

COMPLYING AND CURATING PUBLIC BIOASSAY DATA FOR CHEMICAL
TOXICITY AND ANXIETY DRUG DISCOVERY STUDIES

By

ABENA BOISON

A thesis submitted to the Graduate School – Camden

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Sciences

Graduate Program in Chemistry

written under the direction of

Dr. Hao Zhu

and approved by

Dr. Hao Zhu

Dr. Georgia Arbuckle-Keil

Dr. Jinglin Fu

Camden, New Jersey October 2014

ABSTRACT OF THE THESIS

COMPLYING AND CURATING PUBLIC BIOASSAY DATA FOR CHEMICAL
TOXICITY AND ANXIETY DRUG DISCOVERY STUDIES

BY ABENA BOISON

Thesis Director. Dr. Hao Zhu

Recent investigations suggest that ligands such as steroids inhibit the binding of [^{35}S] t-butylbicyclophosphorothionate ([^{35}S] TBPS) to the convulsant site in the aminobutyric acid type A (GABA_A) receptor complex. Currently, most interest is centered on ligands with [^{35}S] TBPS displacement properties. Ligands binding to the GABA_A receptor, block GABA-gated chloride ion flux in a non-competitive manner, resulting in convulsions. Traditionally, [^{35}S] TBPS inhibition studies are measured using animal tests. Testing compounds, using rat tests, for potentially new ligands are costly and time-consuming. Therefore, developing computational models to predict potential [^{35}S] TBPS displacement could provide many opportunities for the discovery and development of new ligands acting on the GABA_A receptor convulsant site, resulting in the preventions of convulsions.

In this study, Quantitative Structure Activity Relationship (QSAR) approaches were used to develop several computational models for a series of novel and diverse types of compounds (steroids derivatives, Arylsulfonyl derivatives and Propofol analogues). The specific inhibition of [^{35}S] TBPS binding to the GABA_A convulsant site by these compounds was modeled. A database of 266 GABA_A receptor compounds was

compiled. Duplicates, mixtures and salts were removed to prepare the dataset for modeling. The remaining 210 compounds were used for modeling and chemical descriptors for each compound were generated. After calculating descriptors for each compound, computational tools such as *k*-Nearest-Neighbor (*k*NN), Support Vector Machine (SVM) and Random Forest (RF) were used to develop QSAR models. The generated models were validated using five-fold cross validation. Furthermore, predicting the activities of the external set, compounds not used in the modeling set, validated the developed models. The correct classification rates (CCR) for all the models were between 66% and 83%. Prediction values were relatively lower than accepted. However, applying an applicability domain (AD) increased the predictivity (CCR= 77% to 86%) and reduced the coverage (45%). The QSAR models developed in this study could be used to screen chemical libraries and identify potentially new GABA_A receptor convulsant site compounds.

High Throughput Screening (HTS) assays that measure the *in vitro* toxicity of environmental compounds have been widely used as an alternative to *in vivo* animal tests. Current HTS studies provide the community with rich toxicology information that has the potential to be integrated into toxicity research. The available *in vitro* toxicity data is updated daily in structured formats (*e.g.*, deposited into PubChem and other data sharing web portals) or in unstructured ways (papers, laboratory reports, toxicity website updates, etc.) The information derived from the current toxicity data is so large and complex that it becomes difficult to process using available database management tools or traditional data processing applications. For this reason, it is necessary to develop a “Big Data” approach when conducting modern chemical toxicity research.

In-vitro data for a compound, obtained from meaningful bioassays, can be viewed as a response profile that gives detailed information about the compound's ability to affect relevant biological protein/receptors. This information is critical for the evaluation of complex bio-activities (*e.g.*, animal toxicities) and grows rapidly as “big data” in toxicology communities. This review focuses mainly on the existing structured *in vitro* data (*e.g.*, PubChem datasets) as response profiles for compounds of environmental interest (*e.g.*, potential human/animal toxicants). Potential modeling and mining tools used to process big data in chemical toxicity research are also described.

ACKNOWLEDGMENTS

I would like to first thank my mentor, Dr. Hao Zhu for his continuous support, patience and motivation of my MS study and research. I am extremely grateful to Dr. Zhu for this opportunity and I could not imagine having a better advisor.

I would also like to thank Dr. Georgia Arbuckle-Keil and Dr. Jinglin Fu for being on my committee. Marlene Kim for her excellent guidance, knowledge and patience. She always found time from her busy schedule to help me with my research.

I would like to thank my parents, especially my father for his love, support and encouragement throughout all my studies. Finally I would like to thank David Gaston. He was always there to cheer me up through the long nights of studying.

DEDICATION

I would like to dedicate this thesis to my loving father whose support and encouragement has always been and will always be my strongest motivator.

TABLE OF CONTENTS

Title.....	i
Abstract.....	ii
Acknowledgments.....	v
Dedication.....	vi
List of figures.....	viii
List of tables.....	x
Chapter 1: Quantitative Structure Activity Relationship (QSAR) modeling of compounds binding to the GABA _A receptor convulsant site	
Section 1: Introduction.....	1
Section 2: Materials and methods.....	5
Section 3: Results and discussion.....	20
Section 4: Conclusion.....	26
Chapter 2: the use of high throughput screening in chemical toxicity studies	
Section 1: Introduction.....	27
Section 2: High Throughput Screening in chemical toxicology.....	30
Section 3: Current toxicity data sharing projects	32
Section 4: Characterizing toxicants by multiple bioassay data.....	37
Section 5: The use of bioassay data to prioritize animal toxicants.....	42
Section 6: Extracting useful bioassay data from multiple data resources.....	44
Section 4: Conclusion.....	48
References.....	49

LIST OF FIGURES

Figure 1. GABA _A receptor binding sites.....	3
Figure 2. Combinatorial QSAR modeling workflow.	17
Figure 3. Five-fold cross validation workflow.....	18
Figure 4. 3-D Plot of top three principal components using MOE descriptors for 210 modeling set (purple) and 94 external set (red) compounds.....	21
Figure 5 Performance of seven individual and consensus GABA _A binder QSAR models (N=210) using five-fold cross-validation.....	22
Figure 6. Prediction results of external validation set.....	23
Figure 7. Prediction results of external validation set with applicability domain applied.....	24
Figure 8. The scenario of big data for chemical toxicity research.....	29
Figure 9. The increase of compounds recorded in PubChem within five years (from September 2008 to September 2013).....	34
Figure 10. The response space of 962 ToxCast compounds represented by the data obtained from 193 PubChem bioassays. The red dots represent active responses; the blue dots represent inactive responses; and the yellow dots represent no testing data or inconclusive results.....	40
Figure 11. The response spaces of different category ToxCast compounds represented by the data obtained from 193 PubChem bioassays: (a) 171 consumer use chemicals (not including pharmaceuticals or pesticides); (b) 470 pesticides; (c) 245 pharmaceuticals; (d) 34 phthalates, plasticizers and alternatives.....	42
Figure 12. A potential <i>in vitro-in vivo</i> relationship in toxicology studies.....	46

Figure13. The response profiles of 50 PubChem bioassays for 107 compounds that may cause rat fetal growth retardation.....	48
--	----

LIST OF TABLES

Table 1. The data set consisting of 210 GABA _A receptor modulators used in this study	6
Table 2. The predictivity of various models without and with Applicability Domain (AD).....	25
Table 3. Table of the top two important Dragon descriptors.....	26
Table 4. 20 human toxicants with their relevant PubChem bioassay responses.....	39

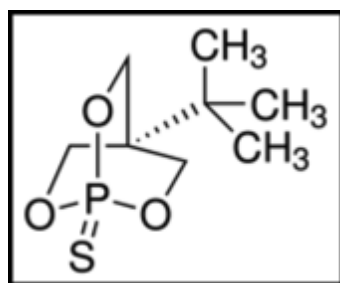
CHAPTER 1: Quantitative Structure Activity Relationship (QSAR) modeling of compounds binding to the Amino butyric acid type A (GABA_A) receptor convulsant site.

Section 1: Introduction

The GABA_A is part of the inhibitory neurotransmitter of ion channels.¹ It is one of the most complicated super families of ligand-gated ion channels and is responsible for anxiety and sleep disorders. The GABA_A receptor is a hetero-oligomeric protein, composed of pentameric protein subunits arranged around a central opening that form a chloride ion channel.^{1,2} A number of different classes of pharmacological agents exert their effects on the GABA_A receptor by binding to recognition sites that are distinct from the active binding site where GABA binds. GABA is the main inhibitory neurotransmitter in the central nervous system. It inhibits neurotransmissions in the brain and calms an anxious person after binding to the GABA site on the GABA_A receptor. The binding of GABA to the GABA_A receptor site activates the opening of the ion channel, which allows chloride anions to go down an electrochemical gradient.³ Compounds such as barbiturates, ethanol, anesthetics and convulsant agents that either directly or allosterically act on the GABA_A receptor can regulate the inhibitory effects of GABA.⁴ This study is centered on compounds binding directly to the GABA_A receptor convulsant site.

The convulsant binding site is located in the central opening that forms the chloride ion channel of the GABA_A receptor.⁵ Compounds binding to this convulsant binding site include picrotoxinin, a number of insecticides (including dieldrine) and [³⁵S] t-butylbicyclopophosphorothionate ([³⁵S] TBPS).⁴ These compounds bind to the GABA_A

receptor, blocking GABA-gated chloride ion flux in a non-competitive manner, resulting in convulsions.⁶ The convulsant [³⁵S] TBPS is considered a GABA_A receptor open channel blocker. The radiolabelled [³⁵S] TBPS binds to GABA_A receptor in the absence of GABA. Furthermore, low concentration of GABA enhances [³⁵S] TBPS binding, while higher concentrations reduce binding. The modulation of [³⁵S] TBPS binding could be used as an indicator of the efficacy of compounds that allosterically modulate GABA_A receptor function.^{4,6,7} Recent studies have shown that steroids are capable of modulating the function of the receptor and displacing [³⁵S] TBPS from the ion channel.⁸ However the interactions between steroids and [³⁵S] TBPS is still not clear.^{8–10} The unknown mechanism of the modification of the GABA_A receptor convulsant site has attracted great attention as information on the GABA_A receptor increases.¹¹ An understanding of how various ligands interact with the convulsant site would act as the starting point for developing potential therapeutics. Recently, the most active area of research has been the search for compounds that act on the convulsant binding site.¹²



[³⁵S] TBPS structure

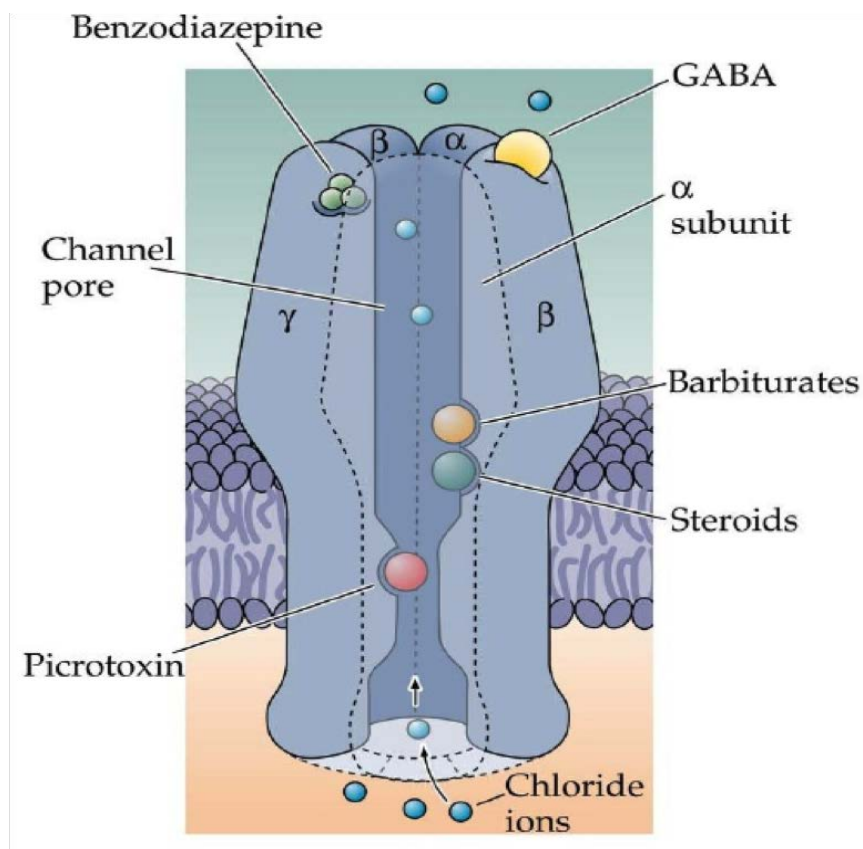


Figure 1 GABA_A receptor with its binding site.

In this study, various QSAR models were developed for a series of compounds (steroids derivatives, Arylsulfonyl derivatives and Propofol analogues) by modeling their ability to inhibit the specific binding of [³⁵S] TBPS to the GABA_A receptor convulsant site. QSAR has gained significant interest of many scientists as a computational method to reveal favorable drug candidates.¹³ Specifically, QSAR allows researchers to build computational models and use the resulting models to virtually screen chemical libraries instead of experimentally testing all the new compounds. Therefore, it could save resources by avoiding testing unfavorable compounds.¹⁴ In QSAR studies, machine learning and statistical approaches are applied to establish quantitative relationships

between chemical structural features and biological activities of compounds that have been tested experimentally.^{14–18}

The traditional methods of experimental testing are expensive, and time-consuming in drug discovery. Implementing computational methods as an alternative method to evaluate these drugs before synthesis would be cost effective and helpful in time management. There have been several reports in the literature on experimental testing of compounds with GABA_A receptor convulsant activity.^{12,19} However, only a few computational models have been developed. Other studies in QSAR modeling used simple linear regression and only one type of descriptor. In this study, QSAR models were developed for the GABA_A convulsant site using different types of descriptors and modeling tools. A database of 266 GABA_A receptor compounds was gathered from literature. Compounds used in this study have been extensively investigated by Rybczynski, *et al.*²⁰ According to the data collected, the biochemical interactions at the GABA_A receptor convulsant site helped to determine the inhibition of [³⁵S] TBPS and the binding affinity of [³⁵S] TBPS was measured. In this study, combinations of different chemical descriptors and modeling approaches were used. Moreover, computational tools such as *k* Nearest Neighbor (*k*NN),²¹ Random Forest (RF)²² and Support Vector Machine (SVM)²³ were used. In addition, model predictivity was validated by an external set and five-fold cross-validation. The validated models could be used to screen chemical libraries.

Section 2: Materials and Methods

Dataset

The series of GABA_A receptor compounds used in this study were obtained from the literature and other public sources.^{19,20,24} The initial data set contained 133 diverse classes of allosteric modulators such as Steroids, Arylsulfonyls and Propofols, which have been tested against ([³⁵S] TBPS) assay. The active compounds had an IC50 value less than 50uM. Since all the compounds are GABA_A receptor binders, in order to develop a model with comparable inactive compounds, we needed to include inactive compounds into the modeling set. Inactive compounds were expected to be the compounds that either do not displace [³⁵S] TBPS or are not known to be neuroactive. Inactive compounds were obtained from an anti-cancer bioassay screen obtained from PubChem (AID 248).²⁵ This anti-cancer dataset contained 55,728 compounds. To generate a balanced modeling set with a similar active/inactive ratio, we had to select a portion of the anti-cancer data set. A similarity search was applied to determine if the compounds in the anti-cancer dataset were structurally similar to the 133 active compounds. Inactive compounds were selected when they were structurally similar to the active compounds. The resulting modeling set contained 266 compounds (133 active, 133 inactive).

To avoid chemical structure errors, commercial software CASE Ultra²⁶ and ChemAxon (www.chemaxon.com) Standardizer and Structure Checker 6.2.2, 2014 were used to curate all the chemical structures into 2D Simplified Molecular-Input Line-Entry System (SMILES). Then, all duplicates, mixtures, inorganics, metalorganics and salts were removed since our modeling tools cannot handle these types of compounds. The

remaining 210 unique compounds were used for QSAR modeling.

Table 1 List of the 210 GABA_A receptor modulators used in this study

CIDS	ACTIVITY	SMILES
1691	0	<chem>COc1cccc2C(=O)c3c(O)c4CC(O)(CC(OC5CC(N)C(O)C(C)O5)c4c(O)c3C(=O)c12)C(=O)CO</chem>
2569	0	<chem>COC(COC(N)=O)C1C(=O)C(N2CC2)=C(C)C(=O)C=1N3CC3</chem>
3228	0	<chem>CCCCCCCCCCCCCCCCCCCCC(=O)NC1C=CN(C2OC(CO)C(O)C2O)C(=O)N=1</chem>
12242	0	<chem>CCCCC(C)(C)O</chem>
61215	0	<chem>COCCOC1C(=O)C(N2CC2)=C(OCCOC)C(=O)C=1N3CC3</chem>
64983	0	<chem>[H][n]1cccc1C2=NCC(=O)Nc3ccc(Cl)cc32</chem>
65702	0	<chem>CICCN(CCCl)P1(=O)OCCCN1CCCl</chem>
65800	0	<chem>C1CN(CCO1)P2(N=P(N=P(N=2)(N3CC3)N4CC4)(N5CC5)N6CC6)N7CC7</chem>
70732	0	<chem>Cc1ccc(O)c(C)n1</chem>
71627	0	<chem>CCCOC1C(=O)C(N2CC2)=C(OCCC)C(=O)C=1N3CC3</chem>
73492	0	<chem>COC(=O)C(O)C(O)(CCC(C)C)C(=O)OC1C2c3cc4OCOc4cc3CCN5CCCC25C=C1OC</chem>
81863	0	<chem>OCN1CC(=O)N(CO)CC1=O</chem>
91467	0	<chem>CC1CC(C)(O)CC(C(O)CC2CC(=O)NC(=O)C2)C1=O</chem>
92114	0	<chem>COc1cc2c3CC4CCCN4Cc3c5cc(OC)c(OC)cc5c2cc1OC</chem>
94737	0	<chem>O=P(N1CCOCC1)(N2CC2)N3CC3</chem>
95892	0	<chem>O=P(Oc1ccc(OP(=O)(N2CC2)N3CC3)cc1)(N4CC4)N5CC5</chem>
99105	0	<chem>CICCN(CCCl)P1(=O)NC(CCO1)OOC2CCOP(=O)(N2)N(CCCl)CCl</chem>
99814	0	<chem>CC(C)C1NC(=O)C(NC(=O)c2ccc(C)c3OC4=C(C)C(=O)C(NCCC O)=C(C(=O)NC5C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C6CCCN6C(=O)C(NC5=O)C(C)C)C4=Nc32)C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C7CCCN7C1=O</chem>
122716	0	<chem>CCC(C)C1NC(=O)C(NC(=O)C2=C(N)C(=O)C(C)=C3Oc4c(C)ccc(C(=O)NC5C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C6CCCN6C(=O)C(NC5=O)C(C)C)c4N=C32)C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C7CCCN7C1=O</chem>
122799	0	<chem>CCC(C)C1NC(=O)C(NC(=O)c2ccc(C)c3OC4=C(C)C(=O)C(N)=C(C(=O)NC5C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C6CCCN6C(=O)C(NC5=O)C(C)CC)C4=Nc32)C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C7CCCN7C1=O</chem>
124693	0	<chem>OCCNC1C(=O)C(N2CC2)=C(NCCO)C(=O)C=1N3CC3</chem>
227087	0	<chem>CCOC1C(=O)C(N2CC2)=C(OCC)C(=O)C=1N3CC3</chem>
227091	0	<chem>CCCC(=O)NC1C(=O)C(N2CC2)=C(NC(=O)CCC)C(=O)C=1N3CC3</chem>
238139	0	<chem>OC(=O)CCOc1cccc1N(CCCl)CCCl</chem>
238141	0	<chem>CCOC(=O)CCOc1cccc1N(CCCl)CCCl</chem>
238912	0	<chem>COc1ccc(OP(=O)(N2CC2)N3CC3)cc1</chem>
239395	0	<chem>CCOC(=O)COc1ccc(cc1)N(CCCl)CCCl</chem>

241659	0	<chem>CC1CN1P(=O)(N2CCCCC2)N3CC3C</chem>
242304	0	<chem>CC(=O)OC(CN1CC1)C=C</chem>
242512	0	<chem>CCC(C)P(=O)(N1CC1)N2CC2</chem>
245527	0	<chem>CN(CCN(C)P(=O)(N1CC1)N2CC2)P(=O)(N3CC3)N4CC4</chem>
245645	0	<chem>CCOC(=O)C(Cc1ccc(cc1)N(CCCl)CCCl)NP2(=O)NCCCCO2</chem>
246259	0	<chem>CC1(C)CN1CC2CO2</chem>
246845	0	<chem>COc1ccc2c3CN4CCCC4Cc3c5cc(OC)c(OC)cc5c2c1OC</chem>
252674	0	<chem>Cc1cc2N=C3C(=O)NC(=O)N=C3N(CCN(CCCl)CCCl)c2cc1C</chem>
261792	0	<chem>COc1cc2c3CN4CCCC4C(O)c3c5ccc(O)cc5c2cc1OC</chem>
262126	0	<chem>CC(C)CCCC(C)C1CCC2C3CCC4CC(CCC4(C)C3CCC12C)NC(=O)N(CCF)N=O</chem>
267924	0	<chem>CN1N=C(Br)C(=O)N(C(c2ccccc2)c3ccccc3)C1=O</chem>
271070	0	<chem>CC(=O)OCC1OC(OC(C)=O)C(NC(=O)N(CCCl)N=O)C(OC(C)=O)C1OC(C)=O</chem>
271100	0	<chem>CC(C)CCCC(C)C1CCC2C3CCC4CN(CCC4C3CCC12C)N=O</chem>
276389	0	<chem>COC(=O)CC(O)(CCC(C)(C)O)C(=O)OC1C2c3cc4OCOc4cc3CCN5CCCC25C=C1OC</chem>
277822	0	<chem>OCC1OC(C(O)C1O)N2C=CC(=O)CC2=O</chem>
280145	0	<chem>[H][n]1c2ccc(OC)cc2c3CCNC(CC4CC5N(CCc6cc(OC)c(OC)cc56)CC4CC)c31</chem>
282479	0	<chem>CCCCCCCCCCCCCCCC(=O)OCC1OC(C(O)C1O)N2C=CC(N)=NC2=O</chem>
285033	0	<chem>COC(=O)CC(O)(CCCC(C)(C)O)C(=O)OC1C2c3cc4OCOc4cc3CCN5CCCC25C=C1OC</chem>
286093	0	<chem>CN(C)N=Nc1ccc(cc1C(N)=O)N(=O)=O</chem>
287401	0	<chem>Clc1ccc2NC(=O)CN=C(c3ccccc3)c2c1</chem>
289158	0	<chem>CC(=O)OCC[n]1c2ccccc2c3c(C)c4cnccc4c(C)c31</chem>
290774	0	<chem>O=C1CN(C(=O)CN2C(=O)c3ccccc3C2=O)C4(CCCCC4)O1</chem>
291125	0	<chem>COc1ccc2c(c1)c3c(C)c4cnccc4c(C)c3[n]2CCO</chem>
292761	0	<chem>Nc1nenc2c1c(c[n]2)C3OC(CO)C(O)C3O)C4NCCCN=4</chem>
296462	0	<chem>Cc1ccc(cc1)C(=O)NC2N=CN(C3CC(O)C(CO)O3)C(=O)N=2</chem>
296555	0	<chem>CN(C)N=Nc1ccc(cc1)S(=O)(=O)Nc2ncccn2</chem>
300071	0	<chem>CCOC(=O)C(=CNc1ccc(OC)cc1N(=O)=O)C(=O)OCC</chem>
302546	0	<chem>ClCCN(CCCl)CCC1OC(=O)c2ccccc2N=1</chem>
302547	0	<chem>ClCCN(CCCl)CCN1C=Nc2ccccc2C1=O</chem>
304437	0	<chem>COc1cc(ccc1N(CCCl)CCCl)C=Nc2ccc3c(c2)nc[n]3C</chem>
304444	0	<chem>COc1cc(ccc1N(CCCl)CCCl)C=Nc2ccc3c(c2)nc(C)[n]3C</chem>
304640	0	<chem>CCOC(=O)N(OCc1ccccc1)P(=O)(N2CC2(C)C)N3CC3(C)C</chem>
304645	0	<chem>CCCOP(=O)(N1CC1(C)C)N2CC2(C)C</chem>
308449	0	<chem>Cc1ccc(cc1C(O)=O)S(Cl)(=O)=O</chem>
311908	0	<chem>CCC(C(C)N1CC(=O)NC(=O)C1)N2CC(=O)NC(=O)C2</chem>
314656	0	<chem>CN1C(O)CCOP1(=O)N(CCCl)CCCl</chem>
315960	0	<chem>OCC(O)CNC1C(=O)C(N2CC2)=C(NCC(O)CO)C(=O)C=1N3CC3</chem>
317400	0	<chem>CN(CCO)C1C(=O)C(N2CC2)=C(N(C)CCO)C(=O)C=1N3CC3</chem>
317858	0	<chem>O=C1C(N2CCOCC2)=C(N3CC3)C(=O)C(N4CCOCC4)=C1N5CC5</chem>
318116	0	<chem>O=C1c2ccccc2C(c3ccccc3)=C1N4CCOCC4</chem>
320696	0	<chem>CC1CCC2(CO)C(OC3C(O)C(O)C2(C)C43CO4)C=1</chem>
409503	0	<chem>CCCCCOP(=O)(N1CC1)N2CC2</chem>

412954	0	CCCCOP(=O)(N1CC1C)N2CC2C
413456	0	CC(C)CCOP(=S)(N1CC1)N2CC2
413529	0	CCN(CC)P(=O)(N1CC1)N2CC2
413614	0	CCCCCCCCOP(=O)(N1CC1)N2CC2
413615	0	CC1CN1P(=O)(OCC=C)N2CC2C
414069	0	CCCOP(=S)(N1CC1(C)C)N2CC2(C)C
414148	0	CC1CN1P(=O)(OCCC#N)N2CC2C
414375	0	CCCCCCCCCCOP(=S)(N1CC1)N2CC2
414525	0	CCOP(=S)(N1CC1(C)C)N2CC2(C)C
419025	0	[H][n]1cc(CC(NC(=O)OC(C)(C)C(=O)N2CCc3cc(OC)c(OC)cc3C2CC4CC5N(CCc6cc(OC)c(OC)cc56)CC4CC)c7ccccc71
419709	0	COC(=O)C(N)Cc1ccc(OC)c(c1)N(CCCl)CCCl
419710	0	CCOC(=O)C(N)Cc1ccc(OC)c(c1)N(CCCl)CCCl
419712	0	CCCCOC(=O)C(N)Cc1ccc(OC)c(c1)N(CCCl)CCCl
420567	0	CC(C)Oc1ccc(CC(N)C(O)=O)cc1N(CCCl)CCCl
420568	0	CCCCCOc1ccc(CC(N)C(O)=O)cc1N(CCCl)CCCl
420569	0	NC(Cc1ccc(OC2CCCC2)c(c1)N(CCCl)CCCl)C(O)=O
421525	0	CCOC(=O)C1C2OC(=O)C(C1C3CCC(CC4CCC(CC=4)C5C(C6OC(=O)C5C(=O)C6CN(CCCl)CCCl)C(=O)OCC)=CC3)C(=O)C2CN(CCCl)CCCl
422111	0	CC(=O)OCC12CCC(C)=CC1OC3C(O)C(OC(C)=O)C2(C)C43CO4
422455	0	COc1cc(ccc1N(CCCl)CCCl)C=Ne2cc(Cl)ccc2[n]3ccnc3C
428565	0	OCCN1CCN(CC1)C2C(=O)C(N3CC3)=C(N4CCN(CCO)CC4)C(=O)C=2N5CC5
429170	0	OC1CCCN(C1)C2C(=O)C(N3CC3)=C(N4CCCC(O)C4)C(=O)C=2N5CC5
429285	0	CN(Cc1cnc2nc(N)nc(N)c2n1)c3ccc(cc3)C(=O)NC(CCC(=O)OC(C)(C)C)C(=O)OC(C)(C)C
429922	0	CC(=O)OCC12CCC(C)=CC1OC3C(O)C(O)C2(C)C43CO4
708507	0	Cc1cc(C)c(NC(=O)C=CC(O)=O)c(C)c1
3978511	0	[H][n]1cnc2ncnc(SC3(CC)C(=O)NC(=O)NC3=O)c21
5351130	0	COC1C(O)C(C)OC(OC2C(O)C(O)C(C)OC2Oc3cccc4c(O)c5C(=O)Oc6ccc(C)c7C(=O)Oc(c34)c5c67)C1O
5357950	0	O=C1C=CC(C=C1)=NNc2cccc(c2)N(=O)=O
5383847	0	CCOc1cc(OCC)c(C=CN(=O)=O)cc1OCC
5384159	0	CC=CC(=O)OCC12CCC(C)=CC1OC3C(O)C(O)C2(C)C43CO4
5798035	0	O=C1OC2CCCCC2C1=CN3CCCCC3
5911723	0	CCOC(=O)C1=CN=C2OC(=CNN3CC(CN4CCOCC4)OC3=O)C=C2C1=O
6711181	0	CC(C)CC(=O)OC1CC2(COC(C)=O)C(OC3C(O)C(OC(C)=O)C2(C)C43CO4)C=C1C
44415057	0	CC(C)C1NC(=O)C(NC(=O)c2ccc(C)c3OC4=C(C)C(=O)C(N)=C(C(=O)NC5C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C6CCCN6C(=O)C(NC5=O)C(C)C)C4=Ne32)C(C)OC(=O)C(C(C)C)N(C)C(=O)CN(C)C(=O)C7CCCN7C1=O
54602262	0	CC(=O)OCC12CCC(C)=CC1OC3C(O)C(OC(=O)CCl)C2(C)C43CO4
54607467	0	CC(=C)C(=O)OCC12CCC(C)=CC1OC3C(O)C(O)C2(C)C43CO4
54610857	0	CC(=O)OCC12CCC(C)=CC1OC3C(=O)C(OC(C)=O)C2(C)C43C

		O4
54704409	0	<chem>CN(C)C1C2CC3C(=C(O)c4c(O)cccc4C3(C)O)C(=O)C2(O)C(O)=C(C(=O)NCN(CCCl)CCCl)C1=O</chem>
9823370	1	<chem>CC(C)(C)OC(=O)c1nc[n]2c3cccc(Br)c3C(=O)N4CCCC4c12</chem>
44375818	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)CN=N#N</chem>
10529558	1	<chem>CC(=O)SCC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCOC(C)(C)C5</chem>
10766614	1	<chem>CC1(C)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(=O)CCl)C5(C)C43)CC2O</chem>
10626038	1	<chem>CCCCC1(CCCC)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O</chem>
10790675	1	<chem>CC(C)CC1CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O</chem>
44375759	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)C#C</chem>
44291795	1	<chem>CCOC1CC2(C)C(CCC3C4CCC(C(C)=O)C4(C)CC(C32)N(C)C)C1O</chem>
44375748	1	<chem>COCC1(O)CCC2(C)C(CCC3C4CCC(C(C)=O)C4(C)CCC32)C1</chem>
10790676	1	<chem>CCC1(CC)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O</chem>
6918305	1	<chem>CC(=O)C1CCC2C3CCC4CC(C)(O)CCC4(C)C3CCC21C</chem>
10717823	1	<chem>Clc1ccc(cc1Cl)C2CCCN(N=2)P(=O)(OC3CCCC3)c4cccc4</chem>
10790110	1	<chem>CC(C)C1CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O</chem>
3016	1	<chem>CN1C(=O)CN=C(c2ccccc2)c3cc(Cl)ccc13</chem>
92786	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)CCC4(C)C3CCC21C</chem>
10765151	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCOC(C)(C)C5</chem>
10813342	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCSC(C)(C)C5</chem>
44375717	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)C=C</chem>
44290903	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CC(C)(C)OC(C)(C)C5</chem>
22868799	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCOC(C5)Cc6cccc6</chem>
44291114	1	<chem>CC12CC(C(O)CC1CCC3C4CCC(C(=O)CBr)C4(C)CCC32)N5CCOCC5</chem>
10003232	1	<chem>CC12CC(C(O)CC1CCC3C4CCC(C(=O)CCl)C4(C)CCC32)N5CCOCC5</chem>
44375980	1	<chem>CCCC1(O)CCC2(C)C(CCC3C4CCC(C(C)=O)C4(C)CCC32)C1</chem>
10672358	1	<chem>CC1(C)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(=O)CCl)C5(C)CC(=O)C43)CC2O</chem>
10623901	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCSC(C)(C)C5</chem>
10454374	1	<chem>CCCCc1ccc(cc1)C2CCCN(N=2)S(=O)(=O)c3ccc(CCCC)cc3</chem>
44375681	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)CC#C</chem>
11957655	1	<chem>CC12CCC(O)CC1CCC3C4CCC(C(=O)CO)C4(C)CCC32</chem>
44375554	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CBr)CCC4(C)C3CCC21C</chem>
44375555	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CF)CCC4(C)C3CCC21C</chem>
44375804	1	<chem>COCC1(O)CCC2(C)C(CCC3C4CCC(C(C)=O)C4(C)CCC32)C1</chem>

44375726	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)CC=C</chem>
44291040	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CC(C)(C)SC(C)(C)C5</chem>
44291209	1	<chem>CC(=O)SCC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCOCC5</chem>
44375661	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCl)CCC4(C)C3CCC21C</chem>
44375742	1	<chem>CCC1(O)CCC2(C)C(CCC3C4CCC(C(C)=O)C4(C)CCC32)C1</chem>
10577865	1	<chem>CC(=O)SCC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCOC(C)(C)C5</chem>
10249569	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)C(F)(F)F</chem>
10598812	1	<chem>CC1CN(CC(C)O1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CCC43)CC2O</chem>
44291039	1	<chem>CC(=O)OCC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCOCC5</chem>
10696819	1	<chem>CC1(C)CN(CC(O1)C2CC3(C)C(CCC4C5CCC(C(=O)CSC#N)C5(C)CC(=O)C43)CC2O</chem>
44375725	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC=C)CCC4(C)C3CCC21C</chem>
44291794	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCOCC5</chem>
44375466	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)CC=C=C</chem>
44375628	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)COCe5cccc5</chem>
10577010	1	<chem>CC(=O)OCC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3CCC21C)N5CCOC(C)(C)C5</chem>
10716882	1	<chem>CC1CN(CC(O1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CCC43)CC2O</chem>
11801168	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCOC(C5)(c6cccc6)c7cccc7</chem>
10523173	1	<chem>CCCCe1ccc(cc1)C2CCCN(N=2)P(=O)(OC)c3cccc3</chem>
9803583	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCOC(C)(C)C5</chem>
44375607	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(CCC4(C)C3CCC21C)CC#N</chem>
44375965	1	<chem>CCCOCC1(O)CCC2(C)C(CCC3C4CCC(C(C)=O)C4(C)CCC32)C1</chem>
44375805	1	<chem>CC(=O)C1CCC2C3CCC4CC(O)(Cl)CCC4(C)C3CCC21C</chem>
10475111	1	<chem>CCCCe1ccc(cc1)C2CCCN(N=2)P(=O)(OCC)c3cccc3</chem>
10767121	1	<chem>CCCCe1ccc(cc1)C2CCCN(N=2)S(=O)(=O)c3ccc(I)cc3</chem>
10837416	1	<chem>CCe1ccc(cc1)C2CCCN(N=2)S(=O)(=O)c3ccc(I)cc3</chem>
53676432	1	<chem>CCC1CN(CC(O1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CCC43)CC2O</chem>
10766360	1	<chem>CCCC1CN(CC(O1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O</chem>
10529493	1	<chem>CC(=O)OCC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCOC(C)(C)C5</chem>
10453392	1	<chem>CCOP(=O)(N1CCCC(=N1)c2ccc(Cl)c(Cl)c2)c3cccc3</chem>
11796910	1	<chem>CCOP(=O)(N1CCCC(=N1)c2ccc(Cl)c(c2)C(F)(F)F)c3cccc3</chem>
10570269	1	<chem>CCCCe1ccc(cc1)C2CCCN(N=2)S(=O)(=O)c3cccc3</chem>
54060826	1	<chem>CC1CN(CC(O1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O</chem>
44291113	1	<chem>CC12CC(C(O)CC1CCC3C4CCC(C(=O)CO)C4(C)CC(=O)C32)N5</chem>

		CCOCC5
44291020	1	CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCSCC5
10549542	1	CC(C)OP(=O)(N1CCCC(=N1)c2ccc(Cl)c(Cl)c2)c3ccccc3
21625943	1	CCC1CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O
10572108	1	CCCCc1ccc(cc1)S(=O)(=O)N2CCCC(=N2)c3ccc(CC)cc3
44375937	1	CC(=O)C1CCC2C3CCC4CC(O)(CO)CCC4(C)C3CCC21C
10624805	1	CC1(C)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)CN=N#N)C5(C)CC(=O)C43)CC2O
10737922	1	CCCCc1ccc(cc1)C2CCCN(N=2)S(=O)(=O)c3ccc(C)cc3
44291500	1	CC(=O)C1CCC2C3CCC4CC(O)C(CC4(C)C3C(=O)CC21C)N5CCOCC5
44291153	1	CC12CC(C(O)CC1CCC3C4CCC(C(=O)CO)C4(C)CCC32)N5CCOCC5
10647003	1	CC1CN(CC(C)O1)C2CC3(C)C(CCC4C5CCC(C(C)=O)C5(C)CC(=O)C43)CC2O
10647712	1	CC1(C)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)CO)C5(C)CC(=O)C43)CC2O
10837999	1	CC(C)c1ccc(cc1)C2CCCN(N=2)S(=O)(=O)c3ccc(I)cc3
10807379	1	COc1ccc(cc1)C2CCCN(N=2)P(=O)(OC)c3ccccc3
10669623	1	COP(=O)(N1CCCC(=N1)c2ccc(Cl)c(c2)C(F)(F)F)c3ccccc3
44291194	1	CC12CC(C(O)CC1CCC3C4CCC(C(=O)CCl)C4(C)CC(=O)C32)N5CCOCC5
10759993	1	CCOP(=O)(N1CCCC(=N1)c2ccc(F)cc2)c3ccccc3
10782784	1	COP(=O)(N1CCCC(=N1)c2ccc(F)cc2)c3ccccc3
10572969	1	CCCCc1ccc(cc1)S(=O)(=O)N2CCCC(=N2)c3ccc(cc3)C(C)C
3000715	1	CCCC(C)C1(CC)C(=O)NC(=S)NC1=O
4737	1	CCCC(C)C1(CC)C(=O)NC(=O)NC1=O
4943	1	CC(C)c1cccc(C(C)C)c1O
9882905	1	CC(C)c1cc(I)cc(C(C)C)c1O
10687388	1	COC(CNCC(=O)Oc1c(ccc1C(C)C)C(C)C)OC
10730904	1	COS(=O)(=O)c1cc(C(C)C)c(O)c(c1)C(C)C
10589500	1	CC(C)c1cc(NC(=O)C(F)(F)F)cc(C(C)C)c1O
10062776	1	CC(C)c1cccc(C(C)C)c1OC(=O)CN2CCCC2
818538	1	CC(C)c1cc(N)cc(C(C)C)c1O
10727719	1	CC(C)c1cc(cc(C(C)C)c1O)N(C)C
10059055	1	CC(C)c1cc(Cl)cc(C(C)C)c1O
10680603	1	CC(C)c1cc(cc(C(C)C)c1O)N(=O)=O
10588978	1	CC(C)c1cc(cc(C(C)C)c1O)C(=O)c2ccccc2
18469877	1	COC(=O)c1cccc(C(C)C)c1O
10422572	1	CC(C)c1cc(Br)cc(C(C)C)c1O
82712	1	CC(C)c1cc(C=O)cc(C(C)C)c1O
600975	1	COc1c(ccc1C(C)C)C(C)C
10779272	1	CC(C)c1cccc(C(C)C)c1OC(=O)c2ccccc2
596091	1	CC(C)c1cccc(C(C)C)c1OC(C)=O
15874	1	CCN(CC)CC(=O)Oc1c(ccc1C(C)C)C(C)C
104845	1	CC(=O)C1CCC2C3CCC4CC(O)CCC4(C)C3C(=O)CC21C
10503609	1	CC1(C)CN(CCO1)C2CC3(C)C(CCC4C5CCC(C(C)=O)CO)C5(C)CC

		C43)CC2O
--	--	----------

Chemical Descriptors

Molecular Operating Environment (MOE) and Dragon version 6.0 were used to calculate 2D descriptors. MOE descriptors included physical properties (such as n-octanol water partition coefficient [Log P], molecular weight and molar refractivity), structural keys, E-state indices, subdivided surface areas, topological indices, topological polar surface area, atom counts and bond counts, Kier & Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors and partial charge descriptors. Dragon descriptors included E-state values and E-state counts, topological descriptors, constitutional descriptors, ring walk and path counts, connectivity indices, information indices, 2D autocorrelations, Burden eigenvalues, molecular distance edge, Kappa, hydrogen bond acceptor/donor counts, chemical fingerprints, molecular fragment counts, 2D matrix-based descriptors, 2D atom pairs, drug-like indices, Chemically Advanced Template Search 2D and geometrical descriptors. The Dragon software generated over 3,000 descriptors. Redundant descriptors were removed. Additionally, descriptors with high correlation coefficients were removed. For example, if two descriptors had a correlation greater than 0.99, one was randomly removed. A total of 186 MOE and 873 Dragon descriptors were used to develop the QSAR models.

Modeling Approaches

The application of RF, *k*NN and SVM algorithms available in R.2.15.1²⁷ were used in this study.

Random Forest (RF)

In machine learning, RF is a predictor, which creates trees from random selection of descriptors. It yields the prediction by combining predictions from individual trees. RF can be used for classification or regression models. The algorithm for inducing a RF was developed by Breiman and Cutler.²² In the RF modeling procedure, n samples are randomly drawn from the training dataset based on the user-defined values. These samples are used to construct n training sets and to build n trees. These trees permit the evaluation of importance of each descriptor, taking into consideration the entire descriptor pool. It allows the isolation of the important descriptors. For each node of the tree, m , which is the total number of input qualities in the dataset, are randomly chosen from all of the available chemical descriptors. The best data split are used to develop a decision tree model based on these m variables in the training set.²² This process repeats for each tree until the nodes are too small to split. RF produces a highly accurate classifier and runs efficiently on large databases. It also provides effective methods for detecting variable interactions.

k Nearest Neighbor (kNN)

In k NN, the program uses a classification algorithm principle and variable selection technique for model development.²⁸ k NN implies that similar compounds display similar activities. This method predicts each activity as the average activity of k (number of compounds) most similar compounds from the training set. The initial step starts with the random selection of a subset of $nvar$ (number of selected variables) descriptors. The $nvar$ is set to different values, and the training set models are developed

with leave-one-out cross validation method, where each compound is eliminated from the training set and its biological activity is predicted as the average activity of the k is optimized as well ($k= 1-5$).²⁹ The similarity is characterized by the Euclidean distance between compounds in multidimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criteria is used to optimize the selection of variables.³⁰ The objective of this method is to obtain the best leave-one-out cross-validated (LOO-CV) Correct Classification Rates (CCR) by optimizing $nvar$ and k . The additional details of the method can be found elsewhere.^{23,30} k NN provides a simple and effective method of modeling.

Following our general QSAR modeling workflow methodology, all of the k NN models were extensively validated. The modeling compounds were divided multiple times into training/test sets. The model acceptability cutoff values of the LOO-CV accuracy of the training sets and the prediction accuracy for test set were both set arbitrarily as 0.7. Models that did not meet both training and test set accuracy cutoff criteria of 0.7 were discarded. The cut-off of 0.7 was selected, because models that perform better than random (cutoff criteria 0.5) were desired.

Support Vector Machine (SVM)

SVM represents a set of supervised learning methods used for classification and regression modeling. SVM was developed by Vapnik²¹ as a general data modeling methodology where both the training set error and the model complexity are incorporated into a special loss function that is minimized during model development. The methodology allows one to regulate the importance of the training set error versus the

model complexity to develop the optimal model that best predicts a test set. Later, SVM was extended to afford the development of SVM regression models for data sets with activities, such as QSAR.²⁹

In SVM, chemical descriptors are mapped onto a high dimensional space using kernel functions that is typically nonlinear. The system then looks for an optimal separation between two classes, such that each in their entirety lies on opposite sides of a separating hyperplane. This is achieved by maximizing the margin between the closest points, known as support vector, and the hyperplane.²³ Advantages of using SVM is that it is effective in high dimensional spaces. SVM is likely to give poor prediction if the number of features is much greater than the number of samples.

Combinatorial QSAR workflow

The complete combinatorial QSAR modeling workflow is shown in Figure 2. Each model was developed using Dragon or MOE descriptors and either RF, SVM or *k*NN modeling approaches, resulting in six different models: Dragon-RF, Dragon-*k*NN, Dragon-SVM, MOE-RF, MOE- *k*NN, and MOE-SVM. In addition resulting models were averaged to generate a consensus model.

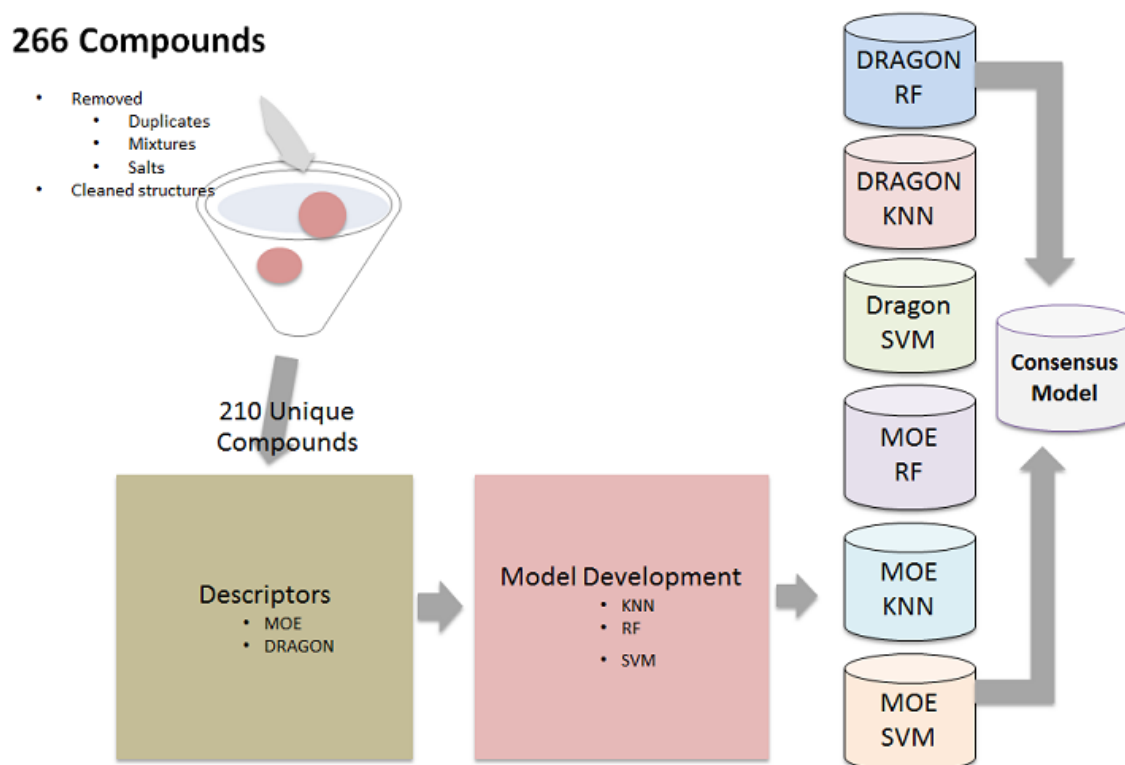


Figure 2 Combinatorial QSAR modeling workflow

Five-fold cross validation

The most common and effective approach to testing models is cross validation. In five-fold cross-validation, the modeling set is distributed randomly into five separate folds (Figure 3). Each fold contained 20% of the compounds in the modeling set; each fold was used as an external set while the remaining four (80%) jointly formed a modeling set. This was done until each fold had been treated as an external set. Our goal here was to evaluate the performance of all the compounds in the modeling set.

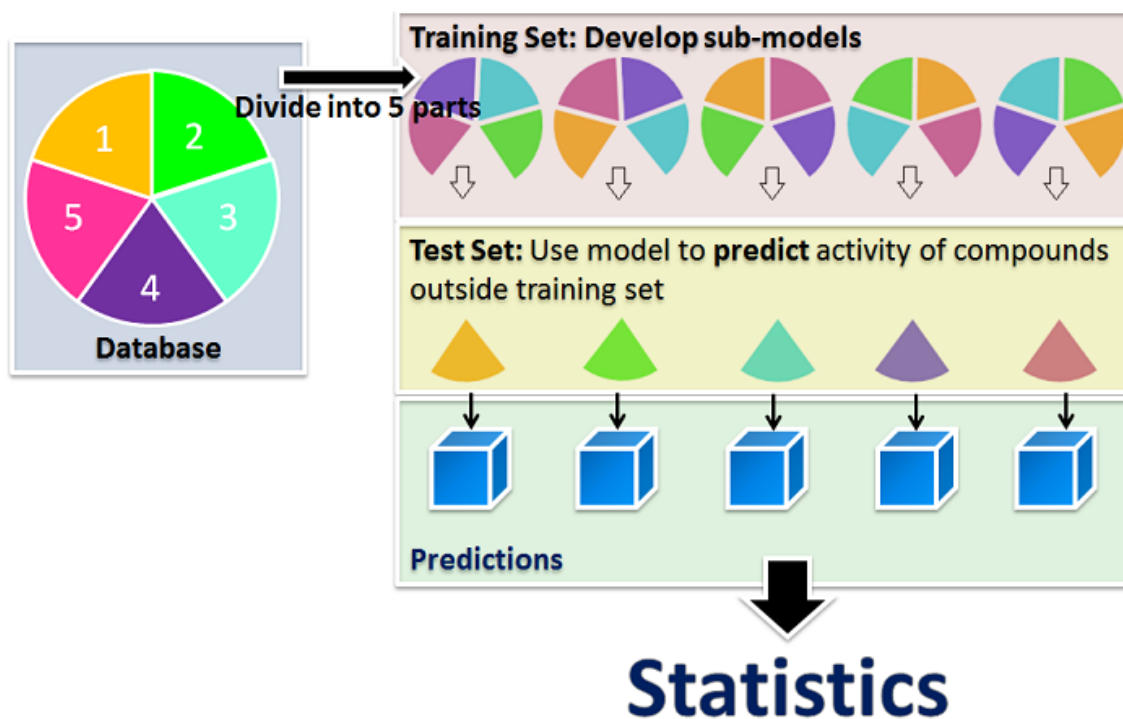


Figure 3 Five-fold cross validation workflow

Universal statistical figures of merit for all models

The implementation of various modeling approaches and descriptors were used in the modeling method. Universal statistical metrics were necessary for the evaluation of model performance. Furthermore, to harmonize the results of this study, the results were analyzed using sensitivity (the percentage of active compounds predicted correctly), specificity (the percentage of inactive compounds predicted correctly) and correct classification rate (CCR) to assess all predictions. These parameters are defined as follows:

$$\% \text{ Sensitivity} = \left(\frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \right) 100$$

$$\% \text{ Specificity} = \left(\frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \right) 100$$

$$\% \text{ CCR} = \left(\frac{\text{sensitivity} + \text{specificity}}{2} \right) 100$$

Applicability Domain (AD)

A QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, if a compound is highly dissimilar to all compounds of the modeling set, the reliable prediction of its activity is unlikely. The concept of the Applicability Domain (AD) was developed and used to avoid an unjustified extrapolation of activity predictions. In this study, the AD was defined as a threshold distance D_T between a compound under prediction and its closest nearest neighbor of the training set, calculated as follows:³¹

$$D_T = \bar{Y} + Z\sigma$$

Here, \bar{Y} is the average Euclidean distance between each compound and its k nearest neighbors in the training set (where k is the parameter optimized in the course of QSAR modeling, and the distances are calculated using all descriptors and descriptors selected by the optimized model only), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. We set the default value of this parameter Z to 0.5, which formally places the allowed distance threshold at the mean plus one-half of the standard deviation. Thus, if the distance of the external

compound from its nearest neighbor in the entire descriptor space or the subspace of descriptors selected in the training set exceeds this threshold, the prediction is not made.³¹

External dataset

For external validation, an external data set (76 compounds) from three bioassays studies from AID 71551, 71835 and 71844 were compiled.³²⁻³⁴ These compounds have been extensively tested for their ability to inhibit the specific displacement of [³⁵S] TBPS at the GABA_A receptor convulsant site. The test set was tested against the modeling set to remove compounds that existed in the training set.

Section 3: Results and Discussion

Overview of modeling set and test set

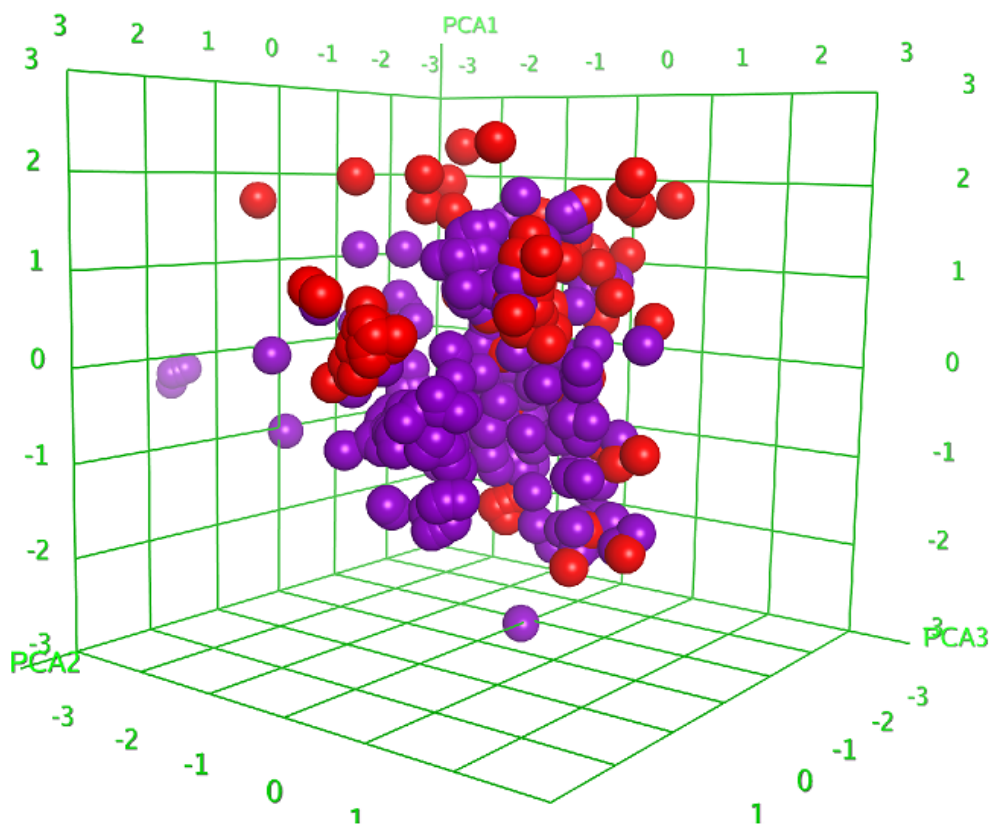


Figure 4. 3-D Plot of top three principal components using MOE descriptors for 210 modeling set (purple) and 76 external set (red) compounds

Principle component analysis (PCA) was used to evaluate the structural similarities in the modeling set. The chemical space of both the modeling set and the external set was analyzed by performing a PCA using MOE chemical descriptors. After calculating principal components using all 186 MOE descriptors, the top three principal components were selected to create a chemical space for 210 modeling set and 76

external set compounds. The chemical space of the modeling set and the external set is shown in Figure 4.

Modeling Results

Six individual and one-consensus category models were developed. The results for all the five-fold-cross validations (refer to methods for description) are shown in Figure 5. The sensitivity, specificity and CCR for all the models ranged from 73-99%, 87-96%, to 80-97%, respectively.

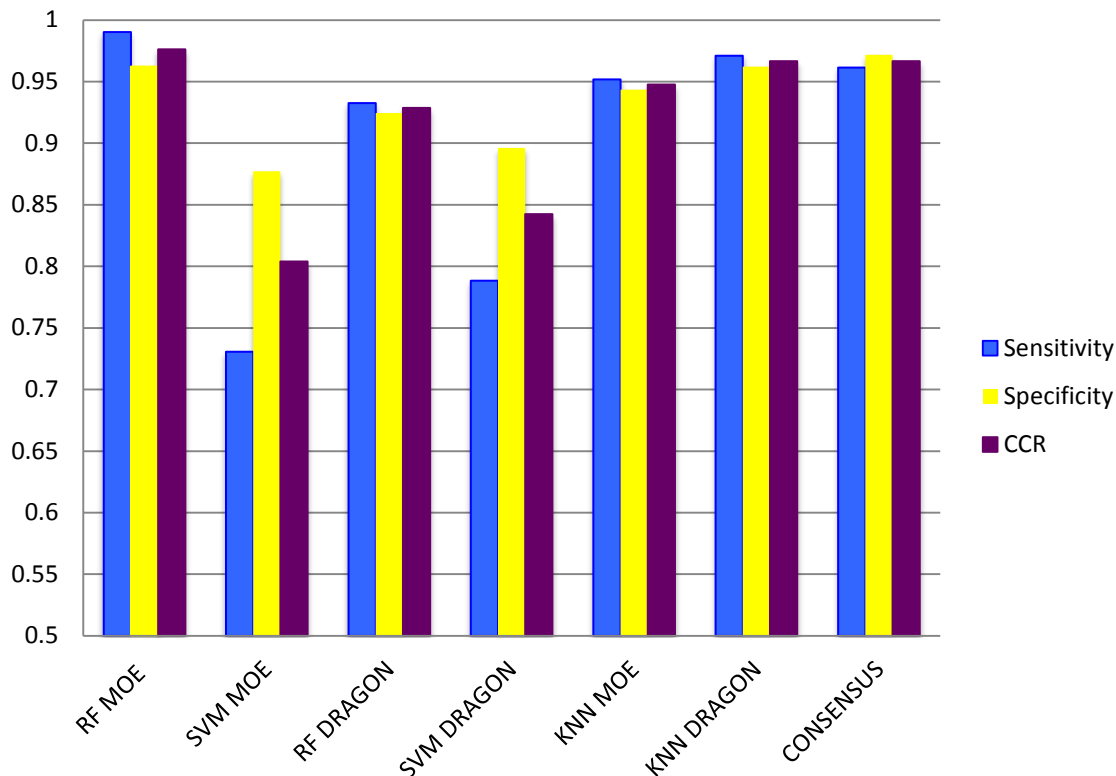


Figure 5 Performance of seven individual and consensus GABA_A binder QSAR models (N=210) using five-fold cross validation

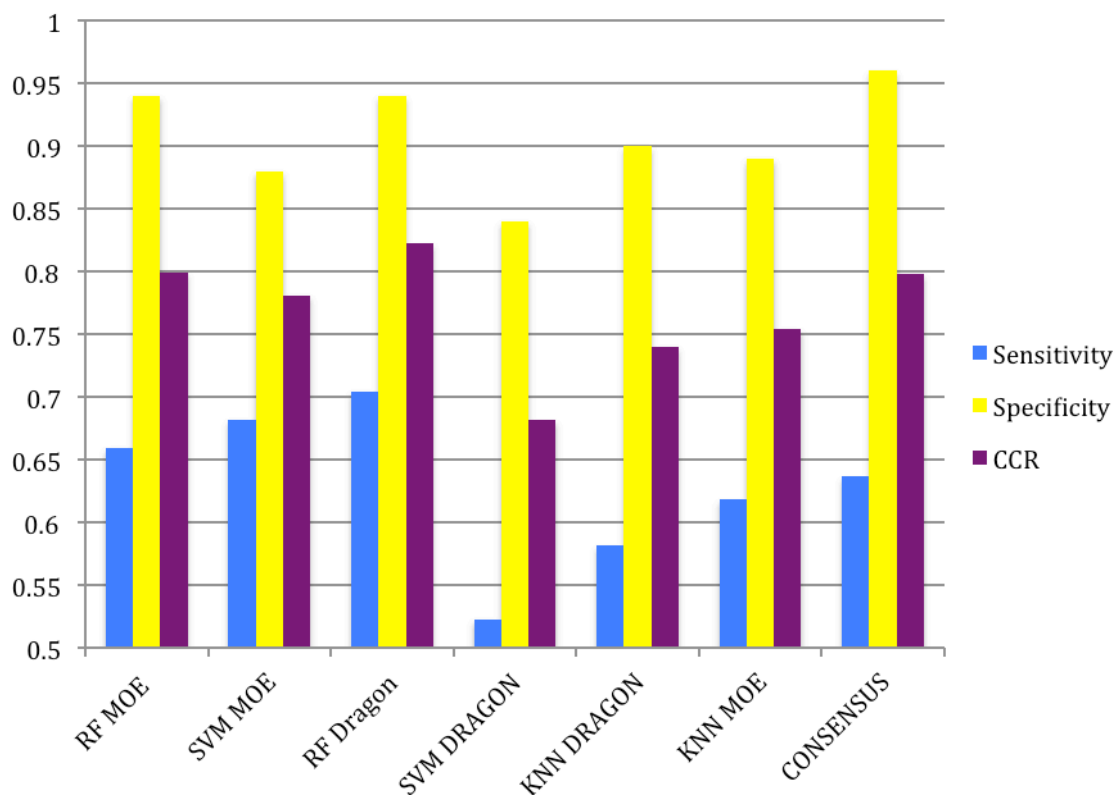


Figure 6 Prediction results of external validation set

Furthermore, the prediction results for the external compounds are shown in Figure 6.

The sensitivity, specificity and CCR for all the models ranged from 52-70%, 84-96%, to 74-82%, respectively. Next, an AD³⁵ was applied to identify outliers in the external set. The AD was defined by a distance between a compound being predicted and its nearest neighbor in the training set.³⁵ For example, a compound “within domain” had a distance less than the defined value. The CCR range for all the individual models and the consensus model was 76-86% (Figure 7). Applying AD, increased the prediction accuracy but decreased the coverage (45%) (Table 2).

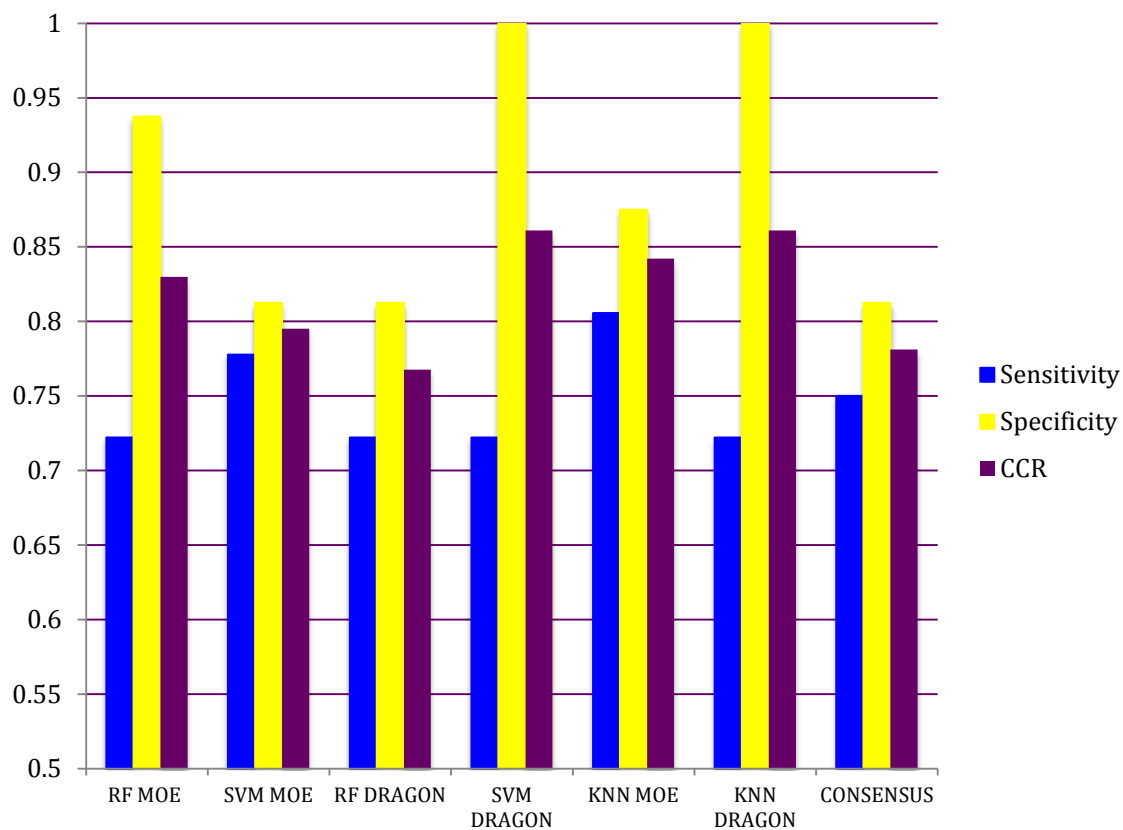


Figure 7 Prediction results of external validation set with applicability domain applied.

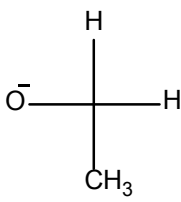
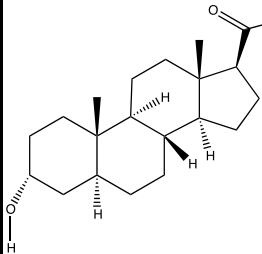
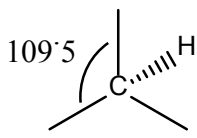
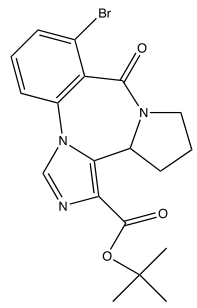
Table 2. The predictivity of various models without and with applicability domain.

AD	STATISTICS %	MODELS						
		Consensus	RF MOE	SVM MOE	RF Dragon	KNN MOE	KNN MOE	KNN Dragon
NO AD N=76	Sensitivity	63	65	68	70	52	61	58
	Specificity	96	94	88	94	84	89	96
	CCR	79	79	78	82	68	75	74
WITH AD N=52	Sensitivity	72	77	72	72	80	72	75
	Specificity	93	81	81	100	87	100	81
	CCR	82	79	76	86	84	86	78

Descriptor Analysis

Chemical descriptors are numerical values that quantitatively represent the characteristic of a molecule. Descriptor analysis helps identify specific behaviors of the molecule and its interactions with the receptor. These descriptors could help provide meaningful explanations into the potential mechanisms since the exact mechanisms are unknown. Dragon descriptors from the *k*NN models were selected for this analysis. Dragon descriptors from the *k*NN models were specifically analyzed, because it contained more diverse descriptors than MOE. The top two descriptors in the *k*NN dragon QSAR models are shown in Table 3, along with their descriptions and frequency (%) out of 801 *k*NN models. Descriptor H-046 had the highest frequency (73%). The H-046 descriptor represents hydrogen attached to carbon with sp^3 hybridization.³⁶ There was also a fragment descriptor (nCt) that was considered to be important in the models. Descriptor nCt represents the number of total tertiary carbons with sp^3 hybridization.^{37–39} It is interesting to notice that the TBPS molecule also has both fragments. Both polar and non-polar characteristics of these molecules (*i.e.* TBPS and most of the active compounds) should be the binding features to the GABA_A receptor convulsant site.

Table 3 Top two Dragon descriptors used to develop *k*NN QSAR models.

Descriptor Name	Description	Illustration	Example of active compound with descriptor	Frequency (%)
H-046	H attached to CO (sp^3) no X attached to next C			73
nCt	Number of total tertiary carbons (sp^3)			25

Section 4: Conclusion

In this study, numerous approaches were used to develop a Combinatorial QSAR model. Models generated in this study showed high predictivity. Statistical techniques used to develop the models demonstrated that the models have the ability to predict new compounds. Resulting models were validated by prediction of the activity of an external set from additional sources. The five-fold cross validation performed better than the external validation. However, applying an AD increased the predictivity (CCR= 76-86%) but reduced coverage.

Overall, a good Combinatorial QSAR model was developed in this study from the data collected. Statistical methods used for modeling demonstrated the ability of the model to predict new chemicals. Descriptor analysis showed that potential GABA_A receptor active compounds, including [³⁵S] TBPS, have some common substructural features (*e.g.*, those described by the H046 and nCt descriptors). These characteristics should be the binding features to the GABA_A receptor convulsant site. Models in this study can be used to screen external chemical libraries and to identify potential active GABA_A receptor convulsant site compounds.

CHAPTER 2: The Use of High Throughput Screening (HTS) in Chemical Toxicity Studies

Section1: Introduction

With the great progress of combinatorial chemistry since the 1990s, large chemical libraries became the major source of modern drug discovery procedure.^{40,41} Over the past 10 years, this effort also stimulated the development of High Throughput Screening (HTS) techniques.^{42,43} Traditional toxicity testing protocols using animal models are expensive and time consuming. Because of the urgent need to use alternative methods in toxicity studies, the US National Research Council (NRC) outlined a new vision and strategies for the increased use of *in vitro* technologies for chemical risk assessment.⁴⁴ With its low cost and short testing time, HTS has been viewed as a potential alternative to animal models.

HTS is a process that screens from thousands to millions of compounds using a rapid and standardized protocol. Current HTS techniques are usually combined with robotic methods. Parallel data processing and biological assay miniaturization has become more and more popular in toxicology studies as they greatly reduce the cost of experimental testing.^{42,45} It is understandable that some “popular” compounds, especially those of toxicity interest (*e.g.*, known human toxicants), have been tested multiple times and in many different bioassays. For this reason, the assay response data from multiple resources and/or multiple testing protocols could be viewed as the “response profile” of the compounds being tested. Figure 8 shows the current data construction of compounds in toxicity testing. Compared to the limited amount of historical animal toxicity data, the chemical-response data space obtained from HTS is much more complex and keeps growing daily.

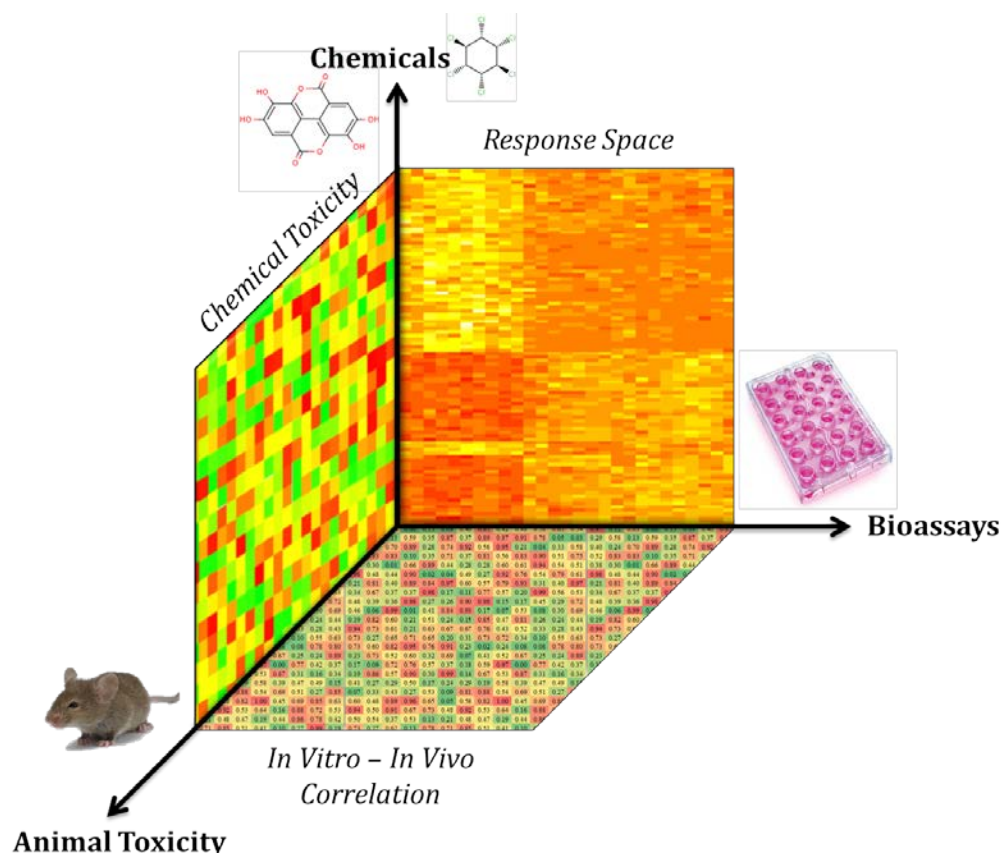


Figure 8 The scenario of big data for chemical toxicity research

The term “big data” describes a collection of data sets that are so large and complex that they are too difficult to process by traditional data analysis tools. Originally the “big data” focus was on advanced data storage and handling techniques, such as cloud-based computing or high-speed heterogeneous computational environments.⁴⁶ Currently, the problem of big data is gaining increasing recognition in clinical studies and other research areas driven by biological data.^{47,48} Clearly the progress of HTS and relevant data sharing projects moved modern chemical toxicity research into the big data era. The need for novel techniques, including data

mining/generation, curation, storage and management, brings new challenges and opportunities to the current toxicology community.

Section 2: High Throughput Screening in Chemical Toxicology

There were several important movements by regulatory agencies for the development of HTS assays, which are potential alternatives to animal testing. The National Institutes of Health (NIH) Roadmap for medical research was launched in 2004.⁴⁹ Fueled by this initiative, several molecular libraries screening centers were developed by the NIH Molecular Libraries Common Fund Program. The NIH Chemical Genomics Center (NCGC), which is now a branch of the National Center for Advancing Translational Sciences (NCATS), was one of them. In 2005, right after NCGC was initiated, the National Toxicology Program (NTP) and NCGC started a collaboration to 1) develop a chemical library suitable for HTS; 2) develop HTS assays potentially informative for *in vivo* toxicity effects; and 3) experimentally test the chemical library by these HTS assays.⁴³ This is one of the early efforts to systemically use the HTS technique within toxicology studies. During the same period, there were many other HTS projects that were performed by other research groups.^{50–54} Although these studies were not specifically designed for chemical toxicity, but for drug discovery and other areas, these HTS efforts also generated numerous bioassay data for large chemical libraries. For the early days of HTS development, several reviews are available.^{42,55–59}

In 2006, the U.S. Environmental Protection Agency (EPA) initiated a research program named toxicity forecaster (ToxCast). The goal of this program was to develop methods for utilizing *in vitro* toxicity tests and various toxicogenomics technologies to quickly evaluate the toxic potential of chemicals and to prioritize candidates for future animal testing.⁵⁵ Phase I of ToxCast employed a chemical library of *ca.* 300 unique compounds, most of which were chemicals for agricultural use, such as pesticides, and

had relevant animal toxicity testing results available.⁶⁰ Around 500 cell-free or cell-based assays were used to screen this chemical library. From these, over 600 *in vitro* end points were measured for each chemical, generating over 200,000 concentration response data points. In ToxCast Phase II, another 767 compounds, including some failed pharmaceuticals, were screened using around 700 HTS assays.⁶¹

In 2008, another big collaborative program, called Toxicity Testing in the 21st century (Tox21), was launched by NTP, NCGC and EPA,⁶²⁻⁶⁴ joined later by the U.S. Food and Drug Administration (FDA). The Tox21 collaboration brought together its partners' expertise in the areas of experimental toxicology, *in vitro* assays, and informatics.⁶⁴ The target chemical library of Tox21 screening contains over 8,000 unique compounds, including commercial compounds, pesticides and all marketed pharmaceuticals.⁶¹ Screening of this extensive chemical library commenced in 2011 at NCGC, with a throughput capacity of approximately 25 assays per year.

Section 3: Current Toxicity Data Sharing Projects

Facilitated by the combined efforts of HTS, as described above, and combinatorial chemical synthesis, modern screening programs produced enormous amounts of biological data, especially the chemical responses on specific targets.⁶⁵ As a result, several data sharing projects, in parallel with the generation of HTS toxicity data, were also initiated in the past ten years. For example, PubChem is a public repository for chemical structures and their biological properties.^{66,67} Most of the HTS data (*e.g.*, those generated from the above toxicology programs) were shared through PubChem. Figure 9 shows the yearly increase of PubChem compounds.⁶⁸⁻⁷³ In the past five years, the number of PubChem compounds increased from 1.9 million in September 2008⁶⁸ to 4.8 million in September 2013.⁷² During the same period, the number of bioassays that were deposited into PubChem increased from 1,197 in September 2008⁶⁸ to over 700,000 in September 2013.⁷² The tremendous amount of PubChem bioassay data, with total size of more than five terabytes, resulted in a big data pool for environmental compounds with various target response information.

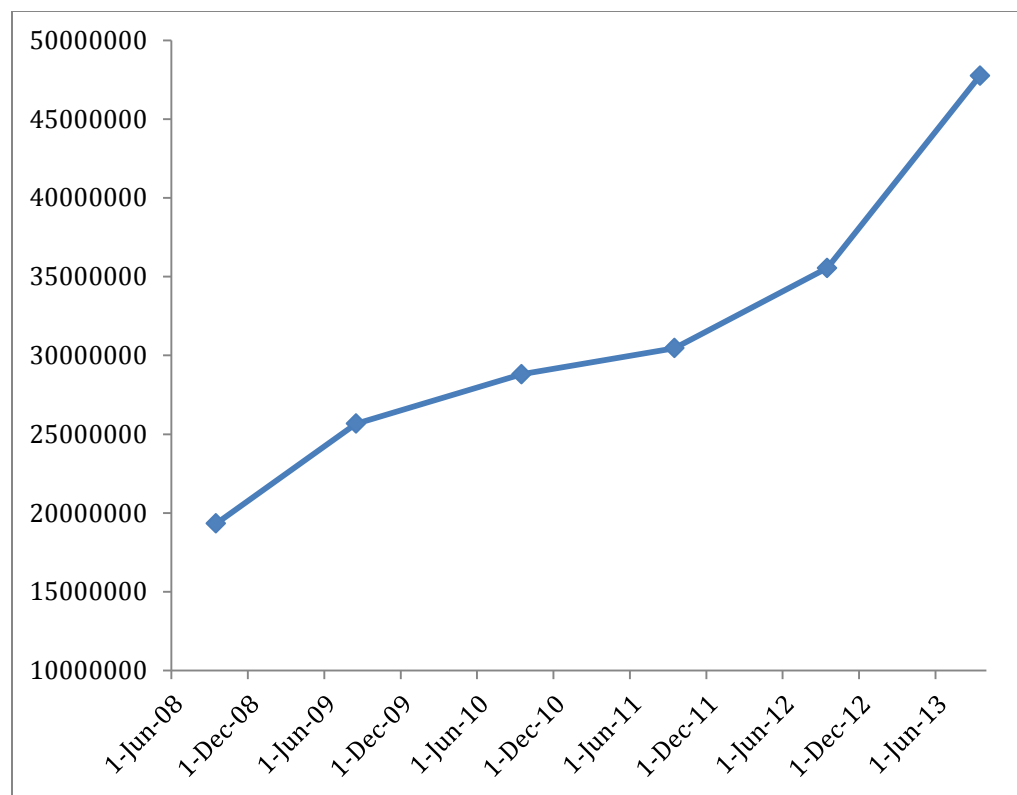


Figure 9 The increase of compounds recorded in PubChem within five years (from September 2008 to September 2013).

A large group of toxicity bioassay data, found in PubChem, but existing also as an individual data sharing project, is from the European Bioinformatics Institute (EBI).⁷⁴ The EBI's goal is to provide freely available data and bioinformatics services to all branches of the scientific community. As a part of this goal, the ChEMBL database was compiled from publicly available data found in scientific publications. In 2011, the ChEMBL version 11(ChEMBL_11) was launched and includes 3.3 million bioassay readout data of 629,943 compounds. This was obtained from curating over 42,500 scientific publications.⁷⁴

There are other programs specifically for sharing chemical toxicity data, including animal testing results. Starting in 2007, the U.S. EPA's National Center for Computational Toxicology (NCCT) program initiated a unique toxicity data search program, named Aggregated Computational Toxicology Resource (ACToR).^{75,76} The mission of ACToR was to develop a central database that links to a set of existing toxicity databases to bring together many types and sources of toxicity data for a large environmental chemical library. Aside from the results of *in vitro* bioassays, the current ACToR portal has the links to over 100 different animal toxicity data sources (*e.g.*, ToxRefDB and DSSTox).⁷⁶ The most recent product of ACToR is the newly launched Chemical Safety for Sustainability Dashboard (<http://actor.epa.gov/dashboard/>). This new function provides an interactive tool to explore rapid, automated (or *in vitro* high-throughput) chemical screening data generated by the ToxCast project and the federal Tox21 collaboration.

Similar to ACToR but with a different mission, the ToxNET program contains and allows navigation through 16 separate databases of much more diverse chemicals.⁷⁷ ToxNET was developed by the National Library of Medicines' (NLM) Division of Specialized Information Services (SIS). By grouping the databases together, ToxNET allows for all information to be accessed from one query form. Although there are 14 separate databases used as the query source of ToxNET, some toxicity data are similar and are grouped together in the integrated report.

In response to the shortage of alternative testing methods, the European Commission and the European Cosmetics Association launched the most recent research initiative, so called Safety Evaluation Ultimately Replacing Animal Testing (SEURAT)

in 2011.⁷⁸ It is called "SEURAT-1", indicating that more steps have to be taken before the final version will be reached. Under the SEURAT-1 initiative, there were six research projects funded and heavy data curation/management analysis involved.⁷⁹ For example, one of these projects, the COSMOS project, is dedicated to the development of freely available tools and workflows to predict the safety of cosmetic ingredients to humans.⁸⁰ In the recently released COSMOS database web portal (<http://cosmosdb.cosmostox.eu/>), there are over 5,500 unique cosmetic-type compounds with their relevant toxicity data.

Toxicogenomics is a field of toxicology that addresses information concerning gene, protein, and metabolite changes within a particular cell or tissue of an organism in response to chemicals. Many modern *in vitro* toxicity studies result in outcomes via relevant toxicity mechanisms and these findings can be translated into biomarkers that could be applied to human exposure studies.⁸¹ Toxicogenomics investigations generate large amounts of “omics” data that are meant to predict toxicity or genetic susceptibility induced by chemicals. The Chemical Effects in Biological Systems (CEBS) database developed by the National Institute of Environmental Health Sciences (NIEHS) is now the public repository for all NTP conventional toxicology and carcinogenicity data as well as NCGC HTS data.⁸² Along with the Comparative Toxicogenomics Database (CTD) at Mount Desert Island Biological Laboratory, CEBS aims to promote comparative studies of genes and proteins across species.^{83–86} Currently, CTD data is searchable through the ToxNET portal and CEBS is available at <http://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm>. In 2010 NIEHS launched DrugMatrix (<https://ntp.niehs.nih.gov/drugmatrix/index.html>), another large-scale data sharing portal that contains *in vitro* and *in vivo* rat gene expression data,

measured after chemical treatment, for over 600 chemicals and 10,000 genes. The Broad Institute since 2006 maintains the Connectivity Map (cmap) project (<http://www.broadinstitute.org/cmap/>), a collection of over 7,000 genome-wide expression profiles from cultured human cell lines for over 1,300 compounds.⁸⁷

Section 4: Characterizing Toxicants by Multiple Bioassay Data

The direct consequence of the HTS testing effort in the past ten years is the massive amount of available biological data for organic compounds, especially those of environmental interest. A significant number of those compounds have been tested multiple times. For example, Table 4 shows 20 common toxicants obtained from the Integrated Risk Information System (IRIS) database (<http://www.epa.gov/IRIS/>). Based on the search result on PubChem (accessed on December of year 2013), these toxicants were reported to be tested in hundreds of PubChem bioassays. For example, chlordecone (CAS 143-50-0), which is an insecticide now banned from the market, showed active responses in 328 bioassays (Table 4). Other toxicants have similarly rich response information on PubChem (Table 4).

The multiple bioassay data of a single compound can be viewed as its biological profile, reflecting its interactions. Profiling compounds, especially the toxicants, to study their toxicity potential is the most straightforward way to use the available bioassay data. ToxCast Phase I screened over 300 unique compounds, mostly food pesticides, in 467 bioassays. The resulting data was used to profile screened compounds for their potential to induce carcinogenicity,⁸⁸ developmental toxicity,^{89,90} reproductive toxicity,⁹¹ and endocrine disruption.^{92,93}

Table 4. 20 human toxicants with their relevant PubChem bioassay responses.

Chemicals	CAS	Number of Active Responses	Number of Inactive Responses
CHLORDECONE	143-50-0	328	539
TOXAPHENE	8001-35-2	294	112
HEXACHLOROCYCLOPENTADIENE	77-47-4	208	262
DICHLORVOS	62-73-7	181	633
PENTACHLOROPHENOL	87-86-5	95	690
HEPTACHLOR	76-44-8	85	624
DDT, P,P'-	50-29-3	76	386
DDD, P,P'-	72-54-8	70	186
ENDOSULFAN	115-29-7	65	259
NAPHTHALENE	91-20-3	61	890
DDD, O,P'-	53-19-0	61	964
1,4-DICHLOROBENZENE	106-46-7	57	362
4,6-DINITRO-O-CRESOL	534-52-1	57	213
PHENOL	108-95-2	53	518
CHLORPYRIFOS	2921-88-2	48	739
METHOXYCHLOR	72-43-5	47	710
2,4-DINITROPHENOL	51-28-5	46	672
TETRACHLOROPHENOL	25167-83-3	45	515
BENZO(A)PYRENE	50-32-8	39	358
4,4'-METHYLENEBIS(2-CHLOROANILINE)	101-14-4	32	431

Besides the bioassay data generated by the ToxCast program, the Tox21 compounds have been tested in other screening projects. In the current big data era, the bioassay response profile can be very large for some compounds (*e.g.* those well-known toxicants shown in Table 4). The initial response space can be large, complex and unorganized. For example, Figure 10 shows the PubChem response space of 962 ToxCast compounds by using 193 PubChem assays (accessed December 2013). By classifying the ToxCast compounds into four major categories,⁹⁴ we could compare the response profiles of different types of compounds (Figure 11). Compared to phthalates

plasticizers, the pharmaceutical compounds and pesticides have been studied in most bioassays and the active response ratios are relatively high.

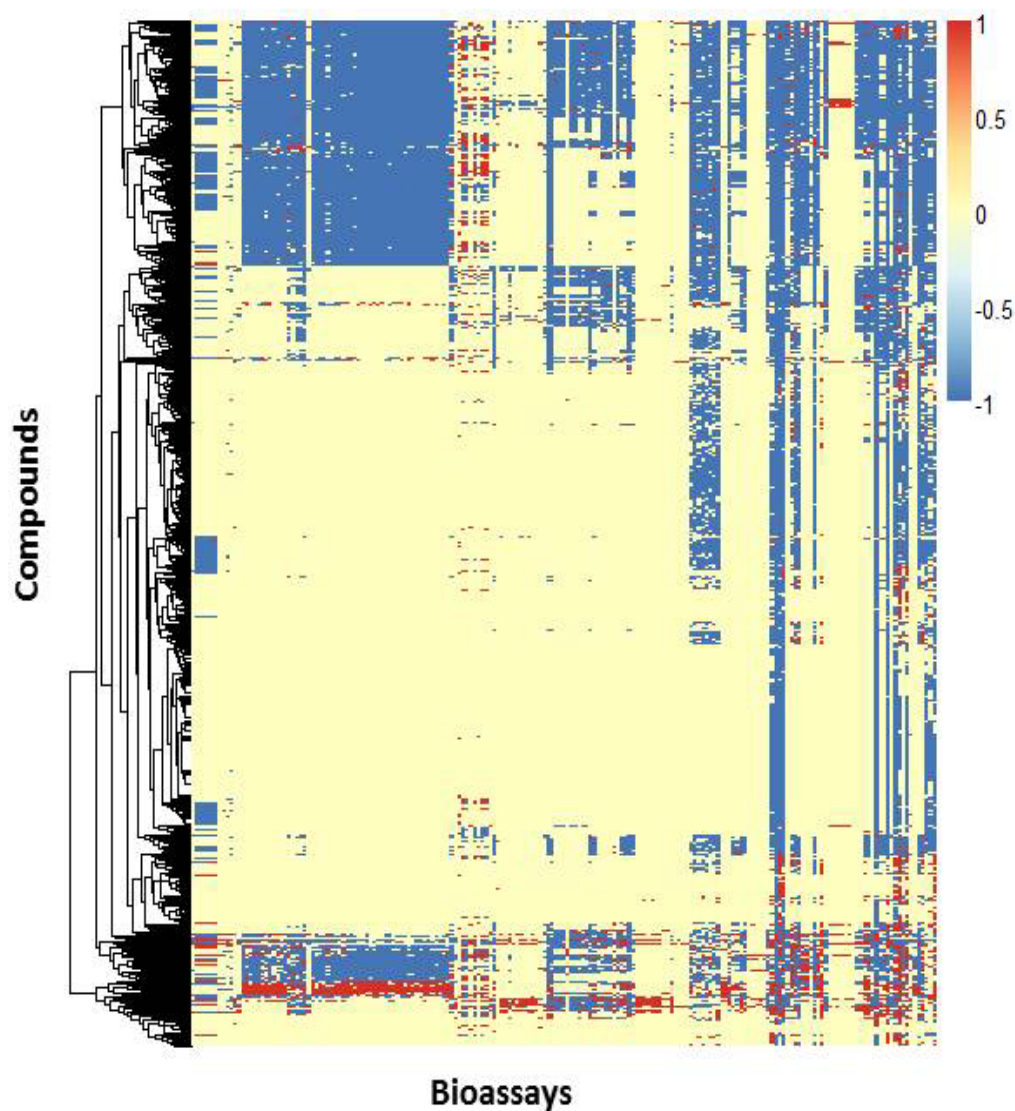


Figure 10 The response space of 962 ToxCast compounds represented by the data obtained from 193 PubChem bioassays. The red dots represent active responses; the blue dots represent inactive responses; and the yellow dots represent no testing data or inconclusive results.

It is understandable that most areas within the initial response map are either “no testing” or “inconclusive” because many bioassays have only been applied to a small portion of this large chemical set. Furthermore, the nature of HTS assays, many of which represent very specific interactions, results in a biased distribution of responses for the target chemicals (many more “inactives” than “active” data entries). Since not all the bioassay data are relevant or useful for a particular type of toxicity, additional rational selection steps are needed to select useful information from the bulk of available big data.

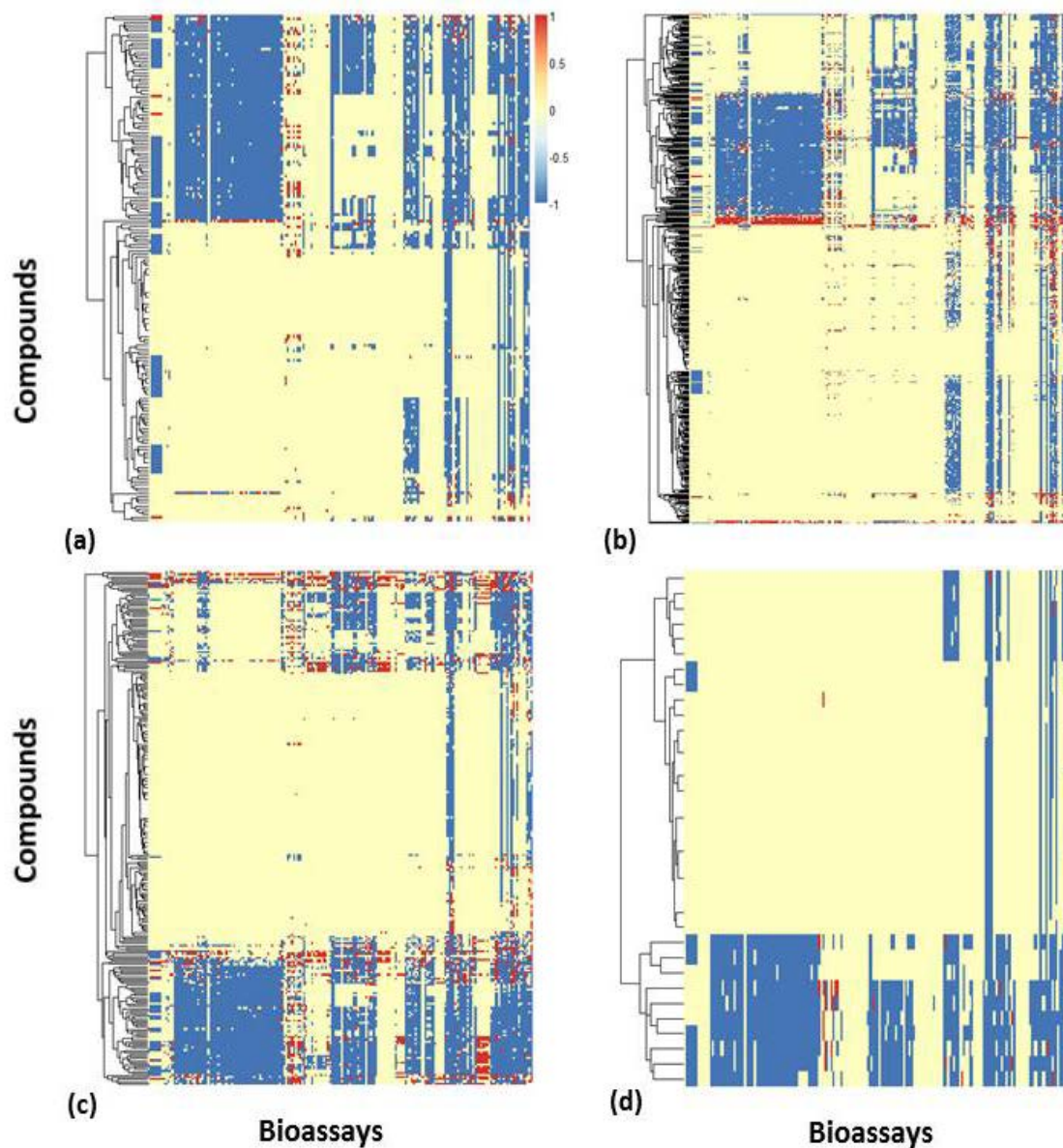


Figure 11 The response spaces of different category ToxCast compounds represented by the data obtained from 193 PubChem bioassays: (a) 171 consumer use chemicals (not including pharmaceuticals or pesticides); (b) 470 pesticides; (c) 245 pharmaceuticals; (d) 34 phthalates, plasticizers and alternatives.

Section 5: The Use of Bioassay Data to Prioritize Animal Toxicants

There have been some studies to use the current bioassay data to identify likely animal toxicants and/or prioritize them for future experimental animal testing. For example, the currently available ToxCast bioassays have been organized into a global scoring system, called ToxPi, to identify potential toxicants by their responses in these assays.^{90,92-95} Furthermore, toxicity pathways could also be generated, linking relevant bioassays together by analyzing their biological targets.^{96,97} ToxCast Phase I is the first time there has been a big data effort to generate and systemically use large scale bioassay data in chemical toxicity studies. In ToxCast Phase II, similar efforts continued with the new 767 target chemicals, including 111 failed pharmaceutical drug molecules.⁹⁴ In the recent Tox21 program, the results obtained from ToxCast were used to select the most useful bioassays as the testing battery for a much larger database.^{63,98,99}

There are other research groups and agencies that use bioassays to study various *in vivo* toxicities, such as acute toxicity,¹⁰⁰⁻¹⁰³ developmental toxicity,¹⁰⁴ and drug-drug interactions.¹⁰⁵ One example is the AcuteTox collaborative project initiated within the European Union. Its purpose is to develop alternative testing strategies that could replace animal testing for predicting human acute oral systemic toxicity.^{103,106-110} Similar to ToxCast, AcuteTox generated large-scale *in vitro* toxicity data from multiple bioassays.¹⁰⁷ All these efforts contributed to the initial pool of big data for chemical toxicants.

Dr Zhu's group has also utilized bioassay data to predict animal toxicity of organic compounds. In the first two of our studies, multiple HTS data from NCGC

bioassays were used as biological descriptors to develop predictive models for various animal toxicity endpoints.^{111,112} The models with hybrid (combination of chemical and biological) descriptors showed better predictivity than the traditional Quantitative Structure-Activity Relationship (QSAR) models using only chemical descriptors. In another study, the biological descriptors obtained from toxicogenomics data were used to model animal hepatotoxicity.¹¹³

Section 6: Extracting Useful Bioassay Data from Multiple Data Resources

The clear limitation of extrapolating results from *in vitro* assays to a whole organism is that each *in vitro* assay generally only considers one or several target sites rather than a comprehensive organism consisting of hundreds of potential targets.^{114,115} The practical solution is to form a large battery of diverse *in vitro* assays for a specific animal toxicity, such as the ToxCast strategy.^{55,61} In the toxicant profiling studies described above, each project was limited to the use of the data generated by its own HTS assays. This lack of data integration across multiple related toxicity databases is clearly a big and open issue. How to integrate large scale datasets from various sources is the key question that needs to be addressed in the current big data scenario. To realize this goal, novel data mining tools need to be developed to extract useful data from different resources. Wild and his coworkers developed a framework called Chem2Bio2RDF to link several data resources, such as DrugBank, PubChem, ChEBL and others.¹¹⁶ This framework, including other similar data mining tools developed in the same group, was used to create complex systems biology models (*e.g.*, for drug adverse effects).^{117–119} Recently Fourches *et al.* reported a newly developed software, named HTS Navigator, to extract, visualize, and analyze HTS data from various resources.¹²⁰

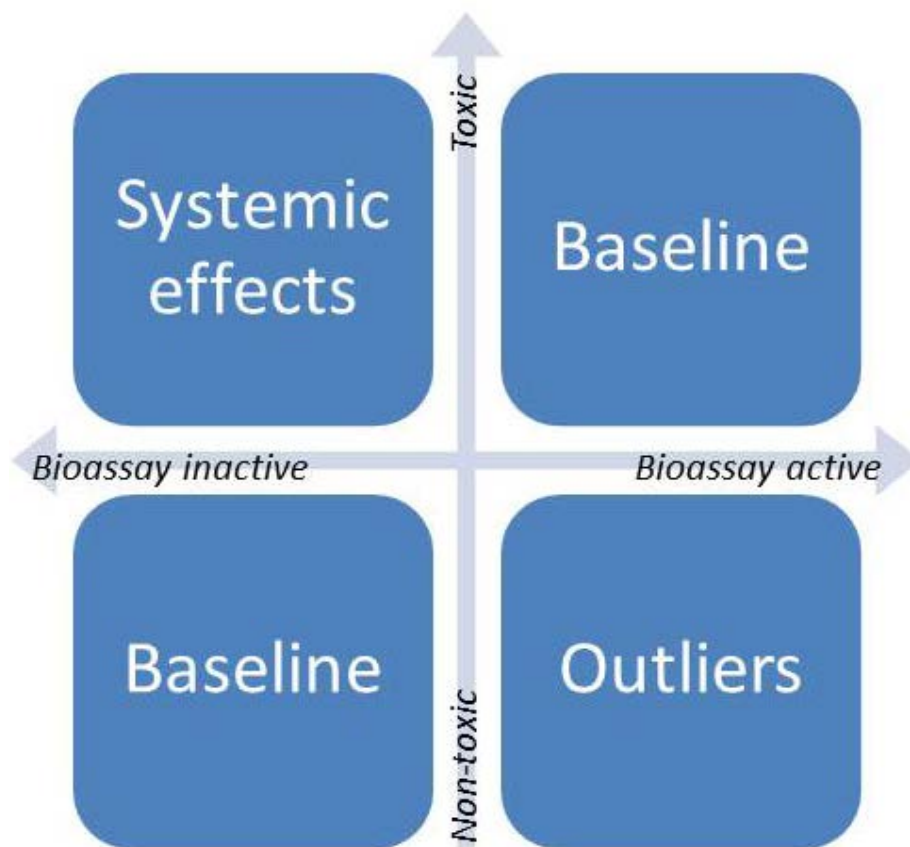


Figure 12 A potential *in vitro-in vivo* relationship in toxicology studies

In the current big data scenario, the most critical issue is to identify useful *in vitro* data. In principle, this could be done by a human expert using the knowledge of the design and quality of each particular bioassay (e.g., “Confidence Score” assigned during manual curation to each assay in ChEMBL). We, however, believe that data-driven approaches would provide more efficient ways. One possible strategy is to select assays based on their *in vitro-in vivo* relationships. Due to multiple mechanisms behind each toxicity phenotype, each bioassay is likely to show only partial correlation with *in vivo* effect. For example, if a bioassay represents a receptor that belongs to a toxicity pathway

relevant to the target animal toxicity, this bioassay should provide useful information, such as receptor/pathway perturbation. However, if compounds show inactive results in a particular bioassay, they can still be toxic since they may bind to other target sites (Figure 12). A previous study showed that the bioassay results had a low false-positive rate to predict the relevant animal toxicity.¹²¹ But the false-negative rate, on the contrary, is high. Based on this study, we recently developed an automatic bioassay system to evaluate and extract the relevant bioassay data based on the *in vitro-in vivo* relationship. For example, we could automatically extract 50 bioassays based on their correlation with the rat fetal growth retardation *in vivo* testing results from PubChem (accessed December 2013). Figure 13 shows the response profile, based on those 50 bioassays, for 107 compounds. The potential toxicants can then be prioritized by ranking the responses from these assays (Figure 13).

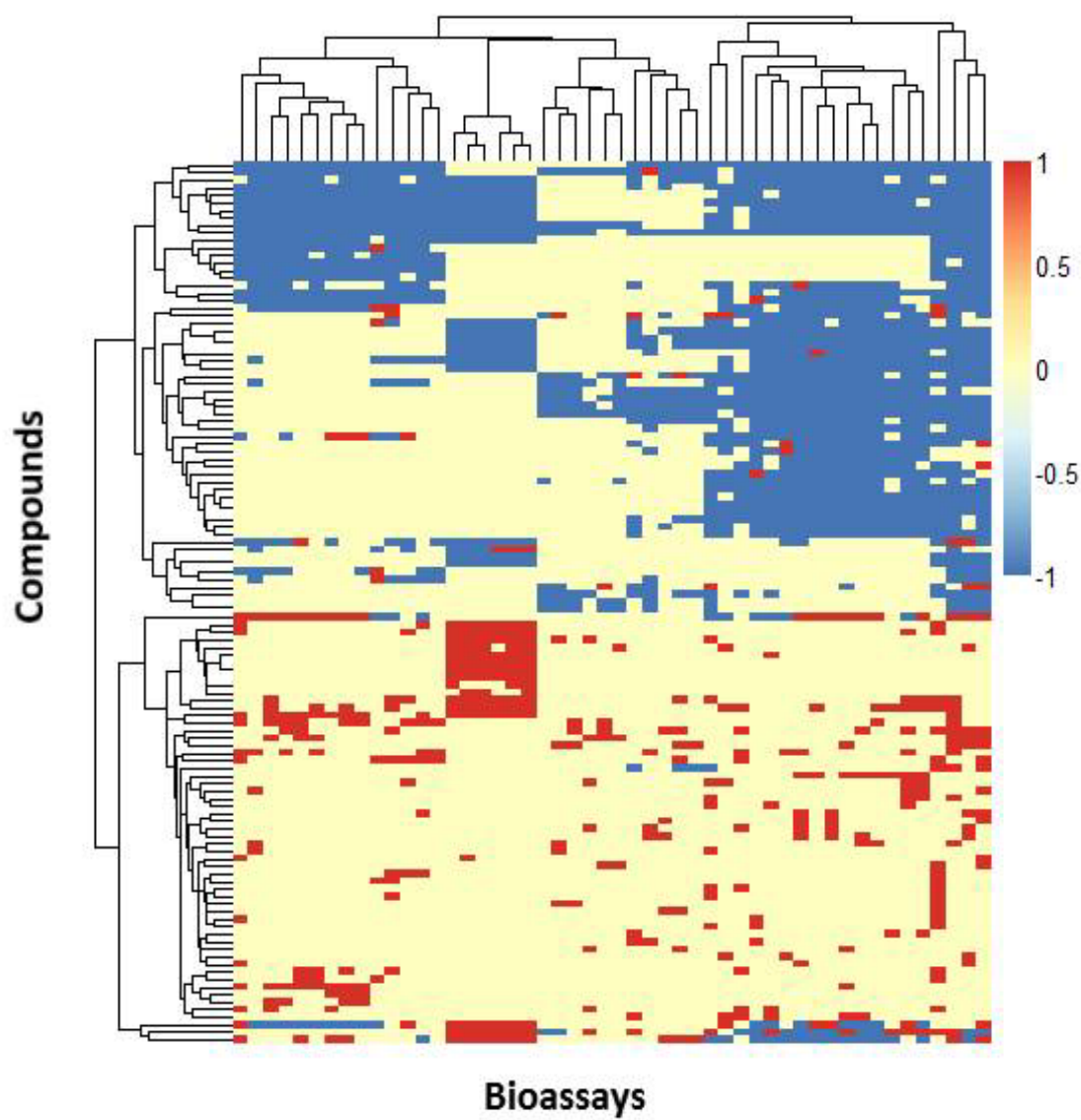


Figure 13 The response profiles of 50 PubChem bioassays for 107 compounds that may cause rat fetal growth retardation.

Section 7: Conclusions

Current innovative technologies enable rapid synthesis and high throughput screening of large libraries of compounds. Daily updated toxicity bioassay data have transformed current toxicology studies into a big data analysis. Fueled by the recent input from US and European governments, there are many ongoing data-generation and data-sharing programs, accompanied with the development of data curation (*e.g.*, Curvep,¹¹¹ <https://github.com/sedykh/curvep>) and automated data management (*e.g.*, ‘EMBL-EBI’ KNIME workflow nodes for ChEMBL, ‘rpubchem’ R package to PubChem) approaches that could be used to sample HTS data in meaningful formats to facilitate chemical toxicity studies. New scoring and modeling methods are also under way to take advantage of the massive amount of bioassay data. Although the use of bioassay data in most current toxicological research projects is still limited to a small portion of well sampled HTS data, several novel approaches have been reported to be able to access and integrate multiple bioassay data resources to profile toxicants. Under the current big data scenario, it is expected that modern toxicology research will be able to better estimate the systemic effects of compounds on the whole organisms and to translate this into better informed regulation of the toxicants for animals and humans.

References

1. Johnston, G. A. R. GABA(A) receptor channel pharmacology. *Curr. Pharm. Des.* **11**, 1867–1885 (2005).
2. Smith, K. S. & Rudolph, U. Anxiety and depression: Mouse genetics and pharmacological approaches to the role of GABA(A) receptor subtypes. *Neuropharmacology* **62**, 54–62 (2012).
3. Chebib, M. & Johnston, G. A. R. GABA-Activated Ligand Gated Ion Channels: Medicinal Chemistry and Molecular Biology. *J. Med. Chem.* **43**, 1429–1447 (2006)
4. Atack, J. R. Development of Subtype-Selective GABAA Receptor Compounds for the Treatment of Anxiety, Sleep Disorders and Epilepsy in *GABA and Sleep*. Editors: Monti, J. M., Pandi-Perumal, S. R., and Möhler, H. 25–73 (2010). doi:10.1007/978-3-0346-0226-6
5. Chebib, M. & Johnston, G. A. R. Proceedings of the Australian Neuroscience Society Symposium GABA and Glycine Receptors : From Neurochemistry to Neural Networks. *Clin. Exp. Pharmacol. Physiol.* **26**, 937–940 (1999).
6. Williams, L. R. *et al.* RDX binds to the GABA(A) receptor-convulsant site and blocks GABA(A) receptor-mediated currents in the amygdala: A mechanism for RDX-induced seizures. *Environ. Health Perspect.* **119**, 357–363 (2011).
7. Qian, M. *et al.* Neurosteroid analogues. 18. Structure-activity studies of ent-steroid potentiators of γ -aminobutyric acid type A receptors and comparison of their activities with those of alphaxalone and allopregnanolone. *J. Med. Chem.* **57**, 171–390 (2014).
8. Hosie, A. M., Wilkins, M. E. & Smart, T. G. Neurosteroid binding sites on GABA(A) receptors. *Pharmacol. Ther.* **116**, 7–19 (2007).
9. Akk, G. *et al.* Mechanisms of neurosteroid interactions with GABA(A) receptors. *Pharmacol. Ther.* **116**, 35–57 (2007).
10. Sousa, A. & Ticku, M. K. Interactions of the neurosteroid dehydroepiandrosterone sulfate with the GABA(A) receptor complex reveals that it may act via the picrotoxin site. *J. Pharmacol. Exp. Ther.* **282**, 827–833 (1997).
11. Maksay, G., Molnár, P. & Simonyi, M. Thermodynamics and kinetics of t-butylbicyclopophosphorothionate binding differentiate convulsant and depressant barbiturate stereoisomers acting via GABAA ionophores. *Naunyn. Schmiedeberg's. Arch. Pharmacol.* **353**, 306–313 (1996).

12. Bhutoria, S. & Ghoshal, N. A Novel Approach for the Identification of Selective Anticonvulsants Based on Differential Molecular Properties for TBPS Displacement and Anticonvulsant Activity: An Integrated QSAR Modelling. *QSAR Comb. Sci.* **27**, 876–889 (2008).
13. Gupta, S. QSAR Studies on Drugs Acting at the Central Nervous System. *Chem.Rev.* **89**, 1765–1800 (1989).
14. Wild, D. J. & Wiggins, G. D. Challenges for chemoinformatics education in drug discovery. *Drug Discov. Today* **11**, 436–439 (2006).
15. Vogt, M. & Bajorath, J. Chemoinformatics: A view of the field and current trends in method development. *Bioorg. Med. Chem.* **20**, 5317–5323 (2012).
16. Gasteiger, J. The central role of chemoinformatics. *Chemom. Intell. Lab. Syst.* **82**, 200–209 (2006).
17. Hrib, N. J. & Peet, N. Chemoinformatics : Are we exploiting this new science? *Drug Discov. Today* **5**, 483–485 (2000).
18. Varnek, A. & Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inform.* **30**, 20–32 (2011).
19. Trapani, G. *et al.* Propofol Analogues. Synthesis , Relationships between Structure and Affinity at GABA(A) Receptor in Rat Brain , and Differential Electrophysiological Profile at Recombinant Human GABA A Receptors. *J. Med. Chem.* **2623**, 1846–1854 (1998).
20. Rybczynski, P. J., Combs, D. W., Jacobs, K., Shank, R. P. & Dubinsky, B. gamma-Aminobutyrate-A receptor modulation by 3-aryl-1-(arylsulfonyl)- 1,4,5,6-tetrahydropyridazines. *J. Med. Chem.* **42**, 2403–2408 (1999).
21. Vapnik, V. & N. *In the nature of statistical learning theory.* (Springer, New York, 2000).
22. Breima, L. *Random Forests. Mach. Learn.* **45**, 5–32 (2001).
23. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* (2014). doi:10.1002/wcms.1183
24. Hogenkamp, D. J. *et al.* Synthesis and in Vitro Activity of 3 -Substituted-3 r -hydroxypregnan-20-ones : Allosteric Modulators of the GABA(A) Receptor. *J. Med. Chem.* **2623**, 61–72 (1997).
25. AID 248 - PubChem BioAssay Summary. at
<<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=248>>

26. Multicase Inc Software. at <<http://www.multicase.com/>>
27. Dalgaard, P. *Introductory statistics with R*. (Springer, 2008). at <http://books.google.com/books?hl=en&lr=&id=zZFCAAAAQBAJ&oi=fnd&pg=PR7&dq=Introductory+Statistics+with+R&ots=k6BF_sG_dc&sig=_YmViVa3gkfF2Ko2RutfObd0F2s>
28. Zheng, W. & Tropsha, A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comp. Sci.* **40**, 185–194 (2000).
29. Zhu, H. *et al.* Combinatorial QSAR modeling of chemical toxicants tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **48**, 766–784 (2008).
30. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **46**, 175–185 (2007).
31. Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A. & Tropsha, A. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* **25**, 1902–1914 (2008).
32. AID 71551 - PubChem BioAssay Summary. at <<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=71551>>
33. AID 71835 - PubChem BioAssay Summary. at <<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=71835>>
34. AID 71844 - PubChem BioAssay Summary. at <<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=71844>>
35. Jaworska, J., Nikolova-Jeliazkova, N. & Aldenberg, T. QSAR applicability domain estimation by projection of the training set descriptor space: A review. *Altern. Lab. Anim.* **33**, 445–459 (2005).
36. Chemical computing group. *Molecular Operating Environment (MOE)*. (Chemical Computing Group Inc., Montreal, Quebec, Canada (2013).
37. Molecular Descriptors Guide Description of the Molecular Descriptors Appearing in the PyChem software package. Electronic bookonline (link: pychem.googlecode.com/files/manual.pdf accessed 2014).
38. Pearlman, R. S. & Smith, K. M. Metric validation and the receptor-relevant Subspace Concept. *J. Chem. Inf. Comp. Sci.* **39**, 28–35 (1999).
39. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug

- Discovery . 1 . A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1**, 55–68 (1999).
40. Patel, D. V & Gordon, E. M. Applications of small-molecule combinatorial chemistry to drug discovery. *Drug Discov. Today* **1**, 134–144 (1996).
 41. Schreiber, S. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**, 1964–1969 (2000).
 42. Malo, N., Hanley, J. & Cerquozzi, S. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167–175 (2006).
 43. Inglese, J. & Auld, D. Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci.* **103**, 11473–11478 (2006).
 44. (NRC), N. R. C. *Toxicity Testing in the 21st Century: A Vision and a Strategy* . (The National Academies Press , 2007).
 45. Szymanski, P., Markowicz, M. & Mikiciuk-Olasik, E. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *Int. J. Mol. Sci.* **13**, 427–452 (2012).
 46. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. *Nat. Rev. Genet.* **12**, 224 (2011).
 47. Marx, V. Biology: The big challenges of big data. *Nature* **498**, 255–260 (2013).
 48. Swarup, V. & Geschwind, D. H. Alzheimer’s disease: From big data to mechanism. *Nature* **500**, 34–35 (2013).
 49. Austin, C., Brady, L., Insel, T. & Collins, F. NIH molecular libraries initiative. *Science* **306**, 1138–1139 (2004).
 50. Fliri, A. F., Loging, W. T., Thadeio, P. F. & Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **1**, 389–397 (2005).
 51. Fliri, A. F., Loging, W. T., Thadeio, P. F. & Volkmann, R. A. Biospectra analysis: Model proteome characterizations for linking molecular structure and biological response. *J. Med. Chem.* **48**, 6918–6925 (2005).
 52. Fliri, A. F., Loging, W. T., Thadeio, P. F. & Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci.* **102**, 261–266 (2005).

53. Smith, S. C., Delaney, J. S., Robinson, M. P. & Rice, M. J. Targeting chemical inputs and optimising HTS for agrochemical discovery. *Comb. Chem. High Throughput Screen.* **8**, 577–587 (2005).
54. Janzen, W. & Hodge, C. A Chemogenomic Approach to Discovering Target-Selective Drugs. *Chem. Biol. Drug Des.* **67**, 85–86 (2006).
55. Dix, D. J. *et al.* The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **95**, 5–12 (2007).
56. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**, 188–195 (2011).
57. Macarrón, R. & Hertzberg, R. P. Design and implementation of high throughput screening assays. *Mol. Biotechnol.* **47**, 270–285 (2011).
58. Stein, R. L. High-throughput screening in academia: The Harvard experience. *J. Biomol. Screen.* **8**, 615–619 (2003).
59. Macarron, R. Critical review of the role of HTS in drug discovery. *Drug Discov. Today* **11**, 277–279 (2006).
60. Judson, R., Houck, K. & Kavlock, R. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ. Health Perspect.* **118**, 485–492 (2010).
61. Kavlock, R. *et al.* Update on EPA's ToxCast program: Providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.* **25**, 1287–1302 (2012).
62. Collins, F. S., Gray, G. M. & Bucher, J. R. Toxicology. Transforming environmental health protection. *Science* **319**, 906–907 (2008).
63. Bucher, J. Regulatory Forum Opinion Piece* Tox21 and Toxicologic Pathology. *Toxicol. Pathol.* **41**, 125–127 (2013).
64. Shukla, S. J., Huang, R., Austin, C. P. & Xia, M. The future of toxicity testing: A focus on in vitro methods using a quantitative high-throughput screening platform. *Drug Discov. Today* **15**, 997–1007 (2010).
65. Klekota, J., Brauner, E., Roth, F. P. & Schreiber, S. L. Using high-throughput screening data to discriminate compounds with single-target effects from those with side effects. *J. Chem. Inf. Model.* **46**, 1549–1562 (2006).
66. Wang, Y. *et al.* PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37**, W623–W633 (2009).

67. Wang, Y. *et al.* An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* **38**, D255–D266 (2010).
68. Wheeler, D. & Barrett, T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **37**, D5–D15 (2007).
69. Wheeler, D. & Barrett, T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **39**, D38–D51 (2007).
70. Wheeler, D. & Barrett, T. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35**, D5–12 (2007).
71. Wheeler, D. & Barrett, T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **41**, D8–D20 (2007).
72. Wheeler, D. & Barrett, T. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **42**, D7–D17 (2007).
73. Wheeler, D. & Barrett, T. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **38**, D5–D16 (2010).
74. Gaulton, A., Bellis, L. & Bento, A. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
75. Judson, R., Martin, M. & Egeghy, P. Aggregating data for computational toxicology applications: the US environmental protection agency (EPA) Aggregated Computational toxicology Resource. *Int. J. Mol. Sci.* **13**, 1805–1831 (2012).
76. Judson, R., Richard, A., Dix, D. & Houck, K. ACToR—aggregated computational toxicology resource. *Toxicol. Appl. Pharmacol.* **233**, 7–13 (2008).
77. Fonger, G. & Stroup, D. TOXNET: A computerized collection of toxicological and environmental health information. *Toxicol. Ind. Health* **16**, 4–6 (2000).
78. Vinken, M., Pauwels, M. & Ates, G. Screening of repeated dose toxicity data present in SCC (NF) P/SCCS safety evaluations of cosmetic ingredients. *Arch. Toxicol.* **86**, 405–412 (2012).
79. Kohonen, P., Benfenati, E. & Bower, D. The ToxBank data warehouse: Supporting the replacement of in vivo repeated dose systemic Toxicity testing. *Mol. Inf.* **32**, 47–63 (2013).
80. Yang, C., Ambrosio, M., Arvidson, K. & Barlow, S. Development of new COSMOS oRepeatDose and non-cancer Threshold of Toxicological Concern

- (TTC) databases to support alternative testing methods. *Toxicol. Lett.* **221**, S80 (2013).
81. McHale, C., Zhang, L., Hubbard, A. & Smith, M. Toxicogenomic profiling of chemically exposed humans in risk assessment. *Mutat. Res.* **705**, 172–183 (2010).
 82. Waters, M. & Stasiewicz, S. CEBS—Chemical Effects in Biological Systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res.* **36**, D892–D900 (2008).
 83. Mattingly, C. & Rosenstein, M. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J. Exp. Zool.* **305**, 689–692 (2006).
 84. Mattingly, C. J. *et al.* The comparative toxicogenomics database: A cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* **92**, 587–595 (2006).
 85. Mattingly, C. & Colby, G. Promoting comparative molecular studies in environmental health research: an overview of the comparative toxicogenomics database (CTD). *Pharmacogenomics J.* **4**, 5–8 (2004).
 86. Mattingly, C. & Colby, G. The Comparative Toxicogenomics Database (CTD). *Environ. Health Perspect.* **111**, 793–795 (2003).
 87. Lamb, J., Crawford, E., Peck, D. & Modell, J. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
 88. Kleinstreuer, N. C. *et al.* In vitro perturbations of targets in cancer hallmark processes predict rodent chemical carcinogenesis. *Toxicol. Sci.* **131**, 40–55 (2013).
 89. Kleinstreuer, N. C. *et al.* Environmental impact on vascular development predicted by high-throughput screening. *Environ. Health Perspect.* **119**, 1596–1603 (2011).
 90. Sipes, N. S. *et al.* Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. *Toxicol. Sci.* **124**, 109–127 (2011).
 91. Martin, M. T. *et al.* Predictive model of rat reproductive toxicity from ToxCast high throughput screening. *Biol. Reprod.* **85**, 327–339 (2011).
 92. Reif, D., Martin, M. & Tan, S. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environ. Health Perspect.* **118**, 1714–1720 (2010).

93. Rotroff, D. M. *et al.* Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ. Health Perspect.* **121**, 7–14 (2013).
94. Sipes, N. S. *et al.* Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays. *Chem. Res. Toxicol.* **26**, 878–895 (2013).
95. Reif, D. M. *et al.* ToxPi GUI: An interactive visualization tool for transparent integration of data from diverse sources of evidence. *Bioinformatics.* **29**, 402–403 (2013).
96. Judson, R. S. *et al.* Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chem. Res. Toxicol.* **24**, 451–462 (2011).
97. Judson, R. S., Mortensen, H. M., Shah, I., Knudsen, T. B. & Elloumi, F. Using pathway modules as targets for assay development in xenobiotic screening. *Mol. Biosyst.* **8**, 531–542 (2012).
98. Betts, K. Tox21 to date: Steps toward modernizing human hazard characterization. *Environ. Health Perspect.* **121**, A228 (2013).
99. Attene-Ramos, M. S. *et al.* The Tox21 robotic platform for the assessment of environmental chemicals - from vision to reality. *Drug Discov. Today* **18**, 716–723 (2013).
100. King, A. & Jones, P. In-house assessment of a modified in vitro cytotoxicity assay for higher throughput estimation of acute toxicity. *Toxicol. In Vitro* **17**, 717–722 (2003).
101. Jones, P. A. & King, A. V. High throughput screening (HTS) for phototoxicity hazard using the in vitro 3T3 neutral red uptake assay. *Toxicol. In Vitro* **17**, 703–708 (2003).
102. Schirmer, K. *et al.* Developing a list of reference chemicals for testing alternatives to whole fish toxicity tests. *Aquat. Toxicol.* **90**, 128–137 (2008).
103. Sjöström, M., Kolman, A., Clemenson, C. & Clothier, R. Estimation of human blood LC50 values for use in modeling of in vitro-in vivo data of the ACuteTox project. *Toxicol. In Vitro* **22**, 1405–1411 (2008).
104. Piersma, A. H. *et al.* Quantitative extrapolation of in vitro whole embryo culture embryotoxicity data to developmental toxicity in vivo using the benchmark dose approach. *Toxicol. Sci.* **101**, 91–100 (2008).

105. Bjornsson, T. D. *et al.* The conduct of in vitro and in vivo drug-drug interaction studies: A Pharmaceutical Research and Manufacturers of America (PhRMA) perspective. *Drug Metab. Dispos.* **31**, 815–832 (2003).
106. Kolman, A. & Clemenson, C. Human in vivo database now on ACuteTox home page. *Toxicol. In Vitro* **27**, 2350–2351 (2013).
107. Kinsner-Ovaskainen, A. & Prieto, P. methods to be included in a testing strategy to predict acute oral toxicity: An approach based on statistical analysis of data collected in phase 1 of the ACuteTox project. *Toxicol. In Vitro* **27**, 1377–1394 (2013).
108. Clothier, R. *et al.* Comparative analysis of eight cytotoxicity assays evaluated within the ACuteTox Project. *Toxicol. In Vitro* **27**, 1347–1356 (2013).
109. Kopp-Schneider, A., Prieto, P., Kinsner-Ovaskainen, A. & Stanzel, S. Design of a testing strategy using non-animal based test methods: Lessons learnt from the ACuteTox project. *Toxicol. In Vitro* **27**, 1395–1401 (2013).
110. Prieto, P. *et al.* The value of selected in vitro and in silico methods to predict acute oral toxicity in a regulatory context: Results from the European Project ACuteTox. *Toxicol. In Vitro* **27**, 1357–1376 (2013).
111. Zhu, H., Rusyn, I., Richard, A. & Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **116**, 506–513 (2008).
112. Sedykh, A., Zhu, H. & Tang, H. Use of in Vitro HTS-Derived ConcentrationaResponse Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ. Health Perspect.* **119**, 364–370 (2010).
113. Low, Y. *et al.* Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem. Res. Toxicol.* **24**, 1251–1262 (2011).
114. Murk, A., Rijntjes, E. & Blaauboer, B. Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals. *Toxicol. In Vito.* **27**, 1320–1346 (2013).
115. Jomaa, B., Aarts, J. M. M. J. & Haan, L. de. In vitro pituitary and thyroid cell proliferation assays and their relevance as alternatives to animal testing. *ALTEX* **30**, 293–307 (2013).

116. Chen, B. *et al.* Chem2Bio2RDF: A semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* **11**, 255 (2010).
117. Chen, B., Ding, Y. & Wild, D. J. Assessing Drug Target Association Using Semantic Linked Data. *Plos Comput. Biol.* **8**, e1002574 (2012).
118. Wild, D. J. *et al.* Systems chemical biology and the Semantic Web: What they mean for the future of drug discovery research. *Drug Discov. Today* **17**, 469–474 (2012).
119. Chen, B., Ding, Y. & Wild, D. Improving integrative searching of systems chemical biology data using semantic annotation. *J. Cheminformatics* **4**, (2012).
120. Fourches, D., Sassano, M. F., Roth, B. L. & Tropsha, A. HTS navigator: Freely accessible cheminformatics software for analyzing high-throughput screening data. *Bioinformatics* **30**, 588–589 (2014).
121. Zhu, H. *et al.* A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environ. Health Perspect.* **117**, 1257–1264 (2009).