Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

69451 Weinheim, Germany

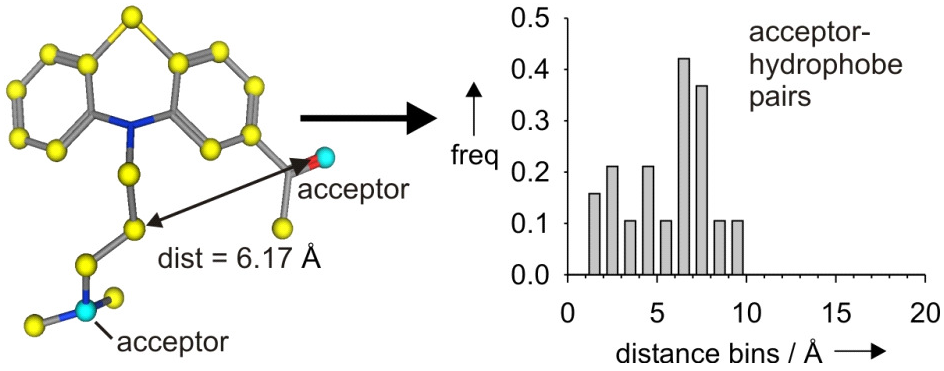
Scaffold-Hopping by 3D-Pharmacophores and Neural Network Ensembles

Steffen Renner, Mirko Hechenberger, Tobias Noeske, Alexander Böcker, Claudia Jatzke, Michael Schmuker, Chris G. Parsons, Tanja Weil, Gisbert Schneider

1. Computational Methods

CATS3D

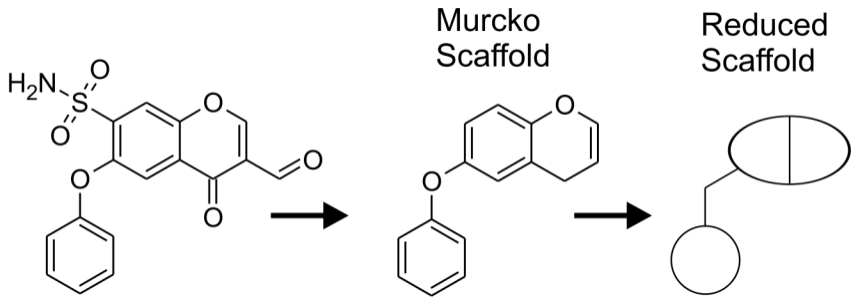
The CATS3D descriptor describes the distribution of potential pharmacophore points (PPPs) in three-dimensional space. Distances between all pairs of PPP are calculated and the frequencies of such pairs falling into defined distance ranges are stored in a correlation vector. For CATS3D 20 distance bins that cover distances from 0 to 20 Å in steps of 1 Å were employed. Distances up to 20 Å were considered to include information of most pairs of atoms in the descriptor, even for large ligands. For atom typing the modified PATTY atom-types available with the pH4_aType function in MOE (Chemical Computing Group, www.chemcomp.com) were used. This function provides six PPP types: cation, anion, hydrogen-bond acceptor, hydrogen-bond donor, polar (hydrogen-bond acceptor AND hydrogen-bond donor) and hydrophobic. Using 20 distance bins for each of the 21 possible combinations of PPP pairs resulted in a descriptor of 420 dimensions. The values of the descriptor dimensions were finally scaled by the added global occurrences of the two respective features of the PPP pair. An illustration of the descriptor calculation is given in Supplementary Figure 1.



Supplementary Figure 1. Calculation of the CATS3D correlation vector. Atoms are colored according to their pharmacophore atom-type (yellow = hydrophobe, cyan = hydrogen-bond acceptor). Distances are measured between all pairs of atoms, and frequencies of pairs are determined for all pairs of pharmacophore types and for defined distance ranges ("bins"). As an example, a section of the resulting descriptor vector representing hydrogen-bond acceptor – hydrophobe pairs is shown.

Scaffold Analysis

The program Meqi (*M*olecular *eq*uivalence *i*ndices, Pannanugget Consulting, www.pannanugget.com) was used to identify scaffolds in molecular databases. Meqi reduces the molecular representation to the scaffold of a molecule and calculates an equivalence number with a modified version of the Morgan algorithm. Three different definitions of scaffolds were used for the equivalence number calculation: "Murcko-Scaffold", "Carbon Scaffold" and "Reduced Scaffold" (Figure 2). Murcko-scaffolds represent molecules without sidechains, carbon scaffolds also discard information about atom and bond types and reduced scaffolds represent molecules without sidechains indifferent for types of atoms and bonds and or the number of heavy atoms in rings. For Murcko-scaffolds the following settings were used: Hydrogen reduced, Subgraphs: CyclicSystem (leads to sidechain-depleted representation), Topology: Unchanged. For carbon scaffolds we used: Hydrogen reduced, Vertex-labeling.list: C ? (transforms all atoms to C), Edge-labeling list: 1 1 2 3 4 (transforms all bonds to single bonds), Subgraphs: CyclicSystem, Topology: Unchanged. For reduced scaffolds we used: Hydrogen reduced, Vertex-labeling.list: C ? (transforms all atoms to C), Edge-labeling list: 1 1 2 3 4 (transforms all bonds to single bonds), Subgraphs: CyclicSystem, Topology: Reduced (leads to simplified ring systems). Other settings were left at default values.



Supplementary Figure 2. "Murcko-Scaffold" and "Reduced Scaffold" of a molecule.

Feed Forward Neural Networks and Data Preparation

For neural network training on "mGluR5/mGluR1 selectivity" 68 mGluR5 allosteric antagonists from literature, patents and from unpublished results of Merz Pharmaceuticals, and 158 allosteric antagonists of mGluR1 from patents and literature were used. Five maximally diverse sets of 100 molecules that were not described as "active" on either mGluR5 or mGluR1 were compiled from the COBRA database (version 3.12) [P. Schneider, G. Schneider, *QSAR Comb. Sci.* **2003**, *22*, 713] to be used as negative data for the training of "mGluR5 likeness". The diverse selection was done with the program *MaxMinSelection* [M. Schmucker, A. Givehchi, G. Schneider, *Mol. Div.* **2004**, *8*, 421.; www.modlab.de], implementing the MaxMin algorithm [R. W. Kennard, L. A. Stone, *Technometrics* **1969**, *11*, 137]. Prior to selection, all molecules were removed from the COBRA database with a substring "mGluR" in the identifier. The five datasets were merged with the 68 mGluR5 actives for the training of "mGluR5-likeness". For all molecules only a single 3D conformation was calculated with CORINA (Molecular Networks GmbH, www.molecular-networks.com), since we have shown that similarity searching with CATS3D produced already near optimal results using only a single conformation [S. Renner, C. H. Schwab, J. Gasteiger, G. Schneider, *J. Chem. Inf. Model.* **2006**, *46*, 2324]. Since neural network training is sensitive to the presence of large numbers of uninformative dependent variables only descriptor dimensions were used with scaled Shannon entropy values above or equal to 0.3 [F. L. Stahura, J. W. Godden, L. Xue, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245; J. Godden, J. Bajorath, *QSAR Comb. Sci.* **2003**, *22*, 487]. Principle components represent new uncorrelated variables that describe the largest amount of variance of the original variables using as few as possible variables. The aim of this procedure was to train two neural networks on each dataset – one supporting the focus on the dominating variance in the dataset and one retaining the potential to find relevant correlations within descriptors with lower variance. Prior to the calculation of principle components the descriptors were autoscaled.

Training of ANNs was done with the software *profi* [G. Schneider, P. Wrede, *Protein Seq. Data Anal.* **1993**, *5*, 419], implementing an evolution strategy for neural network training. To determine an optimal number of hidden neurons, ten times cross-validation was applied splitting the data in equally sized fractions.

Prediction of properties was done using a jury strategy. Such a strategy is assumed to compensate prediction errors of single networks and results in improved prediction accuracy. Since we found that the trained networks had the tendency to predict large numbers of molecules (from the Enamine set) as "mGluR5 selective" and "mGluR5 like", we decided to use a restrictive jury strategy. We used the average output of all networks, *i.e*. all networks had to agree on a compound to be classified as "good".

Self Organizing Maps (SOM)

Self organizing maps (SOMs) of the predicted active molecules were either trained on MACCS keys (as implemented in MOE) or on CATS3D representations. Only CATS3D-dimensions with scaled Shannon entropy values above 0.3 were used to facilitate the SOM training. The training of SOMs was calculated with the program *som_create* by Gisbert Schneider using 5 x 5 = 25 neurons. Visualization of SOMS was done with the program *som_show* (G. Schneider, unpublished). Representative molecules of a particular neuron were selected by the smallest Euclidean distance to the neuron center (Supplementary Table 1).

Supplementary Table 1. SOM-selected molecules that were tested for bioactivity.

SOM cluster ID	*Molecule number*	*SMILES*
CATS3D-SOM		
A2		Clc1cc([nH0]2c(C)cc(c2C)C(=O)COC(=O)C2CC2)ccc1F
A4		Clc1cccc(NC(=O)/C=C/c2cccc(OC)c2)c1C
A5		s1c2ccccc2[nH0]c1COC(=O)c1ccc2[nH]c(C)c(C)c2c1
B1		Clc1ccc(cc1[N+](=O)[O-])C(=O)OCC(=O)c1cc([nH0](c2ccc(C)cc2)c1C)C
B2		O=[N+]([O-])c1ccccc1C(=O)OCC(=O)c1cc([nH0](CCC)c1C)C
B4	**9**	O=[N+]([O-])c1c(C)cccc1C(=O)OCC(=O)N1CCOCC1
B5		S1C=CN2C1=NC(=CC2=O)COC(=O)CCSc1ccc(F)cc1
C1	**10**	Clc1ccc(cc1[N+](=O)[O-])/C=C/c1scc([nH0]1)c1ccc(F)cc1
C3	**6**	s1c2CCC(C)Cc2cc1C(=O)OCc1c(o[nH0]c1C)C
C5		Clc1cccc(c1)C(=O)COC(=O)c1cccc([N+](=O)[O-])c1C
D1	**11**	Fc1ccc(cc1)CN(C)C(=O)/C=C/c1cccc([N+](=O)[O-])c1
D3		Brc1cccc(c1)/C=C/C(=O)c1ccc(F)c(F)c1
D4		Clc1ccc(NN=Cc2ccc(C)cc2C)cc1Cl
E3		Clc1ccc(cc1)C(=O)/C=C/c1cccc(OC)c1
E4		Ic1ccc(NC(=O)c2cc(C)cc(C)c2)cc1
E5		Fc1ccc(NC(=O)/C=C/c2cccc([N+](=O)[O-])c2)c(C)c1
MACCS-SOM		
A5		O=C(Nc1ccc(C)cc1C)CN1CCCCC1
B1		s1cc([nH0]c1NC(=O)CSc1ccccc1)C
B2		s1c2ccccc2[nH0]c1/C=C/c1ccccc1
B3		Clc1ccc(cc1)/C=C/C(=O)c1ccccc1Cl
B4		Clc1ccc(Cl)c(c1)C(=O)Nc1ccc(C)cc1
B5		Clc1cccc(Cl)c1SCC(=O)Nc1ccc(C)cc1
C1		S(Cc1ccc(cc1)c1ccccc1)c1[nH0][nH0]c[nH0]1c1ccccc1
C2	**8**	s1cc([nH0]c1SCC(=O)c1cc([nH0](c2ccccc2)c1C)C)C
C4		Clc1cc(Cl)cc(NC(=O)c2cc([N+](=O)[O-])ccc2Cl)c1
C5		s1c2cc([N+](=O)[O-])ccc2[nH0]c1/C=C/c1ccccc1[N+](=O)[O-]
D1		Clc1[nH0]([nH0]c(C)c1C(=O)OCC(=O)c1cc(C)ccc1C)Cc1ccccc1
D2		O=C(OCC(=O)c1cc([nH0](C)c1C)C)c1cc(C)cc(C)c1
D5		O=[N+]([O-])c1cccc(c1)C(=O)OCC(=O)c1cc([nH0](c2ccccc2)c1C)C
E1		Clc1[nH0]([nH0]c(C)c1C(=O)OCC(=O)N1CCc2ccccc12)Cc1ccccc1
E2	**12**	Clc1[nH0]cccc1C(=O)OCC(=O)c1cc(C)ccc1C
E3	**7**	O=C(OC)c1oc2ccccc2c1COc1ccccc1
E4		O=C(Nc1ccc(OCC)cc1)c1oc2ccccc2c1

2. Experimental

Astrocyte culture

Primary astrocyte cultures were prepared from cortices of newborn rats as described [J. Booher, M. Sensenbrenner, *Neurobiology* **1972**, *2*, 97]. Briefly, Sprague-Dawley rat pups (2 - 4 d old) were decapitated and neocortices were dissected, disintegrated with a nylon filter (poresize 80 µm) and carefully triturated. The cell suspension was plated on poly-D-lysine precoated flasks (Costar) and cultivated in Dulbecco's Modified Eagle's Medium (DMEM, Biochrom) supplemented with 10% heat inactivated fetal calf serum (FCS, Sigma), 4 mM glutamine (Biochrom) and 50 µg/mL gentamycin (Biochrom) at 37°C in a humidified atmosphere of 5% CO_2/95% air for 7 d with exchanging the medium at day 2.
After 7 DIV, cells were shaken overnight at 250 rpm to remove oligodendrocytes and microglia. The next day, astrocytes were rinsed twice with CMF-PBS, trypsinized and subplated on poly-D-lysine precoated 96-well plates (Becton Dickinson) at a density of 40,000 cells/well. 24 h after establishing the secondary culture the astrocytes were rinsed with PBS and fed with astrocyte-defined medium (ADM) consisting of DMEM containing 1x G5-supplement (Invitrogen), 0.5 µg/mL heparan sulfate (Sigma), and 1.5 µg/ mL fibronectin (Sigma) (Miller *et al*., 1993). 3 d later the medium was exchanged and the cells incubated for another 2-3 d, so that at the time of experiments astrocytes were 14-15 DIV.

Calcium FLIPR studies

Functional responses of the receptors were studied using a fluorometric imaging plate reader (FLIPR tetra) in combination with the Ca-Kit (both Molecular Devices) according to the manufacturer's recommendations. Antagonists were pre-incubated for 5 or 10 min (CHO-cells and astrocytes, resp.) before adding the agonist L-quisqualate (final DMSO concentration 0.5%). CHO cells used were transfected with the receptor of interest. Agonist concentration was chosen to give 80% of maximal response (EC_{80}). The transient rise of the fluorescence signal (RFU) reflects an increase of intracellular calcium. Baseline before agonist addition was subtracted from fluorescence maxima after agonist addition. The spatial uniformity correction of the FLIPR software was used as a correction for well to well variations in cell densities. Values were expressed as percentage of control (agonist alone). Dose response curves shown were performed by using Prism 4 (GraphPad Software Inc.).

Chemicals

Unless otherwise stated all chemicals were purchased from Sigma.

IC_{50}-value Estimation

IC_{50} values were estimated from the % of control values from the scintillation assay with a four parameter logistic equation. If both the radio-ligand and the competitor reversibly bind to the same binding site, binding at equilibrium follows Eq. 1.

$$y = \frac{100\%}{1+(\frac{x}{IC_{50}})^{s}} \quad . \qquad \text{(Eq. 1)}$$

If s = 1, Eq. 1 can be reformulated (Eq. 2):

$$IC_{50} = \frac{x}{(\frac{100\%}{y}-1)} \quad \text{, where} \qquad \text{(Eq. 2)}$$

s = slope factor = 1;
x = concentration of test compound [µM] in the assay;
y = result of the binding assay for the test compound [% of control].

Solubility Determination

Two solutions of 10 µM compound with 0.5% DMSO were prepared for the measurement of solubility: solution A with methanol and deionized water (1:1) and solution B with Tris-buffer (50 mM, pH 7.5). HPLC was performed with a Hewlett-Packard Series 1100 HPLC device with diode-array detector (Agilent Technologies, Waldbronn, Germany), and Symmetry C18 column (Waters Corporation, Milford, MA). Both solutions were measured separately with an average pressure of 190 atm. The peaks were compared at the wavelength with the maximum area under the curve (AUC) of the both peaks of solutions A and B. The percentage of the AUC of solution B relative to the AUC of solution A was used as solubility.

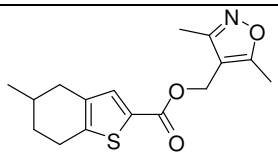
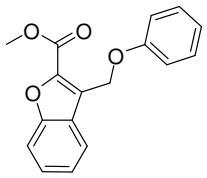
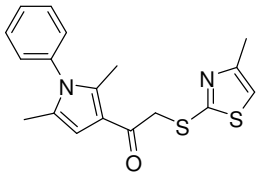
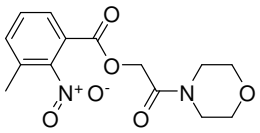
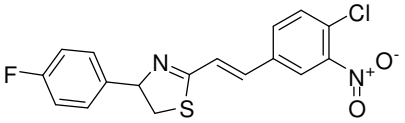
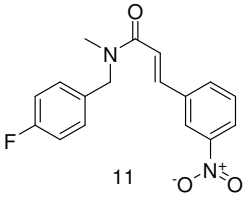
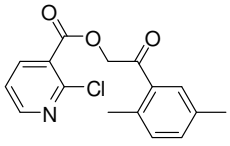
Dynamic Light Scattering

Compounds were prepared as solutions of 10 µM in the buffer used in the mGluR5 functional assay. Measurements were preformed with a Zetasizer Nano S (Malvern Instruments GmbH, Herrenberg, Germany). All measurements were repeated three times at 25°C.

Selection of active molecules and calculation of IC_{50} values

Since we observed that functional activity was not always reproducible using single point measurements, all functional measurements were repeated in independent experiments. Only molecules that showed an effect in more than one experiment were considered as "active". The threshold for activity was set to 85% remaining activity. The data for active molecules is shown in Supplementary Table 2.

Supplementary Table 2. Assay results.
*) IC_{50} values were corrected by solubility; *n.d.*: not determined.

Structure	rmGluR5 Astrocytes	hmGluR1 CHO	rmGluR1 CHO	Solubility (0.5% DMSO)
 6	9.6*	2.5	3.9	44%
 7	14.5*	> 100	> 100	78%
 8	20.7*	4.6	5.4*	47%
 9	41.7*	> 100	> 100	98%
 10	44.4*	> 100	*n.d.*	100%
 11	49.8*	> 100	37.4*	98%
 12	> 100	24.3	59.6*	100%