J Comput Aided Mol Des (2008) 22:59–68
DOI 10.1007/s10822-007-9157-4

The concept of template-based de novo design from drug-derived molecular fragments and its application to TAR RNA

Andreas Schüller · Marcel Suhartono · Uli Fechner · Yusuf Tanrikulu · Sven Breitung · Ute Scheffer · Michael W. Göbel · Gisbert Schneider

Received: 3 October 2007 / Accepted: 19 November 2007 / Published online: 7 December 2007

Abstract Principles of fragment-based molecular design are presented and discussed in the context of de novo drug design. The underlying idea is to dissect known drug molecules in fragments by straightforward *pseudo-retro*-synthesis. The resulting building blocks are then used for automated assembly of new molecules. A particular question has been whether this approach is actually able to perform scaffold-hopping. A prospective case study illustrates the usefulness of fragment-based de novo design for finding new scaffolds. We were able to identify a novel ligand disrupting the interaction between the Tat peptide and TAR RNA, which is part of the human immunodeficiency virus (HIV-1) mRNA. Using a single template structure (acetylpromazine) as reference molecule and a topological pharmacophore descriptor (CATS), new chemotypes were automatically generated by our de novo design software Flux. Flux features an evolutionary algorithm for fragment-based compound assembly and optimization. Pharmacophore superimposition and docking into the target RNA suggest perfect matching between the template molecule and the designed compound. Chemical synthesis was straightforward, and bioactivity of the designed molecule was confirmed in a FRET assay. This study demonstrates the practicability of de novo design to generating RNA ligands containing novel molecular scaffolds.

Andreas Schüller and Marcel Suhartono contributed equally to this work.

A. Schüller · M. Suhartono · U. Fechner · Y. Tanrikulu · S. Breitung · U. Scheffer · M. W. Göbel · G. Schneider (✉)
Institute of Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-University, Max-von-Laue-Straβe 7, Chair for Chem- and Bioinformatics Siesmayerstr. 70, 60323 Frankfurt am Main, Germany
e-mail: g.schneider@chemie.uni-frankfurt.de

Keywords Drug design · Evolutionary algorithm · RNA ligand · Scaffold hopping · Virtual screening

Introduction

The aim of de novo design is to create molecular structures that match the binding pattern of a particular biological target. This binding pattern may be specified by either a known ligand (*ligand-based* design) or the 3D receptor structure (*structure-based* design). Structure-based approaches cannot be exerted when the 3D receptor structure of a biological target is unavailable. This is the case for many membrane-bound receptors such as the large group of G-protein coupled receptors (GPCRs) [1], ion channels [2], and transport proteins [3]. Ligand-based approaches do not require a 3D receptor structure. Instead, at least one known ligand—the reference or template structure—of the biological target is a prerequisite.

Computational de novo design may be regarded as complementary to high-throughput screening (HTS) and virtual screening [4]. Following a fragment-based design concept, "candidate compounds" are built from scratch using a set of predefined building blocks and linkage rules. The search space for de novo design is given by all druglike compounds that are chemically feasible—a number estimated to be in the range of 10^{60} to 10^{100} [5, 6]. Thus, efficient optimization algorithms are required to navigate through chemical space towards regions populated by suitable candidate compounds. Reasonable navigation demands for sign posts. In de novo design, these are given by scoring candidate compounds for receptor binding (structure-based design) or similarity to one or more reference or "template" compounds (ligand-based design). Consequently, the three basic components of a de novo design approach are a method

for virtual compound assembly or structure sampling, an optimization algorithm, and a scoring function.

In the early 1990s, pioneering publications provided the basis for automated de novo design [7–9], and the idea was soon picked up and extended by several research groups in industry and academia. Some strategies proved to be more effective than others; for instance, the exclusive usage of atoms as building blocks is now often superseded by the employment of molecular fragments of varying size [4]. Furthermore, the necessity for multidimensional optimization in drug discovery projects has been realized and taken up by recent de novo design software [10, 11]. Incorporation of secondary target constraints into scoring functions seems to increase the suitability of candidate compounds. In particular, the consideration of ADME properties and synthetic accessibility appears to be viable.

Experimental section

Determination of biological activity: FRET assay

Fluorescence-based binding assays [12] were performed in 96-well microplates (Corning 6860, black, non binding surface) at 37 °C and measured in a Microplate reader (Safire^2, Tecan, Crailsheim, Germany) at excitation wavelength 489 nm and emission wavelength 590 nm. Nuclease-free materials were used in steps involving RNA handling. TAR RNA (BioSpring, Frankfurt, Germany) and dye-labeled Tat49–57 peptide (sequence: fluoresceine-AAARKKRRQRR RAAAC-rhodamine, Thermo Electron Corporation, Ulm, Germany) were both used at concentrations of 10 nM in a final volume of 100 μL in TK buffer (50 mM Tris–HCl, 20 mM KCl, 0.01% Triton-X 100, pH 7.4). Prior to analyses, the RNA (100 nM in 0.25× TK buffer) was heated to 90 °C for 5 min and then immediately placed on ice for an additional 5 min. The emission of the background (TK buffer only), pure peptide and of the Tat–TAR complex was established first. Single-point measurements of potential inhibitors were then carried out in triplicates at varying concentrations. The decrease in fluorescence which corresponds to Tat peptide displacement was calculated as ($\text{fluorescence}_{\text{pot. inhibitor}} - \text{fluorescence}_{\text{Tat peptide}}$)/($\text{fluorescence}_{\text{Tat-TAR complex}} - \text{fluorescence}_{\text{Tat peptide}}$).

Synthesis of compound **11**

General

NMR: Bruker DPX 250 (^{1}H: 250 MHz; ^{13}C: 63 MHz), Bruker Avance 400 (^{1}H: 400 MHz; ^{13}C: 100.6 MHz), or Bruker AM 300 (^{1}H: 300 MHz; ^{13}C: 75.4 MHz; ^{19}F: 282 MHz); ^{1}H chemical shifts (δ) are given ppm relative to $CDCl_3$ (7.26 ppm) or DMSO (2.50 ppm) as internal standard; multiplicities are indicated by 's' for singlet, 'd': doublet, "dd": double doublet, 't': triplet, 'm': multiplet, "bs": broad singlet; J in Hz; ^{13}C chemical shifts (δ) are reported with complete proton decoupling in ppm relative to $CDCl_3$ (t, 77.0 ppm) or DMSO (septet, 39.43 ppm). FT-IR: Perkin-Elmer 1600 Series, peaks are reported in cm^{-1}. Elemental analysis: Heraeus CHN Rapid. Mass spectrometry: Fisons CG Platform II. Melting points (uncorrected): Kofler hot plate microscope. Analytical thin layer chromatography (TLC): alumina plates precoated with silica gel 60 F-254 indicator (Merck), visualization by UV light (254 nm). Flash column chromatography: Merck silica gel 230–400 mesh. Filtration: Celite® 535 (Fluka). All reagents were obtained from commercial suppliers and were used without further purification. THF was distilled from sodium/benzophenone. DCM: dichloromethane. TFA: trifluoroacetic acid.

*(3-Fluoro-phenyl)-(1-methyl-piperidin-4-yl)-amine (**14**)*

3-Fluoroaniline (1.5 g, 13.26 mmol) **12**, and N-methylpiperidon (1.25 g, 11.05 mmol) **13** were dissolved in anhydrous toluene (10 mL). Molecular sieves 3 Å (2 g) were added before heating to reflux for 20 h. After cooling down to room temperature Et_2O (20 mL) was given to the dark yellow reaction mixture, than it was filtered over Celite®. The filtrate was concentrated in vacuo and the crude imine was taken up in EtOH (95%, 10 mL). $NaBH_4$ (209 mg, 5.53 mmol) was added. The reaction mixture was stirred at RT for 5 h. For working-up the reaction was quenched with water and extracted with DCM. The combined organic layers were dried ($MgSO_4$) and concentrated in vacuo to dryness. The light yellow crude product was purified by silica gel column chromatography (EtOAc/*n*-hex 10:1). Colorless crystals (1.36 g, 60%) were obtained after re-crystallization from Et_2O/petroleum ether (1:1).

Analytical: R_f (EtOAc/MeOH 9:1) = 0.05; mp: 51–54 °C (Ref.: 50–59 °C) [13]; ^{1}H-NMR (250 MHz, DMSO-d_6): δ = 7.03 (dd, 1H, J = 15.25, J = 8, aryl-*H*), 6.39–6.19 (m, 3H, aryl-*H*), 5.77 (d, 1H, J = 8, N*H*; exchangeable with D_2O), 3.14 (m, 1H, C*H*), 2.70 (m, 2H, $NCH_{ax}H_{eq}$), 2.15 (s, 3H, CH_3), 1.99 (m, 2H, $NCH_{ax}H_{eq}$), 1.85 (m, 2H, $CH_{ax}H_{eq}$), 1.36 (m, 2H, $CH_{ax}H_{eq}$); ^{13}C-NMR (63 MHz, $CDCl_3$): δ = 164.1 (d, J = 242.9, 1C, aryl-*C3*), 148.8 (d, J = 10.7, 1C, aryl-*C1*), 130.2 (d, J = 10.3, 1C, aryl-*C5*), 109.0 (d, J = 2.2, 1C, aryl-*C6*), 103.4 (d, J = 21.7, 1C, aryl-*C4*), 99.5 (d, J = 25.4, 1C, aryl-*C2*), 54.4 (2C, CH_2), 49.4 (1C, N*C*H), 46.2 (2C, CH_2), 32.4 (1C, CH_3); ^{19}F-NMR (282 MHz, DMSO): δ = −113.39 (1F); IR (KBr):

$\tilde{\nu}$ = 3278, 2943, 2782, 1623, 1533, 1442, 1338, 1272, 1178, 1145, 1066, 960, 752; elemental analysis calcd (%) for $C_{12}H_{17}FN_2$ (208.28): C 69.20, H 8.23, N 13.45; found: C 68.98, H 8.12, N 13.55.

*3-Ethoxy-benzoic acid methyl ester (**16**)*

4-Hydroxy benzoic acid (5.2 g, 34.18 mmol) **15** was dissolved in acetone (30 mL). Then, iodoethane (2.3 mL, 28.48 mmol) and K_2CO_3 (7.48 g, 43.11 mmol) were added to the solution. The reaction mixture was refluxed for 16 h. For the workup procedure water was added to the reaction mixture until K_2CO_3 was dissolved completely. Afterwards the product was extracted twice with Et_2O. The combined organic layers were washed 3 times with aqueous NaOH (10%), twice with H_2O and dried (Na_2SO_4). After removing the solvent in vacuo, the light yellow crude product was purified by column chromatography (*n*-hex/EtOAc 5:1). Colorless oil (5.07 g, 99%) was obtained.

Analytical: R_f (*n*-hex/EtOAc 3:1) = 0.6; ^{1}H-NMR (250 MHz, DMSO-d_6): δ = 7.53 (m, 1H, aryl-*H*), 7.42 (m, 2H, aryl-*H*), 7.20 (m, 1H, aryl-*H*), 4.07 (q, *J* = 7, 2H, C*H*$_2$), 3.84 (s, 3H, OC*H*$_3$), 1.33 (t, *J* = 7, 3H, CH$_2$C*H*$_3$); ^{13}C-NMR (63 MHz, $CDCl_3$): δ = 166.9, 158.9, 131.3, 129.3, 121.7, 119.9, 114.5, 63.6, 52.1, 14.7; IR (film): $\tilde{\nu}$ = 2982, 2598, 2081, 1724, 1586, 1446, 1278, 1223, 1100, 1000, 878, 756, 683; elemental analysis calcd (%) for $C_{10}H_{12}O_3$ (180.20): C 66.65, H 6.71; found: C 66.78, H 6.78.

*(3-Ethoxy-phenyl)-methanol (**17**)*

In a flame dried vessel $LiBH_4$ (0.61 g, 27.76 mmol) was dissolved in abs. THF (30 mL). The solution was cooled in an ice bath. Then, 3-ethoxy-benzoic-acid-methyl-ester **16** (2.5 g, 13.88 mmol), which was dissolved in anhydrous THF (10 mL), was added. The reaction mixture was heated to reflux for 3 hours. For working-up the reaction mixture was cooled to 0 °C, afterwards methanol was added till a colorless solid was generated. To dissolve the solid, water was added. Then the solution was extracted twice with EtOAc. The combined organic layers were dried ($MgSO_4$), concentrated in vacuo and purified by column chromatography (*n*-hex/EtOAc 3:1). Colorless oil (2.09 g, 99%) was obtained.

Analytical: R_f (*n*-hex/EtOAc 3:1) = 0.2; ^{1}H-NMR (250 MHz, DMSO-d_6): δ = 7.21 (m, 1H, aryl-*H*), 6.86 (m, 2H, aryl-*H*), 6.77 (m, 1H, aryl-*H*), 5.16 (bs, 1H, O*H*), 4.46 (s, 2H, C*H*$_2$OH), 4.00 (q, *J* = 7, 2H, C*H*$_2$CH$_3$), 1.32 (t, *J* = 7, 3H, CH$_2$CH$_3$); ^{13}C-NMR (63 MHz, DMSO-d_6): δ = 158.4, 144.1, 129.0, 118.3, 112.5, 112.1, 62.71, 62.69, 14.6; IR (film): $\tilde{\nu}$ = 3375, 2980, 2878, 1601, 1490, 1449, 1263, 1156, 1047, 860, 782, 694; elemental analysis calcd (%) for $C_9H_{12}O_2$ (152.19): C 71.03, H 7.95; found: C 70.94, H 8.08.

*1-Ethoxy-3-iodomethyl-benzene (**18**)*

Iodine (525 mg, 2.07 mmol) was added to a solution of triphenylphosphine (543 mg, 2.07 mmol) in DCM (8 mL) and stirred at RT for 5 min. Then 3-ethoxy benzyl alcohol **17** (300 mg, 1.97 mmol, dissolved in DCM (4 mL)), was given to reaction mixture and was stirred at RT for 1 h. For workup the reaction mixture was diluted with Et_2O (20 mL). The precipitate was removed by filtration over Celite®. The filtrate was successively washed with aq. $NaHCO_3$ (5%), twice with sat. $Na_2S_2O_5$ and water. After drying (Na_2SO_4) and evaporation, the crude product was purified by flash column chromatography (*n*-hex/DCM 1:1). The collected fractions were concentrated till dryness. Colorless needles (490 mg, 95%) were obtained.

Analytical: R_f (*n*-hex/EtOAc 10:1) = 0.9; mp: 26 °C; ^{1}H-NMR (250 MHz, DMSO-d_6): δ = 7.20 (m, 1H, aryl-*H*); 6.98 (m, 2H, aryl-*H*), 6.79 (m, 1H, aryl-*H*), 4.57 (s, 2H, aryl-C*H*$_2$), 4.00 (q, *J* = 7, 2H, OC*H*$_2$), 1.31 (t, *J* = 7, 3H, C*H*$_3$); ^{13}C-NMR (63 MHz, DMSO-d_6): δ = 158.3, 140.9, 129.5, 120.9, 114.7, 113.6, 62.8, 14.5, 7.2; IR (KBr): $\tilde{\nu}$ = 3036, 2973, 2929, 2882, 1929, 1840, 1752, 1602, 1581, 1486, 1446, 1390, 1296, 1265, 1160 1114, 1047, 967, 873, 778, 694, 627; elemental analysis calcd (%) for $C_9H_{11}IO$ (262.09): C 41.24, H 4.23; found: C 41.41, H 4.38.

*(3-Ethoxy-benzyl)-(3-fluoro-phenyl)-(1-methyl-piperidin-4-yl)-amine (**11**)*

K_2CO_3 (645 mg, 4.67 mmol) was suspended in a solution of 3-ethoxy-benzyl-iodide **18** (450 mg, 1.71 mmol) in MeCN (8 mL). The mixture was completely degassed. After flushing with argon (3-fluoro-phenyl)-(1-methyl-piperidin-4-yl)-amine **14** (324 mg, 1.56 mmol) was added and the reaction mixture was degassed once again and backflushed with argon. Afterwards, the reaction was refluxed for 3 h. For workup water was added till K_2CO_3 was completely dissolved. Then the solution was extracted twice with CH_2Cl_2. The combined organic layers were washed with NaOH (1N), then once again with water. The organic phase was dried over $MgSO_4$, concentrated under reduced pressure and dried in vacuo. Colourless foam (515 mg, 96%) was obtained.

Analytical: R_f (EtOAc/MeOH 9:1) = 0.25; ^{1}H-NMR (250 MHz, DMSO-d_6): δ = 7.43 (m, 1H, aryl-*H*); 7.17–7.03 (m, 4H, aryl-*H*), 6.49–6.28 (m, 3H, aryl-*H*), 4.59 and

4.56 (2s, 2H, rotamers of aryl-CH_2), 4.07 (2q, J = 7, 2H, rotamers of OCH_2), 3.64–3.32 (m, 5H, piperidinyl-*H*), 3.00 and 2.95 (2s, 3H, rotamers of NCH_3), 2.20–1.71 (m, 4H, piperidinyl-*H*), 1.35 (2t, J = 7, 3H, rotamers of CH_2CH_3); ^{13}C-NMR (63 MHz, DMSO-d_6): δ = 163.4, 158.4, 149.2, 130.3, 129.9, 128.8, 124.8, 119.0, 116.0, 108.9, 102.1, 98.7, 63.1, 58.3, 57.5, 45.8, 44.5, 25.0, 14.4, some signals are split due to C-F-coupling and the existence of rotamers; ^{19}F-NMR (282 MHz, $CDCl_3$): δ = –112.6 (1F); IR (KBr): $\tilde{\nu}$ = 3277, 2974, 2933, 1618, 1586, 1523, 1494, 1445, 1390, 1336, 1267, 1150, 1113, 1046, 998, 882, 761, 687; MS (ESI): m/z (%) calcd for $C_{21}H_{27}FN_2O$ $[M + H]^+$: 343.22; found: 343.1 (100).

400 mg of the product were purified by HPLC. The product (630 mg, 95%) was obtained as a dihydrotrifluoroacetate.

Preparative: Reprosil AQ, 250 × 20, 0.1% TFA/MeCN 49:51, 10 mL/min;

Analytical: Reprosil AQ, 125 × 4.6, 0.1% TFA/MeCN 55:45, 0.8 mL/min, t_R = 4.78 min;

^{1}H-NMR (250 MHz, DMSO-d_6): δ = 7.43 (m, 1H, aryl-*H*); 7.12–7.08 (m, 4H, aryl-*H*), 6.48–6.27 (m, 3H, aryl-*H*), 4.58 and 4.55 (2s, 2H, rotamers of aryl-CH_2), 4.07 (2q, J = 7, 2H, rotamers of OCH_2), 3.63–3.33 (m, 5H, piperidinyl-*H*), 3.00 and 2.95 (2s, 3H, rotamers of NCH_3), 2.20–1.71 (m, 4H, piperidinyl-*H*), 1.35 (2t, J = 7, rotamers of CH_2CH_3); ^{13}C-NMR (63 MHz, DMSO-d_6): δ = 163.5, 158.5, 149.2, 130.3, 129.9, 128.8, 124.8, 119.0, 116.0, 108.9, 102.1, 98.7, 63.1, 58.4, 57.5, 45.9, 44.6, 25.0, 14.4, some signals are split due to C–F-coupling and the existence of rotamers; ^{19}F-NMR (282 MHz, DMSO-d_6): δ = −74.42 (6F, TFA), −113.05 (1F, aryl-*F*); IR (KBr): $\tilde{\nu}$ = 3447, 3279, 2928, 1618, 1587, 1524, 1494, 1446, 1391, 1336, 1267, 1179, 1151, 1046, 882, 761, 687; MS (ESI): m/z (%) calcd for $C_{21}H_{27}FN_2O$ $[M + H]^+$: 343.22; found: 343.0 (100).

Virtual chemistry: from drugs to building blocks and back again

The fundamental building blocks of de novo design are either atoms or molecular fragments. Atom linking allows for a lot more combinations than the connection of predefined molecular fragments. This, however, often constitutes rather a difficulty than an advantage. The search space for a de novo design exercise is made up by all compounds with a druglike molecular weight (in general, a threshold below 750 Da is employed) that can be assembled from the set of elements typically found in drugs (composition of a typical druglike compound database [14]: C: 75%, O: 12%, N: 10%, halogens: 2%, other elements: 1%). The astronomically large number of possible solutions can be significantly reduced to 10^{20}–10^{30} by using molecular fragments instead of individual atoms for virtual compound assembly. Yet, it is still challenging to find suitable drug candidates in this search space. Noteworthy, early atom-based methods tended to generate candidate compounds that were chemically invalid, unreasonable or synthetically not amenable [9, 15]. But the associated problems were recognized and methods that were later developed used molecular fragments as building blocks [16–18].

Fragments as building blocks only cover a limited region of the chemical space. This region is defined by the number and type of fragments as well as the set of linkage rules. The reduction can be considered meaningful if fragments are chosen that commonly occur in known drugs and druglike molecules. The size of fragments may vary considerably. This enables smaller and larger changes during candidate compound assembling, subject to the particular requirements. Current de novo design methods often use a mixture of building blocks the majority of which are multiple-atom fragments, augmented by fewer single-atom building blocks [19].

There are several ideas of how to assemble candidate compounds in a combinatorial fashion [20]: linking, growing, lattice-based structure sampling, step-wise structure transition, and graph-based assembly.

Linking starts by placing building blocks at interaction sites of a receptor that play a key role. Then, the placed building blocks are joined to each other by other building blocks, the so-called "linkers" [21, 22].

Growing, on the other hand, initiates candidate compound generation by placing a single building block at an interaction site of the receptor. Further building blocks are added one at a time each of which is connected to the growing candidate compound. Building blocks are selected for their ability to satisfy interaction sites of the receptor and match regions between two such sites [23–27].

Lattice-based structure sampling lays out an atomic lattice in the binding pocket of the receptor. Lattice atoms that separate two interaction sites along a shortest path are connected to each other by the formation of bonds [28–32].

Step-wise structure transitions are driven by molecular dynamics simulations. Candidate compounds are dynamically designed by creating bonds between building blocks in a stochastic and reversible way. For example, W. L. Jorgensen and coworkers developed the software BOMB for scoring ligand-receptor complexes and coupled it with extended linear response and free energy perturbation (FEP) calculations using Monte Carlo sampling with explicit inclusion of solvent molecules [33]. At the beginning, building blocks are randomly distributed in the binding pocket. Then, the placed building blocks are moved according to molecular dynamics equilibria. After

several simulation steps some building blocks are randomly chosen, bonds between building blocks are cleaved, and empty valences are filled with bonds to building blocks in the vicinity. New candidate compounds emerge from this alternating procedure of building blocks movement and bond formation.

Ligand-based de novo design does not have information about the binding site. In lieu thereof, most ligand-based approaches directly operate on the topological molecular graph and feature tailored graph-based transition operators.

Graph-based compound assembly shall be demonstrated taking our own fragment-based molecule generator Flux as an example [34, 35]. Flux deploys a set of eleven common bond cleavage types as linkage rules (RECAP, Retrosynthetic Combinatorial Analysis Procedure) [36]. Building blocks are obtained by using this set of linkage rules in a backward fashion: An exhaustive virtual *retro*-synthesis of a drug dataset is carried out. During a de novo design run, fragments are connected by virtual synthesis using the set of linkage rules. This concerted use of virtual *retro*-synthesis and subsequent forward-synthesis by the same set of reaction schemes is assumed to ease the chemical synthesis of candidate compounds [37]. Fragmentation of druglike molecules is expected to result in fragments that cover druglike regions of chemical space. Therefore, the search space of Flux and similar software tools is restricted to areas that are assumed to be relevant for the generation of druglike molecules.

De novo design is faced with a vast search space caused by the numerous possibilities to connect a given set of building blocks. Problems that encounter such a combinatorial explosion are non-deterministic polynomial-time (NP) hard: They cannot be solved with provable optimal solution quality and provable good run times [4]. Combinatorial search algorithms offer solutions by exploring the search space efficiently and concurrently reducing it. Commonly, this is achieved by implementing a *heuristic*. A heuristic algorithm applies empirical rules based on knowledge of a particular problem domain.

Evolutionary algorithms are population-based optimization strategies that simulate the evolutionary pressure of selection and the evolutionary operators "mutation" and "crossover". In case of de novo design individuals are represented by candidate compounds. A mutation event introduces new information into a population. The crossover operator, on the other hand, exploits the information that is present in a population. This is achieved by recombining the information of two individuals of the population. Application of the mutation and crossover operator leads to variation of individuals. In the context of evolutionary algorithms the scoring function is often termed fitness function. It determines which individuals are selected for variation. Basically, an evolutionary algorithm traverses an iterative cycle of variation and selection until a termination criterion is reached. Evolutionary algorithms can be subdivided into genetic algorithms, genetic programming, and evolution strategies.

The de novo design method Flux utilizes an evolution strategy (ES), more specifically a (μ, λ) ES [38]. μ in (μ, λ) indicates that the μ fittest individuals are approved to spawn offspring, λ defines the number of children per generation (Fig. 1). The algorithm employs an adaptive step-size control mechanism that is meant to help escape local optima [39, 40].

Structure assembling in Flux is implemented by two graph-based alteration methods: A molecular mutation operator and a molecular crossover operator. The mutation operator starts with an exhaustive virtual *retro*-synthesis of a candidate compound. One of the resultant fragments is then substituted by another one from the stock of building blocks. The fragments are re-assembled by means of virtual synthesis to obtain the "mutated" candidate compound. The crossover operator takes two candidate compounds as an input. Both are exhaustively dissected by virtual *retro*-synthesis. One fragment is then exchanged between the two dissected structures. Finally, reassembling the fragments yields two altered candidate compounds (Fig. 1).

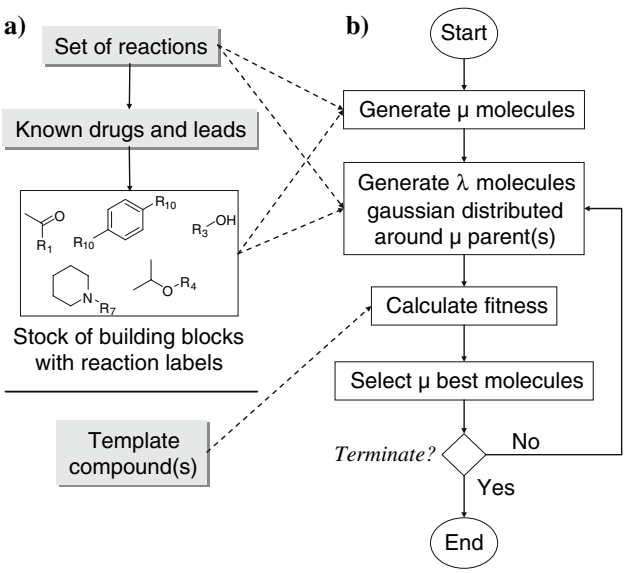


Fig. 1 (μ,λ) Evolution strategy for ligand- and fragment-based de novo design. (**a**) A stock of molecular fragments is generated by virtual *retro*-synthesis of known drugs and lead structures. (**b**) In an iterative process generations of new compounds are bred using mutation and crossover operators together with the stock of building-blocks. One or more template compounds are used for fitness calculation. Parent molecules of a new generation are selected from the offspring only (parents do not participate in the selection; they "die"). For generation of offspring, the reactions for *retro*-synthesis of the parents and forward synthesis of new compounds are randomly selected

Scoring a virtual molecule by topological pharmacophores

The objective of a de novo design exercise is the generation of candidate compounds that exhibit a desired biological or pharmacological effect. A scoring function primarily considers information related to the binding affinity of candidate compounds to a particular biological target (*primary target constraints*). Primary target constraints may be derived from the 3D receptor structure (structure-based design) or from known ligands (ligand-based design).

Scoring functions that are employed in structure-based design are analogous to scoring functions from the field of docking [41–43]. They can be subdivided into explicit force-field approaches, empirical scoring functions, and knowledge-based scoring functions. All of them aim for the estimation of the binding free energy.

Ligand-based design derives primary target constraints from one or more known ligands. Structure-based design unavoidably operates in 3D space and has to deal with the complexity of molecular conformation and conformer sampling. Ligand-based scoring functions, on the other hand, may or may not take into account the 3D information of known ligands. Ligand-based scoring functions quantify the similarity between a candidate compound and one or more known reference ligands. Pharmacophore models [44], QSAR equations [45–47], and pair-wise chemical similarity [48–50] have been employed to assess the potential binding affinity of candidate compounds in de novo design.

Besides primary target constraints many recent scoring functions also consider secondary target constraints such as ADMET properties, adverse effects, and interdependency with other drugs [50]. Their overall score is a weighted sum made up of an estimation of the binding affinity and other terms that account for secondary target constraints. Alternatively, Pareto ranking may be applied [51]. Additional secondary constraints of particular importance are the aqueous solubility, chemical feasibility and synthetic accessibility of the designed structures.

The scoring function of our de novo design program Flux regards primary target constraints by pair-wise chemical similarity between candidate compounds and known ligands that serve as templates. Here, we employed the CATS topological pharmacophore similarity [52]. Although not being part of the scoring function, synthetic accessibility is implicitly considered by the choice of building blocks and their linkage rules.

The potential pharmacophore point (PPP) definitions in CATS are loosely based on the LUDI atom types, which is an exceptionally popular and successful computer program for automated structure-based drug design developed approximately 15 years ago [23, 24, 53]: Hydrogen-bond donor (D), hydrogen-bond acceptor (A), lipophilic (L), positively charged or ionizable (P), and negatively charged or ionizable (N). LUDI originally defined four atom types:

1. Hydrogen-bond donor (all hydrogens that are bound to a nitrogen or an oxygen)
2. Hydrogen-bond acceptor (nitrogen atoms that are not adjacent to hydrogen, all oxygen atoms)
3. Lipophilic-aliphatic (aliphatic carbon atoms)
4. Lipophilic-aromatic (carbon atoms that are members of a phenylalanine, tyrosine, histidine, and tryptophan)

LUDI assigned atom types to protein and ligand atoms whereas CATS assigns them to the atoms of a potential ligand only. The CATS PPPs do not differentiate between aromatic and aliphatic lipophilicity. CATS considers two further atom types to explicitly account for charged or potentially ionized atoms, an addition already seen in earlier cheminformatics tools such as Catalyst [54] and PATTY [55].

The computational protocol of CATS assigns each atom of a molecule to zero, one or two PPPs. A distance matrix whose entries d_{ij} hold the topological shortest path quantified by the number of bonds between PPP i and PPP j is calculated from all PPPs. The distance matrix is computed by a breadth-first algorithm [56]. A mapping function combines the information of the distance matrix and the PPP assignments to obtain a pharmacophoric correlation vector (CV). The mapping function counts the frequencies of all possible PPP pair distances and sums them in ten bins covering the distances of 0 to 9 bonds. The 15 pairing combinations of PPPs (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) and 10 different distances (bins) yield a 150-dimensional CV. The final step of the CATS descriptor calculation is the scaling of the vector [57, 58]. The CATS descriptor was successfully employed in earlier de novo design projects and had proven useful for the purpose of scaffold-hopping [49, 59, 60].

Case study: design of a novel inhibitor of Tat-TAR RNA interaction

For illustration of what can be achieved by template-based de novo design with drug-derived molecular fragments, we demonstrate the application of our de novo design software Flux for generating novel ligands of TAR RNA, which is a potential drug target in the treatment of AIDS [61]. The *trans*-activation response element (TAR) of the human immunodeficiency virus (HIV)-1 mRNA is one of the best studied RNA-based regulatory systems [62]. Specific binding of the Tat protein to TAR is essential for HIV replication. Thus the TAR RNA represents a potential

target for defeating of HIV as well as a model system to deepen the understanding of RNA-small molecule interactions.

In contrast to many protein structures, TAR RNA is a very flexible macromolecule as shown by NMR spectroscopy and molecular dynamics simulations [63, 64]. It thus represents a comparably challenging target. The central bulge region formed by the nucleosides U23, C24, and U25 with the flanking G26 and U40 can accommodate intercalating ligands like, for example, acetylpromazine (Fig. 2) [65]. The aim of our study was to see whether fragment-based design leads to novel TAR RNA binding agents mimicking acetylpromazine.

For de novo design with Flux, we first generated a fragment database containing 6,156 unique building blocks. It was compiled from the COBRA drug and lead database (version 6.1, 7,450 compounds) [14]. The fragments served for molecular assembly during the de novo design process. Both virtual *retro*-synthesis and forward synthesis were performed using the RECAP reactions, excluding the olefin reaction. New structures were bred by applying the mutation operator only (no crossover), the population size was set to 100 individuals, and optimization was followed over 75 generations in each run. A (1,100) evolution strategy with a fixed step-size for mutation was used. The initial parent molecule was randomly assembled, and the entire design process was initialized 100 times to obtain independent results. Fitness of the virtual compounds was assessed by computing the Euclidian distance between their 150-dimensional CATS descriptor vector and the CATS descriptor of the template compound acetylpromazine. The design objective was to obtain a virtual compound with a minimal distance to the template. All designed compounds with a molecular weight between 200 and 750 Da were kept for inspection.

Figure 3 shows the ten best de novo designs obtained. All of these molecules contain a central tertiary amine connected to ring systems via linker chains ("trefoil-like" structure). The three-ring system of the template was redesigned in compound **8**, and a derivative thereof is present in compound **2**. The remaining designs can be regarded as scaffold-hops. The observed high abundance of fluorines and chlorines in the ten best designs is a consequence of a high frequency of halides in terminal building blocks of the fragment database (approx. 20%) and the inability of the scoring function to distinguish fluorides from non-fluorides. Carbohalide groups are generally regarded as lipophilic by the employed scoring function.

We decided to synthesize a variant of the best-ranked molecule **1**. The original design contains a tetrafluoroethoxy group. For easier synthesis, and because the CATS descriptor does not explicitly consider fluorine moieties, we modified molecule **1** yielding compound **11**, where an ethoxy group substitutes for the 1,1,2,2-tetrafluoro-1-ethoxy moiety. Figure 4 shows a conformer superposition of compound **11** with acetylpromazine. Almost perfect fit of every pharmacophoric feature can be observed.

Prior to synthesis and testing, automated docking of compound **11** was performed with GOLD v3.1.1 [67], which relies on a genetic algorithm for ligand positioning. This was done to assess whether this molecule could actually be accommodated in the bulge of TAR RNA. In total, 10 separate docking runs were performed, each with ten initializations of the genetic algorithm. An average GOLDscore of 36.6 ± 7.1 was obtained. The best-ranked pose had a positive GOLDscore of 47.2 indicating

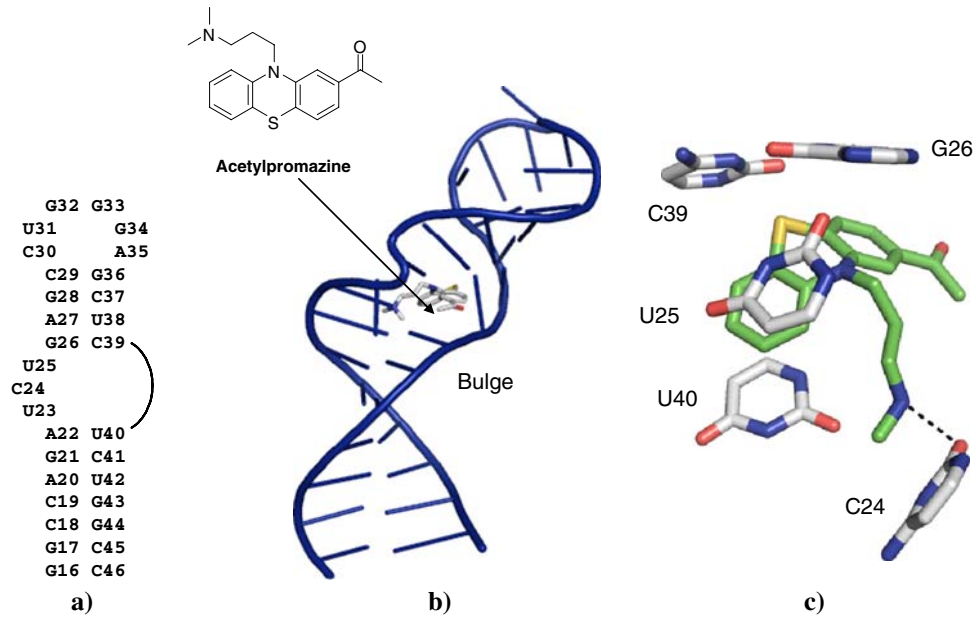


Fig. 2 Stem-loop structure of TAR RNA (**a**), cartoon model (**b**), and close-up of the bulge region with ligand acetylpromazine bound (**c**). The bases that were shown to be major interaction partners for the ligand acetylpromazine are U25, G26, and U40 [63]. Coordinates from NMR model 1 of Protein Database (PDB) entry 1LVJ [63]

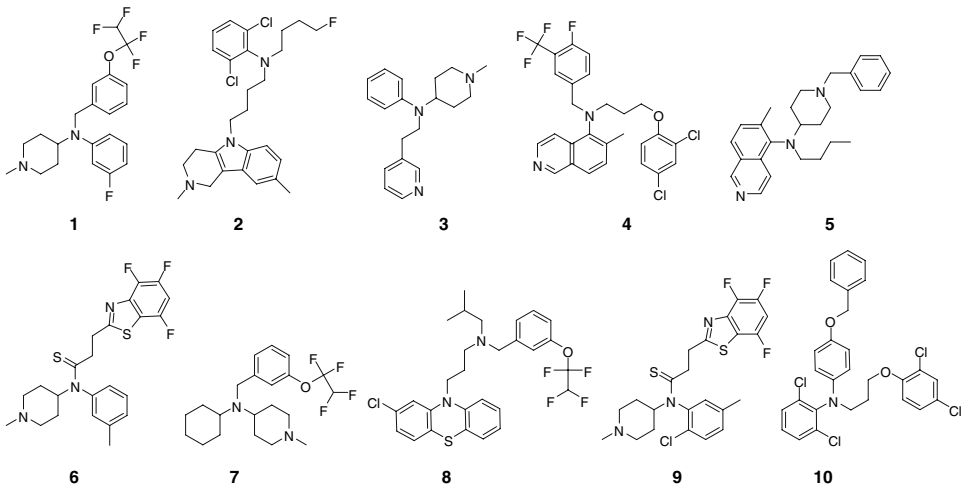


Fig. 3 Top 10 de novo designed virtual molecules with acetylpromazine as template structure, sorted according to decreasing fitness (CATS similarity to acetylpromazine)

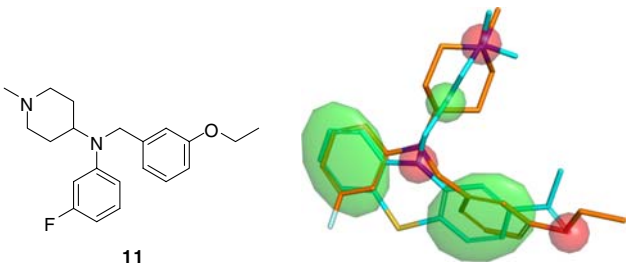


Fig. 4 Superposition of molecule **11** (*gold*) and acetylpromazine (*gray*). Pharmacophoric points are shown as colored ellipsoids (*green*: lipophilic, *red*: hydrogen bond acceptor). Different atoms are denoted by different coloring (nitrogen: *blue*, oxygen: *red*, sulfur: *yellow*, fluorine: *light blue*). Three-dimensional coordinates were gained with CORINA (Molecular Networks GmbH, Erlangen, Germany). The spatial alignment was obtained using the flexible alignment tool in MOE 2005.06 (Chemical Computing Group Inc., Montreal, Canada). Pharmacophore feature calculation was performed with LIQUID [66]

favourable binding to TAR. Essential interaction of high scoring poses with bases U25, G26, and U40 seem possible (Fig. 5). These results confirm the perfect matching of pharmacophoric points in the ligand-based model (Fig. 4). For comparison, re-docking of the template ligand acetylpromazine into the same TAR structure yielded an average GOLDscore of -8.9 ± 1.8, indicating unfavourable binding. This can be readily explained, as the docking algorithm never succeeded in reproducing the bioactive pose of acetylpromazine but rather placed the ligand turned by 180°.

Motivated by these theoretical investigations, molecule **11** was synthesized and subsequently tested in a FRET-assay for its ability to disrupt the Tat-TAR interaction. Assay conditions were as described previously [12, 68]. Chemical synthesis was straightforward (Scheme 1), which can be seen as a consequence of using drug-derived molecular fragments and a *pseudo retro*-synthesis approach for de novo design.

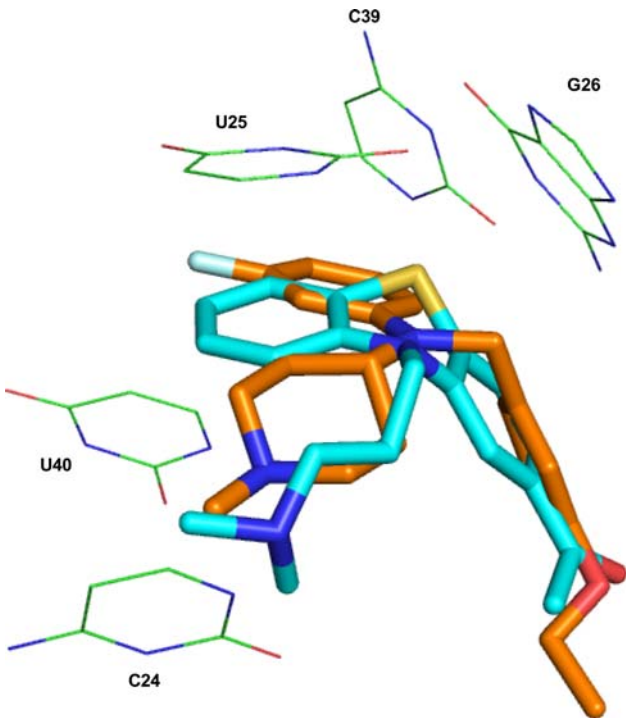
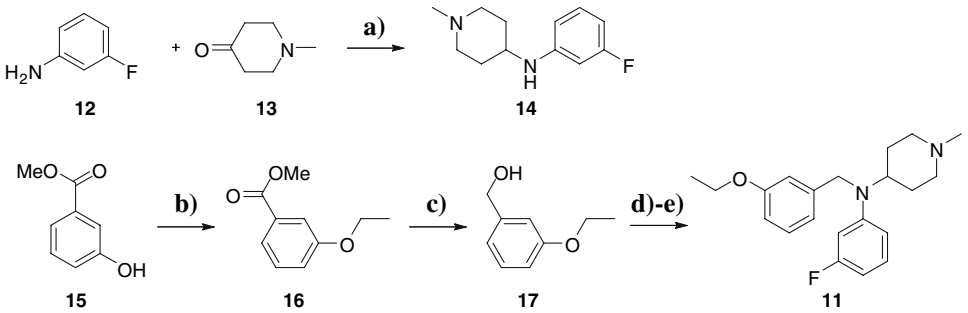


Fig. 5 NMR structure of acetylpromazine (*blue*) and the docking pose of the modified de novo molecule **11** (*gold*) in the bulge of TAR RNA. Atom types are denoted by different coloring (as in Fig. 4). For docking of **11**, acetylpromazine was removed from TAR structure 1LVJ (NMR model 1). GOLD was used with the following parameter setting: *popsize* = 100, *pt_crosswt* = 95, *allele_mutatewt* = 95, *migratewt* = 10. The docking site was defined with a radius of 10 Å centered at $x = -24.8$, $y = 7.9$, $z = 6.5$ with the option *do_cavity_search* = 1

The binding assay confirmed activity of compound **11**, yielding on average 13% decreased Tat-TAR interaction at a concentration of 250 μM, 17% at 500 μM, 21% at 1 mM, and 29% at 3 mM. The template acetylpromazine yields an IC_{50} value of 500 μM in this assay [68]. For comparison, Mayer and James reported a K_d of 270 μM [70], and Du



Scheme 1 Synthesis of compound **11**. (**a**) 1. MS 3Å, PhMe, reflux, 20 h; 2. $NaBH_4$, RT, 5 h [69]; (**b**) C_2H_5I, K_2CO_3, acetone, reflux, 8 h; (**c**) $LiBH_4$, THF, reflux, 3 h; (**d**) PPh_3, I_2, DCM, RT, 1 h; (**e**) **14**, K_2CO_3, MeCN, reflux, 3 h

et al. obtained nanomolar TAR binding of acetylpromazine in a gel-shift assay [63]. Notably, the substance did not have solubility issues, as determined by UV–Vis spectroscopy (not shown). It may therefore be concluded that the designed compound **11** is a weaker inhibitor of the Tat-TAR interaction than the template acetylpromazine, although determination of a reliable IC_{50} or K_d value is still owing. Irrespective of the outcome of such additional activity determination, it is safe to say that compound **11** exhibits the desired function, namely inhibition of the interaction of the Tat peptide with TAR RNA. Clearly, higher assay activity is desirable, and we are working on new designs. Nonetheless, the designed compound represents a novel molecule with anti-TAR RNA activity without comparable analogues in the area of RNA research in present literature as determined by a SciFinder Scholar database search [71].

Conclusions

We demonstrated the applicability of a ligand- and fragment-based concept to de novo design of RNA ligands. The actual design process suggested several chemotypes with strong structural dissimilarity to the template molecule. We conclude that fragment-based design offers the possibility of scaffold-hopping from known bioactive chemotypes to isofunctional "activity islands" in chemical space [52, 72]. Such designs should be regarded as suggestions only. Testing of TAR RNA ligand **11**, a derivative of the de novo designed compound, revealed poorer binding compared to the template structure acetylpromazine. A similar observation has been made in previous de novo design studies that relied on a template or reference structure [49, 59]. Automated de novo design should preferably be used as an idea generator. One cannot expect the optimal lead structure from de novo design in the first place. Notably, after a simple modification of the designed molecule (removal of fluorines), chemical synthesis was straightforward, which is a consequence of using a "simple" virtual reaction scheme and drug-derived building blocks for automated design. In future applications, the stock of fragments should be augmented, in particular by smaller building blocks to allow for fine-tuning of structures, and filtered to remove fragments with undesirable properties, such as overly abundant halides. Additional reactions should be added to the *retro*-synthetic scheme to further increase the chemical variability of the fragment library. We have already begun to compile a list of reactions for this purpose, which are known to often perform with high yields and in few steps, preferably in aqueous solution. "Click chemistry" presents itself as such a potentially ideal concept for automated de novo design [73].

Acknowledgements The authors would like to thank Steffi Becker for technical assistance. M.S. is grateful for a predoctoral fellowship from the Dr. Hilmer Foundation. This work was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, and the Deutsche Forschungsgemeinschaft (SFB 579 "RNA-Ligand Interactions", projects A3 and A11).

References

1. Lowrie JF, Delisle RK, Hobbs DW, Diller DJ (2004) Comb Chem High Throughput Screen 7:495
2. Fontaine B, Plassart-Schiess E, Nicole S (1997) Mol Aspects Med 18:415
3. Oh SJ, Ha H-J, Chi DY, Lee HK (2001) Curr Med Chem 8:999
4. Schneider G, Fechner U (2005) Nat Rev Drug Discov 4:649
5. Walters WP, Stahl MT, Murcko MA (1998) Drug Disc Today 3:160
6. Lipinski C, Hopkins A (2004) Nature 432:855
7. Danziger DJ, Dean PM (1989) Proc R Soc Lond B 236:101
8. Lewis RA (1990) J Comput Aided Mol Des 4:205
9. Nishibata Y, Itai A (1991) Tetrahedron 47:8985
10. Bleicher KH, Böhm H-J, Müller K, Alanine AI (2003) Nat Rev Drug Disc 2:369
11. Keseru GM, Makara GM (2006) Drug Discov Today 11:741
12. Matsumoto C, Hamasaki K, Mihara A, Ueno A (2000) Bioorg Med Chem Lett 10:1857
13. Carabateas PM, Schodack NY (1971) Chem Abstr EN. US patent no. 3679690; 77:151966
14. Schneider P, Schneider G (2003) QSAR Comb Sci 22:713
15. Pearlman DA, Murcko MA (1993) J Comput Chem 14:1184
16. Clark DE, Frenkel D, Levy SA, Li J, Murray CW, Robson B, Waszkowycz B, Westhead DR (1995) J Comput Aided Mol Des 9:13
17. Gillett VJ, Myatt G, Zsoldos Z, Johnson AP (1995) Perspect Drug Discov Des 3:34
18. Todorov NP, Dean PM (1997) J Comput Aided Mol Des 11:175

19. Mauser H, Stahl M (2007) J Chem Inf Model 47:318
20. Leach AR, Bryce RA, Robinson AJ (2000) J Mol Graph Model 18:358
21. Leach AR, Lewis RA (1994) J Comput Chem 15:233
22. Leach AR, Kilvington SR (1994) J Comput Aided Mol Des 8:283
23. Böhm H-J (1992) J Comput Aided Mol Des 6:61
24. Böhm H-J (1992) J Comput Aided Mol Des 6:593
25. Rotstein SH, Murcko MA (1993) J Med Chem 36:1700
26. Murray CW, Clark DE, Auton TR, Firth MA, Li J, Sykes RA, Waszkowycz B, Westhead DR, Young SC (1997) J Comp Aided Mol Des 11:193
27. Eksterowicz JE, Evensen E, Lemmen C, Brady GP, Lanctot JK, Bradley EK, Saiah E, Robinson LA, Grootenhuis PD, Blaney JM (2002) J Mol Graph Model 20:469
28. Lewis RA (1990) J Comput Aided Mol Des 4:205
29. Lewis RA, Roe DC, Huang C, Ferrin TE, Langridge R, Kuntz ID (1992) J Mol Graph 10:66
30. Roe DC, Kuntz ID (1995) J Comput Aided Mol Des 9:269
31. Gehlhaar DK, Moerder KE, Zichi D, Sherman CJ, Ogden RC, Freer ST (1995) J Med Chem 38:466
32. Miranker A, Karplus M (1995) Proteins 23:472
33. Jorgensen WL, Ruiz-Caro J, Tirado-Rives J, Basavapathruni A, Anderson KS, Hamilton AD (2006) Bioorg Med Chem Lett 16:663
34. Fechner U, Schneider G (2006) J Chem Inf Model 46:699
35. Fechner U, Schneider G (2007) J Chem Inf Model 47:656
36. Lewell XO, Budd DB, Watson SP, Hann MM (1998) J Chem Inf Comput Sci 38:511
37. Schneider G, Lee M-L, Stahl M, Schneider P (2000) J Comput Aided Mol Des 14:487
38. Rechenberg I (1994) Evolutionsstrategie '94. Frommann-Holzboog, Stuttgart
39. Schneider G, Schuchhardt J, Wrede P (1994) Comput Appl Biosci 10:635
40. Schneider G, Schuchhardt J, Wrede P (1996) Biol Cybern 74:203
41. Gohlke H, Hendlich M, Klebe G (2000) J Mol Biol 295:337
42. Schneider G, Böhm H-J (2002) Drug Discov Today 7:64
43. Joseph-McCarthy D, Baber JC, Feyfant E, Thompson DC, Humblet C (2007) Curr Opin Drug Discov Devel 10:264
44. Waszkowycz B, Clark DE, Frenkel D, Li J, Murray CW, Robson B, Westhead DR (1994) J Med Chem 37:3994
45. Nachbar RB (2000) Genet Programming Evolvable Machines 1:57
46. Pellegrini E, Field MJ (2003) J Comp Aided Mol Des 17:621
47. Douguet D, Thoreau E, Grassy G (2000) J Comput Aided Mol Des 14:449
48. Globus A, Lawton J, Wipke WT (1999) Nanotechnology 10:290
49. Schneider G, Chomienne-Clement O, Hilfiger L, Kirsch S, Böhm H-J, Schneider P, Neidhart W (2000) Angew Chemie Int Ed 39:4130
50. Brown N, McKay B, Gilardoni F, Gasteiger J (2004) J Chem Inf Comput Sci 44:1079
51. Gillet VJ, Khatib W, Willett P, Fleming PJ, Green DVS (2002) J Chem Inf Comput Sci 42:375
52. Schneider G, Neidhart W, Giller T, Schmid G (1999) Angew Chemie Int Ed 38:2894
53. Böhm H-J (1993) J Mol Recog 6:131
54. Barnum D, Greene J, Smellie A, Sprague P (1996) J Chem Inf Comput Sci 36:563
55. Bush BL, Sheridan RP (1993) J Chem Inf Comput Sci 33:756
56. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms (2nd edn). MIT Press, Cambridge, pp 531–539
57. Fechner U, Schneider G (2004) QSAR Comb Sci 23:19
58. Fechner U, Schneider G (2004) ChemBioChem 5:538
59. Rogers-Evans M, Alanine AI, Bleicher KH, Kube D, Schneider G (2004) QSAR Comb Sci 23:426
60. Schneider G, Schneider P, Renner S (2006) QSAR Comb Sci 25:1162
61. Bannwarth S, Gatignol A (2005) Curr HIV Res 3:61
62. Karn J (1999) J Mol Biol 293:235
63. Du Z, Lind KE, James TL (2002) Chem Biol 9:707
64. Mu Y, Stock G (2006) Biophys J 90:391
65. Lind K, Du Z, Fujinaga K, Peterlin B, James T (2002) Chem Biol 9:185
66. Tanrikulu Y, Nietert M, Proschak E, Grabowski K, Schneider P, Scheffer U, Göbel M, Schneider G (2007) Chembiochem 8:1932
67. Jones G, Willet P, Glen R, Leach A, Taylor R (1997) J Mol Biol 267:727
68. Renner S, Ludwig V, Boden O, Scheffer U, Göbel M, Schneider G (2005) Chembiochem 6:1119
69. Asachi M, Sasakura K, Sugasawa T (1985) Chem Pharm Bull 33:1826
70. Mayer M, James TL (2004) J Am Chem Soc 126:4453
71. (a) SciFinder Scholar 2006, Chemical Abstracts Service, Columbus, Ohio, USA, excessed Oct. 2007 (b) Somerville AN (1998) J Chem Inf Comput Sci 38:1024
72. Noeske T, Sasse BC, Stark H, Parsons CG, Weil T, Schneider G (2006) ChemMedChem 1:1066
73. Kolb HC, Finn MG, Sharpless KB (2001) Angew Chem Int Ed 40:2004