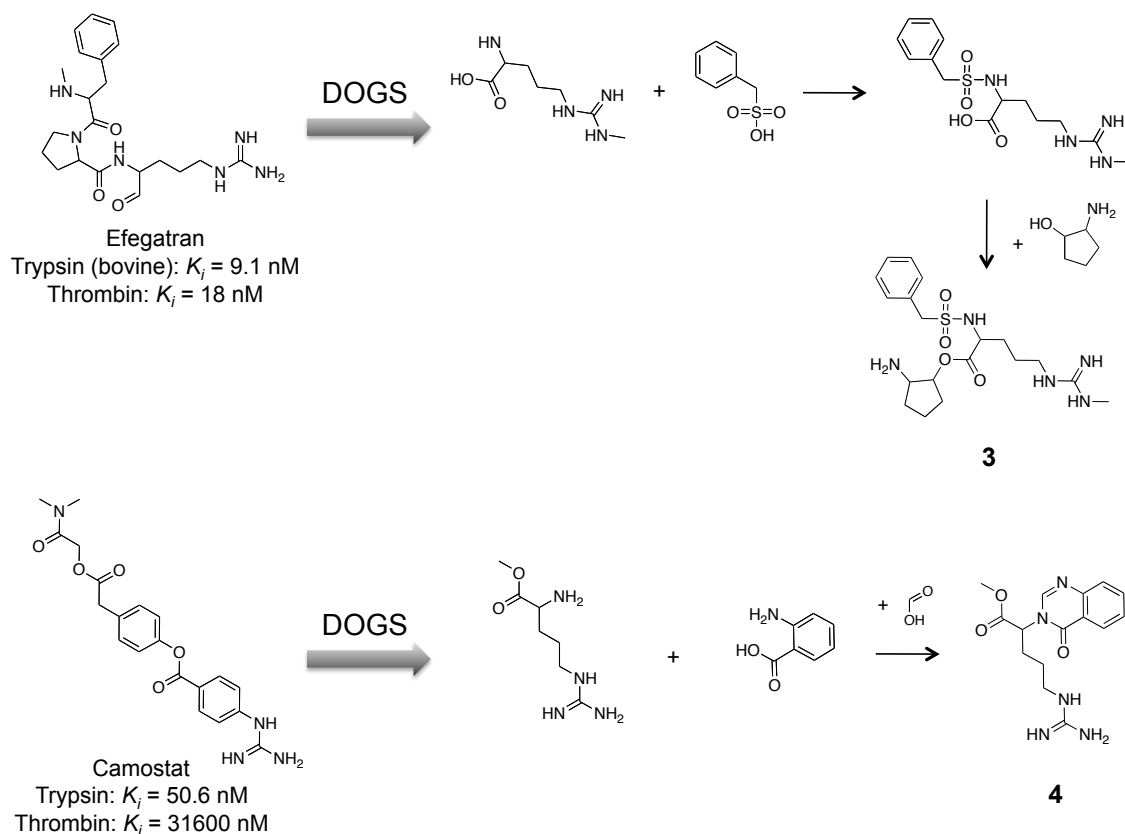


Figure 21. Crystal structure of human trypsin I (pdb-identifier 1trn; only one of the two monomers is shown). An aspartate residue (red surface area) is located at the bottom of the S1 pocket.

Two examples of structures proposed by DOGS as potential trypsin inhibitors are given in Scheme 14. Structures **3** and **4** were obtained from design runs based on Efegatran and Camostat (200 start building blocks, $\alpha = 0.4$ for reduced graph, $\alpha = 0.875$ for molecular graph). Compound **3** originates from the reference ligand Efegatran. It exhibits a central sulfonamide moiety, which is not present in the reference molecule but can be found in other trypsin inhibitors (for example in NAPAMP and UK-156406, Scheme 13). That means DOGS replaced a substructure of the reference by a structurally different but presumably isofunctional fragment, which is present in other known actives. The guanidinium sidechain of Efegatran was exchanged with the close structural analog 3-methylguanidinium.



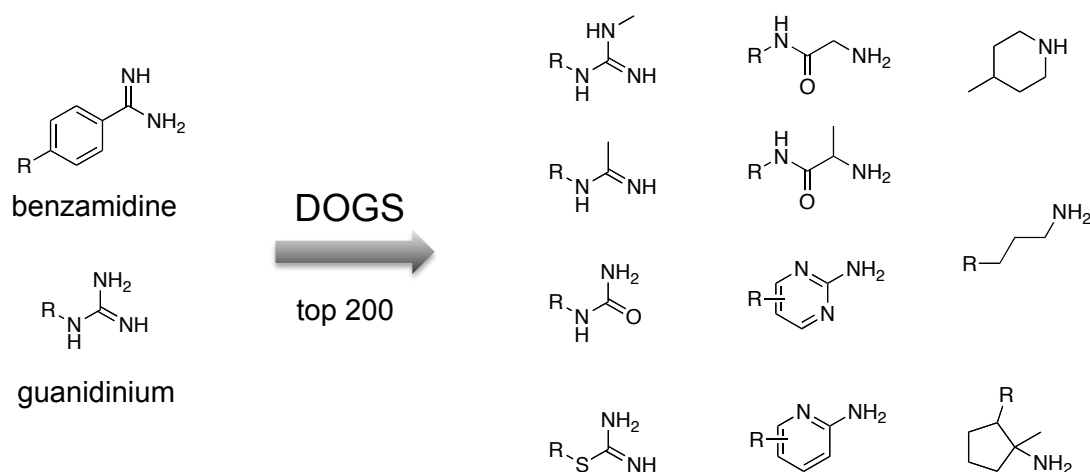
Scheme 14. Compounds **3** and **4** have been proposed by the software as potential trypsin inhibitors. Reference ligands (Efgatran¹⁴⁸, Camostat¹⁴⁶) and suggested synthesis pathways are presented for both candidate structures.

The overall composition of functional groups in **3** resembles the arrangement of the reference. The synthesis route proposed by DOGS will probably have to be augmented by the use of protection groups. For example, the formation of the ester bond in the last synthesis step can be disturbed by the competing formation of an amide bond with the primary amine of educt 2-aminocyclopentanol. Protection of the amine group could remedy this difficulty. Note that DOGS currently does not consider protection groups. Competing side reactions are only addressed by avoiding multiple occurrences of the same functional group in an educt.

Compound **4** has been derived from Camostat. Compared with the former example of molecule **3**, molecule **4** is generally more distinct from its reference compound with respect to the molecular structure. While the guanidinium group of the reference is preserved, it is connected to an alkyl chain instead of a phenyl ring. Alkyl linkers connecting the guanidinium group can also be found in the reference Efgatran and in the sidechain of arginine, a ‘natural’ ligand of the trypsin S1 pocket. An aromatic substructure in distance to the part addressing the S1 pocket is another feature that can be found in other trypsin ligands as well as in compound **4** (compare NAPAMP, Scheme 13). Albeit showing considerable structural difference to the reference compound it originated from, compound **4** represents a

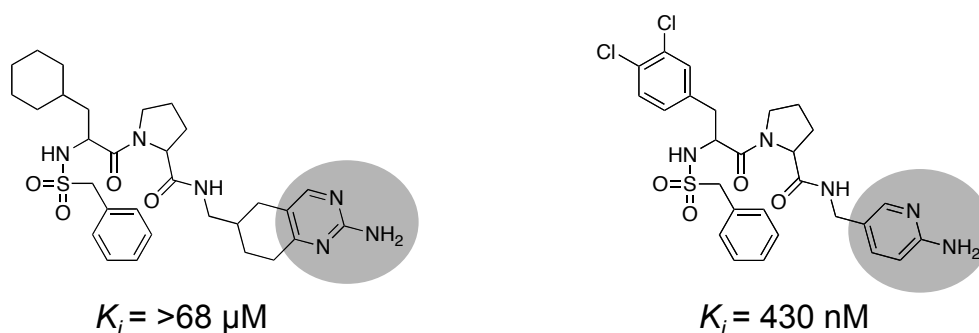
promising candidate structure due to its arginine-like sidechain in combination with the distant aromatic system. In addition, it must be stated that the design of isofunctional but structurally different molecules is one of the goals of *de novo* design. Compound **4** is promising with respect to its potential to fulfill this demand.

Bioisosteric replacement¹³⁵ of functional groups is a key to successful *de novo* design. In order to test DOGS for its ability to perform bioisosteric replacement, the list of 1,767 potential trypsin ligands designed by the software (resulting from ten runs based on five trypsin references) was ranked according to the scores assigned by DOGS. The top 200 molecules were analyzed for functional groups that replace the sidechains of reference compounds addressing the S1 pocket (guanidinium and benzamidine, Scheme 15).



Scheme 15. Sidechains addressing the S1 pocket found in the five reference compounds (*left*) and surrogates suggested by DOGS found in top-scored 200 designs (*right*).

Starting at rank position 78 (compounds on higher ranks exhibit one of the fragments present in the references), DOGS suggested eleven different sidechains replacing the reference fragments in the top 200 designs. Most of them offer the possibility to interact with the negatively charged aspartate sidechain of the S1 binding pocket of trypsin by a positively ionizable nitrogen atom. The terminal urea group and the two aromatic fragments (pyrimidin-2-amine and pyridin-2-amine) represent exceptions, where the nitrogen will likely not carry a positive charge. The formation of this salt bridge is a known key interaction inside the S1 pocket.¹⁵⁵ Albeit the formation of the salt bridge is unlikely for these three fragments, they are still able to form a hydrogen bond to the aspartate sidechain. In fact, both pyrimidin-2-amine and pyridin-2-amine can be found in known trypsin inhibitors as S1 addressing sidechains (Scheme 16).



Scheme 16. Known inhibitors of trypsin showing pyrimidin-2-amine¹⁵⁷ (*left*) and the pyridin-2-amine¹⁵⁸ (*right*) sidechains (grey circles). These moieties were also present in DOGS designs suggested as bioisosters for sidechains of the reference ligands addressing the S1 pocket of trypsin.

In addition, the list of proposed sidechains contains an alkyl chain carrying a terminal nitrogen. This fragment resembles the sidechain of lysine – one of the ‘natural’ ligands filling the S1 pocket during peptide bond cleavage.

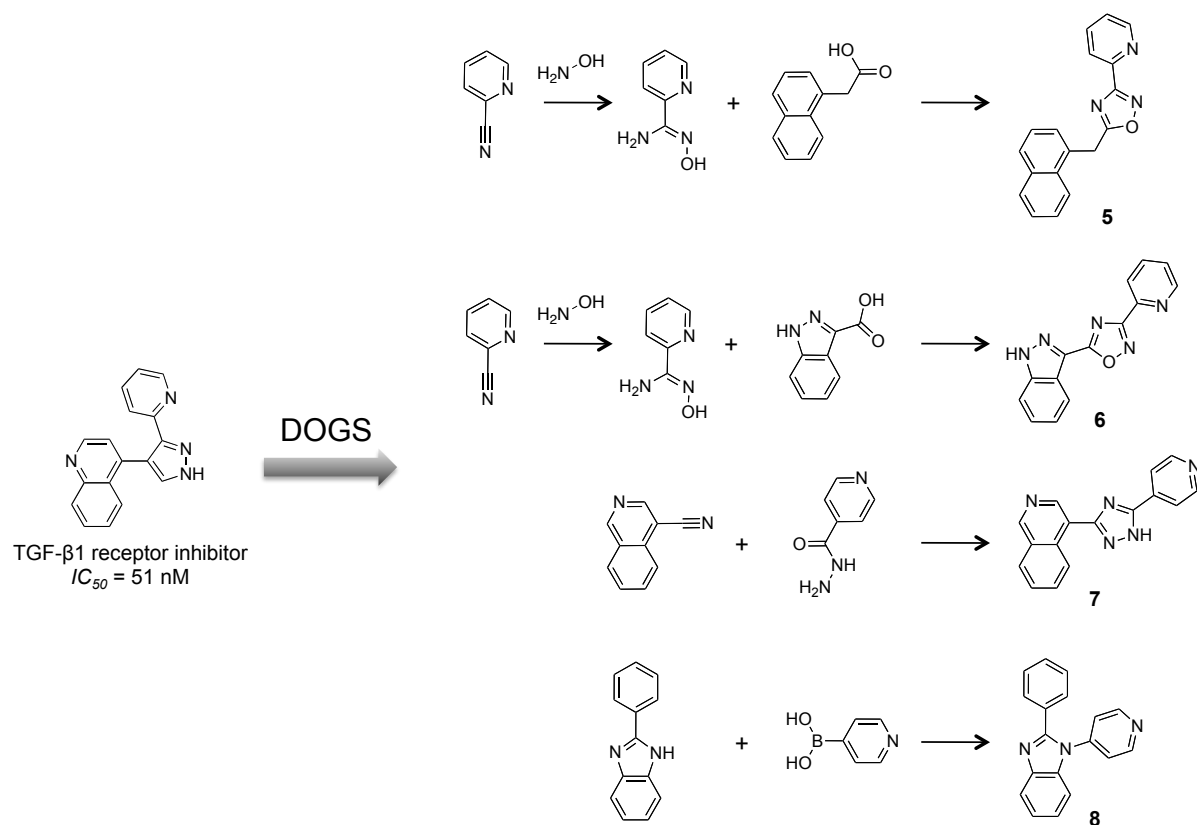
In summary, DOGS was able to suggest some reasonable potential bioisosters for substructures of reference ligands addressing the S1 pocket of trypsin including experimentally validates examples.

3.3.2 Transforming Growth Factor β 1 Receptor

The transforming growth factor (TGF) β 1 receptor is a transmembrane protein involved in the transduction of extracellular signals into the cell.¹⁴³ An intracellular kinase domain is activated upon extracellular binding of the cytokine TGF- β 1.¹⁴³ The receptor is involved in a number of processes like cell differentiation, growth and embryonic development. For this reason, it may play a role in a number of diseases including cancer and wound healing.¹⁴³

The reference ligand already introduced in Scheme 12 (*top*) served as a seed for two runs of DOGS to suggest potential ligands of the human TGF- β 1 receptor kinase domain (run 1: reduced graph, $\alpha = 0.4$; run 2: molecular graph, $\alpha = 0.875$). A selection of designed molecules is presented in Scheme 17.

The overall arrangement of aromatic systems of the reference is kept in the designed molecules, while each structure exhibits a modification compared to the reference. Except for one example, the central ring system is a product of the synthesis pathway. Synthesis routes comprise only one or two steps and can be deemed traceable.



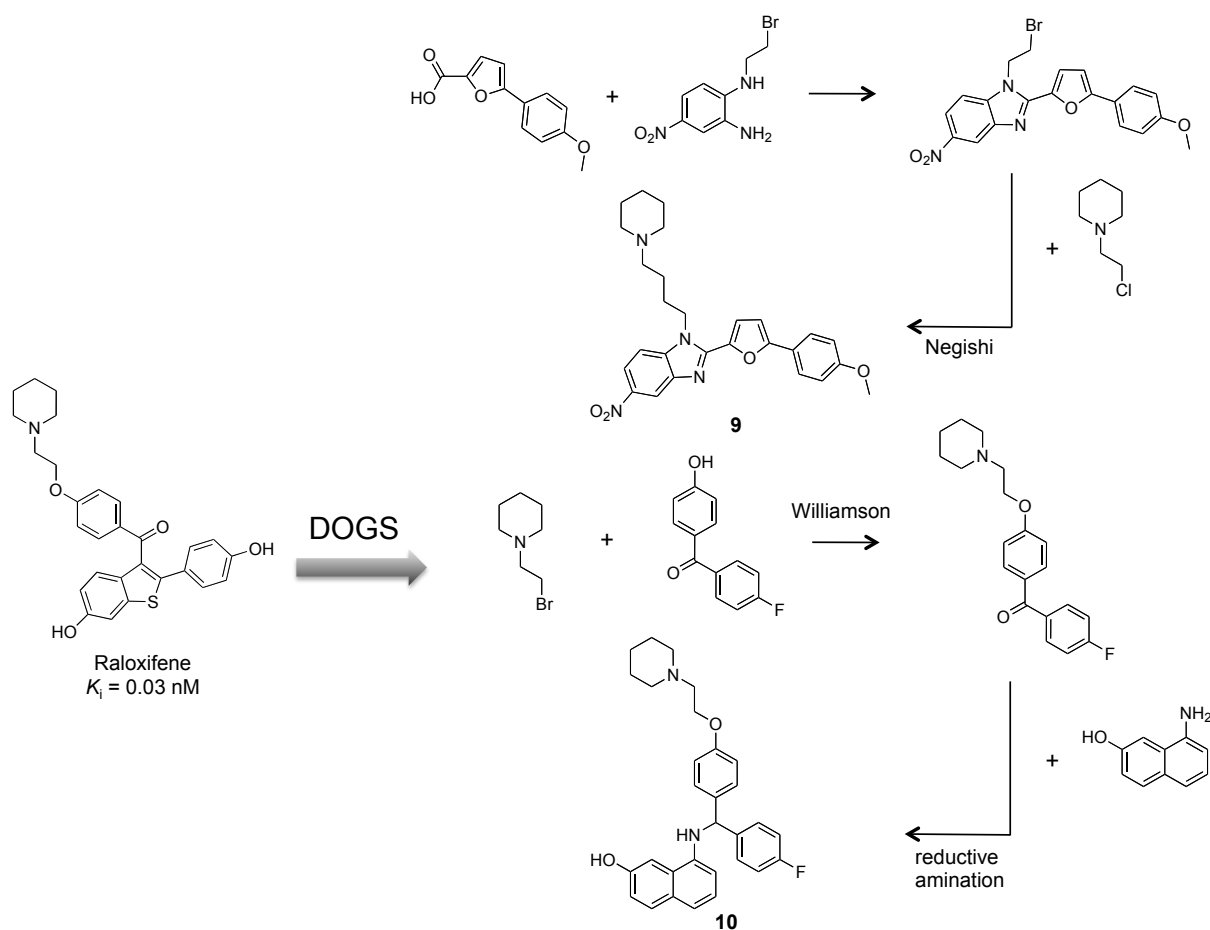
Scheme 17. Molecules **5-8** were proposed by DOGS together with presented synthesis plans based on an inhibitor of the human TGF- β 1 receptor.¹⁴³

3.3.3 Estrogen Receptor

Raloxifene is a potent modulator of the human estrogen receptor (ER).¹⁵⁹ It is approved as a drug for the treatment of osteoporosis in postmenopausal women.¹⁶⁰ DOGS designed two lists of ligand candidates for the human ER based on Raloxifene as a reference (run 1: reduced graph, $\alpha = 0.4$; run 2: molecular graph, $\alpha = 0.8$). Two exemplary structures from these lists are shown in Scheme 18.

As in the former example of TGF- β 1 receptor ligand design, DOGS was able to suggest molecules exhibiting distinct similarity to the reference in the overall composition of structural elements. The number of rings as well as their topological arrangement in designed molecules is comparable to Raloxifene. This is especially the case for compound **10**. Molecule **9** introduces a shift in the localization of an aromatic system (furan) to form a spacer between the benzimidazole and a phenyl ring. Altogether, **10** exhibits a higher structural similarity to the reference ligand than **9**. The sidechain carrying a terminal piperidine is almost identical to the one of Raloxifene (only a carbonyl group is missing), while the linker is completely replaced with an alkyl chain in **9**. This might cause a loss of

potential interactions with the receptor in case atoms of this linker form interactions. While the exchange of a hydroxy group against a methoxy substituent in **9** retains the property of a hydrogen bond acceptor, the second hydroxy group is replaced by a nitro group. Effects of this exchange heavily depend on the kind and energetic contribution of the interaction formed between the replaced hydroxy group and the receptor. Compound **10** replaces a hydroxy group of the reference with a fluorine atom. Fluorine has been reported to act as a hydrogen bond acceptor in some cases, albeit weaker than an oxygen of a hydroxy group.¹⁶¹ Effects of these modifications on the biological activity have to be elucidated by practical synthesis and testing. The synthesis pathway of **10** seems feasible and simple. The ring closing reaction of compound **9** might be difficult because of the highly substituted educts.



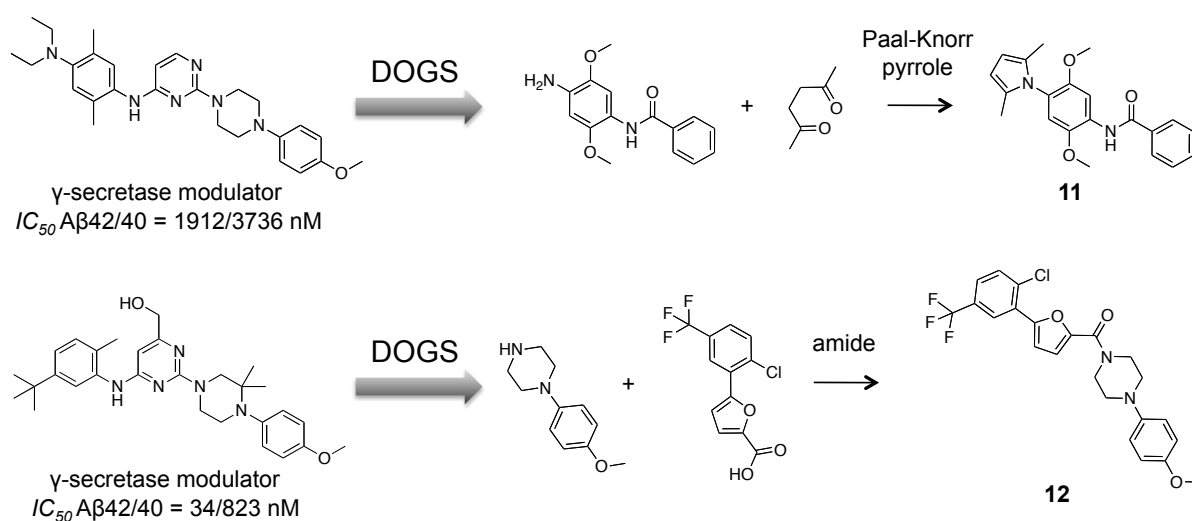
Scheme 18. Molecules **9** and **10** were proposed by DOGS together with presented synthesis plans based on Raloxifene, a modulator of the human estrogen receptor.¹⁵⁹ Where available, general names of reactions are given next to reaction arrows.

3.4 Practical Evaluation of the Software

3.4.1 Human γ -secretase

DOGS was employed to propose candidate structures as new modulators of the human γ -secretase. This target is responsible for the production of potentially toxic amyloid- β (A β) 42 peptides.¹⁶² Extracellular accumulation and formation of amyloid plaques is the primary pathological event in Alzheimer's disease.¹⁶³ Oligomerisation of A β 42 peptides is believed to be a pivotal step in plaque formation.¹⁶² Modulators of the γ -secretase are supposed to shift the product ratio of amyloid precursor protein processing towards shorter, non-toxic A β fragments like A β 38 or A β 40.¹⁶²

Four different reference ligands known to modulate γ -secretase were selected. For each reference compound, two DOGS runs (molecular graph representation, $\alpha = 0.875$; reduced graph representation, $\alpha = 0.4$) were performed. Each of the eight resulting lists of DOGS designs was re-scored after the run by a CATS¹²² similarity analysis (Euclidian distance in the space spanned by the descriptor). Compounds of each list as well as the corresponding reference ligand were encoded by the CATS descriptor and subsequently ranked according to their distance to the reference in order to get an additional criterion for prioritization. Re-ranked lists were visually inspected and two promising ligand candidates **11** and **12** were selected for synthesis (Scheme 19). Criteria for compound selection were (in their order of importance) (i) the subjective rating of the molecular structure by a medicinal chemist, (ii) ease and plausibility of proposed synthesis route, and (iii) CATS as well as ISOAK scores.



Scheme 19. Candidate structures **11** and **12** proposed by DOGS as potential modulators of the human γ -secretase. Synthesis plans were suggested by the software and successfully pursued. Molecules **11** and **12** originate from distinct runs based on different reference ligands.¹⁶³ IC_{50} values are determined by two separate dose response experiments. Concentrations of secreted amyloid peptides are detected separately in cell supernatants by labeled antibodies and electrochemiluminescence.

Synthesis plans were readily traceable as suggested by the software. One-step reactions yielded the products in both cases. DOGS was able to design compounds not only deemed promising by medicinal chemists, but also proved to be synthesizable as suggested.

Synthesized compounds were tested for their ability to modulate the human γ -secretase by measuring the concentrations of amyloid peptides A β 38, A β 40, and A β 42 in cell supernatants. Cell lines overexpressing human γ -secretase and the amyloid precursor protein are treated with the compound. Labeled antibodies specific for each of the three peptides are used to determine their levels of concentration in the cell supernatant in a liquid phase electrochemiluminescence assay.¹⁶⁴ First results report modulation of γ -secretase activity (Figure 22). Both compounds shift the product ratio towards higher levels of A β 42. Although this is not the effect intended for a potential treatment of Alzheimer's disease, this first practical evaluation of DOGS can be deemed successful. For both selected compounds the suggested synthesis plan was readily pursuable and a modulation of target activity could be observed. These ligands can serve as starting points for an optimization of the pharmacological profile by structural modification.

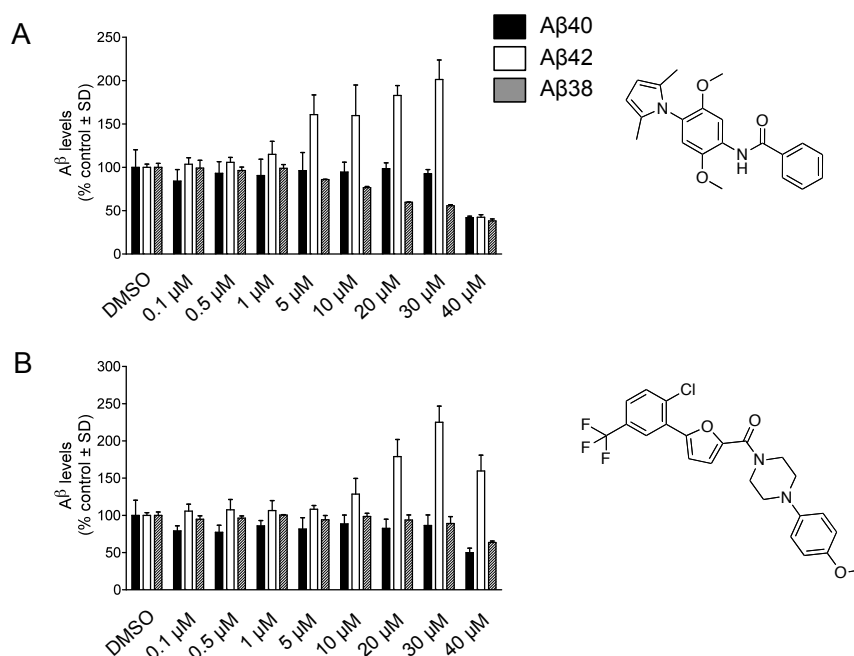
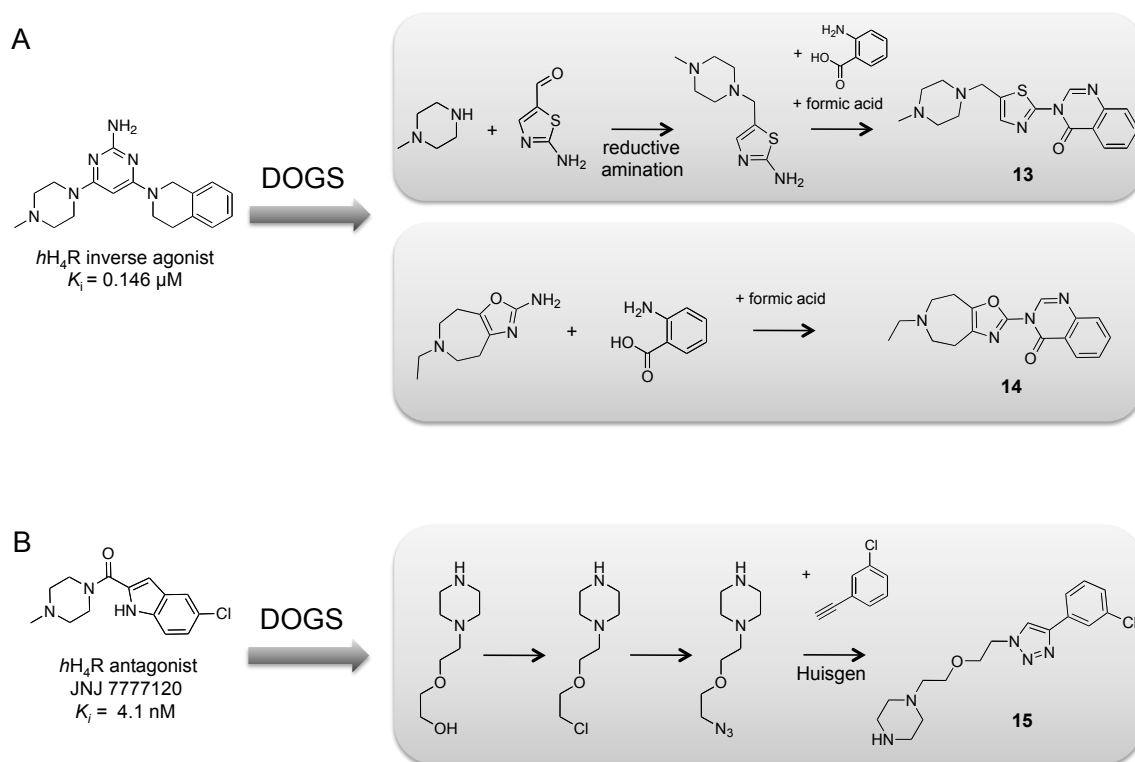


Figure 22. Modulation of γ -secretase activity by designed ligands. Both compounds modulate the activity of γ -secretase by a shift of product ratio towards higher levels of A β 42.

3.4.2 Human Histamine H₄-Receptor

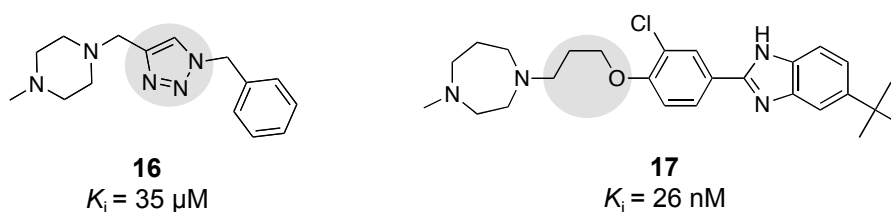
Histamine is a biogenic amine involved in a plethora of signaling pathways as a messenger. Four subtypes of histamine receptors (*hH*₁R – *hH*₄R) are known in human. All subtypes belong to class A (rhodopsin-like) of the GPCR super family.¹⁴² Some antagonists of *hH*₁R and *hH*₂R are approved drugs for the treatment of allergic reactions and ulcer. Clinical trials of *hH*₃R antagonists for the therapy of neuronal diseases like epilepsy, schizophrenia and sleep/wake disorder are currently in progress.¹⁶⁵ Although subtypes 3 and 4 show the highest intra-familial similarity (37% sequence identity), selective *hH*₄R antagonists have been identified. Preclinical trials reveal their potential therapeutic application in allergy, inflammation, autoimmune disorders and cancer.¹⁶⁵

DOGS was applied to give ideas for new selective antagonists or inverse agonists of *hH*₄R. For this purpose, two reference ligands (an inverse agonist and an antagonist) were employed as seed structures (Scheme 20). For each reference, the molecular graph representation ($\alpha = 0.875$) as well as the reduced graph representation ($\alpha = 0.4$) was applied, resulting in four runs. Visual inspection of result lists together with medicinal chemists familiar with the target led to a prioritization of compounds. Three examples of top rated designs are presented in Scheme 20.



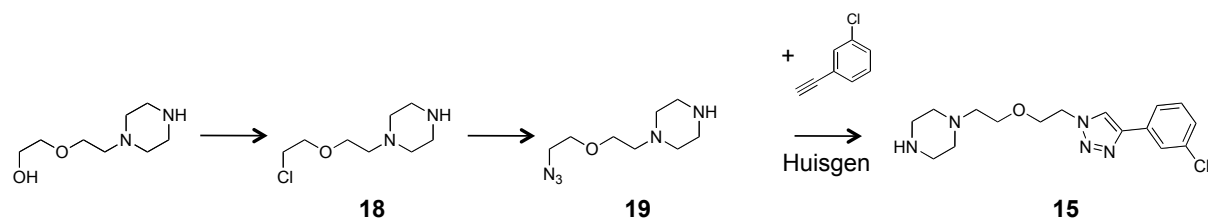
Scheme 20. Molecules **13** and **14** were proposed by DOGS based on an inverse agonist of *hH*₄R¹⁴² (A). Compound **15** is a design originating from the antagonist JNJ 7777120 of the same target¹⁴⁴ (B). General names of reactions are provided below reaction arrows if available.

N-methylpiperazine is present in both references and represents a chemical moiety that is often used as a basic head group in H₄ receptor ligands.¹⁶⁶ The positive charge of basic amines is believed to form a key interaction to a negatively charged sidechain of the protein.¹⁶⁷ While in compound **13** the *N*-methylpiperazine moiety is preserved, it is replaced in **14** and **15** by isofunctional groups. Both of them represent aliphatic rings exhibiting basic amines, which provide the chance to undergo the charge-mediated interaction with the receptor. Localization of aromatic ring systems of reference compounds is also approximately kept within the proposed structures. Compound **15** is of special interest because it combines two structural elements that can be found in reported H₄R ligands: an alkylic linker chain with an ether bridge connecting and a central triazole ring (Scheme 21). Notably, both structural elements are not present in the reference compound. The moderate binding energy of the triazole-carrying ligand **16** ($K_i = 35 \mu\text{M}$) may be caused by a missing hydrogen bond acceptor in the central part. This pharmacophoric feature is also believed to play a role in the interaction with binding pocket of H₄R.¹⁶⁷ The oxygen atom of the ether bridge present in designed compound **15** and H₄R ligand **17** is able to act as a hydrogen bond acceptor. The ISOAK scoring function of DOGS assigns this oxygen to the carbonyl oxygen of the reference, which can act as a hydrogen bond acceptor as well.



Scheme 21. Highlighted features of two *h*H₄R ligands (compound **16**¹⁴²: central triazole ring; compound **17**¹⁶⁸: alkyl linker chain with ether bridge) are combined in designed compound **15**. None of these features is present in the reference ligand.

In order to test for the hypothesis that the combination of features found in compound **15** might lead to affinity to *h*H₄R, compound **15** was selected for synthesis and testing. The synthetic procedure was realized exactly as suggested by the software (Scheme 22). Analytical spectra of intermediate products **18** and **19** as well as of compound **15** can be found in the supplement.



Scheme 22. Synthesis of compound **15** as proposed by the software.

Compound 18 (1-(2-(2-chloroethoxy)ethyl)piperazine). Educt 2-(2-(piperazin-1-yl)ethoxy)ethanol (1 eq.) was precipitated with 5N isopropyl HCl (3 eq.). The salt was filtered off and dried. In order to substitute the hydroxy group with chloride, the salt (1 eq.) was dissolved in toluene, and thionyl chloride (3 eq.) was added slowly under cooling conditions (ice bath). After heating to 70°C for 10 minutes, the mixture was stirred for 3h at 60°C under argon atmosphere. The formed precipitate was filtered off and dried *in vacuo* to yield a yellowish-white solid. MS (ESI⁺): $m/z = 192.91$ [M+H]⁺. ¹H NMR (MeOD, 400.13 MHz): δ 3.57 (t, 2H, $J = 4.9$ Hz), 3.63 (m, 8H), 3.77 (t, 2H, $J = 5.6$ Hz), 3.85 (t, 2H, $J = 5.7$ Hz), 3.97 (t, 2H, $J = 4.9$ Hz).

Compound 19 (1-(2-(2-azidoethoxy)ethyl)piperazine). Compound **18** (1 eq.) and sodium azide (2 eq.) were dissolved in DMSO. The mixture was stirred for 42h at 100°C. The precipitated white solid was removed by filtration. The orange filtrate was diluted with dichloromethane and extracted with 2N NaOH (three times). After removal of the solvent, the brown product (oil) was dried *in vacuo*. MS (ESI⁺): $m/z = 199.93$ [M + H]⁺. ¹H NMR (DMSO-*d*₆, 400.13 MHz): δ 2.34 (t, 4H, $J = 3.9$ Hz), 2.44 (t, 2H, $J = 5.9$ Hz), 2.69 (t, 4H, $J = 4.7$ Hz), 3.38 (t, 2H, $J = 4.9$ Hz), 3.54 (t, 2H, $J = 5.9$ Hz), 3.58 (t, 2H, $J = 4.9$ Hz).

Compound 15 (1-(2-(2-(4-(3-chlorophenyl)-1H-1,2,3-triazol-1-yl)ethoxy)ethyl)piperazine). Compound **19** (1 eq.) and 1-chloro-3-ethynylbenzene (1eq.) were dissolved in a mixture of water and isopropyl alcohol (1:1) and placed in a 5ml microwave vial. Copper(I)-iodide (0.1 eq) was added and the mixture was heated in a microwave oven (Biotage Initiator, 100W, 125°C, 20min, absorption level: high). The mixture was extracted three times with dichloromethane and 2N NaOH. After removal of the solvent, the remaining oil was purified by flash column chromatography (Biotage Isolera One) to yield a light brown oil. MS (ESI⁺): $m/z = 335.82$ [M + H]⁺. ¹H NMR (DMSO-*d*₆, 400.13 MHz): δ 2.25 (t, 4H, $J = 3.9$ Hz), 2.39 (t, 2H, $J = 5.7$ Hz), 2.60 (t, 4H, $J = 5$ Hz), 3.52

(t, 2H, $J = 5.7$ Hz), 3.84 (t, 2H, $J = 5.2$ Hz), 4.57 (t, 2H, $J = 5.2$ Hz), 7.40 (ddd, 1H, $J_1 = 1.1$ Hz, $J_2 = 2.2$ Hz, $J_3 = 8.0$ Hz), 7.49 (t, 1H, $J = 7.9$ Hz), 7.83 (dt, 1H, $J_1 = 1.3$ Hz, $J_2 = 7.9$ Hz), 7.9 (t, 1H, $J = 1.8$ Hz), 8.65 (s, 1H). ^{13}C NMR (DMSO- d_6 , 400.13 MHz): δ 48.22 (2C), 49.46 (2C), 54.74 (1C), 64.42 (1C), 68.57 (2C), 122.72 (1C), 123.69 (1C), 124.67 (1C), 127.58 (1C), 130.86 (1C), 132.87 (1C), 133.68 (1C), 144.93 (1C). HRMS (ESI $^+$): m/z [M + H] $^+$ calculated for C $_{16}$ H $_{23}$ ClN $_5$ O: 336.1586; found: 336.1586. HPLC-MS (MeOH/H $_2$ O): purity 99.68%.

Binding affinity of compound **15** was determined in a competitive binding assay by measuring displacement of radioactive labeled [^3H]histamine bound to H $_4$ R.¹⁶⁹ Membrane preparations of insect Sf9 cells expressing *hH₄R* together with G-protein subunits G α i2 and G β γ $_2$ were performed to yield the protein. A similar assay was used to measure the activity on *hH₃R* (reference ligand: [^3H]N $^{\alpha}$ -methylhistamine).

Compound **15** exhibits only weak affinity to *hH₄R*. From three measurements, a mean K_i of 436 μM (STD: ± 137 μM) was determined. Comparable results were found for the activity of **15** on the H $_3$ receptor ($K_i = 466$ μM (± 209 μM), averaged over four distinct tests).

Although the flexible alignment of compound **15** and the reference ligand does not directly align the central hydrogen bond acceptors, they might still be able to undergo an interaction with the same hydrogen bond donor of the receptor binding site according to the alignment (Figure 23).

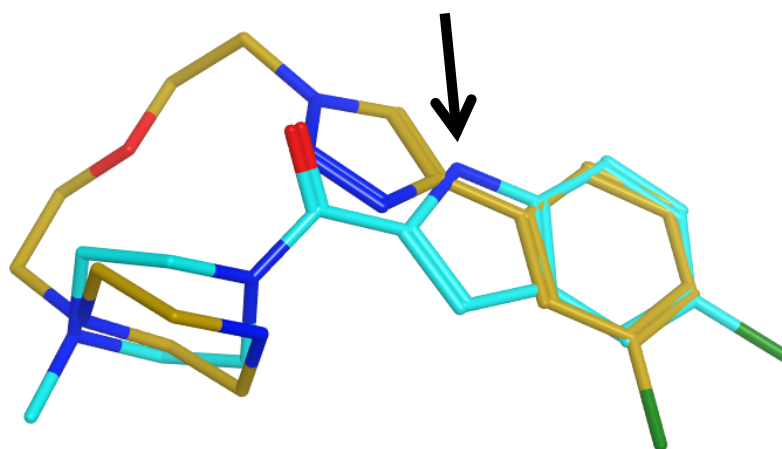


Figure 23. Flexible alignment of compound **15** (gold) and the reference JNJ 777120 (light blue) computed by a component of the software suite MOE. Low activity of compound **15** may be caused by a missing hydrogen bond donor in the central part, which is present in the reference ligand (arrow).

The alignment also suggests a possible reason for the weak activity of compound **15**: JNJ 7777120 exhibits a hydrogen bond donor (nitrogen of the indole scaffold) in the center of the molecule. Compound **15** does not feature an equivalent atom capable to act as a hydrogen bond donor. It has been suggested that this pharmacophoric feature might be important for high affinities to *h*H₄R.¹⁶⁷ Introduction of a hydrogen bond donor in the central part of **15** can therefore be a possible strategy for structural optimization. Another possibility to improve the potency of **15** by a comparably small structural modification could be to replace the piperazine moiety of **15** with *N*-methylpiperazine found in many H₄ reference compounds. It is known that even small changes of the *N*-methylpiperazine group can lead to a considerable decrease of affinity of H₄ receptor ligands.¹⁶⁶

4 Conclusions and Outlook

This work presents a new method for automated *de novo* design of ligand candidates. The program DOGS was evaluated in both theoretical and practical case studies.

The main advantage of DOGS over most of the other existing software tools for *de novo* design is its ability to suggest not only new compounds but also reasonable pathways for their synthesis. The set of reactions as well as the library of synthesis building blocks can be modified by the user to provide flexibility with respect to preferred chemistry and available educts.

Synthesizability of proposed compounds is pivotal for the practical evaluation of *de novo* design tools. It can be argued that lacking ease of synthesis of designed molecules is one of the reasons why only around a dozen of all published algorithms for molecular *de novo* design have been subjected to practical evaluation. The aim of this study was to show that an enhanced build-up strategy employing known chemical reactions can facilitate synthesizability of designed molecules, and hence practical evaluation of *de novo* design software. DOGS was evaluated practically in two realistic case studies on the design of potential ligands for human γ -secretase and human histamine H₄ receptor. In both cases, compounds selected for testing were readily synthesizable. In addition, synthesis routes could be followed up at the bench exactly as suggested by the software. It is clear that this is to some extent a consequence of the selection process by human experts. One criterion for compound selection was a synthesis route deemed simple and pursuable. It cannot be expected that all synthesis pathways suggested by DOGS will directly work when pursued in the lab. Nevertheless, a large part of designed compounds received high synthesizability scores calculated by the software MOE, which likely is a consequence of the reaction-driven compound construction concept of DOGS.

Synthesized candidate ligands for γ -secretase exhibit biological activity. Although target modulation points in the opposite direction as intended, these molecules show affinity to the target and provide a valuable basis for further investigation. The molecule selected and synthesized as a potential ligand for the H₄ receptor only exhibits weak activity, probably due to a missing pharmacophoric feature in the center of the molecule.

Besides these practical evaluations, theoretical investigation of DOGS results was performed and revealed that the software is capable of capturing a calculated biophysical property ($\log P(o/w)$) of reference compounds and reflecting it in the constructed molecules. Notably,

this property is not explicitly accounted for in similarity assessment by the scoring function. A large fraction of molecules designed by DOGS are evaluated to be druglike and mostly violate less than two of Lipinski's 'rule of five'. Again, designed compounds generally reflect the reference compounds in these properties. Features of proposed candidate structures depend on characteristics provided by the references. This behavior is intended for *de novo* design, since property profiles of molecules are thought to be linked to their biological activities and pharmacokinetic characteristic.⁸⁴ DOGS designs originating from druglike reference compounds have an enhanced chance to be druglike themselves.

DOGS introduces graph kernel methods for scoring to the field of *de novo* design. The employed ISOAK method has proven to be well-suited for the DOGS approach. The reason is likely to be found in the concept of ISOAK to compare molecules: it assigns each atom of the smaller molecules to one atom of the larger compound. This renders it possible to compare molecules significantly differing in size, which is a requirement for the stepwise build-up process of the algorithm. Potentially small intermediate products are also scored against the complete reference ligand. In addition, the kernel is not restricted to molecular graphs. This allowed for the implementation of a reduced graph representation for molecules, which extends the pool of meaningful results produced by DOGS. As a consequence of the ligand-based scoring approach, the software can be run with a minimum of available knowledge about the target molecule. A single reference compound known to be active on the target is sufficient to let the software create ideas for new ligands. This feature is of special merit for drug development campaigns for structurally unexplored targets.

For future work, a reduction of the number of duplicate molecules in output lists presents a way to speed-up calculations and save computational power. Duplicates with nearly exact synthesis pathways represent redundant information and could be avoided without loss of important information. A possible remedy to this problem would be to store a list of all dummy products already selected to be pursued in the subsequent reaction step in former synthesis pathways. Whenever the current virtual synthesis chooses a dummy product that has already been selected before, the construction process could be stopped. Because of the deterministic characteristic of the algorithm all subsequent steps would exactly be the same as already calculated. One could either delete the current synthesis and proceed with the next construction pathway or complement it with the remaining part of the pathway already calculated before and store it. The comparison of dummy fragments could be efficiently implemented to work on a prefix tree of canonical SMILES¹⁷⁰ representations.

In order to broaden the scope of DOGS and increase flexibility with respect to implementable reactions, the algorithm could be extended to process reactions with more than two educts. However, one has to keep in mind that this will likely result in increased run times because of the higher number of possible educt combinations.

Detection of isofunctional reactive groups in educts in order to avoid unwanted side reactions would be of great potential for improvements of the method. For example, the presence of an additional secondary amine in one of the reaction partners of a reductive amination between a carbonyl group and a primary amine is prone to lead to unwanted byproducts. Recognition of competing reactive substructures could be used to either flag the corresponding intermediate product with a warning mark and suggest protection groups or to prohibit the reaction. This approach depends on a careful definition of competing reactive substructures for every reaction. Although this step only has to be performed once per reaction during its implementation, it represents a significant effort and is prone to over-regimentation. A more elaborate way to deal with competing reactions would be to computationally estimate substructure reactivity. It is far too demanding to calculate reactions *ab initio* by quantum mechanical methods. However, quantum mechanical calculations could be used to prioritize multiple occurrences of the same reactive group with respect to their reactivity. In fact, this has already been introduced to *de novo* design in the software tool SYNOPSIS⁵⁷ for a special type of reaction. Notably, the authors concluded that the additional computational costs are not justified by the improvements by this approach and excluded it from the final version of the software. However, it could be shown that this method works in principle. It will probably become of higher practical relevance with increasing computational power in the future.

Besides the potential advantages of ligand-based *de novo* design it must be stated that available structural information about the target binding site can be of merit and should be incorporated in the design of potential ligands. For this reason, an extension of the DOGS approach towards receptor-based scoring is expected to be beneficial for targets where structural data exist. The success of this effort depends on finding a receptor-based scoring function capable of preferring small but promising intermediate products over larger ones having less potential to be extended to favorable final solutions. Scoring functions of docking tools have the tendency to favor larger ligands exhibiting more atoms to interact with the receptor binding site.¹⁷¹ Normalization of docking scores by the number of heavy atoms could offer a simple and self-evident solution. Another critical point of 3D scoring functions is their higher computational demand compared to 2D techniques. During a DOGS run a (potentially)

large number of scores needs to be calculated due to the enumeration of complete subspaces in each reaction step. A less demanding scoring function could be employed as a filter to narrow down the number of molecules subsequently scored by the more expensive 3D scoring function to those deemed favorable.

Generating innovative and patentable molecules with biological activity from scratch is an appealing, yet demanding goal. Current software solutions to this problem are far away from being ‘click-and-harvest’ applications that can guarantee to produce readily exploitable results. *De novo* design is still dependent on intervention and support of human expertise. Nevertheless, it can be a valuable source of inspiration and new ideas for drug development projects. In fact, reports about successful application of respective software tools make it safe to say that computational *de novo* design works.^{81,104,114} Incorporating synthesis pathways – as presented in this work – can focus *de novo* design on a more practical standard and adds an important level of information to the output.

The main reason why *de novo* design has not yet grown out of its role as a pure idea generator is our lack of a deeper understanding about interactions taking place between receptors and their ligands upon binding, which is expressed in insufficient scoring functions for molecular docking. Especially entropic contributions to binding energies and solvent effects are still widely ignored. The same holds true for our understanding about chemical similarity between small organic molecules. The special problem of the similarity concept in molecular design is that in reality similarity is ‘measured’ by a binding cavity, which is different for every target. Presence or absence of a structural feature in a ligand might be tolerated by one binding site, but leads to significant changes of binding affinity in the next case. A molecular feature that is important for one target may be less critical in the context of another one. Target dependency of the similarity concept makes it difficult to extract general rules applicable over a wide range of different target molecules.

Broader application of *de novo* design methods has also been hampered for a long time by a lack of accessible and user-friendly software implementations. In most cases, published approaches remain in-house solutions or even never leave proof-of-concept status. The situation has started to change only recently, as today almost every large software suite for molecular modeling offers a *de novo* design module. This can be interpreted as a consequence of a growing interest in this approach.

Regardless of these shortcomings, computer-assisted *de novo* drug design has become an established instrument in the pharmaceutical industry as well as in academia, and will

continue to give valuable impulses to drug design as a complementary tool to other computational approaches.

Summary

A new method for computer-based *de novo* design of drug candidate structures is proposed. DOGS (Design of Genuine Structures) features a ligand-based strategy to suggest new molecular structures. The quality of designed compounds is assessed by a graph kernel method measuring the distance of designed molecules to a known reference ligand. Two graph representations of molecules (molecular graph and reduced graph) are implemented to feature different levels of abstraction from the molecular structure. A fully deterministic construction procedure explicitly designed to facilitate synthesizability of proposed structures is realized: DOGS uses readily available synthesis building blocks and established reaction schemes to assemble new molecules. This approach enables the software to propose not only the final compounds but also to give suggestions for synthesis routes to generate them at the bench. The set of synthesis schemes comprises about 83 chemical reactions. Special focus was put on ring closure reactions forming drug-like substructures. The library of building blocks consists of ~25,000 molecules readily available from a commercial vendor with a molecular mass between 30 and 300 Da.

DOGS builds up new structures in a stepwise process. Each virtual synthesis step adds a fragment to the growing molecule until a stop criterion (molecular mass or number of synthesis steps) is fulfilled.

In a theoretical evaluation, a set of ~1,800 molecules proposed by DOGS is analyzed for critical properties of *de novo* designed compounds. The software is able to suggest drug-like molecules (79% violate less than two of Lipinski's 'rule of five'). In addition, a trained classifier for drug-likeness assigns a score >0.8 to 51% of the designed molecules (with 1.0 being the top score). In addition, most of the DOGS molecules are deemed to be highly synthesizable by a retrosynthesis descriptor (77% of molecules score in the top 10% of the descriptor's value range). Calculated $\log P(o/w)$ values of constructed molecules resemble a unimodal distribution centered close to the mean of $\log P(o/w)$ values calculated for the reference compounds.

A structural analysis of selected designs reveals that DOGS is capable of constructing molecules reflecting the overall topological arrangement of pharmacophoric features found in the reference ligands. At the same time, the DOGS designs represent innovative compounds being structurally distinct from the references. Synthesis routes for these examples are short

and seem feasible in most cases. Some reaction steps might need modification by using protecting groups to avoid unwanted side reactions.

Plausible bioisosters for known privileged fragments addressing the S1 pocket of trypsin were proposed by DOGS in a case study. Some of them can be found in known trypsin inhibitors as S1-addressing side chains.

The software was also tested practically in two realistic drug design scenarios. DOGS was applied to design ligands for human γ -secretase and human histamine receptor subtype 4 (*hH₄R*). Two selected designs for γ -secretase were readily synthesizable as suggested by the software in one-step reactions. Both compounds modulate the activity of the target molecule, although the effect differs from the one suggested as a potential treatment Alzheimer patients. These structures can serve as starting points for structural optimization. In a second case study, a ligand candidate selected for *hH₄R* could again be synthesized exactly following the three-step synthesis plan suggested by DOGS. This compound showed only low activity on the target structure. Nevertheless, these examples represent promising initial results. The concept of DOGS could prove to deliver not only synthesizable compounds but also pursuable synthesis plans. Future practical applications of the software will help to gain a more comprehensive impression of the method's power to contribute to the development of bioactive compounds.

Zusammenfassung

Das Ziel des computergestützten *de novo* Designs ist der Neuentwurf biologisch aktiver Verbindungen, welche als Vorläufer für mögliche Wirkstoffe dienen können. Die zentrale Idee ist es, molekulare Fragmente neu zusammen zu setzen, um maßgeschneiderte Modulatoren für ein gegebenes Zielmolekül (zumeist ein Protein) zu erhalten. Der Fokus von *de novo* Design liegt dabei auf der Innovation und Neuartigkeit der entworfenen Substanzen. Dies unterscheidet die Methode grundsätzlich vom virtuellen Screening (VS), bei dem Sammlungen bereits existierender und beziehbarer Moleküle nach potentiellen Wirkstoffkandidaten durchsucht werden.

Die ersten veröffentlichten *de novo* Design Ansätze konzentrierten sich darauf, neue potentielle Liganden direkt in der Bindetasche zu konstruieren. Dies geschieht unter der Maximierung von sterischer Paßform und unter Berücksichtigung von polaren und elektrostatischen Interaktionsmöglichkeiten mit der Rezeptortasche. Dieser *rezeptorbasierte* Ansatz wurden bald durch *ligandenbasierte* Methoden ergänzt. Hierbei zielt der Entwurf neuer Moleküle auf möglichst hohe Ähnlichkeit zu bereits bekannten Liganden des Zielmoleküls ab. Nach dem zentralen „Ähnlichkeitsprinzip“ sollen ähnliche Moleküle vergleichbare Eigenschaften aufweisen. Der Vorteil von ligandenbasierten Methoden im Vergleich zu rezeptorbasierten Ansätzen liegt darin, dass sie keine Kenntnis über die räumliche Struktur der Bindetasche voraussetzen, welche für eine Vielzahl pharmazeutisch relevanter Zielmoleküle tatsächlich nicht bekannt ist.

Seit den Anfängen des computergestützten *de novo* Design kranken die Methoden daran, dass entworfene Moleküle zwar als potentiell interessant eingestuft werden, aber oft nur schlecht synthetisch zugänglich oder sogar chemisch instabil sind. Neusynthesen sind im Allgemeinen deutlich teurer und aufwendiger als der Bezug fertiger Substanzen von kommerziellen Anbietern oder aus dem eigenen Bestand. Aus diesem Grund sind bisher nur vergleichsweise wenige der beschriebenen Algorithmen zum *de novo* Design überhaupt einer praktischen Evaluation unterzogen worden, in der vorgeschlagene Moleküle synthetisiert und getestet wurden. Trotzdem ist eine Reihe erfolgreicher Anwendungen von entsprechenden Programmen publiziert, und *de novo* Design kann als etablierte Methode angesehen werden.

Ziel dieser Arbeit ist die Entwicklung einer Methode zum computergestützten *de novo* Design (DOGS, Design Of Genuine Structures). Der Fokus von DOGS liegt darauf, Moleküle zu

entwerfen, die eine gute Zugänglichkeit durch chemische Synthese aufweisen. Um dies zu erreichen, greift die Software auf kommerziell verfügbare Synthesebausteine und etablierte chemische Reaktionen zum Aufbau neuer Moleküle zurück. Dies soll zum einen die Wahrscheinlichkeit der guten Synthetisierbarkeit der aufgebauten Moleküle erhöhen, zum anderen aber den Computer in die Lage versetzen, unmittelbar Vorschläge für eine Synthesestrategie zu generieren. Ziel ist die Erhöhung der Akzeptanz der Ergebnisse und die Erleichterung der praktischen Umsetzung.

Die Bibliothek der Synthesebausteine besteht aus etwa 25.000 physikalisch verfügbaren Moleküle mit einer molekularen Masse zwischen 30 und 300 Da. Eine Reihe von Filterkriterien wurde verwendet, um unerwünschter Verbindungen zu entfernen sowie Ladungs- und Protonierungszustände zu standardisieren. Zusätzlich wurde eine Sammlung von Präparations-Reaktionen angewendet, um weitere funktionelle Gruppen in die virtuellen Synthesebausteine einzuführen. Dies dient der Aktivierung von reaktiven Gruppen und damit ihrer späteren Umsetzung durch die Reaktionen zur Kopplung der Bausteine. Abschließend wurde jeder Baustein auf das Vorhandensein aller durch die Reaktionen festgelegten reaktiven Gruppen überprüft und die entsprechende Information zusammen mit dem Baustein in einer MySQL Datenbank gespeichert.

Die Reaktionssammlung umfaßt 83 Reaktionen und wurde durch eine Literaturrecherche zusammengestellt. Insbesondere wurden solche Reaktionen gewählt, die Substrukturen erzeugen, welche häufig in biologisch aktiven und wirkstoffartigen Molekülen vorkommen. Aus diesem Grund befindet sich ein großer Anteil Ringschlußreaktionen in der Reaktionsbibliothek. Weitere Kriterien zur Auswahl der Reaktionen umfaßten hohe beschriebene Ausbeuten, Vermeidung toxischer Reagenzien und Katalysatoren sowie einfache praktische Durchführbarkeit.

DOGS verwendet eine ligandenbasierte Strategie zur Bewertung der entworfenen Moleküle. Eine Kernfunktion vergleicht die erzeugten Moleküle mit einem Referenzliganden anhand ihrer Graphenrepräsentationen. Die berechnete Distanz zum Referenzliganden wird als Gütemaß verwendet. Im Rahmen der Arbeit kommen zwei verschiedene Graphenrepräsentationen zum Einsatz. Der *molekulare Graph* entspricht der topologischen Struktur einer zweidimensionalen Moleküldarstellung. Jedes Atom wird in einen Knoten und jede Bindung in eine Kante des Graphen übersetzt. Im Gegensatz dazu stellt der *reduzierte Graph* eine stärkere Abstraktion von der Molekülstruktur dar. Bestimmte Substrukturen bestehend aus mehreren Atomen (vor allem Ringsysteme, lipophile Bereiche) werden zu

einem einzelnen Knoten zusammengefaßt. Der reduzierte Graph stellt damit nur noch die topologische Anordnung bestimmter Substrukturen des Moleküls dar. Der Anwender legt fest, welche der beiden Moleküldarstellung in einem Konstruktionslauf Verwendung findet. Neue Moleküle werden von DOGS schrittweise aufgebaut, wobei pro Erweiterungsschritt je ein weiterer Baustein an das wachsende Molekül angefügt wird. Das Startfragment einer virtuellen Synthese wird aus allen Fragmenten gemäß seiner Güte ausgewählt. Dazu wird die gesamte Fragmentbibliothek zunächst wie beschrieben mit dem Referenzliganden verglichen. Die Synthesebausteine mit der höchsten Güte werden als Startfragmente verwendet. Ein Erweiterungsschritt besteht aus zwei Unterschritten. Zunächst wird bestimmt, welche der anwendbaren Reaktionen das größte Potential bietet. Dazu werden alle reaktiven Gruppen des zu erweiternden Zwischenprodukts mit passenden Reaktionen und *minimalen Dummy-Fragmenten* als Edukte abreagiert. Das Konzept der minimalen Dummy-Fragmente wird in DOGS eingeführt, um die mindestens zu erwartende strukturelle Veränderung abzuschätzen, die eine Reaktion verursacht. Die Dummy-Fragmente werden durch die Definition der Reaktion festgelegt und weisen ausschließlich jene strukturellen Elemente auf, die für die Durchführung der Reaktion unbedingt notwendig sind. Alle Dummy-Produkte, die aus diesen Pseudoreaktionsschritten hervorgehen, werden mittels der Gütefunktion bewertet. Die Reaktion, welche das beste Dummy-Produkt liefert, wird im zweiten Unterschritt verwendet. In diesem zweiten Schritt wird die Reaktion mit dem zu erweiternden Zwischenprodukt und allen Synthesefragmenten aus der Bibliothek, welche die komplementäre reaktive Substruktur aufweisen, durchgeführt. Aus allen entstehenden Produkten wird abschließend jenes mit der höchsten Güte als neues Zwischenprodukt gewählt, welches im nächsten Erweiterungsschritt bearbeitet wird. Dies wiederholt sich, bis das Molekül entweder eine Mindestmasse überschritten und ein Erweiterungsschritt mit verschlechternder Güte durchgeführt wurde oder das wachsende Molekül eine maximale molekulare Masse überschreitet. Anschließend wird ein neues Startfragment gewählt und die nächste virtuelle Synthese beginnt. Eine vom Benutzer bestimmbare Anzahl von Startfragmenten wird so abgearbeitet. Alle Schritte des Aufbauprozesses sind deterministisch.

DOGS wurde zunächst in einer Reihe von theoretischen Untersuchungen evaluiert. Neben den Faktoren, welche zwangsläufig Einfluß auf die Laufzeit haben (Anzahl Fragmente und Reaktionen, gewählte Anzahl zu bearbeitender Startfragmente, Größe des Referenzmoleküls), ist vor allem die Parametrisierung der Gütefunktion für die Dauer eines DOGS-Laufes verantwortlich. In einem Testszenario erzeugte ein durchschnittlicher DOGS-Lauf mit 200

Startfragmenten und der Standardparametrisierung des molekularen (reduzierten) Molekülgraphen etwa 180 (240) unterschiedliche Molekülstrukturen in 11 (10) Stunden. Diese basieren in beiden Fällen auf ca. 70 unterschiedlichen molekularen Grundgerüsten (Scaffolds). Eine nähere Untersuchung der Ähnlichkeiten zwischen Referenz-Scaffold und von DOGS erzeugten Scaffolds zeigte, dass sich die beiden Graphenrepräsentationen in diesem Punkt für das gewählte Ähnlichkeitsmaß nur unwesentlich differieren. Die durchschnittliche Distanz der erzeugten Scaffolds zur Referenz im gewählten Deskriptorraum unterscheidet sich für beide Moleküldarstellungen kaum. Ein Vergleich auf struktureller Ebene zeigte jedoch, dass sich zwischen den Graphenrepräsentationen nur geringe bis mäßige Überschneidungen in den erzeugten Scaffolds ergeben. Die beiden Graphenrepräsentationen sind somit komplementär und ergeben zusammen eine reichhaltigere Sammlung an entworfenen Scaffolds als jede für sich.

DOGS sollte in der Lage sein, wirkstoffartige Moleküle zu generieren, sofern das Referenzmolekül ebenfalls wirkstoffartig ist. Um dies zu überprüfen, wurden die erzeugten Moleküle aus insgesamt zehn DOGS-Läufen basierend auf fünf verschiedenen Trypsin-Inhibitoren hinsichtlich dieser Eigenschaft untersucht. Ein Großteil (79%) der entworfenen Moleküle verletzt weniger als zwei von Lipinkis „Rule of 5“ Kriterien für bioverfügbare Moleküle. Weiterhin beurteilt ein Klassifizierer für Wirkstoffartigkeit 51% der Moleküle mit einem Wert $>0,8$, wobei 1,0 dem Höchstwert entspricht. Die übrigen 49% der Werte verteilen sich relativ homogen über die Bandbreite möglicher Einschätzungen, mit der Ausnahme, dass auch ein deutlicher Anteil als nicht wirkstoffartig eingestuft wird. Dabei ist zu berücksichtigen, dass auch zwei der Referenzliganden als wirkstoff-untypisch bewertet werden. Die entworfenen Moleküle folgen weiterhin in der Verteilung ihrer berechneten $\log P(o/w)$ Werte der Verteilung dieser Eigenschaft in den Referenzmolekülen. Generell zeigt diese Analyse, dass DOGS in der Lage ist Eigenschaften der Referenzen in die entworfenen Moleküle zu übertragen, die nicht explizit in die Ähnlichkeitsbewertung während der Konstruktion eingehen.

Die Synthetisierbarkeit der von DOGS vorgeschlagenen Moleküle wird für einen Großteil als sehr gut bewertet (77% aller Moleküle liegen in den oberen 10% der Werteskala). Zur Bewertung dieser Eigenschaft wurde ein deskriptorbasiertes Verfahren zur retrosynthetischen Zerlegbarkeit von Molekülen herangezogen. Der verbleibende Anteil verteilt sich auf das obere Mittelfeld des möglichen Wertebereiches. Insgesamt folgt auch hier die Verteilung der DOGS-Moleküle der Werteverteilung der Referenzen. Die reaktionsgetriebene Verknüpfung

von verfügbaren Ausgangsmaterialien resultiert in einer ausgesprochen positiven Bewertung der Synthetisierbarkeit der so konstruierten Moleküle.

Die visuelle Bewertung von ausgewählten DOGS Entwürfe für drei unterschiedliche Zielmoleküle (Trypsin, Östrogen Rezeptor, TGF- β 1 Rezeptor) zeigte, dass das Programm in der Lage ist, die räumliche Anordnung von potentiellen Interaktionszentren der Referenzen in die konstruierten Moleküle zu übertragen. Dabei unterscheiden sich die vorgeschlagenen Moleküle in unterschiedlichem Maße strukturell von der jeweiligen Referenz. Die zugehörigen Reaktionswege sind kurz (ein bis zwei Syntheseschritte) und erscheinen plausibel. Bei einigen Schritten kann der Einsatz von Schutzgruppen zur Vermeidung konkurrierender Nebenreaktionen notwendig sein. Weiterhin wurden die besten 200 für Trypsin entworfenen DOGS-Moleküle auf vorgeschlagene Bioisostere für die S1-Tasche-adressierenden Seitenketten der Referenzen untersucht. Unter den 11 vorgeschlagenen Seitenketten befindet sich unter anderem auch die Seitenkette von Lysin, welche ein natürlicher Ligand der S1-Tasche ist und sich von den Referenzmotiven abhebt. Zwei weitere vorgeschlagene Bioisostere sind in bekannten Trypsin-Inhibitoren als S1-adressierende Seitenkette zu finden. Die meisten der vorgeschlagenen Seitenketten sind positiv ionisierbar und damit in der Lage, eine für die Bindung entscheidende ionische Wechselwirkung mit dem Rezeptor in der S1-Tasche einzugehen.

Schließlich wurde DOGS in zwei realistischen Szenarien zur Identifizierung neuartiger bioaktiver Moleküle eingesetzt und praktisch evaluiert. Für die humane γ -Sekretase wurden aus acht Läufen für vier verschiedene Referenz-Moleküle zwei potentielle Liganden zur Synthese ausgewählt. Beide Verbindungen ließen sich nach dem vom Programm vorgeschlagenen Syntheseweg herstellen. Weiterhin zeigen beide Liganden einen biologischen Effekt am Zielmolekül und modulieren die Aktivität der γ -Sekretase. Die Modulation entspricht dabei allerdings nicht der ursprünglich für therapeutische Zwecke vorgeschlagenen Art und Weise. Die Verbindungen können als Startpunkt weiterer struktureller Optimierungen dienen.

In einer zweiten praktischen Studie mit dem Ziel des Ligandenentwurfs für den humanen Histaminrezeptor Typ 4 wurde aus der Menge der computergenerierten Vorschläge eine Verbindung zur Synthese und Testung ausgewählt. Die dreistufige Synthese konnte wie vorgeschlagen nachvollzogen werden. Der Ligand zeigt mit einem K_i von 436 μ M jedoch nur sehr schwache Aktivität. Grund dafür könnte ein fehlender Wasserstoffbrückendonator im zentralen Teil des Liganden sein, der in anderen Studien als Teil des Pharmakophors angenommen wurde.

Mit DOGS wurde ein neues Werkzeug zum *de novo* Design wirkstoffartiger Moleküle vorgeschlagen. DOGS gehört zu den wenigen Programmen dieser Art, welche einer praktischen Evaluierung unterzogen wurden. Die Ergebnisse der retrospektiven und prospektiven Auswertung zeigen das Potential des Ansatzes auf, Vorschläge von praktischer Relevanz zu generieren. Das Konzept zum Molekülaufbau von DOGS hat gezeigt, dass es nicht nur synthetisierbare Strukturen hervorbringt, sondern zusätzlich auch nachvollziehbare praktikable Vorschläge für deren Synthese liefern kann. Dies ist ein essentieller Vorteil im praktischen Einsatz gegenüber vielen bisher beschriebenen Ansätzen zum *de novo* Design. Zukünftige Verbesserungen in unserem Verständnis von molekularer Ähnlichkeit und Liganden-Rezeptor-Wechselwirkungen können problemlos in Form neuer Gütefunktionen in das Konzept von DOGS eingebunden werden.

References

1. Böhm, H. J., Klebe, G., Kubinyi, H. (1996) "Automatische Konstruktion neuer Proteinliganden" in *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg, Germany.
2. Stumpfe, D., Geppert, H., Bajorath, J. (2010) "In Silico Screening" in *Lead Generation Approaches in Drug Discovery*, Rankovic, Z., Morphy, R., (Ed.), John Wiley & Sons, Hoboken, N.J., USA, pp. 73-103.
3. Eglén, R.M., Schneider, G., Böhm H.-J. (2000) "High-throughput Screening and Virtual Screening: Entry Points to Drug Discovery" in *Virtual Screening for Bioactive Molecules*, Schneider, G., Böhm, H.-J. (Ed.), Wiley-VCH, Weinheim, Germany, pp. 1-14.
4. Schneider, G., Fechner, U. (2005) Computer-based De Novo Design of Druglike Molecules, *Nature Reviews Drug Discovery* 4, 649-663.
5. Danziger, D. J., Dean, P. M. (1989) Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proceedings of the Royal Society of London, Series B*, 236, 101-113.
6. Lewis, R. A., Dean, P. M. (1989) Automated site-directed drug design: the concept of spacer skeletons for primary structure generation, *Proceedings of the Royal Society of London, Series B*, 236, 125-140.
7. Lewis, R.A., Dean, P.M. (1989). Automated site-directed drug design: the formation of molecular templates in primary structure generation, *Proceedings of the Royal Society of London, Series B*, 236, 141-162.
8. Gillett, V. A., Johnson, A. P., Mata, P., Sike, S. (1990) Automated structure design in 3D. *Tetrahedron Computer Methodology* 3, 681-696.
9. Lewis, R. A., Roe, D. C., Huang C., Ferrin T. E., Langridge, R., Kuntz, I. D. (1992) Automated site-directed drug design using molecular lattices- *Journal of Molecular Graphics* 10, 66-78.
10. Böhm, H.-J. (1992) The computer program LUDI: a new simple method for the de-novo design of enzyme inhibitors, *Journal of Computer-Aided Molecular Design* 6, 61-78.
11. Böhm, H.-J. (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads, *Journal of Computer-Aided Molecular Design* 6, 593-606.
12. Böhm, H.-J. (1993) A novel computational tool for automated structure-based drug design, *Journal of Molecular Recognition* 6, 131-137.
13. Böhm, H.-J. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure, *Journal of Computer-Aided Molecular Design* 8, 243-256.
14. Böhm, H.-J. (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *Journal of Computer-Aided Molecular Design*, 12, 309-323.
15. Tschinke, V., Cohen, N. C. (1993) The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypothesis, *Journal of Medicinal Chemistry* 36, 3863-3870.
16. Ho, C. M. W., Marshall, G. R. (1993) SPLICE: a program to assemble partial query solutions from three-dimensional database searches into novel ligands, *Journal of Computer-Aided Molecular Design* 7, 623-647.

17. Rotstein, S. H., Murcko, M. A. (1993) GroupBuild: a fragment-based method for de novo drug design, *Journal of Medicinal Chemistry* 36, 1700-1710.
18. Pearlman, D.A., Murcko, M.A. (1993) CONCEPTS: new dynamic algorithm for de novo design suggestion, *Journal of Computational Chemistry* 14, 1184-1193.
19. Gillet, V. J., Johnson, A. P., Mata P., Sike, S., Williams P. (1993). SPROUT: a program for structure generation, *Journal of Computer-Aided Molecular Design* 7, 127-153.
20. Gillet, V. J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z., Johnson, A. P. (1994) SPROUT: recent developments in the de novo design of molecules, *Journal of Computer-Aided Molecular Design* 34, 207-217.
21. Mata, P., Gillet, V.J., Johnson, A.P., Lampreia, J., Myatt, G. J., Sike, S., Stebbings, A. L. (1995) SPROUT: 3D structure generation using templates, *Journal of Chemical Information and Computer Sciences* 35, 479-493.
22. Gillett, V. J., Myatt, G., Zsoldos, Z. Johnson, A. P. (1995) SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility, *Perspectives in Drug Discovery and Design* 3, 34-50.
23. Eisen, M. B., Wiley, D. C., Karplus, M., Hubbard, R. E. (1994) HOOK: a program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins* 19, 199-221.
24. Miranker, A., Karplus, M. (1991) Functionality maps of binding sites: a multiple copy simultaneous search method, *Proteins* 11, 29-34.
25. Bohacek, R. S., McMartin, C. (1994) Multiple highly diverse structures complementary to enzyme binding sites: results of extensive application of a de novo design method incorporating combinatorial growth, *Journal of the American Chemical Society* 116, 5560-5571.
26. Glen, R. C., Payne, A. W. R. (1995) A genetic algorithm for the automated generation of molecules within constraints, *Journal of Computer-Aided Molecular Design* 9, 181-202.
27. Clark, D. E., Frenkel, D., Levy, S. A., Li, J., Murray, C. W., Robson B., Waszkowycz, B., Westhead, D. R. (1995) PRO LIGAND: an approach to de novo molecular design. 1. Application to the design of organic molecules. *Journal of Computer-Aided Molecular Design* 9, 13-32.
28. Waszkowycz, B., Clark, D. E., Frenkel, D., Li, J., Murray, C.W., Robson, B., Westhead, D. R., (1994) PRO LIGAND: an approach to de novo molecular design. 2. design of novel molecules from molecular field analysis (MFA) models and pharmacophores, *Journal of Medicinal Chemistry* 37, 3994-4002.
29. Westhead, D. R., Clark, D. E., Frenkel, D., Li, J., Murray, C. W., Robson, B., Waszkowycz, B. (1995) PRO LIGAND: an approach to de novo molecular design. 3. A genetic algorithm for structure refinement, *Journal of Computer-Aided Molecular Design* 9, 139-148.
30. Frenkel, D., Clark, D. E., Li, J., Murray, C. W., Robson, B., Waszkowycz, B., Westhead, D. R. (1995) PRO LIGAND: an approach de novo molecular design. 4. Application to the design of peptides, *Journal of Computer-Aided Molecular Design* 9, 213-225.
31. Clark, D. E., Murray, C. W. (1995) PRO LIGAND: an approach to de novo molecular design. 5. Tools for the Analysis of Generated Structures, *Journal of Chemical Information and Computer Sciences*, 35, 914-923.
32. Murray, C. W., Clark, D. E., Byrne, D. G. (1995) PRO LIGAND: an approach to de novo molecular design. 6. Flexible fitting in the design of peptides, *Journal of Computer-Aided Molecular Design* 9, 381-395.
33. DeWitte, R. S., Shakhnovich, E. I. (1996) SMOG de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *Journal of the American Chemical Society* 118, 11733-11744.

34. Ishchenko, A. V., Shakhnovich, E. I. (2002) SMoG2001: an improved knowledge-based scoring function for protein–ligand interactions, *Journal of Medicinal Chemistry* 45, 2770-2780.
35. Grzybowski, B. A., Ishchenko, A. V., Kim, C.-Y., Topalov, G., Chapman, R., Christianson, D. W., Whitesides, G. M., Shakhnovich, E. I. (2002) Combinatorial computational method gives new picomolar ligands for a known enzyme. *Proceedings of the National Academy of Science USA* 99, 1270-1273.
36. Pearlman, D. A., Murcko, M. A. (1996) CONCERTS: dynamic connection of fragments as an approach to de novo ligand design, *Journal of Medicinal Chemistry* 39, 1651-1663.
37. Murray, C. W., Clark, D. E., Auton, T. R., Firth, M. A., Li, J., Sykes, R. A., Waszkowycz, B., Westhead, D. R., Young, S. C. (1997) PRO_SELECT: combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *Journal of Computer-Aided Molecular Design* 11, 193-207.
38. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., Mee, R. P. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design* 11, 425-445.
39. Todorov, N. P., Dean, P. M. (1997) Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design, *Journal of Computer-Aided Molecular Design* 11, 175-192.
40. Todorov, N. P., Dean, P. M. (1998) A branch-and-bound method for optimal atom-type assignment in de novo ligand design, *Journal of Computer-Aided Molecular Design* 12, 335-350.
41. Stahl, M., Todorov, N. P., James, T., Mauser, H., Böhm, H. J., Dean, P. M. (2002) A validation study on the practical use of automated de novo design, *Journal of Computer-Aided Molecular Design* 16, 459-478.
42. Lloyd, D., Buenemann, C. L., Todorov, N. P., Manallack D. T., Dean, P. M. (2004) Scaffold hopping in de novo design. Ligand generation in the absence of receptor information, *Journal of Medicinal Chemistry* 47, 493-496.
43. Roche, O., Sarmiento, R. M. R. (2007) A new class of histamine H3 receptor antagonists derived from ligand based design, *Bioorganic and Medicinal Chemistry Letters* 17, 3670-3675.
44. Stahl, M., Todorov, N. P., James T., Mauser, H., Böhm, H.-J., Dean, P. M. (2002) A validation study on the practical use of automated de novo design, *Journal of Computer Aided Molecular Design*, 16, 459-478.
45. Nachbar, R. B. (1998) Molecular evolution: a hierarchical representation for chemical topology and its automated manipulation. *Proc. 3rd Annual Genetic Programming Conference*, 246-253.
46. Nachbar, R. B. (2000) Molecular evolution: automated manipulation of hierarchical chemical topology and its application to average molecular structures, *Genetic Programming and Evolvable Machines* 1, 57-94.
47. Globus, A., Lawton, J., Wipke, W. T. (1999) Automatic Molecular design using evolutionary algorithms. *Nanotechnology* 10, 290-299.
48. Liu, H., Duan, Z., Luo, Q. Shi, Y. (1999) Structure-based ligand design by dynamically assembling molecular building blocks at binding site, *Proteins* 36, 462-470.
49. Zhu, J., Yu, H., Fan, H., Liu, H., Shi, Y. (2001) Design of selective inhibitors of cyclooxygenase-2 dynamic assembly of molecular building blocks, *Journal of Computer-Aided Molecular Design* 15, 447-463.

50. Douguet, D., Thoreau, E., Grassy, G. (2000) A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm, *Journal of Computer-Aided Molecular Design* 14, 449-466.
51. Wang, R., Gao, Y., Lai, L. (2000) LigBuilder: a multi-purpose program for structure-based drug design, *Journal of Molecular Modeling* 6, 498-516.
52. Schneider, G., Lee, M.-L., Stahl, M., Schneider, P. (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks, *Journal of Computer-Aided Molecular Design* 14, 487-494.
53. Schneider, G., Clément-Chomienne, O., Hilfiger, L., Schneider, P., Kirsch, S., Böhm, H.-J., Neidhart, W. (2000) Virtual screening for bioactive molecules by evolutionary de novo design. *Angewandte Chemie International Edition* 39, 4130-4133.
54. Zhu, J., Fan, H., Liu, H., Shi, Y. (2001) Structure-based ligand design for flexible proteins: application of new F-DycoBlock, *Journal of Computer-Aided Molecular Design* 15, 979-996.
55. Pegg, S. C.-H., Haresco, J. J., Kuntz, I. D. (2001) A genetic algorithm for structure-based *de novo* design. *Journal of Computer-Aided Molecular Design* 15, 911-933.
56. Pellegrini, E., Field, M. J. (2003) Development and testing of a *de novo* drug-design algorithm. *Journal of Computer-Aided Molecular Design* 17, 621-641.
57. Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F., Heeres, J., Koymans, L. M., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., Janssen, P.A. (2003) SYNOPSIS: SYNthesize and OPTimize System in Silico. *Journal of Medicinal Chemistry* 46, 2765-2773.
58. Brown, N., McKay, B., Gilardoni, F., Gasteiger, J. (2004) A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules, *Journal of Chemical Information and Computer Sciences* 44, 1079-1087.
59. Pierce, A. C., Rao, G., Bemis, G. W. (2004) BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease, *Journal of Medicinal Chemistry* 47, 2768-2775.
60. Nikitin, S., Zaitseva, N., Demina, O., Solovieva, V., Mazin, E., Mikhalev, S., Smolov, M., Rubinov, A., Vlasov, P., Lepikhin, D., Khachko, D., Fokin, V., Queen, C., Zosimov, V. (2005) A very large diversity space of synthetically accessible compounds for use with drug design programs, *Journal of Computer Aided Molecular Design* 19, 47-63.
61. Douguet, D., Munier-Lehmann, H., Labesse, G., Pochet, S. (2005). LEA3D: A computer-aided ligand design for structure-based drug design, *Journal of Medicinal Chemistry*, 48, 2457-2468.
62. Fechner, U., Schneider, G. (2006). Flux (1): a virtual synthesis scheme for fragment-based de novo design. *Journal of Chemical Information and Modeling*, 46, 699-707.
63. Fechner, U., Schneider, G. (2007). Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. *Journal of Chemical Information and Modeling*, 47, 656-667.
64. Degen, J., Rarey, M. (2006) FlexNovo: structure-based searching in large fragment spaces, *ChemMedChem* 1, 854-868.
65. Jorgensen, W. L., Ruiz-Caro, J., Tirado-Rives, J., Basavapathruni, A., Anderson, K. S., Hamilton, A. D. (2006) Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase, *Bioorganic and Medicinal Chemistry Letters* 16, 663-667.
66. Feher, M., Gao, Y., Baber, C., Shirley, W. A., Saunders, J. (2008) The use of ligand-based de novo design for scaffold hopping and sidechain optimization: two case studies *Bioorganic and Medicinal Chemistry* 16, 422-427.

67. Dey, F., Caflisch, A. (2008). Fragment-based de novo ligand design by multiobjective evolutionary optimization. *Journal of Chemical Information and Modeling*, 48, 679-690.
68. Hartenfeller, M., Proschak, E., Schüller, A., Schneider, G. (2008). Concept of combinatorial *de novo* design of druglike molecules by particle swarm optimization. *Chemical Biology and Drug Design*, 72, 16-26.
69. Proschak, E., Zettl, H., Tanrikulu, Y., Weisel, M., Kriegl, J. M., Rau, O., Schubert-Zsilavecz, M., Schneider, G. (2009) From molecular shape to potent bioactive agents I: bioisosteric replacement of molecular fragments, *ChemMedChem* 4, 41-4.
70. Proschak, E., Sander, K., Zettl, H., Tanrikulu, Y., Rau, O., Schneider, P., Schubert-Zsilavecz, M., Stark, H., Schneider, G. (2009) From molecular shape to potent bioactive agents II: fragment-based de novo design, *ChemMedChem* 4, 45-48.
71. Hecht, D., Fogel, G. B. (2009) A Novel In Silico Approach to Drug Discovery via Computational Intelligence, *Journal of Chemical Information and Modeling* 49, 1105-1121.
72. Kutchukian, P. S., Lou, D., Shakhnovich, E. I. (2009) FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space, *Journal of Chemical Information and Modeling* 49, 1630-1642.
73. Moriaud, F., Doppelt-Azeroual, O., Martin, L., Oguievetskaia, K., Koch, K., Vorotyntsev, A., Adcock, S. A., Delfaud, F. (2009) Computational Fragment-Based Approach at PDB Scale by Protein Local Similarity, *Journal of Chemical Information and Modeling* 49, 280-294.
74. Nicolaou, C. A., Apostolakis, J., Pattichis, C. S. (2009) De Novo Drug Design Using Multiobjective Evolutionary Graphs, *Journal of Chemical Information and Modeling* 49, 295-307.
75. Nisius, B., Rester, U. (2009) Fragment Shuffling: An Automated Workflow for Three-Dimensional Fragment-Based Ligand Design, *Journal of Chemical Information and Modeling* 49, 1211-1222.
76. Durrant, J. D., Amaro, R. E., McCammon, J. A. (2009) AutoGrow: A Novel Algorithm for Protein Inhibitor Design, *Chem Biol Drug Des* 73, 168-78.
77. Lessel, U., Wellenzohn, B., Lilienthal, M., Claussen, H. (2009) Searching Fragment Spaces with Feature Trees, *Journal of Chemical Information and Modeling* 49, 270-279.
78. Damewood, J. R., Lerman, C. L., Masek, B. B. (2010) NovoFLAP: A Ligand-Based De Novo Design Approach for the Generation of Medicinally Relevant Ideas., *Journal of Chemical Information and Modeling* (published online).
79. Huang, Q., Li, L-L., Yang, S-J.. (2010) PhDD: a new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility, *Journal of Molecular Graphics and Modelling* 28, 775-787.
80. Pfeffer, P., Foer, T., Hüllermeier, E., Klebe, G. (2010) GARLig: A Fully Automated Tool for Subset Selection of Large Fragment Spaces via a Self-Adaptive Genetic Algorithm, *Journal of Chemical Information and Modeling*, published online.
81. Hartenfeller M., Schneider G. (2011) De novo drug design, *Methods in Molecular Biology* 672, 299-323.
82. Wise, A., Gearing, K., Rees, S. (2002) Target validation of G-protein coupled receptors. *Drug Discovery Today* 7, 235-246.
83. Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H., Kuhn, P., Weis, W. I., Kobilka, B. K., Stevens, R. C. (2007) High-Resolution Crystal Structure of an Engineered Human β 2-Adrenergic G Protein-Coupled Receptor, *Science* 318, 1258-1265.
84. Johnson, M. A., Maggiora, G. M. (Eds.) (1990) *Concepts and Applications of Molecular Similarity*, John Wiley & Sons, New York, USA.

85. Schneider, G., Böhm, H. (2002) Virtual screening and fast automated docking methods, *Drug Discovery Today* 7, 64-70.
86. Höltje, H.-D., Sippl, W., Rognan, G., Folkers, G. (2008) *Molecular Modeling. Basic Principles and Applications*, 3rd edition, Wiley-VCH, Weinheim, Germany.
87. Jones, G., Willett, P., Glen, R. C. (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation, *Journal of Molecular Biology* 245, 43-53.
88. Jones, G., Willett, P., Glen, R. C., Leach, A. R., Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking, *Journal of Molecular Biology* 267, 727-748.
89. Rarey, M., Kramer, B., Lengauer, T., Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology* 261, 470-489.
90. Gohlke, H., Hendlich, M., Klebe, G. (2000) Knowledge-based scoring function to predict protein-ligand interactions. *Journal of Molecular Biology* 295, 337-356.
91. Gohlke, H., Hendlich, M., Klebe, G. (2000) Predicting binding modes, binding affinities and 'hot spots' for protein-ligand complexes using a knowledge-based scoring function, *Perspectives in Drug Discovery and Design* 20, 115-144.
92. Lloyd, D. G., Buenemann, C. L., Todorov, N. P., Manallack, D. T., Dean, P. M. (2004) Scaffold Hopping in De Novo Design. Ligand Generation in the Absence of Receptor Information, *Journal of Medicinal Chemistry* 47, 493-496.
93. Tanrikulu, Y., Schneider, G. (2008) Pseudoreceptor models in drug design: bridging ligand- and receptor-based virtual screening, *Nature Reviews Drug Discovery* 7, 667-677.
94. Luo, Z. W., Wang, R. X., Lai, L. H. (1996) RASSE: A new method for structure-based drug design, *Journal of Chemical Informatics and Computer Science* 36, 1187-1194.
95. Markov, A. A., (1971) "Extension of the limit theorems of probability theory to a sum of variables connected in a chain" in *Dynamic Probabilistic Systems, vol. 1*, Markov Chains (reprinted in Appendix B), Howard, R. (Ed.), John Wiley & Sons, Hoboken, N.J., USA.
96. Lewell, X. Q., Judd, D., Watson, S., Hann, M. (1998) RECAP - retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry *Journal of Chemical Information and Computer Science* 38, 511-522.
97. Daylight Chemical Information Systems, Inc., 120 Vantis-Suite 550, Aliso Viejo, CA 92656, USA.
98. Symyx Technology Inc., 2440 Camino Ramon, Suite 300, San Ramon, CA 94583, USA.
99. Bohacek R. S., McMartin C., Guida W. C. (1996) The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective, *Medicinal Research Reviews* 16, 3-50.
100. Rechenberg I. (1973) *Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Frommann-Holzboog, Stuttgart, Germany.
101. Peltason, L., Bajorath, J. (2007) SAR index: quantifying the nature of structure-activity relationships. *Journal of Medicinal Chemistry* 50, 5571-5578
102. Maggiora, G. M. (2006) On Outliers and Activity Cliffs-Why QSAR Often Disappoints, *Journal of Chemical Information and Modeling* 46, 1535.
103. Schneider G., Hartenfeller M., Reutlinger M., Tanrikulu Y., Proschak E., Schneider P. (2008) Voyages to the (un)known: Adaptive Design of Bioactive Compounds. *Trends in Biotechnology* 27, 18-26.

104. Schneider, G., Hartenfeller, M., Proschak, E. "De Novo Design" in *Lead Generation Approaches in Drug Discovery*, Rankovic Z., Morphy R. (Eds.), John Wiley & Sons: Hoboken, N.J., USA (2010), pp. 165-185.
105. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953) Equation of state calculations by fast computing machines, *Journal of Chemical Physics* 21, 1087-1092.
106. Altekari, G., Dwarkadas, S., Huelsenbeck, J., Ronquist, F. (2004) Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference, *Bioinformatics* 20, 407-415.
107. Guo, M., Liu, Y., Malec, J. (2004) A New Q-Learning Algorithm Based on the Metropolis Criterion, *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* 34, 2140-2143.
108. Hiss, J. A., Hartenfeller, M., Schneider, G. (2010) Concepts and applications of "natural computing" techniques in de novo drug and peptide design, *Current Pharmaceutical Design* 16, 1656-1665.
109. Kennedy, J., Eberhart, R. C. (1995) Particle swarm optimization, *Proceedings of the IEEE International Conference on Neural Networks*, 1942-1948.
110. Cormen, T. H., Leiserson, C. E., Rivest, R. L. (1990) "Greedy Algorithms" in *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts, USA.
111. Klebe G., Böhm H. J. (1997) Energetic and entropic factors determining binding affinity in protein-ligand complexes, *Journal of Receptor and Signal Transduction Research* 17, 459-473.
112. Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J., Green, D. V. (2002) Combinatorial library design using a multiobjective genetic algorithm, *Journal of Chemical Information and Computer Sciences* 42, 375-85.
113. Brown, N., McKay, B., Gasteiger, J. (2004) The de novo design of median molecules within a property range of interest, *Journal of Computer-Aided Molecular Design* 18, 761-771.
114. Kutchukian, P. S., Shakhnovich, E. I. (2010) De novo design: balancing novelty and confined chemical space, *Expert Opinion on Drug Discovery* 5, 789-812.
115. Patel, H., Bodkin, M. J., Chen, B., Gillet, V. J. (2009) Knowledge-Based Approach to de Novo Design Using Reaction Vectors, *Journal of Chemical Information and Modeling* 49, 1163-1184.
116. Boda, K., Seidel, T., Gasteiger, J. (2007) Structure and reaction based evaluation of synthetic accessibility, *Journal of Computer-Aided Molecular Design* 21, 311-325.
117. Gillet, V. J., Myatt, G., Zsoldos, Z. and Johnson, A. P. (1995) SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility, *Perspectives in Drug Discovery and Design* 3, 34-50.
118. Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S. Y., Johnson, A. P., Major, S., Wade, R. A., Ando, H. Y. (2009) Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation *Journal of Chemical Information and Modeling* 49, 593-602.
119. Schneider, G., Baringhaus, K.-H. (1998) *Molecular Design: Concepts and Applications*, Wiley-VCH, Weinheim, Germany.
120. Herschhorn, A., Lerman, L., Weitman, M., Gleenberg, I. O., Nudelman, A., Hizi, A. (2007) De novo parallel design, synthesis and evaluation of inhibitors against the reverse transcriptase of human immunodeficiency virus type-1 and drug-resistant variants, *Journal of Medicinal Chemistry* 50, 2370-2384.
121. Firth-Clark, S., Willems, H. M., Williams, A., Harris, W. (2006) Generation and selection of novel estrogen receptor ligands using the de novo structure-based design tool, SkelGen, *Journal of Chemical Information and Modeling* 46, 642-647.

122. Schneider, G., Neidhart, W., Giller, T., Schmid, G. (1999) Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening, *Angewandte Chemie International Edition* 38, 2894-2896.
123. Maass, P., Schulz-Gasch, T., Stahl, M., Rarey, M. (2007) Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations, *Journal of Chemical Information and Modeling* 47, 390-399.
124. Reisen, F. H., Schneider, G., Proschak, E. (2009) Reaction-MQL: Line Notation for Functional Transformation, *Journal of Chemical Information and Modeling* 49, 6-12.
125. Sigma-Aldrich Co., 3050 Spruce St, St. Louis, MO 63103, USA.
126. Irwin, J. J., Shoichet, B. K. (2005) ZINC--a free database of commercially available compounds for virtual screening, *Journal of Chemical Information and Modeling* 45, 177-182.
127. ZINC database: <http://zinc.docking.org/> (accessed 13.05.2010)
128. Chemical Computing Group, Suite 910, 1010 Sherbrooke Street West, Montreal, Quebec, Canada.
129. Hann, M., Hudson, B., Lewell, X., Lively, R., Miller, L., and Ramsden, N. (1999) Strategic Pooling of Compounds for High-Throughput Screening, *Journal of Chemical Information and Computer Sciences* 39, 897-902.
130. Oracle Corporation, 500 Oracle Parkway, Redwood Shores, CA 94065, USA.
131. Rupp, M., Proschak, E., Schneider, G. (2007) Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity, *Journal of Chemical Information and Modeling* 47, 2280-2286.
132. Rupp, M., Schroeter, T., Steri, R., Zettl, H., Proschak, E., Hansen, K., Rau, O., Schwarz, O., Müller-Kuhrt, L., Schubert-Zsilavecz, M., Müller, K., Schneider, G. (2010) From Machine Learning to Natural Product Derivatives that Selectively Activate Transcription Factor PPARgamma, *ChemMedChem* 5, 191-194.
133. Proschak, E., Wegner, J. K., Schüller, A., Schneider, G., Fechner, U. (2007) Molecular Query Language (MQL) - A Context-Free Grammar for Substructure Matching, *Journal of Chemical Information and Modeling* 47, 295-301.
134. Plotkin, M. (1971) Mathematical Basis of Ring-Finding Algorithms at CIDS, *Journal of Chemical Documentation* 11, 60-63.
135. Langdon, S. R., Ertl, P., Brown, N. (2010) Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization, *Molecular Informatics* 5, 366-385.
136. Bemis, G. W., Murcko, M. A. (1996) The Properties of Known Drugs. 1. Molecular Frameworks, *Journal of Medicinal Chemistry* 39, 2887-2893.
137. Hall, L. H., Kier, L. B. (1991) The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure-Property Modeling, *Reviews of Computational Chemistry* 2, 367-422.
138. Petitjean, M. (1992) Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds, *Journal of Chemical Information and Computer Science* 32, 331-337.
139. Schneider, P., Schneider, G. (2003) Collection of Bioactive Reference Compounds for Focused Library Design, *QSAR & Combinatorial Science* 22, 713-718.
140. Steinbeck, C., Han, Y. Q., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E. L. (2003) The Chemistry Development Kit (CDK): An Open-source Java Library for Chemo- and Bioinformatics. *Journal of Chemical Information and Computer Sciences* 43, 493-500.

141. Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., Willighagen, E. L. (2006) Recent Developments of the Chemistry Development Kit (CDK) - an Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design* 12, 2111-2120.
142. Sander, K., Kottke, T., Tanrikulu, Y., Proschak, E., Weizel, L., Schneider, E. H., Seifert, R., Schneider, G., Stark, H. (2009) 2,4-Diaminopyrimidines as histamine H4 receptor ligands – Scaffold optimization and pharmacological characterization, *Bioorganic & Medicinal Chemistry* 17, 7186-7196.
143. Sawyer, J. S., Anderson, B. D., Beight, D. W., Campbell, R. M., Jones, M. L., Herro, D. K., Lampe, J. W., McCowan, J. R., McMillen, W. T., Mort, N., Parsons, S., Smith, E. C. R., Vieth, M., Weir, L. C., Yan, L., Zhang, F., Yingling, J. M. (2003) Synthesis and Activity of New Aryl- and Heteroaryl-Substituted Pyrazole Inhibitors of the Transforming Growth Factor- β Type I Receptor Kinase Domain, *Journal of Medicinal Chemistry* 46, 3953-3956.
144. Thurmond, R. L., Desai, P. J., Dunford, P. J., Hofstra, C. L., Jiang, W., Nguyen, S., Riley, J. P., Sun, S., Williams, K. N., Edwards, J. P., Karlsson, L. (2004) A Potent and Selective Histamine H4 Receptor Antagonist with Anti-Inflammatory Properties, *The Journal of Pharmacology and Experimental Therapeutics* 309, 404-413.
145. Lipinski C. A., Lombardo F., Dominy B. W., Feeney P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Advanced Drug Delivery Reviews* 23, 3-25.
146. Senokuchi, K., Nakai, H., Nakayama, Y., Odagaki, Y., Sakaki, K., Kato, M., Maruyama, T., Miyazaki, T., Ito, H., Kamiyasu, K., Kim, S., Kawamura, M., Hamanaka, N. (1995) New orally active serine protease inhibitors, *Journal of Medicinal Chemistry* 38, 2521-2523.
147. Bergner, A., Bauer, M., Brandstetter, H. (1995) The x-ray crystal structure of thrombin in complex with N-alpha-2-naphthylsulfonyl-L-3-amidino-phenylalanyl-4-methylpiperidide: the beneficial effect of filling out an empty cavity, *Journal of Enzyme Inhibition* 9, 101-110.
148. Costanzo, M. J., Almond, H. R., Hecker, L. R., Schott, M. R., Yabut, S. C., Zhang, H., Andrade-Gordon, P., Corcoran, T. W., Giardino, E. C., Kauffman, J. A., Lewis, J. M., de Garavilla, L., Haertlein, B. J., Maryanoff, B. E. (2005) In-depth study of tripeptide-based alpha-ketoheterocycles as inhibitors of thrombin. Effective utilization of the S1' subsite and its implications to structure-based drug design, *Journal of Medicinal Chemistry* 48, 1984-2008.
149. Miyamoto, K., Hishinuma, I., Nagakawa, J., Nagaoka, N., Yamanaka, T., Wakabayashi, T. (1988) Effects of E-3123, a new protease inhibitor, on several protease activities and on experimental acute pancreatitis, *Nippon Yakurigaku Zasshi* 91, 285-293.
150. Nochi, S., Shimomura, N., Hattori, T., Sato, T., Miyake, Y., Tanizawa, K. (1989) Kinetic study on the mechanism of inhibition of trypsin and trypsin-like enzymes by p-guanidinobenzoate ester, *Chemical & Pharmaceutical Bulletin* 37, 2855-2857.
151. Menear, K. (1999) Expert Opinion on Investigational Drugs Direct thrombin inhibitors: current status and future prospects, *Expert Opinion on Investigational Drugs* 8, 373-1384.
152. Schneider, G., Schneider, P. (2004) "Navigation in chemical space: Ligand-based design of focused compound libraries" in *Chemogenomics in Drug Discovery*, H. Kubinyi, G. Müller; (Eds.), Wiley-VCH, Weinheim, Germany, pp. 341-376.
153. Testa, B., Carrupt, P.-A., Gaillard, P., Billois, F., Weber, P. (1996) Lipophilicity in Molecular Modeling, *Pharmaceutical Research* 13, 335-343.
154. Leo A., Hansch C., Elkins D. (1971) Partition coefficients and their uses, *Chemical Reviews* 71(6), 525-616.
155. Olsen, J.V., Ong, S. E., Mann, M. (2004) Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues, *Molecular and Cellular Proteomics* 3, 608-614.

156. Sanderson, P. E. J. (1999) Small, Noncovalent Serine Protease Inhibitors, *Medicinal Research Reviews* 19, 179-197.
157. Peterlin-Masic, L., Mlinsek, G., Solmajer, T., Trampus-Bakija, A., Stegnard, M., Kikelj, D. (2003) Novel thrombin inhibitors incorporating non-basic partially saturated heterobicyclic P1-Arginine mimetics, *Bioorganic & Medicinal Chemistry Letters* 13, 789-794.
158. Feng, D. M., Gardell, S. J., Lewis, S. D., Bock, M. G., Chen, Z., Freidinger, R. M., Naylor-Olsen, A. M., Ramjit, H. G., Woltmann, R., Baskin, E. P., Lynch, J. J., Lucas, R., Shafer, J. A., Dancheck, K. B., Chen, I. W., Mao, S. S., Krueger, J. A., Hare, T. R., Mulichak, A. M., Vacca, J. P. (1997) Discovery of a novel, selective, and orally bioavailable class of thrombin inhibitors incorporating aminopyridyl moieties at the P1 position, *Journal of Medicinal Chemistry* 40, 3726-3733.
159. Clegg, N. J., Paruthiyil, S., Leitman, D. C., Scanlan, T. S. (2005) Differential response of estrogen receptor subtypes to 1,3-diarylindene and 2,3-diarylindene ligands, *Journal of Medicinal Chemistry* 48, 5989-6003.
160. Prestwood, K. M., Gunness, M., Muchmore, D. B., Lu, Y., Wong, M., Raisz L. G. (2000) A Comparison of the Effects of Raloxifene and Estrogen on Bone in Postmenopausal Women, *The Journal of Clinical Endocrinology and Metabolism* 85 (6), 2197-2202.
161. Carosati, E., Sciabola, S., Cruciani, G. (2004) Hydrogen bonding interactions of covalently bonded fluorine atoms: from crystallographic data to a new angular function in the GRID force field, *Journal of Medicinal Chemistry* 47, 5114-5125.
162. Zettl, H., Weggen, S., Schneider, P., Schneider, G. (2010) Exploring the chemical space of gamma-secretase modulators, *Trends in Pharmacological Sciences*, published online.
163. Rivkin, A., Ahearn, S. P., Chichetti, S. M., Kim, Y. R., Li, C., Rosenau, A., Kattar, S. D., Jung, J., Shah, S., Hughes, B. L., Crispino, J. L., Middleton, R. E., Szewczak, A. A., Munoz, B., Shearman, M. S. (2009) Piperazinyl pyrimidine derivatives as potent gamma-secretase modulators, *Bioorganic & Medicinal Chemistry Letters* 2, 11-13.
164. Czirr, E., Leuchtenberger, S., Dorner-Ciossek, C., Schneider, A., Jucker, M., Koo, E. H., Pietrzik, C. U., Baumann, K., Weggen, S. (2007) Insensitivity to Abeta 42-lowering non-steroidal anti-inflammatory drugs (NSAIDs) and gamma-secretase inhibitors is common among aggressive presenilin-1 mutations, *Journal of Biological Chemistry* 282, 24504-24513.
165. Tiligada, E., Zampeli, E., Sander, K., Stark, H. (2009) Histamine H3 and H4 receptors as novel drug targets, *Expert Opinion in Investigational Drugs* 18(10), 1519-1531.
166. Smits, R. A., Lim, H. D., Hanzer, A., Zuiderveld, O. P., Guaita, E., Adami, M., Coruzzi, G., Leurs, R., de Esch, I. J. (2008) Fragment Based Design of New H4 receptor-ligands with Anti-inflammatory Properties in Vivo, *Journal of Medicinal Chemistry* 51, 2457-2467.
167. Tanrikulu, Y., Proschak, E., Werner, T., Geppert, T., Todoroff, N., Klenner, A., Kottke, T., Sander, K., Schneider, E., Seifert, R., Stark, H., Clark, T., Schneider, G. (2009) Homology model adjustment and ligand screening with a pseudoreceptor of the human histamine H4 receptor, *ChemMedChem* 4, 820-827.
168. Lee-Dutra, A., Arienti, K. L., Buzard, D. J., Hack, M. D., Khatuya, H., Desai, P. J., Nguyen, S., Thurmond, R. L., Karlsson, L., Edwards, J. P., Breitenbucher, J. G. (2006) Identification of 2-arylbenzimidazoles as potent human histamine H4 receptor ligands, *Bioorganic & Medicinal Chemistry Letters* 16, 6043-6048.
169. Schneider, E. H., Schnell, D., Papa, D., Seifert, R. (2009) High constitutive activity and a G-protein-independent high-affinity state of the human histamine H(4)-receptor, *Biochemistry* 48(6), 1424-1438.
170. Weininger D. (1988) SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules, *Journal of Chemical Information and Computer Sciences* 28(1), 31-36.

171. Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T., Murray, C. W., Taylor, R. D., Watson, P. (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment, *Journal of Chemical Information and Computer Sciences* 44, 793-806.

Supplement

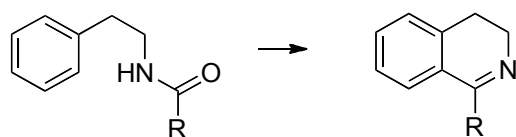
Each coupling and preprocessing reaction is given by the following specifications:

- Reaction-MQL expression,
- Schematic structural representation,
- Minimal structure of educt(s) encoded as SMILES, also representing the dummy fragment used during construction.

Please note that the schematic structural representation not necessarily corresponds completely to the minimal dummy structure given. Schematic representations serve for visualization only.

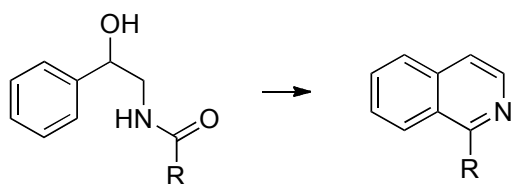
Coupling Reactions

1. c\$1:c4[allHydrogens=1]:c(-C6[allHydrogens=2]-C7[allHydrogens=2]-N2-C3(=O5)-C):c:c:c\$1 >> Bischler-Napieralski >> C6-C7-N2=C3-c4



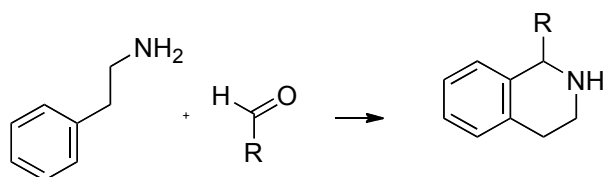
c1cc(CCNC(=O)C)ccc1

2. c\$1:c4[allHydrogens=1]:c(-C6(-O8[allHydrogens=1])-C7[allHydrogens=2]-N2-C3(=O5)-C):c:c:c\$1 >> Pictet-Gams >> C6=C7-N2=C3-c4



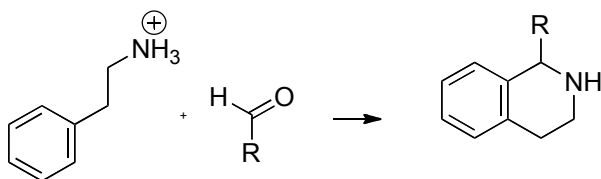
c1cc(C(O)CNC(=O)C)ccc1

3. c2[allHydrogens=1]:c(-C5[sp3 & !ring]-C6[sp3 & !ring]-N7[allHydrogens=2 & charge=0]):c[!bound(-H)] ++ C3[allHydrogens=1](=O4)-C >> Pictet-Spengler (charge 1) >> C5-C6-N7-C3-c2



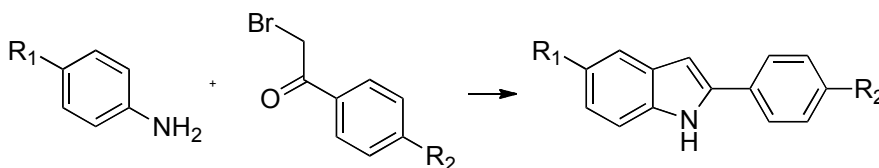
c1cc(CCN)c(C)cc1 + CC(=O)

c2[allHydrogens=1]:c(-C5[sp3 & !ring]-C6[sp3 & !ring]-N7[allHydrogens=3 & charge=1]):c[!bound(-H)]
 ++ C3[allHydrogens=1](=O4)-C >> Pictet-Spengler (charge 2) >> C5-C6-N7-C3-c2



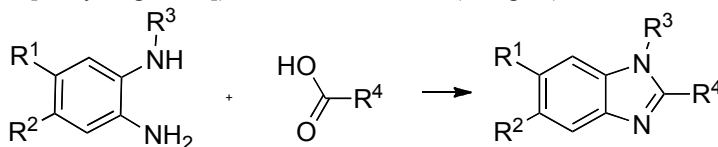
c1cc(CC[NH3+])c(C)cc1 + CC(=O)

4. c\$1:c:c:c(-N3[allHydrogens=2]):c2[allHydrogens=1]:c\$:1 ++ c-C4(=O5)-C6-*7[Cl|Br] >> Bischler Indole
 >> N3-C4=C6-C2



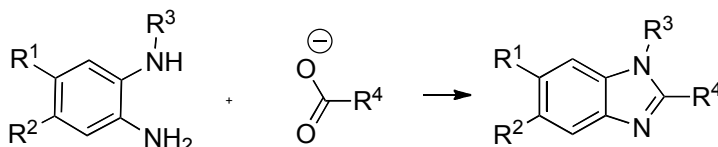
c1c(N)cccc1 + c1c(C(=O)CBr)cccc1

5. c\$1:c[allHydrogens=1]:c(-N7[allHydrogens=2]):c(-N8[bound(-H)]):c[allHydrogens=1]:c\$:1 ++ C3(=O4)(-O5[allHydrogens=1])-C >> Benzimidazol (charge 1) >> N7=C3-N8



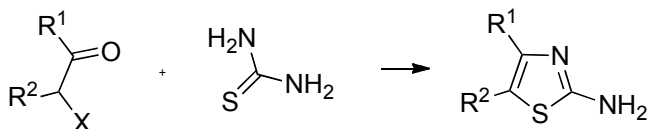
c1cccc(N)c1NC + CC(=O)O

c\$1:c[allHydrogens=1]:c(-N7[allHydrogens=2]):c(-N8[bound(-H)]):c[allHydrogens=1]:c\$:1 ++ C3(=O4)(-O5[charge=-1])-C >> Benzimidazol (charge 2) >> N7=C3-N8



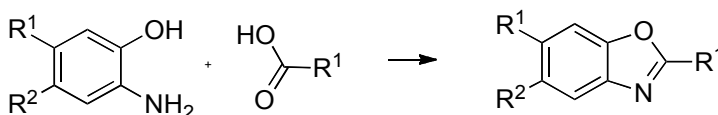
c1cccc(N)c1NC + CC(=O)[O-]

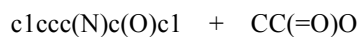
6. C-C4(-*5[Cl|Br])-C6(=O7)-C >> Aminothiazol >> C4\$8-S-C(-N)=N-C6\$=8



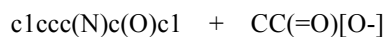
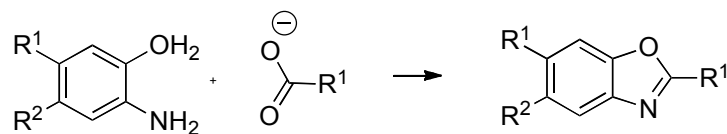
CC(=O)C(Br)C

7. c\$1:c[allHydrogens=1]:c(-O7[allHydrogens=1]):c(-N8[allHydrogens=2]):c[allHydrogens=1]:c\$:1 ++ C3(=O4)(-O5[allHydrogens=1])-C >> Benzoxazol (charge 1) >> O7-C3=N8

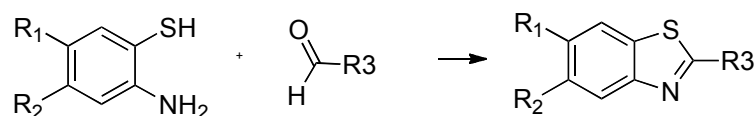




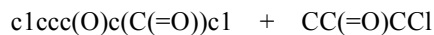
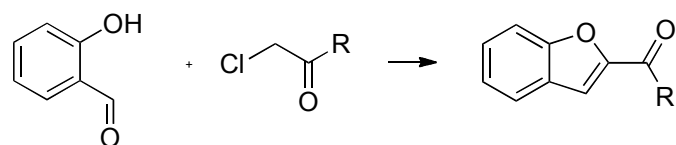
c\$1:c[allHydrogens=1]:c(-O7[allHydrogens=1]):c(-N8[allHydrogens=2]):c[allHydrogens=1]:c\$:1 ++ C3(=O4)(-O5[charge=-1])-C >> Benzoxazol (charge 2) >> O7-C3=N8



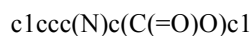
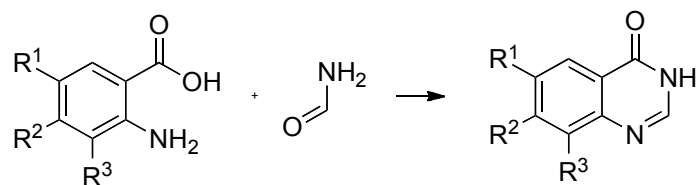
8. c\$1:c:c(-S7[allHydrogens=1]):c(-N8[allHydrogens=2]):c:c\$:1 ++ C3[allHydrogens=1](=O4)-c >> Benzothiazol >> S7-C3=N8



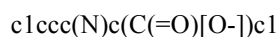
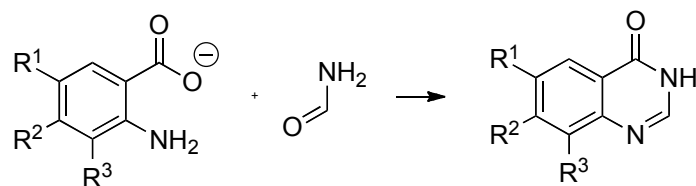
9. c\$1:c:c(-O7[allHydrogens=1]):c(-C8[allHydrogens=1]=O9):c:c\$:1 ++ *5[Cl|Br]-C3[allHydrogens=2]-C(=O)-C >> Rap-Stoermer >> O7-C3=C8



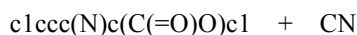
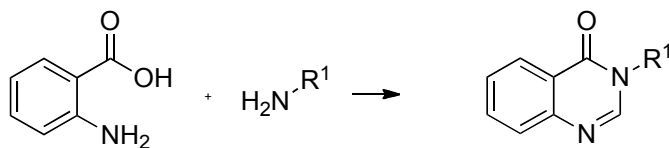
10. c\$1:c:c(-N1[allHydrogens=2]):c(-C2(=O)-O4[allHydrogens=1]):c:c\$:1 >> Niementowski (charge 1) >> C2-N-C=N1



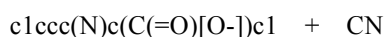
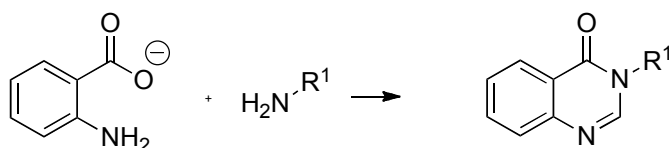
c\$1:c:c(-N1[allHydrogens=2]):c(-C2(=O)-O4[charge=-1]):c:c\$:1 >> classical Niementowski (charge 2) >> C2-N-C=N1



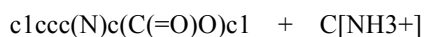
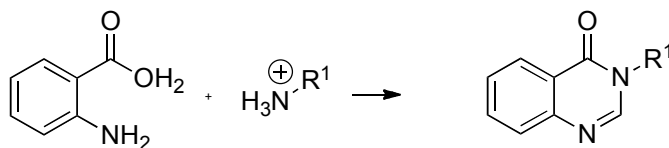
11. c[allHydrogens=1]\$1:c[allHydrogens=1]:c(-N2[allHydrogens=2]):c(-C5(=O)-O4[allHydrogens=1]):c[allHydrogens=1]:c[allHydrogens=1]\$:1 ++ C-N3[allHydrogens=2 & charge=0] >> Quinazolinone (Ladung 1) >> N2=C-N3-C5



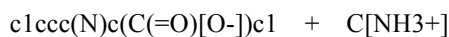
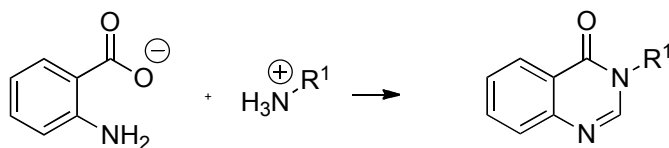
- c[allHydrogens=1]\$1:c[allHydrogens=1]:c(-N2[allHydrogens=2]):c(-C5(=O)-O4[charge=-1]):c[allHydrogens=1]:c[allHydrogens=1]\$:1 ++ C-N3[allHydrogens=2 & charge=0] >> Quinazolinone (Ladung 2) >> N2=C-N3-C5



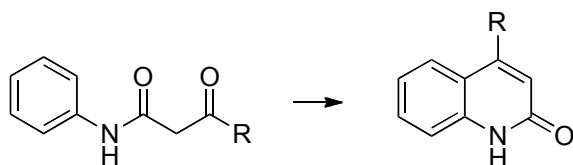
- c[allHydrogens=1]\$1:c[allHydrogens=1]:c(-N2[allHydrogens=2]):c(-C5(=O)-O4[allHydrogens=1]):c[allHydrogens=1]:c[allHydrogens=1]\$:1 ++ C-N3[allHydrogens=3 & charge=1] >> Quinazolinone (Ladung 3) >> N2=C-N3-C5



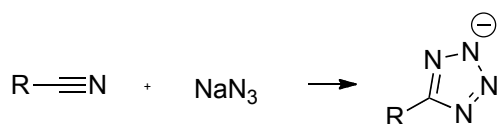
- c[allHydrogens=1]\$1:c[allHydrogens=1]:c(-N2[allHydrogens=2]):c(-C5(=O)-O4[charge=-1]):c[allHydrogens=1]:c[allHydrogens=1]\$:1 ++ C-N3[allHydrogens=3 & charge=1] >> Quinazolinone (Ladung 4) >> N2=C-N3-C5



12. c\$1:c:c(-N-C(=O)-C1[allHydrogens=2 & !ring]-C2[!ring](=O3)-C):c4:c:c\$1 >> Chinolin-2-one intramol. >> C1=C2-c4

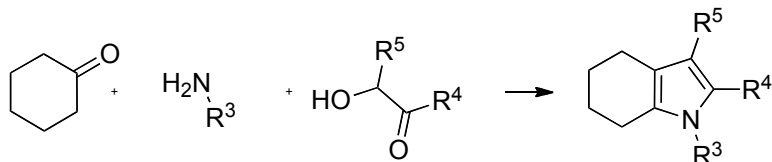


13. C-C1#N2 >> Tetrazol >> C1\$1=N2-N[charge=-1]-N=N\$-1



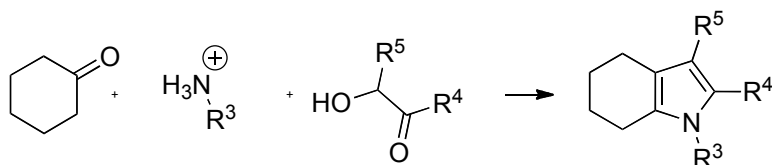
CC#N

14. C-N1[allHydrogens=2 & charge=0] ++ C-C2[allHydrogens=1](-C3(=O4)-C)-O5[allHydrogens=1] >> Tetrahydro-Indole (charge 1) >> C\$1-C-C\$2-N1-C3=C2-C\$=2-C-C\$-1



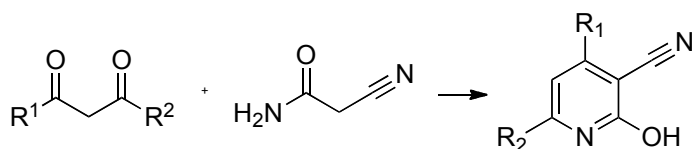
CN + CC(=O)C(O)C

- C-N1[allHydrogens=3 & charge=1] ++ C-C2[allHydrogens=1](-C3(=O4)-C)-O5[allHydrogens=1] >> Tetrahydro-Indole (charge 2) >> C\$1-C-C\$2-N1-C3=C2-C\$=2-C-C\$-1



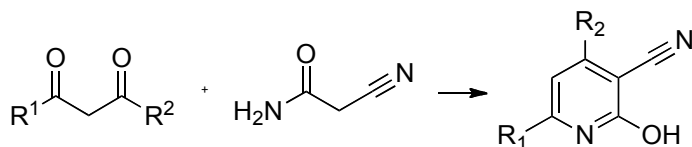
C[NH3+] + CC(=O)C(O)C

15. C1-C2[!ring](=O10)-C3[allHydrogens=2]-C4(=O11)-C5 >> 3-nitrile pyridine (symmetry 1) >> N\$1=C(-O)-C(-C#N)=C2(-C1)-C3=C4\$-1(-C5)



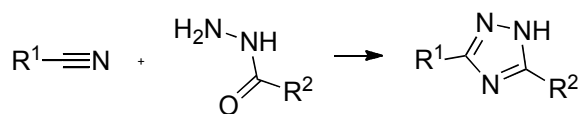
CC(=O)CC(=O)C

- C1-C2[!ring](=O10)-C3[allHydrogens=2]-C4(=O11)-C5 >> 3-nitrile pyridine (symmetry 2) >> N\$1=C(-O)-C(-C#N)=C2(-C5)-C3=C4\$-1(-C1) CC(=O)CC(=O)C



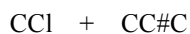
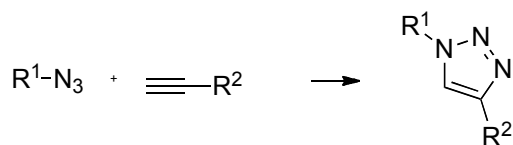
CC(=O)CC(=O)C

16. c-C1#N2[allHydrogens=0] ++ N3[allHydrogens=2]-N6[allHydrogens=1]-C4(=O5)-c >> Triazole >> C1\$8=N3-N6-C4=N2\$-8

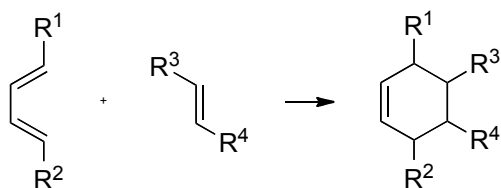


c1ccccc1C#N + NNC(=O)c1ccccc1

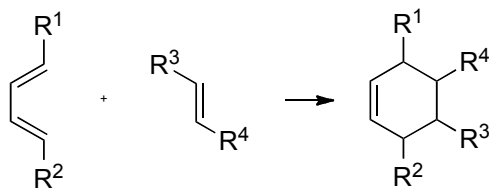
17. C1[sp3]-*2[Cl|Br|I] ++ C3[allHydrogens=1]#C4-C >> Huisgen 1-3 dipolar (azid_in_situ) >> C1-N\$1-N=N-C4=C3[bound(-H)]\$-1



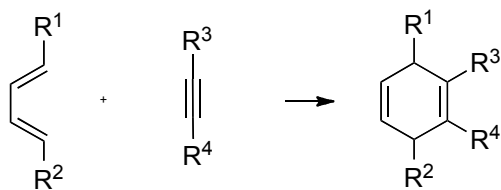
18. C1[!aromatic]=C2[!aromatic]-C3[!aromatic]=C4[!aromatic] ++ C5[!aromatic]=C6[!aromatic] >> Diels-Alder (symmetry 1) >> C1\$1-C2=C3-C4-C5-C6\$-1



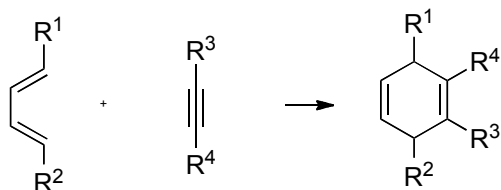
- C1[!aromatic]=C2[!aromatic]-C3[!aromatic]=C4[!aromatic] ++ C5[!aromatic]=C6[!aromatic] >> Diels-Alder (Symmetrie 2) >> C1\$1-C2=C3-C4-C6-C5\$-1



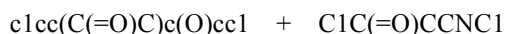
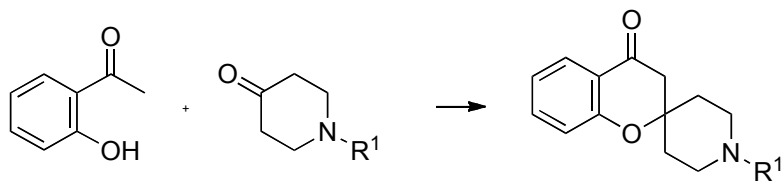
19. C1[!aromatic]=C2[!aromatic & !ring]-C3[!aromatic]=C4[!aromatic] ++ C5#C6 >> Diels-Alder Alkine (symmetry 1) >> C1[!aromatic&!sp2]\$1-C2[!aromatic]=C3[!aromatic]-C4[!aromatic&!sp2]-C5=C6\$-1



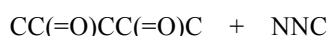
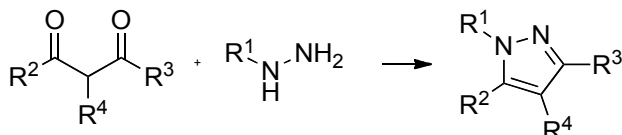
- C1[!aromatic]=C2[!aromatic]-C3[!aromatic]=C4[!aromatic] ++ C5#C6 >> Diels-Alder Alkine (symmetry 2) >> C1[!aromatic&!sp2]\$1-C2[!aromatic]=C3[!aromatic]-C4[!aromatic&!sp2]-C6=C5\$-1



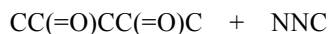
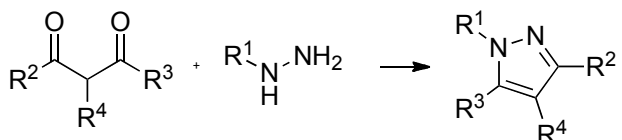
20. c(-O1[allHydrogens=1]):c(-C(=O)-C2[allHydrogens=3]) ++ C[sp3]\$1-C3(=O4)-C[sp3]-C[sp3]-N-C[sp3]\$-1 >> Spiro-piperidine >> O1-C3-C2



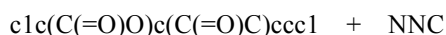
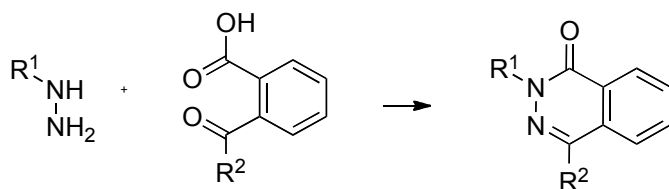
21. C-C1[!ring](=O6)-C2-C3(=O7)-C ++ C-N4[allHydrogens=1]-N5[allHydrogens=2] >> Pyrazole (symmetry 1) >> C1\$1-N4-N5=C3-C2\$=1



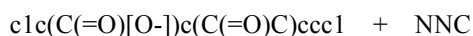
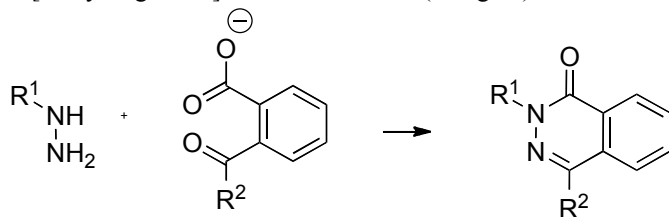
- C-C1[!ring](=O6)-C2-C3(=O7)-C ++ C-N4[allHydrogens=1]-N5[allHydrogens=2] >> Pyrazol (symmetry 2) >> C1\$1-N5-N4=C3-C2\$=1



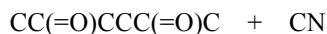
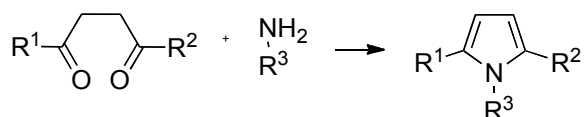
22. C-C1(=O5)-c:c-C2(=O6)-O7[allHydrogens=1] ++ C[!bound(=O) & !bound(=S)]-N3[allHydrogens=1]-N4[allHydrogens=2] >> Phthalazinone (charge 1) >> C2-N3-N4=C1



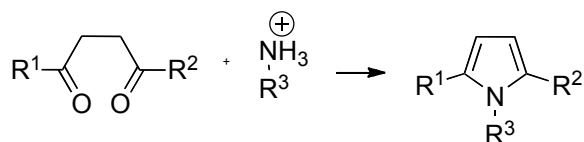
- C-C1(=O5)-c:c-C2(=O6)-O7[charge=-1] ++ C[!bound(=O) & !bound(=S)]-N3[allHydrogens=1]-N4[allHydrogens=2] >> Phthalazinone (charge 2) >> C2-N3-N4=C1



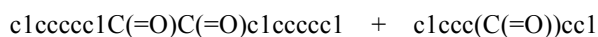
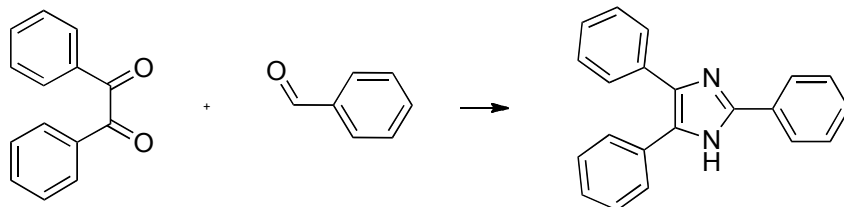
23. C-C1[!ring](=O7)-C4[!aromatic & bound(-H)]-C5[!aromatic & bound(-H)]-C2(=O6)-C ++ C[!bound(=O)]-N3[allHydrogens=2 & charge=0] >> Paal-Knorr pyrrole (charge 1) >> C1\$1-N3-C2=C5-C4\$=1



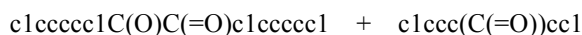
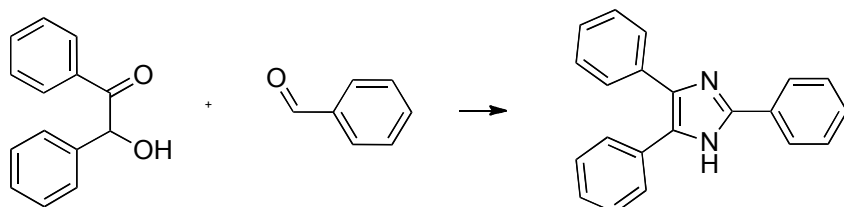
C-C1[!ring](=O7)-C4[!aromatic & bound(-H)]-C5[!aromatic & bound(-H)]-C2(=O6)-C ++ C[!bound(=O)]-N3[charge=1 & allHydrogens=3] >> Paal-Knorr pyrrole (charge 2) >> C1\$1-N3-C2=C5-C4\$=1



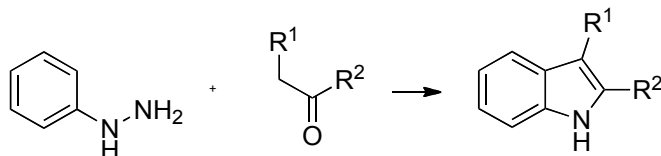
24. c-C1(=O4)-C2(=O5)-c ++ C3[allHydrogens=1](=O6)-c\$1:c:c:c:c\$1 >> Triaryl-imidazol (1,2 diketone)
>> C1\$1-N-C3=N-C2\$=1



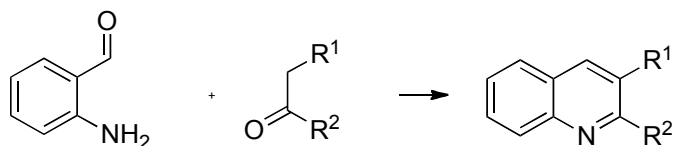
25. c-C1(=O4)-C2(-O5[allHydrogens=1])-c ++ C3[allHydrogens=1](=O6)-c\$1:c:c:c:c\$1 >> Triaryl-imidazol (alpha hydroxy-ketone)
>> C1\$1-N-C3=N-C2\$=1

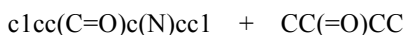


26. c\$1:c4[allHydrogens=1]:c(-N5[allHydrogens=1]-N6[allHydrogens=2]):c:c\$c\$1 ++ C-C1(=O2)-C3[allHydrogens=2]-C >> Fischer indole >> N5=C1=C3-c4

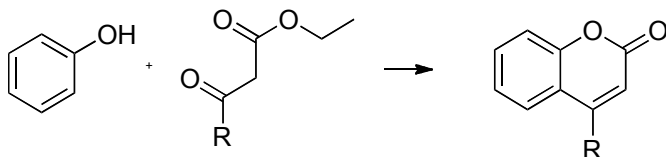


27. c\$1:c(-C4[allHydrogens=1](=O7)):c(-N5[allHydrogens=2]):c:c\$c\$1 ++ C-C1(=O2)-C3[allHydrogens=2]-C >> Friedlaender chinoline >> N5=C1-C3=C4

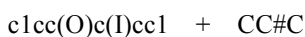
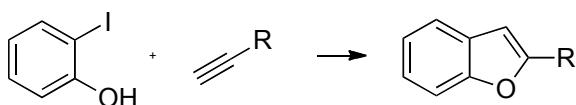




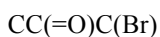
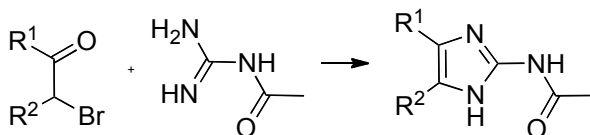
28. c\$1:c1[allHydrogens=1]:c(-O2[allHydrogens=1]):c:c:c\$1 ++ C-C3(=O4)-C5[allHydrogens=2 & !ring]-C6(=O)-O7-C[allHydrogens=2]-C[allHydrogens=3] >> Pechmann coumarine >> c1-C3=C5-C6-O2



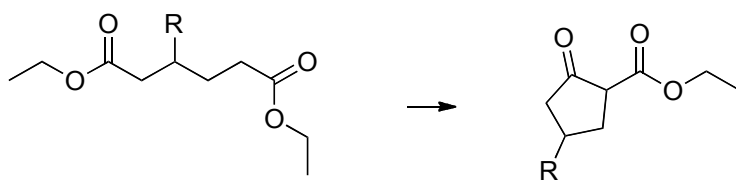
29. c\$1:c(-O1[allHydrogens=1]):c2(-I5):c:c:c\$1 ++ C3[allHydrogens=1]#C4-C >> Benzofuran >> O1-C4=C3-c2



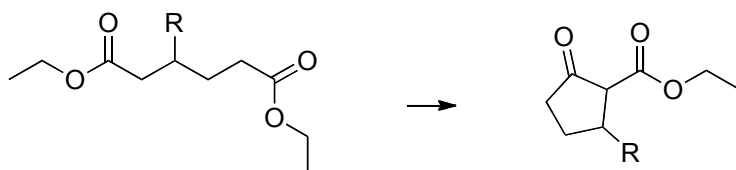
30. C-C1(=O2)-C3[bound(-H)](-Br4) >> Imidazol-Acetamid >> C1\$1=C3-N-C(-N-C(=O)-C)=N\$-1



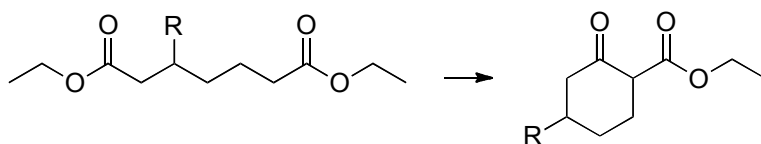
31. C[allHydrogens=3]-C[allHydrogens=2]-O1-C2[!ring](=O)-C[!aromatic]-C[!aromatic]-C[!aromatic]-C3[bound(-H) & !aromatic]-C[!ring](=O)-O1-C[allHydrogens=2]-C[allHydrogens=3] >> Dieckmann 5-ring (symmetry 1) >> C2[ring]-C3



- C[allHydrogens=3]-C[allHydrogens=2]-O-C[!ring](=O)-C2[bound(-H) & !aromatic]-C[!aromatic]-C[!aromatic]-C3[!ring](=O)-O1-C[allHydrogens=2]-C[allHydrogens=3] >> Dieckmann 5-ring (symmetry 2) >> C2-C3[ring]

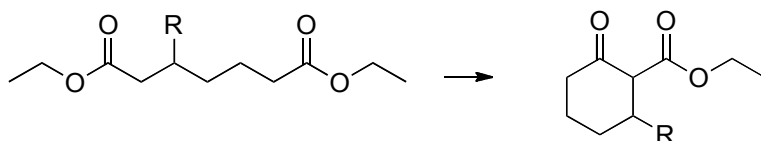


32. C[allHydrogens=3]-C[allHydrogens=2]-O1-C2[!ring](=O)-C[!aromatic]-C[!aromatic]-C[!aromatic]-C[!aromatic]-C3[bound(-H) & !aromatic]-C[!ring](=O)-O-C[allHydrogens=2]-C[allHydrogens=3] >> Dieckmann 6-Ring (symmetry 1) >> C2-C3



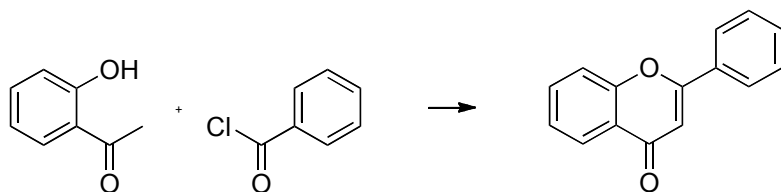
CCOC(=O)CCCCC(=O)OCC

- C[allHydrogens=3]-C[allHydrogens=2]-O-C[!ring](=O)-C2[bound(-H) & !aromatic]-C[!aromatic]-C[!aromatic]-C3[!ring](=O)-O1-C[allHydrogens=2]-C[allHydrogens=3] >> Dieckmann 6-Ring (symmetry 2) >> C2-C3



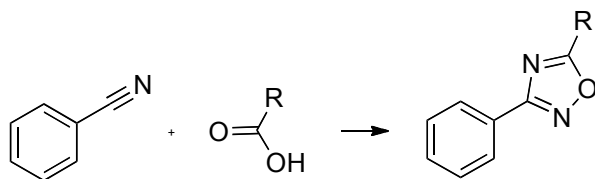
CCOC(=O)CCCCC(=O)OCC

33. c\$1:c:c(-O1[allHydrogens=1]):c(-C(=O)-C2[allHydrogens=3]):c:c\$1 ++ c\$1:c:c(-C3(=O4)-Cl5):c[bound(-H)]:c\$1 >> Flavone >> C2=C3-O1



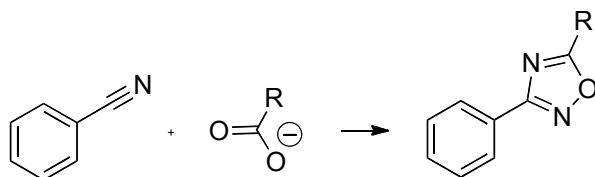
c1cc(O)c(C(=O)C)cc1 + c1ccc(C(=O)Cl)cc1

34. c-C1#N2 ++ C3[allHydrogens=0](=O4)-O5[allHydrogens=1] >> Oxadiazole (charge 1) >> C1\$1=N-O-C3=N2\$-1



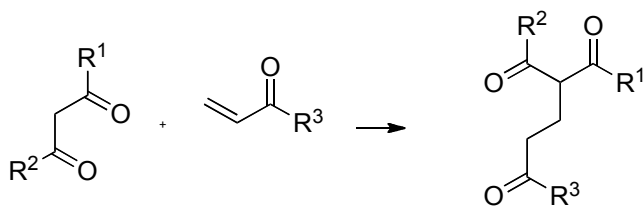
c1cc(C#N)ccc1 + CC(=O)O

- c-C1#N2 ++ C3[allHydrogens=0](=O4)-O5[charge=-1] >> Oxadiazole (charge 2) >> C1\$1=N-O-C3=N2\$-1

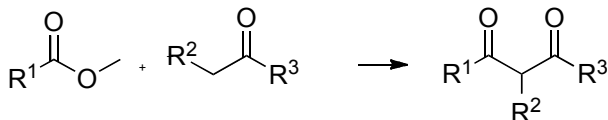


c1cc(C#N)ccc1 + CC(=O)[O-]

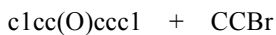
35. C(=O)-*[{O & allHydrogens=0} | C]-C1[allHydrogens=2]-C(=O)-*[{O & allHydrogens=0} | C] ++ C2[!aromatic]=C3[!aromatic]-C4(=O)-C >> Michael addition >> C1-C2-C3-C4



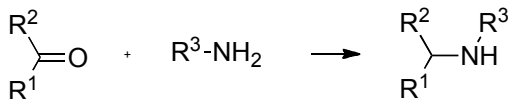
36. `*[C & !bound(-H)]{O & !bound(-H) & charge=0}]-C1[!ring](=O)-O2-C[!ring] ++ C3[allHydrogens=2]-C(=O)-*[C & !bound(-H)]{O & !bound(-H) & charge=0}] >> crossed Claisen >> C3-C1`



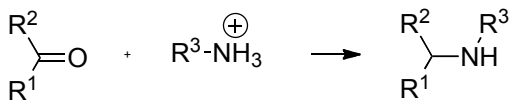
37. `c-O1[allHydrogens=1] ++ C2[allHydrogens=2]-*3[I|Br|Cl] >> Williamson ether >> O1-C2`



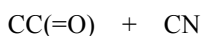
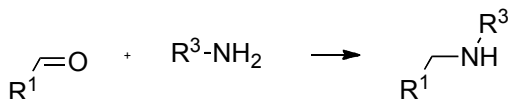
38. `C-C1(=O2)-C ++ N3[allHydrogens=2 & charge=0]-C[!bound(=O) & !bound(=N)] >> red. amination (one step), ketone, prim. amine (charge 1) >> C1-N3[charge=1]`



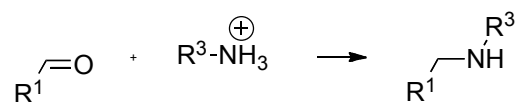
`C-C1(=O2)-C ++ N3[allHydrogens=3 & charge=1]-C[!bound(=O) & !bound(=N)] >> red. amination ketone, prim. amine (charge 2) >> C1-N3[charge=1]`



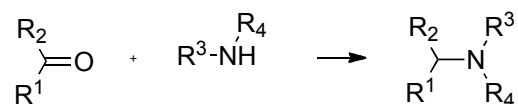
39. `C-C1[bound(-H)](=O2) ++ N3[allHydrogens=2 & charge=0]-C[!bound(=O) & !bound(=N)] >> red. amination, aldehyde, prim. amine (charge 1) >> C1-N3[charge=1]`



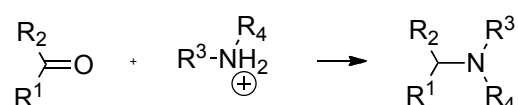
`C-C1[bound(-H)](=O2) ++ N3[allHydrogens=3 & charge=1]-C[!bound(=O) & !bound(=N)] >> red. amination, aldehyde, prim. amine (charge 2) >> C1-N3[charge=1]`



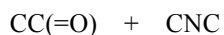
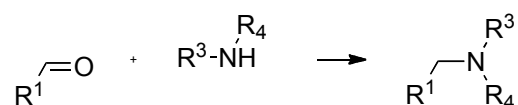
40. C-C1(=O2)-C ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=1 & charge=0 & !aromatic]-C[!bound(=O) & !bound(=N)] >> red. amination, ketone, sec. amine (charge 1) >> C1-N3[charge=1]



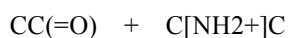
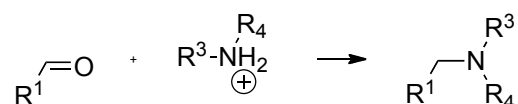
- C-C1(=O2)-C ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=2 & charge=1 & !aromatic]-C[!bound(=O) & !bound(=N)] >> red. amination, ketone, sec. amine (charge 2) >> C1-N3[charge=1]



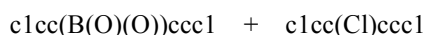
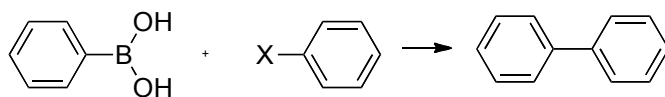
41. C-C1[bound(-H)](=O2) ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=1 & charge=0 & !aromatic]-C[!bound(=O) & !bound(=N)] >> red. amination, aldehyde, sec. amine (charge 1) >> C1-N3[charge=1]



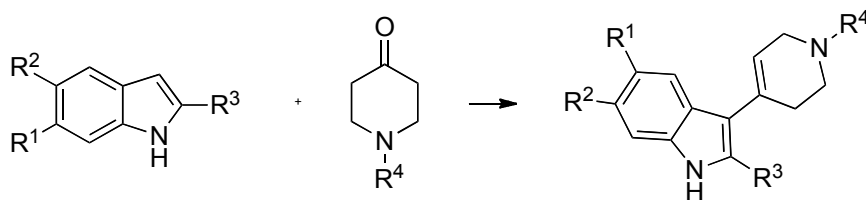
- C-C1[bound(-H)](=O2) ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=2 & charge=1 & !aromatic]-C[!bound(=O) & !bound(=N)] >> red. amination, aldehyde, sec. amine (charge 2) >> C1-N3[charge=1]



42. C1[sp2]-B3(-O)-O ++ C2[sp2 & !bound(=O)]-*4[Cl|Br|I]>> Suzuki >> C1-C2



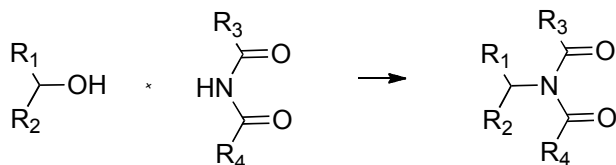
43. c[allHydrogens=1]\$1:c:c[allHydrogens=1]:c\$2-N[allHydrogens=1]-C=C5[allHydrogens=1]-c\$:1\$:2 ++ C[allHydrogens=2]\$3-N-C[allHydrogens=2]-C[allHydrogens=2]-C4(=O7)-C6[allHydrogens=2]\$-3 >> Piperidine+Indole >> C4(=C6)-C5



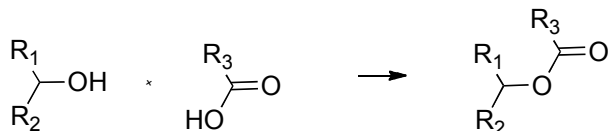
44. C1[!bound(=O)]-*2[Br|Cl|I] ++ *5[Br|Cl]-C4[allHydrogens=2]-C[allHydrogens=2] >> Negishi >> C1-C4



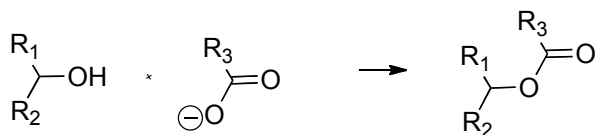
45. C1[bound(-H) & !bound(=O)]-O2[allHydrogens=1] ++ C(=O)-N3[allHydrogens=1]-C(=O) >> Mitsunobu (imide) >> C1-N3



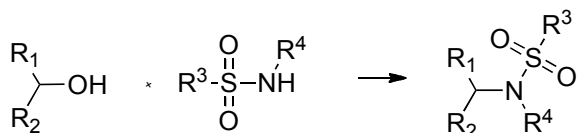
46. C1[bound(-H) & !bound(=O)]-O2[allHydrogens=1] ++ C-C(=O)-O3[allHydrogens=1] >> Mitsunobu Carbonsäure (carbon acid, charge 1) >> C1-O3



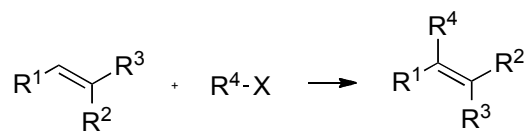
C1[bound(-H) & !bound(=O)]-O2[allHydrogens=1] ++ C-C(=O)-O3[charge=-1] >> Mitsunobu (carbon acid, charge 2) >> C1-O3 CC(O)CCC(=O)[O-]



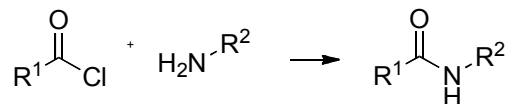
47. C1[bound(-H) & !bound(=O)]-O2[allHydrogens=1] ++ C-N3[bound(-H)]-S(=O)(=O)-C >> Mitsunobu Sulfonic amide >> C1-N3



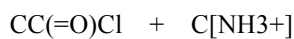
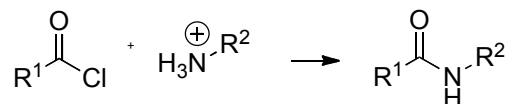
48. C1[!bound(=O)]-*3[Br | I | Cl] ++ C-C2[allHydrogens=1 & !aromatic]=C[!aromatic](-C)-C >> Heck >> C1-C2



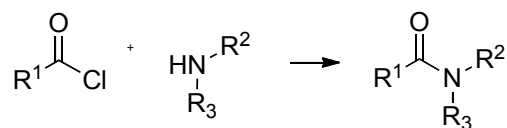
49. C-C1(=O)-Cl2 ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=2 & charge=0] >> Amide, prim. amine (charge 1) >> C1-N3



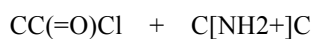
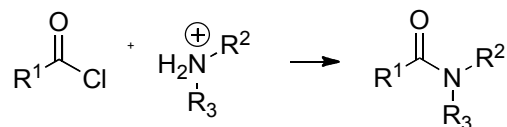
- C-C1(=O)-Cl2 ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=3 & charge=1] >> Amide, prim. amine (charge 2) >> C1-N3



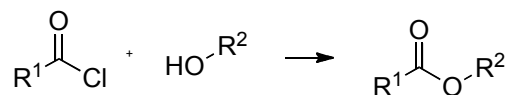
50. C-C1(=O)-Cl2 ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=1 & charge=0]-C[!bound(=O) & !bound(=N)] >> Amide, sec. amine (charge 1) >> C1-N3



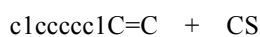
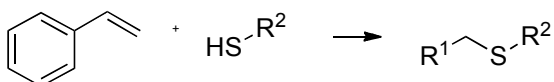
- C-C1(=O)-Cl2 ++ C[!bound(=O) & !bound(=N)]-N3[allHydrogens=2 & charge=1]-C[!bound(=O) & !bound(=N)] >> Amide, sec. amine (charge 2) >> C1-N3



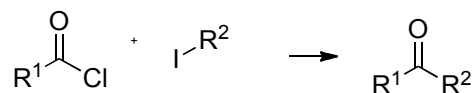
51. C-C1(=O)-Cl2 ++ C[!bound(=O)]-O3[allHydrogens=1] >> Ester >> C1-O3



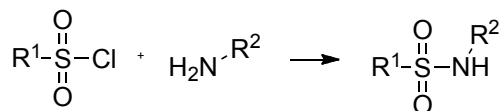
52. c-C1[!aromatic]=C2[!aromatic & allHydrogens=2] ++ S3[allHydrogens=1]-C >> Thioether >> C1-C2-S3



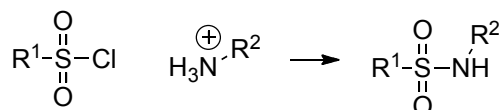
53. C-C1(=O)-Cl2 ++ C3-I4 >> Ketone >> C1-C3



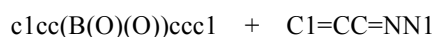
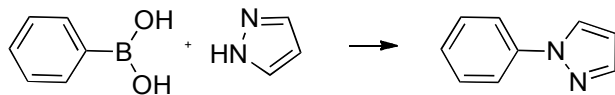
54. C-S1(=O)(=O)-Cl3 ++ N2[allHydrogens=2 & charge=0]-C[!bound(=O) & !bound(=N)] >> Sulfonamid (Ladung 1) >> S1-N2



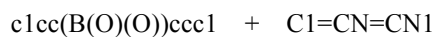
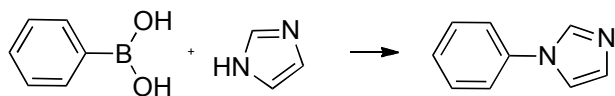
C-S1(=O)(=O)-Cl3 ++ N2[allHydrogens=3 & charge=1]-C[!bound(=O) & !bound(=N)] >> Sulfonamid (Ladung 2) >> S1-N2



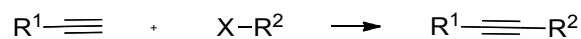
55. c1-B2(-O[allHydrogens=1])(-O[allHydrogens=1]) ++ c\$1:n3[allHydrogens=1]:n:c\$:1 >> Ar-Pyrazole >> c1-N3



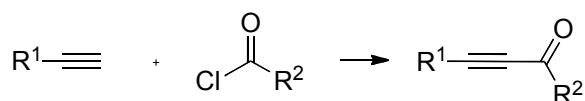
56. c1-B2(-O[allHydrogens=1])(-O[allHydrogens=1]) ++ c\$1:n3[allHydrogens=1]:n:c\$:1 >> Ar-Imidazole >> c1-N3



57. C1[sp3]-*2[Cl|Br|I] ++ C3[allHydrogens=1]#C >> Alkine alkylation >> C1-C3

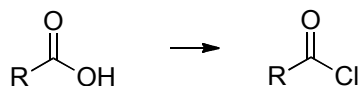


58. C-C2(=O)-Cl4 ++ C3[allHydrogens=1]#C >> Alkine acylation >> C2-C3

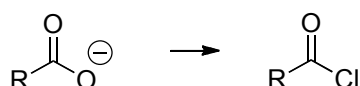


Preprocessing Reactions

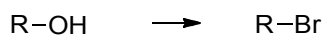
1. C1(=O2)-O3[allHydrogens=1] >> FGI Acyl chloride (charge 1) >> C1(=O2)-Cl



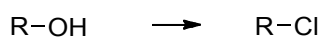
- C1(=O2)-O3[charge=-1] >> FGI acyl Chloride (charge 2) >> C1(=O2)-Cl



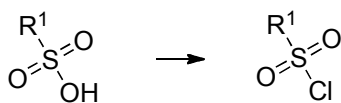
2. C1[aliphatic & !bound(=O) & !bound(=S)]-O2[allHydrogens=1] >> FGI bromination >> C1-Br



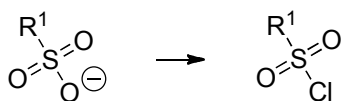
3. C1[aliphatic & !bound(=O) & !bound(=S)]-O2[allHydrogens=1] >> FGI chlorination >> C1-Cl



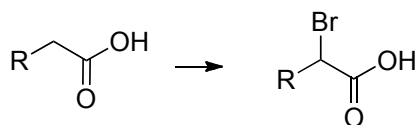
4. C-S1(=O)(=O)-O2[allHydrogens=1] >> FGI sulfonyl chloride (charge 1) >> S1-Cl



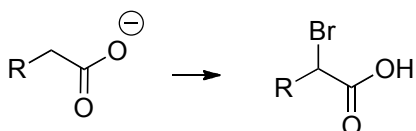
- C-S1(=O)(=O)-O2[charge=-1] >> FGI sulfonyl chloride (charge 2) >> S1-Cl



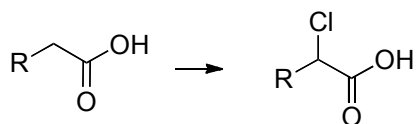
5. C1[!aromatic & allHydrogens=2 & !bound(-Halogen)]-C(=O)-O[allHydrogens=1] >> FGA alpha bromination (charge 1) >> C1-Br



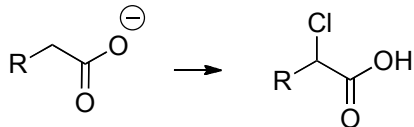
- C1[!aromatic & allHydrogens=2 & !bound(-Halogen)]-C(=O)-O[charge=-1] >> FGA alpha bromination (charge 2) >> C1-Br



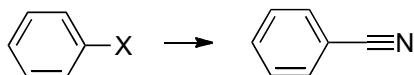
6. C1[!aromatic & allHydrogens=2 & !bound(-Halogen)]-C(=O)-O[allHydrogens=1] >> FGA alpha chlorination (Ladung 1) >> C1-Cl



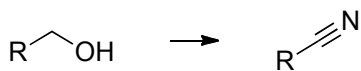
C1[!aromatic & allHydrogens=2 & !bound(-Halogen)]-C(=O)-O[charge=-1] >> FGA alpha chlorination (charge 2) >> C1-Cl



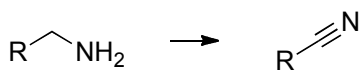
7. c1-*2[Cl|Br] >> FGI Rosenmund-von-Braun >> c1-C#N



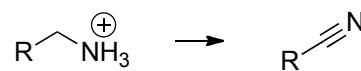
8. C-C1[allHydrogens=2]-O2[allHydrogens=1] >> FGI nitration prim. hydroxy >> C1#N



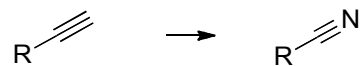
9. C-C1[allHydrogens=2]-N2[allHydrogens=2 & charge=0] >> FGI nitration prim. amine (charge 1) >> C1#N2



C-C1[allHydrogens=2]-N2[allHydrogens=3 & charge=1] >> FGI nitration prim. Aminen (charge 2) >> C1#N2

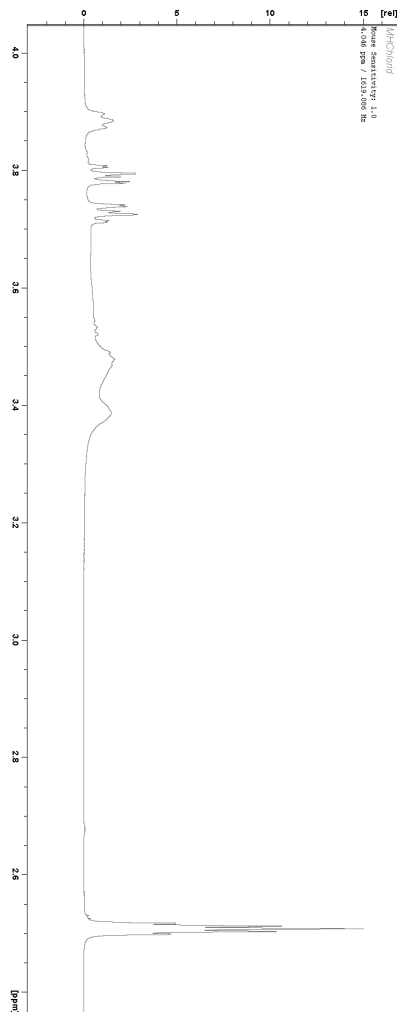


10. C-C1#C2[allHydrogens=1] >> FGI nitration term. alkyne >> C1#N



Analytical Spectra

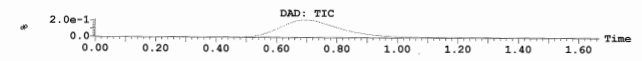
Compound 18: ¹H NMR spectrum



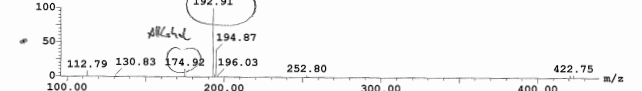
Compound 18: mass spectrum

Openlynx Report - Markus Hartenfeller
 Sample: 1
 File: Schn_MH11-1
 Description:
 Printed: Fri May 21 13:43:17 2010
 Vial: 2.17
 Date: 21-May-2010
 ID: chlroid1412
 Time: 13:40:32
 Page 1

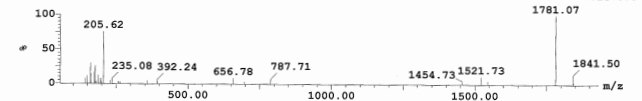
DAD: TIC Smooth (Mn, 2x3) 2.2e+005



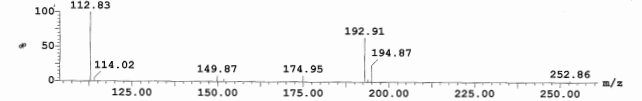
Peak ID 1
 Time 0.70
 Combine (14:18-6:8)
 1:MS ES+
 1.1e+008



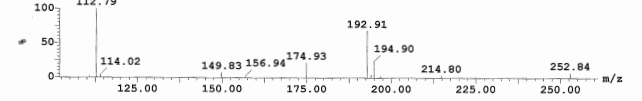
Peak ID 1
 Time 0.70
 Combine (13:18-6:8)
 2:MS ES-
 5.2e+003



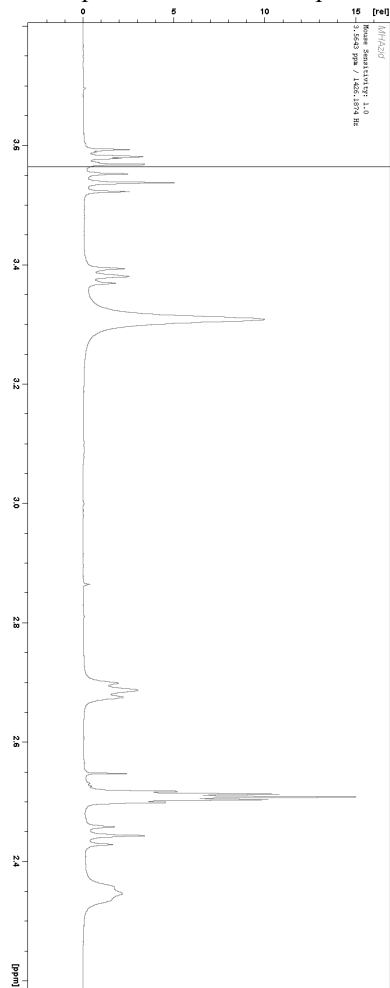
Peak ID 1
 Time 0.70
 Combine (13:17-6:8)
 3:MS ES+
 9.8e+007



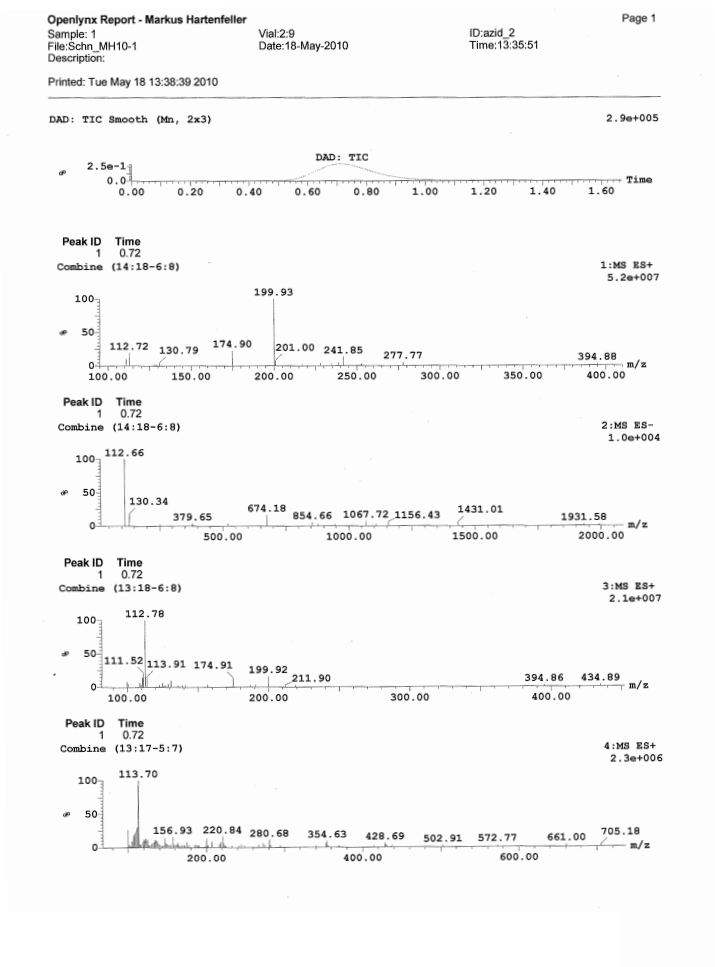
Peak ID 1
 Time 0.70
 Combine (13:17-5:7)
 4:MS ES+
 1.2e+007



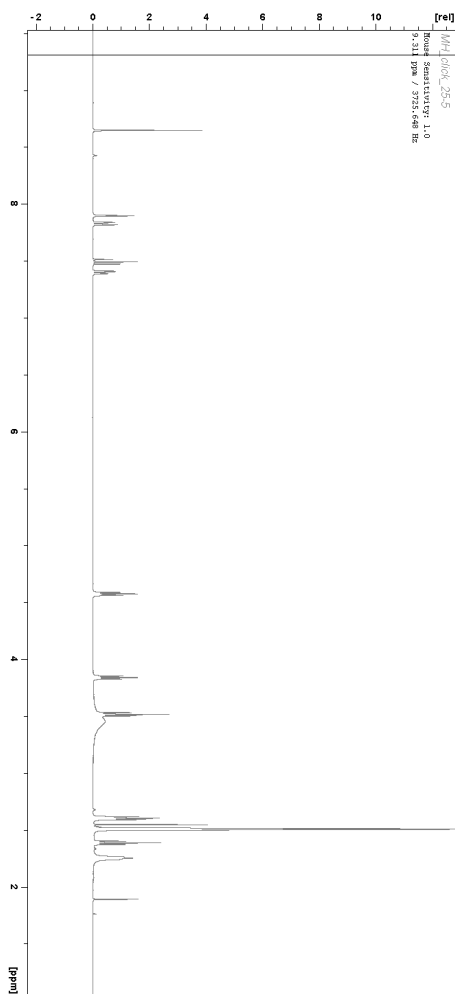
Compound 19: ¹H NMR spectrum



Compound 19: mass spectrum



Compound 15: ¹H NMR spectrum



Compound 15: mass spectrum

Openlynx Report - Markus Hartenfeller

Sample: 1
File: Schn_MH12-1
Description:

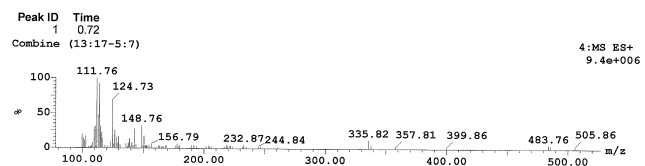
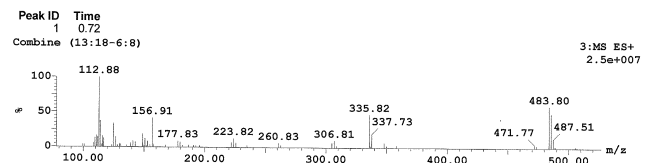
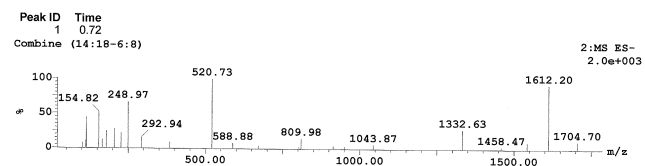
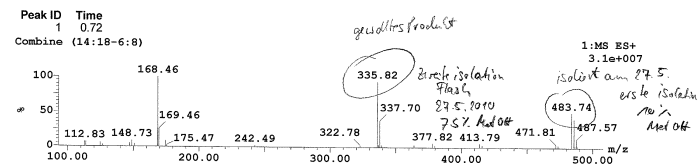
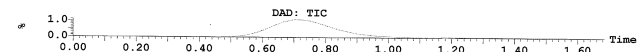
Vial: 120
Date: 27-May-2010

ID: click26-6
Time: 11:51:36

Page 1

Printed: Thu May 27 11:54:29 2010

DAD: TIC Smooth (Mn, 2x3) 1.1e+006

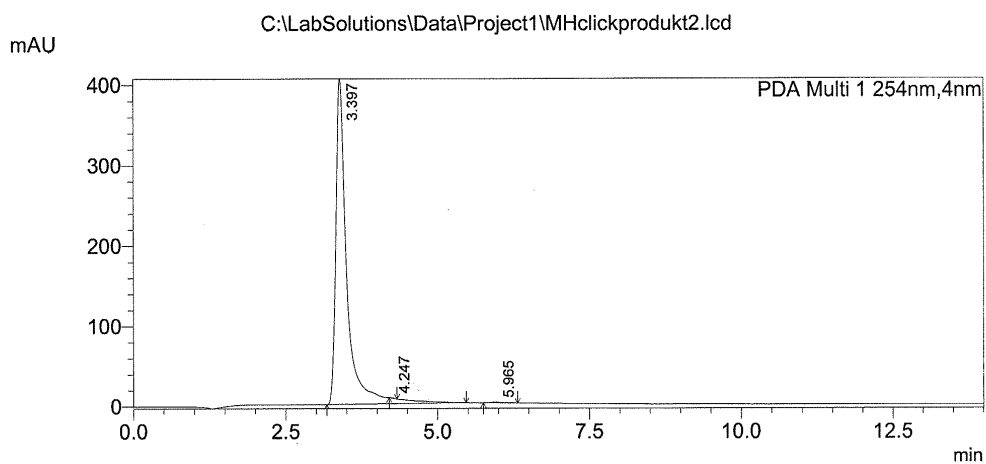


Compound 15: HPLC, UV spectrum

==== Shimadzu LabSolutions Analysis Report ====

C:\LabSolutions\Data\Project1\MHclickprodukt2.lcd
Acquired by : System Administrator
Sample Name : MHclickprodukt
Sample ID :
Tray# : 1
Vial# : 61
Injection Volume : 5
Data File : C:\LabSolutions\Data\Project1\MHclickprodukt2.lcd
Method File : C:\LabSolutions\Data\Project1\HZ142.lcm
Report Format File : C:\LabSolutions\System\DEFAULT.lsr
Month-Day Acquired : 6/2/2010
Month-Day Processed : 6/2/2010

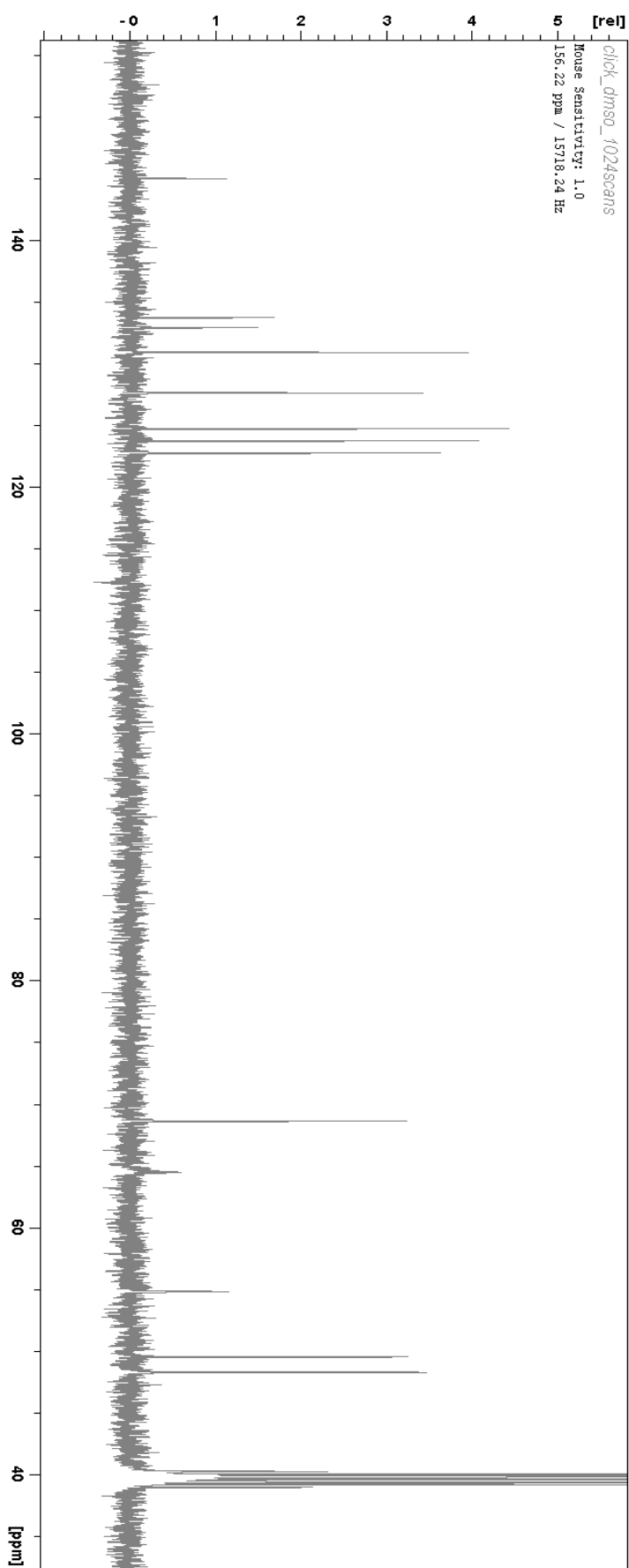
<Chromatogram>



QuantitativeResult

ID#	Name	Ret. Time	Area	Height	Area%
1	RT:3.397	3.397	5164965	404486	99.680
2	RT:4.247	4.247	2285	673	0.044
3	RT:5.965	5.965	14299	909	0.276
Total			5181548	406067	100.000

Compound 15: ^{13}C NMR



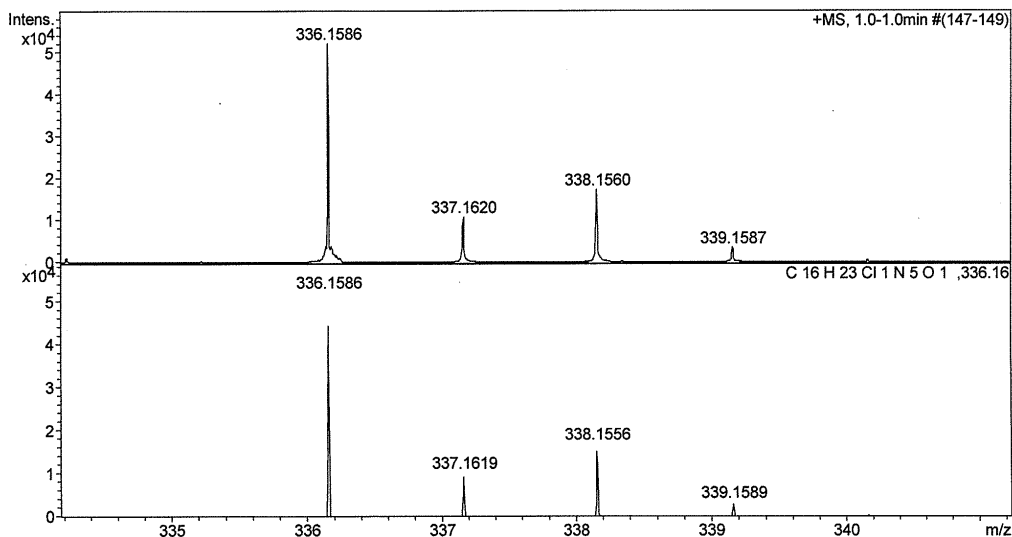
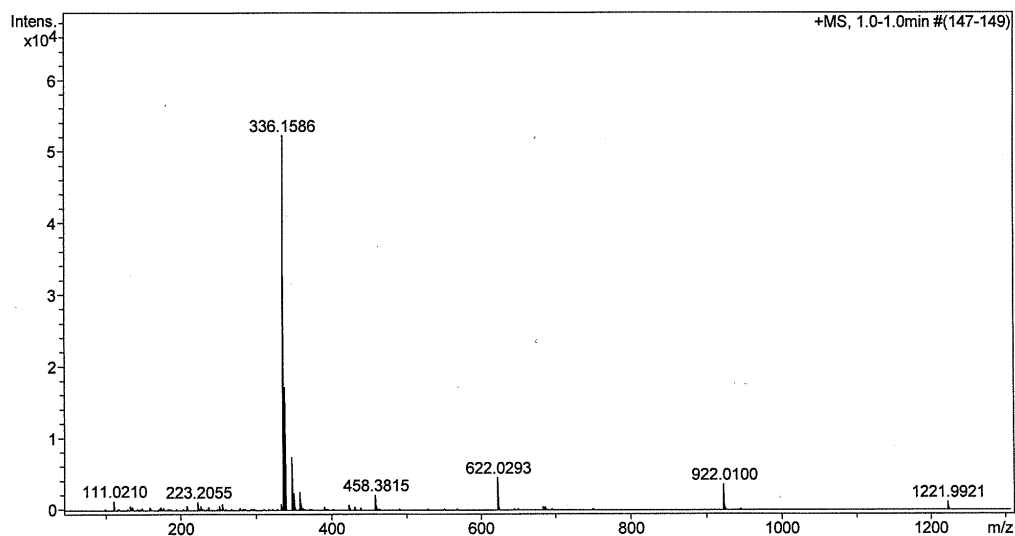
Compound 15: high resolution mass spectrum

MAX04312 Markus Hartenfeller/Schneider - MH1 - DCM/MeOH

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Acquisition Parameter

Method:	ETH_HyStar_HPLC_QTOF_POS_LowMass_Loop-AS.m	Acquisition Date:	15.07.2010 08:18:19		
File Name:	D:\Data\max043xx\MAX04312.d	Operator:	Louis Bertschi		
Source Type	ESI	Ion Polarity	Positive	Set Nebulizer	1.6 Bar
Focus	Active	Set Capillary	4500 V	Set Dry Heater	200 °C
Scan Begin	50 m/z	Set End Plate Offset	-500 V	Set Dry Gas	8.0 l/min
Scan End	1300 m/z	Set Collision Cell RF	200.0 Vpp	Set Divert Valve	Source



Acknowledgement

I want to thank my professor Gisbert Schneider for giving me the opportunity to do my Ph.D. research in his group. Your enthusiasm for scientific research is inspiring, and I am happy that our cooperation will continue after I have finished my PH.D. work in your group.

I am grateful to everyone in our group for creating a stimulating but unstressed climate to work and live in. You made coming to work a pleasure and motivated me. Special thanks go to Matthias Wirth, Volker Hähnke and Alexander Klenner, who have accompanied my way through academia from the very first day and have become good friends.

A lot of people helped me at several points of my work and provided me with support. Here, I would like to name Ewgenij Proschak and Felix Reisen, who implemented MQL and Reaction-MQL, which represent important building blocks of my work. Matthias Rupp implemented the ISOAK scoring function and always had an open ear for discussions on adapting it to my project. Franka Klingler spent her internship on supporting me with the collection of the reaction library.

I would also like to express my gratitude to Merz Pharmaceuticals GmbH for granting a Ph.D. scholarship. Especially Udo Meyer supported my work with advice on compiling the reaction library and on the concept of compound construction.

A special thank also goes to the group of Prof. Holger Stark, who gave me the opportunity to do the main part of the chemical synthesis in their labs. In particular, I would like to express my thankfulness to Kerstin Sander and Miriam Walter, who ventured to give a computational chemist the chance to put his hands on real chemistry. Heiko Zettl continued this job at our own lab at the ETH.

Finally, I want to say thank you to my friends and my family. You created the net that caught me whenever something made me stumble. I am grateful to my parents for their unconditional support in every possible way.

Thank you!

Curriculum Vitae

Dipl. Bioinf. Markus Hartenfeller

Date of birth June 9th, 1981
Place of birth Friedberg, Germany

Education

- Since 08/2010 **Presidential Postdoc** at the Novartis Institutes for Biomedical Research (NIBR) at Novartis AG, Basel, Switzerland, in the group of Dr. Edgar Jacoby under supervision of Dr. Steffen Renner (Novartis), Prof. Dr. Gisbert Schneider (ETH Zürich) and Prof. Dr. Karl-Heinz Altmann (ETH Zürich).
- Since 01/2010 Finalization of **Ph.D. studies** at the Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, with Prof. Dr. Gisbert Schneider.
- 01/2008 – 12/2009 **Ph.D. studies** at the Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University, Frankfurt am Main, Germany, with Prof. Dr. Gisbert Schneider, supported by a **full Ph.D. scholarship** awarded by Merz Pharmaceuticals GmbH.
- Development of a software tool for computer-based *de novo* design of novel lead structure candidates based on established chemical reactions and available molecular building blocks.
 - Focal point of the software: Suggestion of possible synthesis routes for each designed compound.
 - Practical evaluation of suggested synthesis routes by chemical synthesis of selected compounds in a case study for human histamine receptor 4 in the laboratory of Prof Dr. Stark, Johann Wolfgang Goethe-University Frankfurt am Main, Germany.
- 05/2007 – 11/2007 **Diploma thesis** at the Institute for Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University Frankfurt am Main, Germany, with Prof. Dr. Gisbert Schneider.

Title: "Development of a method for combinatorial molecule design".

- Automated design of target structure-tailored ligand candidate libraries based on known privileged scaffolds.
- Design guided by nature-inspired Particle Swarm Optimization.
- Written in Java, including a graphical user interface.

05/2006 – 07/2006

Research internship at Merz Pharmaceuticals GmbH, Frankfurt am Main, Germany, department of Chemical R & D, with Dr. Tanja Weil.

- Development of a three-dimensional QSAR model (CoMFA) for the prediction of binding affinities of ligand candidates for mGluR5 (metabotropic glutamate receptor, subtype 5).
- Implementation of a software tool for calculation of Kolmogorov-Smirnov statistics to guide selection of QSAR descriptors.

10/2002 – 11/2007

Studied Bioinformatics at Johann Wolfgang Goethe-University, Frankfurt am Main, Germany.

University degree: **Diploma in Bioinformatics**.

Average Grade: **1.0** (on a scale between 4.3 and 1.0, with 1.0 being best).

09/2001 – 06/2002

Community service at youth department of city of Niederdorfelden, Germany.

07/1992 – 05/2001

Albert-Einstein **secondary school**, Maintal, Germany

School diploma: University entrance diploma.

Average grade: **1.6** (on a scale between 4.0 and 1.0, with 1.0 being best).