MACHINE LEARNING ALGORITHMS FOR PREDICTION OF BIOLOGICAL ACTIVITY AND CHEMICAL PROPERTIES

By

Ralf Mueller

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

In

Chemistry

August, 2011

Nashville, Tennessee

Approved:

Professor Jens Meiler

Professor Brian Bachmann

Professor Prasad Polavarapu

Professor Dave Weaver

DEDICATION

To my wife, my parents, and my grandfather.

ACKNOWLEDGMENTS

First I would like to acknowledge my advisor Jens Meiler, Ph.D. for giving me the opportunity to work on these projects, valuable input, and great discussions. Furthermore I would like to acknowledge my committee for keeping me on track and my co-workers Kristian Kaufmann, Eric Dawson, and Nils Woetzel for fruitful discussions.

I thank the Conn lab in collaboration with Craig Lindsley and Dave Weaver for providing me the experimental data for the virtual high-throughput screening experiments.

Financial support was granted by the National Institute of Health and the Vanderbilt Institute for Chemical Biology.

Lastly, I would like to thank my wife and my family for their unwavering support over four continents.

TABLE OF CONTENTS

Page

DEDICATION ii

ACKNOWLEDGEMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvi

SUMMARY xviii

Chapter

I. INTRODUCTION 1

Quantitative Structure Activity Relationships 1

Quantitative Structure Property Relationships 1

QSAR/QSPR through machine learning 1

QSAR/QSPR general workflow 2

Identifying publicly and commercially available carbon chemical shift data bases 2

Local high-throughput screens provided small molecule activity data towards mGlu4/5 2

Commercial ADRIANA descriptors were developed into atom-based descriptors 3

Small molecule conformations were computed with CORINA 3

Constitutional chemical shift descriptors based on atom/bond types and ring closure 3

Strong synergy between QSAR and QSPR projects 4

Neural Networks as robust tools for QSAR/QSPR 4

Allosteric modulation of metabotropic glutamate receptors 5

High-throughput screens in drug discovery 5

Virtual high-throughput screens 6

Positive allosteric modulation of metabotropic glutamate receptor subtype 5 7

Positive allosteric modulation of metabotropic glutamate receptor subtype 4 8

Negative allosteric modulation of metabotropic glutamate receptor subtype 5 9

Prediction of carbon chemical shift 10

II. IDENTIFICATION OF METABOTROPIC GLUTAMATE RECEPTOR SUBTYPE 5 POTENTIATORS USING VIRTUAL HIGH-THROUGHPUT SCREENING 13

Introduction 13

Activators of mGlu5 may provide a novel approach to treatment of schizophrenia 14

High-throughput screening in drug discovery 15

Quantitative structure activity relations in drug discovery 15

Numerical descriptors of chemical structure for QSARs 16

Fragment-independent transformation-invariant descriptor schemes 17

Application of machine learning algorithms to establish QSARs 18

Quantitative structure activity relation models for mGlu5 positive allosteric modulation 18

Results and Discussion 19

Discussion of concentration response curves in the experimental high-throughput screen 19

Input sensitivity is a reliable measure to prioritize descriptors 20

Optimization of molecular descriptor set improves prediction accuracy of the ANN model 21

Balancing the datasets through oversampling yields better results than two undersampling strategies 26

Radial distribution functions and electronegativity contribute most to an accurate prediction 28

Virtual screening of ChemBridge compound library 30

Analysis of the newly identified set of mGlu5 potentiators 31

Major scaffolds are evenly distributed throughout training, monitoring, and independent datasets 31

The majority of hit compounds share a scaffold with previously identified potentiator compounds 32

Benzamides, benzoxazepines, and MPEP-like compounds are enriched among active compounds in the post-screen 32

Inactive compounds in the post-screen library contain 53 % benzamides, benzoxazepines, and MPEP-like compounds 33

Significant numbers of hit compounds are non-trivial modifications of original HTS screen hits 34

High potency cutoff may introduce bias to close derivatives of original HTS screen hit compounds 34

Fragment-independent numerical description deals efficiently with multiple scaffolds 35

Conclusions 35

Methods 36

Experimental high-throughput screen for mGlu5 potentiators and hit validation 36

Generation of numerical descriptors for training of QSAR models 39

Oversampling was used for balanced training 40

A monitoring dataset was introduced to early terminate ANN training 41

Artificial neural network (ANN) architecture and training 42

Selection of the optimal set of descriptors of chemical structure 43

Enrichment and area under the curve ($\boldsymbol{auc}$) as quality measures 45

Implementation 47

III. VIRTUAL HIGH-THROUGHPUT SCREENING AS A ROBUST TOOL TO IDENTIFY METABOTROPIC GLUTAMATE RECEPTOR SUBTYPE 4 POTENTIATORS 48

Introduction 48

Results and Discussion 50

Descriptor categories were selected according to input sensitivity 50

Optimization of molecular descriptor set improves prediction results 51

Jury model combines favorable features of all previous models 53

Radial Distribution Functions (RDFs) carry most of the input sensitivity 55

Virtual screening of ChemBridge compound library 56

Hits and misses of virtual screening overlap with known actives 57

Second round of predictions focused on scaffold-hopping 59

Non-trivial modifications of known actives constitute 59% of newly identified PAMs 60

Conclusions 62

Methods 62

Balancing the data by oversampling 62

Translating molecular structures into input numerical descriptors 63

From biological data to output 64

Monitoring data set determines progress and termination of training 64

A three-layered ANN was trained implementing Resilient Propagation 64

Jury system combines output of three best networks 65

Selection of the optimal set of descriptors of chemical structure 65

Enrichment and area under the curve complement rmsd as quality measures 66

Implementation 68

IV. IDENTIFICATION OF METABOTROPIC GLUTAMATE RECEPTOR SUBTYPE 5 NEGATIVE ALLOSTERIC MODULATORS USING VIRTUAL HIGH-THROUGHPUT SCREENING 69

Introduction 69

Biology and pharmacology of metabotropic glutamate receptors 69

Negative allosteric modulation of mGlu5 could allow treatment of fragile X syndrome ... 70
Negative allosteric modulators of metabotropic glutamate receptor subtype 5 ... 70
High-Throughput Screening in Drug Discovery ... 71
Virtual High-Throughput Screening ... 72
Describing Chemical Structure for QSAR ... 72
Establishing QSAR in Drug Discovery ... 73
Machine learning and QSAR ... 73
Results and Discussion ... 73
Optimization of molecular descriptor set improves prediction results ... 73
Enrichment is the critical quality measure for virtual screening ... 74
Predictions focused on scaffold-hopping ... 77
Results of virtual screening of ChemDiv compound library ... 77
Virtual High-Throughput Screening identified new scaffold of mGlu5 NAMs ... 77
Conclusions ... 78
Methods ... 78
Translating molecular structures into input numerical descriptors ... 78
Balancing the data by oversampling ... 79
From biological data to output ... 79
Monitoring data set determines progress and termination of training ... 79
A three-layered ANN was trained implementing Resilient Propagation ... 80
Selection of the optimal set of descriptors of chemical structure ... 80

Enrichment and area under the curve as binary quality measures 81

Implementation 82

V. PREDICTING CARBON CHEMICAL SHIFTS EMPLOYING MACHINE-LEARNING METHODS 83

Introduction 83

Ab initio methods are computationally expensive and lack accuracy 83

Database methods rely on large sets of stored spectra. 84

Incremental methods are fast but lack accuracy 85

Artificial neural networks combine speed and accuracy 86

Atom environment code by sphere 86

Iterative Partial equalization of orbital electronegativity determines σ-charges.. 87

Hueckel molecular orbital method determines π-charges of conjugated systems 88

Results and Discussion 88

A combination of resilient and simple back-propagation is used to achieve optimal training results. 88

Chemical shifts with a deviation above 17ppm are excluded from final training. 89

The trained ANN predicts ^{13}C chemical shifts with a mean average error of 2.95ppm. 89

The ANN performs well for predicting ^{13}C NMR spectra of organic compounds. 90

Predicting natural products with a mae of 3.29ppm. 92

Descriptors were sorted by input sensitivity. 95

RDF SigChg, PiChg, TotChg are the best conformational descriptors. 96

A combination of resilient and simple back-propagation is used to achieve optimal training results. 96

The trained ANN predicts ^{13}C chemical shifts with a mean average error of 2.84ppm. 96

Including radial distribution function descriptors allows differentiation between configurations. 97

Growing ^{13}C databases will improve prediction accuracy. 99

Conclusions 99

Methods 100

Programming and data processing. 100

Artificial neural network architecture. 100

Training was performed with simple and resilient back-propagation of errors. 101

Close to 185,000 ^{13}C chemical shifts were available for training. 102

Atoms are sorted in 35 distinct groups. 102

Special descriptors are introduced for π-conjugated systems 103

The numerical description of the atom environment considers five spheres defined by molecule constitution. 104

Frequencies of atom and bond types in each sphere are input for the ANN. 105

Constitutional and conformational descriptors utilize up to eight chemical properties. 106

Descriptors are sorted by input sensitivity to improve prediction accuracy. 107

A single ANN is trained to predict the chemical shift of all ^{13}C atoms. 107

Training was accelerated using recently optimized training algorithm. 108

VI. DISCUSSION 109

Conclusions and future directions 109

Quantitative Structure Activity Relations 109

Metabotropic glutamate receptor 5 PAMs 112

Metabotropic glutamate receptor 4 PAMs 113

Metabotropic glutamate receptor 5 NAMs 114

Carbon chemical shift prediction 114

LIST OF TABLES

Table 1: Summary of fit statistics 20

Table 2: Summary of 1,252 molecular descriptors 23

Table 3: The ***rmsd***, ***auc***, and ***enrichment*** values for all mGlu5 PAMs QSAR models 24

Table 4: Summary of 1,252 molecular descriptors 51

Table 5: The ***rmsd***, ***auc***, and ***enrichment*** values for all round 1 mGlu4 QSAR models 53

Table 6: The ***rmsd***, ***auc***, and ***enrichment*** values for all round 2 mGlu4 QSAR models 59

Table 7: Summary of 1,252 molecular descriptors 75

Table 8: The ***rmsd***, ***auc***, and ***enrichment*** values for all mGlu5 NAMs QSAR models 76

Table 9: The *mae* and *rmsd* of the shift prediction by the different carbon atom types 90

Table 10: The *rmsd* between reported and predicted ^{13}C chemical shifts for some natural products. 92

Table 11: Models with three different descriptor sets were trained 95

Table 12: Atom types: 103

Table 13: 2D/3D auto-correlation and radial distribution functions 106

LIST OF FIGURES

Figure 1: Concentration response curves for (A) phenylethynyl-phenyl, (B) benzamide, and (C) benzoxazepine PAMs.....20

Figure 2: Schematic view of an ANN.....22

Figure 3: Receiver Operating Characteristic (ROC) curve plot:.....25

Figure 4: Receiver Operating Characteristic (ROC) curve plots for undersampling methods comparison:.....27

Figure 5: Scaffold category analysis:.....29

Figure 6: FDSS measurement of intracellular Ca2+ release in response to mGlu5 activation and potentiation by allosteric modulator compounds:.....37

Figure 7: Overall model generation workflow.....40

Figure 8: Correlation plot between measured and predicted $LnEC_{50}$ values. 45

Figure 9: Receiver Operating Characteristic (ROC) curve plot for classical, all, and jury approach (round 1) 52

Figure 10: Schematic view of the jury system.....54

Figure 11: Receiver Operating Characteristic (ROC) curve plot for 415, 578, 741, and jury approach (round 1).....55

Figure 12: Schematic view of an ANN.....56

Figure 13: Scaffold category analysis (round 1).....58

Figure 14: Receiver Operating Characteristic (ROC) curve plot for 415, 578, 741 descriptors, and jury approach (round 2) 59

Figure 15: Scaffold category analysis (round 2).....61

Figure 16: Overall model generation workflow.....63

Figure 17: Receiver Operating Characteristic (ROC) curve plot for 416, 555, 683, 763, 972, and all descriptors.....76

Figure 18: Two novel mGlu5 NAMs sharing a previously unknown scaffold. 78

Figure 19: The training of the ANN with 317 descriptors 89

Figure 20: Highest chemical shift prediction deviation per molecule 91

Figure 21: Average chemical shift prediction deviation per molecule 91

Figure 22: The set of 12 natural products. 94

Figure 23: The training of the ANN with 701 descriptors 97

Figure 24: The difference between the σ-charge radial distribution functions for carbon 1 98

Figure 25: The difference between the σ-charge radial distribution functions for carbon 4 98

Figure 26: Comparison between experimental (bold) and predicted (italic) ^{13}C chemical shifts for two isomers. 99

Figure 27: Principal scheme for the spherical code (5 spheres) and the artificial neural network. 101

Figure 28: An example for the sphere code for *n=3* spheres 105

Figure 29: Influence of the stereochemical descriptors on the overall quality of the trained ANNs 124

LIST OF ABBREVIATIONS

ANN – artificial neural network

BCL – BioChemistryLibrary

CATS-2D – Chemically Advanced Template Search-2D

CDPPB – 3-cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamide

cLogP – calculated log of n-octanol/water partition coefficient

CMR – calculated molecular refractivity

CNS – central nervous system

CoMFA – Comparative Molecular Field Analysis;

CoMSIA – Comparative Molecular Similarity Analysis

CPPHA – N-{4-chloro-2-[(1,3-dioxo-1,3-dihydro-2H-isoindol-2-yl)methyl]phenyl}2-hydroxybenzamide

CSD – Cambridge Structural Database

DFB – 3,3-difluorobenzaldazine

EC_{50} – half maximal effective concentration

FEPOPS – feature point pharmacophores

GPCR – G protein coupled receptor

G proteins – guanine nucleotide binding proteins

HOSE – hierarchically ordered spherical description of environment

HTS – high-throughput screening

IC_{50} – half maximal inhibitory concentration

iGlus – ionotropic glutamate receptors

logP - log of n-octanol/water partition coefficient

mGlus – metabotropic glutamate receptors

mGlu4 – metabotropic glutamate receptor subtype 4

mGlu5 – metabotropic glutamate receptor subtype 5

MPEP – 2-methyl-6-(phenylethynyl)-pyridine

NMDAR – N-methyl D-aspartate receptor

NMRshiftDB – publicly available web data base of chemical shifts

PAM – positive allosteric modulation/modulator

PCP – phencyclidine

QSAR – quantitative structure activity relationship

QSPR – quantitative structure property relationship

ROC – receiver operating characteristic

SD file/SDF – structure data file

TPSA – topological polar surface area

SUMMARY

The focus of this work was to establish quantitative structure activity (QSAR, potency of allosteric modulators) and property (QSPR, carbon chemical shifts) relations for molecules with known structure and activity/property by means of machine learning. These trained machine learning models were then employed to predict biological activity and carbon chemical shifts of molecules with known structure but unknown biological activity or carbon chemical shifts. All described algorithms were implemented in the BioChemistry library (BCL). The BCL is an in-house object-oriented library providing functionality to manipulate small molecules and proteins.

Chapter I gives an introduction to the field of QSAR/QSPR with specific consideration of machine learning and descriptor development. In chapter II, the application of these principles to determine positive allosteric modulators (PAMs) of metabotropic glutamate receptor subtype 5 (mGlu5) from a database of commercially available compounds is described. Chapter III details the expansion of the methods established in chapter II to promote scaffold-hopping in the search for metabotropic glutamate receptor subtype 4 (mGlu4) PAMs. The refinement of the described methods led to the discovery of two compounds representing a new scaffold of negative allosteric modulators (NAMs) of mGlu5. These results are reported in chapter IV. The prediction of carbon chemical shifts by means of a QSPR is reported in chapter V. Chapter VI concludes the main part of this document with a discussion of the results presented. The appendix details protocols, file structures on the accompanying DVD, and applications employed in establishing the QSARs/QSPRs.

Chapter II is based on 'Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening' which was published 2010 in the American Chemical Society journal 'Chemical Neuroscience'[1]. The content of chapter III will be expanded with biological data for the newly identified PAMs for mGlu4 and submitted to the same journal for

publication. Chapter IV will be incorporated into a paper characterizing the two mGlu5 NAMs representing a previously unknown scaffold. Chapter V details research for predicting carbon chemical shifts. The first part was submitted as a manuscript to Journal of Chemical Information and Modeling. Based on the reviews, it was expanded to include stereochemical descriptors. It has not yet been resubmitted.

CHAPTER I

INTRODUCTION

Quantitative Structure Activity Relationships

Structure activity relationships are built on the paradigm that structurally similar compounds have similar biological activity. Quantitative structure activity relationships (QSARs) connect activity to structure by fitting a mathematical function through molecules with known structure and activity: $f(structure) = activity$. Once this relationship is established, it can be employed to predict the activity of molecules with known structure but unknown activity.

Quantitative Structure Property Relationships

Quantitative structure property relationships (QSPRs) expand the idea of QSAR to predict any specific chemical or biological property of a molecule. In this work, the term QSAR is always connected to the prediction of biological activity (half maximal effective/inhibitor concentration EC_{50}/IC_{50}) of molecules towards allosteric modulation of metabotropic glutamate receptor subtype 4 (mGlu4, chapter III) and 5 (mGlu5, chapter II and IV). Alternatively, QSPR describes the attempts to predict chemical shifts for carbon atoms of small molecules (<50 heavy atoms) based on the structure of the molecule around them.

QSAR/QSPR through machine learning

Linear regression analysis is commonly used to establish QSARs/QSPRs. However, this assumes that the activity can be described as a linear combination of all the structural descriptors which is not necessarily the case. Machine learning tools like artificial neural networks (ANNs) and support vector machines can alleviate this problem, since they employ non-linear functions to relate input (structure) and output (activity/property).

QSAR/QSPR general workflow

Establishing all four QSARs/QSPRs described here includes the same four steps: i) a database connecting molecules with known structure and activity/property needs to be obtained, ii) the structure of the molecules is described in a way that is independent from spatial orientation to avoid lengthy superimposition of the molecules, iii) training of models (ANNs) through supervised back-propagation of errors, iv) predicting properties of molecules without prior knowledge of said property. In the case of the allosteric modulators, molecules from commercially available databases (ChemBridge/ChemDiv) which were predicted to have favorable activity were ordered and tested; for the carbon chemical shift prediction, small molecules from the literature were tested towards correct description of their chemical shifts.

Identifying publicly and commercially available carbon chemical shift data bases

NMRshiftDB[2] is a publicly available database that allows scientists to upload NMR spectra with assigned shifts. It contains approximately 290,000 unique carbon chemical shifts in 27,800 spectra (July 2010). A subset of these molecules comprising 178,000 unique carbon chemical shifts in 16,500 spectra is provided with three-dimensional coordinates for the atoms. The SpecInfo database (2002) contains approximately 1.5 million unique carbon chemical shifts in 105,000 spectra. Another publicly available chemical shift database is the Spectral Database for Organic Compounds (SDBS, approximately 130,000 carbon chemical shifts). Examples for commercial shift databases include BIORAD KnowItAll based on the CSEARCH database (~4,000,000 carbon chemical shifts), MODGRAPH NMRPredict (~3,500,000 ^{13}C chemical shifts), and ACD/CNMR (~2,160,000 ^{13}C chemical shifts).

Local high-throughput screens provided small molecule activity data towards mGlu4/5

The biological activity data employed in chapters II to IV was collected from two high-throughput screens performed at the Vanderbilt high-throughput screening center. Both screens

were based on triple-add fluorescence calcium flux assays. The results of the mGlu5 screen were published by Rodriguez et al in October 2010[3]. The Vanderbilt high-throughput screen results for mGlu4 based on work by Niswender et al have not been published yet; however, several papers based on this screen came out in the last few years[4].

Commercial ADRIANA descriptors were developed into atom-based descriptors

ADRIANA[5] is commercially available software that generates descriptors for small molecules. These fall into four main categories (eight scalar descriptors, eight 2D/3D-auto correlations each, eight radial distribution functions, and three surface auto-correlations). However, in the commercial ADRIANA software these descriptors are summed up over all pairs of atoms in the molecule. To generalize these descriptors for single atoms they were implemented based on the references provided in the ADRIANA manual in the local BioChemistryLibrary (BCL) software package. A detailed description of this implementation is given in chapter V.

Small molecule conformations were computed with CORINA

Some of the ADRIANA descriptors (3D and surface auto-correlations, radial distribution functions) needed low-energy conformations to be computed correctly. These conformations were generated with the help of the CORINA[6] software. Since these conformations only needed to be determined once, the CORINA software was not re-implemented.

Constitutional chemical shift descriptors based on atom/bond types and ring closure

To describe the constitution surrounding an atom of interest in the carbon chemical shift prediction, the code in chapter V is based on atoms sharing a certain bond distance from the atom of interest (atom spheres). All atoms with the same atom type (based on element type, hybridization, charge) were summed up for each sphere. Furthermore, the code contains

information about the bonds that connect any given atom to its previous sphere and the number and type of rings closed in each sphere.

Strong synergy between QSAR and QSPR projects

The combination of two seemingly different projects like carbon chemical shift prediction and determination of biological activity still allows for a strong synergy between both. The number of data points is comparable (50,000 – 250,000 for most of the models), the number of inputs is similar (300 – 2,000), and the descriptors could be re-implemented to accommodate both whole molecules and single atoms. Both data sets were stored in SD files which could be processed with a SDF reading function specifically implemented in the BCL.

Neural Networks as robust tools for QSAR/QSPR

For a review of the role of neural networks in drug discovery see Winkler et al[7]. It not only discusses the way QSARs are established and how neural networks can be employed in this process but also provides a general introduction to neural networks and their training.

Another review from Walters and Murcko[8] concentrates on the automated prediction of 'drug-likeness', i. e. to select compounds which can be developed into actual drugs from the vast combinatorial space of all small molecules.

Tetko et al[9] combine the advantages of comparative molecular field analysis (CoMFA) with neural networks. CoMFA field variables that describe the 3D structure of the small molecule are clustered by a self-organizing map to determine the most relevant descriptors. A neural network is trained with back-propagation of errors to predict the activity of each small molecule based on the most relevant field variables.

The QSAR/QSPR models reported in this work are neural networks employing one hidden layer with a variable number of inputs and hidden neurons trained with simple back-propagation of

error, or resilient propagation[10]. However, the input to the QSAR models (virtual HTS, chapters II to IV) consists only of molecular descriptors, while the QSPR model (carbon chemical shift prediction, chapter V) only employs atom descriptors derived from the molecular environment of each carbon atom of interest.

Allosteric modulation of metabotropic glutamate receptors

Metabotropic gluatamate receptors (mGlus) are G-protein coupled receptors divided into eight subtypes in three classes[11]. They are connected to a number of disorders of the central nervous system like schizophrenia[12], Parkinson's disease[13], and fragile X syndrome[14]. Selectively targeting specific mGlu subtypes would allow treatment of these diseases without interfering with other glutamate receptors, which is believed to reduce adverse effects. Pin et al[15] gave an overview of the overall structure of mGlus, pharmacophore models, allosteric binding sites, and available selective ligands for each group of mGlus in 1999. Since then, other modulators for the different subgroups have been discovered. Gasparini[12b] et al specifically describe allosteric modulators of group I (mGlu1 and mGlu5) mGlus. For a recent overview on allosteric modulation of G-protein coupled receptors and its therapeutic potential, see Conn et al[16].

High-throughput screens in drug discovery

The virtual high-throughput screens described in chapters II and IV are based on the high-throughput screening data published by Rodriguez et al[3] in 2010. A local collection of 160,000 small molecules was screened in a single-concentration triple-add calcium flux assay for modulators of mGlu5. Potentiators (2,403) and antagonists (624) were tested in a concentration response curve experiment with ten concentrations ranging from 0.1nM to 0.1mM. Sixty percent (1,387) of the potentiators and 55% (345) of the antagonists were confirmed. Curves were fitted to the concentration responses to determine EC_{50} and IC_{50} values.

Virtual high-throughput screens

QSAR/QSPR models can be employed for virtual high-throughput screens, i.e., the search for (virtual) molecules with predicted properties in a given acceptance range. These molecules can be ordered or synthesized and tested for the desired property.

In 2001 Harper et al[17] reported the prediction of biological activities based on binary kernel discrimination. Two datasets (1,650 monoamine oxidase inhibitors and 101,437 compounds from an enzyme essay) were described by binary descriptors based on atom pairs and topological torsions (APTT). Kernel discrimination was compared to merged similarity search and neural networks. While it outperformed merged similarity search, neural networks proved a better method in some of the experiments, especially since learning rate, momentum, and hidden neurons were not optimized leaving room for improvement.

Jorissen et al[18] employed support vector machines to virtually screen for 50 inhibitors each of different target like cyclooxygenase-2, cyclin-dependent kinase 2, etc. The inactive molecules were drawn from the National Cancer Institute diversity set of chemical compounds. The chemical descriptors were computed with the DRAGON program based on CORINA conformations. Enrichment factors were calculated in the same fashion, as reported in this work. While the method outperforms binary kernel discrimination, the data sets are rather small and known active compounds were retrieved from a set of unrelated inactive compounds. It is not immediately clear if these results will hold in an experiment where external compounds are ordered and tested as in the virtual high-throughput experiments described here.

In 2007 Noeske et al[19] described a similarity search based on CATS-2D descriptors for finding novel allosteric mGlu1 antagonists. After discovering a highly active antagonist based on a coumarine scaffold and developing it into a small molecule library, this approach was further developed in the 2009 paper[20] where self-organizing maps were trained on the CATS-2D

description of pharmacologically active molecules in the COBRA collection. A set of 357 known allosteric antagonists of mGlu1 was projected onto the trained map to identify neurons connected to allosteric mGlu1 antagonists. The actual vHTS consisted in mapping the Asinex Gold Collection 2003 (194,563 compounds) onto the trained map and identifying 28 screening candidates from the 60 top-ranked compounds. One compound showed activity below 1µM, five between 1 and 15µM in an rmGlu1 assay.

In this work, vHTS for mGlu5 PAMs, NAMs, and mGlu4 PAMs are reported. The trained QSAR models were employed to screen commercially available databases (ChemBridge and ChemDiv).

Positive allosteric modulation of metabotropic glutamate receptor subtype 5

For a review of the role of positive allosteric modulation of glutamate receptor subtype 5 as a treatment strategy for schizophrenia see Conn et al[12a]. This explains the great interest in identifying novel positive allosteric modulators of mGlu5 during the last years. Several publications present novel PAMs and their function in cooperation with the receptor.

CPPHA was one of the earlier potent and selective PAMs for mGlu5 reported by O'Brien et al[21]. It had no agonist activity on its own and acted as a PAM in human and rat mGlu5 assays. Being active in both assays strengthens the claim that it works directly on mGlu5. It was an extension of earlier work[22] from the same group describing a whole family of benzaldazines functioning as allosteric modulators of mGlu5 including the negative allosteric modulator 3,3′-dimethoxybenzaldazine with an IC_{50} of 3µM.

Kinney et al[23] described CDPPB as a brain penetrating drug that can reverse amphetamine-induced locomotor activity and deficits in prepulse inhibition in rats. These animal models can be employed to assess antipsychotic drug treatment. It supports the hypothesis that allosteric modulation of mGlu5 can assist in developing antipsychotic drugs.

Lindsley et al[24] reported the discovery of the first centrally active allosteric modulators of mGlu5. A small molecule library was developed around a set of highly active benzamides. This compound class also plays a prominent role in chapter II, since it is one of the larger classes of compounds identified by the QSAR model.

The connection between positive and negative allosteric modulation of mGlu5 is made by Chen et al[25] showing that the PAMs 3,3-difluorobenzaldazine and CDPPB bind to the same site as the NAM MPEP. Other PAMs not binding to this site are reported. Eliminating MPEP binding through mutation of mGlu5 or binding a neutral ligand to this site reduces or antagonizes positive allosteric modulation of the receptor.

Additional information on positive allosteric modulation of mGlu5 can be found in the introduction to chapter II.

Positive allosteric modulation of metabotropic glutamate receptor subtype 4

Positive allosteric modulation of mGlu4 is thought of as a possible treatment for Parkinson's disease. This is based on observations of Valenti[26] et al that activation of mGlu4 created antiparkinsonsian actions in behavioral rodent models via inhibition of transmission at the striatopallidal synapse. Furthermore, activation of mGlu4 could have neuroprotective effects by reducing the release of glutamate in the substantia nigra[27].

Maj et al[28] showed the neuroprotective effects of (-)-PHCCC against NMDA and β-amyloid protein toxicity. They reported (-)-PHCCC before as a positive allosteric modulator of mGlu4.

Allosteric modulation of human mGlu4 was reported for MPEP and SIB-1893 by Mathiesen et al[29] in 2003. The selectivity of these compounds for mGlu4 can be derived from the fact that neither of them had any effect on either mGlu2 expressing cells or the parent cell line.

In 2008 Niswender et al[4b] published data derived from the Vanderbilt high-throughput screen for mGlu4 PAMs which was also employed for the work reported in chapter III. A set of 434 mGlu4 PAMs was identified in the screen. Potencies, maximal glutamate responses, and fold shifts in concentration response curves were reported for some of these compounds. Several experiments were conducted to establish the selectivity of the novel mGlu4 PAMs with respect to other mGlus. The authors point out that the aforementioned MPEP and SIB-1893 are potent mGlu5 antagonists rendering them unfeasible as selective mGlu4 PAMs. PHCCC works as an mGlu1 antagonist and has low potency and solubility. The lead compound highlighted in this reference (VU0155041) potentiated glutamate at mGlu4 by a factor of eight, is highly selective, and soluble in an aqueous vehicle.

Additional information on positive allosteric modulation of mGlu4 can be found in the introduction to chapter III.

Negative allosteric modulation of metabotropic glutamate receptor subtype 5

Negative allosteric modulation of metabotropic glutamate receptor subtype 5 has potential in the treatment of fragile-x syndrome. For an overview of literature supporting this connection see Dölen et al[14].

In 1999 Varney et al[30] employed high-throughput screening to find the first two selective antagonists of mGlu5, SIB-1757 and SIB-1893. While showing μM activity at human mGlu5a, there was no measurable effect on other mGlu subtypes except a minute agonist activity of SIB-1893 on mGlu4 (26.4μM). Furthermore, both compounds were inactive for recombinant ionotropic glutamate receptors.

Gasparini et al[31] modified SIB-1757 and SIB-1893 to generate 2-Methyl-6-(phenylethynyl)-pyridine (MPEP), a selective mGlu5 NAM with even higher potency (36nM for human mGlu5a). No activity was measured for other metabotropic or ionotropic glutamate receptors. Also, MPEP

showed no activity in the absence of agonists. Additional work was carried out to show that MPEP acts at a novel pharmacological site of the mGlu5 transmembrane region[32] and to develop it into a mGlu5 selective radioligand[33].

Based on the development of MPEP, Yan et al[34] could suppress two of the symptoms of a fragile X syndrome mouse model, sensitivity to audiogenic seizures and the tendency to spend more time in the center of an open field. They conclude that this could reduce symptoms of fragile X syndrome.

Rodriguez et al[35] published two close derivatives of MPEP (Br-5MPEPy and M-5MPEP) which are partial antagonists of mGlu5. A third compound (5MPEP) exhibits no effect on mGlu5 on its own but blocks both MPEP (NAM) and CDPPB (PAM) from modulating mGlu5.

In 2009 the same group reported a set of novel mGlu5 NAMs representing scaffolds different from MPEP[36]. Three structurally diverse compounds show sub-µM activity towards inhibition of mGlu5. One of the NAM leads could also be developed into weak PAMs for the first time in a non-MPEP-like ligand.

Prediction of carbon chemical shift

Empirical methods for general carbon chemical shift prediction rely on large databases of compounds with known chemical shifts. Three options are common to predict chemical shifts for a small molecule: i) look up a similar compound with known chemical shifts in the database, ii) derive addition rules from known compounds in the database to describe how certain substituents and fragments influence the chemical shifts, and iii) train machine learning methods on the known chemical shifts.

Predicted chemical spectra can be employed to distinguish between different structural proposals for an unknown compound, assisting with its structure elucidation[37]. In a similar fashion possible

stereoisomers can be filtered to reduce the number of structure verification experiments[38]. Another interesting application is the calculation of chemical properties like the natural logarithm of the n-octanol/water partition coefficient (logP) based on predicted chemical shifts, molar volume, and hydrogen bonds[39].

Kalchhauser and Robien[40] introduced the computer program CSEARCH for prediction and automatic assignment of carbon chemical shifts. The original version was based on 8,000 spectra from the literature. The chemical shift prediction is based on the HOSE[41] code with increasing number of spheres (1..5). The automatic assignment is based on constructing an isomorphism between predicted and experimental chemical shifts. At each level of HOSE code prediction experimental shifts are sorted out which can be definitively assigned. The remaining shifts constitute a smaller sub-matrix. With increasing level of HOSE code prediction, the number of unassigned shifts should go to zero.

An example of structure elucidation based on addition rules is Tusar et al[42]. Here both proton and carbon chemical shifts are predicted as the sum of a basic shift depending on the environment of the atom of interest, cross-correlation terms for substituents, and correction terms for configuration and conformation which are not described by the first two terms. A structure generator creates all possible molecules under given constraints for which chemical spectra are predicted. The one closest to the experimental spectrum is selected as the solution.

The CAST/CNMR system introduced by Satoh et al in 2003[43] is based on a database combining structural information with NMR chemical shift data. The CAST method encodes the environment of each carbon atom including stereochemistry. This information allows looking up similar atoms in a database of 733 compounds. The predicted shift is the average of all hits found.

This approach was expanded later to include the size of rings which will influence the chemical shifts significantly[44]. The examples show improved chemical shift predictions for several natural

products, but again the overall error cannot be assessed. The size of the database doubled, making direct comparison of the quality of the predicted shifts difficult.

Blinov et al[45] reported a partial least square regression model trained on 2 million chemical shifts. The encoding is similar to Bremser[41] and Meiler[46]. The number of latent variables was optimized on approximately 10% of the dataset. The optimal amount is reported as being between 20 and 70. Furthermore, the optimal number of spheres around the atom of interest was determined to be four. Optimization of carbon atom types for the atom of interest led to eleven classes based on hybridization, number of hydrogen atoms attached, and participation in an aromatic system. To further improve the predictions cross-correlation terms between all atoms in the first three spheres around the atom of interest which were not more than three bonds apart from each other were introduced. This lowered the $rmsd$ between actual and predicted chemical shifts to 2.76ppm (average deviation 1.85ppm).

The focus of this work is to utilize machine learning to establish quantitative structure-activity or structure-property relationships. These trained models were successfully employed to identify positive allosteric modulators of metabotropic glutamate receptor subtypes 4 (chapter III) and 5 (chapter II), negative allosteric modulators of subtype 5 (chapter IV), and to predict carbon chemical shifts (chapter V).

CHAPTER II

IDENTIFICATION OF METABOTROPIC GLUTAMATE RECEPTOR SUBTYPE 5 POTENTIATORS USING VIRTUAL HIGH-THROUGHPUT SCREENING

Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305.

Introduction

Glutamate is the primary excitatory neurotransmitter in the mammalian central nervous system (CNS) and activates metabotropic glutamate receptors (mGlus), which are coupled to downstream effector systems through guanine nucleotide binding proteins (G proteins)[47]. The mGlus 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 mGlu subtypes, the opportunity exists for developing therapeutic agents that selectively interact with mGlus involved in only one or a limited number of CNS functions. Such drugs could have a dramatic impact on development of novel treatment strategies for a variety of psychiatric and neurological disorders including depression[48], anxiety disorders[49], schizophrenia[12a, 50] , chronic pain[51], epilepsy[52], Alzheimer's disease[53], and Parkinson's disease[54]. The mGlu5 receptor subtype is a closely associated signaling partner of the ionotropic N-Methyl D-Aspartate receptor (NMDAR) and may play a significant role in setting the tone of NMDAR function in forebrain regions containing neuronal circuits important for cognitive behavior and for reporting on efficacy of antipsychotic agents[50c].

Activators of mGlu5 may provide a novel approach to treatment of schizophrenia

Activation of mGlu5 potentiates NMDAR function in forebrain circuits thought to be disrupted in schizophrenia. The mGlu5 selective allosteric antagonist 2-methyl-6-(phenylethynyl)-pyridine (MPEP) potentiates the effect of the non-competitive NMDAR antagonist phencyclidine (PCP) in behavioral phenotypic assays[55] and mGlu5 knockout mice have deficits in pre-pulse inhibition in acoustic startle response behavioral assays compared with wild-type mice[55c, 56]. Positive allosteric modulators of mGlu5 have recently been developed and reported[21-24, 57]. Four well-characterized structural classes of mGlu5 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-2H-isoindol-2-yl)methyl]phenyl}-2-hydroxybenzamide] (CPPHA) and [3-cyano-N-(1,3-diphenyl-1H-pyrazol-5-yl)benzamide] (CDPPB), and an oxadiazole chemotype represented by ADX-47273[58]. Despite striking functional similarities, radioligand binding studies revealed different mGlu5 binding profiles for DFB and CDPPB compared with CPPHA[21, 23, 59]. Both CDPPB[23-24] and ADX-47273[58a, b, 58d] 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 lack of solubility in many vehicles[57]. However, some improvement of physicochemical properties was recently reported for the mGlu5 ago-potentiator ADX-47273[58c]. 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[22, 60]. For these reasons, further validation of mGlu5 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 generation of ligand-target interaction datasets[61]. 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 GPCR screening capabilities featuring enhanced sample handling and increased throughput via miniaturization up to 1,536 well format have recently been reported for targets such as M1 acetylcholine receptor[62] and 5HT2b serotonin receptor[63]. However, current literature suggests that one marketable drug emerges from the information gained by screening approximately one million compounds[61a]. If fewer compounds could be tested without compromising the probability of success, library purchasing cost and screening time as well as failure rates in clinical testing may be reduced[61, 64].

Quantitative structure activity relations in drug discovery

Quantitative structure activity relations (QSAR) attempt to model complex non-linear relationships between the chemical and physical properties of molecules and their biological activity (reviews[65]). Hansch et al. established classical QSAR analysis as a paradigm by reporting the use of Hammett substituent constants to establish a quantitative relation between electron density and biological activity[66]. At the same time, they introduced a new hydrophobic parameter, the partition coefficient (P) of the compound in a 1-octanol-water system (logP). Variations and extensions of Hansch's analysis have been applied to drug discovery for over 40 years and rely on well-studied scalar or 2D descriptors such as calculated logP (cLogP), 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[7, 67]. High-resolution methods such as Comparative Molecular Field Analysis (CoMFA)[68] and Comparative Molecular Similarity Indices Analysis (COMSIA)[69] 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 (size of a substituent in position A, presence of a negatively charged group in position B, atom type in heterocyle C, etc.). Examples of fragment-based strategies include MACCS[70], binary structural keys based on occurrence/counts of up to 166 different chemical features found in a compound; HQSAR[71], a 2D method for capturing chiral information based on a Molecular Hologram hashing algorithm without the requirement for generation of 3D coordinates; and SKEYS/FRED, a combination of MDL structural key based fingerprints with an evolutionary algorithm[72].

Traditional 2D- and 3D-QSAR methods often require fragment-based structural encoding schemes[70b, 73] or conformational superposition of biologically active conformations of the chemical structures[68b, 74] that may restrict the utility of resulting models to predictions related to single chemotypes[73] or single protein binding sites[73, 74b]. While suitable for optimization of a lead structure in a small focused library, such encoding schemes often preclude analysis of large, diverse databases as the 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. 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[74b] and to identify novel chemotypes with biological activity at query targets from virtual screens[75]. 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[76]. 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[77] and in an investigation to separate the activity of carcinogenic and non-carcinogenic compounds in a rodent toxicity model[78]. Autocorrelation functions are fragment independent, invariant to translation and rotation, and encode identity and electronic attributes of molecular structure including atom types, partial atomic charges, electronegativity and polarizability into vector representations[5b]. Several studies have employed autocorrelation descriptors for training machine learning algorithms for applications including separation of dopamine agonists and benzodiazepine receptor agonists[79], virtual screening for chemical library enumeration[80], and identification of novel chemotypes[81]. Surface area correlation functions store molecular shape geometry for molecules with known biological activity into neural networks for shape-based molecular recognition in external datasets, as reported for the analysis of corticosteroid-binding globulin activity of steroids[82]. 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[83].

Application of machine learning algorithms to establish QSARs

Machine Learning algorithms have proven to be of practical value for approximating non-linear separable data, especially for classifying biological target data[67b, 84]. Recently, a machine learning approach was applied to generate a model for the tubulin polymerization activities of a library of 250 analogs of the anti-cancer drug Epothilone[67a]. ANNs have been successfully applied for many years in chemistry and biochemistry to generate QSAR models[7,85,86]. Studies were reported involving prediction of dihydrofolate reductase inhibition based on data derived from high-throughput screening using pre-clustering and evolved neural networks[87] as well as applications for prescreening compounds for HIV inhibition while optimizing specificity and potency[88]. Our group recently published a theoretical comparison of machine learning techniques for identification of compounds that are predicted allosteric modulators of the mGlu5 glutamate response[89].

Quantitative structure activity relation models for mGlu5 positive allosteric modulation

The objective of the present research is to employ ANNs to develop QSAR models for mGlu5 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 mGlu5 glutamate response. Such models may also be useful for 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[22, 35, 60]. Activity data for mGlu5 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 selection of an optimal descriptor subset yields QSAR models that enrich active compounds by a factor of up to 38 in independent datasets. The method is applied to a virtual screen of a commercial library of ~450,000 available compounds. A set of

824 compounds with predicted mGlu5 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 mGlu5 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 mGlu5 positive allosteric modulators (PAMs).

Discussion of 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, $a + \frac{b}{1+(\frac{x}{c})^d}$. 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 µM), confidence intervals were within a range of 300 nM. For a PAM with low potency (EC_{50} value above 1 µM), 95% confidence intervals were generally within a range of 1.5 µM. 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, phenylethynl-phenyl, and benzamide PAMs is detailed in Table 1 and Figure 1.

Table 1: Summary of fit statistics for one example of each of the major scaffolds identified using a virtual high-throughput screen. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

		95% Confidence Limit		
Scaffold	**IC_{50} (nM)**	**Upper Limit (nM)**	**Lower Limit (nM)**	**r^2**
Phenylethynyl-phenyl	28.0	43.1	12.8	0.99
Benzamide	344	382	306	0.99
Benzoxazepine	4120	4650	3590	0.99

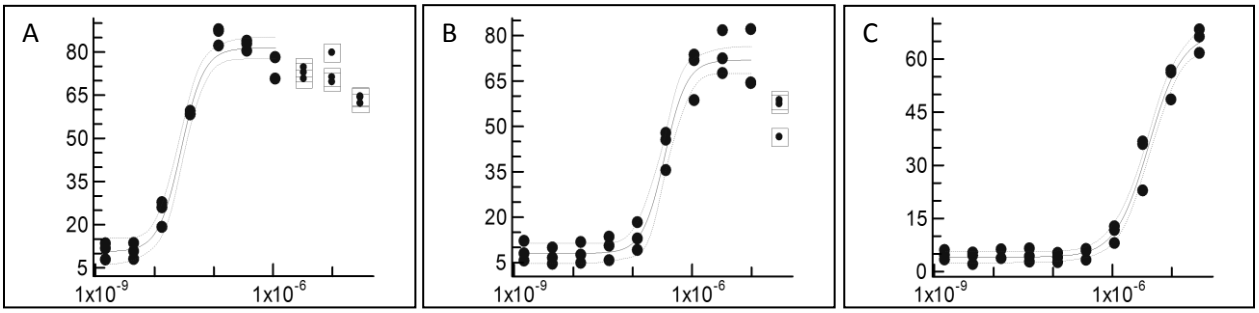


Figure 1: Concentration response curves for (A) phenylethynyl-phenyl, (B) benzamide, and (C) benzoxazepine PAMs. Solid circles indicate individual data points. Solid lines indicate fit derived from averaged data. Dotted lines indicate 95% confidence limits. Points within squares were excluded from analysis due to agonist activity. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

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 reduction of noise

through the increased ratio of datasets versus weights. An increased ratio of datasets 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 doesn't add to the determination of the optimal solution.

Optimization of molecular descriptor set improves 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 2). The *root mean square deviation* ($rmsd$, see Equation 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 3). For a detailed discussion of these measures see Methods. The individual sensitivity value for 'XlogP' (0.97) remained the highest in the baseline network with the remaining input sensitivity distributed across the other scalar descriptors (Figure 2a). Keeping the scalar descriptors in the following models allowed comparing their sensitivity with this baseline.

Equation 1:

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

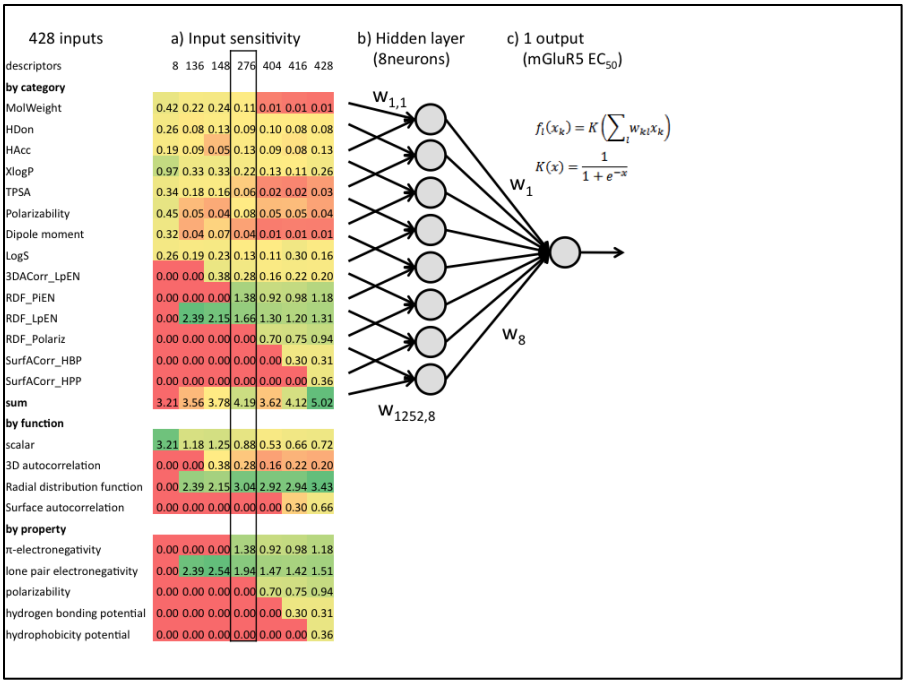


Figure 2: Schematic view of an ANN: a) Up to 1,252 descriptors (from 35 categories) are fed into the ANN input layer. b) The weighted sum of the input data is modified by the activation function and serves as input to the next layer. c) The output predicts the biological activity of the input molecule based on complex non-linear 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. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

Table 2: Summary of 1,252 molecular descriptors in 35 categories computed with ADRIANA. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

	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		σ atom charges	2DA_SigChg	11
11		π atom charges	2DA_PiChg	11
12		total charges	2DA_TotChg	11
13		σ atom electronegativities	2DA_SigEN	11
14		π 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		σ atom charges	3DA_SigChg	12
19		π atom charges	3DA_PiChg	12
20		total charges	3DA_TotChg	12
21		σ atom electronegativities	3DA_SigEN	12
22		π 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		σ atom charges	RDF_SigChg	128
27		π atom charges	RDF_PiChg	128
28		total charges	RDF_TotChg	128
29		σ atom electronegativities	RDF_SigEN	128
30		π 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

Table 3: The *rmsd*, *auc*, and *enrichment* values for all mGlu5 PAMs QSAR models.

Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

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

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 3). To further optimize the set of descriptors, the least sensitive descriptor categories were systematically removed in an iterative process (Table 3 iterations 2-6; Figure 2a). 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 3.

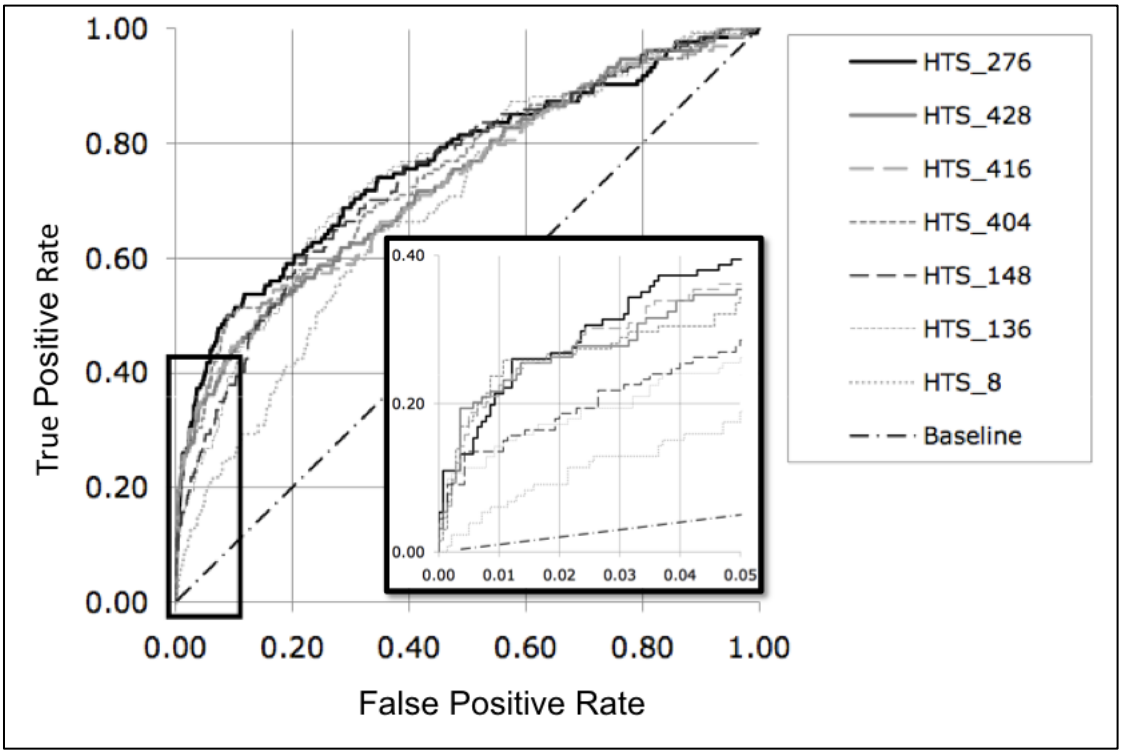


Figure 3: Receiver Operating Characteristic (ROC) curve plot: Traditional (8) scalar QSAR descriptors (HTS_8, dotted grey line) were compared to groups of ADRIANA scalar and vector descriptor sets from the input sensitivity analysis (see Figure 2a) by plotting ROC curves to examine the initial slope. 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 mGlu5 experimental HTS dataset. Based on the ROC curve analysis, HTS_276 descriptors (heavy black line) and HTS_428 descriptors (heavy grey line) displayed the best signal to noise profiles. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

Iterations 1-4 remove 152 descriptors to yield a set of 276 descriptors including the eight scalar descriptors, the 3D auto-correlation lone pair electronegativity, and the radial distribution functions for lone pair electronegativity and π-electronegativity (Table 3; Figure 2a). Retraining

of the ANN with 276 descriptors yields a $rmsd$ 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 π-electronegativity and the 3D autocorrelation function for lone-pair electronegativity were removed (Figure 2a; Figure 3). 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, auc value of 0.738 and an $enrichment$ of 25 (Table 3). 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 dataset combined with the smallest descriptor set. This network was used in all *in silico* screening experiments described below.

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'[90] and therefore 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 the datasets through oversampling yields better results than two undersampling strategies

The oversampling strategy employed throughout the study was compared with two approaches that undersample inactive compounds when using the optimized 276 input descriptors (see Methods, Figure 4). Usage of randomly chosen inactive compounds resulted in a $rmsd$ value for the independent data of 0.221, 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, auc value of 0.654 and an $enrichment$ of 2 (Table 3).

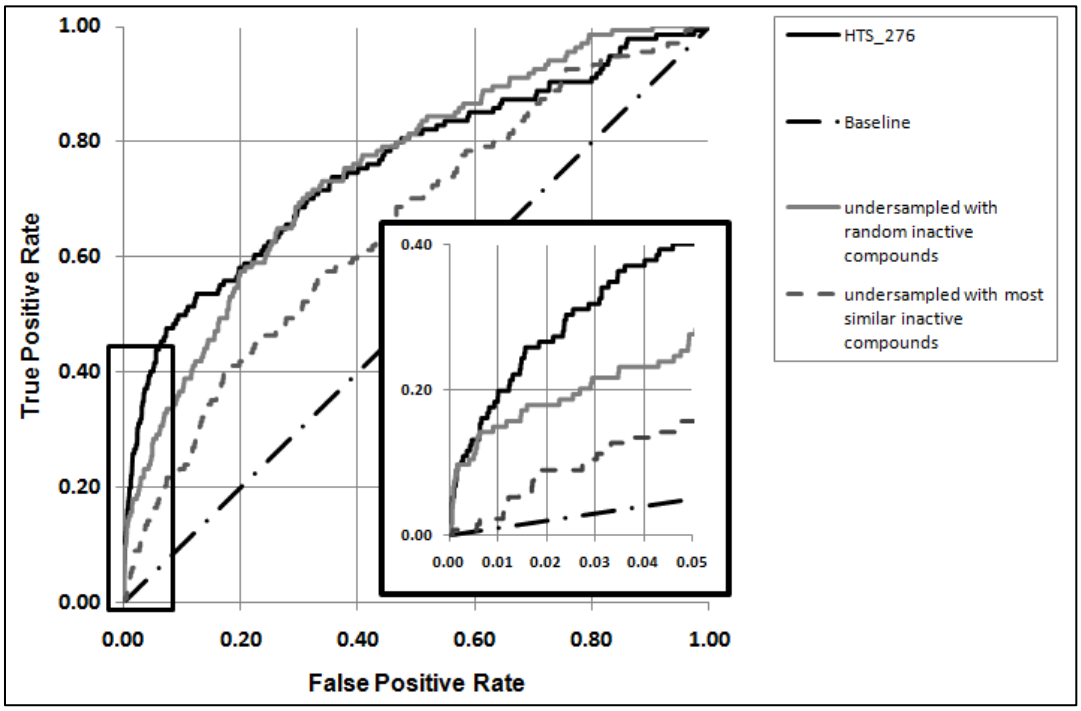


Figure 4: Receiver Operating Characteristic (ROC) curve plots for undersampling methods comparison: ROC curve analysis showing optimized descriptor set HTS_276 based on oversampling (solid black line) compared to undersampling using a random selection of inactive compounds for monitoring and training dataset (solid grey line) as well as a selection of most similar inactive to active compounds (dashed grey line). Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

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

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, Surface Autocorrelation) reveals superior performance of radial distribution functions across the six ANN models tested (Figure 2a). Surface Autocorrelation functions were only tested in the first two models (428 and 416 descriptors) due to lower sensitivity scores (Figure 2a). Analysis of input sensitivity by property revealed high sensitivities for π 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 5) that are well represented in the training dataset contain extended π 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 prediction of mGlu5 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 performance of the QSAR model (Table 3; Figure 2a).

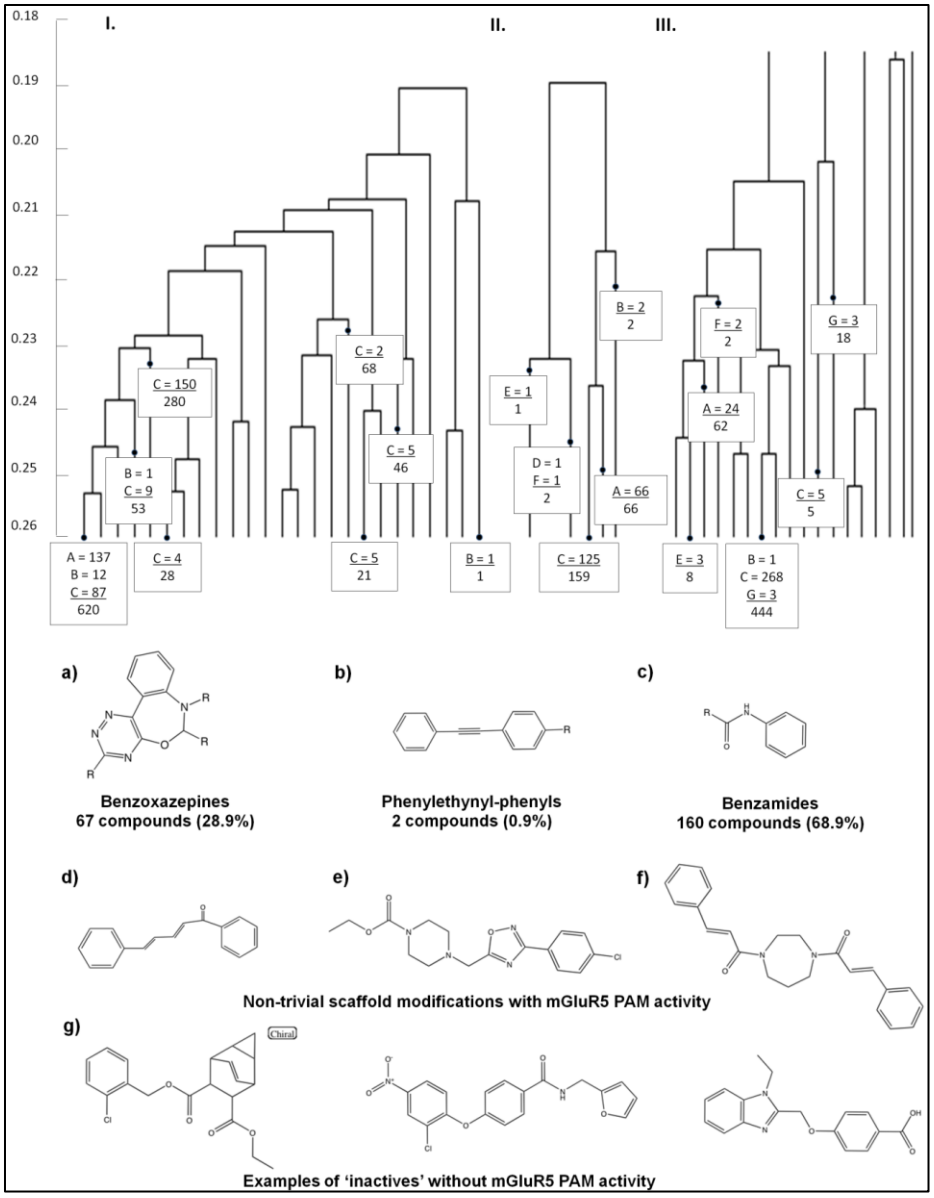


Figure 5: Scaffold category analysis: I) 1,382 mGlu5 PAMs from HTS were clustered with the Mathematica package using the Tanimoto coefficient of the largest common substructure as distance measure. Three major scaffolds are constituted by 137 benzoxazepines (9.9%, a), 14 phenylethynyls (1.0%, b), and 267 benzamides (19.3%, c). II) Scaffold composition of active compounds in post-screen. III) Scaffold composition of inactive compounds in post-screen. Compounds d), e) and f) are non-trivial mGlu5 PAM scaffold modifications identified by the virtual screen using the ANN QSAR model. Panel g) highlights representative compounds found inactive in the post-screen. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

A recent study demonstrated the necessity for optimizing molecular descriptor types for each individual dataset to yield optimal QSAR models[91]. Other studies independently reported the radial distribution function as most robust molecular descriptor category in capturing the structure activity signal from experimental HTS datasets[77-78].

Virtual screening of 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 μM for mGlu5 PAM activity were selected. An additional 11 compounds were chosen based on visual inspection by an expert medicinal chemist (C.W.L.) from clusters at a lower potency cutoff of 10 μM 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 one antagonist. Follow-up CRC assays confirmed 232 compounds with various activities at mGlu5. The compounds were classified as pure potentiators (177), and partial agonists (55). The remaining 27 compounds were either inactive (21, see Figure 5g), fluorescent (2) or showed increased baseline measurements in the fluorescent assays (4). This result reflects an $enrichment = 232/824 \times 144{,}475/1{,}356 = 30$ relative to the initial experimental HTS hit rate. The experimental enrichment is consistent with the enrichment values predicted from analysis of an independent dataset during development of the QSAR model (Table 3).

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 mGlu5 potentiators

According to MACCS fingerprint-based clustering[92] using a Tanimoto coefficient[92b] of 0.75 for similarity, of the 232 compounds with confirmed mGlu5 activity identified in our virtual screen of the ChemBridge commercial library: 67 compounds (28.9%) were classified as benzoxazepines with pure potentiator activity (Figure 5a); 2 compounds (0.9%) were structurally similar to MPEP (containing a phenylethynyl-phenyl moiety) and displayed partial agonist activity (Figure 5b); 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 5c); and 3 compounds (1.3%) contained other non-trivial scaffold modifications (Figure 5d-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 μM cutoff).

Major scaffolds are evenly distributed throughout training, monitoring, and independent datasets

The library of 1,382 compounds identified as active in the original HTS screen was analyzed using a clustering approach. At a cutoff of 25 % similarity, 25 different scaffolds were identified (see Methods, Figure 5).

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; 137 compounds were classified as benzoxazepines, with 114 compounds (83.2%) in the training set, 13 compounds (9.5%) monitoring set, and ten compounds (7.3%) in the independent set; and lastly, the mGlu5 PAM library contained 14 compounds structurally similar to MPEP (containing a phenylethynyl-phenyl moiety), and distributed throughout the data sets as followed: twelve compounds (85.7%), one compound (7.1%), and one compound (7.1%).

The 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 datasets (Figure 5a-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 relation for each of the scaffolds. In the early stages of drug discovery, time can be saved through acquisition of commercially available compounds to enumerate focused libraries around confirmed HTS hit compounds. The results can help planning of synthetic chemistry efforts.

Benzamides, benzoxazepines, and MPEP-like compounds are enriched among active compounds in the post-screen

The post-screen 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 5). The majority of benzoxazepine and benzamide derivatives form a single cluster at this cutoff containing 125 and 66 compounds, respectively. The non-trivial scaffold modifications with mGlu5 PAM activity were found in two separate clusters. Each non-

trivial scaffold modification was observed once in the post-screen 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 post-screen 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 post-screen library contain 53 % benzamides, benzoxazepines, and MPEP-like compounds

The remainder of the post-screen library was shown to be inactive towards 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 5a-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 towards mGlu5. Derivatives of the non-trivial compounds (Figure 5d-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 mGlu5. 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 screen 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 post-screen library we find 229 derivatives of these scaffolds with activity and 312 without. The enrichment of active compounds that share one of these scaffolds is 44 and 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 the rate of active compounds would have been lower.

Significant numbers of hit compounds are non-trivial modifications of original HTS screen hits

While 99% of the newly identified mGlu5 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 5a-c). The remaining 28% had multiple modifications with respect to any of the hit compounds in the original HTS screen (Figure 5d-f). These compounds would have been difficult to identify with a similarity search as discussed above.

High potency cutoff may introduce bias to close derivatives of original HTS screen hit compounds

As part of our virtual screen, several different potency cutoffs (300nM, 1μM, 2μM, 5μM, 10μM) were employed to identify a compound library size that was tractable for experimental ordering and testing. Selection of a predicted potency cutoff of 1.0μM for mGlu5 PAM activity might have biased the majority of the 824 compounds towards molecules with similar chemotypes to the compound classes that represented the majority of the known active compounds included in our training dataset (benzoxazepines, phenyl ethynyls and benzamide-containing scaffolds) (Figure 5a-c). However, with the identification of three non-trivial modifications of known chemotypes having mGlu5 PAM activity (EC_{50} > 2.5 μM), "scaffold hopping" appears to be possible using this method (Figure 5d-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 (10-30μM) combined with filters to remove compounds with similar chemotypes to those in the training dataset. 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 mGlu5 virtual screen at a lower potency cutoff (10μM). This small subset of compounds was chosen by visual inspection. We did not discover mGlu5 PAM activity in any of these compounds. The compounds were either fluorescent or inactive in

our experiments. However, this result is inconclusive due to the very small number of compounds selected according to these criteria.

Fragment-independent numerical description deals efficiently with multiple scaffolds

The observation of three non-trivial chemotype modifications underscores the ability of fragment-independent numerical descriptions to map chemical structure of a diverse compound library into a numerical fingerprint. Different classes of compounds displaying mGlu5 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 underlines 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.

Conclusions

In conclusion, machine learning methods (ANN) were used to generate QSAR models from an HTS experimental dataset in virtual screens of an external commercial compound collection for the purpose of enrichment of our local library for compounds with mGlu5 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 minimized noise by excluding less sensitive descriptors from training inputs to maximize signal for mGlu5 and proved to be a crucial step for increasing enrichment for active compounds. Oversampling of actives was included in dataset generation to balance the training of our models and an independent dataset 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 mGlu5 receptor. Model validity was assessed based on 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 previously identified mGlu5 PAM compounds, a significant fraction of these compounds are non-trivial 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 similar chemotypes to those in the training dataset. We would expect substantially reduced enrichment factors in such a scenario.

Methods

Experimental high-throughput screen for mGlu5 potentiators and hit validation

In the initial HTS experiment, 144,475 compounds were tested for allosteric potentiation of mGlu5 using full automation in conjunction with the Vanderbilt HTS facility[3]. 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 6a).

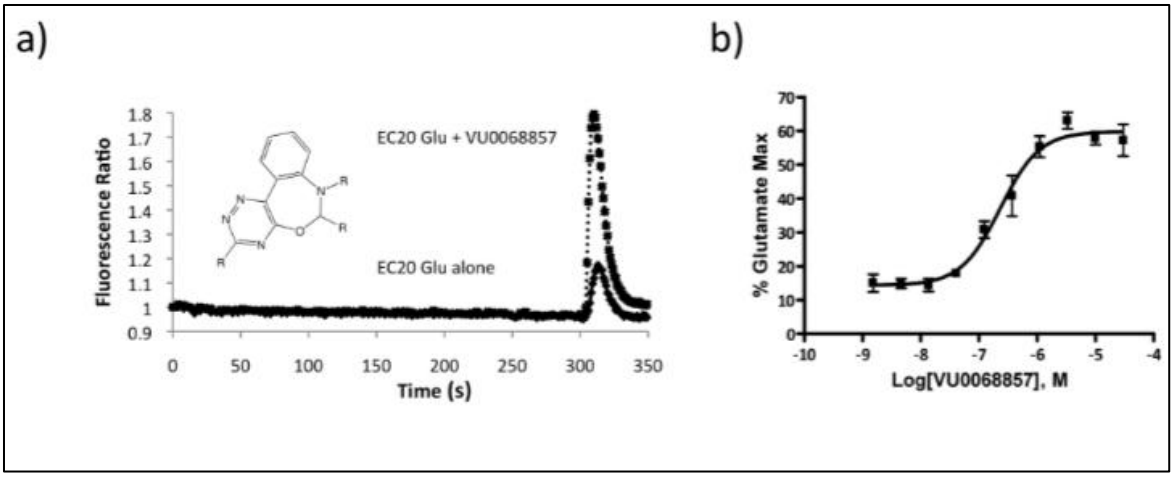


Figure 6: FDSS measurement of intracellular Ca2+ release in response to mGlu5 activation and potentiation by allosteric modulator compounds: a) Agonist induced Ca^{2+} transients were quantified based on the fluorescence change observed in cells treated with an EC_{20} concentration of glutamate plus 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 mGlu5 over a range of 4 log units with 10 point concentration response curves (30 µM – 1 nM final concentration). Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

HEK 293A cells stably expressing mGlu5 were plated in black-walled, clear-bottomed, poly-D-lysine coated 384-well plates (BD Biosciences, San Jose, CA) in 20 µL 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, medium was removed and the cells incubated with 20 µL of 2 µM 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 µL 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 mGlu5 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 6a), phenylethynyl, and benzamide derivatives (Figure 5a-c; manuscript in preparation).

For further analysis the mGlu5 PAM library of active compounds in the original HTS screen as well as the compounds selected for post-screening were clustered using the Mathematica package[93]. The Tanimoto coefficient based on number atoms in the maximum common substructure served as distance metric:

Equation 2:

$$T(molecule_1, molecule_2) = \frac{\#atoms_{substructure}}{\#atoms_1 + \#atoms_2 - \#atoms_{substructure}}$$

In an initial primary screen of ANN selected compounds, single concentrations of compounds (30 µM final) were transferred to daughter plates using the Echo acoustic plate reformatter (Labcyte, Sunnyvale, CA). Compounds were diluted into assay buffer to a 2x stock using a Thermo Fisher Combi (Thermo Fisher, Waltham, MA) which was applied to cells at t = 3s. 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 based on 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 mGlu5 over a range of 4 log units (Figure 6b). Compounds were serially diluted 1:3 into 10 point concentration response curves (30µM – 1nM 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 formula $a + \frac{b}{1+(\frac{x}{c})^d}$ was utilized.

Generation of numerical descriptors for training of QSAR models

For input to machine learning methods the chemical structure of each molecule needs to be described numerically (see Figure 7a). Initially, 3D models of all 144,475 small molecules are generated using the CORINA software package[6]. From the 3D structural models a set of 1,252 numerical descriptors is computed using the ADRIANA software[5]. The descriptors can be classified into 35 categories including eight scalar descriptors, eight 2D and eight 3D auto-correlation functions, eight radial distribution functions, and three surface-auto-correlation functions (see Table 2).

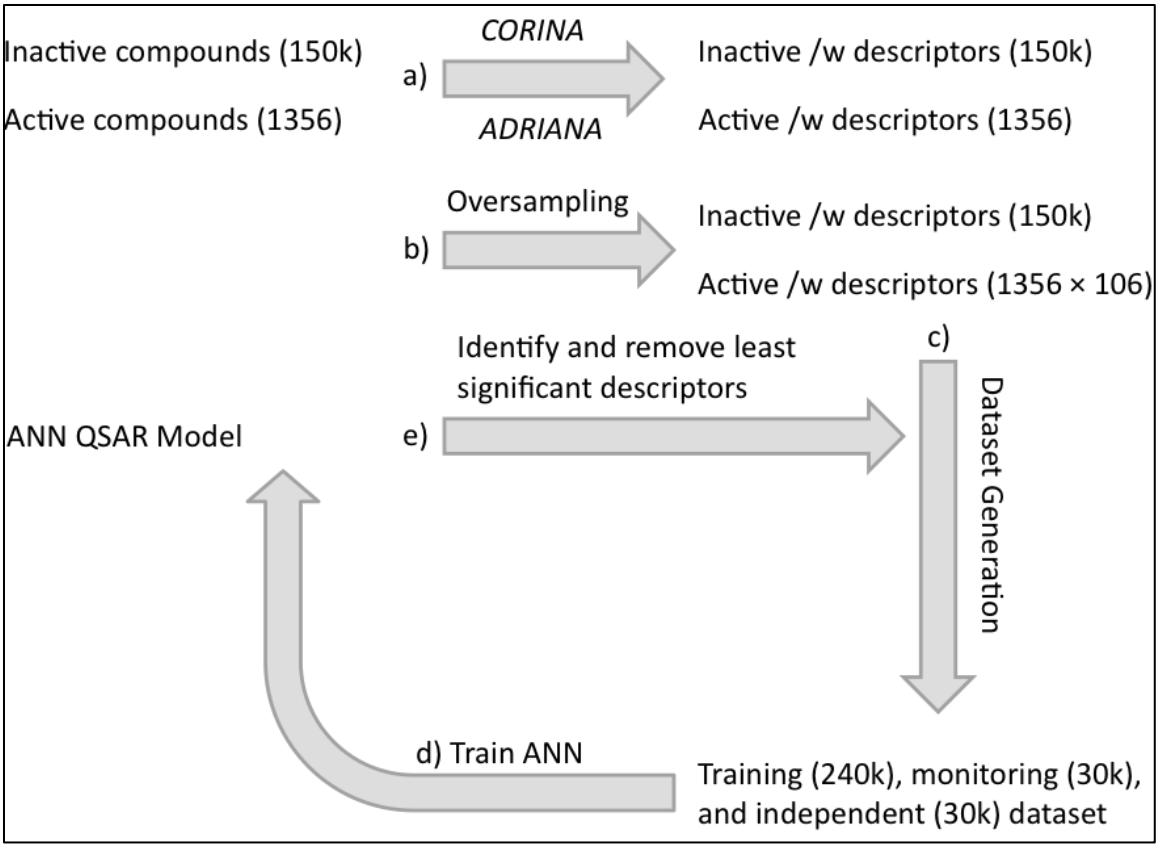


Figure 7: Overall model generation workflow: a) active and inactive molecules were retrieved as MDL SD files from experimental collaborators; 3D structures were generated with CORINA and used as input for calculation of molecular descriptors using ADRIANA; b) active molecules were oversampled 106 times to balance data sets; c) molecules were randomly included in training dataset (80%), monitoring dataset (10%), and independent dataset (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 dataset was achieved. Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

Oversampling was used for balanced training

As detailed above, 1,382 compounds were confirmed to be active potentiators of the mGlu5 glutamate response (0.94% hit rate). Of these, only 1,356 compounds were used as 'actives' in model generation due to difficulty of encoding charged molecules with ADRIANA (see Figure 7a). We refer to the active dataset as these 1,356 compounds. All other compounds were

classified as inactive. In order to maximize information content of the final prediction method, the dataset 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 7b). 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 7b,c,d).

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 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 dataset was kept identical to the oversampling scenario to enable direct comparison. For training and monitoring datasets (i) a random selection of inactive compounds was selected and (ii) the inactive compounds most similar to active compounds were chosen using MACCS fingerprint keys[71] and Tanimoto coefficient as similarity measure.

A monitoring dataset was introduced to early terminate ANN training

The natural logarithm of the experimentally determined EC_{50} value of each compound i was used as output for the ANN models ($exp_i = LnEC_{50,i}$). 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 Equation 1) is used as objective function when training the ANN models.

For training ANNs the dataset is split. Of the total experimental dataset, 115,581 (80%) data points were used for the ANN training (Figure 7c,d). 14,448 (10%) data points were set aside for monitoring during ANN training and initiate early termination (Figure 7c,d). After each training iteration the $rmsd$ of the monitoring dataset was computed. Training was terminated once the $rmsd$ value of the monitoring dataset was minimized. The final 14,448 data points (10%) were reserved for independent testing of QSAR models (see Table 3). Care was taken to avoid any overlap between training, monitoring, and independent dataset. All results reported were obtained for the independent dataset 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 of neurons $j = 1, 2, \dots, n$ containing N_j neurons each. In a pair-wise fashion neurons in neighboring layers are interlinked by weighted connections w_{kl} (Figure 2b). 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:

Equation 3:

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

The output $f_l(x_k)$ then serves as input to the l-th neuron of the next layer (Figure 2b).

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 2a), 8 hidden neurons (Figure 2b), and 1 output (Figure 2c). The sigmoid function

Equation 4:

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

is applied as activation function K of the neurons. The training method used is resilient back-propagation of errors[10b], 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 ($rmsd$, Equation 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 dataset achieved its minimum $rmsd$. The training took up to 13 hours per network using eight cores of a core2 quad 2.33GHz 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 least significant for prediction of PAM activity (Figure 7e; Figure 2a). Objective of this procedure is to reduce the total number of inputs and therefore the total number of weights of the ANN (Figure 7d, e; Figure 2a). 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 is accelerated. Further, 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 2). After completion of the training the ANN represents a multidimensional function:

Equation 5:

$$y = f\left(x_1, x_2, \dots, x_{N_0}\right) = f(\langle x \rangle)$$

with input values $x_1, x_2, \dots, x_{N_0}$ and output y. The partial derivative of each input with respect to the output can be determined numerically and is introduced as "input sensitivity":

Equation 6:

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

For this purpose each input value x_k is altered by a small Δx_k in an independent experiment and the change Δy is monitored. Following this procedure the input sensitivity is determined for each input k by selecting randomly 100 compounds from the independent dataset. The input x_k is perturbed by a small number Δx_k = ±5 % of the input range. The output change Δy is recorded[94]. The input sensitivity of input k is the average ratio observed (Equation 6).

The input sensitivity of each of the 27 non-scalar 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 2a). 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 7c-e; Figure 2a).

Enrichment and area under the curve (***auc***) *as quality measures*

As mentioned before, the *rmsd* between predicted experimental $LnEC_{50}$ was used as objective function for training the ANNs. EC_{50} values for compounds classified as inactive were assumed to be 1mM. Analysis of the *rmsd* proved to be a poor indicator for model quality (see Table 3) as the correlation coefficients between experimental and predicted $LnEC_{50}$ values are typically smaller than 0.5 (see Figure 8).

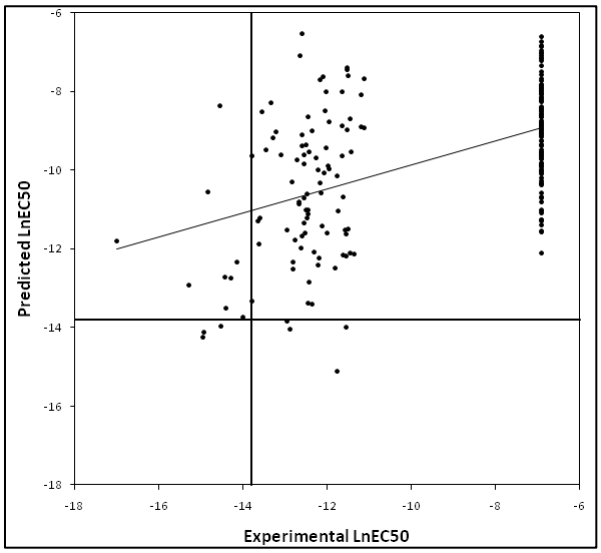


Figure 8: Correlation plot between measured and predicted $LnEC_{50}$ values. Inactive compounds were set to an EC_{50} of 1mM ($LnEC_{50}$= -6.9). The solid lines represent the cutoff used for acquisition of compounds (EC_{50} of 1μM / $LnEC_{50}$= -13.82). Reprinted with permission from Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305. Copyright 2010 American Chemical Society.

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 binary classification based on enrichment and area under the curve (auc).

Receiver operating characteristic (ROC) curves were generated as a measure to evaluate 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 binary classifier (Figure 3). 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 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:

Equation 7:

$$enrichment = \frac{TP}{TP + FP} \Big/ \frac{P}{P + N}$$

When computed for the independent dataset 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 dataset (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 3 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 $LnEC_{50}$ 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 dataset, 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 $LnEC_{50}$ values (see Table 3). 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[10]. 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[5] was used for generation of chemical descriptors. CORINA[6] was used for generation of three-dimensional structures.

CHAPTER III

VIRTUAL HIGH-THROUGHPUT SCREENING AS A ROBUST TOOL TO IDENTIFY METABOTROPIC GLUTAMATE RECEPTOR SUBTYPE 4 POTENTIATORS

Introduction

This study implements a machine learning approach (Artificial Neural Networks) to virtually screen commercially available compounds for potentiators of metabotropic glutamate receptor subtype 4 (mGlu4). Marino and Conn[13b, 95] showed that activation of mGlu4 is a viable option in treating Parkinson's disease. Parkinson's disease (PD) is a debilitating movement disorder that afflicts more than 1 million people in North America. In Parkinson's patients, there is a decrease in GABAergic transmission at the inhibitory striatopallidal synapse within the basal ganglia; this abnormality is thought to contribute to the motor dysfunctions observed in PD patients. Current PD treatments that are focused on dopamine-replacement strategies ultimately fail in most patients because of loss of efficacy and severe adverse effects that worsen as the disease progresses[13a, 96]. Selective activation of mGlu4 could provide palliative benefit in PD. Further, selectively targeting mGlu4 avoids the loss of efficacy and severe side-effects of long-term dopamine replacement therapy. In 2003 Maj[28] et al. reported on the discovery of (-)-PHCCC, the first positive allosteric modulator of mGlu4 with demonstrated selectivity for group III mGlus, but also a partial antagonist for mGlu1 (group I). Around the same time, Mathiesen[29] et al. showed that SIB-1893 and MPEP (a known mGlu5 antagonist[31]) are mGlu4 potentiators.

Despite their tractability as drug targets, the majority of GPCR-based drug discovery programs have failed in the past to yield highly selective compounds. The traditional approach to target the endogenous ligand orthosteric binding site has suffered from a paucity of suitably subtype-selective ligands as orthosteric binding sites are highly conserved between GPCR subtypes. An

alternative approach is to target allosteric sites that are topographically distinct from the orthosteric site, either enhancing or inhibiting receptor activation[97]. Discovery and characterization of allosteric modulators of GPCRs has gained significant momentum over the last few years, especially since the clinical validity of GPCR allosteric modulators was demonstrated with two allosteric modulators entering the market[98]. Thus, allosteric modulation represents an exciting novel means of targeting GPCRs particularly for CNS disorders, a therapeutic area with one of the highest rates of attrition in drug discovery[99].

Recently, Niswender[4b] et al. reported the discovery of 434 potentiators of mGlu4 from a high-throughput screen of approximately 155,000 compounds. The study highlighted a series of cyclohexyl amides joined to a substituted phenyl ring. The structures of these tested molecules and their experimentally determined EC_{50} towards mGlu4 potentiation were employed in the model described in this paper. Engers et al.[4a] discuss the synthesis and evaluation of a set of heterobiarylamides that were derived from the aforementioned compounds and optimized for penetrating the central nervous system. Around the same time, several pyrazolo[3,4-d]pyrimidines were also described to be novel mGlu4 positive allosteric modulators[4c]. Two challenges in further developing potentiators of mGlu4 as a PD treatment strategy are the low hit-rate of 0.3% in the original high-throughput screen resulting in a small number of available ligands and the "flat" structure activity relationship (SAR) around the 'proof of concept' compound PHCCC. Even slight structural modifications lead to complete loss of activity for the reported compounds[4d]. The present study addressed both challenges by identifying additional allosteric potentiators from commercially available compound libraries and exploring the chemical space around the known active compounds.

Quantitative structure activity relationship (QSAR) models describe the often complex, non-linear relation between the chemical and physical properties of molecules and their biological activity; for a review of different methods see Todeschini et al. or Hansch[65]. Classical QSAR was

introduced by Hansch et al. by deriving biological activity from electron density[66]. Modern QSAR techniques employ advanced 2D molecular fingerprints and 3D molecular descriptors coupled with machine learning[7, 67]. In this study, artificial neural networks (ANNs) were trained on descriptors computed with the software package ADRIANA[5] linking chemical properties of small molecules to their potency as potentiators of mGlu4.

Fragment-independent scalar descriptors, 2D and 3D surface and auto-correlation functions, and radial distribution functions are employed to encode a large diversity of chemotypes into comparable mathematical representations[1].

ANNs have been successfully applied in biochemistry to generate QSAR models[7, 85-88]. Our group recently published a theoretical comparison of machine learning techniques for identification of compounds that are predicted allosteric modulators of the mGlu5 glutamate response[1, 89].

Results and Discussion

Artificial Neural Networks (ANNs) were trained to predict the capability of drug-like molecules for allosteric potentiation of the metabotropic glutamate receptor subtype 4 (mGlu4) based on a high-throughput screen as reported by Niswender et al[4b]. Commercially available databases of small molecules were virtually screened for novel potentiators of mGlu4. Hit compounds were verified experimentally.

Descriptor categories were selected according to input sensitivity

Several of the descriptor categories (see Table 4) employ the same chemical property but different encoding functions (2D vs. 3D auto-correlation and Radial Distribution Functions). Therefore, a descriptor optimization strategy was employed to identify the smallest set of descriptors needed for optimal QSAR models. Using this technique, the number of parameters (weights) in the ANNs is reduced, improving the signal-to-noise ratio for the trained models. To

determine the 'least necessary' descriptor categories, the input sensitivity (see Methods) of each input with respect to the output of the ANN, i.e. biological activity prediction, was determined.

Table 4: 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		σ atom charges	2DA_SigChg	11
11		π atom charges	2DA_PiChg	11
12		total charges	2DA_TotChg	11
13		σ atom electronegativities	2DA_SigEN	11
14		π 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		σ atom charges	3DA_SigChg	12
19		π atom charges	3DA_PiChg	12
20		total charges	3DA_TotChg	12
21		σ atom electronegativities	3DA_SigEN	12
22		π 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		σ atom charges	RDF_SigChg	128
27		π atom charges	RDF_PiChg	128
28		total charges	RDF_TotChg	128
29		σ atom electronegativities	RDF_SigEN	128
30		π 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

Optimization of molecular descriptor set improves prediction results

An ANN was trained using only the scalar descriptors 1-8 to report a baseline performance using only naïve descriptors (Figure 9 and Table 5). The *auc* value of 0.631 serves as a basis for comparison in model optimization (Table 5). The *enrichment* equals 1.2 at a compound cutoff

of 2%. The relative *root mean square deviation* ($rmsd$) value for the independent data set is 0.238. 'Total Polarizable Surface Area' was the input with the highest sensitivity (0.87) in this model with the other descriptor sensitivities ranging from 0.07 ('Dipole Moment') to 0.48 ('Hydrogen Bond Acceptors'). The second baseline model involved all 1,252 descriptors as inputs. The auc and $enrichment$ values improved to 0.708 and 7.3, respectively, while the $rmsd$ dropped to 0.234.

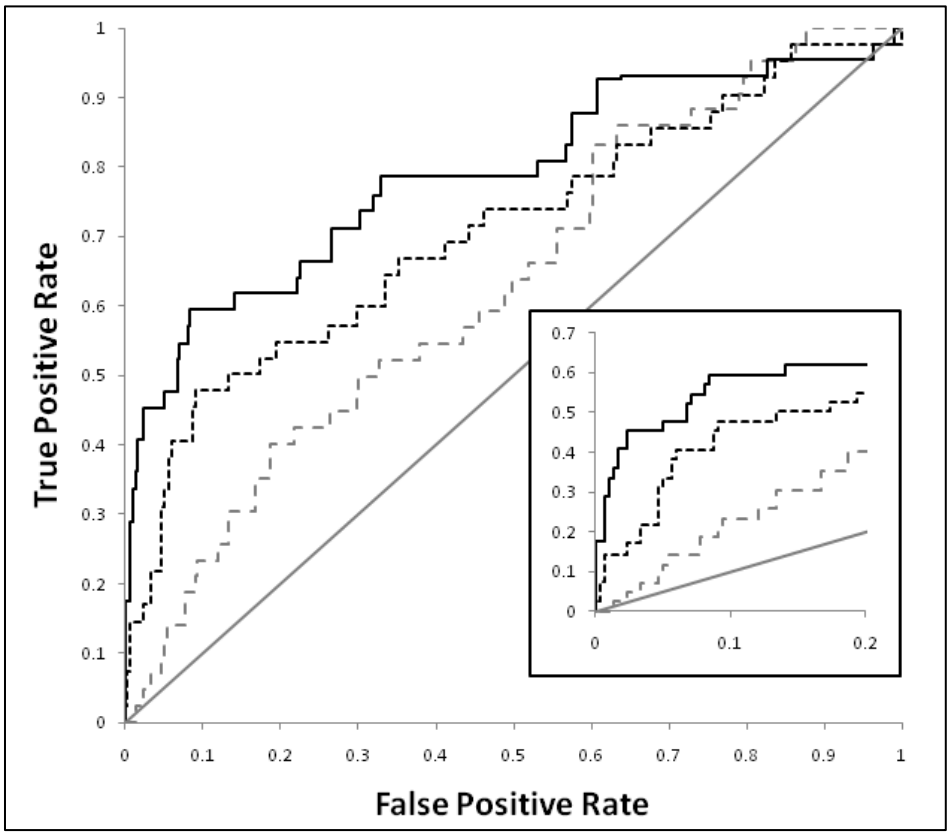


Figure 9: Receiver Operating Characteristic (ROC) curve plot for classical, all, and jury approach (round 1): This plot compares traditional QSAR (eight scalar descriptors, dotted grey line) with utilizing all (1,252, dotted black) available ADRIANA descriptors and a jury approach (solid black line). It demonstrates that ADRIANA descriptors add to the classical approach and that a jury approach improves performance even further. The inset shows the first 20% of the False Positive rate.

In a first round of descriptor optimization, one third of the descriptor categories with the lowest input sensitivity (Equation 11) were removed. Note that the scalar descriptors were kept in all models to facilitate comparison with the baseline. This procedure leads to a final model containing 741 descriptors in 21 categories (see Table 5) with a slightly worse statistical profile: auc: 0.703, $rmsd$: 0.227, $enrichment$: 7.1. The second round yielded a model with 578 descriptors in 17 categories. The quality measures were better than the model with all descriptors with an auc of 0.706, $rmsd$ of 0.229, and an improved $enrichment$ of 13.0. The last iteration left 415 descriptors in 13 categories. While the auc value (0.804) and $rmsd$ value (0.222) improved, the $enrichment$ (10.7) dropped.

Table 5: The $rmsd$, auc, and $enrichment$ values for all round 1 mGlu4 QSAR models

Iteration	Number and type of descriptors		$rmsd$			auc	$enrichment$
			train	monitor	independent		at 2%
All	1252	1-35	0.204	0.232	0.234	0.708	7.3
Scalar	8	1-8	0.232	0.236	0.239	0.631	1.2
1	741	1–8, 14–16, 21–23, 25, 29–33, 35	0.224	0.224	0.227	0.703	7.1
2	578	1–8, 15–16, 23, 25, 30 – 33, 35	0.187	0.212	0.229	0.706	13.0
3	415	1–8, 15, 25, 30–31, 35	0.192	0.211	0.222	0.804	10.7
Jury	**-**	**-**	**0.159**	**0.214**	**0.207**	**0.732**	**15.4**

Jury model combines favorable features of all previous models

As these QSAR models have a comparable quality and enrichment values are affiliated with high uncertainties, a jury approach was tested to combine models. An ANN was trained on the output of the three ANNs with the reduced descriptor sets (see Figure 10). This procedure improved the critical $enrichment$ value to 15.4 and reduced the $rmsd$ to 0.207 (see Table 5). The auc value is with 0.732, a value lower than the 0.804 value reported for the ANN model with 415 descriptors.

However, the improved performance of this model results from the second half of the ROC curve, which is not employed when predicting molecules with high activities (Figure 11).

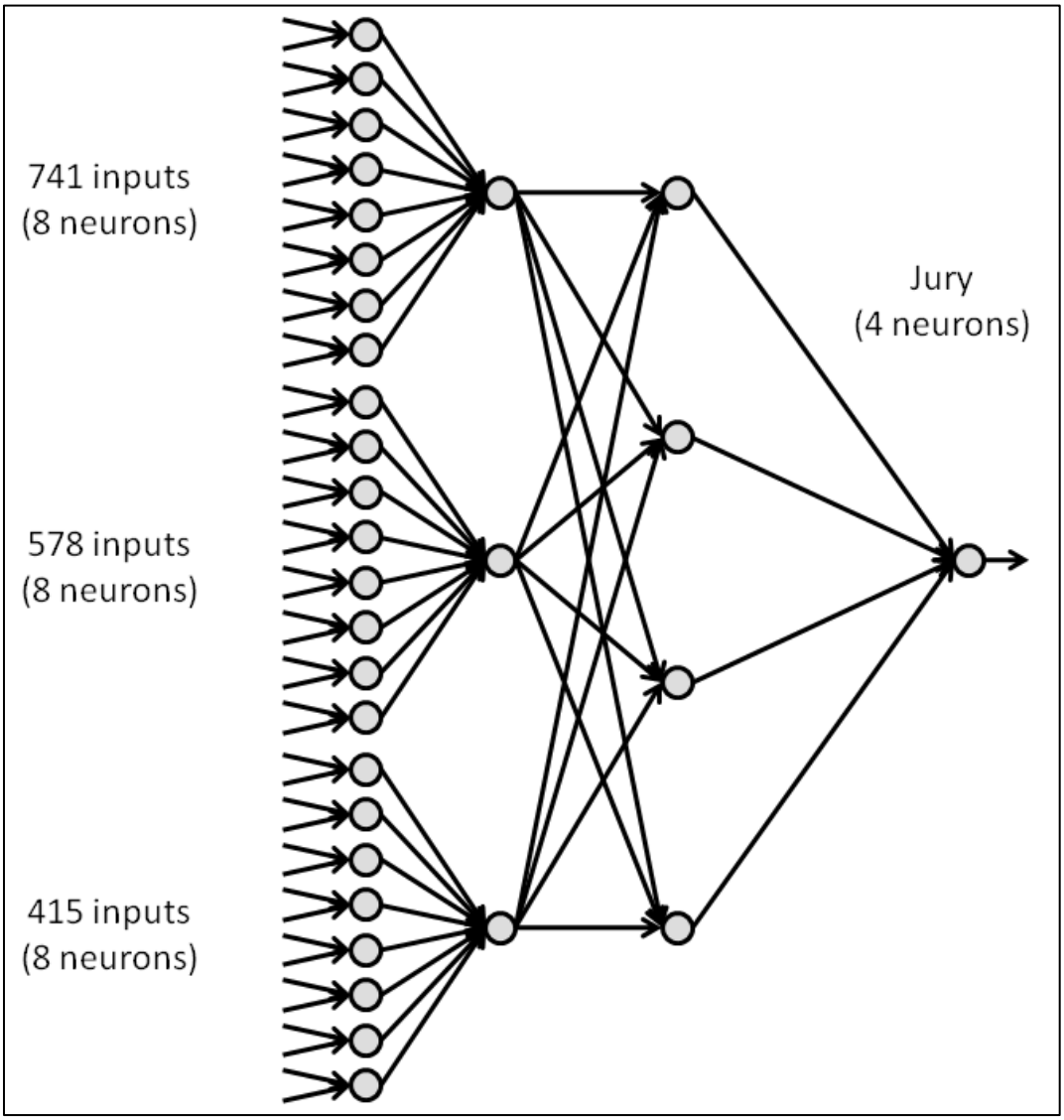


Figure 10: Schematic view of the jury system: The output of the best three ANNs according to the quality measures reported in Table 5 were employed as inputs for a jury ANN with four hidden neurons.

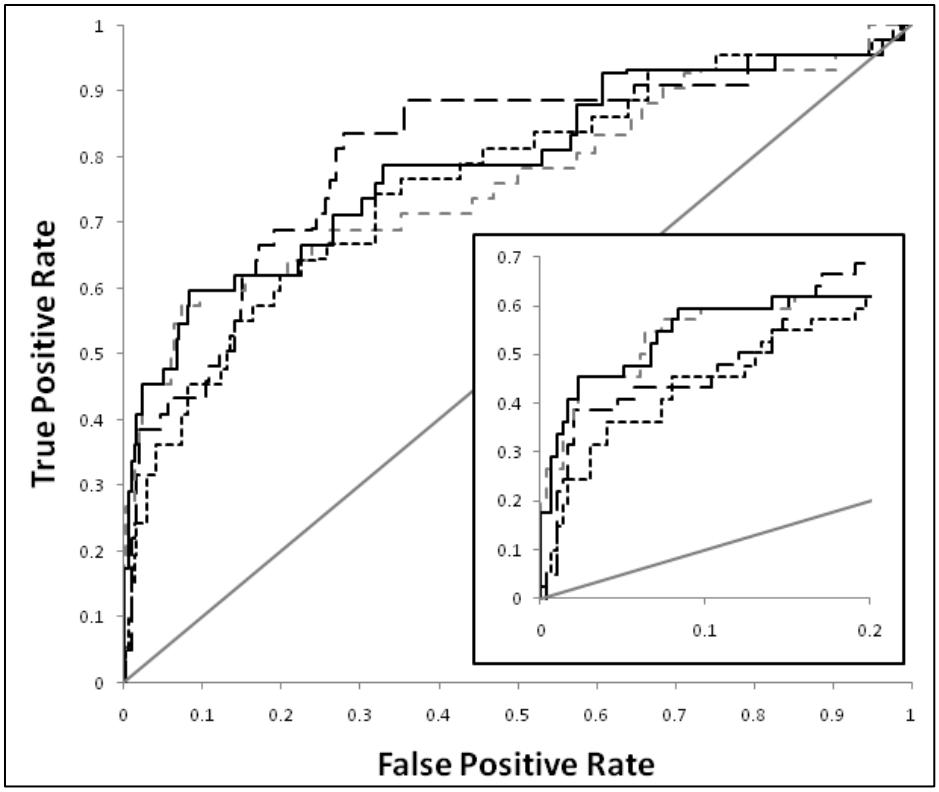
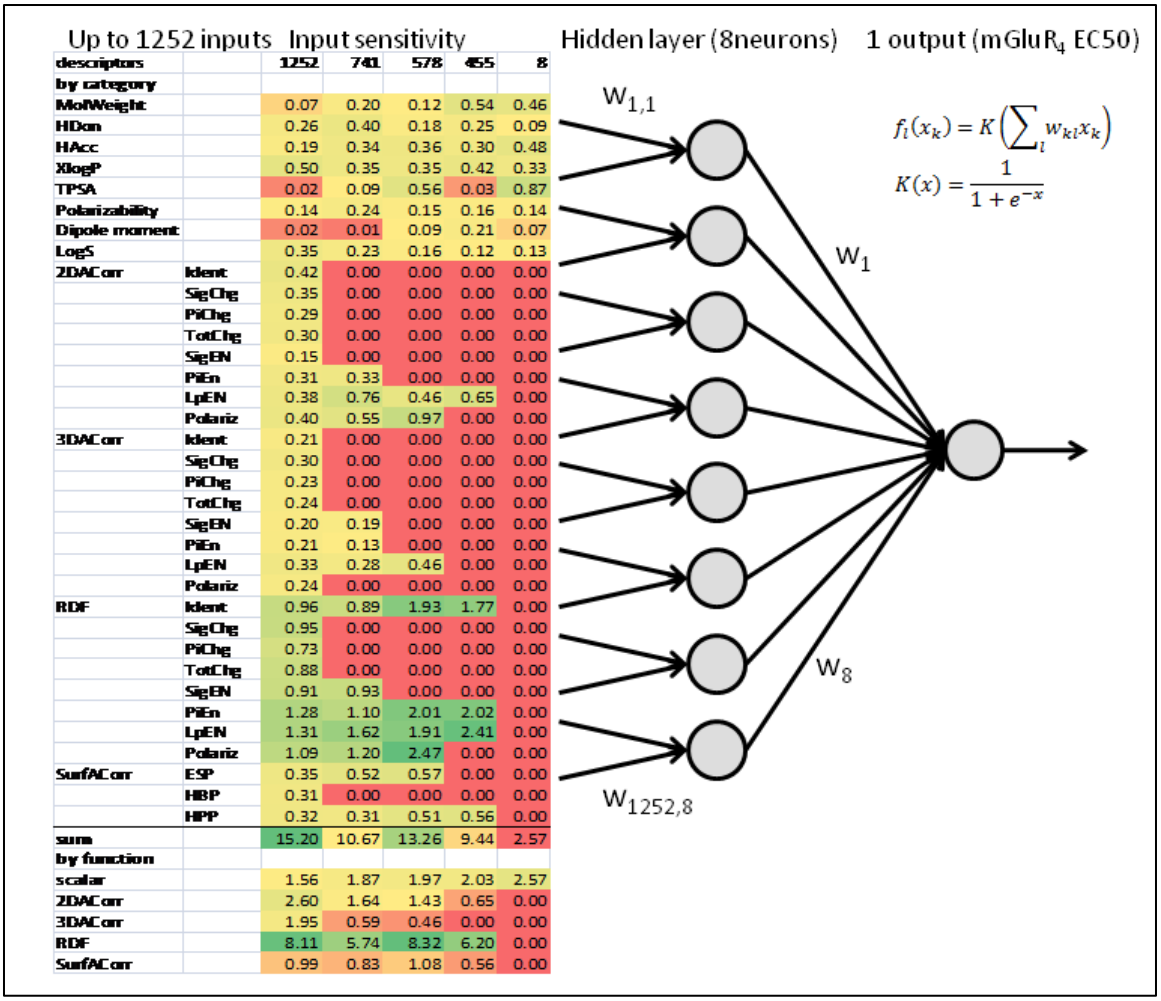


Figure 11: Receiver Operating Characteristic (ROC) curve plot for 415, 578, 741, and jury approach (round 1): The three optimized descriptor sets (415: long dotted black line, 578: dotted gray line, 741: dotted black line) perform similarly well as the jury approach (black line). However, the jury approach is more stable compared to the three other ANNs, as can be seen in Table 5.

Radial Distribution Functions (RDFs) carry most of the input sensitivity

As more descriptors are removed from the inputs, the input sensitivity values increase for RDFs (see Figure 12). Specifically, RDFs for π- and lone pair electronegativity always play an important role. RDFs for identity and polarizability are featured most prominently in the model with 578 descriptors which is the best non-jury network (see Table 5). The importance of these descriptors immediately makes sense, since the active compounds of the original high-throughput screen often feature phenyl rings and amide substructures that are well described by such RDFs.



descriptors		1252	741	578	455	8
by category						
MolWeight		0.07	0.20	0.12	0.54	0.46
HDon		0.26	0.40	0.18	0.25	0.09
HAcc		0.19	0.34	0.36	0.30	0.48
XlogP		0.50	0.35	0.35	0.42	0.33
TPSA		0.02	0.09	0.56	0.03	0.87
Polarizability		0.14	0.24	0.15	0.16	0.14
Dipole moment		0.02	0.01	0.09	0.21	0.07
LogS		0.35	0.23	0.16	0.12	0.13
2DACorr	Ident	0.42	0.00	0.00	0.00	0.00
	SigChg	0.35	0.00	0.00	0.00	0.00
	PiChg	0.29	0.00	0.00	0.00	0.00
	TotChg	0.30	0.00	0.00	0.00	0.00
	SigEN	0.15	0.00	0.00	0.00	0.00
	PiEn	0.31	0.33	0.00	0.00	0.00
	LpEN	0.38	0.76	0.46	0.65	0.00
	Polariz	0.40	0.55	0.97	0.00	0.00
3DACorr	Ident	0.21	0.00	0.00	0.00	0.00
	SigChg	0.30	0.00	0.00	0.00	0.00
	PiChg	0.23	0.00	0.00	0.00	0.00
	TotChg	0.24	0.00	0.00	0.00	0.00
	SigEN	0.20	0.19	0.00	0.00	0.00
	PiEn	0.21	0.13	0.00	0.00	0.00
	LpEN	0.33	0.28	0.46	0.00	0.00
	Polariz	0.24	0.00	0.00	0.00	0.00
RDF	Ident	0.96	0.89	1.93	1.77	0.00
	SigChg	0.95	0.00	0.00	0.00	0.00
	PiChg	0.73	0.00	0.00	0.00	0.00
	TotChg	0.88	0.00	0.00	0.00	0.00
	SigEN	0.91	0.93	0.00	0.00	0.00
	PiEn	1.28	1.10	2.01	2.02	0.00
	LpEN	1.31	1.62	1.91	2.41	0.00
	Polariz	1.09	1.20	2.47	0.00	0.00
SurfACorr	ESP	0.35	0.52	0.57	0.00	0.00
	HBP	0.31	0.00	0.00	0.00	0.00
	HPP	0.32	0.31	0.51	0.56	0.00
sum		15.20	10.67	13.26	9.44	2.57
by function						
scalar		1.56	1.87	1.97	2.03	2.57
2DACorr		2.60	1.64	1.43	0.65	0.00
3DACorr		1.95	0.59	0.46	0.00	0.00
RDF		8.11	5.74	8.32	6.20	0.00
SurfACorr		0.99	0.83	1.08	0.56	0.00

Figure 12: Schematic view of an ANN: The input to the ANN consists of up to 1,252 descriptors in 35 categories. The weighted sum of the inputs is propagated through the activation function and fed into the hidden layer (8 neurons). The output is the predicted value of the logarithm of the EC_{50} of the small molecule towards potentiation of glutamate response at the metabotropic glutamate receptor 4. The heat map shows the input sensitivity of each category from lowest (red) to highest (green).

Virtual screening of 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 ten hours on a regular personal computer. A total of 1,108 compounds with predicted EC_{50} values below 3μM for mGlu4 PAM activity were selected.

The compounds identified in the virtual screen were ordered from ChemBridge and tested at the Vanderbilt HTS facility. These compounds were screened in single point at a 10 μM (nominal) concentration using the human mGlu4/Gqi5 calcium mobilization assay as well as the rat mGlu4 thallium flux assay described in Niswender et al. Compounds that scored as 3 standard deviations higher than the control EC_{20} response population (168 compounds) were then moved to screening in concentration-response curve format in both assays. 67 compounds were confirmed as potentiators in both assays, representing an $enrichment = 67/1{,}108 \times 156{,}184/434 = 22$ relative to the initial experimental HTS hit rate. The experimental enrichment is consistent with the enrichment values predicted from analysis of an independent dataset during development of the QSAR model, given the large uncertainty of these values (Table 5).

Hits and misses of virtual screening overlap with known actives

The 67 newly identified mGlu4 PAM compounds contained eight benzo-oxazoles, 42 furan-amides including 22 thioureas, and three phenylbenzamides. All three compound classes were represented in the original HTS hits and featured only trivial R-group modifications (53 total compounds, Figure 13). Experimentally inactive compounds from these classes included 72 phenylbenzamides, 193 benzo-oxazoles, and 213 furanamides . A second round of training was attempted incorporating the results from the first round to increase the chemical diversity in the model.

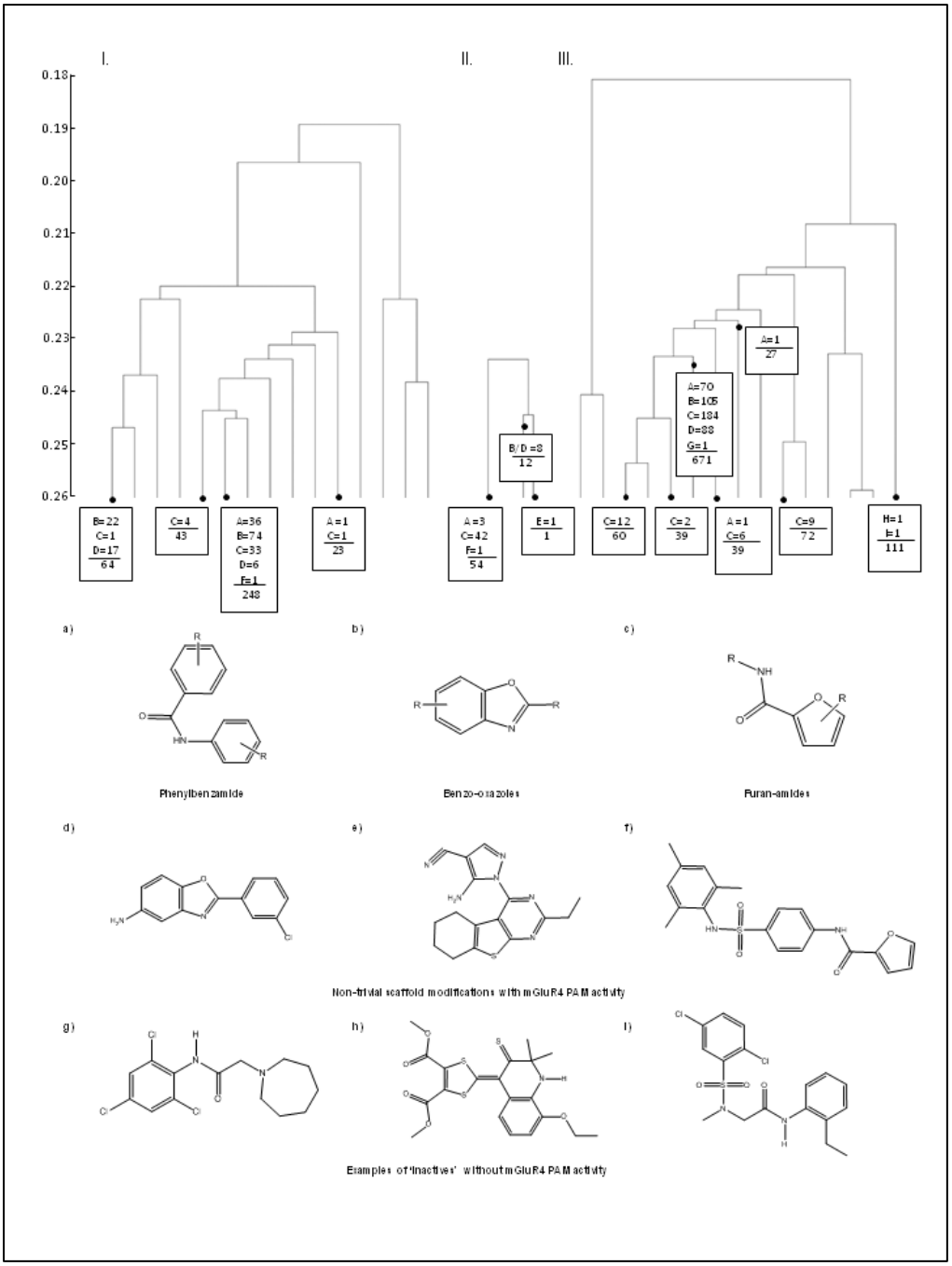


Figure 13: Scaffold category analysis (round 1): (I) Scaffold composition of 432 mGlu4 PAMs from HTS. mGlu4 PAMs were clustered with the Mathematica package using the Tanimoto coefficient of the largest common substructure as distance measure. Three major scaffolds are constituted by 37 phenylbenzamides (8.5%, a), 96 benzo-oxazoles (22.2%, b), and 39 (9.0%, c) furan-amides. (II) Scaffold composition of 67 active compounds in the postscreen. (III) Scaffold composition of inactive compounds in the postscreen. Compounds d, e, and f are examples for active compounds identified by the virtual HTS, where g, h, and i were found to be inactive.

Second round of predictions focused on scaffold-hopping

The known actives from the initial HTS combined with the 67 newly identified mGlu4 PAMs and a set of additional bromofurans comprised the set of 504 active compounds used to generate the second round model. The data set also contained 149,778 inactive compounds. Models were trained employing the same sets of descriptors and jury setup (see Figure 14 and Table 6).

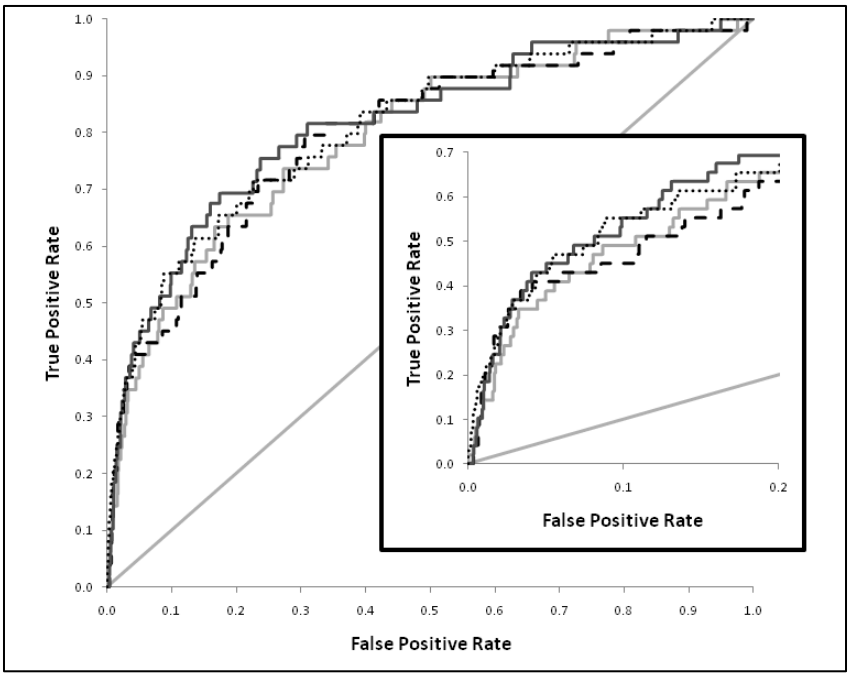


Figure 14: Receiver Operating Characteristic (ROC) curve plot for 415, 578, 741 descriptors, and jury approach (round 2): The models for 415 (gray), 578 (dashed black), 741 descriptors (black), and jury (dotted black) achieve almost identical performance. The inset shows the first 20% of the False Positive rate.

Table 6: The *rmsd*, *auc*, and *enrichment* values for all round 2 mGlu4 QSAR models

Iteration	Number and type of descriptors		*rmsd*			*auc*	*enrichment*
			train	monitor	independent		at 2%
1	741	1–8, 14–16, 21–23, 25, 29–33, 35	0.191	0.225	0.224	0.657	12.1
2	578	1–8, 15–16, 23, 25, 30 – 33, 35	0.189	0.224	0.232	0.649	15.1
3	415	1–8, 15, 25, 30–31, 35	0.186	0.231	0.236	0.647	11.1
Jury	**-**	**-**	**0.172**	**0.228**	**0.236**	**0.655**	**14.1**

However, to improve the diversity of compound predictions, a higher cutoff of EC_{50} = 20 μM was chosen and all compounds identified by any of the models 1-3 were considered. Out of a total of 36,930 compounds predicted to be active, compounds with TPSA < 130Å^2 and XlogP <4.0 were retained (leaving 17,268 molecules). Compounds containing reactive fragments were then removed leaving 12,218 compounds. Finally, compounds that had a large overlap with a known active compound (>0.65 similarity based on molecular graph) were removed leaving 2,015 compounds. Additional compounds predicted at a potency cutoff of 3 μM were then combined with this set (2,777 compounds total) and vetted for removal of any remaining promiscuous and reactive groups leading to a final order of 2,630 compounds for experimental testing.

These compounds were tested using the hmGlu4 Gqi5 assay, again in single point with 166 compounds moving to CRC testing. These studies identified 70 previously unknown mGlu4 active compounds giving an $enrichment = 70/2{,}630 \times 149{,}778/508 = 8$ relative to the initial experimental HTS hit rate. The enrichment was expected to be substantially lower than in the first round due to the higher EC_{50} cutoff of 20 μM compared to 3 μM (Table 6).

Non-trivial modifications of known actives constitute 59% of newly identified PAMs

Out of the 70 newly identified active compounds, 61 constituted mGlu4 PAMs. The set of the 61 PAMs contained four benzisoxazoles, and 21 furanyl amides, classes which were represented in the original HTS hits (Figure 15). The remaining 36 PAMs consisted of ten biaryl anilines, two aryl ketones, seven aryl ethers, three aryl thioethers, ten phenyl sulfonylamides, and four novel singletons with diverse scaffolds. Hence, 59% (36/61) of the newly identified PAMs showed non-trivial modifications of hits of the original HTS. One of the biaryl anilines (Figure 15d) is similar to compounds patented by ADDEX [100]. The biaryl compound shown in Figure 15e has not been reported in the literature thus far as an mGlu4 PAM.

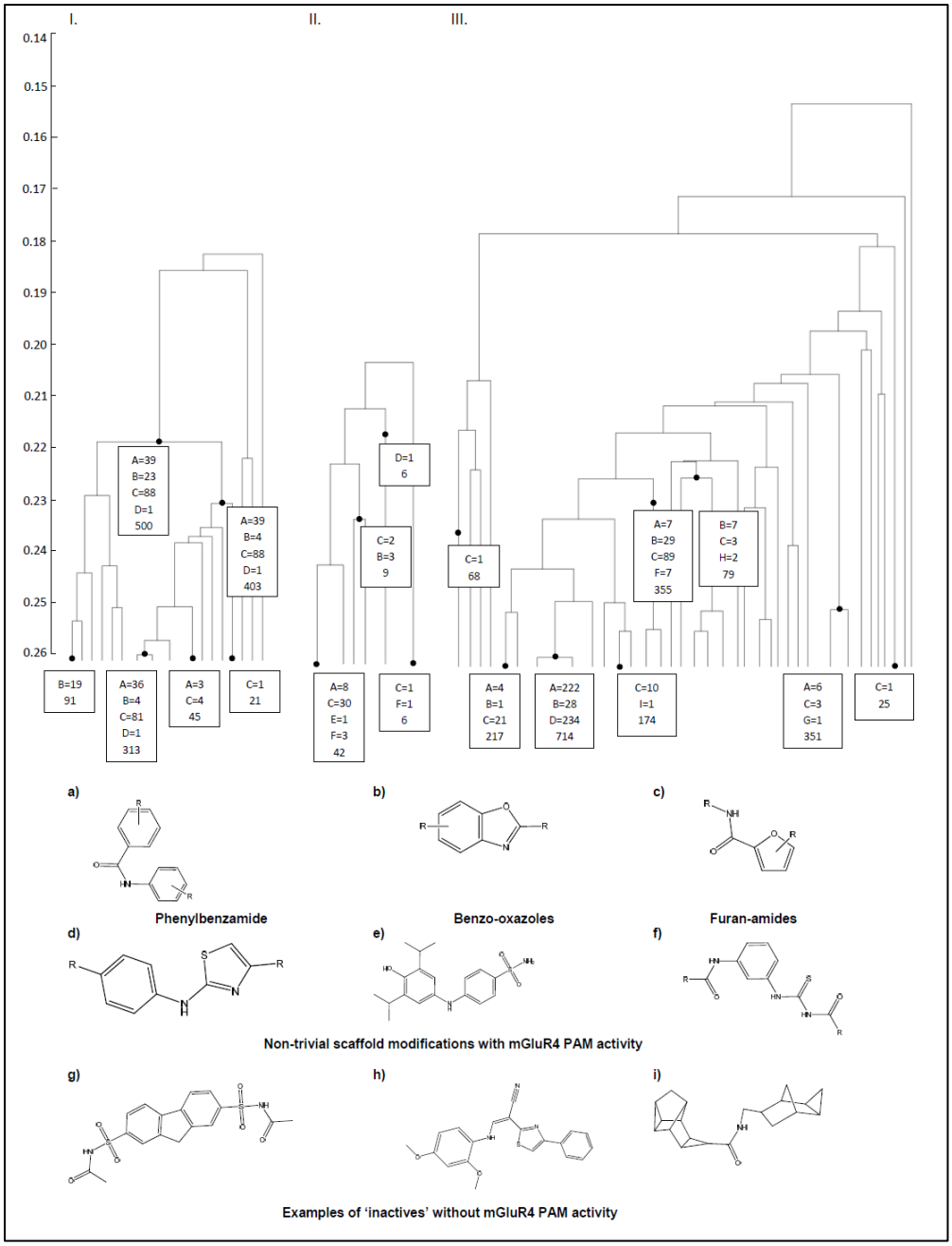


Figure 15: Scaffold category analysis (round 2): (I) Scaffold composition of 504 mGlu4 PAMs from HTS and first round vHTS. mGlu4 PAMs were clustered with the Mathematica package using the Tanimoto coefficient of the largest common substructure as distance measure. The three major scaffolds from round 1 are represented by 39 phenylbenzamides (7.7%, a), 23 benzo-oxazoles (4.6%, b), and 88 (17.5%, c) furan-amides. (II) Scaffold composition of 70 active compounds in the postscreen. (III) Scaffold composition of inactive compounds in the postscreen. Compounds d, e, and f are examples for active compounds identified by the virtual HTS, where g, h, and i were found to be inactive.

Conclusions

Artificial Neural Networks were trained to generate QSAR models from an HTS experimental dataset of compounds with mGlu4 positive allosteric modulator activity. A jury system, based on the three ANN models, generated improved enrichments when compared to each individual model. The enrichment factor of 22 determined from biological testing of 1,108 compounds prioritized from a commercial library of ~450,000 substances demonstrates the predictive power of the method. This enrichment factor agrees well with the theoretically expected enrichment of 36. A second round of virtual screening using models that had been refined with the results from the first screen identified 36 compounds with novel chemotypes that have mGlu4 potentiation activity. This approach therefore allows screening of external libraries for target-specific lead-like candidate molecules around known scaffolds but also has the potential to identify novel chemotypes.

Methods

Balancing the data by oversampling

Only 0.3% (432 molecules) of the whole data set (156,146 molecules) was active. For training the dataset was oversampled by a factor of 360 (see Figure 16)[1]. This lead to an oversampled data set with 311,234 molecules where approximately half of the data points were active and the other half inactive.

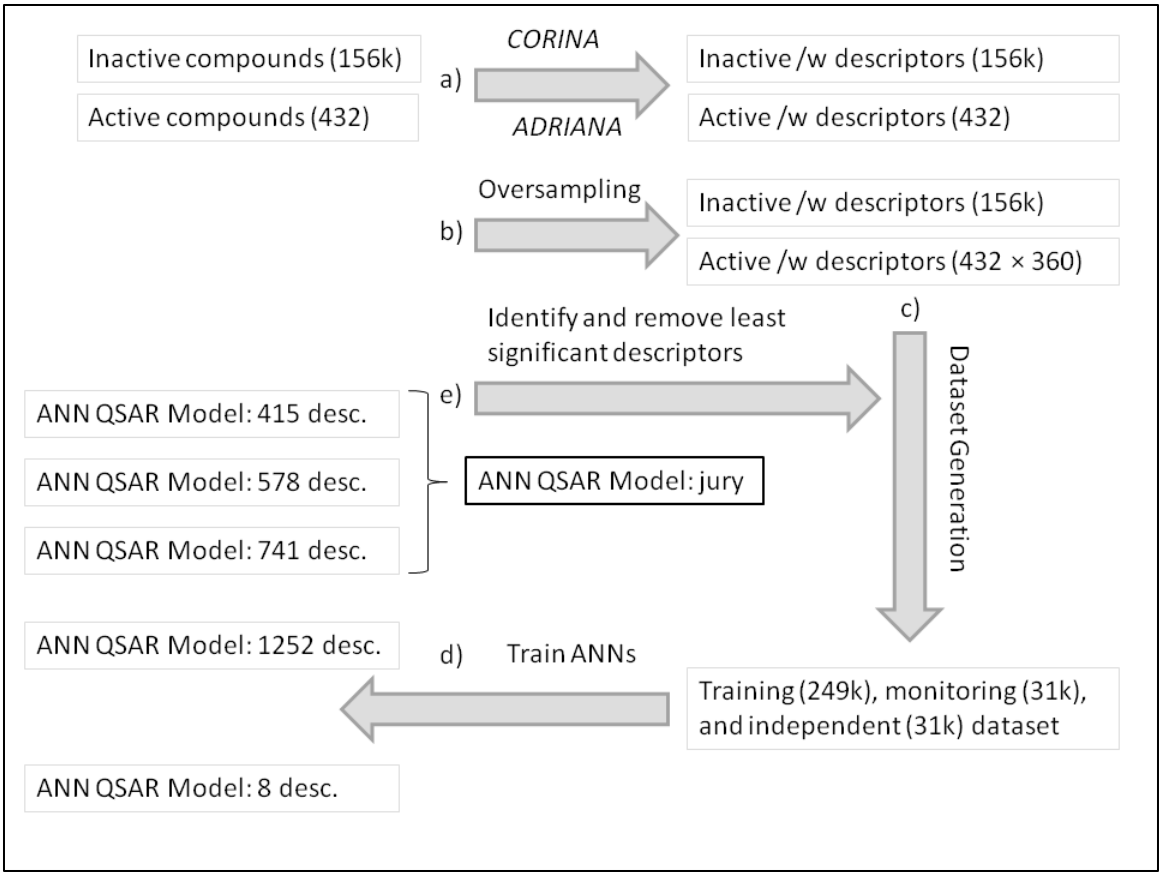


Figure 16: Overall model generation workflow: a) Colleen Niswender provided SD files with active and inactive compounds towards mGlu4 determined by HTS and CRC; CORINA and ADRIANA were employed to generate 3D structures and molecular descriptors; b) active molecules were oversampled 360 times to balance data sets; c) molecules were randomly distributed between training (80%), monitoring (10%), and independent (10%) datasets; d) ANNs were trained and e) low sensitivity descriptors were removed until the quality measures (see Table 5) no longer improved; the best three ANNs were combined into a jury network.

Translating molecular structures into input numerical descriptors

To determine the input for the ANNs, 3D models of all 156,146 molecules from the original HTS were generated using CORINA[6]. These models served as input for the ADRIANA[5] software package. All 35 categories (scalar, 2D/3D auto-correlation, RDF (eight each), surface auto-correlation (three), see Table 4) were computed implementing the default values in each category. Approximately 4% of all molecules were not properly encoded by ADRIANA and removed from the data set. The final data set consisted of 298,914 data points.

From biological data to output

The experimentally determined EC_{50} values of the active compounds ranged from 94.4nM to 17.8µM. To distinguish between active and inactive compounds, all inactive compounds were set to an arbitrary potency of 1mM. The output for training the ANN consisted of the natural logarithm of the $ln(EC_{50})$ values ranging from -16.2 (most active) over -10.9 (least active) to -6.9 (inactive). The *root mean square deviation (rmsd)* between experimental and predicted EC_{50} values was employed as objective function in training the ANNs:

Equation 8:

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

Monitoring data set determines progress and termination of training

From the 298,914 data points in the oversampled data set, 239,132 (80%) were employed in the actual training of the ANN. The monitoring data set consisted of 29,891 data points (10%). The *rmsd* between experimental and predicted $ln(EC_{50})$ was computed for the monitoring data set after each iteration over the full training data set. Once the *rmsd* stabilized, the training was terminated, and the *rmsd* of the remaining 10% (independent data set) computed (see Table 5). Care was taken to avoid overlap between training, monitor, and independent data set.

A three-layered ANN was trained implementing Resilient Propagation

The trained ANNs consisted of the input layer with up to 1,252 chemical descriptors, the hidden layer consisting of eight neurons, and one neuron in the output layer predicting the $ln(EC_{50})$ of the described molecule. The sigmoid function

Equation 9:

$$S(x) = \frac{1}{1 + e^{-x}}$$

served as activation function of the neurons. The ANNs were trained by implementing resilient back-propagation of errors[10b], a supervised learning approach. The training was terminated after up to 40,000 iterations when the monitoring dataset achieved its minimum *rmsd*. It took up to 13 hours per network using eight cores of a core2 quad 2.33GHz Intel Xeon microprocessor in parallel on the 64-bit version of Red Hat Enterprise Linux 5.2.

Jury system combines output of three best networks

The outputs of the three best ANNs were used as input for a jury ANN that consisted of three inputs, four hidden neurons, and one output. The training of the jury ANN terminated after 290 steps.

Selection of the optimal set of descriptors of chemical structure

It is crucial to select the optimal set of descriptors from the 35 available categories. In a top-down approach, the least significant categories for predicting $ln(EC_{50})$ were successively removed to increase the predictive power of the according ANNs. The advantage lies in removing degrees of freedom from the ANN by reducing the number of inputs. Since the number of data points stays the same, the signal-to-noise ratio improves. This procedure is described in detail elsewhere[1].

The ANN can be thought of as a multidimensional function:

Equation 10:

$$y = f\left(x_1, x_2, \ldots, x_{N_0}\right) = f(\langle x \rangle)$$

with input values $x_1, x_2, \ldots, x_{N_0}$ and output y. The partial derivative of each input with respect to the output can be determined numerically and is introduced as "input sensitivity":

Equation 11:

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

For this purpose each input value x_k is altered by a small Δx_k = ±0.05 * x_k in an independent experiment and the change Δy is monitored[94]. Following this procedure the input sensitivity is determined for each input k by selecting 100 random compounds from the independent dataset. The input sensitivity of input k is the average ratio observed (Equation 11).

The input sensitivity of each of the 27 non-scalar descriptor categories was determined as norm over the individual input sensitivity values within this category. The descriptor categories were sorted by input sensitivity. In each step, categories comprising the least 10% of input sensitivity were removed. This process was repeated until the quality measures did no longer improve (see Table 5 and Table 6).

Enrichment and area under the curve complement rmsd as quality measures

Analysis of the *rmsd* proved to be a poor indicator for model quality (see Table 5 and Table 6). Hence, all models were also assessed in terms of their 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 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 binary classifier (see Figure 9, Figure 11, and Figure 14). 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 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. 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). 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:

Equation 12:

$$enrichment = \frac{TP}{TP + FP} \Big/ \frac{P}{P + N}$$

When computed for the independent dataset 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 dataset (here 0.28%). Note that enrichment values are always coupled to a certain cutoff, the fraction of molecules retained after filtering. The enrichments reported in Table 5 and Table 6 were determined for a cutoff of 2%. As an example, this would correspond to filtering 9,000 compounds out of a library of about 450,000.

Implementation

The ANN algorithm was implemented in the BioChemistryLibrary (BCL). The training method used is Resilient Propagation, a supervised learning approach[10]. Further detail is given above. The BCL is an in house developed, object-oriented library written in the C++ programming language. It currently consists of approximately 400 classes and 300,000 lines of code. ADRIANA[5] was used for generation of chemical descriptors. CORINA[6] was used for generation of three-dimensional structures.

CHAPTER IV

IDENTIFICATION OF METABOTROPIC GLUTAMATE RECEPTOR SUBTYPE 5 NEGATIVE ALLOSTERIC MODULATORS USING VIRTUAL HIGH-THROUGHPUT SCREENING

Introduction

Artificial Neural Networks (ANNs) were trained to predict the capability of drug-like molecules for allosteric inhibition of the metabotropic glutamate receptor subtype 5 (mGlu5) based on a high-throughput screen of 345 confirmed negative allosteric modulators (NAMs) and 155,774 compounds showing no activity towards mGlu5. Commercially available databases of small molecules were virtually screened for novel inhibitors of mGlu5. Hit compounds were verified experimentally. For an overview and introduction to allosteric modulation of mGlu5 see Gasparini et al[12b].

Biology and pharmacology of metabotropic glutamate receptors

Pin et al[47b] gave a review of the metabotropic glutamate receptors in 1995. They discuss the different subtypes (mGlu1 to mGlu8) and their organization into three groups based on cloning results, furthermore transduction mechanisms for the cloned receptors, and pharmacology including certain ligands. In the chapter on structure and function they describe the seven transmembrane helices, the glutamate binding site, and the G-protein coupling. Multiple alignments of the mGlus and the closely related PCaR1 show conserved residues. Residues affecting glutamate affinity are marked. The last two chapters discuss transduction mechanisms of native mGlus and their physiological roles.

This work was updated and expanded by Conn and Pin[11] in 1997. Potential clinical uses are mentioned as well as the role of mGlus as presynaptic autoreceptors and in regulating ion channels.

After updating this information in Pin et al (1999)[15], the authors move on to discuss the development of selective ligands for each group of mGlus. Pharmacophore models for the glutamate binding site are reported. The remainder of the paper highlights additional regulatory sites on the mGlus which could be utilized for allosteric modulation.

Negative allosteric modulation of mGlu5 could allow treatment of fragile X syndrome

Dölen and Bear[14] give a review of studies connecting negative allosteric modulation to the treatment of fragile X syndrome which can cause mental retardation and autism. Fragile X syndrome is triggered by mutational deactivation of the fragile X mental retardation protein (FMRP) which inhibits protein synthesis. It is therefore thought of as the natural opponent of mGlu5 regulating translation of mRNA at the synapse. Negative allosteric modulation of mGlu5 could reestablish this balance between mGlu5 potentiation and FMRP inhibition of protein synthesis, therefore ameliorating symptoms of fragile X syndrome.

For a general overview of the therapeutic potential of allosteric modulation of G-protein coupled receptors see Conn, Christopoulos, and Lindsley[16].

Negative allosteric modulators of metabotropic glutamate receptor subtype 5

The first potent and selective negative allosteric modulator 2-Methyl-6-(phenylethynyl)-pyridine (MPEP) was reported by Gasparini et al[31] in 1999. Several papers were published analyzing the binding pocket of MPEP[25, 101], its augmentation of PCP-induced cognitive deficits[55a, b], and the development of MPEP into a radioligand[33].

Rodriguez et al[35] show that an MPEP-like ligand could block the effect of multiple allosteric modulators and therefore function as a neutral allosteric ligand. In 2009 the same group reported the discovery and structure-activity relationship of mGlu5 antagonists different from the MPEP scaffold[36].

High-Throughput Screening in Drug Discovery

High-throughput screening is the process where a large (several thousand to millions) library of small molecule ligands is screened in an automated fashion for a beneficial property in an assay. The 120 GPCR-based HTS assays published in PubChem (pubchem.ncbi.nlm.nih.gov, accessed April 2011) address targets like RGS16-$G_{\alpha o}$ (AID1441, primary screen, 826 active out of 218,535 tested) and 5-hydroxytryptamine (serotonin) receptor subtype 1a (5HT1a) (AID567, primary screen, 366/64,907). However, the hit rate in these examples is always around 0.5%, meaning 99.5% of the screened compounds are inactive towards the target. Increasing the hit rate to 5% through virtual high-throughput screening would enrich the screen by a factor of 10, meaning only 10% of the molecules need to be tested to get the same number of active compounds and henceforth reducing screening cost and time.

Rodriguez et al[3] screened approximately 160,000 small molecules to identify modulators of mGlu5. The primary triple-add calcium flux assay revealed 624 potential antagonists. In the confirmatory screen employing full concentration response curves, 345 antagonists were verified. This HTS data is the basis for the virtual HTS experiment described in this chapter. Additional experiments determining functional diversity and in vivo activity of the novel allosteric modulators were reported.

Virtual High-Throughput Screening

Virtual HTS allows searching for modulators of a given biological target by means of a quantitative structure activity relationship. A model is trained on the structure of small molecules with known activity towards the target. Under the assumption that structurally similar targets have similar activity, a database of virtual or commercially available small molecules is screened for compounds with similar structural features resulting in similar activity.

Noeske et al[19] utilized virtual HTS to search for mGlu1 NAMs. The top five compounds of similarity searches on the Asinex Gold Collection (194,563 compounds in 2003) with six known inhibitors of mGlu1were ordered together with eight compounds that were identified in at least three of the six runs regardless of rank. Among the 23 compounds that were delivered one had sub-μM activity, five were between 1μM and 15μM, and the remaining 17 compounds were above 15μM. The same group expanded this approach[20] by training a self-organizing map (SOM) on the COBRA database of pharmacologically relevant compounds. Therefore they described the structure of each molecule by CATS-2D topological atom-pair descriptors which describe the molecule by binning bond distances between chemically relevant groups of atoms. Known inhibitors of mGlu1 were clustered near two neurons in the SOM. Molecules in the Asinex Gold Collection were projected onto the SOM. Compounds in the vicinity of these two neurons were manually inspected and 28 screening candidates ordered. One molecule showed activity below 1μM, five between 1 and 15μM in an rmGlu1 assay.

Describing Chemical Structure for QSAR

Recently, the focus for describing chemical structure for QSARs has been more on fragment-independent descriptors because of the higher flexibility in describing small molecules. Other examples besides the aforementioned CATS-2D topological descriptors are radial basis and surface auto-correlation functions[5].

Establishing QSAR in Drug Discovery

The importance and examples of QSAR in drug discovery are discussed in detail in the introductions of chapters II and III. For an overview on QSAR methods see Hansch et al[65a].

Machine learning and QSAR

Winkler et al[7] reviews how neural networks can be employed in establishing QSARs. Furthermore, it provides a general introduction to neural networks and their training.

Machine learning models were trained on locally available high-throughput screening data to predict biological activity of small molecule ligands towards negative allosteric modulation of metabotropic glutamate receptor subtype 5. A set of 749 compounds was ordered and tested in a triple-add calcium flux assay. Negative allosteric modulators were enriched by a factor of seven and a new scaffold of NAMs was identified.

Results and Discussion

Optimization of molecular descriptor set improves prediction results

The ANN trained on all 35 descriptor categories available in the ADRIANA descriptor set (see Table 7) achieved an *root mean square deviation (rmsd)* of the independent data set of 0.209, *area under the curve (auc)* of 0.83, and *enrichment* of 18.8. For the first optimization the four least sensitive descriptor categories were removed leading to an *rmsd* of 0.199, *auc* of 0.87, and *enrichment* of 28.2. Removing eight more descriptor categories led to the optimal model with 763 descriptors and to an *rmsd* of 0.201, *auc* of 0.86, and *enrichment* of 37.6. Further reduction of the number of descriptors led to slightly worse models which were not considered for further predictions (see Figure 17).

Enrichment is the critical quality measure for virtual screening

While *rmsd* and *auc* identified the general quality of the trained ANNs, these quality measures captured the overall quality of the models including prediction of inactive compounds. However, the success of a model is based on identification of active compounds for ordering. So the emphasis of the quality measures was placed on the *enrichment* that predicts the ratio of correctly identified NAMs in the final order.

Table 7: 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		σ atom charges	2DA_SigChg	11
11		π atom charges	2DA_PiChg	11
12		total charges	2DA_TotChg	11
13		σ atom electronegativities	2DA_SigEN	11
14		π 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		σ atom charges	3DA_SigChg	12
19		π atom charges	3DA_PiChg	12
20		total charges	3DA_TotChg	12
21		σ atom electronegativities	3DA_SigEN	12
22		π 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		σ atom charges	RDF_SigChg	128
27		π atom charges	RDF_PiChg	128
28		total charges	RDF_TotChg	128
29		σ atom electronegativities	RDF_SigEN	128
30		π 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

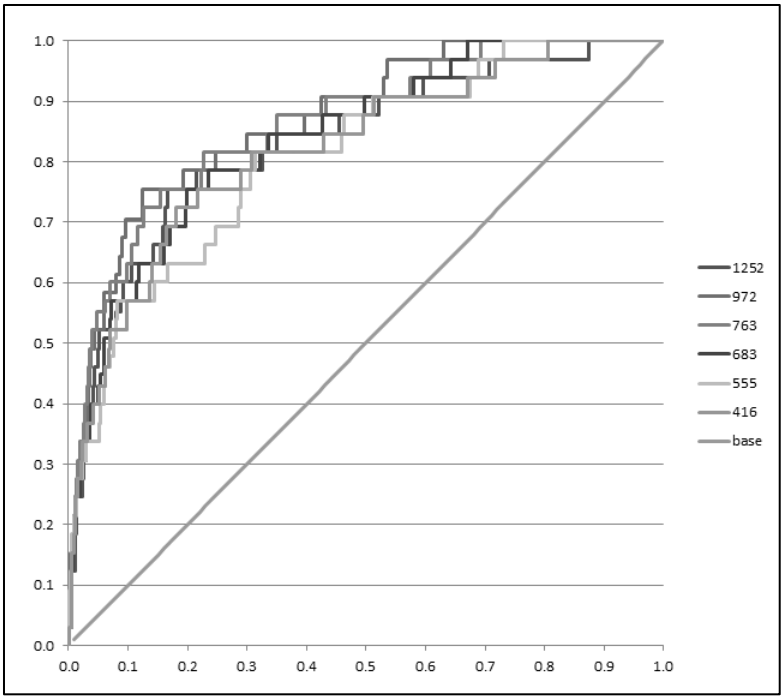


Figure 17: Receiver Operating Characteristic (ROC) curve plot for 416, 555, 683, 763, 972, and all descriptors: While the models with 972 and 763 descriptors perform well over the whole ROC curve, the other models clearly show reduced performance in the middle part of the ROC curve. However, in the beginning of the curve all of the models look very similar. Based on the enrichment (see Table 8) the model with 763 descriptors was chosen for predicting active compounds to be ordered.

Table 8: The *rmsd*, *auc*, and *enrichment* values for all mGlu5 NAMs QSAR models

Iteration	Number and type of descriptors		*rmsd*			*auc*	*enrichment*
			train	monitor	independent		(at 0.3%)
All	1252	1-35	0.184	0.203	0.209	0.83	18.8
1	972	1–19, 21–26, 29–32, 34– 35	0.168	0.202	0.199	0.87	28.2
2	**763**	**1–9, 11, 12, 14, 15, 17, 21–23, 25, 29 – 32, 35**	**0.157**	**0.201**	**0.201**	**0.86**	**37.6**
3	683	1–8, 14, 23, 25, 29 – 32, 35	0.178	0.204	0.210	0.84	9.4
4	555	1–8, 14, 23, 25, 29 – 31, 35	0.189	0.204	0.218	0.81	28.2
5	416	1–8, 23, 25, 30, 31, 35	0.180	0.210	0.215	0.82	9.4

Predictions focused on scaffold-hopping

The ANN with the highest enrichment for mGlu5 NAMs (Table 8, iteration 2) was employed to virtually screen the ChemDiv (San Diego, CA) Discovery Chemistry database of 708,416 (May 2008) commercially available drug-like compounds. Only compounds which had a Tanimoto coefficient based on size of substructure of less than 0.6 compared to locally known mGlu5 NAMs were considered for ordering. This emphasized identifying new scaffolds with low similarity to known inhibitors.

Results of virtual screening of ChemDiv compound library

The model predicted a set of 42,041 small molecules at a 10μM potency cutoff from the virtual screen. Molecules with a weight above 600Da, 130$Å^2$ total polarizable surface area, cLogP of 4.0, or labile or reactive fragments were sorted out. Removal of molecules with a substructure similarity above 0.6 led to the final order of 749 compounds. These compounds were subsequently tested at the Vanderbilt Institute for Chemical Biology high-throughput screening center to reveal a single point concentration hit rate of 12% (88/749 compounds) that included 51 antagonists (NAMs), 18 positive allosteric modulators (PAMs) and 19 agonists. Concentration response curves at 10 concentrations (1nM to 30μM range) confirmed 12 NAMs, 14 PAMs and 1 partial agonist compound (3.6% hitrate) representing an enrichment factor of 15.7 for mGlu5 activity compared with the original mGlu5 experimental screening data (0.22% hit rate). Compounds with confirmed antagonist (NAM) activity (1.6%) were enriched by a factor of 7.0 with novel mGlu5 NAM scaffolds.

Virtual High-Throughput Screening identified new scaffold of mGlu5 NAMs

Two mGlu5 NAMs with an IC_{50} of 75 and 124nM, respectively, were identified resembling a formerly unknown scaffold of mGlu5 inhibitors.

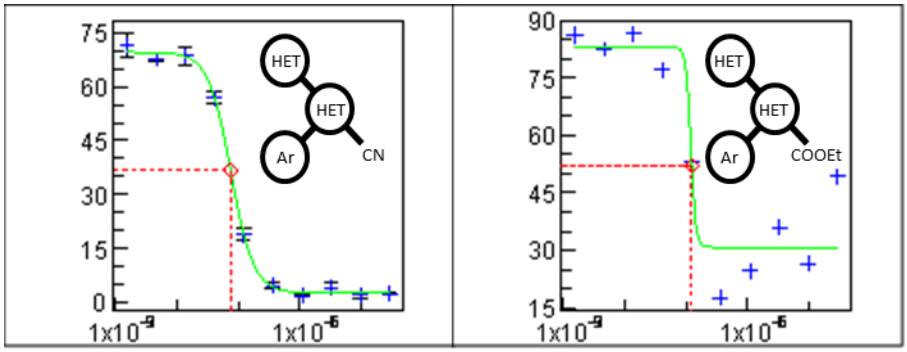


Figure 18: Two novel mGlu5 NAMs sharing a previously unknown scaffold. Two of the 51 NAMs identified in an order of 749 compounds. These compounds were confirmed in concentration response curves revealing activities of 75 (left) and 124nM (right).

Conclusions

The latest virtual HTS experiment reported here had the lowest ratio of active compounds in the original screen (0.22%). Nevertheless, it was possible to train stable machine learning QSARs on this sparse data. The enrichment of 7.0 for mGlu5 NAMs was lower compared to the enrichment for mGlu5 PAMs. This was expected because of the emphasis on scaffold hopping in generating the compound set to order. These efforts proved finally successful by identifying two mGlu5 NAMs with activities around 100nM sharing a novel scaffold compared to known mGlu5 NAMs.

Methods

Translating molecular structures into input numerical descriptors

To determine the input for the ANNs, low energy conformations of all 156,146 molecules from the original HTS were generated using CORINA. These models served as input for the ADRIANA software package. All 35 categories (scalar, 2D/3D auto-correlation, RDF (eight each), surface auto-correlation (three), see Table 7) were computed implementing the default values in each category. Eighteen (5.2%) of the 345 known inhibitors were not properly encoded by ADRIANA and removed from the data set leading to a set of 327 properly encoded inhibitors.

ADRIANA failed to encode a similar rate of 7,291out of 155,774 (4.7%) inactive compounds leaving 148,483 correctly encoded inactive compounds.

Balancing the data by oversampling

To keep the ANN from predicting all compounds as being inactive, the training data needs to be balanced. Each of the 327 properly encoded inhibitors was represented 451 times in the training data set (147,477 entries compared to 148,483 inactive compounds for a total of 296,936 entries).

From biological data to output

The experimentally determined EC_{50} values of the active compounds ranged from 4.8nM to 20μM. To distinguish between active and inactive compounds, all inactive compounds were set to an arbitrary potency of 1 mM. The output for training the ANN consisted of the natural logarithm of the $ln(EC_{50})$ values ranging from -19.2 (most active) over -10.8 (least active) to -6.9 (inactive). The *root mean square deviation (rmsd)* between experimental and predicted EC_{50} values was employed as objective function in training the ANNs:

Equation 13:

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

Monitoring data set determines progress and termination of training

From the 296,936 data points in the oversampled data set, 237,550 (80%) were employed in the actual training of the ANN. The monitoring data set consisted of 29,693 data points (10%). The *rmsd* between experimental and predicted $ln(EC_{50})$ was computed for the monitoring data set after each iteration over the full training data set. Once the *rmsd* stabilized, the training was terminated, and the *rmsd* of the remaining 10% (independent data set) computed (see Table 8). Care was taken to avoid overlap between training, monitor, and independent data set.

A three-layered ANN was trained implementing Resilient Propagation

The trained ANNs consisted of the input layer with up to 1,252 chemical descriptors, the hidden layer consisting of eight neurons, and one neuron in the output layer predicting the $ln(EC_{50})$ of the described molecule. The sigmoid function

Equation 14:

$$S(x) = \frac{1}{1 + e^{-x}}$$

served as activation function of the neurons. The ANNs were trained by implementing resilient back-propagation of errors[10], a supervised learning approach. The training took up to 40,000 iterations of Resilient Propagation. However, training was terminated early when the monitoring dataset achieved its minimum *rmsd*. The training took up to 13 hours per network using eight cores of a core2 quad 2.33GHz 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

It is crucial to select the optimal set of descriptors from the 35 available categories. In a top-down approach, the least significant categories for predicting $ln(EC_{50})$ were successively removed to increase the predictive power of the according ANNs. The advantage lies in removing degrees of freedom from the ANN by reducing the number of inputs. Since the number of data points stays the same, the signal-to-noise ratio improves. This procedure is described in detail elsewhere[1].

The ANN can be thought of as a multidimensional function:

Equation 15:

$$y = f\left(x_1, x_2, \dots, x_{N_0}\right) = f(\langle x \rangle)$$

with input values $x_1, x_2, \dots, x_{N_0}$ and output y. The partial derivative of each input with respect to the output can be determined numerically and is introduced as input sensitivity:

Equation 16:

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

For this purpose each input value x_k is altered by a small Δx_k = ±0.05 * x_k in an independent experiment and the change Δy is monitored[94]. Following this procedure the input sensitivity is determined for each input k by selecting 100 random compounds from the independent dataset. The input sensitivity of input k is the average ratio observed (Equation 16).

The input sensitivity of each of the 27 non-scalar descriptor categories was determined as norm over the individual input sensitivity values within this category. The descriptor categories were sorted by input sensitivity. In each step, categories comprising the least 10% of input sensitivity were removed. This process was repeated until the quality measures did no longer improve (see Table 8).

Enrichment and area under the curve as binary quality measures

The $enrichment$ measures the quotient of the rate of active compounds (true positives over all predicted active) in the predicted hits of the virtual HTS over the rate of actives (positives over all compounds) in the original HTS. It depends on the cutoff of compounds in the virtual HTS which here was set to 0.3% of all compounds. This is lower than the actual cutoff utilized for driving the

scaffold hopping. Therefore, the actual enrichment of seven was lower than the theoretical enrichments reported in Table 8.

Equation 17:

$$enrichment = \frac{TP}{TP + FP} \Big/ \frac{P}{P + N}$$

The $area\ under\ the\ ROC\ curve$ measures the overall quality of the model. It is a real number between (usually) 0.5 and 1 with 0.5 representing random prediction of activity and 1 describing perfect prediction.

Both measures are based on the correct binary prediction of a compound being active or not in contrast to a non-discrete measure like $rmsd$.

Implementation

The ANN algorithm was implemented in the BioChemistryLibrary (BCL). The training method used is Resilient Propagation, a supervised learning approach[10]. Further detail is given above. The BCL is an in house developed, object-oriented library written in the C++ programming language. It currently consists of approximately 400 classes and 300,000 lines of code. ADRIANA was used for generation of chemical descriptors[5]. CORINA was used for generation of three-dimensional structures[6].

CHAPTER V

PREDICTING CARBON CHEMICAL SHIFTS EMPLOYING MACHINE-LEARNING METHODS

Introduction

^{13}C NMR spectroscopy is a powerful tool in structure elucidation of natural products and product validation in organic chemistry. In particular, the comparison of experimentally determined NMR chemical shifts with predicted chemical shifts can provide critical information to determine the constitution of an unknown compound[102]. Similarly, in synthetic organic chemistry, comparison of the ^{13}C NMR spectrum of the synthesized compound with a predicted spectrum can reveal the success of the synthesis.

Widely used approaches for predicting chemical shifts include *ab initio* calculations, empirical methods (e.g., lookup of similar compounds in databases, incremental correction systems), machine learning approaches such as artificial neural networks, and combinations thereof.

Ab initio methods are computationally expensive and lack accuracy

Ab initio calculations determine the chemical shifts by computing magnetic properties directly from a given conformation of the substance. Recently, Mulholland et al. utilized the Logic for Structure Determination (LSD) program to suggest a composition for the natural product angelon [103]. The LSD program allows the scientist to input constraints derived from NMR data and from this data, suggests an ensemble of structures that fulfill these constraints. The correctness of the solution was demonstrated by comparing the experimental ^{13}C chemical shifts with values calculated from the predicted structure by means of GAUSSIAN[104], a quantum chemistry software program. The *rmsd* between the nine experimental and calculated ^{13}C chemical shifts is 2.71ppm. Another example is given by Bagno[105], who summarizes computational work using

GAUSSIAN 03 and the Amsterdam Density Functional (ADF) suite to determine chemical shifts for a number of different chemical elements. Perez et al.[106] compare the experimental values of two chloropyrimidines with four carbon atoms each to the ACD database approach (see below, *rmsd* 3.75ppm) and different *ab initio* methods, e.g. HF(bs1)s1//HF(bs1) (*rmsd* 4.53ppm). For a general overview of *ab initio* methods to predict ^{13}C chemical shifts see Cimino et al.[107]. The authors report the *corrected mean absolute error (cmae)* for about 50 different chemical shift calculation setups ranging from 1.49 to 3.35ppm.

Database methods rely on large sets of stored spectra.

In contrast to *ab initio* methods, empirical approaches derive chemical shifts not from first principles but rely on large data sets of known compounds with assigned chemical shifts. Database approaches employ similarity searches, while incremental methods and machine learning approaches derive rules to compute the chemical shifts from similar substances.

The capacities of modern computer systems make it feasible to collect large numbers of chemical structures together with assigned ^{13}C NMR chemical shifts. Examples include the BIORAD KnowItAll database which is based on the CSEARCH database (~4,000,000 ^{13}C chemical shifts), MODGRAPH NMRPredict (~3,500,000 ^{13}C chemical shifts), SpecInfo database (~1,500,000 ^{13}C chemical shifts), ACD/CNMR (~2,160,000 ^{13}C chemical shifts), NMRShiftDB (~200,000 ^{13}C chemical shifts, www.nmrshiftdb.org), and the Spectral Database for Organic Compounds (SDBS, ~ 130,000 ^{13}C chemical shifts).

The prediction of the ^{13}C chemical shifts of a compound based on such a database usually happens one atom at a time. For each carbon atom the database is searched for structures that contain carbon atoms in similar chemical settings with known chemical shifts. The most difficult challenge is to encode the chemical environment of an atom in a way that can be easily searched. Examples are the Hierarchically Ordered Spherical description of Environment code (HOSE) or

the Simplified Molecular Input Line Entry System (SMILES) code. Satoh et al.[44] describe the CAnonical representation of STereochemistry code (CAST) which explicitly accounts for the stereochemistry of the atom .

However, database approaches have certain drawbacks: The storage space required increases linearly with the number of molecules represented. The access-time increases logarithmic with the size of the database. Further, use of these methods is problematic for novel compound classes of e.g. natural products that are not well represented in the database. In Perdue et al.[108] the commercial ACD software gives an *rmsd* of 1.52ppm for a set of 29 differently substituted aromatic natural compounds. In Meiler[109] a HOSE code prediction on the SpecInfo database reaches an *rmsd* of 2.60ppm for 100 compounds that were previously not in the database.

Incremental methods are fast but lack accuracy

Incremental methods determine the influence of substituents or structural elements like rings or double bonds on the chemical shift of an atom. Usually the increments are determined from databases of chemical shifts using multiple linear regressions (MLR). Assuming that all these influences are uncoupled, all terms are summed to estimate the chemical shift of a given carbon atom. This simple mathematical model allows for rapid computation. Current implementations of this method work well for many classes of organic substances. However, the deviations between predicted shift and experimental value increase for highly substituted and sterically restrained compounds. The interaction between two or more of these structural elements makes it necessary to introduce cross-correlation terms into the model. Due to the large number of potential cross-correlation terms, such an approach becomes quickly intractable.

Cheeseman et al.[104] compared GAUSSIAN and ChemNMR Pro predictions for Taxol, an anti-cancer drug. The *mean absolute errors* in this case are 3.80ppm for ChemNMR and 4.20ppm for GAUSSIAN. Perdue et al.[108] developed an incremental approach (SPARIA) for distinguishing

substitution patterns in natural aromatic compounds. In this case, eight different substituents in five different positions around an aromatic ring lead to 16,640 unique substitution patterns. The *rmsd* of 2.27ppm is slightly worse than the *rmsd* of 1.52ppm for the ACD database approach on the same compounds discussed above.

Artificial neural networks combine speed and accuracy

Machine learning approaches such as artificial neural networks (ANNs) can combine the advantages of database and incremental approaches due to their inherent capability to model correlations between different substituents. This allows the ANN to reach a higher accuracy than incremental methods without requiring storage of a large database of chemical shifts as the chemical shift databases are only used in the training step of the ANN. The ANN can be re-trained as new databases become available. For an overview on applications of ANNs in chemistry and more specifically in NMR see[110]. The accuracy level of ANNs can reach that of database approaches[109] (*rmsd* 2.7ppm ANN vs. 2.6ppm SpecInfo for a set of 100 newly formed compounds). ANNs were used to predict chemical shifts for several classes of compounds, like acrylonitrile copolymers[111] (*rmsd* 1.47ppm) or trisaccharides[112] (*rmsd* 1.30-1.85ppm). Other implementations cover the complete space of small organic molecules, like [109] (*rmsd* 2.1ppm) and [113] (*rmsd* 4ppm). Blinov[114] trained an ANN on the aforementioned ACD/Labs database and tested it by predicting ^{13}C shifts for the NMRShiftDB (*rmsd* 1.59ppm).

Atom environment code by sphere

An important precondition to train and test ANNs is the encoding of the chemical environment of each carbon atom as a vector of numerical input values. Meiler previously introduced a modification of the aforementioned HOSE code to translate chemical information into input for an ANN[46].

This chapter describes the introduction of a novel encoding scheme for the constitution of a small molecule and training of an ANN to predict carbon chemical shifts. This new encoding scheme widens the prediction capabilities of the ANN by including a larger variety of atom types, in particular charged atoms and solvent dependencies. Furthermore, through the use of exclusively publicly available databases for training the ANN, it is possible to provide a free-access tool (bcl::shift) to the scientific community. To address the smaller size of publicly available carbon chemical shift databases (~10% of commercial databases), the substituents around the atom of interest were summed up in spheres based on the bond distance to the atom of interest reducing the number of inputs for the ANN. This ANN was expanded in a second experiment to include stereochemical descriptors based on calculated partial charges.

The flexible encoding scheme implemented in the BCL allows for the easy incorporation of new atom types to describe molecules with novel features. The ANN allows the prediction of ^{13}C chemical shifts for molecules containing C, H, N, O, S, F, Cl, Br, I, P, B, Se, and Si.

Iterative Partial equalization of orbital electronegativity determines σ-charges

Gasteiger and Marsili[115] described a method for determining the distribution of σ-charge in a small molecule. It is based on iterative partial equalization of orbital electronegativity (PEOE). First, all atoms in the molecule get assigned a base electronegativity according to their element type and hybridization and a σ-charge of zero. Based on their electronegativity, connected atoms will exchange partial charge. This in turn changes the electronegativity which is described by a quadratic polynomial fitted through orbital electronegativity data by Hinze and Jaffe[116]. Electronegativity and σ-charges stabilize after six iterations to give the final σ-charges.

Gasteiger and Marsili[117] used PEOE in 1981 to predict proton chemical shifts from the charge at the corresponding hydrogen atom. They showed a linear correlation between charge and shift of

the hydrogen. However, no linear correlation could be found between the proton chemical shift and the charge at the heavy atom (mostly carbon) the hydrogen is connected to.

Hueckel molecular orbital method determines π-charges of conjugated systems

The π-charges of conjugated systems can be determined by the Hueckel molecular orbital (HMO) method (see for instance[118]). It minimizes the energy of the conjugated system taking only the energy of an electron in the field of the atom core (Coulomb integral) and interactions between neighboring atoms (exchange integral).

Marsili and Gasteiger[119] provide a parameterization to describe π-electronegativity as a quadratic function of the π-charge in a similar fashion as described earlier for σ-charges and σ-electronegativity. They also suggest including σ-charges into the Coulomb integral of the HMO to distinguish different substituents which are not parts of the conjugated system themselves.

Results and Discussion

A combination of resilient and simple back-propagation is used to achieve optimal training results.

The training of the ANN with 317 inputs started with a fast decrease in the *rmsd* of the monitoring dataset using resilient back-propagation (Figure 19). However, after 20,000 training periods the *rmsd* showed oscillating behavior. This was due to a well-known effect of the resilient propagation training algorithm (discontinuity around zero)[120]. Switching to the slower simple back-propagation training algorithm stabilized the training behavior and allowed the *rmsd* to be minimized. The monitoring data set showed very similar results to the training data indicating that overtraining had been avoided. As expected, the independent test data showed slightly weaker results.

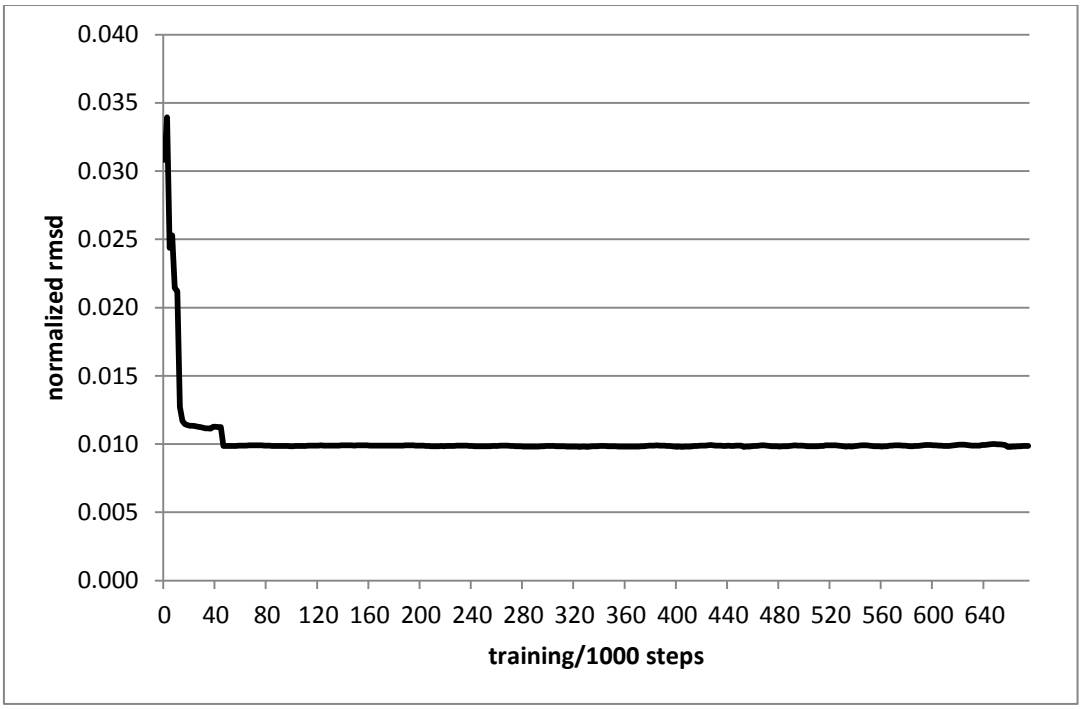


Figure 19: The training of the ANN with 317 descriptors started with approximately 20,000 steps using the resilient propagation algorithm (fast training, steep decline). The rest of the training (~660,000 steps) was accomplished with simple propagation. The first jump in the curve denotes the point after ca. 44,000 steps when the full data set was used for training (before: 50% of the data). The second dent denotes the point where a subset of the data was removed to improve training.

Chemical shifts with a deviation above 17ppm are excluded from final training.

To address the fact that some molecules have wrongly assigned shifts, all small molecules containing an atom with a difference between measured and predicted ^{13}C chemical shift exceeding 17ppm were removed utilizing the model with 317 descriptors trained 20,000 periods of resilient propagation on the full dataset. This left 13,229 molecules out of 14,598 (90.6%).

The trained ANN predicts ^{13}C chemical shifts with a mean average error of 2.95ppm.

The best training results were achieved for an ANN with 317 inputs, 48 hidden neurons and 1 output neuron. The input data consisted of 142,243 carbon atoms with ^{13}C chemical shift values

in 13,229 molecules. The dataset was split into three smaller datasets: 113,795 shifts (80%) for training of the ANN, 14,224 (10%) for monitoring, and 14,224 (10%) for independent testing.

The mean absolute error *mae* and the root mean square deviation *rmsd* in each of the three data sets was computed between all experimental chemical shifts and the shifts predicted by the ANN. The results are 2.44/3.42ppm for the training data, 2.51/3.54ppm for the monitoring data, and 2.95/3.95ppm for the independent data.

The ANN performs well for predicting ^{13}C NMR spectra of organic compounds.

To illustrate the reliability of the prediction method for a whole spectrum, the highest (Figure 20) and average (Figure 21) difference between measured and predicted chemical shifts in a molecule was computed. For 89.3% (11,815) of the molecules the *mae* was below 4ppm which is slightly worse than UpSol NMRPrediction (95% below 3.8ppm). Another 9.1% (1,209) had an *mae* between 4 and 6ppm, leaving 1.6% (205) molecules with a *mae* above 6ppm. Only 14 molecules had a *mae* above 10ppm. The ANN shows differences in prediction accuracy for the different carbon atom types (Table 9). The prediction accuracy decreases when going from *sp3*-hybridized carbon atoms over *sp2*-hybridized carbon atoms to *sp*-hybridized carbon atoms due to the more complex chemical environment and larger influence of π-conjugated systems.

Table 9: The *mae* and *rmsd* of the shift prediction by the different carbon atom types

element type	#bonds	Charge	geometry	#shifts	mae/ppm	rmsd/ppm
Carbon	4	0	tetrahedral	58536	2.25	3.12
Carbon	3	0	trigonal	81850	2.67	3.73
Carbon	2	0	linear	1804	2.53	3.59
Carbon	1	-1	-	22	1.38	2.07
Carbon	3	-1	trigonal	4	1.50	1.56
Total					2.50	3.49

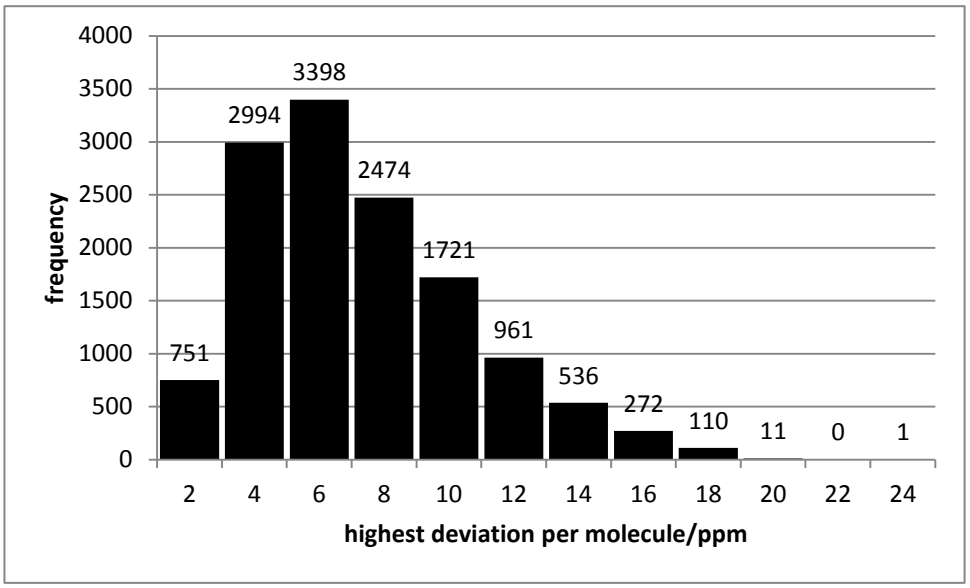


Figure 20: Highest chemical shift prediction deviation per molecule. The histogram shows the distribution of the highest deviation between a single experimental and predicted ^{13}C shift in each of all 13,229 molecules.

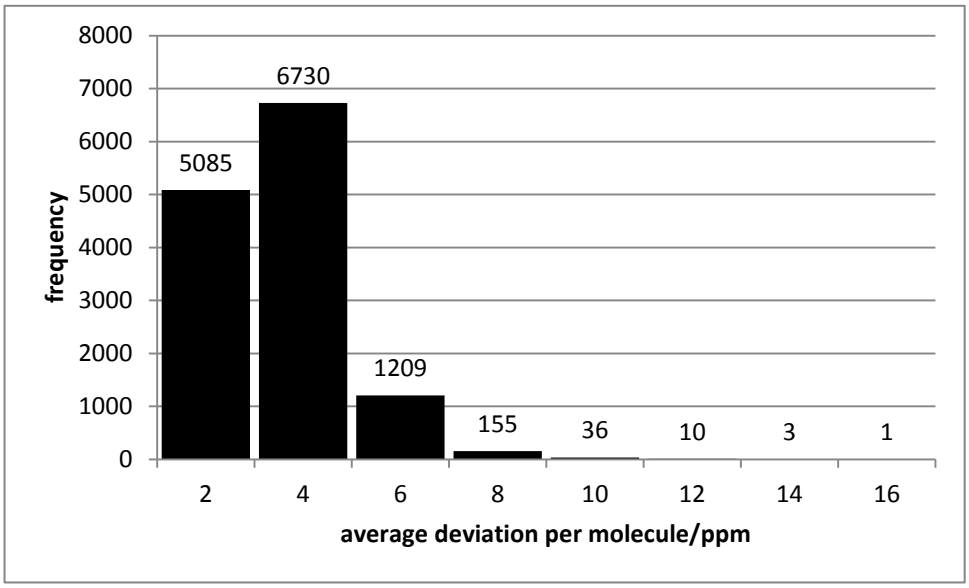


Figure 21: Average chemical shift prediction deviation per molecule. The histogram shows the distribution of the *mae* between experimental and predicted ^{13}C shifts per molecule for all 13,229 molecules.

Predicting natural products with a mae of 3.29ppm.

A small test set of 12 molecules with known ^{13}C chemical shifts was used to determine the performance of the approach for natural products (Figure 22). The overall *mae* for the 263 shifts was 3.29ppm and the *rmsd* was 4.50ppm (see Table 10).

Table 10: The *rmsd* between reported and predicted ^{13}C chemical shifts for some natural products.

structure	Reference	*mae*/ppm	*rmsd*/ppm
strychnine	Singh	3.61	4.52
	Verpoorte 1976	3.23	4.06
	Srinivasan	4.00	5.36
	Leung	3.30	4.21
	Verpoorte 1977	3.23	4.08
	Wenkert	3.09	3.88
brucine	Singh	3.39	4.30
	Wehrli	2.85	3.67
	Verpoorte	3.00	3.79
	Srinivasan	3.74	5.15
	Wenkert	3.09	3.72
anthraquinone1	Xia	3.22	4.34
anthraquinone2	Xia	2.84	4.61
anthraquinone3	Xia	3.06	4.58
subergorgia1	Qi	2.27	2.82
subergorgia2	Qi	4.31	5.48
subergorgia3	Qi	2.96	3.63
5α-androstane	Kalinowski	4.56	6.17
5β-androstane	Kalinowski	4.57	6.27
cholesterol	Kalinowski	3.14	4.41
testosterone	Kalinowski	2.73	3.50

Verpoorte[121] gives a table of different data sets for strychnine and brucine. Xia[122] reports shifts for a new set of anthraquinones found in *Halorosselinia*, Qi[123] for three new polyhydroxylated sterols from *Subergorgia suberosa*. The shifts for the steroids were taken from Kalinowski [124]. The overall *mae* and *rmsd* for all 12 compounds (using the best *mae/rmsd* for strychnine/brucine) are 3.29ppm and 4.50ppm, respectively.

The C_SHIFT program introduced in [109] gave an overall *rmsd* of 3.42ppm. We attribute the difference in performance to the significantly smaller dataset available for training the method.

C_SHIFT was trained with ~1.5 million ^{13}C chemical shifts of the Specinfo database compared to the ~185,000 ^{13}C chemical shifts used for bcl::shift. Looking only at taxol, the *mae* and *rmsd* are 2.7ppm and 4.1ppm, respectively (maximal deviation 14.9ppm). This compares well to other methods discussed in [109] which achieved *maes* between 1.0ppm and 4.0ppm, *rmsds* between 1.5ppm and 5.8ppm, and maximal deviations between 4.1ppm and 23.6ppm.

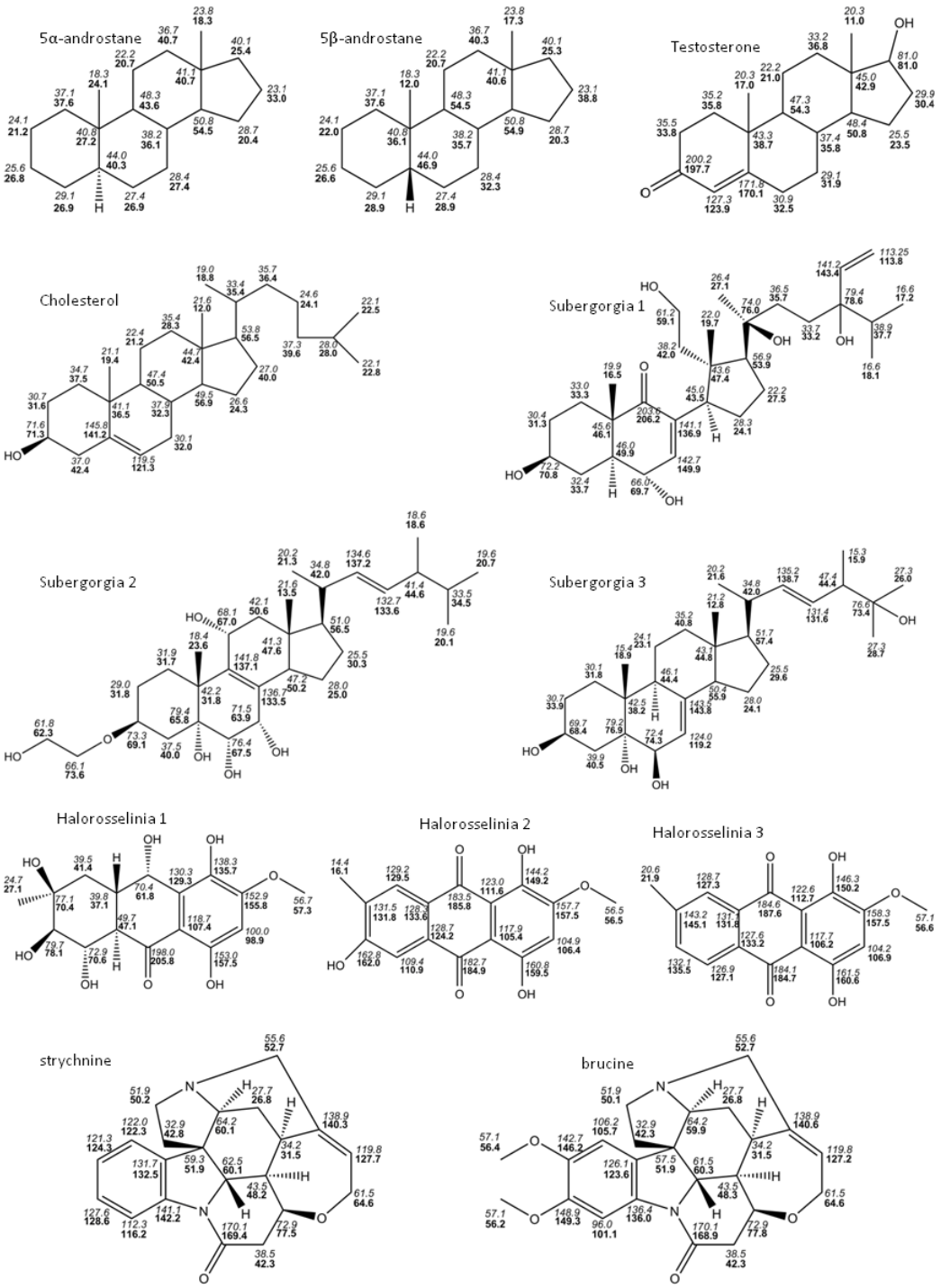


Figure 22: The set of 12 natural products. Brucine and strychnine were taken from Verpoorte[121]. Three anthraquinones found in *Halorosselinia* were reported by Xia[122], Qi[123] is the source for three polyhydroxylated sterols from *Subergorgia suberosa*. The steroids were taken from Kalinowski [124]. The experimental ^{13}C shifts are given in **bold** letters, the predicted in *italic*.

Descriptors were sorted by input sensitivity.

After the conclusion of these experiments stereochemical descriptors were introduced into the model as described in Methods (see Table 13). An input sensitivity analysis determined the importance of each descriptor category. Therefore, ANNs with 16 hidden neurons were trained on 10% of the total data set (18,306 chemical shifts). The input sensitivities are shown in

Table 11. An input sensitivity of 0 means that the descriptor was not employed in this model.

Table 11: Models with three different descriptor sets were trained: 1,525 (all), 1,038 (2DAutoCorr_*Chg, 3DAutoCorr_Ident, _*Chg, RDF_*Chg, _PiEN, _Polariz), 701 (RDF_*Chg). The descriptors on the left-hand side of the table were always employed.

#descriptors	1,525	1,038	701	#descriptors		1,525	1,038	701
atom_of_interest	0.051	0.078	0.123	2DAutoCorr	Ident	0.047	0.000	0.000
sphere1	0.052	0.052	0.071	2DAutoCorr	SigChg	0.129	0.079	0.000
sphere2	0.027	0.027	0.045	2DAutoCorr	PiChg	0.159	0.093	0.000
sphere3	0.025	0.024	0.039	2DAutoCorr	TotChg	0.134	0.129	0.000
sphere4	0.023	0.021	0.034	2DAutoCorr	SigEN	0.042	0.000	0.000
conjugated_sphere1	0.045	0.023	0.035	2DAutoCorr	PiEN	0.023	0.000	0.000
conjugated_sphere2	0.026	0.024	0.039	2DAutoCorr	LpEN	0.009	0.000	0.000
conjugated_sphere3	0.024	0.022	0.036	2DAutoCorr	Polariz	0.021	0.000	0.000
conjugated_sphere4	0.023	0.022	0.033	3DAutoCorr	Ident	0.087	0.070	0.000
bond_types1	0.027	0.042	0.053	3DAutoCorr	SigChg	0.125	0.083	0.000
bond_types2	0.034	0.030	0.043	3DAutoCorr	PiChg	0.148	0.128	0.000
bond_types3	0.009	0.014	0.030	3DAutoCorr	TotChg	0.144	0.113	0.000
bond_types4	0.006	0.007	0.010	3DAutoCorr	SigEN	0.039	0.000	0.000
bond_types5	0.006	0.007	0.008	3DAutoCorr	PiEN	0.023	0.000	0.000
ring_closure	0.016	0.015	0.024	3DAutoCorr	LpEN	0.008	0.000	0.000
solvent_properties	0.047	0.027	0.032	3DAutoCorr	Polariz	0.022	0.000	0.000
temperature	0.005	0.003	0.005	RDF	Ident	0.049	0.000	0.000
				RDF	SigChg	0.430	0.277	0.353
				RDF	PiChg	1.519	1.725	2.880
				RDF	TotChg	0.502	0.399	0.580
				RDF	SigEN	0.039	0.000	0.000
				RDF	PiEN	0.079	0.053	0.000
				RDF	LpEN	0.041	0.000	0.000
				RDF	Polariz	0.056	0.049	0.000
Sum						1.705	1.819	2.965

RDF SigChg, PiChg, TotChg are the best conformational descriptors.

Since the *rmsd* of the models trained on 10% of the data was inconclusive, each of the descriptor sets was employed to train a model on the full dataset. The model with 701 descriptors (317 constitutional descriptors + 3*128 RDF SigChg, PiChg, TotChg) had the smallest overall *rmsd* after 20,000 periods of training utilizing resilient back-propagation of errors.

A combination of resilient and simple back-propagation is used to achieve optimal training results.

The model with 701 descriptors was trained to completion on the pruned dataset. The training of the ANN started with a fast decrease in the *rmsd* of the monitoring dataset using resilient back-propagation (Figure 23). After 20,000 training periods all chemical shifts with a deviation greater 17ppm between experimental and predicted values were removed. Another 20,000 steps of resilient propagation yielded no improvement. Switching to the slower simple back-propagation training algorithm allowed the final minimization of the *rmsd*. The monitoring data set showed very similar results to the training data indicating that overtraining had been avoided. As expected, the independent test data showed slightly weaker results.

The trained ANN predicts ^{13}C chemical shifts with a mean average error of 2.84ppm.

The best training results were achieved for an ANN with 701 inputs, 48 hidden neurons and 1 output neuron. The input data consisted of 181,786 carbon atoms with ^{13}C chemical shift values in 17,336 molecules. The dataset was split into three smaller datasets: 146,112 shifts (80%) for training of the ANN, 18,402 (10%) for monitoring and 18,152 (10%) for independent testing.

The mean absolute error *mae* and the root mean square deviation *rmsd* in each of the three data sets was computed between all experimental chemical shifts and the shifts predicted by the ANN.

The results are 2.55/3.42ppm for the training data, 2.79/3.65ppm for the monitoring data, and 2.84/3.73ppm for the independent data.

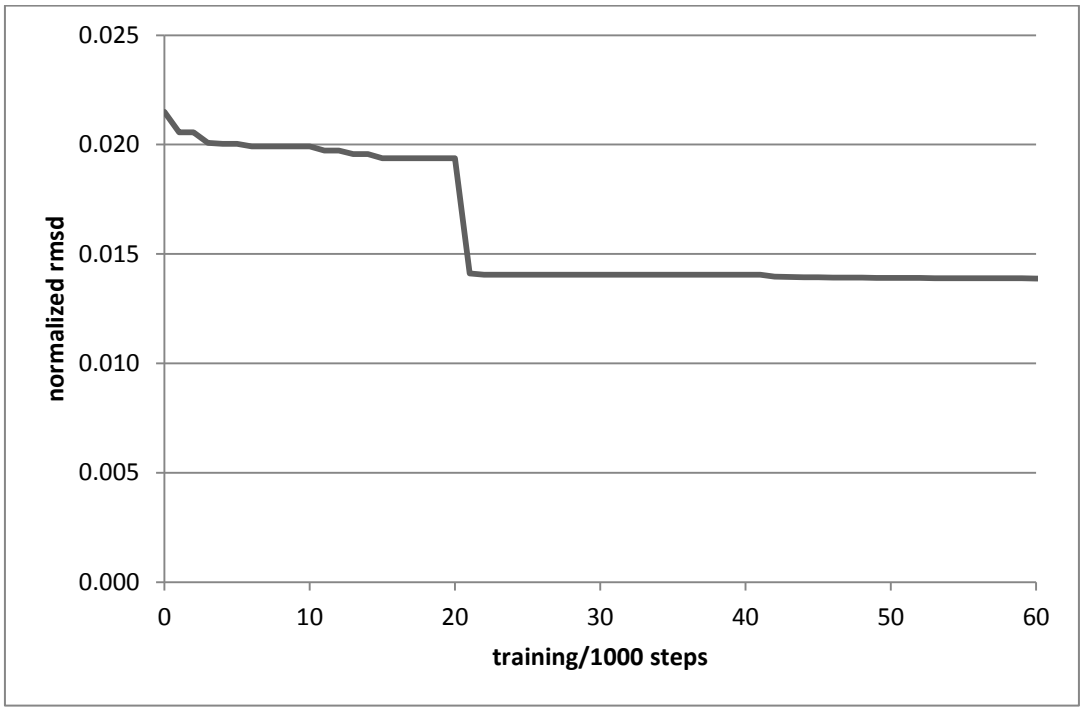


Figure 23: The training of the ANN with 701 descriptors. The training started with approximately 20,000 steps using the resilient propagation algorithm (fast training, steep decline). After 20,000 steps, all chemical shifts with a deviation greater 17ppm between experimental and predicted values were removed. Another 20,000 steps of resilient propagation yielded no improvement; the training was finished with simple back-propagation of errors.

Including radial distribution function descriptors allows differentiation between configurations.

To illustrate the capability of the trained ANN to differentiate between different configurations, two isomers (cis-/trans-2-buten-1-ol) were predicted with the final model. The code is still identical for the two sp2-carbons, since all distances to other atoms are the same for both isomers. However, the two sp3-carbons are described differently due to the changes in distance to (some of) the other atoms in the molecule.

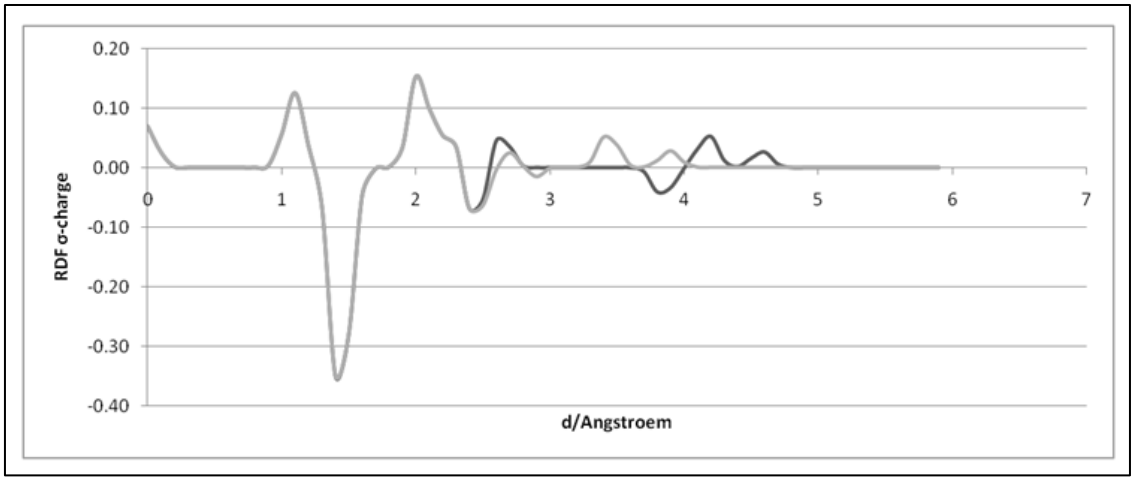


Figure 24: The difference between the σ-charge radial distribution functions for carbon 1 in cis- (light gray) and trans-2-buten-1-ol (dark gray). The configurational changes occur between 2.5 and 5Å.

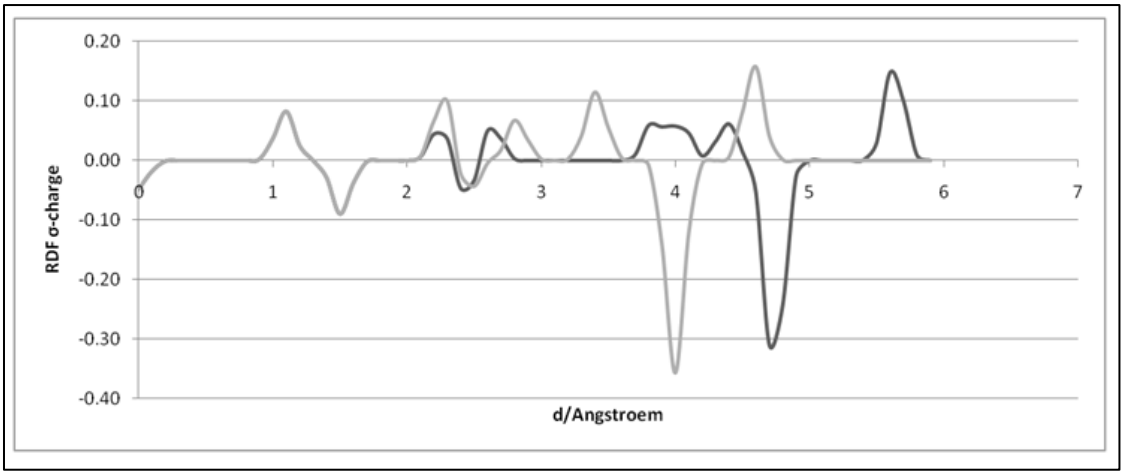


Figure 25: The difference between the σ-charge radial distribution functions for carbon 4 (methyl) in cis- (light gray) and trans-2-buten-1-ol (dark gray). The peaks between 3 and 6Å are shifted due to the change in distance.

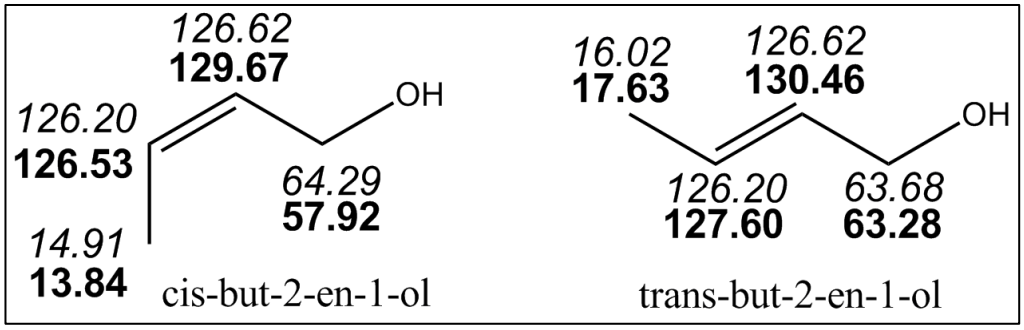


Figure 26: Comparison between experimental (bold) and predicted (italic) ^{13}C chemical shifts for two isomers. While carbon 4 is predicted along the experimental values, the order of the shifts for carbon 1 is inversed. The experimental values were taken from SDBS[4c]. rmsd=3.6ppm(cis), 2.2ppm(trans), mae=2.7ppm(cis), 1.8ppm(trans)

Growing ^{13}C databases will improve prediction accuracy.

Given the small size and the diversity of the trained dataset, the results compare well to similar works that incorporated larger and more homogenous databases like SpecInfo. Currently approximately 185,000 ^{13}C shifts were used for training the ANN. This is only about 10% of the 1,500,000 shifts in the SpecInfo database used in previous approaches. As publicly available databases grow, the ANN will be retrained to improve prediction accuracy.

Conclusions

The publicly available data of the NMRShiftDB can be used to train an ANN to predict ^{13}C chemical shifts with a *mae* of 2.95ppm (*rmsd* of 3.95ppm). For a subset of 12 natural products a *mae* of 3.29ppm (*rmsd* of 4.50ppm) was determined demonstrating the ability of the methods to predict the ^{13}C chemical shifts of newly discovered natural products. The successful introduction of configurational and conformational descriptors was shown by an improved mae on the independent data set of 2.84ppm and comparing the predictions for cis-/trans-but-2-en-1-ol. The most current version of the bcl::shift prediction tool is available at www.meilerlab.org for academic use.

Methods

Programming and data processing.

All of the procedures described below are executed utilizing the object-oriented C++ BioChemistry Library (BCL) developed in the Meiler laboratory. The SD files provided by the NMRShiftDB were read, converted into BCL molecule and spectra objects, and stored in a MySQL database. The ANN is trained by back-propagation of errors (see below).

Artificial neural network architecture.

The ANN used is a feed-forward network trained with back-propagation of error. The ANN consists of 317 inputs for the numerical description of the chemical environment (see below). Those inputs are fully connected to a hidden layer with 48 neurons. To compute the output of each of these 48 hidden neurons the input data x_i to the first layer are summed up according to their neuron-specific weights and modified by the sigmoid activation function:

$$f_j(x_i) = \left[1 + exp\left(\sum_i w_{ij} x_i\right)\right]^{-1}$$

The output f_j then serves as input to single neuron of the final layer predicting the ^{13}C chemical shift (Figure 27):

$$o = \left[1 + exp\left(\sum_j w_j f_j(x_i)\right)\right]^{-1}$$

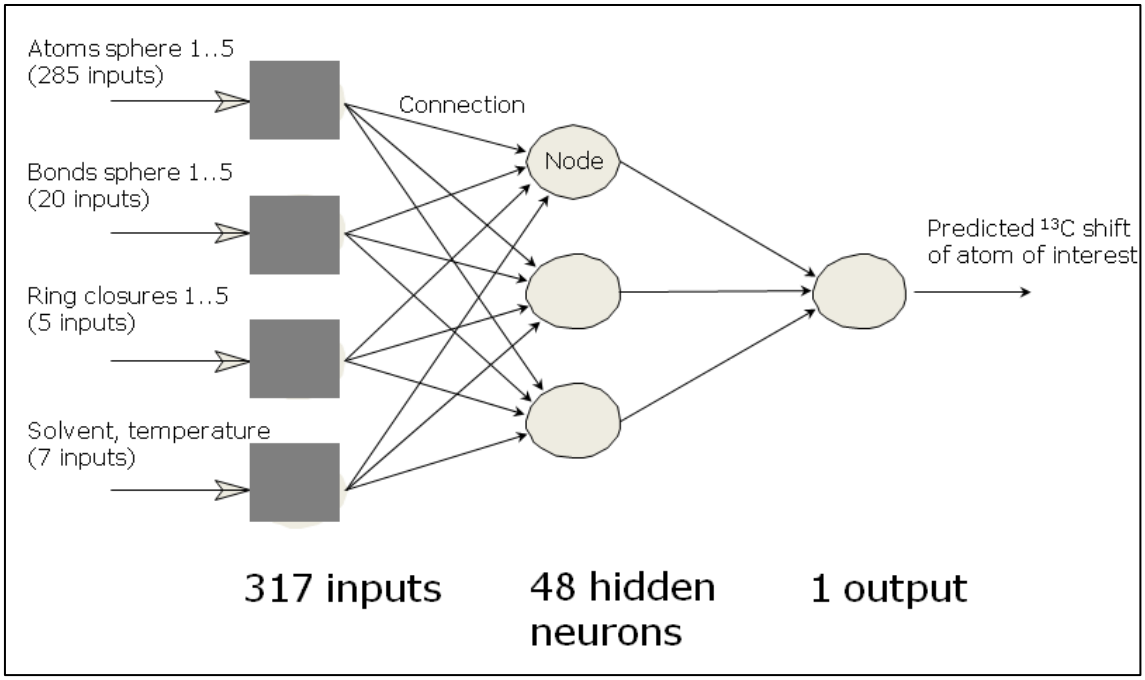


Figure 27: Principal scheme for the spherical code (5 spheres) and the artificial neural network.

Training was performed with simple and resilient back-propagation of errors.

Two back-propagation algorithms were applied, simple back-propagation [110a] and resilient back-propagation[10a] of errors. Back-propagation starts with a normal forward run through the ANN determining the output for a given input vector. The difference between ANN output and target output is recorded as prediction error which is now "back-propagated". Starting with the output layer – i.e. backward – the error determines an adjustment to the weights layer by layer. Simple back-propagation is a gradient-based update algorithm with two adjustable parameters, the learning rate η and the momentum α. Resilient back-propagation is a recent improvement of the algorithm where the weight adjustment depends on the sign but not on the magnitude of the gradient.

Close to 185,000 ^{13}C chemical shifts were available for training.

An ensemble of 17,336 molecules with 17,913 ^{13}C spectra was extracted from the NMRShiftDB [104] (www.nmrshiftdb.org) database which provided ^{13}C chemical shifts for 185,099 carbon atoms. In a first step, 87 molecules missing bonds or containing atom types not considered in our encoding scheme (e.g., sodium) were removed. In a second step, the remaining 17,249 molecules were rectified by adding missing hydrogen atoms and charges.

Atoms are sorted in 35 distinct groups.

To describe each atom a list of atom types was derived empirically from the ensemble of 13,229 molecules. Each atom type was defined by the element type, the number of bonds, the nominal charge, the number of electrons in bonds (i.e. double bonds, triple bonds, aromatic bonds), and the geometry at the center. The possible bond types are only given implicitly: The second carbon atom type describes both double-bonded and aromatic carbons, the third allenic and acetylenic carbons. This procedure yielded 35 distinct atom types summarized in Table 12. In the process of atom type determination all bonds are categorized into four groups: single, double, triple, or aromatic.

Table 12: Atom types: Each atom type is determined by the element type, the number of bonds, the charge, the number of electrons in bonds (to describe double, triple and aromatic bonds), and the geometry (linear, trigonal, and tetrahedral).

	bonds	charge	#e- in bonds	geometry	frequency		bonds	charge	#e- in bonds	geometry	frequency
H	1	0	1	-	236295	O	2	0	2	tetra	21214
C	4	0	4	tetra	80884		1	0	2	-	15005
	3	0	4	trig	109761		1	-1	2	-	40
	2	0	4	lin	2416		2	1	2	trig	15
	1	-1	4	-	27	F	1	0	1	-	1469
	3	-1	4	trig	7	P	3	0	3	tetra	67
N	3	0	3	tetra	1623		4	0	5	tetra	170
	2	0	3	trig	4898		4	1	3	tetra	11
	1	0	3	-	1167	S	2	0	2	tetra	2012
	4	1	3	tetra	75		1	0	2	-	247
	3	1	3	trig	61		3	1	2	tetra	18
	2	1	3	lin	37		4	0	6	tetra	440
	3	0	3	trig	6616	Cl	1	0	1	-	3145
	4	0	5	tetra	9	Br	1	0	1	-	1618
	3	0	5	trig	1013	I	1	0	1	-	339
	2	0	5	lin	46		2	1	1	tetra	10
Si	4	0	4	tetra	310	Se	2	0	2	tetra	133
B	3	0	3	tetra	270						

Special descriptors are introduced for π-conjugated systems

Within conjugated systems a substitution on one side of the molecule can trigger a shift in the electron density distribution over large distances and change the shielding of a carbon atom many bonds away. To address this issue, all atoms of a molecule were distinguished as in contact or not in contact with the atom of interest via a conjugated π-system. One π-system is defined as a set of connected atoms that have a planar geometry and either lone-pair electrons or double, triple, or aromatic bonds. All these atoms in the direct neighborhood of the carbon atom of interest are grouped together and yield an additional set of numerical descriptors for input into the ANN. This procedure is similar to the one described by Meiler[109].

The numerical description of the atom environment considers five spheres defined by molecule constitution.

The numerical code describes the constitutional environment of each carbon atom. Since the ANN has a defined number of inputs, the environment code must be described by a vector of constant length. Furthermore, it is advantageous if every input of the ANN always encodes the same property.

Centered on the atom of interest, *n* spheres in the network of chemical bonds are encoded (Figure 28). The *i*-th sphere consists of all atoms that have a minimal distance of *i* bonds to the atom of interest. For a small molecule the outermost spheres with $i > 5$ are often sparsely inhabited. This was addressed by allowing those inputs to be zero if the distinct atom type or bond type is not present (see below). The encoding scheme relies only on the constitution of the molecule. The optimal number of spheres was determined to be five by systematically increasing *n* until no further improvement in the accuracy of ^{13}C chemical shift prediction was observed.

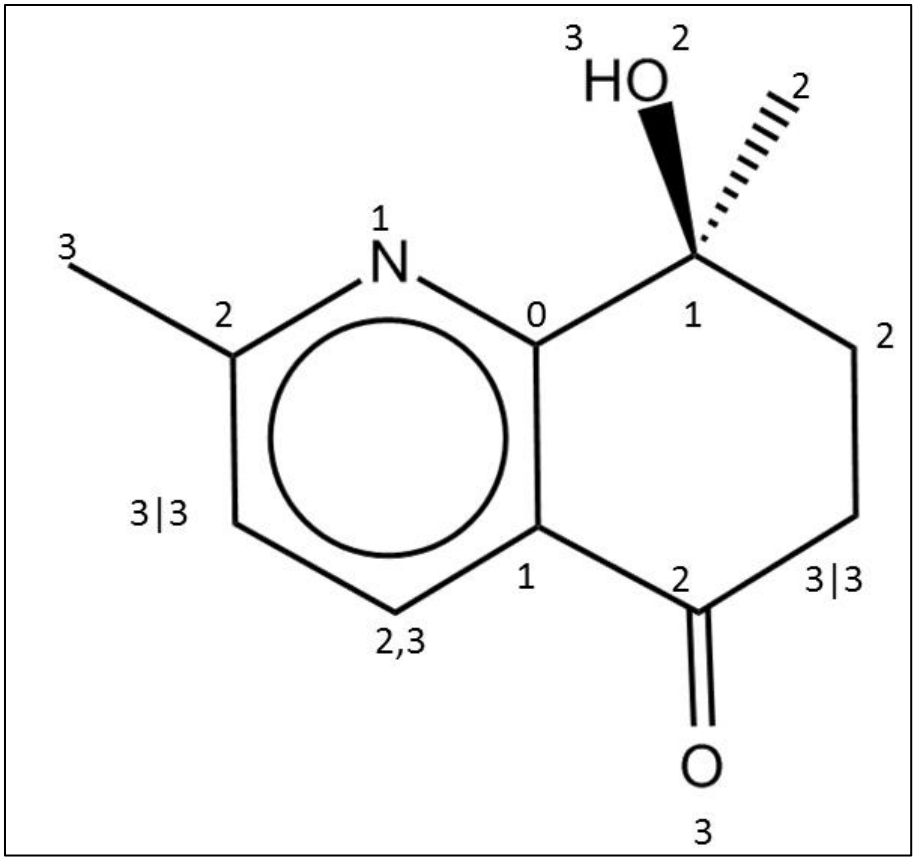


Figure 28: An example for the sphere code for *n=3* spheres: The atom of interest is the only element in sphere 0. The first sphere consists of nitrogen, *sp2*- and *sp3*-hybridized carbon. Their substituents constitute sphere 2. Sphere 3 contains two ring closures indicated by 3|3.

Frequencies of atom and bond types in each sphere are input for the ANN.

For the numerical code, the frequency of atom types in each sphere is determined. The first sphere – i.e. the atom of interest – can only contain one of five carbon atom types. Since 35 atom types are considered in a π-conjugated and non-π-conjugated state, this leads to 2*35*(*n*-1) inputs for the remaining spheres. Furthermore, the number of single, double, triple and aromatic bonds per sphere (4**n* inputs) and the number of ring closures happening in this sphere are added (*n* inputs). Six inputs describe the physical properties of the solvent including molecular weight, boiling point, melting point, density, dielectric constant, and dipole moment. If the solvent was

unknown, chloroform would be assumed. The last input is the temperature at which the spectrum was taken. For *n*=5 spheres this leads to a total of 317 inputs.

Constitutional and conformational descriptors utilize up to eight chemical properties.

We implemented 2D/3D auto-correlation and radial distribution functions to describe the constitution and conformation of the molecule (Table 13) based on a similar approach to ADRIANA[5].

Table 13: 2D/3D auto-correlation and radial distribution functions

Descriptor Name	Description
2D Autocorrelation (**2DACorr**) 3D Autocorrelation (**3DACorr**) Radial Distribution Function (**RDF**)	11 descriptors (d=11) 12 descriptors (d=12) 128 descriptors (d=128)
***_Ident	weighted by atom identities
***_SigChg	weighted by σ atom charges
***_PiChg	weighted by π atom charges
***_TotChg	weighted by sum of σ and π charges
***_SigEN	weighted by σ atom electronegativities
***_PiEN	weighted by π atom electronegativities
***_LpEN	weighted by lone pair electronegativities
***_Polariz	weighted by effective atom polarizabilities

σ-charge was implemented utilizing Partial Equalization of Orbital Electronegativity (PEOE, [115]), π-charges were calculated by identifying conjugated systems and solving the according Hueckel-matrices. The Hueckel parameters were taken from [125]. Total charge is the sum of σ- and π-charge. The electronegativities were determined from the charges by means of a quadratic polynomial fitted through charges -1, 0, and +1[115, 119]. The determination of polarizability is detailed in [126]. The codes are all centered on the atom of interest and sum up a certain property of each atom in a certain distance (interval) from the atom of interest.

The formula for 2D auto-correlation: $h(d) = \sum p_d$, *d*=0, …, 10, where p_d is the property of each atom that is *d* bonds away.

The formula for 3D auto-correlation: $h(d) = \sum p_d$, *d*=0, …, 11, where p_d is the property of each atom with a Euclidean distance between *d* and *(d+1)*Å.

The formula for RDF: $h(d) = \sum p_d \, e^{-100(0.1d-d')^2}$, *d*=0, …, 127, where p_d is the property of each atom with a Euclidean distance *d'* from the atom of interest, effectively measuring distances between 0 and 12.7Å.

This leads to a total of 1208 descriptors. For an illustration how configurational changes are captured by the radial distribution functions see Figure 24.

Descriptors are sorted by input sensitivity to improve prediction accuracy.

To determine the significance of each input the ANN is considered to be a multidimensional function

$$y = f(x_1, x_2, \dots, x_n)$$

with input values $x_1,\dots,x_n$ and output *y*. Then the sensitivity of each input can be measured by determining the partial derivative $\frac{\partial^i f}{\partial x_i}$.

For categories containing multiple descriptors, the square norm of the sensitivity of all inputs is considered as total sensitivity.

A single ANN is trained to predict the chemical shift of all ^{13}C atoms.

The advantage of this encoding scheme is its flexibility: Each molecule and each atom of interest can be encoded with the same scheme and therefore predicted by a single ANN. This is in contrast to previous work [109], where a distinct encoding scheme and ANN was used for each

carbon atom type. The switch to a single ANN was prompted by the significantly smaller database of spectra available for determining the parameters of the ANNs. The more generic encoding scheme allows for a single ANN with fewer parameters.

Training was accelerated using recently optimized training algorithm.

To train the network, Resilient Propagation[10], an adaptive gradient-based algorithm, was implemented. It constitutes a faster learning algorithm than standard back-propagation of errors. However, due to oscillating behavior of the resilient back-propagation after an initial optimization of 20,000 steps, the final optimization was carried out with up to 700,000 steps of simple back-propagation. During the simple back-propagation, the learning rate η was reduced from 0.1 to 0.001, the momentum α was increased from 0.5 to 1.0. The training took place on a RedHat Scientific Linux workstation with 2GB RAM and 2 Intel Pentium D processors (3.2GHz). The total time required for training the ANN amounted to 2 weeks.

CHAPTER VI

DISCUSSION

Conclusions and future directions

Quantitative Structure Activity Relations

The trained neural networks established very stable quantitative structure-activity relationships for metabotropic glutamate receptors. The quality measures of root mean square deviation, area under the ROC curve and enrichment are comparable for all three models. The strength of the models lies more in predicting if a compound will be active or not (a binary decision) based on a certain cut-off rather than in exactly predicting the potency of a given molecule (hence the large overall *rmsd* values). This could be based on the fact that all inactive compounds were set to an arbitrary potency of 1mM where the correct values are spread out over a certain range. However, the enrichment is only based on picking the right compounds below the aforementioned cut-off, making it the most reliable quality measure employed in these experiments. An interesting question would be if setting the inactive compounds to a different value, e.g., 100µM, influences the quality of the models. While reducing the distance between the least active and the inactive compounds, it would augment the differences in activity between the active compounds. Another improvement to the model quality could consist in finding a way to distinguish inactive compounds by their level of inactivity. The first thing that comes to mind is expanding the range of the concentration response curve experiments. Furthermore, if a different assay for the same target and small molecules is available, its data could be incorporated in the same model. Also, if additional experimental data like solubility of the compound could be collected during the high-throughput screen, these data could reinforce or even replace some of the computed descriptors like calculated logP. One way to enhance the model quality without performing additional high-

throughput experiments consists of training models on other data collected or derived in the original HTS like minimum and maximum of glutamate response or efficacy. Taking this idea to its natural conclusion, all thirty experimental data points per compound (ten concentrations in triplicate) could be predicted allowing the differences in the quality of the curve and its confidence levels to inform the model. In such a model, positive and negative allosteric modulation of the same target could be described simultaneously, since they are represented by differently shaped curves in the same concentration response curve experiment, where EC50 and IC50 cannot be encoded by the same output due to their different nature. This would substantially increase the rate of actives over inactive possibly leading to better models. Training times for the models should be similar if only the number of output is increased because the number of degrees of freedoms (weights) will only be slightly larger. However, it could be necessary to employ more hidden neurons due to the larger dimension of the output space which would increase training times.

Encoding the drug molecules through chemical descriptors ensures that the activity of each possible molecule can be determined by the trained models independent from the molecule's orientation in space. The actual conformation of the molecule is captured by the 3D descriptors which consistently play an important role in each final model. However, each molecule is only represented by one conformation computed by the CORINA software package without any knowledge of the target protein. Therefore, the actual conformation of the bound molecule will be substantially different from this one computed conformation. One way to alleviate this problem without *a priori* knowledge of the protein is to compute a conformational ensemble of the compound in question. This works especially well for the active compounds which are oversampled to begin with: Each instance of the balanced data set could represent a different conformation of the small molecule.

All three models allowed the identification of previously unknown compounds. The virtual high-throughput screen was performed on the external ChemBridge and ChemDiv compound libraries that were also the building blocks for generating the Vanderbilt high-throughput screening center collection of molecules. Screening these larger libraries could be biased towards predicting similar compounds to the known active ones. A future experiment should be conducted to determine if screening compound libraries unrelated to the ones employed in creating the library the model is trained on would lead to a higher percentage of compounds representing novel scaffolds.

Automated feature selection and training of the neural networks allows greater flexibility in selecting novel targets and collaborating with other scientists. So far, the overall experiment involves several manual steps. Integrating the overall work flow in an existing database solution for high-throughput screening like Pipeline Pilot would make this approach an integral part of many drug discovery projects.

Final models for all three QSARs were integrated into a web server predicting the activity of a given molecule towards positive and negative allosteric modulation of metabotropic glutamate receptor subtype 5 and positive allosteric modulation of subtype 4. Currently, the web server is only available in the Vanderbilt Drug Discovery program and is employed to routinely check compounds for their usability as allosteric modulators of mGlu4 and mGlu5.

All neural networks were trained on a subset of ADRIANA descriptors. The inclusion of other sets of descriptors could improve the neural networks' capability to correctly describe additional chemical features that increase the enrichment of active compounds in virtual high-throughput screens. The overall size of the model could be kept similar to the existing models by removing more descriptors during the feature selection process.

Different methods were assessed to determine the importance of each descriptor category. In a first implementation of the feature selection, the mean of the descriptor sensitivities in each category was employed to rank them. In a second approach, the Euclidean sum of the sensitivities was the decisive criterion. This put more emphasis on the larger descriptor categories like radial distribution functions (128 inputs each). Other methods of ranking the descriptors could better evaluate each category's importance. Reducing the number of bins for radial distribution functions would provide another way to balance the descriptors.

Four of the biggest questions remaining are: (a) Can this approach be easily transferred to other receptors? An attempt to answer this question can be found in an upcoming paper that benchmarks this method for such targets as the M1 muscarinic receptor. The general setup of the computational workflow allows for easy application to different HTS. (b) What is the smallest set of compounds with known activities needed for this approach? I. e., what size of a high-throughput screen would allow virtual high-throughput screening? Leading directly to (c) How can the iterative approach be expanded towards an automated way of improving the overall quality of the neural network prediction? Based on the observation that the 3D descriptors play an important role in each of the final models another interesting question is (d) Does the quality of the models improve if an ensemble of low-energy conformations is employed in describing the molecules? This could help capturing differences in the conformations of bound and unbound molecules, even without necessarily knowing the conformation of bound molecules in a target/drug complex.

Metabotropic glutamate receptor 5 PAMs

The high ratio of confirmed metabotropic glutamate receptor subtype 5 positive allosteric modulators in the high-throughput screen allowed for the training of very stable models, resulting in a high enrichment that is similar over the range of models trained.

It is known that there are at least two allosteric sites for positive modulation of mGlu5 representing two or more different scaffolds. The general approach allowed describing modulators for both sides in a single model. This led to a broad range of novel mGlu5 PAMs being discovered.

Even if the majority of the high-throughput screening results are published, results of small screens for mGlu5 PAMs should be incorporated into the existing models by training the ANNs on this additional data. Additional PAMs could reduce the need to balance the data set between active and inactive compounds. After the drug discovery efforts for mGlu5 PAMs are finalized, the trained model should be made publicly available to allow incorporation into future drug discovery efforts toward this target.

Metabotropic glutamate receptor 4 PAMs

The set of positive allosteric modulators towards metabotropic glutamate receptor subtype 4 identified by high-throughput screening was significantly smaller than the set of PAMs for subtype 5. Furthermore, the 434 PAMs showed a high degree of similarity. This could be one of the reasons for the lower enrichments compared to the model for mGlu5 PAMs.

The designed experiment consisted of two rounds of collecting data, training models, and predicting mGlu4 PAMs. This allowed for the possibility of incorporating the results of a virtual high-throughput screen in an iterative fashion to improve the quality of the trained model. Furthermore, the second round aimed at increasing the possibility for scaffold hopping by choosing a higher cutoff for predicting compounds as being active. This further reduced the enrichment.

The mGlu4 PAM vHTS experiment clearly was the most difficult one in terms of the available data. The main result here is the transferability of the approach from mGlu5 to a subtype of a

different group of metabotropic glutamate receptors. Even while active molecules with non-trivial modifications from known actives were found, true scaffold hopping did not occur.

New screening results for mGlu4 PAMs should be incorporated into the existing model. After the drug discovery efforts for mGlu4 PAMs are finalized, the model should be disclosed.

Metabotropic glutamate receptor 5 NAMs

Being the last of the three experiments to predict modulators of metabotropic glutamate receptors, this model greatly profited from the experiences with the first two. Descriptor selection and iterative approach proved valuable in generating true scaffold hopping. Two molecules with approximately 100nM potency each were discovered in the set of molecules predicted to be active by the model.

The SAR around these two molecules proved to be very narrow which disqualifies them as lead compounds for developing new drugs. The compounds could still be employed as probe molecules.

The experimental data will be published together with animal model behavioral studies and strong biological data for the two novel NAMs.

Carbon chemical shift prediction

Artificial neural networks were trained in a quantitative structure property relation experiment (QSPR) to predict carbon chemical shifts. All neighbors of an atom of interest were organized in spheres. The distribution of atom types in these spheres together with bond and ring descriptors served as input for the neural networks. A web server was implemented based on this universal model which is publicly available.

Based on the ADRIANA descriptors, radial distribution functions describing the distribution of atoms in the three-dimensional space around the atom of interest were developed. These allowed the neural network to distinguish between different configurations and conformations. Including these descriptors into the input for each atom of interest greatly improved the quality of the model.

Increasing the number of data points by identifying freely available chemical shift databases will improve the quality of the trained models. Another source could be locally collected spectral data through a web tool designed by Bill Graham after discussion with Don Stec and Ralf Mueller. This tool could also help in transferring carbon spectra from the literature to a local database.

In the future, the influence of different machine learning approaches (e.g., support vector machines) on the quality of the models should be evaluated.

Other steps to improve the quality and the usability of the QSPR models could be the inclusion of other spectra, especially proton spectral data, expansion of the descriptor base drawing on the overlap with the metabotropic glutamate receptor experiments, and applying descriptor optimization techniques described in the QSAR chapters.

APPENDIX

General comments

The computational work presented here is based on applications programmed in the BCL. Command lines will usually employ a compiled version of the BCL ("apps_release.exe"). A recent version of the BCL executable is available by executing "bcl.exe". However, to modify existing applications it is necessary to check out a local copy of the BCL repository.

The command lines and applications reported here span five years of development in the BCL. The description of the data in the files was changed several times to reflect improvements in the object design in the BCL. Therefore it can become necessary to adjust headers or structures of data files to the most current version represented in the BCL. A general understanding of a scripting language like awk, perl, or python will prove beneficial in this task.

General explanations for each subdirectory are in READ_ME.txt files which can be looked at for instance with less or any text editor like vi. Output of specific command lines was redirected into log files usually starting with "fun" and containing the specific BCL application in their name.

QSAR data structure

The QSAR projects described in chapters II to IV share a very similar data structure: under /QSAR/ are subdirectories named after each of the QSAR chapters: mGlu5_PAM, mGlu4, mGlu5_NAM. Each folder contains subdirectories for all trained models distinguished by the number of descriptors, e.g. "1252descriptors/" for a model employing all of the ADRIANA descriptors. Executable files compiled specifically for the given data are collected in "executables/". These should especially be employed in cases were the current version of the

BCL fails to process data files. Collections of molecules that were predicted with the trained models are stored in SD files under "sdf/".

QSAR applications

After collecting SD files for the active and inactive molecules for a given drug target, the active compounds needed to be oversampled due to the usually low rate of active compounds in the set of all molecules. To remove active molecules from the data set of all molecules tools like 'sdsort' can be utilized. The 'combine_sm_ensembles.exe' application loads the SD files containing the active and inactive compounds into small molecule ensembles, determines the enrichment factor, randomizes both ensembles, oversamples the smaller small molecule ensemble to a similar size to the larger ensemble and folds the two ensembles together before writing it out to a combined SD file. 'Folding' means alternating between active and inactive compounds which ensures an even distribution of active and inactive compounds in the training, monitoring, and independent data set. The randomization has to be done before the oversampling and folding to avoid the distribution of the independent active molecules to the monitoring and training data sets.

The 'apps_release.exe GenerateSmallMoleculeCode' application computes the input for training an ANN or other type of model from a given SD file and code object. The code object is a list of descriptors available in the BCL, mostly based on re-implementations of ADRIANA descriptors. The output is stored in .dat-files which are lists of pairs of vectors. Each pair of vectors represents the input and according output calculated from one molecule (molecule-based descriptors).

The training of the ANNs was realized through the 'apps_release.exe TrainANN' application which allows the specification of the architecture/geometry of the ANN, the number of training steps, the training algorithm, etc.

Trained ANNs can be employed in applications like the web server 'PredictQSAR' predicting the activity of a given small molecule towards mGlu5_NAMs/PAMs and mGlu4_PAMs, and 'ComputeQSARRange' determining all small molecules with predicted activities in a specified range.

QSAR/mGlu5_PAM

The mGlu5 PAM project is the oldest QSAR project described in this document. Most of the tools described here were developed while undertaking it. Input and output were described in the same vector. Subsets of 20,000 small molecules can be found in the 'generate_data_*' subdirectories where dat-files were created from the ADRIANA descriptors stored in the SD file. Enrichment was determined manually and the overall dat-file generated by randomizing and folding the data vectors employing awk lines developed by Kristian Kaufmann.

The descriptor selection process was performed in 'mGlu5' where csv-files for different sets of descriptors represent the different stages of descriptor optimization. The files are not designated by the number of descriptors but instead list all descriptor categories employed in the actual model.

Eric Dawson trained the ANNs from 428 down to 136 descriptors and the scalar model with 8 descriptors. The results can be found in '/home/dawsones/qsar/encode/Adriana/HTS/mol_descriptor'. A local copy can be found under 'models_eric'.

The 'ChemBridge' folder contains all results for the virtual HTS described in chapter II. Furthermore, the folder holds scaffold hopping experiments which were abandoned due to the wealth of known mGlu5 PAMs and focus on the other QSAR experiments described in chapter III and IV.

QSAR/mGlu4_PAM

The active compounds are stored in sdf/mGlu4_Vandy_confirmed_Nov08.sdf.bz2. The file with the ADRIANA descriptors is in the same directory. The inactive compounds of the HTS at Vanderbilt are stored in minus_confirmed/all.sdf.

Purging describes the process of removing mostly charged small molecules which were missing ADRIANA descriptors. Descriptors were generated from all_combined.sdf.

The trained models and the jury can be found in the respective directories. The first trained jury had different independent data sets for the 415, 578, and 741 descriptor models which were fixed in old_same_independent. The scaffold hopping experiments were performed together with Thuy Nguyen (/home/nguyent8/Projects). A local copy can be found in scaffold_hopping/Projects.

QSAR/mGlu5_NAM

The SD files with the active (HTS_mGlu5_NAM_adriana.sdf.bz2) and inactive compounds (Vandy_Library_minus_mGlu5_NAM_adriana.sdf.bz2) can be found in the sdf directory. The randomized, balanced, and folded SD file is code.sdf.bz2.

The trained models are in their respective subdirectories together with the corresponding dat-files. The prediction of active compounds at different cut-offs, the filtering of these small molecule sets by different criteria, and the generation of the compound orders was done by Eric Dawson (/home/dawsones/mGlu5_NAM). A local copy can be found under analysis_eric.

Predict_chemical_shifts data structure

The predict_chemical_shifts subdirectory contains the work connected to chapter V. Each 'datagen200*' subdirectory consists of (modifications to) an implementation of an atom-centered code, data generation based on this code, ANN training, and optimization of code parameters like number of spheres around the atom of interest and ANN parameters like number of hidden neurons. For comparison, the models reported in Meiler 2002 were retrained in the 'ANNnew100nmr*' subdirectories.

The two main data bases employed in the training of the carbon chemical shift models can be found under 'data_bases/NMRshiftDB' and 'data_bases/CSD'. Dihedral angle histograms for all pairs of atom types found in the CSD were generated in 'histos'.

Test sets of different small molecules are stored under 'sdf'. References can be found in the 'references' subdirectory.

Additional information can be found in the READ_ME.txt and fun*.log files in the according subdirectories.

Predict_chemical_shifts/ applications

The overall approach is similar to the QSAR projects. After generation of the input data (*.dat.bz2) from the SD files ANNs are trained employing the 'apps_release.exe TrainANN' application. However, due to the different approaches for code generation different 'apps_release GenerateCSCode' applications exist. The details will be discussed in the following paragraphs. Usually, it is unnecessary to balance and fold the data set. In one experiment it was tested to balance molecules with high deviations (small data set) against molecules with low deviations (large data set) but improvements were negligible.

quality_assurance/

To provide conformational data for the small molecules in NMRshiftDB, the overlap between NMRshiftDB and Cambridge Structural Database was determined employing the 'own' library written by Jens Meiler. The local copy under 'small_molecule/own' was modified to include specific 'WriteInMDLFile' functions in the molecule and ensemble classes. A total set of 6,418 small molecules was identified being in both databases. This approach was later replaced by employing CORINA to determine a low-energy conformation for all small molecules in the NMRshiftDB.

histos/

In this directory dihedral histograms were determined for all atom type pairs in the CSD. A list of 38 common atom types in the CSD was compiled. All occurrences for each combination of these 38 atom types with each other were measured for the dihedral angles including them and their substituents. The resulting angles were binned by 10° steps leading to a sparsely populated 38 x 38 x 36 tensor ('dihe_csd_20071029.csv'). Another histogram was generated excluding atoms in rings to remove bias towards common ring dihedrals ('dihe_csd_no_rings_20071029.csv'). This work was utilized in Kaufmann 'Small Molecule Rotamers Enable Simultaneous Optimization of Small Molecule and Protein Degrees of Freedom in RosettaLigand Docking'.

datagen20*/

The carbon chemical shift prediction work presented in chapter V is organized in folders called 'datagen20*' usually including the date of their creation. They include modifications of earlier implementations of the chemical shift code and ANNs testing these modifications.

datagen20070321/

Before the generation of the actual code the ‘nmrshiftdb_csd_part.sdf’ was prepared by fixing missing hydrogens, removing molecules missing bonds or having undetermined atom types. Aromaticity was determined and a list of solvents created to be represented in the BCL. A subset of seventeen uncharged atom types was employed to generate a smaller data set without charged atoms. Executables were created to prepare the small molecule ensemble (BCL), sort out given atom types (BCL), combine the NMRshiftDB and CSD (own), determine the $rmsd$ for individual molecules and the whole ensemble (own), and analyze the occurrence of given fragments in the small molecule ensemble.

datagen20071011

A new 2D code was introduced to provide more detail how atoms are connected to the atom of interest.

General structure of the 1D-code

The code contains the sum of atom types for a certain amount of spheres around the atom of interest, first all atom types (n per sphere), second only atom types conjugated with the atom of interest (n per sphere). Furthermore the sum of bond types (8 per sphere) in each sphere and the ring closure number (1 per sphere). The last 7 inputs are solvent properties. The output is the chemical shift of the atom of interest.

General structure of the 2D-code

The code differs from the 1D code by providing a matrix combining bond and atom types. This gives number bond types * number atom types per sphere, once for all and once for conjugated

atom types. For the first sphere (atom of interest) only the carbon atom type (2*5 inputs) was given. Ring closure numbers, solvent properties and output are the same like 1D.

ANNs were trained for both codes over a hidden neuron range from eight to 128 and for the full set of atom types and the reduced set of uncharged atom types. The number of spheres around the atom of interest was varied between five, seven, and eight.

datagen20080110

Here outliers above 20ppm were removed and ANNs with 64 hidden neurons trained on the 1D and 2D code. Afterwards, the small molecules were split into molecules having maximal chemical shift deviation per atom above 12ppm and below 12ppm. The smaller set above 12ppm was balanced against the larger set below 12ppm to train the models on these outliers. However, no substantial improvement in overall $rmsd$ was found.

datagen20080110

Building on the earlier results, the cut-off was optimized to 17ppm. Outliers were scrutinized and different atom types of phosphorus tested. The number of hidden neurons was set to 48.

datagen20081205

Codes based only on connectivity and codes implementing conformational descriptors (RDF_Ident, 64 descriptors, and 0.1Å) were created from NMRshiftDB to test the influence of the stereochemical descriptors. Both codes consist of 5 spheres around the atom of interest. Both networks have 48 hidden neurons. Even if the ANNs were still improving, it is clearly visible from Figure 29 that adding stereochemical descriptors improves the overall quality of the ANN.

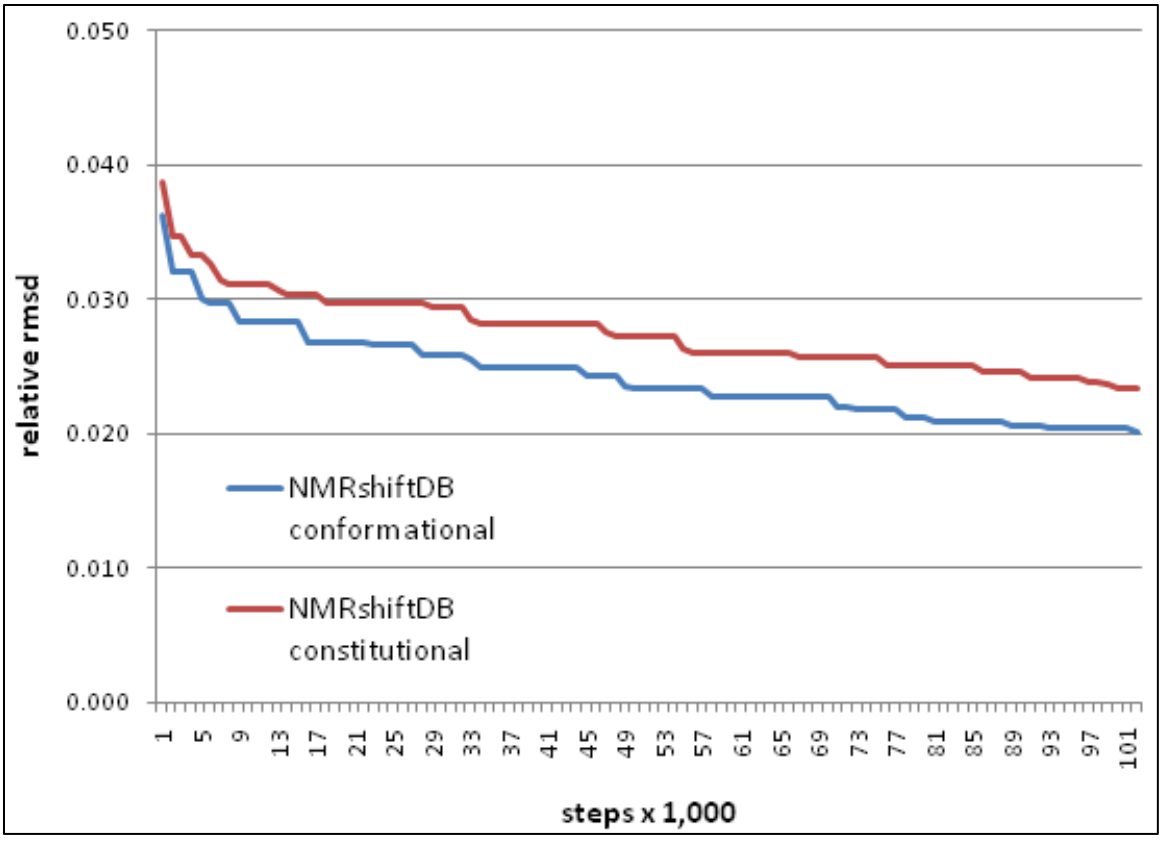


Figure 29: Influence of the stereochemical descriptors on the overall quality of the trained ANNs: ANNs improve after inclusion of the RDF_Ident stereochemical descriptors.

datagen20090225/

To identify problems with the existing code an ANN was trained on random 10% of the NMRshiftDB data (18,607 data points including RDF_Ident stereochemical descriptor) revealing several short-comings of the protocol established so far. Atoms of interest containing silicon in their environment showed large deviations probably due to the small number of molecules with silicon.

The geometry of some molecules was not correct leading to wrongly placed hydrogens and therefore erroneous stereochemical descriptors. One molecule had a carbon chemical shift assigned to a hydrogen atom. The BCL method adding missing hydrogens produced clashes if a

sp3-carbon already had three substituents with 120° which can be fixed by generating low-energy conformations with CORINA.

Several shifts reported in the NMRshiftDB are wrong compared to similar structures in the literature.

It proved problematic to describe quinone-like structures as being aromatic.

datagen20090226

Based on the results so far, a whole workflow was designed aimed at improving the quality of the input files to the ANNs. A new version of NMRshiftDB was downloaded. CORINA was employed to create low-energy conformations and add missing hydrogen atoms. The solvents reported in the NMRshiftDB were mapped to a small set of known solvents in the BCL. Aromaticity was determined through a script created by Kristian Kaufmann.

Eight molecules missing hydrogen atoms were checked revealing steric clashes that apparently were not resolved by CORINA. These molecules were sorted out. Approximately 200 molecules missing charges could be fixed through BCL methods.

A random 1% subset of molecules was checked manually for obvious errors: (i) hydroxyl groups are sometimes not in plane with benzene, (ii) conjugated systems are sometimes not planar, (iii) two benzenes connected by a sp3-hybridized carbon have overlapping π-clouds, (iv) carbon-sulfur bonds in one ring were too long, (v) para-substituted pyrine between two benzenes is not planar, and (vi) two nitrogen neighbors should be sp3-hybridized.

After addressing these errors the whole ensemble was randomized (ensemble_random_100pct.sdf). The first 10% of all molecules were set aside as independent, second 10% as monitor, and the remaining 80% as training data set.

datagen20090423

After implementing the atom-based descriptors an ANN was trained based on the conformational code (2D/3DAutoCorr, RDF) on 10% of the data (18,306 data points). The input sensitivity analysis showed that the lone-pair electronegativity descriptors carried the least information. Checking the code revealed an error in the implementation which was fixed subsequentially.

datagen20090519

Repeating the analysis with the fixed LpEN descriptors generated the input sensitivity data reported in chapter V.

datagen20090604

A descriptor optimization experiment was performed based on input sensitivity analysis. Models with 1,525, 1,166, 701, and 509 descriptors were trained. The results were inconclusive.

datagen20100714

A first implementation of the substituent code was developed in collaboration with Laura Wiley. It included descriptors for atom types, bonds, and ring closure. However, conjugated systems near the atom of interest were not described. Furthermore, the substituents were not ordered according to their bond order and atomic weight.

datagen20110226

After adding the ordering of the substituents and the description of conjugated systems, the final models so far were trained. Each combination of atom type and number of hydrogen atoms got its own subdirectory. The optimal number of neurons was determined at 32. Three and four sphere codes (datagen20110306/) were tested, but for the larger data sets like sp2_c the training of the ANNs remained incomplete due to the exponential increase of the number of descriptors.

TrainANN.pbs scripts were employed to perform a five-fold cross-validation (cv_0-0 ... cv_0-4) on the piranha cluster in the Center for Structural Biology at Vanderbilt. The first 10% of the data set always represented the independent data set. The monitoring data set was varied from the second to sixth 10%. Taking out approximately 10% of the outliers greatly improved the overall predictions. Different cutoffs were determined for all atom type/#hydrogen atoms combinations. The results can be found in 'reduced_*ppm' subdirectories. Under 'reduced_*ppm_new_nn' new models were trained based on the reduced data sets.

similarity_search/

The clusters of small molecules shown in Figure 5, Figure 13, and Figure 15 are based on determining the largest connected common substructure between two given molecules. An algorithm published by Krissinel in 2004 was implemented in the BCL. However, since this algorithm allows unconnected substructures (bond order 0), it was modified to allow specifying if the substructure should be connected or not. Testing results for correctness and speed can be found under 'similarity_search'. The BCL classes for this algorithm are graph::CommonSubgraphIsomorphismBase and CommonSubgraphIsomorphismWithUserGraph. The application for generating the small molecule distance matrices utilized to generate the clusters is 'apps_release.exe DetermineCSIMAtrix'.

REFERENCES

1. Mueller, R.; Rodriguez, A. L.; Dawson, E. S.; Butkiewicz, M.; Nguyen, T. T.; Oleszkiewicz, S.; Bleckmann, A.; Weaver, C. D.; Lindsley, C. W.; Conn, P. J.; Meiler, J., Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening. *ACS Chem. Neurosci.* **2010,** *1* (4), 288-305.

2. Steinbeck, C.; Kuhn, S., NMRShiftDB - compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **2004,** *65* (19), 2711-7.

3. Rodriguez, A. L.; Grier, M. D.; Jones, C. K.; Herman, E. J.; Kane, A. S.; Smith, R. L.; Williams, R.; Zhou, Y.; Marlo, J. E.; Days, E. L.; Blatt, T. N.; Jadhav, S.; Menon, U. N.; Vinson, P. N.; Rook, J. M.; Stauffer, S. R.; Niswender, C. M.; Lindsley, C. W.; Weaver, C. D.; Conn, P. J., Discovery of Novel Allosteric Modulators of Metabotropic Glutamate Receptor Subtype 5 Reveals Chemical and Functional Diversity and In Vivo Activity in Rat Behavioral Models of Anxiolytic and Antipsychotic Activity. *Mol. Pharmacol.* **2010,** *78* (6), 1105-1123.

4. (a) Engers, D. W.; Niswender, C. M.; Weaver, C. D.; Jadhav, S.; Menon, U. N.; Zamorano, R.; Conn, P. J.; Lindsley, C. W.; Hopkins, C. R., Synthesis and Evaluation of a Series of Heterobiarylamides That Are Centrally Penetrant Metabotropic Glutamate Receptor 4 (mGluR4) Positive Allosteric Modulators (PAMs). *J. Med. Chem.* **2009,** *52* (14), 4115-4118; (b) Niswender, C. M.; Johnson, K. A.; Weaver, C. D.; Jones, C. K.; Xiang, Z.; Luo, Q.; Rodriguez, A. L.; Marlo, J. E.; de Paulis, T.; Thompson, A. D.; Days, E. L.; Nalywajko, T.; Austin, C. A.; Williams, M. B.; Ayala, J. E.; Williams, R.; Lindsley, C. W.; Conn, P. J., Discovery, Characterization, and Antiparkinsonian Effect of Novel Positive Allosteric Modulators of Metabotropic Glutamate Receptor 4. *Mol. Pharmacol.* **2008,** *74* (5), 1345-1358; (c) Niswender, C. M.; Lebois, E. P.; Luo, Q.; Kim, K.; Muchalski, H.; Yin, H.; Conn, P. J.; Lindsley, C. W., Positive allosteric modulators of the metabotropic glutamate receptor subtype 4 (mGluR4): Part I. Discovery of pyrazolo[3,4-d]pyrimidines as novel mGluR4 positive allosteric modulators. *Bioorg. Med. Chem. Lett.* **2008,** *18* (20), 5626-5630; (d) Williams, R.; Niswender, C. M.; Luo, Q.; Le, U.; Conn, P. J.; Lindsley, C. W., Positive allosteric modulators of the metabotropic glutamate receptor subtype 4 (mGluR4). Part II: Challenges in hit-to-lead. *Bioorg. Med. Chem. Lett.* **2009,** *19* (3), 962-966.

5. (a) ADRIANA. http://www.molecular-networks.com/ (accessed November 16, 2009); (b) Computerchemie, M. N. G.; Schwab, C. H.; Gasteiger, J., *ADRIANA.Code; Algorithms for the Encoding of Molecular Structures; Version 2.0; Program Description*. 2006.

6. (a) 3D Structure Generator CORINA, developed and distributed by Molecular Networks GmbH, Erlangen, Germany. www.molecular-networks.com (accessed November 16, 2009); (b) Gasteiger, J.; Rudolph, C.; Sadowski, J., Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Meth.* **1990,** *3* (6c), 537-47.

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

8. Walters, W. P.; Murcko, M. A., Prediction of 'drug-likeness'. *Adv. Drug Delivery Rev.* **2002,** *54* (3), 255-271.

9. Tetko, I. V.; Kovalishyn, V. V.; Livingstone, D. J., Volume Learning Algorithm Artificial Neural Networks for 3D QSAR Studies. *J. Med. Chem.* **2001,** *44* (15), 2411-2420.

10. (a) Riedmiller, M.; Braun, H., A direct adaptive method for faster backpropagation learning: The Rprop algorithm. *Proc. - Int. Conf. Neural Networks* **1993**, 586-591; (b) Riedmiller, M.; Braun, H. In *Rprop - A Fast Adaptive Learning Algorithm.*, International Symposium on Computer and Information Science VII, 1992.

11. Conn, P. J.; Pin, J.-P., Pharmacology and Functions of Metabotropic Glutamate Receptors. *Annu. Rev. Pharmacol. Toxicol.* **1997,** *37* (1), 205-237.

12. (a) Conn, P. J.; Lindsley, C. W.; Jones, C. K., Activation of metabotropic glutamate receptors as a novel approach for the treatment of schizophrenia. *Trends Pharmacol. Sci.* **2009,** *30* (1), 25-31; (b) Gasparini, F.; Kuhn, R.; Pin, J.-P., Allosteric modulators of group I metabotropic glutamate receptors: novel subtype-selective ligands and therapeutic perspectives. *Curr. Opin. Pharmacol.* **2002,** *2* (1), 43-49.

13. (a) Marino, M.; Valenti, O.; Conn, P. J., Glutamate receptors and Parkinson's disease : opportunities for intervention. *Drugs Aging* **2003,** *20* (5), 377-97; (b) Marino, M. J.; Williams, D. L.; O'Brien, J. A.; Valenti, O.; McDonald, T. P.; Clements, M. K.; Wang, R.; DiLella, A. G.; Hess, J. F.; Kinney, G. G.; Conn, P. J., Allosteric modulation of group III metabotropic glutamate receptor 4: A potential approach to Parkinson's disease treatment. *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100* (23), 13668-13673.

14. Dölen, G.; Bear, M. F., Role for metabotropic glutamate receptor 5 (mGluR5) in the pathogenesis of fragile X syndrome. *J. Physiol.* **2008,** *586* (6), 1503-1508.

15. Pin, J.-P.; De Colle, C.; Bessis, A.-S.; Acher, F., New perspectives for the development of selective metabotropic glutamate receptor ligands. *Eur. J. Pharmacol.* **1999,** *375* (1-3), 277-294.

16. Conn, P. J.; Christopoulos, A.; Lindsley, C. W., Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discov.* **2009,** *8* (1), 41-54.

17. Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R., Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001,** *41* (5), 1295-1300.

18. Jorissen, R. N.; Gilson, M. K., Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005,** *36* (32), no-no.

19. Noeske, T.; Jirgensons, A.; Starchenkovs, I.; Renner, S.; Jaunzeme, I.; Trifanova, D.; Hechenberger, M.; Bauer, T.; Kauss, V.; Parsons, C. G.; Schneider, G.; Weil, T., Virtual screening for selective allosteric mGluR1 antagonists and structure-activity relationship investigations for coumarine derivatives. *ChemMedChem* **2007,** *2* (12), 1763-1773.

20. Noeske, T.; Trifanova, D.; Kauss, V.; Renner, S.; Parsons, C. G.; Schneider, G.; Weil, T., Synergism of virtual screening and medicinal chemistry: Identification and optimization of allosteric antagonists of metabotropic glutamate receptor 1. *Bioorg. Med. Chem.* **2009,** *17* (15), 5708-5715.

21. 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.; Williams, D. L., Jr., A novel selective allosteric modulator potentiates the activity of native metabotropic glutamate receptor subtype 5 in rat forebrain. *J. Pharmacol. Exp. Ther.* **2004,** *309* (2), 568-77.

22. 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.; Williams, D. L., Jr., A family of highly selective allosteric modulators of the metabotropic glutamate receptor subtype 5. *Mol. Pharmacol.* **2003,** *64* (3), 731-40.

23. 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.; Williams, D. L., Jr., 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.* **2005,** *313* (1), 199-206.

24. 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.; Huff, J. R., 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.* **2004,** *47* (24), 5825-8.

25. Chen, Y.; Nong, Y.; Goudet, C.; Hemstapat, K.; de Paulis, T.; Pin, J.-P.; Conn, P. J., Interaction of Novel Positive Allosteric Modulators of Metabotropic Glutamate Receptor 5 with the Negative Allosteric Antagonist Site Is Required for Potentiation of Receptor Responses. *Mol. Pharmacol.* **2007,** *71* (5), 1389-1398.

26. Valenti, O.; Marino, M. J.; Wittmann, M.; Lis, E.; DiLella, A. G.; Kinney, G. G.; Conn, P. J., Group III Metabotropic Glutamate Receptor-Mediated Modulation of the Striatopallidal Synapse. *J. Neurosci.* **2003,** *23* (18), 7218-7226.

27. Valenti, O.; Mannaioni, G.; Seabrook, G. R.; Conn, P. J.; Marino, M. J., Group III Metabotropic Glutamate-Receptor-Mediated Modulation of Excitatory Transmission in Rodent Substantia Nigra Pars Compacta Dopamine Neurons. *J. Pharmacol. Exp. Ther.* **2005,** *313* (3), 1296-1304.

28. Maj, M.; Bruno, V.; Dragic, Z.; Yamamoto, R.; Battaglia, G.; Inderbitzin, W.; Stoehr, N.; Stein, T.; Gasparini, F.; Vranesic, I.; Kuhn, R.; Nicoletti, F.; Flor, P. J., (-)-PHCCC, a positive allosteric modulator of mGluR4: characterization, mechanism of action, and neuroprotection. *Neuropharmacology* **2003,** *45* (7), 895-906.

29. Mathiesen, J. M.; Svendsen, N.; Bräuner-Osborne, H.; Thomsen, C.; Ramirez, M. T., Positive allosteric modulation of the human metabotropic glutamate receptor 4 (hmGluR4) by SIB-1893 and MPEP. *Br. J. Pharmacol.* **2003,** *138* (6), 1026-1030.

30. Varney, M. A.; Cosford, N. D. P.; Jachec, C.; Rao, S. P.; Sacaan, A.; Lin, F.-F.; Bleicher, L.; Santori, E. M.; Flor, P. J.; Allgeier, H.; Gasparini, F.; Kuhn, R.; Hess, S. D.; Veliçelebi, G.; Johnson, E. C., SIB-1757 and SIB-1893: Selective, Noncompetitive Antagonists of Metabotropic Glutamate Receptor Type 5. *J. Pharmacol. Exp. Ther.* **1999,** *290* (1), 170-181.

31. Gasparini, F.; Lingenhöhl, K.; Stoehr, N.; Flor, P. J.; Heinrich, M.; Vranesic, I.; Biollaz, M.; Allgeier, H.; Heckendorn, R.; Urwyler, S.; Varney, M. A.; Johnson, E. C.; Hess, S. D.; Rao, S. P.; Sacaan, A. I.; Santori, E. M.; Veliçelebi, G.; Kuhn, R., 2-Methyl-6-(phenylethynyl)-pyridine (MPEP), a potent, selective and systemically active mGlu5 receptor antagonist. *Neuropharmacology* **1999,** *38* (10), 1493-1503.

32. Gasparini, F.; Floersheim, P.; Flor, P. J.; Heinrich, M.; Inderbitzin, W.; Ott, D.; Pagano, A.; Stierlin, C.; Stoehr, N.; Vranesic, I.; Kuhn, R., Discovery and characterization of non-competitive antagonists of group I metabotropic glutamate receptors. *Farmaco* **2001,** *56* (1-2), 95-99.

33. Gasparini, F.; Andres, H.; Flor, P. J.; Heinrich, M.; Inderbitzin, W.; Lingenhöhl, K.; Müller, H.; Munk, V. C.; Omilusik, K.; Stierlin, C.; Stoehr, N.; Vranesic, I.; Kuhn, R., [3H]-M-MPEP, a Potent, Subtype-Selective Radioligand for the Metabotropic Glutamate Receptor Subtype 5. *Bioorg. Med. Chem. Lett.* **2002,** *12* (3), 407-409.

34. Yan, Q. J.; Rammal, M.; Tranfaglia, M.; Bauchwitz, R. P., Suppression of two major Fragile X Syndrome mouse model phenotypes by the mGluR5 antagonist MPEP. *Neuropharmacology* **2005,** *49* (7), 1053-1066.

35. Rodriguez, A. L.; Nong, Y.; Sekaran, N. K.; Alagille, D.; Tamagnan, G. D.; Conn, P. J., 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.* **2005,** *68* (6), 1793-1802.

36. Rodriguez, A. L.; Williams, R.; Zhou, Y.; Lindsley, S. R.; Le, U.; Grier, M. D.; David Weaver, C.; Conn, P. J.; Lindsley, C. W., Discovery and SAR of novel mGluR5 non-competitive antagonists not based on an MPEP chemotype. *Bioorg. Med. Chem. Lett.* **2009,** *19* (12), 3209-3213.

37. Meiler, J.; Sanli, E.; Junker, J.; Meusinger, R.; Lindel, T.; Will, M.; Maier, W.; Kock, M., Validation of structural proposals by substructure analysis and 13C NMR chemical shift prediction. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (2), 241-8.

38. Elyashberg, M. E.; Blinov, K. A.; Williams, A. J., The application of empirical methods of 13C NMR chemical shift prediction as a filter for determining possible relative stereochemistry. *Magn. Reson. Chem.* **2009,** *47* (4), 333-341.

39. Schnackenberg, L. K.; Beger, R. D., Whole-Molecule Calculation of Log P Based on Molar Volume, Hydrogen Bonds, and Simulated 13C NMR Spectra. *J. Chem. Inf. Model.* **2005,** *45* (2), 360-365.

40. Kalchhauser, H.; Robien, W., CSEARCH: a computer program for identification of organic compounds and fully automated assignment of carbon-13 nuclear magnetic resonance spectra. *J. Chem. Inf. Comput. Sci.* **1985,** *25* (2), 103-108.

41. Bremser, W., HOSE - a novel substructure code. *Anal. Chim. Acta* **1978,** *103* (4), 355-365.

42. Tusar, M.; Tusar, L.; Bohanec, S.; Zupan, J., Proton and carbon-13 NMR spectra simulation. *J. Chem. Inf. Comput. Sci.* **1992,** *32* (4), 299-303.

43. Satoh, H.; Koshino, H.; Uzawa, J.; Nakata, T., CAST/CNMR: highly accurate 13C NMR chemical shift prediction system considering stereochemistry. *Tetrahedron* **2003,** *59* (25), 4539-4547.

44. Satoh, H.; Koshino, H.; Uno, T.; Koichi, S.; Iwata, S.; Nakata, T., Effective consideration of ring structures in CAST/CNMR for highly accurate 13C NMR chemical shift prediction. *Tetrahedron* **2005,** *61* (31), 7431-7437.

45. Blinov, K. A.; Smurnyy, Y. D.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J., Development of a fast and accurate method of 13C NMR chemical shift prediction. *Chemom. Intell. Lab. Syst.* **2009,** *97* (1), 91-97.

46. Meiler, J.; Meusinger, R.; Will, M., Fast determination of 13C NMR chemical shifts using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **2000,** *40* (5), 1169-76.

47. (a) Conn, P. J.; Pin, J. P., Pharmacology and functions of metabotropic glutamate receptors. *Annu. Rev. Pharmacol. Toxicol.* **1997,** *37*, 205-37; (b) Pin, J. P.; Duvoisin, R., The metabotropic glutamate receptors: Structure and functions. *Neuropharmacology* **1995,** *34* (1), 1-26.

48. Palucha, A.; Pilc, A., On the role of metabotropic glutamate receptors in the mechanisms of action of antidepressants. *Pol. J. Pharmacol.* **2002,** *54* (6), 581-6.

49. (a) Chojnacka-Wojcik, E.; Klodzinska, A.; Pilc, A., Glutamate receptor ligands as anxiolytics. *Curr. Opin. Investig. Drugs* **2001,** *2* (8), 1112-9; (b) Pilc, A., LY-354740 (Eli Lilly). *IDrugs* **2003,** *6* (1), 66-71.

50. (a) 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.; Conn, P. J., mGluR5 Positive Allosteric Modulators Facilitate both Hippocampal LTP and LTD and Enhance Spatial Learning. *Neuropsychopharmacology* **2009**; (b) Chavez-Noriega, L. E.; Schaffhauser, H.; Campbell, U. C., Metabotropic glutamate receptors: potential drug targets for the treatment of schizophrenia. *Curr. Drug Targets CNS Neurol. Disord.* **2002,** *1* (3), 261-81; (c) Marino, M. J.; Conn, P. J., Direct and indirect modulation of the N-methyl D-aspartate receptor. *Curr. Drug Targets CNS Neurol. Disord.* **2002,** *1* (1), 1-16.

51. Varney, M. A.; Gereau, R. W. t., Metabotropic glutamate receptor involvement in models of acute and persistent pain: prospects for the development of novel analgesics. *Curr. Drug Targets CNS Neurol. Disord.* **2002,** *1* (3), 283-96.

52. Doherty, J.; Dingledine, R., The roles of metabotropic glutamate receptors in seizures and epilepsy. *Curr. Drug Targets CNS Neurol. Disord.* **2002,** *1* (3), 251-60.

53. Wisniewski, K.; Car, H., (S)-3,5-DHPG: a review. *CNS Drug Rev.* **2002,** *8* (1), 101-16.

54. Marino, M. J.; Conn, J. P., Modulation of the basal ganglia by metabotropic glutamate receptors: potential for novel therapeutics. *Curr. Drug Targets CNS Neurol. Disord.* **2002,** *1* (3), 239-50.

55. (a) Campbell, U. C.; Lalwani, K.; Hernandez, L.; Kinney, G. G.; Conn, P. J.; Bristow, L. J., The mGluR5 antagonist 2-methyl-6-(phenylethynyl)-pyridine (MPEP) potentiates PCP-induced cognitive deficits in rats. *Psychopharmacology (Berl)* **2004,** *175* (3), 310-8; (b) Henry, S. A.; Lehmann-Masten, V.; Gasparini, F.; Geyer, M. A.; Markou, A., The mGluR5 antagonist MPEP, but not the mGluR2/3 agonist LY314582, augments PCP effects on prepulse inhibition and locomotor activity. *Neuropharmacology* **2002,** *43* (8), 1199-1209; (c) Kinney, G. G.; Burno, M.; Campbell, U. C.; Hernandez, L. M.; Rodriguez, D.; Bristow, L. J.; Conn, P. J., Metabotropic glutamate subtype 5 receptors modulate locomotor activity and sensorimotor gating in rodents. *J. Pharmacol. Exp. Ther.* **2003,** *306* (1), 116-23.

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

57. de Paulis, T.; Hemstapat, K.; Chen, Y.; Zhang, Y.; Saleh, S.; Alagille, D.; Baldwin, R. M.; Tamagnan, G. D.; Conn, P. J., 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.* **2006,** *49* (11), 3332-44.

58. (a) Bessis, A.-S.; Bonnet, B.; Le Poul, E.; Rocher, J.-P.; Epping-Jordan, M. Preparation of piperidine derivatives as modulators of metabotropic glutamate receptors (mGluR5) WO 044797, 2005; (b) Bugada, P.; Gagliardi, S.; Le Poul, E.; Mutel, V.; Palombi, G.; Rocher, J.-P. 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, 2006; (c) Engers, D. W.; Rodriguez, A. L.; Williams, R.; Hammond, A. S.; Venable, D.; Oluwatola, O.; Sulikowski, G. A.; Conn, P. J.; Lindsley, C. W., Synthesis, SAR and Unanticipated Pharmacological Profiles of Analogues of the mGluR5 Ago-potentiator ADX-47273. *ChemMedChem* **2009,** *4* (4), 505-511; (d) 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.; Marquis, K. L., 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.* **2008,** *327* (3), 827-39.

59. Chen, Y.; Goudet, C.; Pin, J. P.; Conn, P. J., 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.* **2008,** *73* (3), 909-18.

60. Sharma, S.; Rodriguez, A. L.; Conn, P. J.; Lindsley, C. W., 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.* **2008,** *18* (14), 4098-101.

61. (a) Carnero, A., High throughput screening in drug discovery. *Clin. Transl. Oncol.* **2006,** *8* (7), 482-90; (b) Liu, B.; Li, S.; Hu, J., Technological advances in high-throughput screening. *Am. J. Pharmacogenomics* **2004,** *4* (4), 263-76.

62. Hodder, P.; Mull, R.; Cassaday, J.; Berry, K.; Strulovici, B., Miniaturization of intracellular calcium functional assays to 1536-well plate format using a fluorometric imaging plate reader. *J. Biomol. Screen.* **2004,** *9* (5), 417-26.

63. Gilchrist, M. A., 2nd; Cacace, A.; Harden, D. G., Characterization of the 5-HT2b receptor in evaluation of aequorin detection of calcium mobilization for miniaturized GPCR high-throughput screening. *J. Biomol. Screen.* **2008,** *13* (6), 486-93.

64. Posner, B. A., High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics. *Curr. Opin. Drug. Discov. Devel.* **2005,** *8* (4), 487-94.

65. (a) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D., Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem. Rev.* **2002,** *102* (3), 783-812; (b) Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. WILEY - VCH: 2000; Vol. 11.

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

67. (a) Bleckmann, A.; Meiler, J., Epothilones: Quantitative Structure Activity Relations Studied by Support Vector Machines and Artificial Neural Networks. *QSAR Comb. Sci.* **2003,** *22* (7), 719-721; (b) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A., Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008,** *48* (9), 1733-46.

68. (a) Cramer, R. D., 3rd; Patterson, D. E.; Bunce, J. D., Recent advances in comparative molecular field analysis (CoMFA). *Prog. Clin. Biol. Res.* **1989,** *291*, 161-5; (b) Cramer, R. D.; Patterson, D. E.; Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988,** *110* (18), 5959 - 5967.

69. Klebe, G.; Abraham, U.; Mietzner, T., Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994,** *37* (24), 4130-46.

70. (a) Krasowski, M. D.; Siam, M. G.; Iyer, M.; Ekins, S., Molecular Similarity Methods for Predicting Cross-Reactivity With Therapeutic Drug Monitoring Immunoassays. *Ther. Drug Monit.* **2009**; (b) Wild, D. J.; Blankley, C. J., Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.* **2000,** *40* (1), 155-62.

71. (a) Heritage, T. W.; Hurst, T., HQSAR - a highly predictive QSAR technique based on molecular holograms. *Book of Abstracts, 214th ACS National Meeting, Las Vegas, NV, September 7-11* **1997**, COMP-080; (b) Moda, T. L.; Montanari, C. A.; Andricopulo, A. D., Hologram QSAR model for the prediction of human oral bioavailability. *Bioorg. Med. Chem.* **2007,** *15* (24), 7738-7745; (c) Salum, L. B.; Andricopulo, A. D., Fragment-based QSAR: perspectives in drug design. *Mol. Divers.* **2009**.

72. Waller, C. L., 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.* **2004,** *44* (2), 758-65.

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

74. (a) Brown, R. D.; Martin, Y. C., Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J Chem Inf Comput Sci* **1996,** *36* (3), 572-84; (b) Jenkins, J. L.; Glick, M.; Davies, J. W., A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004,** *47* (25), 6144-6159.

75. Nettles, J. H.; Jenkins, J. L.; Williams, C.; Clark, A. M.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M., Flexible 3D pharmacophores as descriptors of dynamic biological space. *J. Mol. Graph. Model.* **2007,** *26* (3), 622-33.

76. Marrero-Ponce, Y., 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.* **2004,** *44* (6), 2010-26.

77. Gonzalez, M. P.; Puente, M.; Fall, Y.; Gomez, G., In silico studies using Radial Distribution Function approach for predicting affinity of 1 alpha,25-dihydroxyvitamin D(3) analogues for Vitamin D receptor. *Steroids* **2006,** *71* (6), 510-27.

78. Morales, A. H.; Cabrera Perez, M. A.; Gonzalez, M. P., A radial-distribution-function approach for predicting rodent carcinogenicity. *J. Mol. Model.* **2006,** *12* (6), 769-80.

79. Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J., Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (6), 1205-13.

80. Hristozov, D. P.; Oprea, T. I.; Gasteiger, J., Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput. Aided Mol. Des.* **2007,** *21* (10-11), 617-40.

81. Hristozov, D.; Oprea, T. I.; Gasteiger, J., Ligand-based virtual screening by novelty detection with self-organizing maps. *J. Chem. Inf. Model.* **2007,** *47* (6), 2044-62.

82. Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J., 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.* **1996,** *10* (6), 521-34.

83. Holzgrabe, U.; Wagener, M.; Gasteiger, J., Comparison of structurally different allosteric modulators of muscarinic receptors by self-organizing neural networks. *J. Mol. Graph.* **1996,** *14* (4), 185-93, 217-21.

84. Teckentrup, A.; Briem, H.; Gasteiger, J., Mining high-throughput screening data of combinatorial libraries: development of a filter to distinguish hits from nonhits. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (2), 626-34.

85. Zupan, J.; Gasteiger, J., *Neural Networks for Chemists*. VCH Verlagsgesellschaft mbH: Weinheim, 1993.

86. Burton, J.; Ijjaali, I.; Barberan, O.; Petitet, F.; Vercauteren, D. P.; Michel, A., Recursive Partitioning for the Prediction of Cytochromes P450 2D6 and 1A2 Inhibition: Importance of the Quality of the Dataset. *J. Med. Chem.* **2006,** *49* (21), 6231-6240.

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

88. Hecht, D.; Fogel, G., High-Throughput Ligand Screening via Preclustering and Evolved Neural Networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2007,** *4* (3), 476-484.

89. Butkiewicz, M.; Mueller, R.; Selic, D.; Dawson, E.; Meiler, J., Application of Machine Learning Approaches on Quantitative Structure Activity Relationships. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Wiese, K. C., Ed. Nashville, 2009.

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

91. Gedeck, P.; Rohde, B.; Bartels, C., QSAR - how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **2006,** *46* (5), 1924-36.

92. (a) Brown, R. D.; Martin, Y. C., Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996,** *36* (3), 572-84; (b) Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006,** *11* (23-24), 1046-53.

93. Wolfram Research, I., *Mathematica*. Version 7 ed.; Wolfram Research, Inc.: Champaigne, Illinois, 2008.

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

95. Marino, M. J.; Conn, P. J., Glutamate-based therapeutic approaches: allosteric modulators of metabotropic glutamate receptors. *Curr. Opin. Pharmacol.* **2006,** *6* (1), 98-102.

96. Marino, M. J.; Williams, D. L., Jr.; O'Brien, J. A.; Valenti, O.; McDonald, T. P.; Clements, M. K.; Wang, R.; DiLella, A. G.; Hess, J. F.; Kinney, G. G.; Conn, P. J., Allosteric modulation of group III metabotropic glutamate receptor 4: a potential approach to Parkinson's disease treatment. *Proc Natl Acad Sci U S A* **2003,** *100* (23), 13668-73.

97. Gregory, K. J.; Dong, E. N.; Meiler, J.; Conn, P. J., Allosteric Modulation of Metabotropic Glutamate Receptors: Structural Insights and Therapeutic Potential. *Neuropharmacology* **2010**.

98. (a) Lindberg, J. S.; Culleton, B.; Wong, G.; Borah, M. F.; Clark, R. V.; Shapiro, W. B.; Roger, S. D.; Husserl, F. E.; Klassen, P. S.; Guo, M. D.; Albizem, M. B.; Coburn, J. W., Cinacalcet HCl, an Oral Calcimimetic Agent for the Treatment of Secondary Hyperparathyroidism in Hemodialysis and Peritoneal Dialysis: A Randomized, Double-Blind, Multicenter Study. *J Am Soc Nephrol* **2005,** *16* (3), 800-807; (b) Dorr, P.; Westby, M.; Dobbs, S.; Griffin, P.; Irvine, B.; Macartney, M.; Mori, J.; Rickett, G.; Smith-Burchnell, C.; Napier, C.; Webster, R.; Armour, D.; Price, D.; Stammen, B.; Wood, A.; Perros, M., Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine Receptor CCR5 with Broad-Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity. *Antimicrob. Agents Chemother.* **2005,** *49* (11), 4721-4732.

99. Kola, I.; Landis, J., Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* **2004,** *3* (8), 711-716.

100. (a) Bolea, C. Novel thiazoles derivatives and their use as positive allosteric modulators of metabotropic glutamate receptors and preparation. 2010; (b) Bolea, C.; Celanire, S. Preparation of novel heteroaromatic derivatives and their use as positive allosteric modulators of metabotropic glutamate receptors. 2009.

101. (a) Malherbe, P.; Kratochwil, N.; Mühlemann, A.; Zenner, M.-T.; Fischer, C.; Stahl, M.; Gerber, P. R.; Jaeschke, G.; Porter, R. H. P., Comparison of the binding pockets of two chemically unrelated allosteric antagonists of the mGlu5 receptor and identification of crucial residues involved in the inverse agonism of MPEP. *J. Neurochem.* **2006,** *98* (2), 601-615; (b) Malherbe, P.; Kratochwil, N.; Zenner, M.-T.; Piussi, J.; Diener, C.; Kratzeisen, C.; Fischer, C.; Porter, R. H. P., Mutational Analysis and Molecular Modeling of the Binding Pocket of the Metabotropic Glutamate 5 Receptor Negative Modulator 2-Methyl-6-(phenylethynyl)-pyridine. *Mol. Pharmacol.* **2003,** *64* (4), 823-832; (c) Pagano, A.; Rüegg, D.; Litschig, S.; Stoehr, N.; Stierlin, C.; Heinrich, M.; Floersheim, P.; Prezèau, L.; Carroll, F.; Pin, J.-P.; Cambria, A.; Vranesic, I.; Flor, P. J.; Gasparini, F.; Kuhn, R., The Non-competitive Antagonists 2-Methyl-6-(phenylethynyl)pyridine and 7-Hydroxyiminocyclopropan[b]chromen-1a-carboxylic Acid Ethyl Ester Interact with Overlapping Binding Pockets in the Transmembrane Region of Group I Metabotropic Glutamate Receptors. *J. Biol. Chem.* **2000,** *275* (43), 33750-33758.

102. (a) Staerk, D.; Chapagain, B. P.; Lindin, T.; Wiesman, Z.; Jaroszewski, J. W., Structural analysis of complex saponins of Balanites aegyptiaca by 800 MHz 1H NMR spectroscopy. *Magn. Reson. Chem.* **2006,** *44* (10), 923-8; (b) dos Santos, C. C.; Sousa Lima, M. A.; Braz-Filho, R.; de Simone, C. A.; Silveira, E. R., NMR assignments and X-ray diffraction spectra for two unusual kaurene diterpenes from Erythroxylum barbatum. *Magn. Reson. Chem.* **2005,** *43* (12), 1012-5.

103. Mulholland, D. A.; Langlois, A.; Randrianarivelojosia, M.; Derat, E.; Nuzillard, J. M., The structural elucidation of a novel iridoid derivative from Tachiadenus longiflorus (Gentianaceae) using the LSD programme and quantum chemical computations. *Phytochem. Anal.* **2006,** *17* (2), 87-90.

104. Cheeseman, J. R.; Frisch, A., Predicting Magnetic Properties with ChemDraw and Gaussian. *Gaussian, Inc.* **2000**.

105. Bagno, A.; Saielli, G., Computational NMR spectroscopy: reversing the information flow. *Theor. Chem. Acc.* **2007,** *117* (5), 603-619.

106. Perez, M.; Peakman, T. M.; Alex, A.; Higginson, P. D.; Mitchell, J. C.; Snowden, M. J.; Morao, I., Accuracy vs time dilemma on the prediction of NMR chemical shifts: a case study (chloropyrimidines). *J. Org. Chem.* **2006,** *71* (8), 3103-10.

107. Cimino, P.; Gomez-Paloma, L.; Duca, D.; Riccio, R.; Bifulco, G., Comparison of different theory models and basis sets in the calculation of 13C NMR chemical shifts of natural products. *Magn. Reson. Chem.* **2004,** *42 Spec no*, S26-33.

108. Perdue, E. M.; Hertkorn, N.; Kettrup, A., Substitution Patterns in Aromatic Rings by Increment Analysis. Model Development and Application to Natural Organic Matter. *Anal. Chem.* **2007,** *79* (3), 1010-1021.

109. Meiler, J.; Maier, W.; Will, M.; Meusinger, R., Using neural networks for (13)c NMR chemical shift prediction-comparison with traditional methods. *J. Magn. Reson.* **2002,** *157* (2), 242-52.

110. (a) Zupan, J.; Gasteiger, J., *Neural networks for chemists : an introduction*. VCH: Weinheim ; New York, 1993; p xix, 305 p; (b) Meusinger, R.; Himmelreich, U., Neural networks and genetic algorithms applications in nuclear magnetic resonance spectroscopy. *Data Handl. Sci. Technol.* **2003,** *23*, 281-321; (c) *Handbook of Chemoinformatics*. Wiley-VCH Verlag GmbH: Weinheim, 2003; p 1295-1299.

111. Kaur, J.; Brar, A. S., An approach to predict the 13C NMR chemical shifts of acrylonitrile copolymers using artificial neural network. *Eur. Polym. J.* **2007,** *43* (1), 156-163.

112. Clouser, D. L.; Jurs, P. C., Simulation of the 13C nuclear magnetic resonance spectra of trisaccharides using multiple linear regression analysis and neural networks. *Carbohydr. Res.* **1995,** *271* (1), 65-77.

113. Le Bret, C., A general 13C NMR spectrum predictor using data mining techniques. *SAR QSAR Environ. Res.* **2000,** *11* (3-4), 211-34.

114. Blinov, K. A.; Smurnyy, Y. D.; Elyashberg, M. E.; Churanova, T. S.; Kvasha, M.; Steinbeck, C.; Lefebvre, B. A.; Williams, A. J., Performance Validation of Neural Network Based 13C NMR Prediction Using a Publicly Available Data Source. *J. Chem. Inf. Model.* **2008,** *48* (3), 550-555.

115. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity--a rapid access to atomic charges. *Tetrahedron* **1980,** *36* (22), 3219-3228.

116. (a) Hinze, J.; Jaffe, H. H., Electronegativity. I. Orbital Electronegativity of Neutral Atoms. *J. Am. Chem. Soc.* **1962,** *84* (4), 540-546; (b) Hinze, J.; Jaffe, H. H., Electronegativity. IV. Orbital Electronegativities of the Neutral Atoms of the Periods Three A and Four A and of Positive Ions of Peridos One and Two. *J. Phys. Chem.* **1963,** *67* (7), 1501-1506.

117. Gasteiger, J.; Marsili, M., Prediction of proton magnetic resonance shifts: The dependence on hydrogen charges obtained by iterative partial equalization of orbital electronegativity. *Org. Magn. Reson.* **1981,** *15* (4), 353-360.

118. Streitwieser, A., *Molecular Orbital Theory for Organic Chemists*. John Wiley & Sons, Inc.: New York - London, 1961; p 489.

119. Marsili, M.; Gasteiger, J., Pi Charge Distribution from Molecular Topology and pi Orbital Electronegativity. *Croat. Chem. Acta* **1980,** *53* (4), 601-614.

120. Anastasiadis, A. D.; Magoulas, G. D.; Vrahatis, M. N., New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing* **2005,** *64*, 253-270.

121. Verpoorte, R., Assignment of ^{13}C-NMR spectra of strychnine and brucine. *J. Pharm. Sc.* **1980,** *69* (7), 865-867.

122. Xia, X.-K.; Huang, H.-R.; She, Z.-G.; Shao, C.-L.; Liu, F.; Cai, X.-L.; Vrijmoed, L. L. P.; Lin, Y.-C., ^{1}H and ^{13}C NMR assignments for five anthraquinones from the mangrove endophytic fungus Halorosellinia sp. (No. 1403). *Magn. Reson. Chem.* **2007,** *45* (11), 1006-1009.

123. Qi, S.-H.; Zhang, S.; Wang, Y.-F.; Li, M.-Y., Complete ^{1}H and ^{13}C NMR assignments of three new polyhydroxylated sterols from the South China Sea gorgonian *Subergorgia suberosa*. *Magn. Reson. Chem.* **2007,** *45* (12), 1088-1091.

124. Kalinowski, H.-O., Berger, St., Braun, S., Carbon-13 NMR Spectroscopy. **1988**, 792.

125. Abraham, R. J.; Smith, P. E., Charge calculations in molecular mechanics IV: A general method for conjugated systems. *J. Comput. Chem.* **1988,** *9* (4), 288-297.

126. Gasteiger, J.; Hutchings, M. G., Empirical Models of Substituent Polarisability and their Application to Stabilisation Effects in Positively Charged Species. *Tetrahedron* **1983,** *24*, 2537-2540.