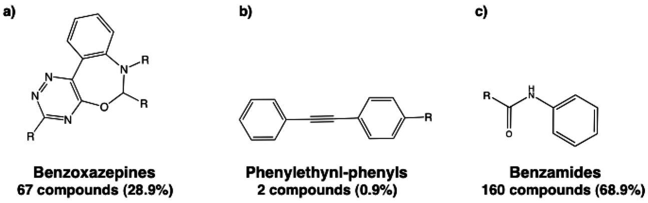
# Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening

Ralf Mueller,[†,⊥] Alice L. Rodriguez,[‡,⊥] Eric S. Dawson,[‖] Mariusz Butkiewicz,[†] Thuy T. Nguyen,[‡] Stephen Oleszkiewicz,[†] Annalen Bleckmann,[†] C. David Weaver,[‡,§] Craig W. Lindsley,[†,‡,§] P. Jeffrey Conn,[‡,§] and Jens Meiler*[,†,‡,§,‖]

[†]Department of Chemistry, [‡]Department of Pharmacology, [§]Institute for Chemical Biology, and [‖]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37232-6600

## Abstract



Selective potentiators of glutamate response at metabotropic glutamate receptor subtype 5 (mGluR5) have exciting potential for the development of novel treatment strategies for schizophrenia. A total of 1,382 compounds with positive allosteric modulation (PAM) of the mGluR5 glutamate response were identified through high-throughput screening (HTS) of a diverse library of 144,475 substances utilizing a functional assay measuring receptor-induced intracellular release of calcium. Primary hits were tested for concentration-dependent activity, and potency data ($EC_{50}$ values) were used for training artificial neural network (ANN) quantitative structure−activity relationship (QSAR) models that predict biological potency from the chemical structure. While all models were trained to predict $EC_{50}$, the quality of the models was assessed by using both continuous measures and binary classification. Numerical descriptors of chemical structure were used as input for the machine learning procedure and optimized in an iterative protocol. The ANN models achieved theoretical enrichment ratios of up to 38 for an independent data set not used in training the model. A database of ∼450,000 commercially available drug-like compounds was targeted in a virtual screen. A set of 824 compounds was obtained for testing based on the highest predicted potency values. Biological testing found 28.2% (232/824) of these compounds with various activities at mGluR5 including 177 pure potentiators and 55 partial agonists. These results represent an enrichment factor of 23 for pure potentiation of the mGluR5 glutamate response and 30 for overall mGluR5 modulation activity when compared with those of the original mGluR5 experimental screening data (0.94% hit rate). The active compounds identified contained 72% close derivatives of previously identified PAMs as well as 28% nontrivial derivatives of known active compounds.

**Keywords:** Metabotropic glutamate receptor (mGluR), allosteric modulator, high-throughput screen (HTS), artificial neural network (ANN), cheminformatics, virtual screen, quantitative structure−activity relationship (QSAR)

Glutamate is the primary excitatory neurotransmitter in the mammalian central nervous system (CNS) and activates metabotropic glutamate receptors (mGluRs), which are coupled to downstream effector systems through guanine nucleotide binding proteins (G proteins) (*1*, *2*). The mGluRs provide a mechanism by which glutamate can modulate or fine-tune activity at the same synapses on which it elicits fast synaptic responses. Because of the wide diversity, heterogeneous distribution, and diverse physiological roles of mGluR subtypes, the opportunity exists for developing therapeutic agents that selectively interact with mGluRs involved in only one or a limited number of CNS functions. Such drugs could have a dramatic impact on the development of novel treatment strategies for a variety of psychiatric and neurological disorders including depression (*3*), anxiety disorders (*4*, *5*), schizophrenia (*6*−*9*), chronic pain (*10*), epilepsy (*11*), Alzheimer's disease (*12*), and Parkinson's disease (*13*). The mGluR5 receptor subtype is a closely associated signaling partner of the ionotropic NMDA receptor (NMDAR) and may play a significant role in setting the tone of NMDAR function in the forebrain regions containing neuronal circuits important for cognitive behavior and for reporting on the efficacy of antipsychotic agents (*6*).

## Activators of mGluR5 May Provide a Novel Approach to the Treatment of Schizophrenia

Activation of mGluR5 potentiates NMDAR function in forebrain circuits thought to be disrupted in

**Received Date:** August 14, 2009
**Accepted Date:** January 4, 2010
**Published on Web Date:** January 28, 2010

schizophrenia. The mGluR5 selective allosteric antagonist [2-methyl-6-(phenylethynyl)-pyridine] MPEP potentiates the effect of the noncompetitive NMDAR antagonist phencyclidine (PCP) in behavioral phenotypic assays (*14*−*16*), and mGluR5 knockout mice have deficits in prepulse inhibition in acoustic startle response behavioral assays compared with those of wild type mice (*14*, *17*). Positive allosteric modulators of mGluR5 have recently been developed and reported (*18*−*22*). Four well-characterized structural classes of mGluR5 allosteric potentiators have been identified, including benzaldazine derivatives [3,3-difluorobenzaldazine] (DFB), two types of benzamides, [*N*-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2*H*-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide] (CPPHA) and [3-cyano-*N*-(1,3-diphenyl-1*H*-pyrazol-5-yl)benzamide] (CDPPB), and an oxadiazole chemotype represented by ADX-47273 (*23*−*26*). Despite striking functional similarities, radioligand binding studies revealed different mGluR5 binding profiles for DFB and CDPPB compared with those of CPPHA (*19*, *21*, *27*). Both CDPPB (*20*, *21*) and ADX-47273 (*24*−*26*) have displayed *in vivo* efficacy in behavioral models. Unfortunately, lead optimization of the CDPPB scaffold was unable to address a number of issues including poor physiochemical properties due to the lack of solubility in many vehicles (*22*). However, some improvement of physicochemical properties was recently reported for the mGluR5 ago-potentiator ADX-47273 (*23*). Recent reports have also shown that small structural modifications to related compounds in a series including benzaldazine and (phenethynyl)pyrimidine scaffolds can bind to a single allosteric site to exert effects ranging from partial to full antagonism to positive allosteric modulation (*18*, *28*, *29*). For these reasons, further validation of mGluR5 potentiation as a therapeutic approach to Schizophrenia requires the discovery of novel chemotypes possessing improved physiochemical and pharmacological properties.

## High-Throughput Screening in Drug Discovery

High-throughput screening (HTS) is the process of testing a large number of diverse chemical structures against potential disease targets to identify new potential lead compounds by taking a rapid, high efficiency approach to the generation of ligand−target interaction data sets (*30*, *31*). More than 120 GPCR-based HTS assays have been published in PubChem (pubchem.ncbi.nlm.nih.gov). For example, 63,676 compounds were screened at Vanderbilt in an assay for allosteric agonist activity at acetylcholine Muscarinic M1 Receptor to identify 309 confirmed M1 agonists (PubChem Bioassay number AID626 (primary screen) and AID1488 (confirmatory screen)). Increased throughput GPCR screens using 1,536 well format have recently been reported for targets such as M1 acetylcholine receptor (*32*) and 5HT2b serotonin receptor (*33*). However, the current literature suggests that one marketable drug emerges from the information gained by screening approximately one million compounds (*31*). If fewer compounds could be tested without compromising the probability of success, screening cost and time as well as failure rates in clinical testing may be reduced (*30*, *31*, *34*).

## Quantitative Structure Activity Relations in Drug Discovery

Quantitative structure activity relations (QSAR) attempt to model complex nonlinear relationships between the chemical and physical properties of molecules and their biological activity (*35*, *36*). Hansch et al. established classical QSAR analysis as a paradigm by reporting the use of Hammett substituent constants to establish a quantitative relationship between electron density and biological activity (*37*). At the same time, they introduced a new hydrophobic parameter, the partition coefficient (*P*) of the compound in a 1-octanol−water system (log *P*). Variations and extensions of the Hansch analysis have been applied to drug discovery for over 40 years and rely on well-studied scalar or 2D descriptors such as calculated log *P* (*c* log *P*), molecular refractivity (CMR), and topological polar surface area (TPSA). Modern QSAR techniques employ advanced 2D molecular fingerprints and 3D molecular descriptors coupled with machine learning (*38*−*40*). High-resolution methods such as comparative molecular field analysis (CoMFA) (*41*, *42*) and comparative molecular similarity indices analysis (COMSIA) (*43*) require the alignment of biologically relevant 3D conformations of molecules with a common substructure to generate a map of regions important for the structure−activity profile of a given related series of molecules.

## Numerical Descriptors of Chemical Structure for QSARs

Encoding schemes that are fragment-based usually identify a common fragment in small focused chemical libraries, and chemical modifications to that fragment (common substructure) are numerically encoded (the size of a substituent in position A, the presence of a negatively charged group in position B, atom type in heterocycle C, etc.). Examples of fragment-based strategies include MACCS (*44*, *45*), binary structural keys based on occurrence/counts of up to 166 different chemical features found in a compound; HQSAR (*46*−*48*), a 2D method for capturing chiral information based on a molecular hologram hashing algorithm without the requirement for the generation of 3D coordinates; and SKEYS/FRED, a combination of MDL structural key based fingerprints with an evolutionary algorithm (*49*).

Traditional 2D- and 3D-QSAR methods often require fragment-based structural encoding schemes (*44*, *50*) or conformational superposition of biologically active conformations of the chemical structures (*42*, *51*, *52*) that may restrict the utility of resulting models to predictions related to single chemotypes (*50*) or single protein binding sites (*50*, *52*). While suitable for optimization of a lead structure in a small focused library, such encoding schemes often preclude the analysis of large, diverse databases as a large majority of the substances in such a database will not share a large common fragment.

### Fragment-Independent Transformation-Invariant Descriptor Schemes

Fragment-independent molecular descriptors have the potential to encode a large diversity of chemical scaffold information into mathematical representations not sensitive to scaffold size, composition, and rotation/translation of 3D coordinate molecule representations. The use of feature point pharmacophores (FEPOPS), an automated method that simplifies flexible 3D chemical descriptions, was recently reported to outperform traditional 2D- and 3D-QSAR methods for enrichment of actives taken from high-throughput screening compound collections (*52*) and to identify novel chemotypes with biological activity at query targets from virtual screens (*53*). A recent study of HIV-1 integrase inhibitors introduced atom-type linear indices of the molecular pseudograph atom adjacency matrix as fragment-independent indices containing important structural information to be used in QSAR and drug design studies (*54*). Radial distribution functions have recently been shown to outperform traditional fragment-based molecular descriptors in a study of the chick intestinal vitamin D receptor affinity of 49 vitamin D analogues (*55*) and in an investigation to separate the activity of carcinogenic and noncarcinogenic compounds in a rodent toxicity model (*56*). Autocorrelation functions are fragment independent, invariant to translation and rotation, and encode the identity and electronic attributes of molecular structure including atom types, partial atomic charges, electronegativity, and polarizability into vector representations (*57*). Several studies have employed autocorrelation descriptors for training machine learning algorithms for applications including separation of dopamine agonists and benzodiazepine receptor agonists (*58*), virtual screening for chemical library enumeration (*59*), and identification of novel chemotypes (*60*). Surface area correlation functions store molecular shape geometry for molecules with known biological activity into neural networks for shape-based molecular recognition in external data sets, as reported for the analysis of corticosteroid-binding globulin activity of steroids (*61*). Self-organizing neural networks using molecular electrostatic potential as the structural encoding scheme were also successfully applied to study structurally different classes of muscarinic acetylcholine receptor allosteric modulators (*62*).

### Application of Machine Learning Algorithms to Establish QSARs

Machine Learning algorithms have proven to be of practical value for approximating nonlinear separable data, especially for classifying biological target data (*39*, *63*). Recently, a machine learning approach was applied to generate a model for the tubulin polymerization activities of a library of 250 analogues of the anticancer drug Epothilone (*38*). ANNs have been successfully applied for many years in chemistry and biochemistry to generate QSAR models (*40*, *64*, *65*). Studies were reported involving the prediction of dihydrofolate reductase inhibition based on data derived from high-throughput screening using preclustering and evolved neural networks (*66*) as well as applications for prescreening compounds for HIV inhibition while optimizing specificity and potency (*67*). Our group recently published a theoretical comparison of machine learning techniques for the identification of compounds that are predicted allosteric modulators of the mGluR5 glutamate response (*68*).

### Quantitative Structure-Activity Relation Models for mGluR5 Positive Allosteric Modulation

The objective of the present research is to employ ANNs to develop QSAR models for mGluR5 PAM activity. QSAR models capable of combining the structural diversity of different chemical scaffolds into a single model could inform the discovery of new chemotypes for allosteric potentiation of the mGluR5 glutamate response. Such models may also be useful for the identification of compounds with a spectrum of activity (agonists, antagonists, and allosteric potentiators) by analogy to the well-documented activities of agonists, inverse agonists, and neutral antagonists at orthosteric binding sites on a broad range of receptors (*18*, *28*, *29*). Activity data for mGluR5 PAMs obtained from a high-throughput screen of ~150,000 compounds is used to develop the QSAR model. A set of fragment-independent and transformation-invariant chemical descriptors serves as input for the ANN. A novel strategy for the selection of an optimal descriptor subset yields QSAR models that enrich active compounds by a factor of up to 38 in independent data sets. The method is applied to a virtual screen of a commercial library of ~450,000 available compounds. A set of 824 compounds with predicted mGluR5 PAM activity containing multiple chemical scaffolds was experimentally tested.

## Results and Discussion

Machine learning techniques were applied to generate specific QSAR models for allosteric potentiation of the mGluR5 glutamate response. These models were then used to prioritize compounds for acquisition with the aim of enhancing both the speed and diversity of hit-to-lead discovery efforts for mGluR5 positive allosteric modulators (PAMs).

## Concentration Response Curves in the Experimental High-Throughput Screen

Concentration response curves were generated from the averaged data of three experiments using a four point logistical equation, $a + b/[1 + (x/c)^d]$. No parameters were constrained and no values were weighted. Points corresponding to concentrations of PAM exhibiting an agonist effect were excluded from the analysis. For a PAM with excellent potency ($EC_{50}$ value below 100 nM), 95% confidence intervals were on average within a range of 30 nM. For a PAM with moderate potency ($EC_{50}$ value roughly 100 nM to 1 $\mu$M), confidence intervals were within a range of 300 nM. For a PAM with low potency ($EC_{50}$ value above 1 $\mu$M), 95% confidence intervals were generally within a range of 1.5 $\mu$M. Weak PAMs whose concentration response curve did not reach a plateau but did significantly enhance a glutamate $EC_{20}$ were categorized as PAMs, but fit statistics were not determined. A summary of fit statistics and a concentration response curve for one example of each of the major scaffolds identified including benzoxazepine, phenylethynyl-phenyl, and benzamide PAMs is detailed in the Supporting Information (Figure S1 and Table S1).

## Input Sensitivity Is a Reliable Measure to Prioritize Descriptors

The selection of input descriptors with highest input sensitivity reduces the degrees of freedom within the ANN model and results in models with substantially improved prediction capability. The input sensitivity can be understood as the partial derivative of each input with respect to the output of the ANN (see Methods). The main reason for this improvement is the reduction of noise through the increased ratio of data sets versus weights. An increased ratio of data sets versus weights leads to more information available to fit every degree of freedom. Each degree of freedom can be determined more precisely despite the intrinsic noise of HTS data used for training. Since several of the *ADRIANA* molecular descriptors (see Methods) encode the same chemical property with different encoding functions, it seems plausible that information in these descriptors is redundant and therefore does not add to the determination of the optimal solution.

## Optimization of Molecular Descriptor Set Improves the Prediction Accuracy of the ANN Model

To obtain a baseline for descriptor optimization, an ANN was trained using only the scalar descriptors 1−8 (Table 1). The root-mean-square deviation (*rmsd*) (see eq 1) value for the independent data set of 0.228, area under the receiver operating characteristic curve (*auc*) value of 0.673, and enrichment of active compounds relative to inactive compounds value of 6 served as a basis for comparison in model optimization (Table 2). For a definition of these measures, see Methods. The individual sensitivity value for $X \log P$ (0.97) remained the highest in the baseline network with the remaining input sensitivity distributed across the other scalar descriptors (Figure 1a). Keeping the scalar descriptors in the following models allowed one to compare their sensitivity with this baseline.

$$rmsd = \sqrt{\frac{\sum_{i=1}^{n} (\exp_i - pred_i)^2}{n}} \qquad (1)$$

The most sensitive 428 descriptors in 14 categories were retained for additional iterations of descriptor optimization. Retraining of the ANN with 428 descriptors (iteration 1) yields significantly improved metrics relative to the baseline model (scalar only) including an *rmsd* value for the independent data of 0.214, an *auc* value of 0.731, and an *enrichment* of 36 (Table 2). To further optimize the set of descriptors, the least sensitive descriptor categories were systematically removed in an iterative process (Table 2 iterations 2−6; Figure 1a). In particular, the enrichment measure is substantially improved with respect to the scalar only baseline ANN as emphasized by the initial slope of the ROC curves in Figure 2.

Iterations 1−4 remove 152 descriptors to yield a set of 276 descriptors including the eight scalar descriptors, the 3D autocorrelation lone pair electronegativity, and the radial distribution functions for lone pair electronegativity and $\pi$-electronegativity (Table 2 and Figure 1a). Retraining of the ANN with 276 descriptors yields an *rmsd* value for the independent data of 0.212, an *auc* value of 0.757, and an *enrichment* of 38.

In the last two iterations 5 and 6, the radial distribution function for $\pi$-electronegativity and the 3D autocorrelation function for lone-pair electronegativity were removed (Figure 1a and Figure 2). In iteration 5, the ANN with 148 descriptors failed to improve the model as indicated by an *rmsd* value for the independent data of 0.217, an *auc* value of 0.738, and an *enrichment* of 25 (Table 2). In iteration 6, the ANN with 136 descriptors had similar quality measures.

At this point, the iterative descriptor optimization procedure was terminated. The ANN model from iteration 4 with 276 input descriptors is considered to be the optimal model as it displays optimal performance on the independent data set combined with the smallest descriptor set. This network was used in all of the *in silico* screening experiments described below.

Table 1. Summary of 1,252 Molecular Descriptors in 35 Categories Computed with *ADRIANA*

| | description method | description property | abbreviation | number |
|---|---|---|---|---|
| 1 | scalar descriptors | molecular weight of compound | Weight | 1 |
| 2 | | number of hydrogen bonding acceptors | HDon | 1 |
| 3 | | number of hydrogen bonding donors | HAcc | 1 |
| 4 | | octanol/water partition coefficient in [log units] | XlogP | 1 |
| 5 | | topological polar surface area in [Å$^2$] | TPSA | 1 |
| 6 | | mean molecular polarizability in [Å$^3$] | Polariz | 1 |
| 7 | | dipole moment in [Debye] | Dipol | 1 |
| 8 | | solubility of the molecule in water in [log units] | LogS | 1 |
| 9 | 2D autocorrelation | atom identities | 2DA_Ident | 11 |
| 10 | | $\sigma$ atom charges | 2DA_SigChg | 11 |
| 11 | | $\pi$ atom charges | 2DA_PiChg | 11 |
| 12 | | total charges | 2DA_TotChg | 11 |
| 13 | | $\sigma$ atom electronegativities | 2DA_SigEN | 11 |
| 14 | | $\pi$ atom electronegativities | 2DA_PiEN | 11 |
| 15 | | lone pair electronegativities | 2DA_LpEN | 11 |
| 16 | | effective atom polarizabilities | 2DA_Polariz | 11 |
| 17 | 3D autocorrelation | atom identities | 3DA_Ident | 12 |
| 18 | | $\sigma$ atom charges | 3DA_SigChg | 12 |
| 19 | | $\pi$ atom charges | 3DA_PiChg | 12 |
| 20 | | total charges | 3DA_TotChg | 12 |
| 21 | | $\sigma$ atom electronegativities | 3DA_SigEN | 12 |
| 22 | | $\pi$ atom electronegativities | 3DA_PiEN | 12 |
| 23 | | lone pair electronegativities | 3DA_LpEN | 12 |
| 24 | | effective atom polarizabilities | 3DA_Polariz | 12 |
| 25 | radial distribution function | atom identities | RDF_Ident | 128 |
| 26 | | $\sigma$ atom charges | RDF_SigChg | 128 |
| 27 | | $\pi$ atom charges | RDF_PiChg | 128 |
| 28 | | total charges | RDF_TotChg | 128 |
| 29 | | $\sigma$ atom electronegativities | RDF_SigEN | 128 |
| 30 | | $\pi$ atom electronegativities | RDF_PiEN | 128 |
| 31 | | lone pair electronegativities | RDF_LpEN | 128 |
| 32 | | effective atom polarizabilities | RDF_Polariz | 128 |
| 33 | surface autocorrelation | molecular electrostatic potential | Surf_ESP | 12 |
| 34 | | hydrogen bonding potential | Surf_HBP | 12 |
| 35 | | hydrophobicity potential | Surf_HPP | 12 |
| | total | | | 1252 |

The rationale for keeping the scalar descriptors with lower sensitivity throughout descriptor optimization is to maintain comparability with the baseline established by training with these eight descriptors alone (read below). These parameters relate to Lipinski's Rule of Five (*69*) and therefore are widely accepted criteria for drug-like compounds. Note that the scalar descriptors represent only 0.6% of all descriptors. Removal of scalar descriptors will therefore not decrease the complexity of the ANN model.

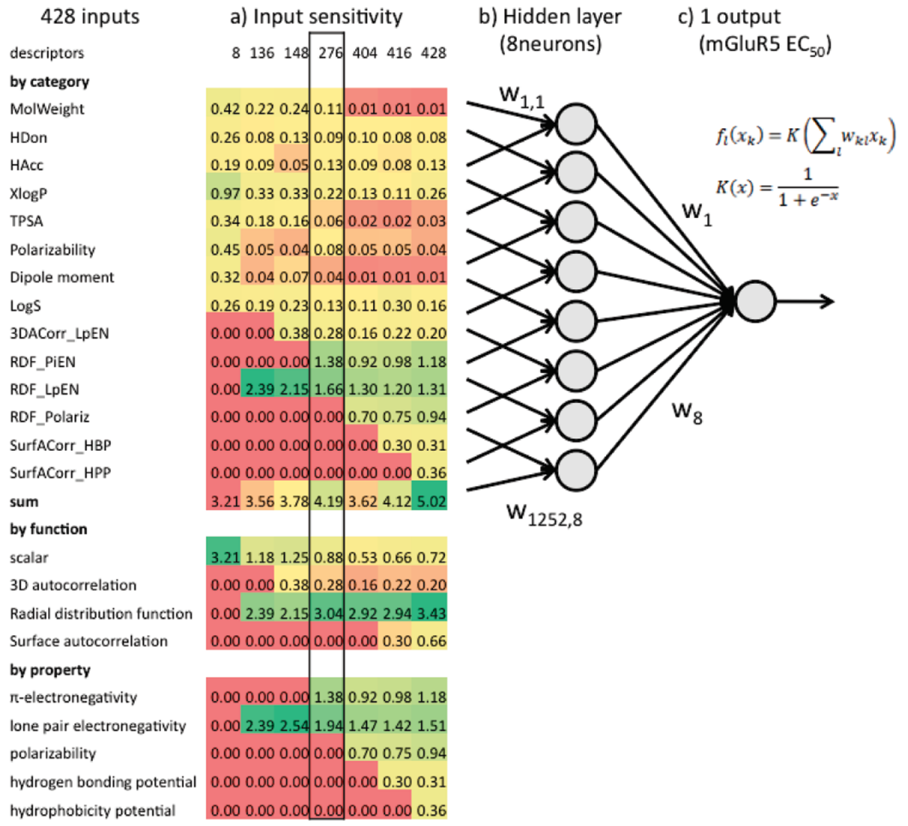
## Balancing through Oversampling Yields Better Results than Two Undersampling Strategies

The oversampling strategy employed throughout the study (see Methods) was compared with two approaches that undersample inactive compounds when using the optimized 276 input descriptors (Figure 3). Usage of randomly chosen inactive compounds resulted in an *rmsd* value for the independent data of 0.221, an *auc* value of 0.753, and an *enrichment* of 8. Determining inactive compounds for undersampling maximally similar to the active compounds yields in an *rmsd* value for the independent data of 0.261, an *auc* value of 0.654, and an *enrichment* of 2 (Table 2).

Our interpretation of this finding is that our models do not so much recognize active compounds but rather filter out inactive compounds. Hence, detailed knowledge of the entire space of inactive compounds improves performance of the models in binary classification settings. Random selection of a small fraction of inactive compounds reduces the space of inactive compounds

Table 2. The *rmsd*, *auc*, and *enrichment* Values for All QSAR Models

| iteration | number and type of descriptors | | *rmsd* | | | *auc* | *enrichment* |
|---|---|---|---|---|---|---|---|
| | | | train | monitor | independent | | |
| all | 1252 | 1−35 | 0.196 | 0.248 | 0.248 | 0.701 | 10 |
| scalar | 8 | 1−8 | 0.223 | 0.224 | 0.228 | 0.673 | 6 |
| 1 | 428 | 1 − 8, 23, 30−32, 34, 35 | 0.196 | 0.212 | 0.214 | 0.731 | 36 |
| 2 | 416 | 1−8, 23, 30−32, 34 | 0.193 | 0.213 | 0.216 | 0.742 | 38 |
| 3 | 404 | 1−8, 23, 30−32 | 0.191 | 0.214 | 0.214 | 0.731 | 36 |
| **4** | **276** | **1−8, 23, 30, 31** | **0.185** | **0.215** | **0.212** | **0.757** | **38** |
| 5 | 148 | 1−8, 23, 31 | 0.203 | 0.215 | 0.217 | 0.738 | 25 |
| 6 | 136 | 1−8, 31 | 0.204 | 0.214 | 0.217 | 0.742 | 25 |
| method | | | | | | | |
| binary | 276 | 1−35 | 0.334 | 0.370 | 0.385 | 0.744 | 26 |
| undersampled | | | | | | | |
| -random | 276 | 1−8, 23, 30, 31 | 0.202 | 0.226 | 0.221 | 0.757 | 8 |
| -MACCS | 276 | 1−8, 23, 30, 31 | 0.171 | 0.195 | 0.217 | 0.654 | 2 |



428 inputs

a) Input sensitivity

| descriptors | 8 | 136 | 148 | 276 | 404 | 416 | 428 |
|---|---|---|---|---|---|---|---|
| **by category** | | | | | | | |
| MolWeight | 0.42 | 0.22 | 0.24 | 0.11 | 0.01 | 0.01 | 0.01 |
| HDon | 0.26 | 0.08 | 0.13 | 0.09 | 0.10 | 0.08 | 0.08 |
| HAcc | 0.19 | 0.09 | 0.05 | 0.13 | 0.09 | 0.08 | 0.13 |
| XlogP | 0.97 | 0.33 | 0.33 | 0.22 | 0.13 | 0.11 | 0.26 |
| TPSA | 0.34 | 0.18 | 0.16 | 0.06 | 0.02 | 0.02 | 0.03 |
| Polarizability | 0.45 | 0.05 | 0.04 | 0.08 | 0.05 | 0.05 | 0.04 |
| Dipole moment | 0.32 | 0.04 | 0.07 | 0.04 | 0.01 | 0.01 | 0.01 |
| LogS | 0.26 | 0.19 | 0.23 | 0.13 | 0.11 | 0.30 | 0.16 |
| 3DACorr_LpEN | 0.00 | 0.00 | 0.38 | 0.28 | 0.16 | 0.22 | 0.20 |
| RDF_PiEN | 0.00 | 0.00 | 0.00 | 1.38 | 0.92 | 0.98 | 1.18 |
| RDF_LpEN | 0.00 | 2.39 | 2.15 | 1.66 | 1.30 | 1.20 | 1.31 |
| RDF_Polariz | 0.00 | 0.00 | 0.00 | 0.00 | 0.70 | 0.75 | 0.94 |
| SurfACorr_HBP | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.31 |
| SurfACorr_HPP | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.36 |
| **sum** | 3.21 | 3.56 | 3.78 | 4.19 | 3.62 | 4.12 | 5.02 |
| **by function** | | | | | | | |
| scalar | 3.21 | 1.18 | 1.25 | 0.88 | 0.53 | 0.66 | 0.72 |
| 3D autocorrelation | 0.00 | 0.00 | 0.38 | 0.28 | 0.16 | 0.22 | 0.20 |
| Radial distribution function | 0.00 | 2.39 | 2.15 | 3.04 | 2.92 | 2.94 | 3.43 |
| Surface autocorrelation | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.66 |
| **by property** | | | | | | | |
| π-electronegativity | 0.00 | 0.00 | 0.00 | 1.38 | 0.92 | 0.98 | 1.18 |
| lone pair electronegativity | 0.00 | 2.39 | 2.54 | 1.94 | 1.47 | 1.42 | 1.51 |
| polarizability | 0.00 | 0.00 | 0.00 | 0.00 | 0.70 | 0.75 | 0.94 |
| hydrogen bonding potential | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.31 |
| hydrophobicity potential | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.36 |

Figure 1. Schematic view of an ANN: (a) Up to 1,252 descriptors (from 35 categories) are fed into the ANN input layer. (b) The weighted sum of the input data is modified by the activation function and serves as input to the next layer. (c) The output predicts the biological activity of the input molecule on the basis of complex nonlinear relationships derived from machine learning through iterative ANN model training. Panel (a) displays input sensitivities for iterations 1−6 as a heat map from least sensitive (red) to most sensitive (green). The final optimized ANN model with 276 descriptors is highlighted by a black frame.

substantially; targeted selection of inactive compounds similar to active compounds reduces the space even more. The model loses the ability to classify molecules dissimilar from active compounds.

## Radial Distribution Functions and Electronegativity Contribute Most to an Accurate Prediction

Analysis of input sensitivity by encoding functions (3D autocorrelation, radial distribution functions, and

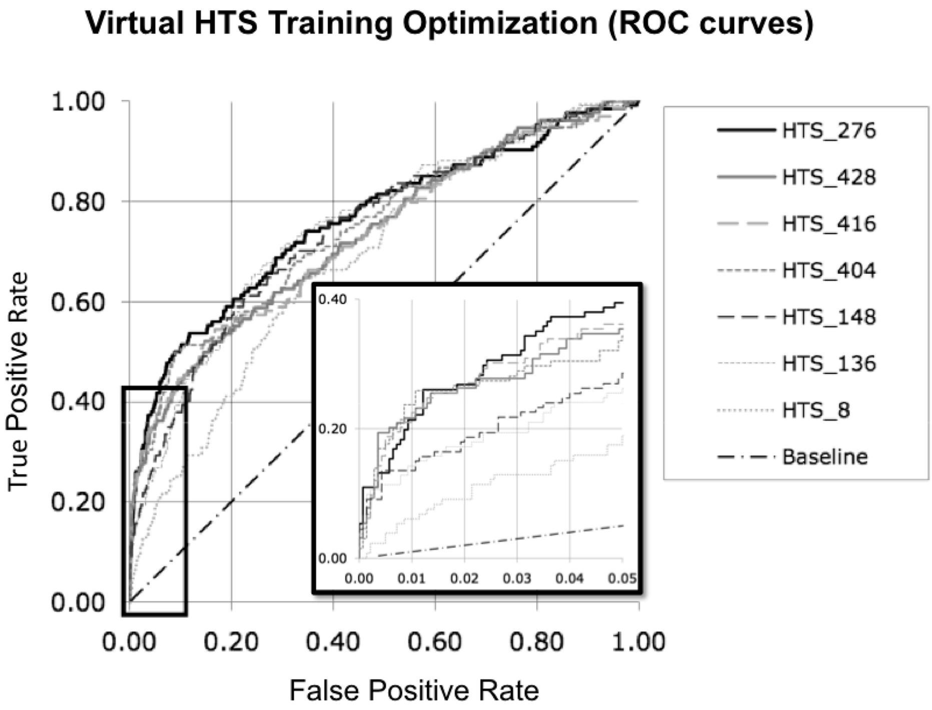


**Figure 2.** Receiver operating characteristic (ROC) curve plots. Traditional (8) scalar QSAR descriptors (HTS_8, dotted gray line) were compared to groups of *ADRIANA* scalar and vector descriptor sets from the input sensitivity analysis (see Figure 4a) by plotting ROC curves to examine the initial slope. The descriptor set was systematically reduced in size in sequential steps using oversampled data from HTS_428 to HTS_8 to statistically optimize the final QSAR model of the mGluR5 experimental HTS data set. On the basis of the ROC curve analysis, HTS_276 descriptors (heavy black line) and HTS_428 descriptors (heavy gray line) displayed the best signal-to-noise profiles.

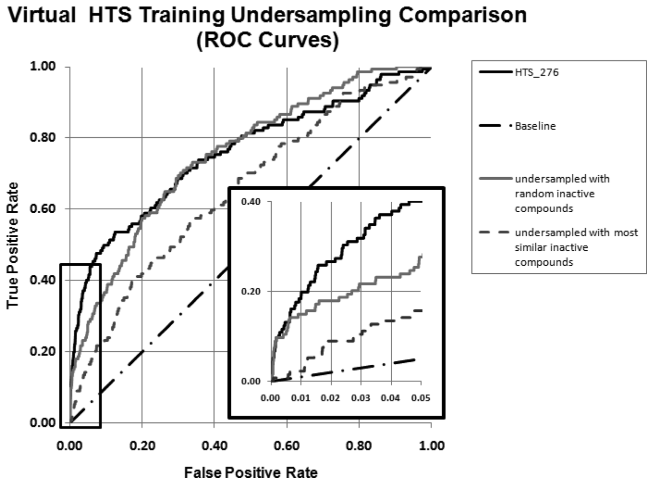


**Figure 3.** Receiver operating characteristic (ROC) curve plots for undersampling methods comparison. ROC curve analysis showing optimized descriptor set HTS_276 based on oversampling (solid black line) compared to undersampling using a random selection of inactive compounds for monitoring and training data sets (solid gray line) as well as a selection of the most similar inactive to active compounds (dashed gray line).

surface autocorrelation) reveals the superior performance of radial distribution functions across the six ANN models tested (Figure 1a). Surface autocorrelation functions were only tested in the first two models (428 and 416 descriptors) because of lower sensitivity scores (Figure 1a). Analysis of input sensitivity by property revealed high sensitivities for $\pi$ atom (0.92−1.38) electronegativity, lone pair (1.42−2.54) electronegativity, and for polarizability (0.70−0.94).

The impact of these descriptors makes intuitive sense as active compounds such as benzoxazepines and benzamides (Figure 4) that are well represented in the training data set contain extended $\pi$ conjugated systems as well as hetero atoms with lone pair electrons. However, we expect overlap in the description of chemical structure by various groups of descriptors. Hence, while the current descriptor set is optimal for the prediction of mGluR5 PAM activity, other suitable combinations of descriptors can yield similarly good results as demonstrated in iterations 1, 2, and 3. Nevertheless, descriptor optimization is important as usage of the maximum number of descriptors or usage of a small set of scalar descriptors will hamper the performance of the QSAR model (Table 2 and Figure 1a).

A recent study demonstrated the necessity for optimizing molecular descriptor types for each individual data set to yield optimal QSAR models (*70*). Other

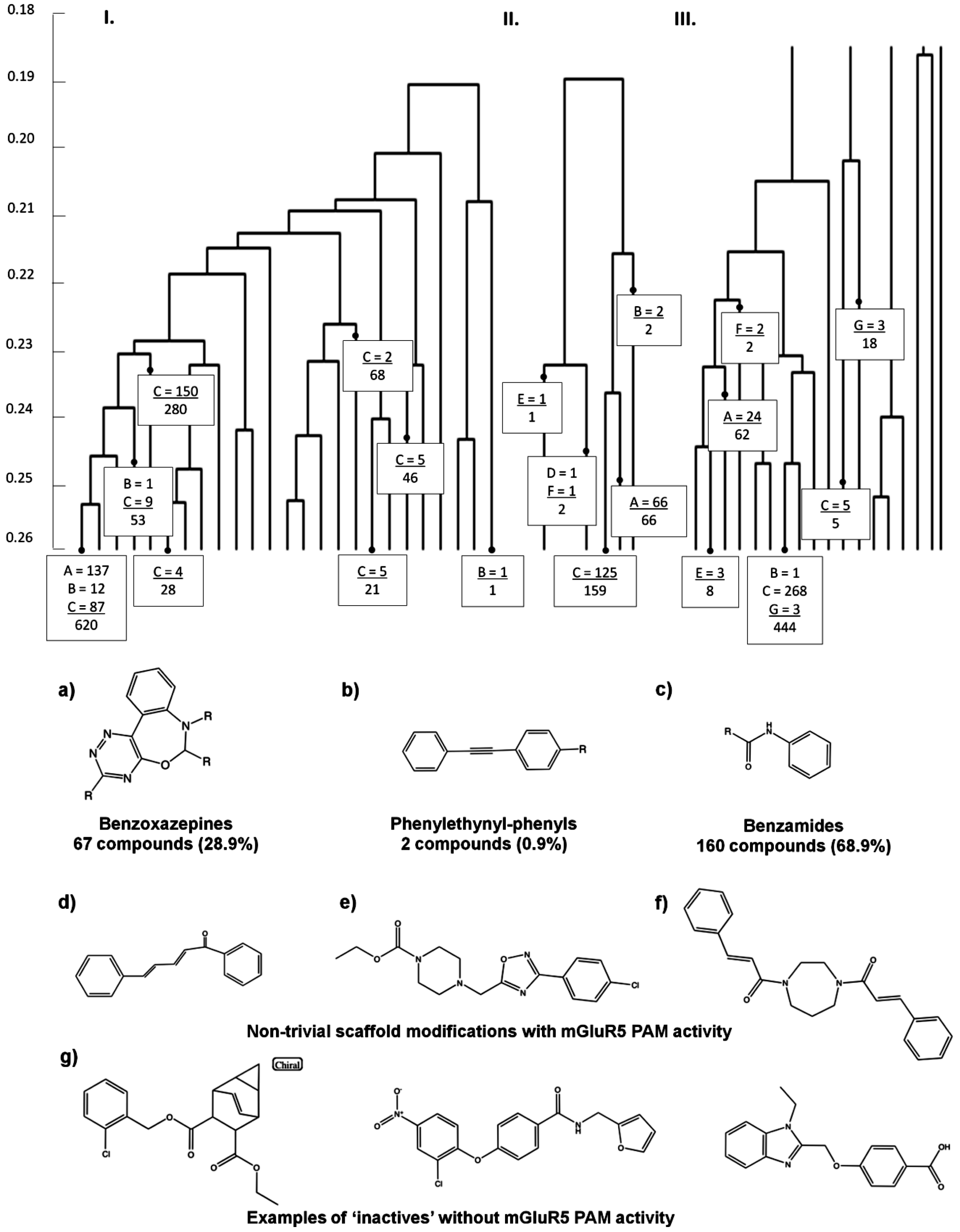


**Figure 4.** Scaffold category analysis. (I) Scaffold composition of 1,382 mGluR5 PAMs from HTS. mGluR5 PAMs were clustered with the Mathematica package using the Tanimoto coefficient of the largest common substructure as distance measure. Three major scaffolds are constituted by 137 benzoxazepines (9.9%, a), 14 phenylethynyls (1.0%, b), and 267 benzamides (19.3%, c). (II) Scaffold composition of active compounds in the postscreen. (III) Scaffold composition of inactive compounds in the postscreen. Compounds d, e, and f are nontrivial mGluR5 PAM scaffold modifications identified by the virtual screen using the ANN QSAR model. Panel (g) highlights representative compounds found inactive in the postscreen.

studies independently reported the radial distribution function as the most robust molecular descriptor category in capturing the structure activity signal from experimental HTS data sets (*55*, *56*).

## Virtual Screening of the ChemBridge Compound Library

The ANN QSAR model was applied in a virtual screen of the ChemBridge database of commercially available compounds. *In silico* screening of the entire library of ~450,000 compounds took approximately one hour on a regular personal computer. A total of 813 compounds with predicted $EC_{50}$ values below 1.0 $\mu$M for mGluR5 PAM activity were selected. An additional 11 compounds were chosen on the basis of visual inspection by an expert medicinal chemist (C.W.L.) from clusters at a lower potency cutoff of 10 $\mu$M for a total of 824 compounds.

The compounds identified in the virtual screen were ordered from the vendor (ChemBridge) and tested at the Vanderbilt HTS facility. In an initial primary screen (see Methods) of the predicted compound collection from our virtual screen, 260 compounds were identified and classified as 210 PAMs, 49 partial agonists, and 1 antagonist. Follow-up CRC assays confirmed 232 compounds with various activities at mGluR5. The compounds were classified as pure potentiators (177) and partial agonists (55). The remaining 27 compounds were either inactive (Figure 4g) (21), fluorescent (2), or showed increased baseline measurements in the fluorescent assays (4). This result reflects an *enrichment* = 232/824 × 144,475/1,356 = 30 relative to the initial experimental HTS hit rate. The experimental enrichment is consistent with the enrichment values predicted from an analysis of an independent data set during the development of the QSAR model (Table 2).

To assess whether the active compounds identified by the present virtual screening approach could have been identified through simpler procedures, a similarity search was performed on the ChemBridge database using MACCS structural keys as molecular fingerprints. Implementation of a Tanimoto coefficient cutoff of 99% for similarity between known actives from the high-throughput screen and compounds with unknown activity yielded a total of 1204 novel hits including 849 benzamides, 91 benzoxazepines, and two phenylethynyl-phenyls. The overlap between this set and the 232 active compounds identified by the ANN approach is 74 compounds (32%). This result demonstrates that our method identified 158 compounds that would have been missed in a naïve similarity search.

## Analysis of the Newly Identified Set of mGluR5 Potentiators

According to MACCS fingerprint-based clustering (*51*, *71*) using a Tanimoto coefficient (*71*) of 0.75 for similarity, of the 232 compounds with confirmed mGluR5 activity identified in our virtual screen of the ChemBridge commercial library 67 compounds (28.9%) were classified as benzoxazepines with pure potentiator activity (Figure 4a); 2 compounds (0.9%) were structurally similar to MPEP (containing a phenylethynyl-phenyl moiety) and displayed partial agonist activity (Figure 4b); 53 compounds (22.8%) were classified as benzamide derivatives with partial agonist activity, and 107 compounds (46.1%) from the same scaffold were classified with pure potentiator activity (Figure 4c); and 3 compounds (1.3%) contained other nontrivial scaffold modifications (Figure 4d−f) with weaker potentiator activity ($EC_{50} \geq 2.5$ $\mu$M). The latter 3 compounds were contained in the 813 compounds predicted at the higher potency (1.0 $\mu$M cutoff).

## Major Scaffolds Are Evenly Distributed Throughout Training, Monitoring, and Independent Data Sets

The library of 1,382 compounds identified as active in the original HTS screen was analyzed using a clustering approach (see Methods). At a cutoff of 25% similarity, 25 different scaffolds were identified (Figure 4).

All large scaffold clusters were equally represented throughout the training, monitoring, and independent data sets. Of the 267 compounds classified as benzamides, 214 compounds (80.1%), 21 compounds (7.9%), and 32 compounds (12.0%) were found in the training, monitoring, and independent data sets, respectively; and 137 compounds were classified as benzoxazepines, with 114 compounds (83.2%) in the training set, 13 compounds (9.5%) monitoring set, and 10 compounds (7.3%) in the independent set. Last, the mGluR5 PAM library contained 14 compounds structurally similar to MPEP (containing a phenylethynyl-phenyl moiety) which were distributed throughout the data sets as follows: 12 compounds (85.7%), 1 compound (7.1%), and 1 compound (7.1%).

## Majority of Hit Compounds Share a Scaffold with Previously Identified Potentiator Compounds

The majority of the compounds recovered contained chemotypes that were the major component of the training data sets (Figure 4a−c). Therefore, our results demonstrate a powerful method for hit explosion, the enumeration of compounds around scaffolds from a HTS experiment. The results build a detailed picture of structure−activity relationships for each of the scaffolds. In the early stages of drug discovery, time can be saved through the acquisition of commercially available compounds to enumerate focused libraries around confirmed HTS hit compounds. The results can help in the planning of synthetic chemistry efforts.

## Benzamides, Benzoxazepines, and MPEP-Like Compounds Are Enriched in the Postscreen

The postscreen library of 824 compounds identified 232 compounds with potentiating activity; compounds were analyzed with a clustering approach and yielded five unique scaffolds at a cutoff of 25% similarity (Figure 4). The majority of benzoxazepine and benzamide derivatives form a single cluster at this cutoff containing 125 and 66 compounds, respectively. The nontrivial scaffold modifications with mGluR5 PAM activity were found in two separate clusters. Each nontrivial scaffold modification was observed once in the postscreen library. One cluster consisted of the only two MPEP-derivatives found in the active compounds. Note that while benzamides, benzoxazepines, and MPEP-like compounds made up only 30% of active compounds in the original HTS experiment, 99% of all active compounds identified in the postscreen belong to one of these three substance classes. We conclude that the machine learning method excelled in recognizing these three scaffolds while other active compounds might have been predicted only at a reduced potency cutoff.

## Inactive Compounds in the Postscreen Library Contain 47% Benzamides

The remainder of the postscreen library was shown to be inactive toward the receptor, and a clustering approach was utilized to identify 18 unique scaffolds at a cutoff of 25% similarity. The major scaffolds seen throughout the training sets (Figure 4a−c) distributed as follows: 24 compounds were identified as benzoxazepines, benzamide derivatives yielded 278 compounds, and 10 compounds structurally similar to MPEP were observed in the compounds confirmed with inactivity toward mGluR5. Derivatives of the nontrivial compounds (Figure 4d−f) were represented among the inactive compounds five times. We conclude that by far not all benzamides, benzoxazepines, and MPEP-like compounds are active PAMs of mGluR5. While the ANN enriches for these scaffolds, it also collects a number of inactive compounds that share this chemotype. In fact, in our original HTS experiment, a total of 42,588 compounds with these scaffolds were found inactive and only 418 were found active, which mirrors our overall rate of active compounds (0.97%). In the postscreen library, we found 229 derivatives of these scaffolds with activity and 312 without. The enrichment of active compounds that share one of these scaffolds is 44 and is therefore somewhat higher than the overall enrichment observed. Note that a naïve similarity search for these scaffolds would have failed to produce these enrichment rates and therefore resulted in a lower rate of active compounds.

## Twenty-eight Percent of the Active Compounds Are Nontrivial Modifications of Original HTS Hits

While 99% of the newly identified mGluR5 PAM compounds have a scaffold that has been previously identified, only 72% of the 232 compounds were trivial derivatives, i.e., have a single functional group added or removed (Figure 4a−c). The remaining 28% had multiple modifications with respect to any of the hit compounds in the original HTS experiment (Figure 4d−f). These compounds would have been difficult to identify with a similarity search as discussed above.

## High Potency Cutoff Introduces Bias to Close Derivatives of Original HTS Hits

As part of our virtual screen, several different potency cutoffs (300nM, 1 $\mu$M, 2 $\mu$M, 5 $\mu$M, and 10 $\mu$M) were employed to identify a compound library size that was tractable for experimental ordering and testing. Selection of a predicted potency cutoff of 1.0 $\mu$M for mGluR5 PAM activity might have biased the majority of the 824 compounds toward molecules with similar chemotypes to the compound classes that represented the majority of the known active compounds included in our training data set (benzoxazepines, phenyl ethynyls, and benzamide-containing scaffolds) (Figure 4a−c). However, with the identification of three nontrivial modifications of known chemotypes having mGluR5 PAM activity ($EC_{50}$ > 2.5 $\mu$M), scaffold hopping appears to be possible using this method (Figure 4d−f). The identification of 158 compounds missed by a naïve similarity search demonstrates the complementary chemical space sampled in a hit explosion setup. For this purpose, more compounds should be selected from a lower potency cutoff (10−30 $\mu$M) combined with filters to remove compounds with chemotypes similar to those in the training data set. We would expect substantially reduced enrichment factors in such a scenario. We included 11 compounds in the 824 compounds ordered from several chemotypes that were identified from a cluster analysis of our mGluR5 virtual screen at a lower potency cutoff (10 $\mu$M). This small subset of compounds was chosen by visual inspection. We did not discover mGluR5 PAM activity in any of these compounds. The compounds were either fluorescent or inactive in our experiments. However, this result is inconclusive because of the very small number of compounds selected according to these criteria.

## Fragment-Independent Numerical Description Deals Efficiently with Multiple Scaffolds

The observation of three nontrivial chemotype modifications underscores the ability of fragment-independent numerical descriptions to map the chemical structure of a diverse compound library into a numerical fingerprint. Different classes of compounds displaying mGluR5 PAM activity (phenylethynyls, benzoxazepines, benzamides, etc.) were used in training the ANN models, and all of those same classes of compounds are recovered in the library of 232 hit compounds. This emphasizes the ability of our machine learning based QSAR model to efficiently deal with biologically complex and little understood phenomena in a black-box-like fashion.

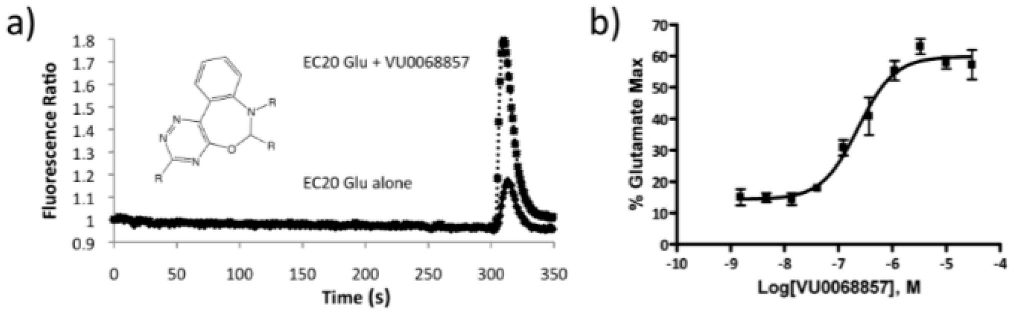


**Figure 5.** FDSS measurement of intracellular $Ca^{2+}$ release in response to mGluR5 activation and potentiation by allosteric modulator compounds. (a) Agonist induced $Ca^{2+}$ transients were quantified on the basis of the fluorescence change observed in cells treated with an $EC_{20}$ concentration of glutamate plus candidate allosteric potentiator compounds (dashed line trace) versus with glutamate alone (solid line trace). (b) Putative primary screen hits showed potentiation of the glutamate response and were confirmed by testing for concentration-dependent activity on mGluR5 over a range of 4 log units with 10 point concentration response curves (30 $\mu$M−1 nM final concentration).

## Conclusions

In conclusion, machine learning methods (ANN) were used to generate QSAR models from an HTS experimental data set in virtual screens of an external commercial compound collection for the purpose of enrichment of our local library for compounds with mGluR5 allosteric activity. A combination of 2D- and 3D-molecular descriptors spanning 35 categories was implemented to encode a broad range of physical and chemical data for each compound. Optimization of the molecular descriptors used to encode chemical structures using oversampled data sets minimized noise by excluding less sensitive descriptors from training inputs to maximize the signal for mGluR5 and proved to be a crucial step for increasing enrichment for active compounds. Oversampling of active compounds was included in data set generation to balance the training of our models, and an independent data set representing a randomly selected 10% of the experimental HTS data was reserved for model cross-validation purposes. Fragment-independent numerical description deals efficiently with multiple scaffolds and (potentially) multiple allosteric sites at the mGluR5 receptor. Model validity was assessed on the basis of multiple measures including *rmsd* between predicted and experimental activity, enrichment of active compounds in a virtually screened compound library, and *auc* value of ROC curves. The enrichment factor of 30 determined from biological testing of 824 compounds prioritized from a library of ~450,000 substances demonstrates the predictive power of the method. This enrichment factor also agrees with the theoretically predicted enrichment of 38. While the majority of hit compounds share a chemical scaffold with the previously identified mGluR5 PAM compounds, a significant fraction of these compounds are nontrivial modifications of hit compounds in the original HTS screen. The high potency cutoff used in the virtual screen might have introduced the bias to close derivatives of hit compounds in the original HTS screen. To attempt identification of novel scaffolds (scaffold hopping), lower potency cutoffs should be combined with filters to remove compounds with chemotypes similar to those in the training data set. We would expect substantially reduced enrichment factors in such a scenario.

## Methods

### Experimental High-Throughput Screen for mGluR5 Potentiators and Hit Validation

In the initial HTS experiment, 144,475 compounds were tested for allosteric potentiation of mGluR5 using full automation in conjunction with the Vanderbilt HTS facility (manuscript in preparation). The Vanderbilt screening library is composed of commercially available compounds selected for maximum structural diversity from ChemBridge and ChemDiv vendors. Receptor-induced intracellular release of calcium in response to agonist treatment was measured in a fluorometric assay by utilizing an imaging-based plate reader (FDSS6000, Hamamatsu, Japan) that makes simultaneous measurements of calcium levels in each well of a 384 plate (Figure 5a). HEK 293A cells stably expressing mGluR5 were plated in black-walled, clear-bottomed, poly-D-lysine coated 384-well plates (BD Biosciences, San Jose, CA) in 20 $\mu$L assay medium (DMEM containing 10% dialyzed FBS, 20 mM HEPES, and 1 mM sodium pyruvate) at a density of 20K cells/well. The cells were grown overnight at 37 °C in the presence of 6% $CO_2$. The next day, the medium was removed and the cells incubated with 20 $\mu$L of 2 $\mu$M Fluo-4, AM (Invitrogen, Carlsbad, CA) prepared as a 2.3 mM stock in DMSO and mixed in a 1:1 ratio with 10% (w/v) pluronic acid F-127 and diluted in assay buffer (Hank's balanced salt solution, 20 mM HEPES, and 2.5 mM Probenecid (Sigma-Aldrich, St. Louis, MO)) for 45 m at 37 °C. Dye was removed, 20 $\mu$L assay buffer was added, and the plate was incubated for 10 m at room temperature. $Ca^{2+}$ flux was measured using the functional drug screening system. As a result, 1,382 compounds were confirmed as potentiators of the mGluR5 glutamate response and used to build QSAR models. Interestingly, several scaffolds with substantial differences in their chemical structures resulted from this experimental screen including benzoxazepine (Figure 5a), phenylethynyl, and benzamide derivatives (Figure 4a−c; manuscript in preparation).

For further analysis, the mGluR5 PAM library of active compounds in the original HTS screen as well as the compounds selected for postscreening were clustered using the Mathematica package (*72*). The Tanimoto coefficient based on the number of atoms in the maximum common substructure served as distance metric:

$$T(\text{molecule}_1, \text{molecule}_2) = \frac{\text{no. of atoms}_{\text{substructure}}}{\text{no. of atoms}_1 + \text{no. of atoms}_2 - \text{no. of atoms}_{\text{substructure}}} \tag{2}$$

In an initial primary screen of ANN selected compounds, single concentrations of compounds (30 $\mu$M final) were transferred to daughter plates using the Echo acoustic plate reformatter (Labcyte, Sunnyvale, CA). Compounds were diluted into assay buffer to a 2× stock using a Thermo Fisher Combi (Thermo Fisher, Waltham, MA), which was applied to cells at $t$ = 3 s. Cells were incubated with test compounds for 140 s, stimulated for 74 s with an $EC_{20}$ concentration of glutamate, and then stimulated for 32 s with an $EC_{80}$ concentration of glutamate. Data were collected at 1 Hz. Agonist induced $Ca^{2+}$ transients were quantified on the basis of the fluorescence change observed in cells treated with an $EC_{20}$ concentration of agonist (glutamate) ± concentrations of candidate allosteric potentiator compounds. Putative hits from the primary screen were confirmed by testing for concentration-dependent activity on mGluR5 over a range of 4 log units (Figure 5b). Compounds were serially diluted 1:3 into 10 point concentration response curves (30 $\mu$M−1 nM final), transferred to daughter plates using the Echo acoustic plate reformatter, and tested as described in the primary screen. Concentration response curves were generated using a four point logistical equation with XLfit curve fitting software for Excel (IDBS, Guildford, UK). Within this software suite, equation number 200 under the category "Dose Response One Site" with the formula $a + b/[1 + (x/c)^d]$ was utilized.

## Generation of Numerical Descriptors for the Training of QSAR Models

For input to machine learning methods, the chemical structure of each molecule needs to be described numerically (see Figure 6a). Initially, 3D models of all 144,475 small molecules are generated using the CORINA software package (*73*). From the 3D structural models, a set of 1,252 numerical descriptors is computed using the ADRIANA software (*57*, *74*). The descriptors can be classified into 35 categories including eight scalar descriptors, eight 2D and eight 3D autocorrelation functions, eight radial distribution functions, and three surface-autocorrelation functions (see Table 1).

## Oversampling Was Used for Balanced Training

As detailed above, 1,382 compounds were confirmed to be active potentiators of the mGluR5 glutamate response (0.94% hit rate). Of these, only 1,356 compounds were used as actives in model generation because of the difficulty in encoding charged molecules with ADRIANA (see Figure 6a). We refer to the active data set as these 1,356 compounds. All other compounds were classified as inactive. In order to maximize the information content of the final prediction method, the data set needs to contain an equal number of active and inactive compounds when training, i.e., its entropy is maximized. Otherwise, a method that would predict all compounds as inactive would be right 99% of the time but completely useless. Balancing was achieved through oversampling (Figure 6b). Active compounds were used in training the ANNs 106 times more frequently to account for their smaller number compared to the inactive set of compounds (0.94% hit rate, see Figure 6b−d).

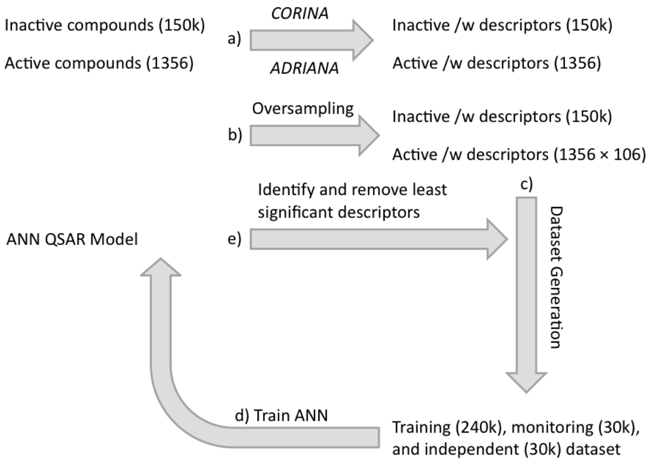

Inactive compounds (150k)
Active compounds (1356)
a)
CORINA
ADRIANA
Inactive /w descriptors (150k)
Active /w descriptors (1356)
Oversampling
b)
Inactive /w descriptors (150k)
Active /w descriptors (1356 × 106)
Identify and remove least significant descriptors
c)
ANN QSAR Model
e)
Dataset Generation
d) Train ANN
Training (240k), monitoring (30k), and independent (30k) dataset

**Figure 6.** Overall model generation workflow: (a) active and inactive molecules were retrieved as MDL SD files from experimental collaborators; 3D structures were generated with CORINA and used as input for the calculation of molecular descriptors using ADRIANA; (b) active molecules were oversampled 106 times to balance data sets; (c) molecules were randomly included in the training data set (80%), monitoring data set (10%), and independent data set (10%); (d) iterative training of ANN models coupled with (e) input sensitivity analysis was used to reduce and optimize the descriptor set until no further improvement in the quality criteria for the independent data set was achieved.

In principle, balancing of the training data can be achieved by two approaches: oversampling of active compounds or undersampling of inactive compounds. Oversampling approaches avoid the removal of part of the inactive compounds, hence utilize all available information for model development, and should therefore yield better results. However, undersampling has the advantage that models can be trained more quickly as only a fraction of the data needs to be fitted. To validate that oversampling gives optimal QSAR models for the present application, two models were developed with different strategies of undersampling inactive compounds and the optimized descriptor set (276 descriptors). The independent data set was kept identical to the oversampling scenario to enable direct comparison. For training and monitoring data sets, (1) a random selection of inactive compounds was selected, and (2) the inactive compounds most similar to active compounds were chosen using MACCS fingerprint keys (*75*) and Tanimoto coefficient as a similarity measure.

## Monitoring Data Set Was Introduced to Terminate ANN Training Early

The natural logarithm of the experimentally determined $EC_{50}$ value of each compound $i$ was used as output for the

ANN models ($exp_i$ = ln $EC_{50,i}$). Compounds classified as inactive were assumed to have an $EC_{50} \geq$ 1mM. The root-mean-square deviation (*rmsd*) between experimental activity $exp_i$ and predicted activity $pred_i$ (see eq 1) is used as the objective function when training the ANN models.

For training ANNs, the data set is split. Of the total experimental data set, 115,581 (80%) data points were used for the ANN training (Figure 6c,d); 14,448 (10%) data points were set aside for monitoring during ANN training and initiating early termination (Figure 6c,d). After each training iteration, the *rmsd* of the monitoring data set was computed. Training was terminated once the *rmsd* value of the monitoring data set was minimized. The final 14,448 data points (10%) were reserved for independent testing of QSAR models (see Table 2). Care was taken to avoid any overlap between training, monitoring, and the independent data set. All results reported were obtained for the independent data set unless noted differently.

## Artificial Neural Network (ANN) Architecture and Training

ANNs are machine learning algorithms that reflect characteristics of biological neural systems in a much simplified fashion. The simplest ANN consists of several layers $j$ = 1,2, ..., $n$ containing $N_j$ neurons each. $N_o$ corresponds to the number of inputs. In a pairwise fashion, neurons in neighboring layers are interlinked by weighted connections $w_{kl}$ (Figure 1b). These connections represent the degrees of freedom of the ANN which are optimized during the training procedure. The input data $x_k$ to every neuron are summed up according to their weights and modified by the activation function $K$:

$$f_l(x_k) = K\left(\sum_l w_{kl}x_k\right) \qquad (3)$$

The output $f_l(x_k)$ then serves as input to the $l$-th neuron of the next layer (Figure 1b).

For the present setup, the input vector $\langle x\rangle$ to the first layer consists of the chemical descriptors introduced above. The single output number of the last layer that contains only one neuron is the experimentally determined biological activity $exp_i$. The present ANNs have up to 1,252 inputs (Figure 1a), 8 hidden neurons (Figure 1b), and 1 output (Figure 1c). The sigmoid function shown in eq 4 is applied as activation function $K$ of the neurons.

$$K(x) = \frac{1}{1+e^{-x}} \qquad (4)$$

The training method used is resilient back-propagation of errors (*76*), a supervised learning approach. The difference between the experimental activity $exp_i$ and predicted activity $pred_i$ determines the change of each weight within the back-propagation of errors. Ultimately, the root-mean-square deviation (*rmsd*, eq 1) between experimental and predicted biological activity is minimized. The ANNs were trained with up to 40,000 iterations of resilient propagation. However, training was terminated early when the monitoring data set achieved its minimum *rmsd*. The training took up to 13 h per network using 8 cores of a core2 quad 2.33 GHz Intel Xeon microprocessor in parallel on the 64-bit version of Red Hat Enterprise Linux 5.2.

## Selection of the Optimal Set of Descriptors of Chemical Structure

Optimization of the descriptor set was achieved by systematic removal of molecular descriptor groups that were the least significant for the prediction of PAM activity (Figure 6e and Figure 1a). The objective of this procedure is to reduce the total number of inputs and therefore the total number of weights of the ANN (Figure 6d,e and Figure 1a). It is advantageous to remove obsolete descriptors in order to minimize the number of degrees of freedom (weights) that need to be determined. In the process, training and prediction of ANNs are accelerated. Furthermore, noise is reduced while the ratio of data points versus degrees of freedom increases.

To determine the significance of each input, the ANN is first trained using the complete set of 1,252 descriptors (Table 1). After the completion of training, the ANN represents a multidimensional function with input values $x_1$,$x_2$,..., $x_{N_0}$ and output $y$:

$$y = f(x_1, x_2, ..., x_{N_0}) = f(\langle x\rangle) \qquad (5)$$

The partial derivative of each input with respect to the output can be determined numerically and is introduced as input sensitivity:

$$\textit{input sensitivity} = \left(\frac{\partial^k y}{\partial x_k}\right)_{x_{l\neq k}} \approx \frac{1}{100}\sum_{i=1}^{100}\frac{\Delta y}{\Delta x_k} \qquad (6)$$

For this purpose, each input value $x_k$ is altered by a small $\Delta x_k$ in an independent experiment, and the change $\Delta y$ is monitored. Following this procedure, the input sensitivity is determined for each input $k$ by randomly selecting 100 compounds from the independent data set. The input $x_k$ is perturbed by a small number $\Delta x_k$ = ± 5% of the input range. The output change $\Delta y$ is recorded(*77*) . The input sensitivity of input $k$ is the average ratio observed (eq 6).

The input sensitivity of each of the 27 nonscalar descriptor categories was determined as norm over the individual input sensitivity values within this category. The descriptor categories were sorted by input sensitivity. All 3D autocorrelation, radial distribution function, and surface autocorrelation descriptors with an input sensitivity above 0.06 were used to train an oversampled model with 428 descriptors, while descriptors with a smaller input sensitivity were removed. Approximately 2/3 (65%) of the total input sensitivity were maintained by implementing approximately 1/3 (34%) of the total number of descriptors. This reduction sped up the training process by a factor of 3. The least significant descriptor category was removed in subsequent iterations of descriptor optimization (see Figure 1a). This procedure was repeated until further removal of descriptors did not result in an increase of prediction accuracy for the independent data set (see Figure 6c−e and Figure 1a).

## Enrichment and Area under the Curve (*auc*) As Quality Measures

As mentioned before, the *rmsd* between predicted and experimental ln $EC_{50}$ was used as the objective function for training the ANNs. $EC_{50}$ values for compounds classified as inactive were assumed to be 1 mM. Analysis of the *rmsd* proved to be a poor indicator for model quality (see Table 2) as the

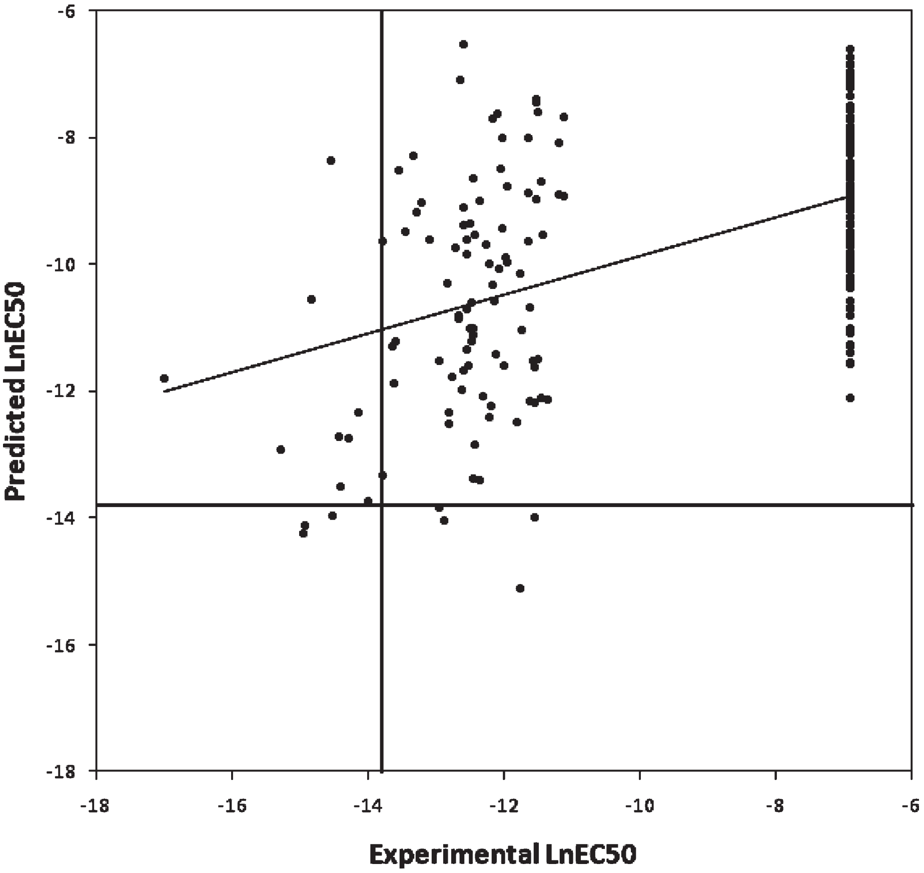


**Figure 7.** Correlation plot between measured and predicted $\ln EC_{50}$ values shows only weak correlation ($R = 0.4805$, $y = 0.305x - 6.8251$). Inactive compounds were set to an $EC_{50}$ of 1 mM ($\ln EC_{50} = -6.9$). The solid lines represent the cutoff used for the acquisition of compounds ($EC_{50}$ of 1 $\mu$M/$\ln EC_{50} = -13.82$).

correlation coefficients between experimental and predicted $\ln EC_{50}$ values are typically smaller than 0.5 (see Figure 7). Note that for the application of these models as tools in virtual screening (read below) binary classification is the important criteria, as in the end, a binary decision for compound acquisition is made. Hence, all models were also assessed in terms of the binary classification power using enrichment and area under the curve (*auc*) quality measures. Receiver operating characteristic (ROC) curves were generated as a measure to evaluate the predictive power of the machine learning approaches. ROC curves plot the rate of true positives $TP$ or $sensitivity = TP/P$ versus the rate of false positives $FP$ or $(1 - specificity) = 1 - TN/N = FP/N$ of a binary classifier. $TP$ represents the number of true positives and $FP$ the number of false positives within this subset. $P$ represents the total number of positives and $N$ the total cases known to be negative. Here, biological activity was used as the binary classifier (Figure 2). The diagonal represents the performance expected from a random predictor. The larger the *auc* of a ROC curve, the larger is the predictive power of the model.

For the prediction of biological activity, often only the very initial part of the ROC curve is of interest. This is the area containing the compounds with the highest predicted biological activity. As after a virtual screen of a compound library, only a small percentage (typically 0.1−1.0%) of compounds predicted to be maximally active will enter biological tests (only this fraction of the ROC curve will be actually used in the virtual screen). The *auc* value is a poor measure of predictive power in this region of the ROC curve as it measures overall performance.

Therefore, often the initial slope of the ROC curve is analyzed using so-called *enrichment* values. Enrichment measures the factor by which active compounds (positives) are increased relative to inactive compounds (negatives) when selecting a subset of data predicted with the highest confidence levels by a model:

$$enrichment = \frac{TP}{TP+FP} \Big/ \frac{P}{P+N} \qquad (7)$$

When computed for the independent data set, the *enrichment* represents the expected factor by which the fraction of active compounds is increased in an *in silico* virtual screen when compared to the chance of finding active compounds in an unbiased data set (here 0.94%). Note that enrichment values are always coupled to a certain cutoff, the fraction of molecules retained after filtering. The enrichments reported in Table 2 were determined for a cutoff of 0.35%. As an example, this would correspond to filtering 1,000 compounds out of a library of about 300,000.

As the models were trained with continuous ln $EC_{50}$ values but largely applied in a binary classification setting, we tested if training of the models as pure binary classifiers offered any advantages. A model was trained where all active compounds were given an activity of 1, and all inactive compounds were set to 0. For the independent data set, an *auc* of 0.744 and an *enrichment* of 26 were calculated. However, this procedure did not yield an improvement over models trained with continuous ln $EC_{50}$ values (see Table 2). This approach was not pursued further.

### Implementation

The ANN algorithm was implemented in the BioChemistryLibrary (BCL). The training method used is resilient propagation, a supervised learning approach (*76*). Further detail is given above. The BCL is an in house developed object-oriented library written in the C++ programming language. It consists currently of approximately 400 classes and 300,000 lines of code. ADRIANA (*57*, *74*) was used for the generation of chemical descriptors. CORINA (*73*) was used for the generation of three-dimensional structures.

## Supporting Information Available

A summary of fit statistics and a concentration response curve for one example of each of the major scaffolds identified. Furthermore, a correlation plot between XlogP and $EC_{50}$ for the independent data set. This material is available free of charge via the Internet at http://pubs.acs.org.

## Author Information

### Corresponding Author

* Vanderbilt University Department of Chemistry, 465 21st Ave. South, BIOSCI/MRBIII, Room 5144B, Nashville, TN 37232-8725. Phone: +1 (615) 936-5662. Fax: +1 (615) 936-2211. E-mail: jens.meiler@vanderbilt.edu. URL: www.meilerlab.org.

### Author Contributions

⊥ Authors contributed equally to this work.

### Funding Sources

This work was supported by National Institutes of Health Grants (R21 MH082254 to J.M., R01 NS031373, R01 MH062646, and R01 MH074953 to P.J.C., R01 DA023947 to C.W.L, T32 GM07628 to T.T.N.).

## Abbreviations

mGluR5, metabotropic glutamate receptor subtype 5; HTS, high-throughput screening; ANN, artificial neural network; QSAR, quantitative structure−activity relationship; CNS, central nervous system; iGluRs, ionotropic glutamate receptors; mGluRs, metabotropic glutamate receptors; G proteins, guanine nucleotide binding proteins; NMDAR, *N*-methyl D-aspartate receptor; MPEP, 2-methyl-6-(phenylethynyl)-pyridine; CPPHA, *N*-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2*H*-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide; CDPPB, 3-cyano-*N*-(1,3-diphenyl-1*H*-pyrazol-5-yl)benzamide; PCP, phencyclidine; DFB, 3,3-difluorobenzaldazine; GPCR, G protein coupled receptor; PAM, positive allosteric modulation/modulator; *c* log *P*, calculated log of *n*-octanol/water partition coefficient; CMR, calculated molecular refractivity; TPSA, topological polar surface area; CoMFA, comparative molecular field analysis; CoMSIA, comparative molecular similarity analysis; FEPOPS, feature point pharmacophores; ROC, receiver operating characteristic; BCL, BioChemistry-Library.

## References

1. Conn, P. J., and Pin, J. P. (1997) Pharmacology and functions of metabotropic glutamate receptors. *Annu. Rev. Pharmacol. Toxicol. 37*, 205–237.

2. Pin, J. P., and Duvoisin, R. (1995) The metabotropic glutamate receptors: structure and functions. *Neuropharmacology 34* (1), 1–26.

3. Palucha, A., and Pilc, A. (2002) On the role of metabotropic glutamate receptors in the mechanisms of action of antidepressants. *Pol. J. Pharmacol. 54* (6), 581–586.

4. Chojnacka-Wojcik, E., Klodzinska, A., and Pilc, A. (2001) Glutamate receptor ligands as anxiolytics. *Curr. Opin. Invest. Drugs 2* (8), 1112–1119.

5. Pilc, A. (2003) LY-354740 (Eli Lilly). *IDrugs 6* (1), 66–71.

6. Marino, M. J., and Conn, P. J. (2002) Direct and indirect modulation of the N-methyl D-aspartate receptor. *Curr. Drug Targets CNS Neurol. Disord. 1* (1), 1–16.

7. Chavez-Noriega, L. E., Schaffhauser, H., and Campbell, U. C. (2002) Metabotropic glutamate receptors: potential drug targets for the treatment of schizophrenia. *Curr. Drug Targets CNS Neurol. Disord. 1* (3), 261–281.

8. Conn, P. J., Lindsley, C. W., and Jones, C. K. (2009) Activation of metabotropic glutamate receptors as a novel approach for the treatment of schizophrenia. *Trends Pharmacol. Sci. 30* (1), 25–31.

9. Ayala, J. E., Chen, Y., Banko, J. L., Sheffler, D. J., Williams, R., Telk, A. N., Watson, N. L., Xiang, Z., Zhang, Y., Jones, P. J., Lindsley, C. W., Olive, M. F., and Conn, P. J. mGluR5 positive allosteric modulators facilitate both hippocampal LTP and LTD and enhance spatial learning. *Neuropsychopharmacology* 2009, *34* (9), 2057–2071.

10. Varney, M. A., and Gereau, R. W. t. (2002) Metabotropic glutamate receptor involvement in models of acute and persistent pain: prospects for the development of novel analgesics. *Curr. Drug Targets CNS Neurol. Disord. 1* (3), 283–296.

11. Doherty, J., and Dingledine, R. (2002) The roles of metabotropic glutamate receptors in seizures and epilepsy. *Curr. Drug Targets CNS Neurol. Disord. 1* (3), 251–260.

12. Wisniewski, K., and Car, H. (2002) (S)-3,5-DHPG: a review. *CNS Drug Rev. 8* (1), 101–116.

13. Marino, M. J., and Conn, J. P. (2002) Modulation of the basal ganglia by metabotropic glutamate receptors: potential for novel therapeutics. *Curr. Drug Targets CNS Neurol. Disord. 1* (3), 239–250.

14. Kinney, G. G., Burno, M., Campbell, U. C., Hernandez, L. M., Rodriguez, D., Bristow, L. J., and Conn, P. J. (2003) Metabotropic glutamate subtype 5 receptors modulate

locomotor activity and sensorimotor gating in rodents. *J. Pharmacol. Exp. Ther. 306* (1), 116–123.

15. Henry, S. A., Lehmann-Masten, V., Gasparini, F., Geyer, M. A., and Markou, A. (2002) The mGluR5 antagonist MPEP, but not the mGluR2/3 agonist LY314582, augments PCP effects on prepulse inhibition and locomotor activity. *Neuropharmacology 43* (8), 1199–1209.

16. Campbell, U. C., Lalwani, K., Hernandez, L., Kinney, G. G., Conn, P. J., and Bristow, L. J. (2004) The mGluR5 antagonist 2-methyl-6-(phenylethynyl)-pyridine (MPEP) potentiates PCP-induced cognitive deficits in rats. *Psychopharmacology* (*Berlin*) *175* (3), 310–318.

17. Brody, S. A., Dulawa, S. C., Conquet, F., and Geyer, M. A. (2004) Assessment of a prepulse inhibition deficit in a mutant mouse lacking mGlu5 receptors. *Mol. Psychiatry 9* (1), 35–41.

18. O'Brien, J. A., Lemaire, W., Chen, T. B., Chang, R. S., Jacobson, M. A., Ha, S. N., Lindsley, C. W., Schaffhauser, H. J., Sur, C., Pettibone, D. J., Conn, P. J., and Williams, D. L., Jr. (2003) A family of highly selective allosteric modulators of the metabotropic glutamate receptor subtype 5. *Mol. Pharmacol. 64* (3), 731–740.

19. O'Brien, J. A., Lemaire, W., Wittmann, M., Jacobson, M. A., Ha, S. N., Wisnoski, D. D., Lindsley, C. W., Schaffhauser, H. J., Rowe, B., Sur, C., Duggan, M. E., Pettibone, D. J., Conn, P. J., and Williams, D. L., Jr. (2004) A novel selective allosteric modulator potentiates the activity of native metabotropic glutamate receptor subtype 5 in rat forebrain. *J. Pharmacol. Exp. Ther. 309* (2), 568–577.

20. Lindsley, C. W., Wisnoski, D. D., Leister, W. H., O'Brien, J, A., Lemaire, W., Williams, D. L., Jr., Burno, M., Sur, C., Kinney, G. G., Pettibone, D. J., Tiller, P. R., Smith, S., Duggan, M. E., Hartman, G. D., Conn, P. J., and Huff, J. R. (2004) Discovery of positive allosteric modulators for the metabotropic glutamate receptor subtype 5 from a series of N-(1,3-diphenyl-1H- pyrazol-5-yl)benzamides that potentiate receptor function in vivo. *J. Med. Chem. 47* (24), 5825–5828.

21. Kinney, G. G., O'Brien, J. A., Lemaire, W., Burno, M., Bickel, D. J., Clements, M. K., Chen, T. B., Wisnoski, D. D., Lindsley, C. W., Tiller, P. R., Smith, S., Jacobson, M. A., Sur, C., Duggan, M. E., Pettibone, D. J., Conn, P. J., and Williams, D. L., Jr. (2005) A novel selective positive allosteric modulator of metabotropic glutamate receptor subtype 5 has in vivo activity and antipsychotic-like effects in rat behavioral models. *J. Pharmacol. Exp. Ther. 313* (1), 199–206.

22. de Paulis, T., Hemstapat, K., Chen, Y., Zhang, Y., Saleh, S., Alagille, D., Baldwin, R. M., Tamagnan, G. D., and Conn, P. J. (2006) Substituent effects of N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamides on positive allosteric modulation of the metabotropic glutamate-5 receptor in rat cortical astrocytes. *J. Med. Chem. 49* (11), 3332–3344.

23. Engers, D. W., Rodriguez, A. L., Williams, R., Hammond, A. S., Venable, D., Oluwatola, O., Sulikowski, G. A., Conn, P. J., and Lindsley, C. W. (2009) Synthesis, SAR and unanticipated pharmacological profiles of analogues of the mGluR5 Ago-potentiator ADX-47273. *ChemMedChem 4* (4), 505–511.

24. Liu, F., Grauer, S., Kelley, C., Navarra, R., Graf, R., Zhang, G., Atkinson, P. J., Popiolek, M., Wantuch, C., Khawaja, X., Smith, D., Olsen, M., Kouranova, E., Lai, M., Pruthi, F., Pulicicchio, C., Day, M., Gilbert, A., Pausch, M. H., Brandon, N. J., Beyer, C. E., Comery, T. A., Logue, S., Rosenzweig-Lipson, S., and Marquis, K. L. (2008) ADX47273 [S-(4-fluoro-phenyl)-{3-[3-(4-fluoro-phenyl)-[1,2,4]-oxadiazol-5-yl]-piper idin-1-yl}-methanone]: a novel metabotropic glutamate receptor 5-selective positive allosteric modulator with preclinical antipsychotic-like and procognitive activities. *J. Pharmacol. Exp. Ther. 327* (3), 827–839.

25. Bessis, A.-S. B. B., Le Poul, E., Rocher, J.-P., and Epping-Jordan, M. (2005) Preparation of piperidine derivatives as modulators of metabotropic glutamate receptors (mGluR5), WO 044797.

26. Bugada, P. G. S., Le Poul, E., Mutel, V., Palombi, G., and Rocher, J.-P. (2006) Novel oxadiazole derivatives and their use as positive allosteric modulators of metabotropic glutamate receptors and their preparation, pharmaceutical compositions and use in the treatment of central and peripheral nervous system disorders, WO 6123249.

27. Chen, Y., Goudet, C., Pin, J. P., and Conn, P. J. (2008) N-{4-Chloro-2-[(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)-methyl]phenyl}-2-hy droxybenzamide (CPPHA) acts through a novel site as a positive allosteric modulator of group 1 metabotropic glutamate receptors. *Mol. Pharmacol. 73* (3), 909–918.

28. Sharma, S., Rodriguez, A. L., Conn, P. J., and Lindsley, C. W. (2008) Synthesis and SAR of a mGluR5 allosteric partial antagonist lead: unexpected modulation of pharmacology with slight structural modifications to a 5-(phenylethynyl)pyrimidine scaffold. *Bioorg. Med. Chem. Lett. 18* (14), 4098–4101.

29. Rodriguez, A. L., Nong, Y., Sekaran, N. K., Alagille, D., Tamagnan, G. D., and Conn, P. J. (2005) A close structural analog of 2-methyl-6-(phenylethynyl)-pyridine acts as a neutral allosteric site ligand on metabotropic glutamate receptor subtype 5 and blocks the effects of multiple allosteric modulators. *Mol. Pharmacol. 68* (6), 1793–1802.

30. Liu, B., Li, S., and Hu, J. (2004) Technological advances in high-throughput screening. *Am. J. Pharmacogenomics 4* (4), 263–276.

31. Carnero, A. (2006) High throughput screening in drug discovery. *Clin. Trans. Oncol. 8* (7), 482–490.

32. Hodder, P., Mull, R., Cassaday, J., Berry, K., and Strulovici, B. (2004) Miniaturization of intracellular calcium functional assays to 1536-well plate format using a fluorometric imaging plate reader. *J. Biomol. Screening 9* (5), 417–426.

33. Gilchrist, M. A.II, Cacace, A., and Harden, D. G. (2008) Characterization of the 5-HT2b receptor in evaluation of aequorin detection of calcium mobilization for miniaturized GPCR high-throughput screening. *J. Biomol Screening 13* (6), 486–493.

34. Posner, B. A. (2005) High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics. *Curr. Opin. Drug Discovery Dev. 8* (4), 487–494.

35. Todeschini, R., and Consonni, V. (2000) *Handbook of Molecular Descriptors*, Vol. 11, Wiley-VCH, Weinheim, Germany.

36. Hansch, C., Hoekman, D., Leo, A., Weininger, D., and Selassie, C. D. (2002) Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem. Rev. 102* (3), 783–812.

37. Hansch, C. M., Peyton, P., Fujita, Toshio, and Muir, Robert M. (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature 194*, 178–180.

38. Bleckmann, A., and Meiler, J. (2003) Epothilones: quantitative structure activity relations studied by support vector machines and artificial neural networks. *QSAR Comb. Sci. 22* (7), 719–721.

39. Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., and Varnek, A. (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model. 48* (9), 1733–1746.

40. Winkler, D. (2004) Neural networks as robust tools in drug lead discovery and development. *Mol. Biotechnol. 27* (2), 139–167.

41. Cramer, R. D.III, Patterson, D. E., and Bunce, J. D. (1989) Recent advances in comparative molecular field analysis (CoMFA). *Prog. Clin. Biol. Res. 291*, 161–165.

42. Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc. 110* (18), 5959–5967.

43. Klebe, G., Abraham, U., and Mietzner, T. (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem. 37* (24), 4130–4146.

44. Wild, D. J., and Blankley, C. J. (2000) Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci. 40* (1), 155–162.

45. Krasowski, M. D., Siam, M. G., Iyer, M., and Ekins, S. (2009) Molecular Similarity Methods for Predicting Cross-Reactivity With Therapeutic Drug Monitoring Immunoassays. *Ther. Drug Monit. 31* (3), 337–344.

46. Moda, T. L., Montanari, C. A., and Andricopulo, A. D. (2007) Hologram QSAR model for the prediction of human oral bioavailability. *Bioorg. Med. Chem. 15* (24), 7738–7745.

47. Salum, L. B., and Andricopulo, A. D. (2009) Fragment-based QSAR: perspectives in drug design. *Mol. Divers. 13*, 277–285.

48. Heritage, T. W. and Hurst, T. (1997) HQSAR: a highly predictive QSAR technique based on molecular holograms, *Book of Abstracts*, 214th ACS National Meeting, Las Vegas, NV, Sep 7−11, COMP-080.

49. Waller, C. L. (2004) A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *J. Chem. Inf. Comput. Sci. 44* (2), 758–765.

50. Vogt, I., Ahmed, H. E., Auer, J., and Bajorath, J. (2008) Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. *Mol. Divers. 12* (1), 25–40.

51. Brown, R. D., and Martin, Y. C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci. 36* (3), 572–584.

52. Jenkins, J. L., Glick, M., and Davies, J. W. (2004) A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem. 47* (25), 6144–6159.

53. Nettles, J. H., Jenkins, J. L., Williams, C., Clark, A. M., Bender, A., Deng, Z., Davies, J. W., and Glick, M. (2007) Flexible 3D pharmacophores as descriptors of dynamic biological space. *J. Mol. Graphics Modell. 26* (3), 622–633.

54. Marrero-Ponce, Y. (2004) Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *J. Chem. Inf. Comput. Sci. 44* (6), 2010–2026.

55. Gonzalez, M. P., Puente, M., Fall, Y., and Gomez, G. (2006) In silico studies using Radial Distribution Function approach for predicting affinity of 1 alpha,25-dihydroxyvitamin D(3) analogues for Vitamin D receptor. *Steroids 71* (6), 510–527.

56. Morales, A. H., Cabrera Perez, M. A., and Gonzalez, M. P. (2006) A radial-distribution-function approach for predicting rodent carcinogenicity. *J. Mol. Model. 12* (6), 769–780.

57. Computerchemie, M. N. G., Schwab, C. H., and Gasteiger, J. (2006) *ADRIANA.Code; Algorithms for the Encoding of Molecular Structures*, version 2.0, program description, Molecular Networks GmbH, Erlangen, Germany.

58. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J., and Gasteiger, J. (1996) Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci. 36* (6), 1205–1213.

59. Hristozov, D. P., Oprea, T. I., and Gasteiger, J. (2007) Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput.-Aided Mol. Des. 21* (10−11), 617–640.

60. Hristozov, D., Oprea, T. I., and Gasteiger, J. (2007) Ligand-based virtual screening by novelty detection with self-organizing maps. *J. Chem. Inf. Model. 47* (6), 2044–2062.

61. Anzali, S., Barnickel, G., Krug, M., Sadowski, J., Wagener, M., Gasteiger, J., and Polanski, J. (1996) The comparison of geometric and electronic properties of molecular surfaces by neural networks: application to

the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput.-Aided Mol. Des. 10* (6), 521–534.

62. Holzgrabe, U., Wagener, M., and Gasteiger, J. (1996) Comparison of structurally different allosteric modulators of muscarinic receptors by self-organizing neural networks. *J. Mol. Graphics 14* (4), 185–193.

63. Teckentrup, A., Briem, H., and Gasteiger, J. (2004) Mining high-throughput screening data of combinatorial libraries: development of a filter to distinguish hits from nonhits. *J. Chem. Inf. Comput. Sci. 44* (2), 626–634.

64. Zupan, J., and Gasteiger, J. (1993) *Neural Networks for Chemists*, VCH Verlagsgesellschaft mbH, Weinheim, Germany.

65. Burton, J., Ijjaali, I., Barberan, O., Petitet, F., Vercauteren, D. P., and Michel, A. (2006) Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *J. Med. Chem. 49* (21), 6231–6240.

66. Hecht, D., Cheung, M., and Fogel, G. B. (2008) QSAR using evolved neural networks for the inhibition of mutant PfDHFR by pyrimethamine derivatives. *Biosystems 92* (1), 10–15.

67. Hecht, D., and Fogel, G. (2007) High-throughput ligand screening via preclustering and evolved neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 4* (3), 476–484.

68. Butkiewicz, M., Mueller, R., Selic, D., Dawson, E., and Meiler, J. (2009) Application of Machine Learning Approaches on Quantitative Structure Activity Relationships, in *IEEE CIBCB 2009: 2009 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp 255−262, IEEE Computational Intelligence Society, Nashville, TN.

69. Lipinski, C. A. (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies 1* (4), 337–341.

70. Gedeck, P., Rohde, B., and Bartels, C. (2006) QSAR--how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model. 46* (5), 1924–1936.

71. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today 11* (23−24), 1046–1053.

72. Wolfram Research, I. (2008) *Mathematica*, version 7, Wolfram Research, Inc., Champaigne, IL.

73. Gasteiger, J., Rudolph, C., and Sadowski, J. (1990) Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol. 3* (6c), 537–457.

74. Schwab, C. H. (2007) *ADRIANA.Code, 2.0*, Molecular Networks GmbH, Erlangen, Germany.

75. Schoelkopf, B., and Smola, A. J. (2002) *Learning with Kernels*, The MIT Press, Cambridge, MA.

76. Riedmiller, M., and Braun, H. (1992) *Rprop - A Fast Adaptive Learning Algorithm*, Proceedings of the International Symposium on Computer and Information Science VII, pp. 279−286, IEEE, Antalya, Turkey.

77. Meiler, J., Müller, M., Zeidler, A., and Schmäschke, F. (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J. Mol. Model. 7* (9), 360–369.