PREDICTION OF CHEMICAL ACTIVITY AGAINST VARIOUS DISEASE-RELATED TARGETS WITH MACHINE LEARNING METHODS

by

Srilatha Sakamuru
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

______________________________ Dr. Iosif Vaisman, Committee Chair

______________________________ Dr. Ruili Huang, Committee Member

______________________________ Dr. Dmitri Klimov, Committee Member

______________________________ Dr. Donald Seto, Committee Member

______________________________ Dr. Iosif Vaisman, Director, School of Systems Biology

______________________________ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

______________________________ Dr. Fernando R. Miralles-Wilhelm, Dean, College of Science

Date: _________________________ Fall Semester 2020
George Mason University
Fairfax, VA

Prediction of Chemical Activity against Various Disease-Related Targets with Machine Learning Methods

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Srilatha Sakamuru
Master of Science
Johns Hopkins University, 2008

Director: Iosif Vaisman, Professor
Director, School of Systems Biology

Fall Semester 2020
George Mason University
Fairfax, VA

DEDICATION

This dissertation is dedicated to my beloved parents.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Dr. Iosif Vaisman for his guidance, valuable advices and I’m very grateful to him for being very patient in explaining every concept in detail. I would like to thank Dr. Ruili Huang for her amazing mentorship and always being available. I express my profound thanks to Dr. Dmitri Klimov and Dr. Donald Seto for their assistance and valuable time. I’m so privileged to have Drs. Huang, Klimov and Seto in my committee.

I would like to thank my husband Srinivas and daughter Nandini for their support and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

Quantitative High Throughput Screening	qHTS
Quantitative Structure Activity Relationship	QSAR
Machine Learning	ML
Opioid Receptor	OR
G-Protein Coupled Receptor	GPCR
Cyclic Adenosine Monophosphate	cAMP
Dimethyl Sulfoxide	DMSO
Cytochrome P450 3A4	CYP3A4
Estrogen Receptor 1	ESR1
Adrenoceptor Alpha 1A	ADRA1A
Mu Opioid Receptor	OPRM
Kappa Opioid Receptor	OPRK
Delta Opioid Receptor	OPRD
Centers for Disease Control and Prevention	CDC
Random Forests	RF
Naïve Bayes	NB
Support Vector Machines	SVM
Neural Networks	NN
eXtreme Gradient Boosting	XGBoost
National Center for Advancing Translational Sciences	NCATS
NCATS Pharmaceutical Collection	NPC
NCATS Pharmacologically Active Chemical Toolbox	NPACT
Molecular Access System	MACCS
Extended Connectivity FingerPrints	ECFP
Half-Maximal Activity	AC_{50}
Receiver Operating Characteristic	ROC
Area Under the ROC Curve	AUC-ROC
Mathews Correlation Coefficient	MCC
Positive Predictive Value	PPV
Applicability Domain	AD
Protein Data Bank	PDB

ABSTRACT

PREDICTION OF CHEMICAL ACTIVITY AGAINST VARIOUS DISEASE-RELATED TARGETS WITH MACHINE LEARNING METHODS

Srilatha Sakamuru, Ph.D.

George Mason University, 2020

Dissertation Director: Dr. Iosif Vaisman

High throughput screening (HTS) technologies led to the accumulation of large amount of biological data for a broad range of chemical compounds. The main goal of this study is to build machine learning models of chemical compounds activity against various disease related targets, including cytochrome P450 3A4 (CYP3A4), estrogen receptor 1 (ESR1), adrenoceptor alpha 1A (ADRA1A), and opioid receptors (OPRs) like mu (OPRM), kappa (OPRK) and delta (OPRD). The training sets consist of ~3,000 investigational and approved drugs for animal and human use with experimentally determined in vitro activity. For CYP3A4, ESR1, and ADRA1A targets, the compounds are represented both by their bioactivity and structural features and the models were validated using internal test set and the best performed models achieved an AUC-ROC of 0.90 (ESR1), 0.89 (ADRA1A), and 0.81 (CYP3A4). For OPRM, OPRK, and OPRD targets the compounds are represented only by their structural features and the best performing models have AUC above 0.85. This approach produced robust OPR

prediction models that can be applied to prioritize compounds from large libraries which will match or exceed the opioid analgesic properties, but will have lower addiction potential. The models identified several novel potent compounds as activators/inhibitors of OPRs that were confirmed experimentally. This work can open novel ways to address important biomedical and public health needs.

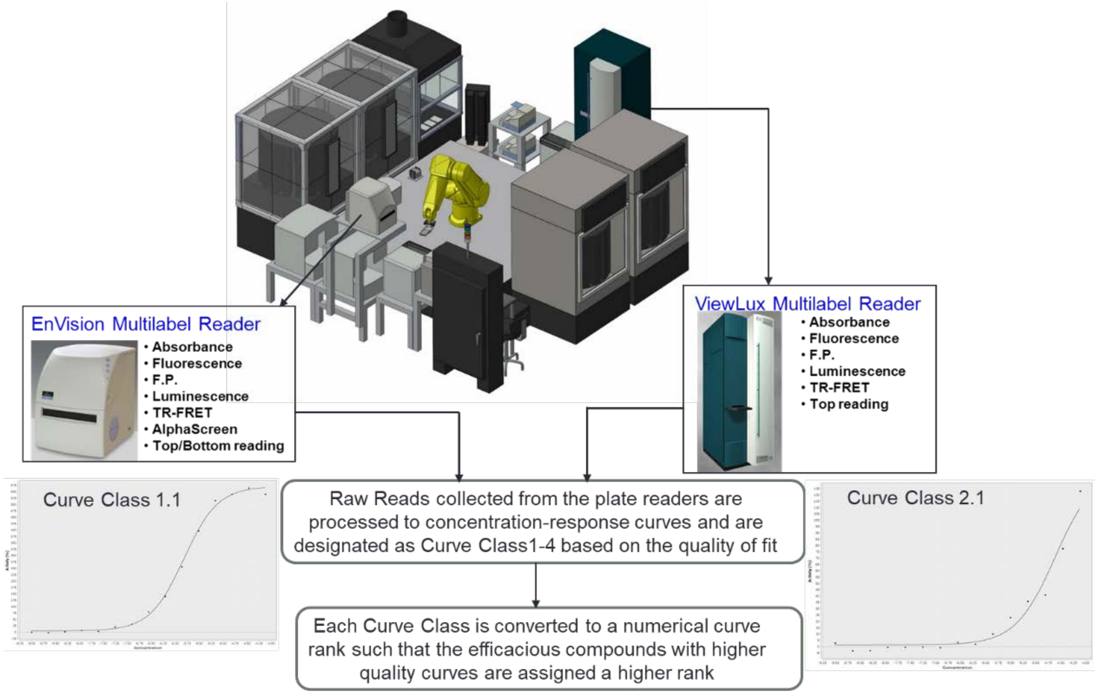
BACKGROUND

High throughput screening (HTS) is a method especially used in drug discovery, which enables testing thousands of small molecule, or large-scale in silico designed compounds against various biological targets of interest in a 1,536-well microplate format on a fully automated robotic platform [1]. HTS has emerged as an efficient alternative to animal testing in recent years especially in reducing time and costs involved in traditional *in vivo* studies. HTS is useful for identifying ligands for receptors, pharmacological targets and profiling the relationship between chemical structures and biological activities etc. Many of the *in vitro* assays like testing compound's activity on a particular biological target, pathway, disease, or cytotoxicity can be easily converted to HTS format [2]. The primary goal of HTS is to identify the active hits for the target of interest, and these hits are called as lead molecules that have a potential effect on the specific target at a very low concentration, to reduce the undesired chemical toxic effects. The raw reads from HTS are processed into a meaningful data for each compound whether it is active/inactive for a specific target/pathway.

One such HTS facility from National Institutes of Health (NIH) is at National Center for Advancing Translational Sciences (NCATS), which started more than a decade ago. NCATS runs HTS in a titration based approach called quantitative HTS (qHTS) methodology in which compound collection consists of compounds in serial dilutions and tests each compound at multiple concentrations against specific target [3]. Quantitative HTS generates a concentration-response curve for every compound from the large chemical library in a single assay and these results can be used for profiling the compounds based on their end targets. Quantitative HTS data yields half maximal effective concentration (EC50), Hill coefficient, and maximal response for each compound in the entire library enabling the assessment of structure activity relationship (SAR) [4]. This quantitative methodology is mainly to produce high quality data by reducing the frequency of false positives and false negatives which is the limiting step in traditional HTS and will be a crucial in computational modeling.

The qHTS platform has become a central aspect of the “Toxicology testing in the 21st Century” (Tox21) program [5]. Tox21 program is a federal collaboration involving NCATS, the National Toxicology Program (NTP), the Environmental Protection Agency (EPA), and the Food and Drug Administration (FDA) and is mainly aimed at developing better toxicity assessment methods by advancing the toxicity testing of chemical compounds from traditional animal toxicology studies to high throughput *in vitro* assays [6]. The goal of Tox21 project is to quickly and efficiently test whether certain chemicals have the potential to disrupt processes in the human body that leads to the adverse effects. The Tox21 chemical library consists of around 10,000 structurally diverse compounds. Most of these chemicals are those to which humans are exposed through the environment including drugs, industrial chemicals, pesticides, household products, food additives, etc. [7]. The compound library has each compound in fifteen concentrations starting with ~50µM final concentration and in 2-fold dilutions made in DMSO which results in a concentration range of four orders of magnitude. Several *in vitro* cell-based assays like target-specific and mechanism-based have been screened against Tox21 chemical libraries so far [8]. Initially all these assays have been optimized and miniaturized to a 1,536-well plate format. Then the assays will be transferred to the primary screening against Tox21 collection on the robotic platform.

A typical robotic platform (Fig. 1) consists of assay and compound plate incubators, micro-well plate dispensers, compound transfer station, and plate readers. In Fig. 1, the capabilities of the various detection technologies for the two readers are shown in the inserts.



Raw Reads collected from the plate readers are processed to concentration-response curves and are designated as Curve Class1-4 based on the quality of fit

Each Curve Class is converted to a numerical curve rank such that the efficacious compounds with higher quality curves are assigned a higher rank

Figure 1. A robotic platform pictured here. Quantitative HTS assay results in producing a concentration-response curves of each compound, different classes of curves shown in the inserts.

From the primary screen results, concentration-response curves for each compound were analyzed [9]. The data analysis from primary screening yields potency and efficacy values of each compound. Based on the concentration response data, each compound is assigned an activity outcome in each assay [4]. Activities of a compound across different assays form an activity profile. These activity profiles can be used to build models to predict compound target or mechanism of action. All quantitative HTS data from different primary screenings and follow-up studies are available and easily accessible from the NCATS in house database or publicly available on chemical databases like PubChem (https://pubchem.ncbi.nlm.nih.gov/), maintained by NCBI, a database for chemical compounds with their bioactivities. These publicly available big data have the potential to accelerate the drug discovery by incorporating machine learning methods.

Till date we have concentration-response data available from screening the Tox21 libraries against a large panel of cell based assays. From these entire cell based assay screenings, data can be aggregated for further analysis of the relationship between the drug activity profiles and their therapeutic indications or targets. In other words, models can be built using these in vitro activity profiles to predict other possible targets and/or indications of drugs from the collection across a broad array of human diseases.

The models built can be validated in our lab by selecting a few drugs and designing assays to test them against their predicted, potentially novel drug targets. The novelty of this study is mainly using the Tox21 collection for numerous targets that have been screened so far at our lab for developing new predictive models. The study is feasible by accessing the data that is available from NCATS database or on the public domains (PubChem). For further confirmation studies, *in vitro* (cell based) assays can be performed in our lab as an experimental validation to test the predictive models.

CHAPTER 1: PREDICTIVE MODELING OF CHEMICAL ACTIVITY AGAINST VARIOUS GENE TARGETS

Abstract

High throughput screening (HTS) tests thousands of small molecule compounds on a fully automated robotic system in a high density microplate format. The HTS facility at National Center for Advancing Translational Sciences (NCATS/NIH) produced millions of data points in terms of concentration-response for each chemical compound that has been tested on different biological targets and pathways. Such quantitative HTS data for a compound library called as NCATS Pharmaceutical Collection (NPC), comprising of ~ 3000 small molecule investigational drugs that are approved for human/animal use were used to develop models for CYP3A4, ESR1, and ADRA1A. These compounds were represented by activity and structural features. Activity features consists of each compound's bioactivity from HTS against a large panel of cell based assays, as the chemical compounds possess some moieties/patterns that make them active for a multi-targets. Structural features include molecular descriptors and fingerprints. Machine learning algorithms, including Naïve Bayes, Random Forest, Support Vector Machines, and Extreme Gradient Boosting were used to generate predictive models for three gene targets and the models were validated using internal test set. Models using the combination of *in vitro* activity-structural features performed better when compared to the activity or structural features alone.

Introduction

Using state-of-the-art HTS resources at NCATS, several thousands of small molecule compounds were screened on different cellular targets/pathways so far and thus resulted in generating millions of data points for all these small molecule compounds against different cellular targets/pathways. All these data will be uploaded in a timely manner on to the public databases like PubChem, from where the datasets on specific targets can be easily downloadable. In silico drug repurposing can be carried out in several potential ways in order to identify new indications of drugs, including drug (target-based), disease (knowledge-based) and treatment (pathway or network-based) oriented [10]. A drug target is usually a protein that is associated with a particular disease which could be addressed by a drug to produce a therapeutic effect. The most common drug targets include: G protein-coupled receptors, enzymes (like proteases, phosphatases, protein kinases, esterases etc.), ion channels, nuclear hormone receptors, and structural and membrane transport proteins. Target-based method is more significant for drug repurposing, as most targets link directly to disease mechanism. Several computational modeling studies were carried out to predict the structure activity relationship on target-specific like androgen and estrogen receptors [11, 12].

The compound collection used is NPC, consisting of around 3,254 small molecule drugs [13] and has fifteen titration points of each compound starting from 3.0nM to ~50µM final concentration and in 2-fold dilutions made in DMSO which results in a concentration range of four orders of magnitude. The NPCs comprises of small molecules (molecular weight < 1,500) (Fig. 2) which are soluble in DMSO.

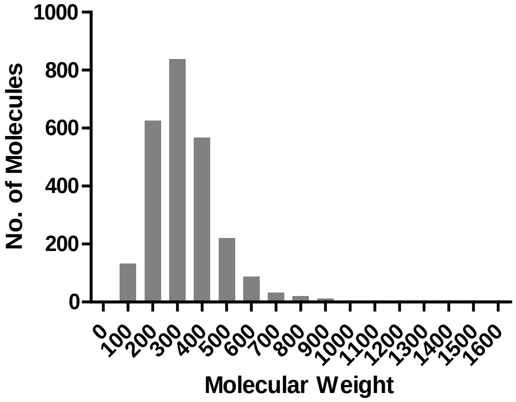


Figure 2. Distribution of drugs in NPC collection based on the molecular weight.

In this study, we propose the use of *in vitro* activity data from HTS of NPCs to build models for predicting the compounds activity on new targets. So, predicting the compound's activity through computational modeling for a drug repurposing study, means discovering novel therapeutic indications of existing/approved drugs and this approach reduces risks in terms of safety as the drugs are pre-approved for human and animal use, and also significantly reduces cost by bypassing the initial steps of drug development that normally takes years [14]. As a proof-of-concept, for our current study we first attempted to build predictive models for three targets: Cytochrome P450 3A4 (CYP3A4), Estrogen Receptor 1 (ESR1), Adrenoceptor Alpha 1A (ADRA1A), and the information for compounds to be active against these three targets was extracted from DrugBank [15].

Cytochrome P450 Family 3 Subfamily A Member 4 (CYP3A4) gene encodes CYP3A4 enzyme, a member of the cytochrome P450 superfamily. The cytochrome P450 (CYP) enzymes are membrane-bound hemeproteins that play a key role in metabolism of majority of drugs, steroids, carcinogens, and xenobiotics [16, 17]. Identifying the compounds that effect on CYP isozymes is useful to minimize the adverse drug reactions and toxicities in drug development process. CYP3A4 is the most important of all CYP450 enzymes. It metabolizes about half of all the drugs and it is predominantly found in the liver.

Estrogen receptor 1 (ESR1) gene encodes an estrogen receptor which is a nuclear receptor for estrogen hormone binding, and plays an important role in development, metabolic homeostasis and reproduction. Estrogen receptors (ERs) are involved in breast cancer and endometrial cancer [18]. Endocrine disrupting chemicals (EDCs) and their interactions with steroid hormone receptors like ER disrupts normal endocrine function. It is important to understand the effect of compound on ER.

Adrenoceptor Alpha 1A (ADRA1A) encodes for alpha-1A-adrenergic receptor, which is a G protein-coupled transmembrane receptor that binds to the catecholamine, epinephrine and norepinephrine and mediates the actions of peripheral and central nervous system [19, 20]. The adrenergic receptors are the targets for many therapeutic drugs like those for cardiovascular diseases, prostatic hypertrophy, asthma etc. [21].

The Four classification algorithms: Random Forest, Support vector machines, Naïve Bayes, and Extreme Gradient Boosting methods were employed to fit models for predicting compound’s activity for the three targets (Fig. 3).

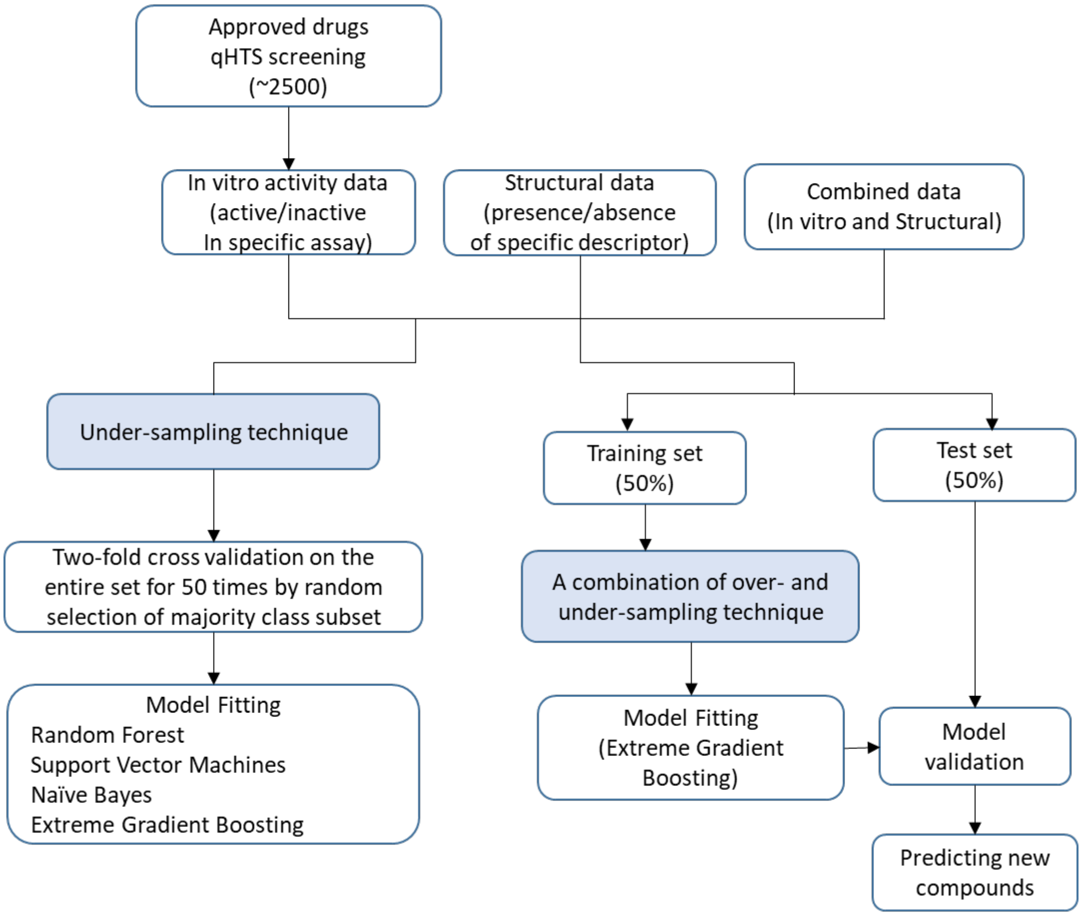


Figure 3. Flowchart of the study.

Materials and Methods

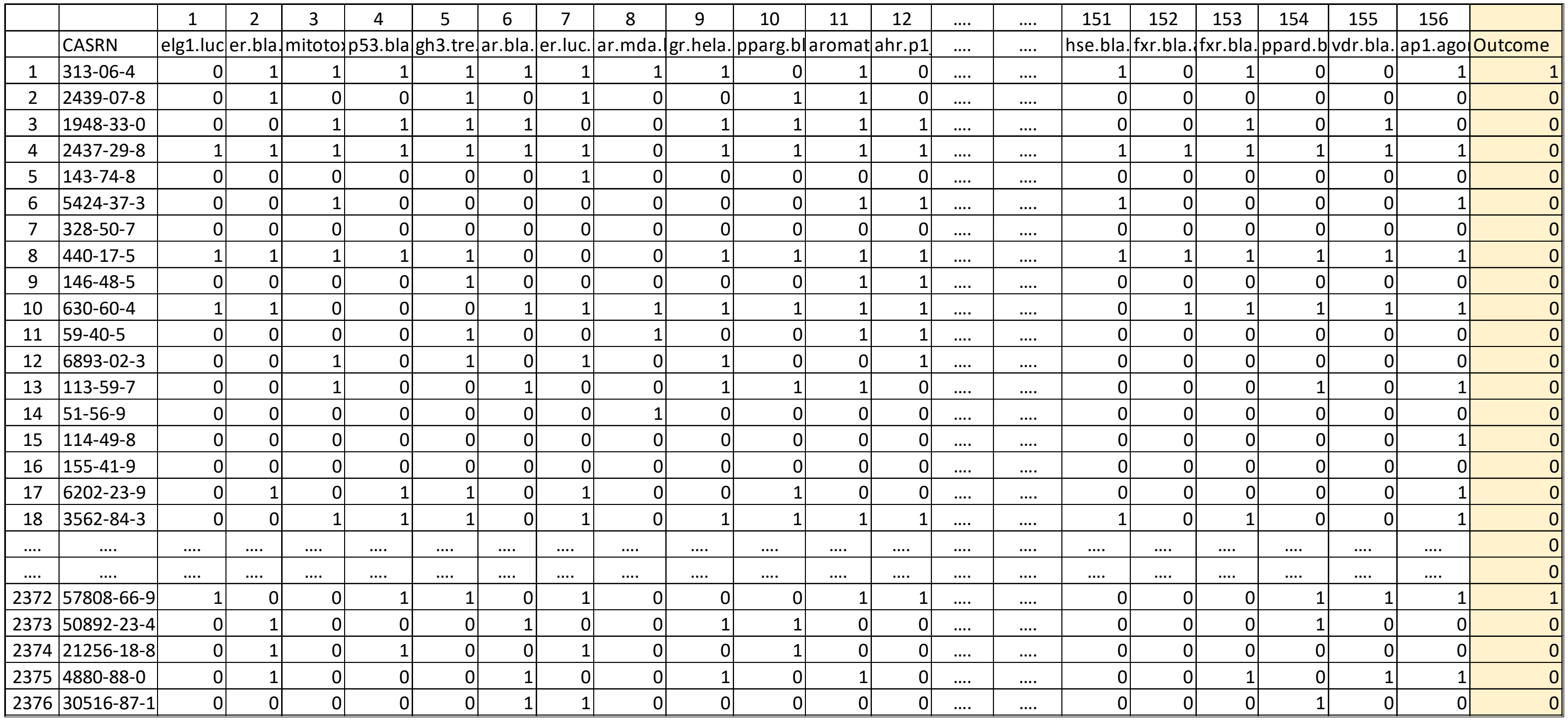
Datasets

The dataset used for our current study comprises of the results from HTS of NPCs against cell-based assays [9]. These include a large panel of *in vitro* assays covering a broad-array of pathways, including nuclear receptor signaling, stress response pathways, and developmental toxicology with fluorescence and luminescence detection readouts. The raw reads from quantitative HTS are processed into concentration response curves

[4] and are deposited into the NCATS database. Compounds are classified to classes 1-4 based on the type of concentration-response curves obtained. Curve classes are further converted to curve ranks, which are numeric measures of compound activity. For our current study, compounds with curve ranks > 0.5 are considered to be active. Each compound from the NPC collection shows its activity or inactivity in a particular assay and this *in vitro* assay data is termed as activity dataset in our current study and is of binaries (active/inactive) as shown in the data sheet of Table 1A. In addition to the *in vitro* activity data, structural features were extracted from Dragon 7 [22] software. This include Extended Connectivity Fingerprints (ECFP) in binaries with size 128 and molecular descriptors (constitutional indices, Table 2) in values with size 47. The structure dataset is represented in binaries (presence/absence of a particular structural feature) along with molecular descriptor values as shown in the data sheet of Table 1B. Also a combination of both activity and structure datasets were used to fit the models for predicting the compound's activity on new targets. In summary a total of 156 assay readouts (curve ranks for each compound from different cell based assays) were included in the activity dataset, 175 structural features were included in the structure dataset, and 331 combined features were included in the combination dataset. NPC collection consists of 3,254 small molecule compounds of which 817 were omitted due to missing *in vitro* assay data and in addition 61 were omitted due to missing structural molecular descriptors. A final of 2,376 compounds with activity, structure, and combination data were used in our current study.

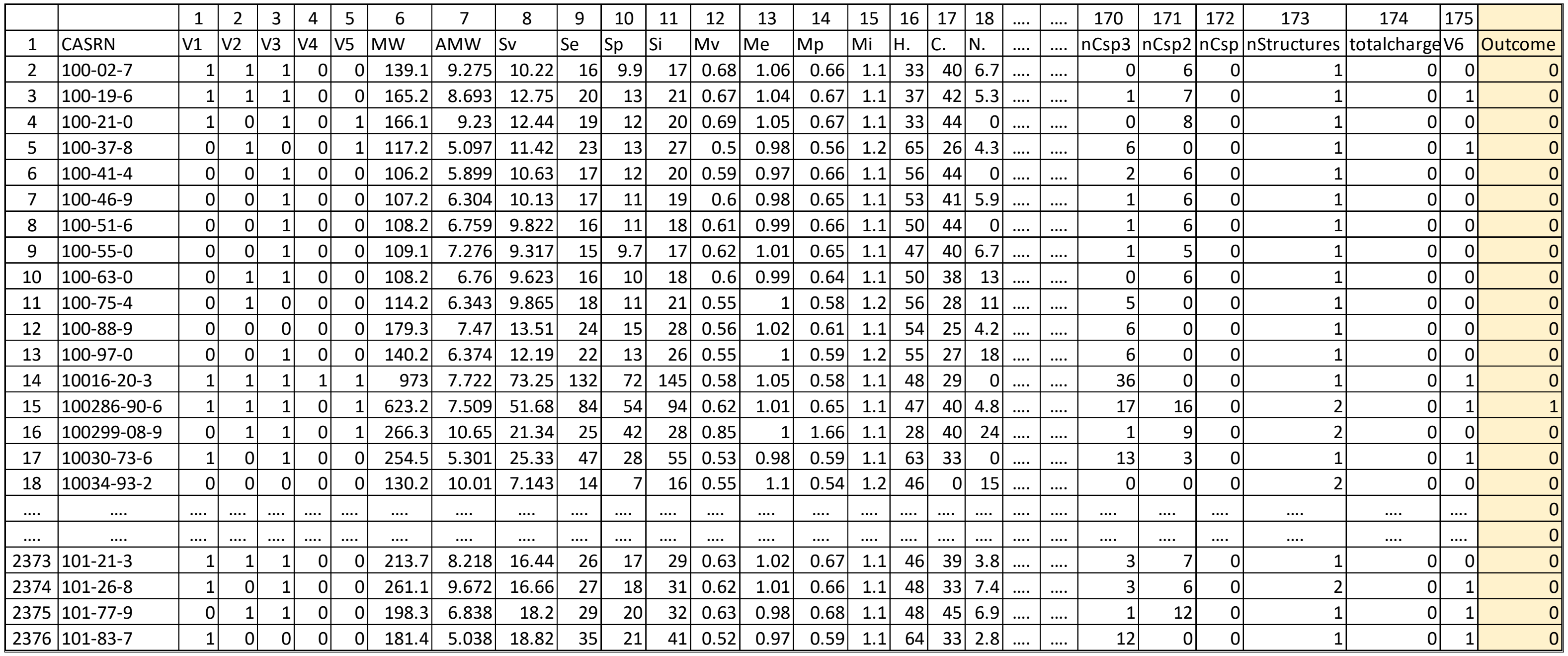
Table 1. A subset of the activity (A) and structure (B) datasets. Each row represents a specific compound (given is IDs) and each column is categorical (active/inactive) from a particular assay (A) or values (molecular descriptors) and categorical (presence/absence of structure fingerprints) (B). The class (Outcome) variable is of binaries (compounds active/inactive)

A.



		1	2	3	4	5	6	7	8	9	10	11	12	....	....	151	152	153	154	155	156	
	CASRN	elg1.luc	er.bla.	mitotox	p53.bla	gh3.tre.	ar.bla.	er.luc.	ar.mda.	gr.hela.	pparg.bl	aromat	ahr.p1	....	....	hse.bla.	fxr.bla.a	fxr.bla.	ppard.b	vdr.bla.	ap1.ago	Outcome
1	313-06-4	0	1	1	1	1	1	1	1	1	0	1	0	....	....	1	0	1	0	0	1	1
2	2439-07-8	0	1	0	0	1	0	1	0	0	1	1	0	....	....	0	0	0	0	0	0	0
3	1948-33-0	0	0	1	1	1	1	0	0	1	1	1	1	....	....	0	0	1	0	1	0	0
4	2437-29-8	1	1	1	1	1	1	1	0	1	1	1	1	....	....	1	1	1	1	1	1	0
5	143-74-8	0	0	0	0	0	0	1	0	0	0	0	0	....	....	0	0	0	0	0	0	0
6	5424-37-3	0	0	1	0	0	0	0	0	0	0	1	1	....	....	1	0	0	0	0	1	0
7	328-50-7	0	0	0	0	0	0	0	0	0	0	0	0	....	....	0	0	0	0	0	0	0
8	440-17-5	1	1	1	1	1	0	0	0	1	1	1	1	....	....	1	1	1	1	1	1	0
9	146-48-5	0	0	0	0	1	0	0	0	0	0	1	1	....	....	0	0	0	0	0	0	0
10	630-60-4	1	1	0	0	0	1	1	1	1	1	1	1	....	....	0	1	1	1	1	1	0
11	59-40-5	0	0	0	0	1	0	0	1	0	0	1	1	....	....	0	0	0	0	0	0	0
12	6893-02-3	0	0	1	0	1	0	1	0	1	0	0	1	....	....	0	0	0	0	0	0	0
13	113-59-7	0	0	1	0	0	1	0	0	1	1	1	0	....	....	0	0	0	1	0	1	0
14	51-56-9	0	0	0	0	0	0	0	1	0	0	0	0	....	....	0	0	0	0	0	0	0
15	114-49-8	0	0	0	0	0	0	0	0	0	0	0	0	....	....	0	0	0	0	0	1	0
16	155-41-9	0	0	0	0	0	0	0	0	0	0	0	0	....	....	0	0	0	0	0	0	0
17	6202-23-9	0	1	0	1	1	0	1	0	0	1	0	0	....	....	0	0	0	0	0	1	0
18	3562-84-3	0	0	1	1	1	0	1	0	1	1	1	1	....	....	1	0	1	0	0	1	0
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	0
....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	0
2372	57808-66-9	1	0	0	1	1	0	1	0	0	0	1	1	....	....	0	0	0	1	1	1	1
2373	50892-23-4	0	1	0	0	0	1	0	0	1	1	0	0	....	....	0	0	0	1	0	0	0
2374	21256-18-8	0	1	0	1	0	0	1	0	0	1	0	0	....	....	0	0	0	0	0	0	0
2375	4880-88-0	0	1	0	0	0	1	0	0	1	0	1	0	....	....	0	0	1	0	1	1	0
2376	30516-87-1	0	0	0	0	0	1	1	0	0	0	0	0	....	....	0	0	0	1	0	0	0

B.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	….	….	170	171	172	173	174	175	
1	CASRN	V1	V2	V3	V4	V5	MW	AMW	Sv	Se	Sp	Si	Mv	Me	Mp	Mi	H.	C.	N.	….	….	nCsp3	nCsp2	nCsp	nStructures	totalcharge	V6	Outcome
2	100-02-7	1	1	1	0	0	139.1	9.275	10.22	16	9.9	17	0.68	1.06	0.66	1.1	33	40	6.7	….	….	0	6	0	1	0	0	0
3	100-19-6	1	1	1	0	0	165.2	8.693	12.75	20	13	21	0.67	1.04	0.67	1.1	37	42	5.3	….	….	1	7	0	1	0	1	0
4	100-21-0	1	0	1	0	1	166.1	9.23	12.44	19	12	20	0.69	1.05	0.67	1.1	33	44	0	….	….	0	8	0	1	0	0	0
5	100-37-8	0	1	0	0	1	117.2	5.097	11.42	23	13	27	0.5	0.98	0.56	1.2	65	26	4.3	….	….	6	0	0	1	0	1	0
6	100-41-4	0	0	1	0	0	106.2	5.899	10.63	17	12	20	0.59	0.97	0.66	1.1	56	44	0	….	….	2	6	0	1	0	0	0
7	100-46-9	0	0	1	0	0	107.2	6.304	10.13	17	11	19	0.6	0.98	0.65	1.1	53	41	5.9	….	….	1	6	0	1	0	0	0
8	100-51-6	0	0	1	0	0	108.2	6.759	9.822	16	11	18	0.61	0.99	0.66	1.1	50	44	0	….	….	1	6	0	1	0	0	0
9	100-55-0	0	0	1	0	0	109.1	7.276	9.317	15	9.7	17	0.62	1.01	0.65	1.1	47	40	6.7	….	….	1	5	0	1	0	0	0
10	100-63-0	0	1	1	0	0	108.2	6.76	9.623	16	10	18	0.6	0.99	0.64	1.1	50	38	13	….	….	0	6	0	1	0	0	0
11	100-75-4	0	1	0	0	0	114.2	6.343	9.865	18	11	21	0.55	1	0.58	1.2	56	28	11	….	….	5	0	0	1	0	0	0
12	100-88-9	0	0	0	0	0	179.3	7.47	13.51	24	15	28	0.56	1.02	0.61	1.1	54	25	4.2	….	….	6	0	0	1	0	0	0
13	100-97-0	0	0	1	0	0	140.2	6.374	12.19	22	13	26	0.55	1	0.59	1.2	55	27	18	….	….	6	0	0	1	0	0	0
14	10016-20-3	1	1	1	1	1	973	7.722	73.25	132	72	145	0.58	1.05	0.58	1.1	48	29	0	….	….	36	0	0	1	0	1	0
15	100286-90-6	1	1	1	0	1	623.2	7.509	51.68	84	54	94	0.62	1.01	0.65	1.1	47	40	4.8	….	….	17	16	0	2	0	1	1
16	100299-08-9	0	1	1	0	1	266.3	10.65	21.34	25	42	28	0.85	1	1.66	1.1	28	40	24	….	….	1	9	0	2	0	0	0
17	10030-73-6	1	0	1	0	0	254.5	5.301	25.33	47	28	55	0.53	0.98	0.59	1.1	63	33	0	….	….	13	3	0	1	0	1	0
18	10034-93-2	0	0	0	0	0	130.2	10.01	7.143	14	7	16	0.55	1.1	0.54	1.2	46	0	15	….	….	0	0	0	2	0	0	0
….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	0
….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	….	0
2373	101-21-3	1	1	1	0	0	213.7	8.218	16.44	26	17	29	0.63	1.02	0.67	1.1	46	39	3.8	….	….	3	7	0	1	0	0	0
2374	101-26-8	1	0	1	0	0	261.1	9.672	16.66	27	18	31	0.62	1.01	0.66	1.1	48	33	7.4	….	….	3	6	0	2	0	1	0
2375	101-77-9	0	1	1	0	0	198.3	6.838	18.2	29	20	32	0.63	0.98	0.68	1.1	48	45	6.9	….	….	1	12	0	1	0	1	0
2376	101-83-7	1	0	0	0	0	181.4	5.038	18.82	35	21	41	0.52	0.97	0.59	1.1	64	33	2.8	….	….	12	0	0	1	0	1	0

Table 2. List of molecular descriptors included along with structural features

No.	Name	Mean	Std.dev.	Maximum	Minimum	Not zero	Description
1	MW	316.259	169.6026	2663.85	32.05	2437	molecular weight
2	AMW	8.368	3.82077	126.9	4.31	2437	average molecular weight
3	Sv	24.944	12.88803	109.035	2.768	2421	sum of atomic van der Waals volumes (scaled on Carbon atom)
4	Se	40.739	22.21555	195.227	2.022	2434	sum of atomic Sanderson electronegativities (scaled on Carbon atom)
5	Sp	27.259	14.80594	193.352	2.977	2437	sum of atomic polarizabilities (scaled on Carbon atom)
6	Si	45.645	25.12785	224.754	1.856	2437	sum of first ionization potentials (scaled on Carbon atom)
7	Mv	0.632	0.08853	1.794	0.461	2421	mean atomic van der Waals volume (scaled on Carbon atom)
8	Me	1.014	0.03275	1.304	0.85	2434	mean atomic Sanderson electronegativity (scaled on Carbon atom)
9	Mp	0.705	0.31629	6.983	0.46	2437	mean atomic polarizability (scaled on Carbon atom)
10	Mi	1.131	0.02437	1.36	0.899	2437	mean first ionization potential (scaled on Carbon atom)
11	GD	0.135	0.08477	1	0.015	2437	graph density
12	nAT	40.376	22.18571	196	2	2437	number of atoms
13	nSK	21.138	11.06353	110	2	2437	number of non-H atoms
14	nTA	4.903	3.35966	40	0	2383	number of terminal atoms
15	nBT	41.212	23.02974	196	1	2437	number of bonds
16	nBO	21.974	12.04486	109	1	2437	number of non-H bonds
17	nBM	9.085	6.34763	63	0	2287	number of multiple bonds
18	SCBO	27.501	14.73264	135	1	2437	sum of conventional bond orders (H-depleted)
19	RBN	4.518	4.02509	33	0	2121	number of rotatable bonds
20	RBF	0.104	0.07073	0.429	0	2121	rotatable bond fraction
21	nDB	1.86	1.83545	20	0	1876	number of double bonds
22	nTB	0.036	0.23092	6	0	74	number of triple bonds
23	nAB	7.189	6.1347	49	0	1732	number of aromatic bonds
24	nH	19.238	12.05375	111	0	2409	number of Hydrogen atoms
25	nC	15.1	8.54712	70	0	2417	number of Carbon atoms
26	nN	1.751	1.78139	20	0	1757	number of Nitrogen atoms
27	nO	3.192	3.13071	55	0	2104	number of Oxygen atoms
28	nP	0.037	0.21031	2	0	78	number of Phosphorous atoms
29	nS	0.285	1.0232	40	0	495	number of Sulfur atoms
30	nF	0.175	0.88072	24	0	197	number of Fluorine atoms
31	nCL	0.444	0.94071	10	0	681	number of Chlorine atoms
32	nBR	0.038	0.24267	4	0	73	number of Bromine atoms
33	nI	0.035	0.32798	6	0	34	number of Iodine atoms
34	nB	0.001	0.02864	1	0	2	number of Boron atoms
35	nHM	0.92	1.61268	50	0	1241	number of heavy atoms
36	nHet	6.038	4.31362	72	0	2414	number of heteroatoms
37	nX	0.692	1.33187	24	0	903	number of halogen atoms
38	H%	45.968	10.26065	70	0	2409	percentage of H atoms
39	C%	36.631	8.39871	61.538	0	2417	percentage of C atoms
40	N%	4.966	5.68005	75	0	1757	percentage of N atoms
41	O%	8.664	7.6996	75	0	2104	percentage of O atoms
42	X%	2.427	6.70519	100	0	903	percentage of halogen atoms
43	nCsp3	6.605	6.16383	56	0	2177	number of sp3 hybridized Carbon atoms
44	nCsp2	8.447	5.94185	49	0	2198	number of sp2 hybridized Carbon atoms
45	nCsp	0.046	0.27096	4	0	79	number of sp hybridized Carbon atoms
46	nStructures	1.372	1.04951	30	1	2437	number of disconnected structures
47	totalcharge	0.002	0.05357	2	0	4	total charge

Rebalancing the classes

The class variables are highly imbalanced because of the less percentage of compounds known to have activity against CYP3A4, ESR1 and ADRA1A targets and the percentages are at 7.9%, 1.1%, and 2.7% respectively (Table 3). For the initial model fitting, under sampling technique was used in order to rebalance the class variable, where a random subset of majority class for each target was selected equal to the size of the minority class. This process of random selection of majority class subset is repeated 50 times so that the unknown class will be sufficiently sampled.

Table 3. Distribution of different targets in the dataset

	Known actives	Unknowns
CYP3A4	187	2189
ESR1	25	2351
ADRA1A	64	2312

For the later part of model fitting in our current study, a combination of over- and under-sampling was carried out to rebalance the heavily skewed class variable of the 50% training set. This combination of sampling was done by random over-sampling of minority class (compounds known to have activity on these targets) and under-sampling of majority class (compounds with unknown activity for a given target) and thus leading to a more balanced dataset. So, the proportion of minority class in the resulting dataset was set to 0.5.

Machine learning methods

Four different classification algorithms namely, Random Forest, Support Vector Machines, Naïve Bayes, and Extreme Gradient Boosting have been employed to build the predictive models. The implementation of these algorithms were carried out using R 3.4.2.

Random Forest (RF), an ensemble-based method focuses on ensemble of decision trees. This method combines the base principles of bagging with random feature selection to add additional diversity to the decision tree models [23]. RF adds an additional layer of randomness to bagging where each node is split with the best subset of randomly chosen predictors. RF can handle extremely large datasets, as the ensemble uses only a small, random portion of the full feature set. Also RF can tend to be easier to use and less prone to overfitting. In our current study for building the RF classifier on training set, the number of trees to grow were specified to 100, and the parameter for the number of features to randomly select at each split was left to default settings (by default uses √number of features in the data). For making predictions on the test set, the type parameter selected was predicted probabilities.

Support Vector Machines (SVM) is a discriminative classifier defined by creating a flat boundary called a hyperplane, which divides the data points plotted in multidimensional representing feature values [24]. Like RF, SVM also deals well with a large number of features. For our current study, C-classification was implemented with cost of constraints violation set to 100 to build the models for calculating class probabilities and with rest all parameters set to default values.

Naïve Bayes (NB) is a simple probabilistic classification method based on the Bayes theorem given in eq. (1) and computes the conditional probabilities of a class (target) given independent features [25]. NB classifier assumes that all of the features in the dataset are equally important and independent and predicts the probability of the class based on the prior probability distribution of the class variable in training set.

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} \tag{1}$$

Where *P(A|X)* is the posterior probability of target *A* for a given variable *X*, *P(A)* and *P(X)* are the prior probabilities of target and variable, *P(X|A)* is the likelihood which is the probability of a feature for a given target. For our current study, the NB model built was used to predict the probability that a compound from the test dataset represented by either activity or structural features is active or not and returned the conditional probabilities for each target.

eXtreme Gradient Boosting (XGBoost) is also a tree-based technique, but differs from RF as this algorithm additionally tries to find optimal linear combination of trees (final model is the weighted sum of predictions of individual trees) in relation to given train data. This algorithm executes at faster speed resulting in excellent model performance. Most of the winning models in data mining competitions are built using gradient boosting algorithms [26]. For our current study, binary classification model was trained with a maximum depth of trees set at 2 and the number of cpu threads used were 2 for hundred passes on the data.

Feature Selection

To improve the performance of the predictive models, feature selection method is incorporated in the current study. It is a preprocessing stage for building the predictive models which involves the selection of a subset of relevant features [27]. The number of features (predictors) are reduced in order to increase the prediction accuracies, and reducing the variabilities.

A feature ranking and selection algorithm called Boruta, which is based on random forests is used to select the relevant features for a specific target [28]. It is a wrapper method that remove the predictors to find the optimal combination to maximize the model performance and selects features that are statistically significant. The strictness of the algorithm can be adjusted by adjusting the p value (default is 0.01) and maxRuns (number of times the algorithm is run, and the default value is 100). As shown in Figs. 4A-B, the columns in green are 'confirmed' features to be included and the ones in red are not. There are two blue bars representing ShadowMax and ShadowMin, which are not actual features but are used by the algorithm to decide if a particular variable is important or not. For each specific target, the number of features that are considered to be important are varying as given in Table 4.

Table 4. Number of features selected for each target

	Activity data (156)	Structure data (175)	Combined data (331)
CYP3A4	97	69	69
ESR1	38	51	42
ADRA1A	77	67	60

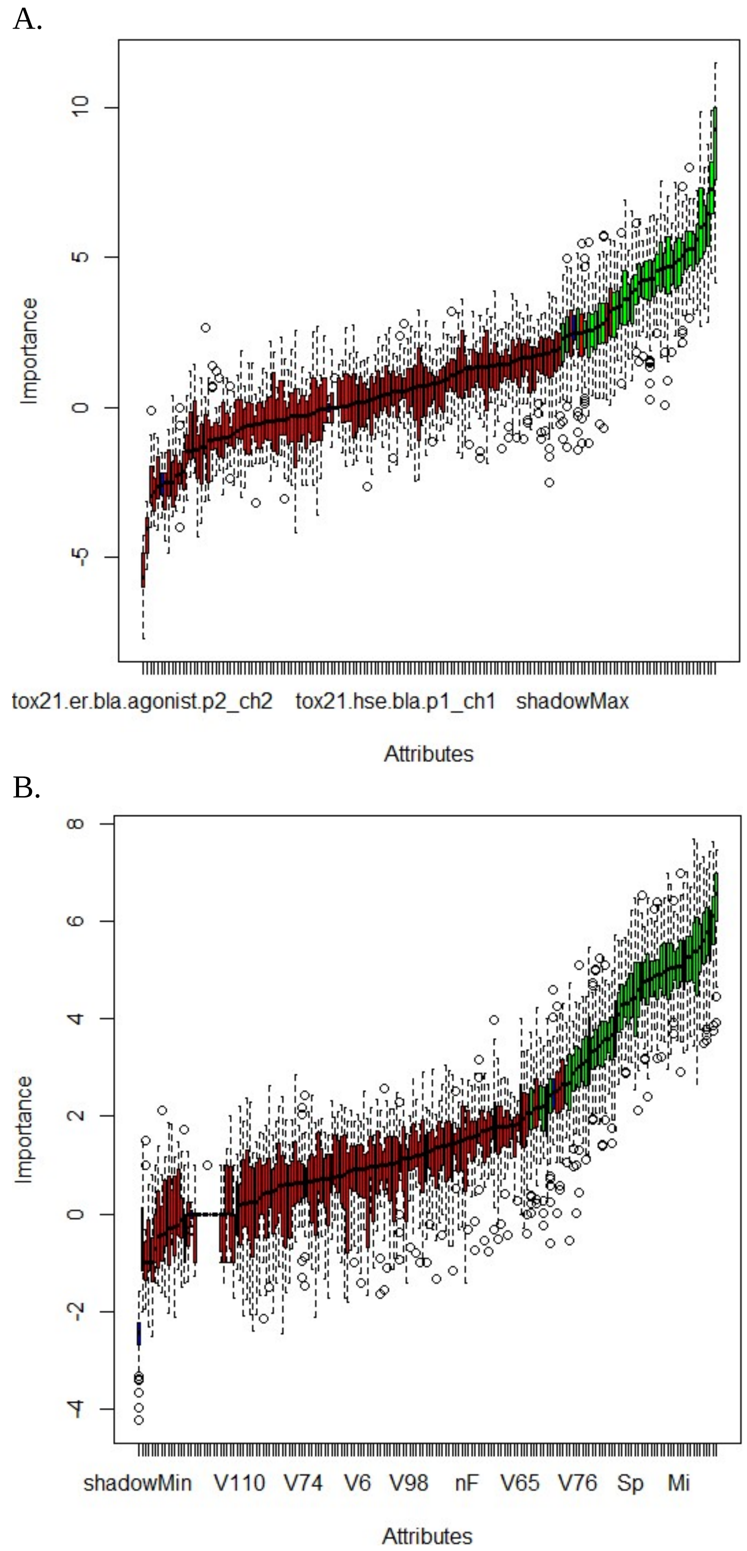


Figure 4. Plots reveal the importance of each of the features for Activity (A) and Structure (B) datasets of ESR1 target.

Evaluation of model performances

Internal test set was used for validation of the models. For our study the class of interest are the compounds known to have activity for the given target and termed as active compounds, while all others are inactive compounds. The predictions fall into the four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). All models were evaluated by sensitivity, specificity, and accuracy [29]. Sensitivity (true positive rate) measures the proportion of active compounds that were correctly classified as given in eq. (2) and specificity (true negative rate) measures the proportion of inactive compounds that were classified correctly in eq. (3). Accuracy in eq. (4) and Matthews Correlation Coefficient (MCC) in eq. (5) were calculated to evaluate the performance of the classification models. The calculation equations are:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$\text{MCC} = \frac{(TP\ X\ TN)-(FP\ X\ FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

In addition Area under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC) were used to assess the performance of the classification models [30]. ROC curve is defined on a plot with proportion of TP on the vertical axis and the proportion of FP on the horizontal axis, as these values are equivalent to sensitivity and (1-specificity) respectively. AUC-ROC ranges from 0.5 (classifier with no predictive value) to 1.0 (perfect classifier).

Results

Two-fold cross validation

In our current study as a preliminary step, four ML algorithms namely; RF, SVM, NB, and XGBoost were employed to build models for predicting the compounds activity for three targets (CYP3A4, ESR1 and ADRA1A). Both the *in vitro* activity and structural datasets were used for this preliminary study. The predictive performances were evaluated using a two-fold cross validation method by random sampling of the majority class for 50 times. A summary of the performances of the models, for the three targets are shown in the boxplots of Fig. 5A-C, which represents the averaged area under the curve (AUC) values of the 100 receiver operating characteristic (ROC) curves. RF and XGBoost classifiers slightly outperformed when compared to the other two models and of the three targets, ESR1 has resulted in higher AUC-ROC values. The average AUC-ROC values for ESR1 target with RF model are 083±0.07 and 0.83±0.11 and with XGBoost model are 0.81±0.07 and 0.80±0.1 for activity and structure datasets respectively. Whereas the average AUC-ROC values for CYP3A4 target with RF model are 0.74±.0.3 and 0.78±0.03 and with XGBoost model are 0.71±0.04 and 0.76±0.03 for activity and structure datasets respectively. Though the CYP3A4 target has higher percentage of known active compounds, when compared to ESR1, but in terms of performances the predictions are biased towards the majority class due to heavily skewed data of ESR1 target variable.

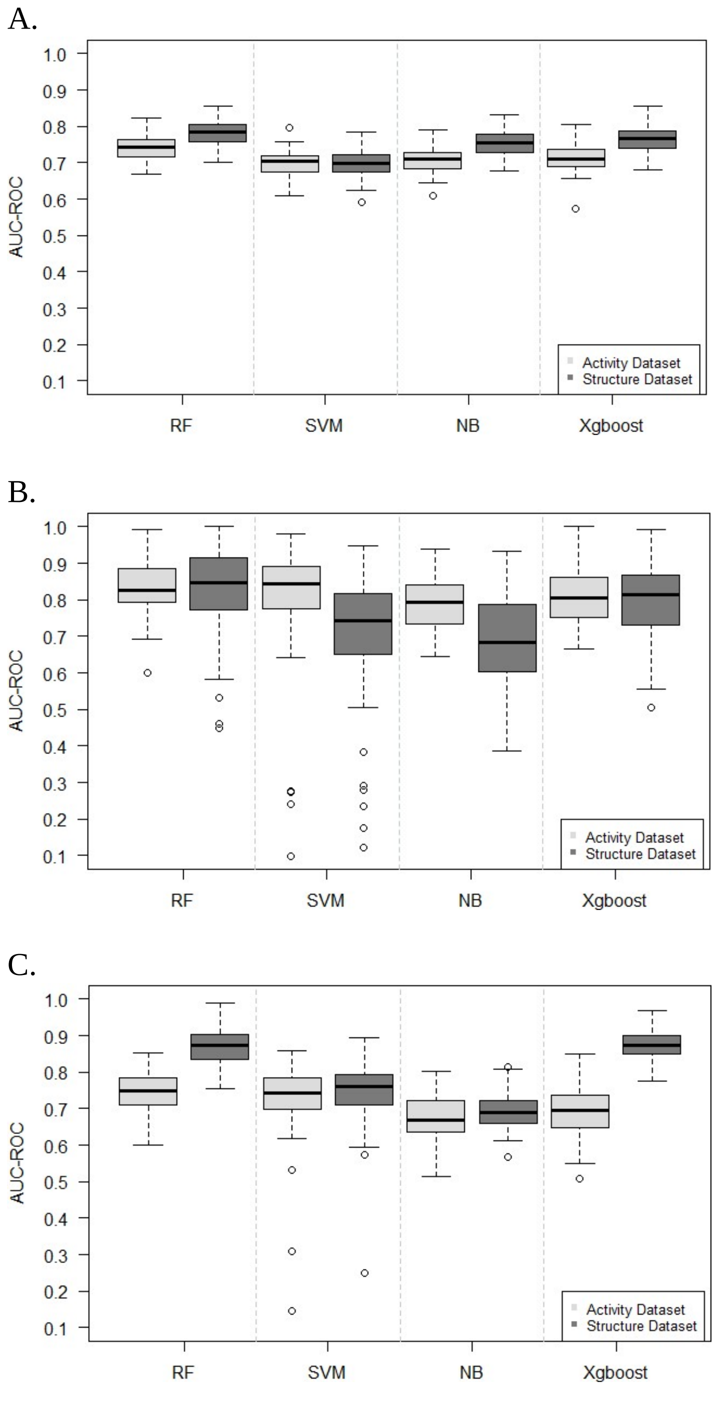
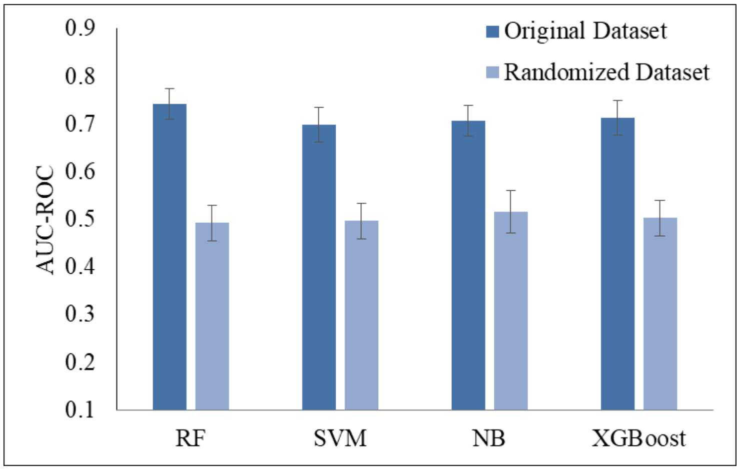


Figure 5. Boxplots of 100 AUC-ROCs predicted for CYP3A4 (A), ESR1 (B), and ADRA1A (C) targets. The horizontal lines inside the boxes represent median values.

Label randomization

The labels (active/inactive) for CYP3A4 target were randomly inter-replaced, such that the 50% actives were converted to inactives and 50% inactives were converted to actives. This label randomization procedure was carried out to check if our predictions were better than random predictions. The activity and structure datasets of CYP3A4 target was subjected to label randomization and the results were shown as bar graphs in Figs. 6A-B with comparison to the original datasets. The AUC-ROC values are equal to ~0.5 for the randomized datasets, which indicates that our models have an acceptable/excellent discrimination in classifying the two classes (active or inactive) of the three targets.

A.



B.

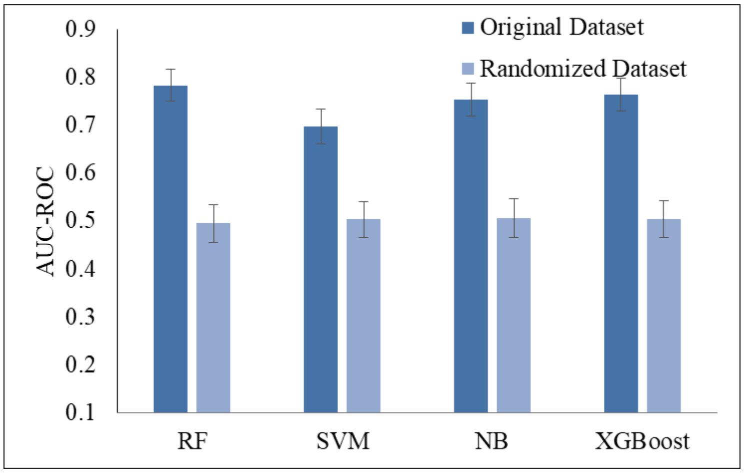


Figure 6. The bar graphs for activity (A) and structure datasets (B) of CYP3A4 target. The labels were randomized and decrease in AUC-ROC observed for randomized labels.

External validation

Further model fitting was carried out using chemical structural features with/without *in vitro* activity data and by using the XGBoost classification algorithm. All the three datasets (activity, structure, and combined) were divided into 50% train set and 50% test set. The results provided in Table 5, shows the predictive powers of XGBoost classifier when predicted on their respective test sets.

Table 5. Performance metrics of XGBoost classification algorithm for three targets

	CYP3A4			ESR1			ADRA1A		
	AUC-ROC	MCC	Accuracy	AUC-ROC	MCC	Accuracy	AUC-ROC	MCC	Accuracy
Activity	0.75	0.40	0.81	0.85	0.10	0.93	0.73	0.23	0.89
Structure	0.79	0.47	0.74	0.88	0.34	0.93	0.88	0.29	0.86
Combined	0.81	0.48	0.75	0.90	0.35	0.94	0.89	0.34	0.86

A total of nine XGBoost classification models for three targets (CYP3A4, ESR1 and ADRA1A) using three datasets (activity, structure and combined) were generated. For all three targets, sensitivity (true positive rate) and specificity values are varying proportionately based on the probability value cut-off. For positive predictions the cut-off value set for our current study is 0.5. The models built with activity data alone for CYP3A4 and ADRA1A yielded higher prediction accuracies of 0.81 and 0.89 respectively. Whereas predictive accuracies are higher for the models built with combined dataset for ESR1 which yielded 0.94, and not much significant difference observed in the prediction accuracies for the models built using activity alone (0.93) or in combination with structure data (0.93) (Table 5). ROC curves generated from predictions made on internal test set for all these nine XGBoost classification models are shown in Fig. 7A-C. The AUC-ROC values are higher for the models built with combined dataset except for ADRA1A target. The predicted models for CYP3A4 target yielded higher MCC values of 0.48 with the combination dataset of both activity and structural features. Whereas the predicted models for ESR1 and ADRA1A targets yielded an MCC of just 0.35 and 0.34 respectively, and this low score of MCC is due to the high imbalance of the class variables.

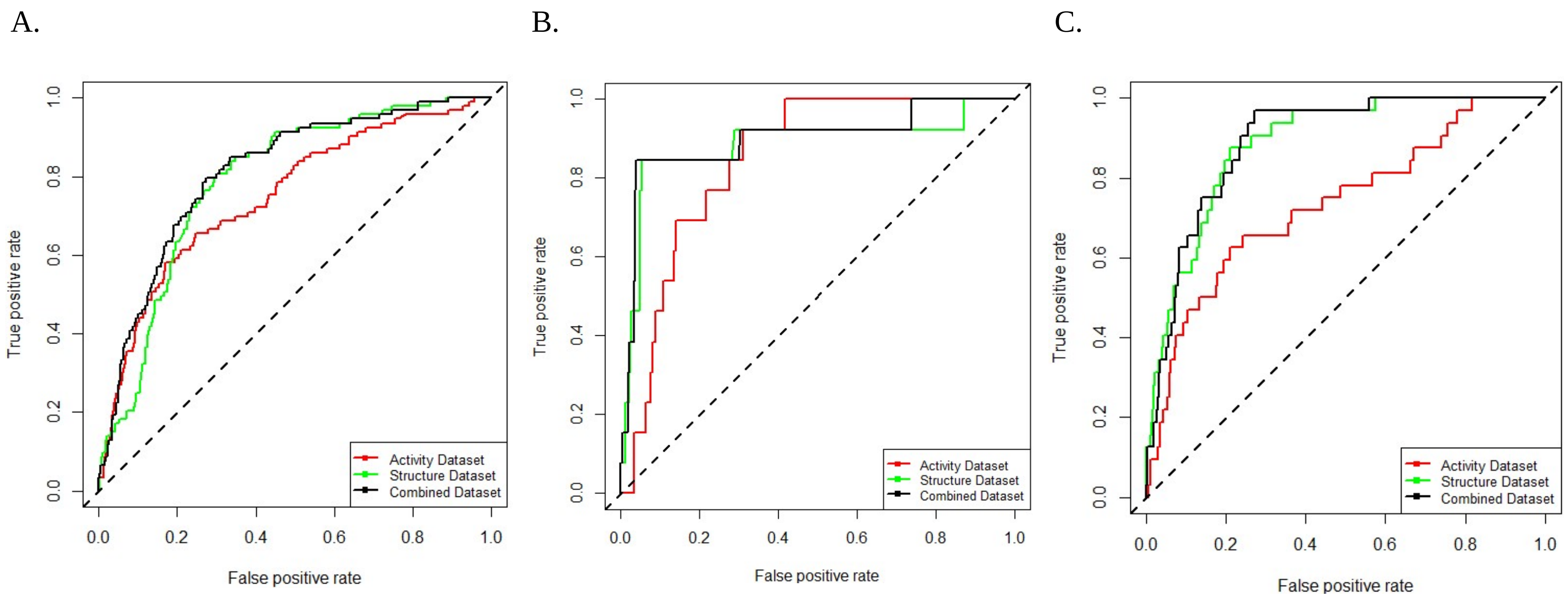


Figure 7. ROC curves generated by XGBoost classifier for CYP3A4 (A), ESR1 (B), and ADRA1A (C) targets.

In our current study, a predicted probability value of ≥ 0.5 is considered to be an active class for a particular target. The predicted activity for each model generated using all three datasets are clustered and shown in heat map (Fig. 8). Different clusters of compounds were predicted to be active for three targets and majority of those cluster compounds for each target were predicted to be active by three datasets, especially in CYP3A4 and ADRA1A. Whereas for ESR1, from different clusters were predicted to be active by activity-alone and structure alone datasets. So, in summary most of the compounds predicted by using structure dataset were also predicted to be active when activity data features are included (combination dataset). So, the classification models using structural features of the compounds along with *in vitro* activity features can help us to assess the impact of predictive modeling as a tool for drug repurposing.

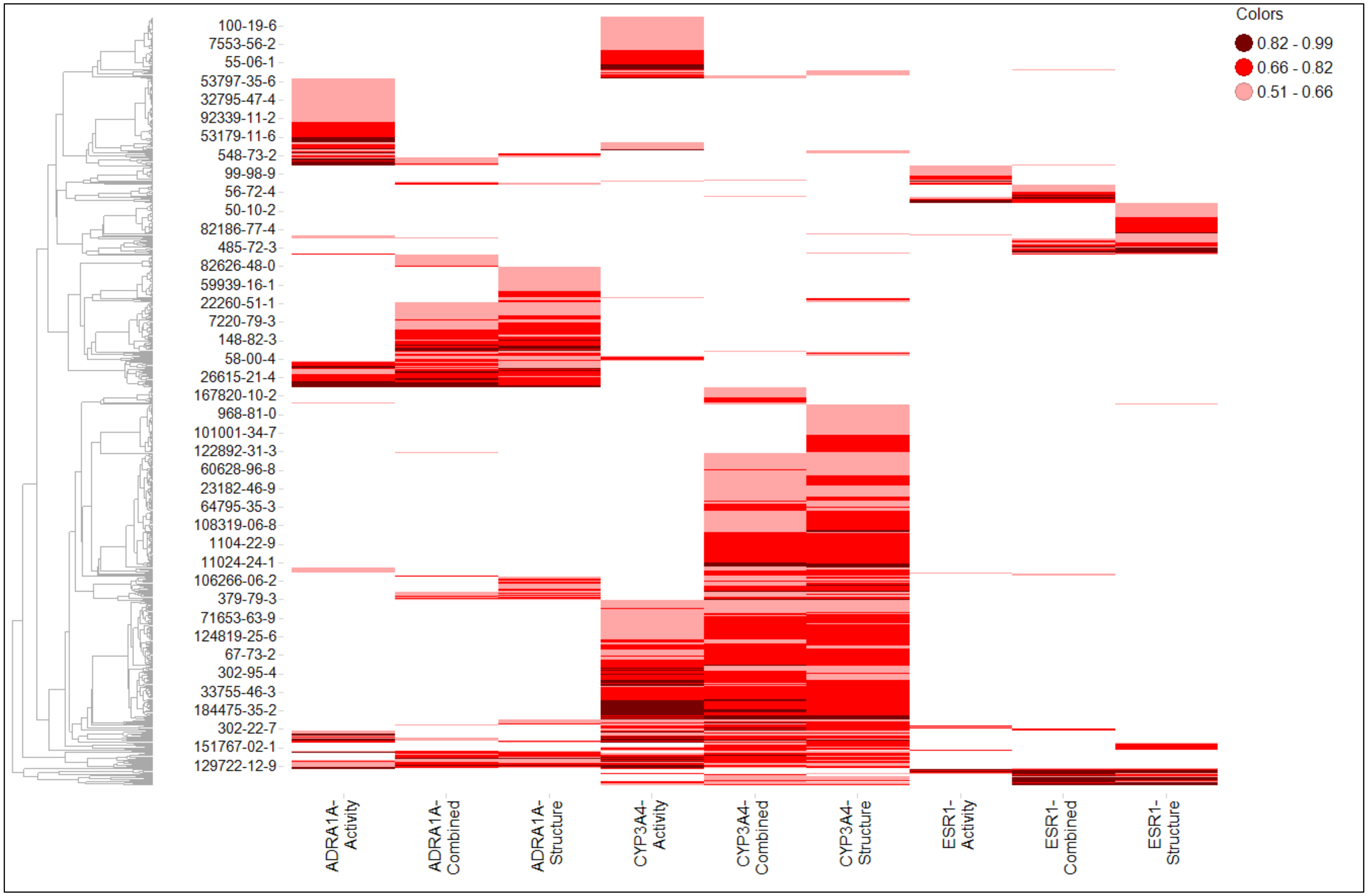


Figure 8. Predictions made for three targets. In the heat map, each row is a compound and each column is different dataset. The heat maps is colored by the probabilities of the predictions made. For a particular compound, dark red indicates the prediction to be ≥0.5. Compounds are grouped into clusters based on their active predictions.

Discussion

Computational models are built for three targets, including CYP3A4, ESR1, and ADRA1A using the bioactivity and structural data for representing each compound. The class information (active/inactive) for each compound against the specific target is obtained from DrugBank [15]. The *in vitro* data for compound's bioactivity included were from the assays against various targets like nuclear receptors, stress response signaling, phospholipidosis, DNA damage, cellular toxicity/apoptotic, ion channel and GPCR signaling pathways [7, 31]. The models performed slightly better when the structural data is combined with bioactivities (Fig. 7 and Table 5), rather than using activity or structure data alone. Hence using a wide-covered cell-based assay's data along with structural features aid in improved predictive performances of the compound's effect on new targets. The data generated from experimental screening of chemicals for biological activities were being used in computational modeling studies for rapid prioritization of chemicals [32, 33]. Using this *in vitro* assay data, studies on building predictive models for *in vivo* toxicity endpoints using a cluster-based approach [8] and on human toxicity based on drug adverse effects [34] were published from our center.

CYP3A4 was one of the extensively studied target for building QSAR models for predicting compound's effect on CYP inhibitions using quantitative HTS data and these models were built on of 5 different CYP isozymes (1A2, 2C9, 2C19, 2D6, and 3A4) by using molecular descriptors and SVM algorithms and also developed support vector classification (SVC) models for the 5 isozymes by a set of generic atom types and these both studies involved QSAR models [35, 36]. Prediction models have been developed for accurately predicting the metabolism of xenobiotics mediated by CYP3A4, 2D6, and 2C9 isozymes by using a novel concept of microsomal metabolic reaction system, which integrates the information for the site of metabolism and enzyme [37]. However our current study included the bioactivities data covering a large pool of cellular targets. Though the bioactivities data didn't include any of the CYP assays, but CYP3A4 is induced via nuclear receptors like aryl hydrocarbon receptor (AHR), pregnane X receptor (PXR), constitutive androstane receptor (CAR), retinoid X receptor (RXR), estrogen receptor (ER), and glucocorticoid receptor (GR) [38]. Our current datasets include the bioactivities of each compound against all these nuclear receptors. The combined datasets for CYP3A4 has shown improved predictive performance when compared to structure dataset alone and therefore we hypothesize that the compounds inducing these aforementioned nuclear receptors can aid in predicting each compound's effect on CYP3A4 activity.

Majority of our datasets consists of the activity from the assays of nuclear hormone receptors, therefore our current dataset can help to predict ESR1 target and this answers the question of yielding high predictive performances for this particular target when compared to CYP3A4 and ADRA1A, even though the class variable is highly biased (very less percentage of active compounds, Table 4). Several in silico QSAR models have been developed using machine learning and deep learning methods to predict endocrine-disrupting chemical (EDCs) binding with AR/ER and their effect on human health [39, 40]. The ER has been implicated in breast/ovarian cancers and there are 2 subtypes of ER, α and β. Both have similar expression patterns with some uniqueness in both types. EDCs and their interactions with steroid hormone receptors like ER disrupts the normal endocrine function. Therefore, it is important to understand the effect of environmental chemicals in ER signaling pathway.

Also for ADRA1A target, high AUC-ROCs are obtained for combined dataset rather than the bioactivities or structural data alone. ADRA1A is the member of the GPCR superfamily and is mainly involved in cell growth and regulation and the diseases associated with it include Horner's syndrome and benign prostatic hyperplasia (BPH) [41]. QSAR models have identified phthalimide-phenyl piperazines as a novel series of potent and selective ADRA1A antagonists [42]. In our bioactivities dataset, only GPCR-cAMP assays included are thyroid stimulating hormone receptor (TSHR) and thyrotropin-releasing hormone receptor (TRHR), which are the receptors for thyrotropin (thyroid stimulating hormone) and tripeptide thyrotropin releasing hormone respectively. But the high predictive performances for ADRA1A target may be due to the fact that GPCR pathways involve with multiple signaling cascades and networks in the cells and these signal transductions can lead to several cellular responses including transcription, growth, modulation within the pathway through downstream and toxicities as well [43].

The significance of our current study is that data from such wide covered cell-based bioassays along with structural features can be leveraged for computational predictive modeling studies to provide a valuable information regarding the target-specific and pathway based biological activities. Collective usage of data from such wide covered cell-based bioassays is primarily of a unique type of research when compared to the studies that have been published elsewhere [39, 44] and so can be of promising especially in predicting the chemical’s effect on new targets and furthermore the predicted models have the potential to predict the novel chemicals and their effect against CYP3A4, ESR1, and ADRA1A activities.

CHAPTER 2: PREDICTIVE MODELING OF CHEMICAL ACTIVITY AGAINST OPIOID RECEPTOR TARGETS

Abstract

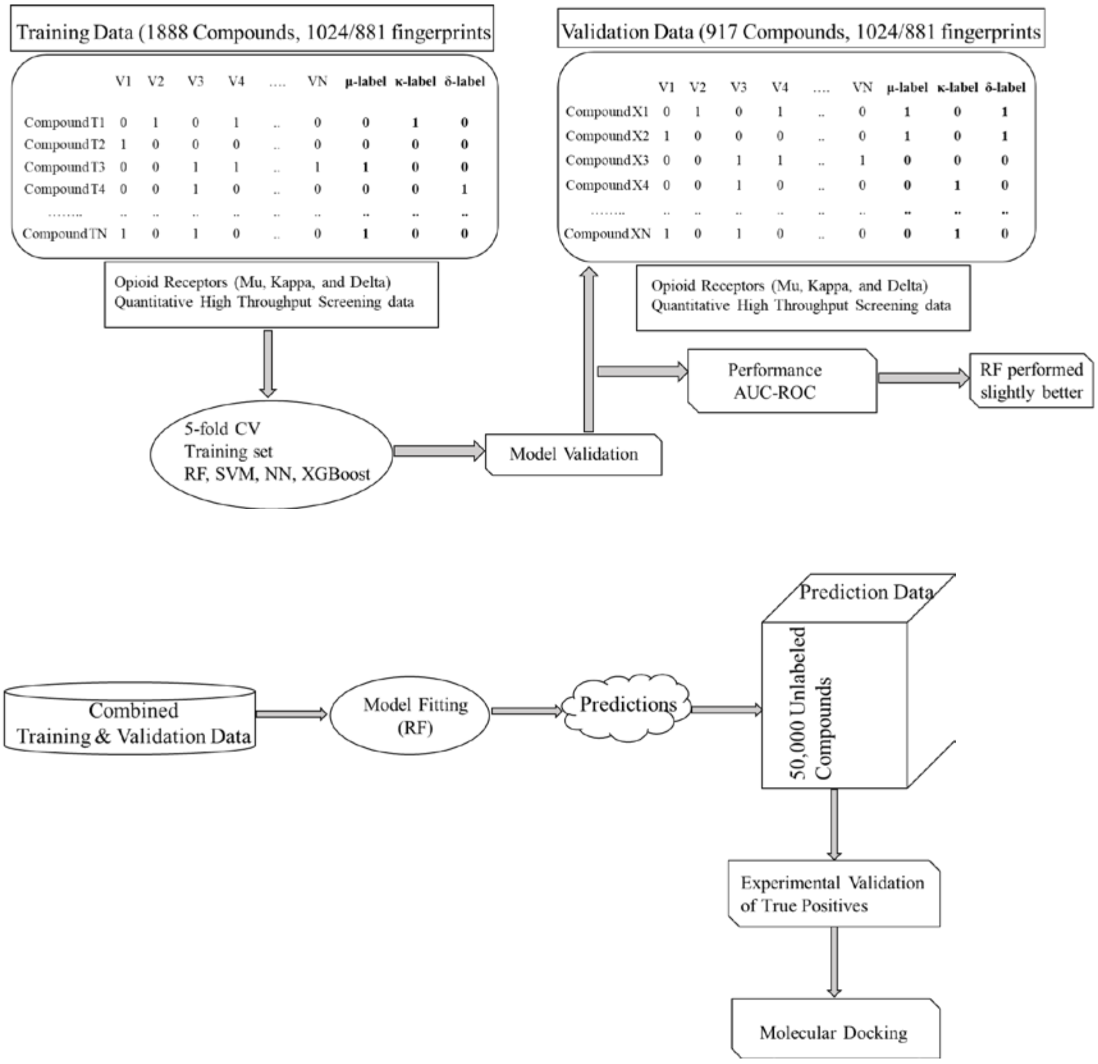
Opioid receptors (OPRs) are the main targets for the treatment of pain and related disorders. The opiate compounds that activate these receptors are effective analgesics but leads to adverse side effects and are highly addictive drugs of abuse. Search for alternative chemical structures that are analgesic, and reducing/avoiding the unwanted effects are urgent to relieve the public health crisis of opioid addiction. Here, we aim to develop models to predict the OPR activity of small molecule compounds based on chemical structure and apply these models to identify novel OPR active compounds. We used four different machine learning algorithms to build models based on quantitative high throughput screening (quantitative HTS) datasets of three OPRs in both agonist and antagonist mode. The best performing models were applied to virtually screen a large collection of compounds. The model predicted active compounds were experimentally validated using the same quantitative HTS assays that generated the initial training data. Random forest was the best classifier with the highest performance metrics and the OPRM-agonist model achieved the best performance with AUC-ROC (0.88) and MCC (0.7) values. The model predicted actives resulted in hit rates ranging from 2.3% (OPRD-agonist) to 15.8% (OPRM-agonist) after experimental validation. Comparing to the original assay hit rate, all models enriched % active by ≥ 2-fold. Our approach produced robust OPR prediction models that can be applied to prioritize compounds from large libraries for experimental validation. The models identified several novel potent

compounds as activators/inhibitors of OPRs that were confirmed experimentally. The potent hits were further investigated using molecular docking to find the interactions of the novel ligands in the active site of the corresponding OPR.

Introduction

Opioid receptors (OPRs) belong to the superfamily of G protein-coupled receptors (GPCR), consisting of 3 main classical types: mu (OPRM), kappa (OPRK), and delta (OPRD). These receptors are important for expressing pain transmission and modulation pathways, and are largely distributed in the central nervous system, while to a less extent in the periphery including gastrointestinal tract, heart and immune system etc. [45]. OPRs are activated both endogenously and exogenously. The endogenous ligands include the peptides: endorphins, dynorphins, and enkephalins for OPRM, OPRK, and OPRD respectively [46]. Whereas the exogenous opioid drugs include codeine, fentanyl, hydrocodone, methadone, morphine, oxycodone, buprenorphine, naloxone, naltrexone etc. with varying effect on different receptor types [47]. Most of these drugs that are administered as analgesics have side effects leading to addiction and drug abuse [48, 49]. In recent years there is a statistically significant increase in drug overdose death rate. According to the Centers for Disease Control and Prevention (CDC), more than 67,000 drug overdose deaths occurred in the United States in 2018 and opioid-involved overdose accounted for ~70% of the total drug overdose deaths [50]. Mainly the compounds targeting OPRM are known to produce several side effects that could be fatal. The search for analgesics with fewer side effects and/or for compounds targeting OPRK and OPRD has emerged as an alternative to produce safer drugs [51].

The vast amount of data generated from high-throughput screens are commonly used as training data for developing quantitative structure-activity relationship (QSAR) models to predict the activity of novel chemicals on biological targets using machine learning techniques [33, 52]. In order to build predictive models for the identification of novel activators/inhibitors of OPRs, we screened a collection of ~3000 approved drugs against three OPRs, OPRD, OPRK and OPRM, in a quantitative HTS format in both agonist and antagonist mode. Quantitative HTS generates a concentration-response for every compound in the primary screen producing high quality data that are ideal for training machine learning models. Several research works were published in the past using quantitative HTS data to build predictive models for various endpoints using machine learning algorithms and produced robust models [53-55]. In this study, we developed predictive models to identify activators/inhibitors of OPRs using the quantitative HTS assay data as training datasets. Six QSAR models were developed, which were trained with the experimental quantitative HTS datasets of agonist/antagonist modes of OPRM, OPRK, and OPRD. The models with good performance were applied to virtually screen our large in-house collections of 49,018 compounds to identify potential new OPR actives. The model predicted active compounds were validated experimentally. The potent actives were further evaluated by docking them to the crystal structures of the respective OPRs to study their interactions. Several independent research groups have identified novel hits through docking, such as discovery of active molecules against mu [56], and kappa [57] OPRs with new scaffolds that are unrelated to the known opioids. Through our current study (workflow shown in Fig. 9), we identified several potent compounds with novel structures that have the ability to activate/inhibit OPRs. Our models can be applied to make predictions on large chemical libraries, which lack experimental data, to prioritize the rapidly increasing drug-like new compounds for further testing.



Training Data (1888 Compounds, 1024/881 fingerprints

	V1	V2	V3	V4	….	VN	**µ-label**	**κ-label**	**δ-label**
CompoundT1	0	1	0	1	..	0	**0**	**1**	**0**
CompoundT2	1	0	0	0	..	0	**0**	**0**	**0**
CompoundT3	0	0	1	1	..	1	**1**	**0**	**0**
CompoundT4	0	0	1	0	..	0	**0**	**0**	**1**
……..	..	..	..	..	..	..	..	..	..
CompoundTN	1	0	1	0	..	0	**1**	**0**	**0**

Validation Data (917 Compounds, 1024/881 fingerprints

	V1	V2	V3	V4	….	VN	**µ-label**	**κ-label**	**δ-label**
CompoundX1	0	1	0	1	..	0	**1**	**0**	**1**
CompoundX2	1	0	0	0	..	0	**1**	**0**	**1**
CompoundX3	0	0	1	1	..	1	**0**	**0**	**0**
CompoundX4	0	0	1	0	..	0	**0**	**1**	**0**
……..	..	..	..	..	..	..	..	..	..
CompoundXN	1	0	1	0	..	0	**0**	**1**	**0**

Figure 9. The workflow of the study

Materials and Methods

In vitro qHTS assay

The CHO-K1 cells that express full-length human recombinant µ-, κ-, and δ-OPRs (HMOR, HKOR, and HDOR respectively) were purchased from Multispan, Inc. (Hayward, CA). The cells were cultured in DMEM/F12, 10% FBS, 100U/ml penicillin-100µg/ml streptomycin, and 10µg/mL puromycin (HMOR and HKOR) or 10µg/mL puromycin + 250µg/mL hygromycin (HDOR). The cell culture was maintained at 37°C, 5% CO_2, and 99% humidity. The cells were plated at 1,000/well in 3µL of the culture medium without the antibiotic marker in a 1,536-well white solid-bottom plates (Greiner Bio-One North America, NC) using Multidrop combi dispenser (Thermo Fisher Scientific Inc., Waltham, MA). The assay plates were incubated at 37°C for 18hr for cell adhesion to the plates, then 23nL of the positive control and test compounds were transferred to each well of the assay plates using Pintool station (Wako, San Diego, CA). The agonist positive controls used were DAMGO (Abcam, Cambridge, MA) for HMOR, and Dynorphin B peptide (Abcam) for HKOR and HDOR. The antagonist positive controls used were naloxone for HMOR and HKOR (Sigma-Aldrich, St. Louis, MO), and naltrindole (Sigma-Aldrich) for HDOR. Compound transfer was followed by the addition of 1µL of 0.5mM IBMX (3-Isobutyl-1-methylxanthine, Sigma-Aldrich) to each well of the assay plates using a Flying Reagent Dispenser (FRD, Aurora Discovery, Carlsbad, CA). Whereas for antagonist mode, 1µL of a mixture of 0.5mM IBMX and agonist positive control (2nM DAMGO for HMOR, 0.6nM or 2nM dynorphin B for HKOR or HDOR respectively) was added to each well of the assay plates using an FRD. The assay plates were incubated at 37°C for 20min and followed by the addition of 1µL of 1.0µM NKH477 (Sigma-Aldrich) to each well of the assay plates using an FRD. The assay plates were incubated at 37°C for 30min. Then the detection reagents were added at 2.5µL of cAMP-Cryptate (cAMP-Gi kit, Cisbio US, Inc., Bedford, MA), followed by 2.5µL of Anti-cAMP-d2 (cAMP-Gi kit, Cisbio) to each well of the assay plates using an FRD. After 1hr incubation at room temperature, the fluorescence intensity was quantified using Envision plate reader (PerkinElmer, Waltham, MA) at excitation 340nm and emissions at 665 and 620 nm. Data were expressed as ratio of 665nm/620nm.

qHTS data analysis

For primary data analysis, raw plate reads for each titration point were first normalized relative to positive control (agonist mode: 100%, antagonist mode: 0%) and DMSO only wells (agonist mode: 0%, antagonist mode: -100%). Percent activity is then calculated as: % Activity = ((Vtest compound–V_{DMSO})/(Vpositive control–V_{DMSO})) × 100, where Vtest compound are the values of compound wells, Vpositive control is the median value of the positive control wells, and V_{DMSO} is the median value of DMSO-only wells, and then corrected by applying a pattern correction algorithm using compound-free control plates (DMSO plates). Concentration-response titration points for each compound were fitted to the Hill equation and concentrations of half-maximal activity (AC_{50}) and maximal response (efficacy) values were calculated [9]. Compounds were designated as class 1–4 according to the type of concentration–response curve observed. Class 4 compounds were considered inactive. Compounds with class 1.1, 1.2, 2.1 curves or 2.2 curves with >40% efficacy in the agonist mode assays or >50% efficacy in the antagonist mode assays, and inactive or >6-fold less potent in the wild type counter screen were considered active. All other classes of compounds were considered inconclusive and excluded from modeling [4].

Compound library and datasets for modeling

The training set used in our study consists of 2805 compounds from the NCATS Pharmaceutical Collection (NPC) of approved and investigational drugs [13, 58]. The prediction set used in our study consists of 49,018 compounds from both Sytravon (a library of retired pharma screening collection containing a diversity of novel small molecules with an emphasis on medicinal chemistry-tractable scaffolds) and NPACT (NCATS Pharmacologically Active Chemical Toolbox; a library of annotated compounds that inform on novel phenotypes, cellular processes, and biological pathways) collections. The qHTS data obtained from *in vitro* assay of HMOR, HKOR, and HDOR cells screened against NPC were used to train and test models. *In vitro* qHTS assay data were randomly split into two sets, roughly two-thirds (1888) for model training and testing (cross-validation) and one-third (917) for external validation. Three different binary fingerprints: MACCS, PubChem, and ECFP were used to represent the compound structures, which are in 166, 881, and 1024-bit length, respectively. MACCS and PubChem fingerprints were generated from a workflow-based cheminformatics tool, KNIME-CDK [59] and ECFP from Dragon 7 software.

Supervised machine learning algorithms

Data processing was performed using R 3.5.3. QSAR models were developed using four machine learning algorithms: Random Forests (RF), Support Vector Machines

(SVM), Neural Networks (NN), and eXtreme Gradient Boosting (XGBoost) [23, 24, 26, 60] to classify the compounds based on their chemical structures for a given target. The packages used in R were randomForest, kernlab, nnet, and xgboost for implementing RF, SVM, NN, and XGBoost methods respectively, to run a 5-fold cross validation for 20 iterations. The key parameters chosen for RF were a default value of 500 for number of trees, and randomly selected variables for each split was set to the square root of the number of predictors. SVM algorithm implemented was a kernel-based method for classification with cost of constraints violation set to 100. A feed-forward NN with a single hidden layer of unit size 4 and decay of 0.1 were set as parameters. The boosting parameters for XGBoost were set to 0.05 (control the learning rate), 2 (maximum depth of trees) and the objective specified for the learning task was a logistic regression for binary classification, with 200 boosting iterations. For making predictions on the validation and prediction datasets, the Caret package was used for model fitting on the training set using RF algorithm and the hyper-parameters were selected based on the optimal model with the largest “ROC” metric [61].

Class rebalancing and feature selection

Random under- and over-sampling methods were employed to the training set. Under-sampling was implemented by random selection of majority class at each iteration to balance with the active class compounds and over-sampling was implemented using Rose package in R [62]. Features that have zero variance were eliminated and the significant features that show bivariate relationships were identified via a chi-square test [63]. Only those features that has p-value <0.05 were used in our current study [64].

Evaluation of model performances

The 5-fold cross validation performance of the training set was evaluated by computing area under the ROC (receiver operating characteristic) curve (AUC-ROC) values and were computed using “ROCR” package in R. The models generated using the training set were validated on the hold-out test set. The predictions fall into the four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The following measures were used to evaluate the model performances in addition to AUC-ROC [29, 65, 66]:

Sensitivity = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{FP+TN}$

Balanced Accuracy (BA) = $\frac{Sensitivity+Specificity}{2}$

Matthews Correlation Coefficient (MCC) = $\frac{(TP\ X\ TN)-(FP\ X\ FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Positive Predictive Value (PPV) = $\frac{TP}{TP+FP}$

The boxplots and ROC curves were generated using “ggplot2” and “ggpubr” packages in R. Compounds represented by PubChem fingerprints (categorical data) were clustered using k-modes method, an extension to k-means algorithm especially developed for categorical datasets [67]. The clustering was performed in R using klaR package by a simple-matching distance method. The structural similarity between two compounds in our study was measured by the Tanimoto score [68, 69].

Molecular docking

The three dimensional structures of OPRs bound to a ligand molecule were retrieved from Protein Data Bank (PDB) with the following PDB codes: 5C1M (OPRM-BU72 agonist), 4DKL (OPRM-βFNA antagonist), 6B73 (OPRK-MP1104 agonist), 4DJH (OPRK-JDTic antagonist), 6PT3 (OPRD-DPI287 agonist), and 4EJ4 (OPRD-naltrindole antagonist). Molecular docking was performed using Autodock Vina, an open source docking program [70]. Using AutoDock Tools, the ligands and the protein molecule (after the addition of hydrogen atoms) were saved in pdbqt format and the binding sites of the receptors were identified and a grid box was defined [71]. From the correctly aligned grid box that covers the entire active site of the receptor's binding pocket, the coordinates were saved. The center coordinates are different for each target, but the size of the grid box for each target is same (size_x: 40, size_y: 40, and size_z: 40). The active site amino acid residues had been indicated in the published crystal structures of each protein-ligand complexes like OPRM-BU72 [72], OPRM-βFNA [73], OPRK-MP1104 [74], OPRK-JDTic [75], OPRD-DPI287 [76], and OPRD-naltrindole [77]. The potent active ($AC_{50} \leq 2.5\mu M$) compounds from each set were docked into the active site of the respective OPRs. The pose of the docked ligand with the least binding affinity (kcal/mol) was selected and the structures of the protein-ligand docked complexes were analyzed and visualized using PyMOL tool.

Results

Assay performances and activity distribution

The compounds from the NPC library were screened at 7 concentrations in agonist and antagonist modes against three cells-based assays (OPRM, OPRK, and OPRD) to identify activators/inhibitors of OPRs. All the assays performed well with coefficient of variances (CV) <7.0%, and the signal/background (S/B) ratios were ≥2.0 for OPRM-agonist and OPRK-antagonist, and Z' factors were >0.40 for OPRM- & OPRD-agonists and OPRK-antagonist. The rest of the assays had lower S/B and Z' factors but the overall performances were compensated by the good CV values. The compound activity distributions in terms of active, inactive, or inconclusive for each OPR assay are shown in Fig. 10 [4]. The assays produced the highest active (hit) rates were OPRM-agonist (15%) and OPRK-antagonist (12%), and the OPRD-agonist (3%) assay yielded the lowest hit rate.

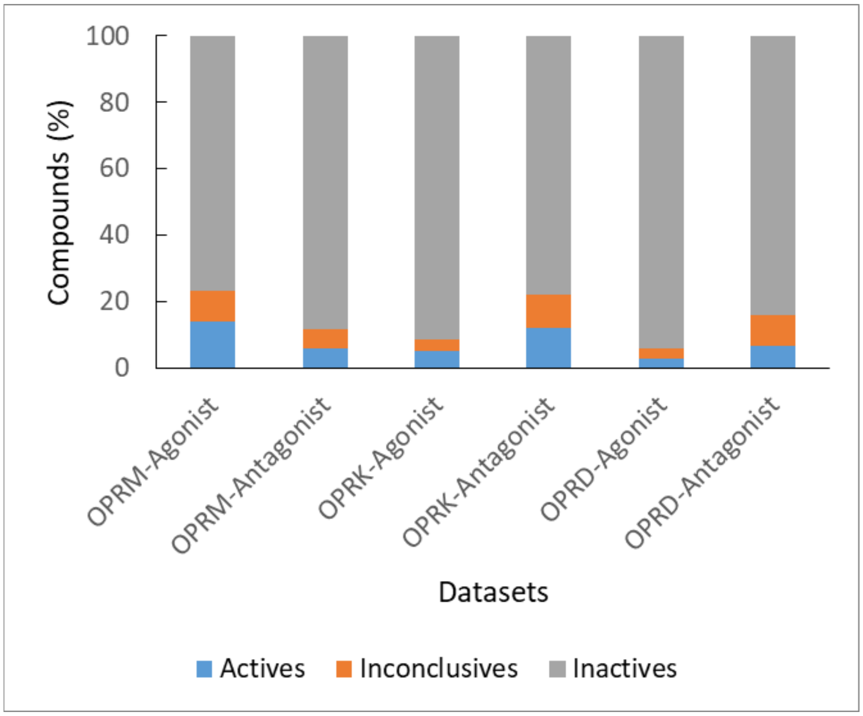


Figure 10. The class distribution of compounds in training set (*in vitro* assay data)

Model training and cross-validation

Four machine learning algorithms: RF, SVM, NN, and XGBoost were applied to develop QSAR models for the 6 OPR targets (agonist or antagonist of OPRM, OPRK, and OPRD) using the *in vitro* assay data. These models were trained using 3 different structural fingerprints for compounds representation: ECFP, MACCS, and PubChem. All 6 datasets were severely unbalanced as the hit (active) rates were low (Fig. 10). An under-sampling technique was applied to balance the active/inactive classes by random selection of the majority class (inactive compounds) to maintain a ratio of 1:2 (active: inactive). A 5-fold cross-validation with 20 iterations was implemented on the training set and iterating ensures that the majority class was sufficiently sampled to cover the whole inactive set of compounds. The model performances from the 5-fold cross-validation with 20 iterations were reported as mean ± standard deviation of a total of 100 AUC-ROC values (Table 6), and the summary of distributions are shown as box plots in Fig. 11. The RF classifier generated high AUC-ROC values, and the highest was 0.87 ± 0.03 for OPRM-agonist with PubChem fingerprints. The next best scores were for OPRK-agonist (0.81 ± 0.07), and OPRK-antagonist (0.82 ± 0.04) with PubChem and ECFP fingerprints respectively. The next best classifier was SVM, which yielded an AUC-ROC value of 0.85 ± 0.04 for OPRM-agonist.

Table 6. AUC-ROC values (mean ± standard deviation) for 20 iterations of a 5-fold cross-validation on training set.

		OPRM-agonist	OPRK-agonist	OPRD-agonist	OPRM-antagonist	OPRK-antagonist	OPRD-antagonist
RF	ECFP	0.84 ± 0.03	0.81 ± 0.07	0.76 ± 0.10	0.78 ± 0.06	0.82 ± 0.04	0.77 ± 0.07
	MACCS	0.84 ± 0.04	0.79 ± 0.06	0.78 ± 0.11	0.67 ± 0.07	0.79 ± 0.05	0.71 ± 0.07
	PubChem	0.87 ± 0.03	0.81 ± 0.07	0.76 ± 0.10	0.69 ± 0.07	0.79 ± 0.04	0.73 ± 0.07
SVM	ECFP	0.83 ± 0.04	0.75 ± 0.09	0.77 ± 0.11	0.77 ± 0.07	0.81 ± 0.04	0.77 ± 0.06
	MACCS	0.79 ± 0.04	0.74 ± 0.08	0.75 ± 0.12	0.59 ± 0.09	0.74 ± 0.05	0.65 ± 0.07
	PubChem	0.85 ± 0.04	0.76 ± 0.07	0.72 ± 0.13	0.66 ± 0.08	0.76 ± 0.04	0.69 ± 0.07
NN	ECFP	0.74 ± 0.04	0.71 ± 0.07	0.64 ± 0.10	0.68 ± 0.06	0.72 ± 0.05	0.68 ± 0.06
	MACCS	0.71 ± 0.05	0.69 ± 0.7	0.67 ± 0.10	0.61 ± 0.07	0.66 ± 0.06	0.62 ± 0.07
	PubChem	0.75 ± 0.04	0.69 ± 0.07	0.69 ± 0.10	0.61 ± 0.08	0.68 ± 0.05	0.62 ± 0.07
XGBoost	ECFP	0.74 ± 0.04	0.70 ± 0.07	0.67 ± 0.09	0.65 ± 0.07	0.70 ± 0.05	0.66 ± 0.06
	MACCS	0.75 ± 0.04	0.69 ± 0.07	0.71 ± 0.09	0.64 ± 0.07	0.69 ± 0.05	0.64 ± 0.07
	PubChem	0.79 ± 0.04	0.72 ± 0.07	0.68 ± 0.09	0.63 ± 0.07	0.71 ± 0.04	0.65 ± 0.07

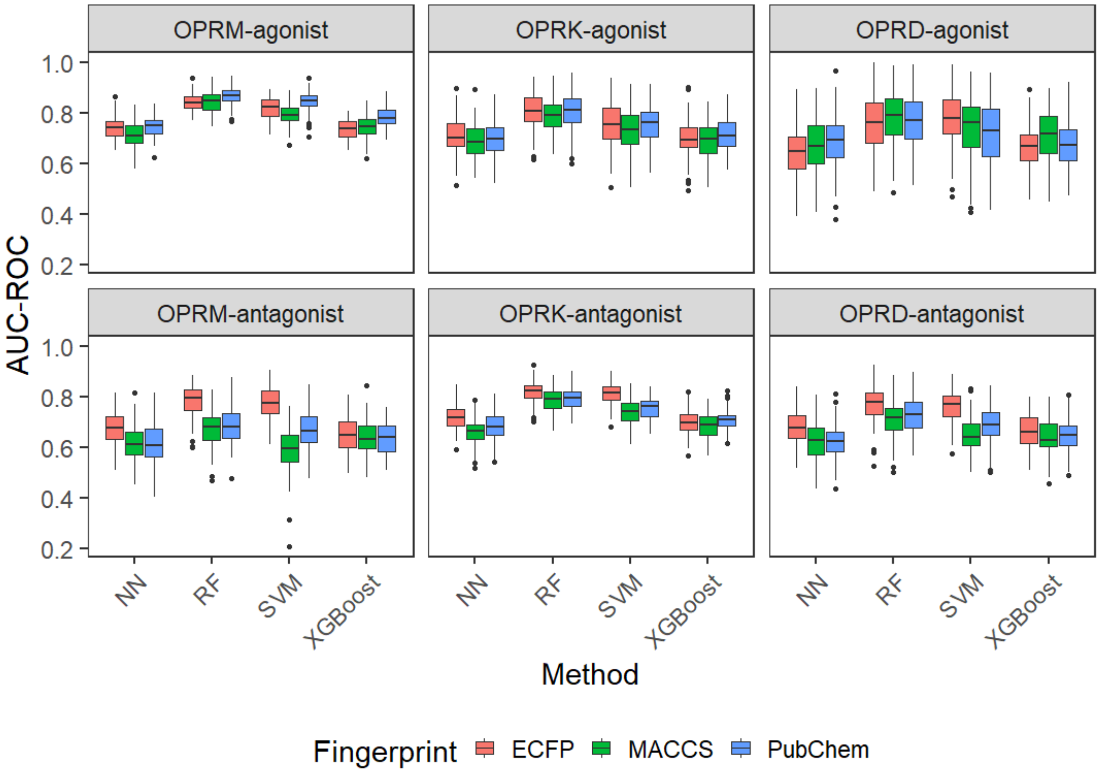


Figure 11. Box plots for the AUC-ROC values showing distributions of 4 machine learning algorithms for different fingerprint types.

External validation and predictions

Dealing with unbalanced classes, an over-sampling strategy was applied by random duplication of the minority class (active compounds) to maintain both classes at 50% each for OPRK-agonist/antagonist or 30% for OPRM-agonist/antagonist, and OPRD-agonist/antagonist. Both ECFP and PubChem fingerprints were used as predictors to train the 6 models using two-thirds of the NPC compounds by the RF algorithm, and the models were applied to make predictions on the remaining one-third of the compounds that served as the external validation set. The predictive performances of all 6 models on the external validation set in terms of balanced accuracy, AUC-ROC and MCC were calculated (Table 7) and the ROC curves were plotted (Fig. 12). The ROC curves for agonist (Fig. 12.A) and antagonist modes (Fig. 12.B) for each fingerprint type showed slight variations at different thresholds of true positive and false positive rates. PubChem fingerprints produced slightly better performance metrics when compared to ECFP in terms of AUC-ROC and MCC values (Table 7).

Table 7. Performance measures of sensitivity, AUC-ROC and MCC from the evaluation on the validation set with 6 models

	PubChem fingerprints			ECFP fingerprints		
	BA	AUC	MCC	BA	AUC	MCC
OPRM-agonist	0.73	0.88	0.7	0.67	0.84	0.59
OPRK-agonist	0.62	0.76	0.33	0.61	0.76	0.31
OPRD-agonist	0.58	0.8	0.17	0.52	0.78	0.07
OPRM-antagonist	0.61	0.76	0.31	0.58	0.72	0.26
OPRK-antagonist	0.62	0.79	0.39	0.61	0.78	0.39
OPRD-antagonist	0.62	0.76	0.32	0.58	0.73	0.27

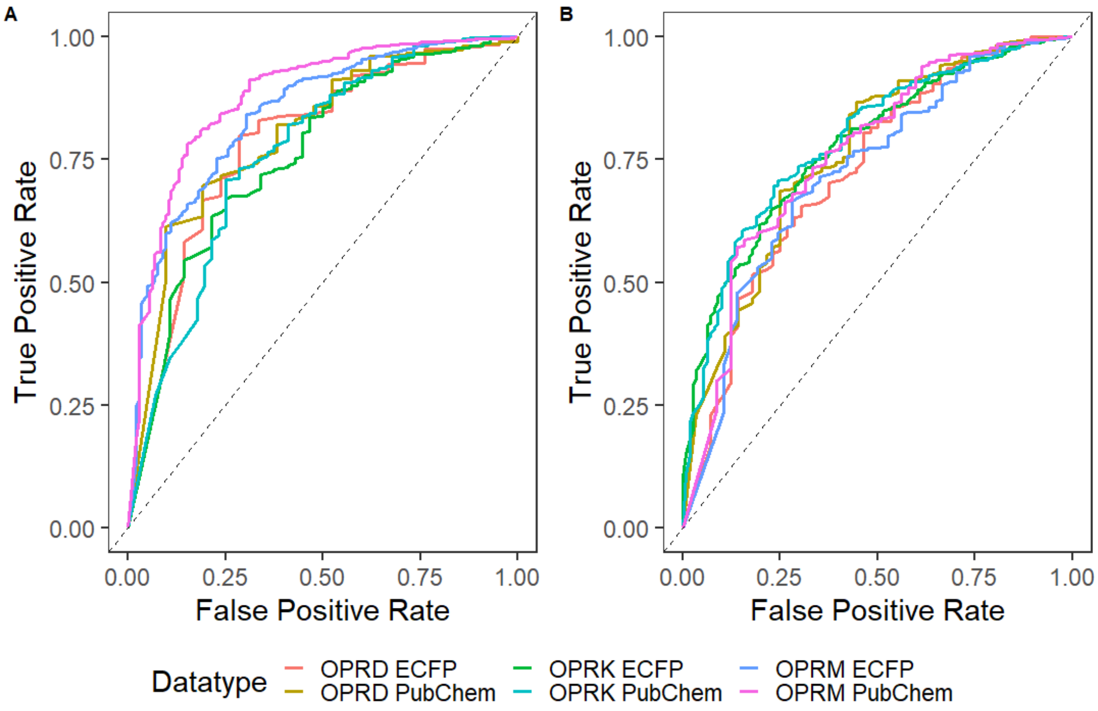


Figure 12. ROC curves for RF classifier testing on the validation set for agonists (A) and antagonists (B) of 3 opioid receptors.

The training and external validation sets were combined, and only PubChem fingerprints were used, to build the final models, which were applied to make predictions on the larger prediction set of 49,018 compounds that have no experimental data.

Compound selection and experimental validation

Compounds with predicted probability of 0.5 or higher were classified as active (Fig. 13). The whole prediction set of compounds were clustered using the k-modes algorithm resulting in 2450 clusters, with cluster sizes ranging from one to 112 compounds. The compounds from each dataset were ranked based on the probability score and the cluster size. For the current study, to reduce false positives from our predictions, we implemented a precise prediction probability cutoff for selection of compounds from each dataset (OPRM-agonist > 0.7, OPRK-antagonist > 0.6, and others > 0.5). From each cluster, the compounds with the highest probability scores were selected if the cluster size was < 10, and the top two were selected if the cluster size was > 10. Based on the in-house availability, 2816 compounds, which fit exactly into two screening plates, were selected for experimental validation.

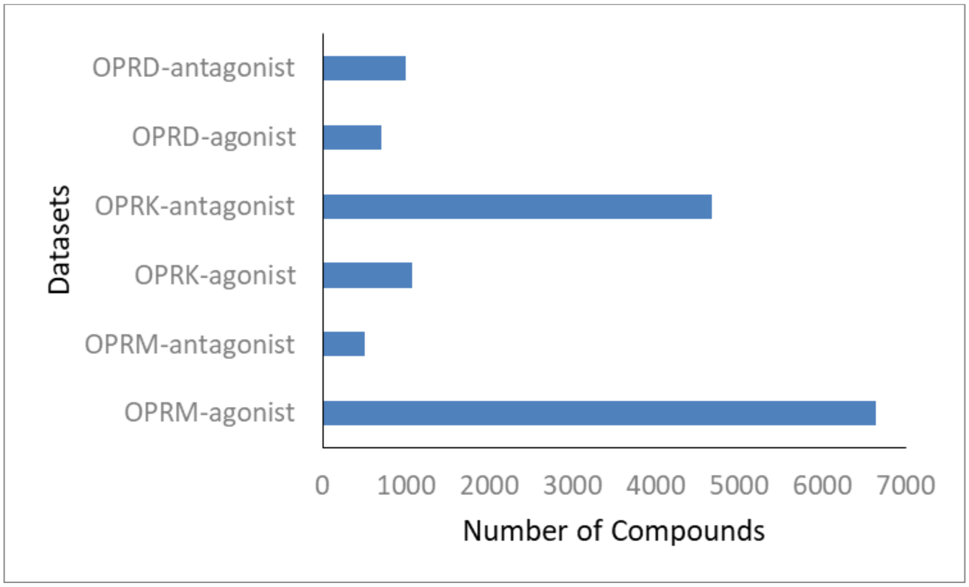


Figure 13. Number of compounds predicted (probabilities > 0.5) to be active in a particular dataset.

These predicted active compounds were tested in the same qHTS assays that generated the corresponding training data. Experimentally validated compounds were counted as true positive (TP) and false positive (FP) otherwise. The confusion matrices describing the performance of the models based on the experimental validation results are given in Table 8.

Table 8. Confusion matrices for the experimental validation of 6 models

A. OPRM-agonist

	Predicted: active	Predicted: inactive
Actual: active	164	369
Actual: inactive	476	1807

B. OPRM-antagonist

	Predicted: active	Predicted: inactive
Actual: active	30	193
Actual: inactive	245	2348

C. OPRK-agonist

	Predicted: active	Predicted: inactive
Actual: active	156	551
Actual: inactive	380	1729

D. OPRK-antagonist

	Predicted: active	Predicted: inactive
Actual: active	100	183
Actual: inactive	584	1949

E. OPRD-agonist

	Predicted: active	Predicted: inactive
Actual: active	40	207
Actual: inactive	403	2166

F. OPRD-antagonist

	Predicted: active	Predicted: inactive
Actual: active	39	99
Actual: inactive	464	2214

The performances were assessed by PPV measures, shown as histograms in Fig. 14A. In our current study, the highest PPVs were obtained for OPRM-agonist (0.31), OPRK-agonist (0.30), and –antagonist (0.24), whereas OPRD-agonist and –antagonist obtained the lowest PPV of 0.18 and 0.15 respectively. To assess the applicability domain (AD) of the models, the Tanimoto similarity score was calculated between each predicted active compound and all active compounds in the training set for every model. The Tanimoto score (Tmax) between the predicted active and the compound most similar to it in the training set was recorded. The predicted actives with Tmax>0.8 (fall within the model AD) was selected to re-evaluate the PPV for each model. The histograms representing the PPV in comparison with the active hit rate from the original training set is shown with and without AD consideration in Figs. 14A and B respectively. From the initial analysis (Fig. 14A), only two models enriched the % active rate by ≥ 2-fold, which are OPRK- and OPRD-agonists, with 5- and 3.5-fold enrichment, respectively. When an AD was defined with a similarity of Tmax>0.8, all 6 models enriched the % active by ≥ 2-fold and the enrichment by the OPRK- and OPRD-agonist models even increased up to 6- and 7-fold, respectively (Fig. 14B).

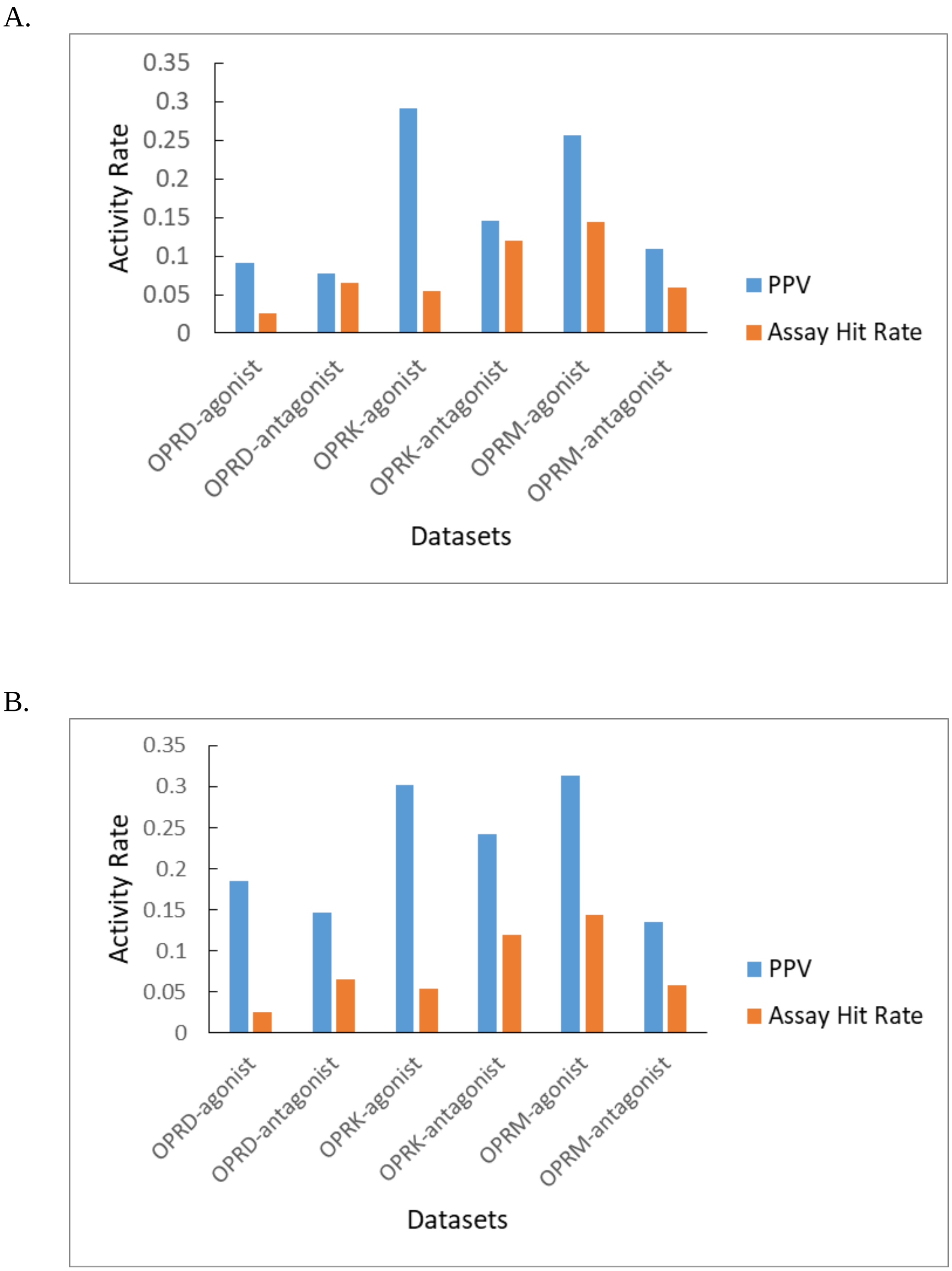


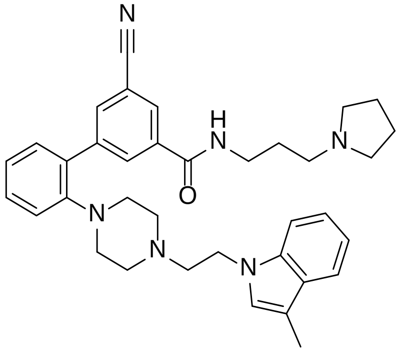
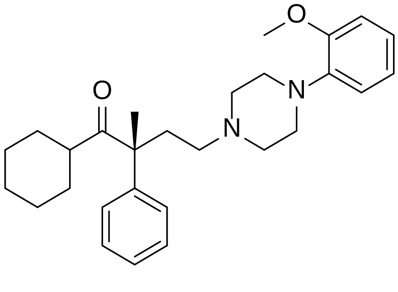
Figure 14. Comparison of initial (A) and final (B: Tanimoto score consideration) data analysis of PPV with % active hit rate from the original training set.

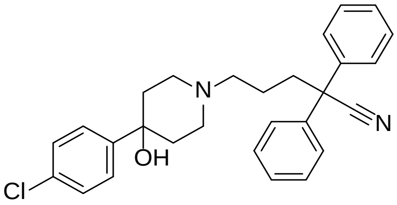
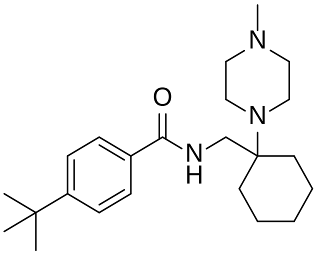
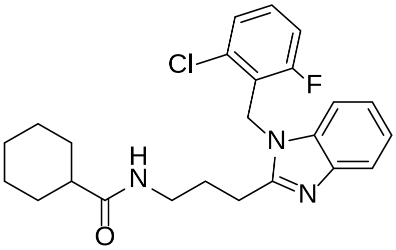
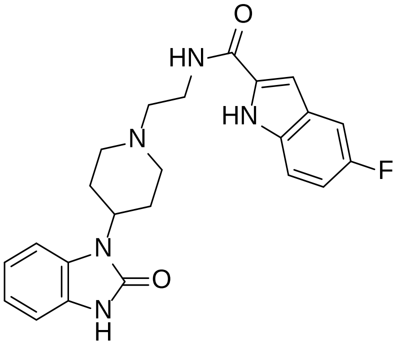
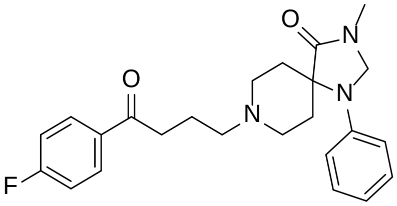
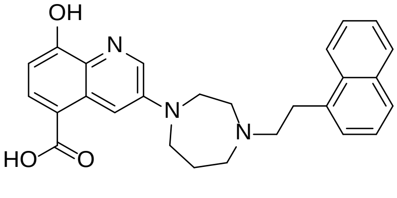
Docking analysis

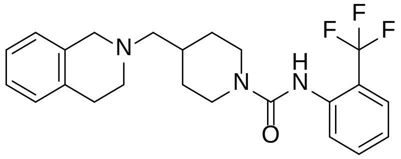
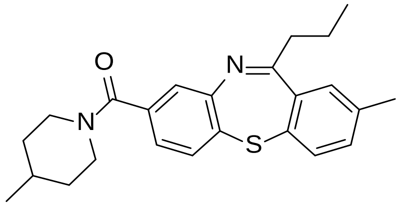
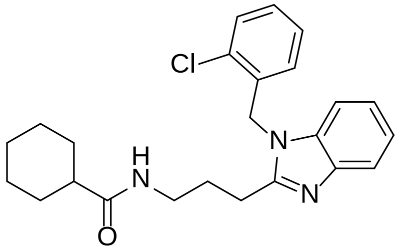
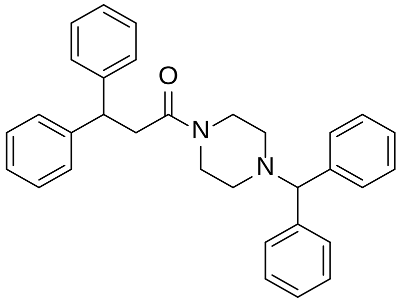
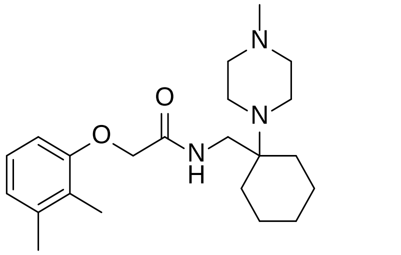
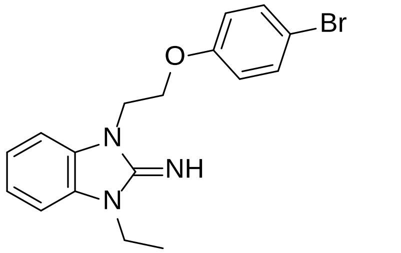
To further evaluate the binding potential of the experimental hits to their respective OPR, all the novel potent active ($AC_{50} \leq 2.5\mu M$) compounds were docked into the binding pocket of the respective active targets. All the ligands yielded binding affinities ranging from -6.7 to -11.1 kcal/mol (Table 9), and these values are comparable to the binding affinities of the known OPR ligands with their corresponding receptors, namely OPRM-BU72 agonist (-8.2 kcal/mol), OPRM-βFNA antagonist (-9.0 kcal/mol), OPRK-MP1104 agonist (-10.3 kcal/mol), OPRK-JDTic antagonist (-10.1 kcal/mol), OPRD-DPI287 agonist (-9.6 kcal/mol), and OPRD-naltrindole antagonist (-10.5 kcal/mol) . From the experimental validation, the OPRM-agonist model produced the highest number of potent compounds and the docking interactions of these OPRM-agonist positive compounds showed the highest affinities (majority of them are ≤ -9.5 kcal/mol). The most potent compound for each target in complex with the protein is shown in Fig. 15A-F, and the active site amino acid residues are indicated as a single letter code followed by their positional number. All the potent compounds are well embedded in the binding pocket, except for ridaforolimus in the OPRK-agonist target, for which the second most potent compound NCGC00135974 is shown in the best docked pose (Fig. 15B). The interaction with the agonist targets are: the amide oxygen atom of LLY-507 forms hydrogen bond with His54 residue of OPRM, NCGC00135974 has three polar interactions with Tyr139, Ser211, and Tyr312 residues of OPRK, and NCGC00139128 forms a hydrogen bond with His301 residue of OPRD. Whereas for antagonist targets, the most potent ligand for all three targets is adenosine 3',5'-

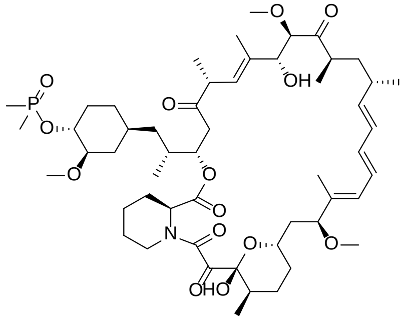
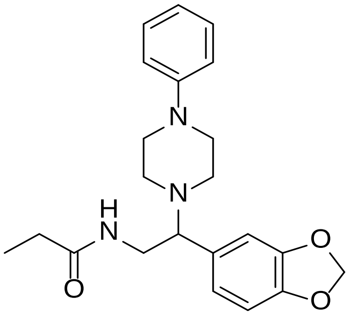
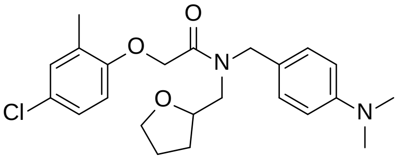
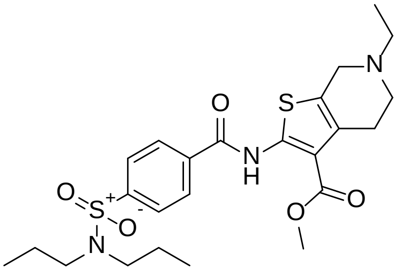
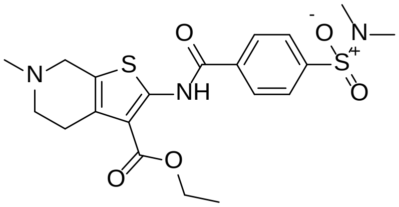
cyclothiophosphate, which showed interactions with Asp147, and a non-polar covalent bond with His297 of OPRM, three polar interactions with Thr111, Gln115, and Tyr320 residues of OPRK, and forms a hydrogen bond with Lys108 and a non-polar interaction with Tyr129 residues of OPRD. The novel ligands were shown to have interactions with the amino acid residues present in the binding pockets of crystal structures of the opioid targets that are published [72-77].

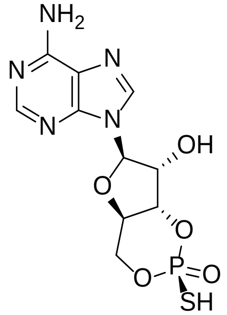
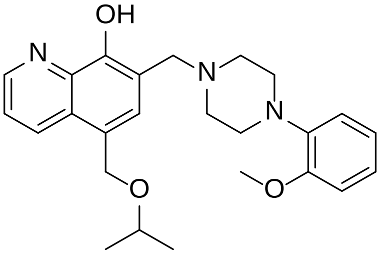
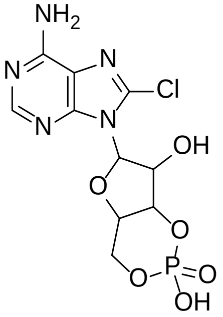
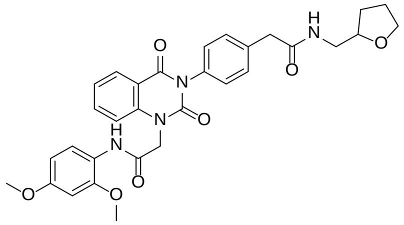
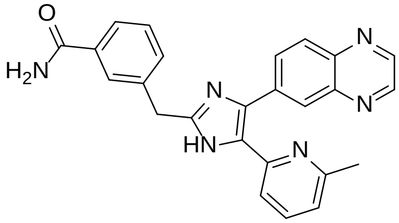
Table 9. List of novel active compounds ($AC_{50} \leq 2.5$ μM)

ID	Name	AC50 (μM)	Binding affinity (kcal/mol)	Target	Structure
NCGC00356417-05	LLY-507	0.13	-9.5	OPRM-agonist	
NCGC00386484-01	LY 426965	0.18	-9.6	OPRM-agonist	

NCGC00378719-01	VUF-2274	0.41	-10.3	OPRM-agonist	
NCGC00118549-01		0.46	-7.2	OPRM-agonist	
NCGC00114741-01		0.52	-9.6	OPRM-agonist	
NCGC00370807-05	FIPI	0.58	-11.1	OPRM-agonist	
NCGC00485045-01	N-Methylspiperone	0.65	-9.5	OPRM-agonist	
NCGC00247751-02		0.65	-10.4	OPRM-agonist	

NCGC00123899-01		0.92	-10.2	OPRM-agonist	
NCGC00107633-01		0.92	-7.5	OPRM-agonist	
NCGC00114743-01		1.64	-10.1	OPRM-agonist	
NCGC00378879-01	NP-118809	1.83	-10.9	OPRM-agonist	
NCGC00118605-01		1.83	-8.3	OPRM-agonist	
NCGC00141762-01		2.06	-8.1	OPRM-agonist	

NCGC00346 481-02	Ridaforolimus	0.13	-8.1	OPRK-agonist	
NCGC00135 974-01		1.16	-9.7	OPRK-agonist	
NCGC00117 293-01		2.50	-9.1	OPRK-agonist	
NCGC00139 128-01		2.06	-7.7	OPRD-agonist	
NCGC00136 158-01		2.31	-7.9	OPRD-agonist	

NCGC00485585-01	Adenosine 3',5'-cyclothiophosphate	0.06 0.07 0.04	-6.7 -7.2 -6.8	OPRM-antagonist OPRK-antagonist OPRD-antagonist	
NCGC00114968-02		0.65	-8.5	OPRM-antagonist	
NCGC00485288-01	Tocladesine	1.03 0.92 0.46	-6.9 -7.6 -7.3	OPRM-antagonist OPRK-antagonist OPRD-antagonist	
NCGC00106344-01		2.06	-10.2	OPRK-antagonist	
NCGC00386726-01	IN-1130	2.06	-11.0	OPRK-antagonist	

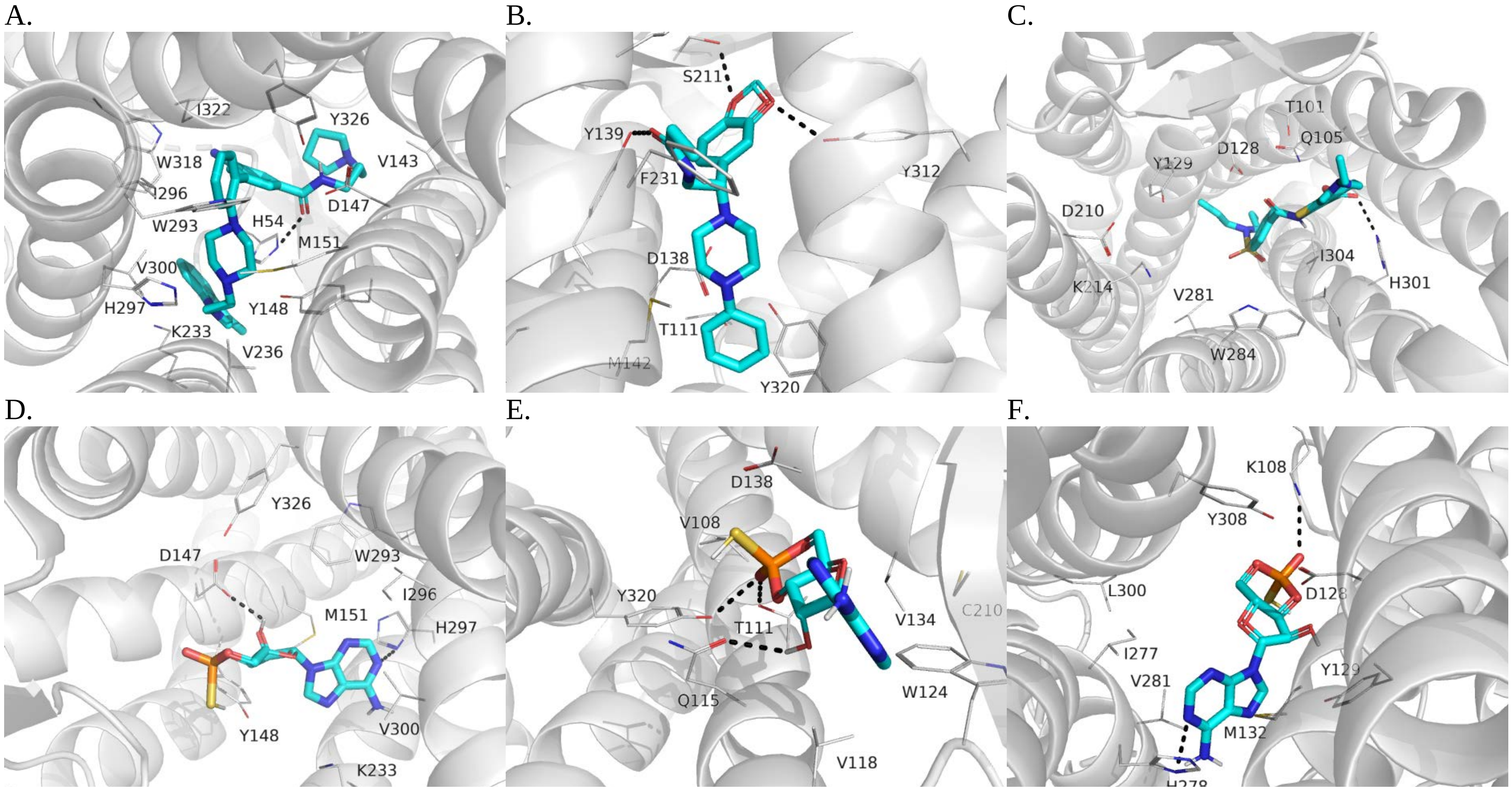


Figure 15. Docked poses of the most potent active compounds. The target receptors are shown as grey helices, the active site amino acid residues are represented as lines in grey, the potent compounds shown as sticks with carbons colored in cyan, and the interactions are shown as black dashed lines. Docked poses in top row are for opioid receptor agonists (A-OPRM; B-OPRK; C-OPRD) and the bottom row is for antagonists (D-OPRM; E-OPRK; F-OPRD).

Discussion

The aim of this study was to develop QSAR models to predict a compound's agonistic and/or antagonistic effect on OPRM, OPRK, and OPRD targets. A total of 6 QSAR models were developed based on *in vitro* assay data. The active compounds constituted only a small percentage in all six qHTS datasets (Fig. 10). The HTS hit rates from a diverse compound library are typically ≤ 1%, unless there are some exceptions where the compounds selected are for multi-targets, for which the hit rates can go beyond or up to 10% [8]. Two approaches were used in our study to balance the classes: random under-sampling and over-sampling. For evaluating the model performances on a cross-validation of training set, an under-sampling strategy was applied due to its computational inexpensiveness, and for the rest of the analysis over-sampling was applied. The two strategies showed comparable performances in terms of AUC-ROC (Tables 6 and 7). The classifiers that were generated without employing any class-balancing strategy were more biased toward the majority class (inactive compounds), for example, a true-positive rate of 0.01 was observed for the OPRM-agonist model though it had the highest percentage of the minority class (active compounds). Based on the performance metrics given in Tables 6 and 7, the RF algorithm was adopted as the method of classification, and PubChem fingerprints were chosen to develop the final predictive models.

The experimental validation of the predictions generated from the RF classifier, resulted in a large number of false positives (Table 8). Even though the QSAR models developed in our study were trained on a structurally diverse set of compounds, the compounds in the NPC are all drugs whereas most of the compounds in the prediction set are novel synthetic molecules that may fall out of the model's AD [78]. Predictions made outside of a model's AD are often unreliable [79]. We found this to be true in this study as well. When the compounds that fell outside of the model AD were excluded based on a structure similarity cutoff, the models showed significant improvement in performance on the experimental validation set (Fig. 14B) [80].

Herein, the OPRM-agonist and OPRK-antagonist models predicted more active compounds than the other models, and the OPRM-agonist model predictions yielded the largest number of novel potent active compounds that were experimentally validated. Most opioids in use for pain treatment are OPRM agonists, and with some activity exerting on OPRK as well [81]. The lack of mu OPRs in mice has demonstrated that they are the sole receptors in mediating morphine's analgesic and addictive properties [82]. From our study, 19 novel potent compounds were identified to have effect on the OPRs with the majority active against OPRM (Table 9), these compounds could be developed into new therapies to combat the opioid crisis. Only a few of these compounds have previously reported targets, for example, LLY-507 is a potent and selective inhibitor of protein-lysine methyltransferase SMYD2 [83]; LY 426965 is an aryl piperazine compound that acts as a serotonin$_{1A}$ (5-hydroxytryptamine$_{1A}$) antagonist [84]; VUF 2274 is a human cytomegalovirus encoded US28 (a GPCR) inhibitor [85]; FIPI (a halopemide derivative) is a potent phospholipase D inhibitor [86]; NP-118809 (39-1B4) is a potent N-type calcium channel blocker [87]; and ridaforolimus (MK 8669) is a selective inhibitor of the mammalian target of rapamycin (mTOR) and has an anti-tumor activity [88].

The OPR antagonist compounds compete with the agonists and block the receptor thus reverse the agonistic effects, so they are used in the clinic for partial/complete reversal of opioid toxicity, and to relieve opioid-related adverse effects etc. [89]. The most commonly used antagonists for reversing the opioid toxicity are naloxone, naltrexone (both compounds inhibit all types of OPRs), and naltrindole (OPRD specific) [47] . For our initial assay optimization, naloxone was used as a positive control compound, and the IC_{50} was 1.2nM and 0.22μM for OPRM and OPRK, respectively. Our models also identified novel compounds that exhibited similar potent inhibitory effect on OPRs. Two compounds that are analogs of cyclic adenosine monophosphate (cAMP) showed antagonistic effects on mu, kappa, and delta OPRs, and they are adenosine 3',5'-cyclothiophosphate and 8-chloro cAMP (tocladesine, an anticancer drug). The first compound was the most potent against all three receptors (IC_{50} in range of 40-70nM) (Table 9). Another novel compound, NCGC00114968, that contains the 8-hydroxyquinoline (8HQ) moiety, showed potent inhibition against OPRM (IC_{50} = 0.65μM). 8HQ derivatives have been used as fungicides and a few of its derivatives with the piperazine ring (like in NCGC00114968) were reported to exert antineurodegenerative effect [90]. Two other novel compounds, which were shown to have a specific inhibitory effect on OPRK, both with IC_{50} of 2.06μM, are NCGC00106344, a quinazolineacetamide derivative with no known target reported yet and NCGC00386726 (IN-1130), a well-studied drug for its potency in inhibiting the TGF-beta signaling pathway [91].

We performed docking to get insights on the interactions of the novel potent ($AC_{50} \leq 2.5\mu M$) compounds that were validated experimentally, with the active sites of the respective crystal structures of OPRs. The experimentally validated, most potent compound for each OPR was shown to have interactions with residues in the active site (Fig. 15), except for ridaforolimus ($EC_{50} = 0.13\mu M$), which could not fit into the binding pocket of OPRK in the docking study. Ridaforolimus is a rapamycin analog, which are macrolides known to form complexes with the intracellular receptor FK506-binding protein (FKBP12), and interfere with the mTOR activity [92, 93]. The linkage of OPRK with mTOR system has not been well understood yet, but the OPRK-mediated mTOR signaling was shown in the mouse brain as the activation of the mTOR pathway occurs in neurons expressing the OPRK [94]. The research for safer drugs to alleviate pain without exerting severe adverse effects was mostly *in silico* driven, and these approaches have provided important information regarding the structural determinants that are responsible for binding affinity and selectivity of newly identified ligands [95]. Also through pharmacophore-based modeling, novel antagonists for the mu OPR were identified and evaluated in *in vitro* [96] and *in vivo* [97] for its significant inhibition of morphine-induced antinociception. Docking approaches to predict the binding affinities of fentanyl derivatives to the mu-OPR have been developed recently [98, 99]. Thousands of fentanyl analogues were identified, and a strong correlation was found between the docking scores and experimental binding affinities. These approaches are exploited when *in vitro* data are not available and may facilitate temporary scheduling of those substances that pose risks to the public. Other studies included the design and synthesis of analogues of known

OPR agonists and antagonists, which were evaluated in *in vitro* pharmacological assays. Such efforts involved modifying the 6th position of the morphinan that plays a key role in the mu OPR activity [100], and replacing the hydroxyl groups with other groups in JDTic to see their effect on mu, kappa, and delta OPRs [101]. These target structure-based virtual screening approaches often have limited capacity in identifying novel chemical scaffolds, whereas models developed based on assay data may discover compounds with more diverse structures. Our study presented the first predictive models built on *in vitro* assay data, which were generated from a large, diverse set of known drugs against OPRs. These models could be applied to virtually screen large compound libraries to identify novel OPR active compounds.

In summary, we developed models based on qHTS data for the prediction of compound activity on three different OPRs. The models identified a number of novel compounds, which were validated experimentally. The potent active compounds were shown to have interactions within the receptor's binding pocket via molecular docking. All models were able to enrich active hit rate by $\geq$ 2-fold. These models have the potential to be used for larger collections to predict the compound's effect on OPRs.

CONCLUSION AND FUTURE WORK

We built different computational models for predicting the chemical activity on various disease targets and these targets are usually the proteins that are intrinsically associated with a particular disease. The chemicals involved in our study are the small drug-like molecules (molecular weight < 500 for majority of the chemicals) and they are synthesized with an aim that they represent all theoretically possible combinations of different scaffolds and their collections in large numbers are called as chemical or compound libraries that are ultimately used in high-throughput screening. For our study we included the targets like CYP3A4, ESR1, ADRA1A, OPRM, OPRD, and OPRK.

For CYP3A4, ESR1, and ARA1A targets, the compounds were represented by their bioactivities (data obtained from various *in vitro* assays tested against numerous targets/endpoints) along with their structural features. The aim of using the bioactivity data of the compounds is to show that the small molecule drug-like compounds have a potential to activate new targets, and/or therapeutic indications across a broad array of human diseases and thus facilitates the discovery of novel therapeutic uses of approved drugs for repurposing. To the best of our knowledge, using bioactivity data for representing the compounds is the first of its kind in building computational predictive models.

For OPRM, OPRK, and OPRD, a qHTS data was used to train the models for predicting compound’s effect on different ORs and these are QSAR models for finding the relationship between structure and activity of the compounds. For this study the models built using RF algorithm were observed to have high performance rates and PubChem fingerprints were shown to have better predictability of the classes. The generated models were used for prediction of novel compounds and these predictions were validated experimentally. A molecular docking of the experimentally validated true positive compounds have shown to have interactions with in the receptor's binding pocket. These models have the potential to be used for larger collections to predict the compound's effect on ORs, taken into consideration if the external compounds share some extent of structural similarity to the training set.

Drug discovery is a complex process which can take 12-15 years and costs more than $1 billion [102]. The identification and validation of each target in order to get prepared for HTS itself may take several years. The total time for target identification and lead optimization may take several years (3-5 years) [14]. Our current approach of developing computational predictive models for chemical activity against new targets can be leveraged to identify the leads at much faster speed and inexpensive when compared to running the HTS against all the existed targets. Whereas our study helped in rapid identification of the potential novel compounds that are capable of modulating the ORs. The comparison of the timelines in the traditional drug discovery process and our current approach in identifying the novel potent compounds right from the start of the target of interest and database (compound library) selection is shown in Fig. 16.

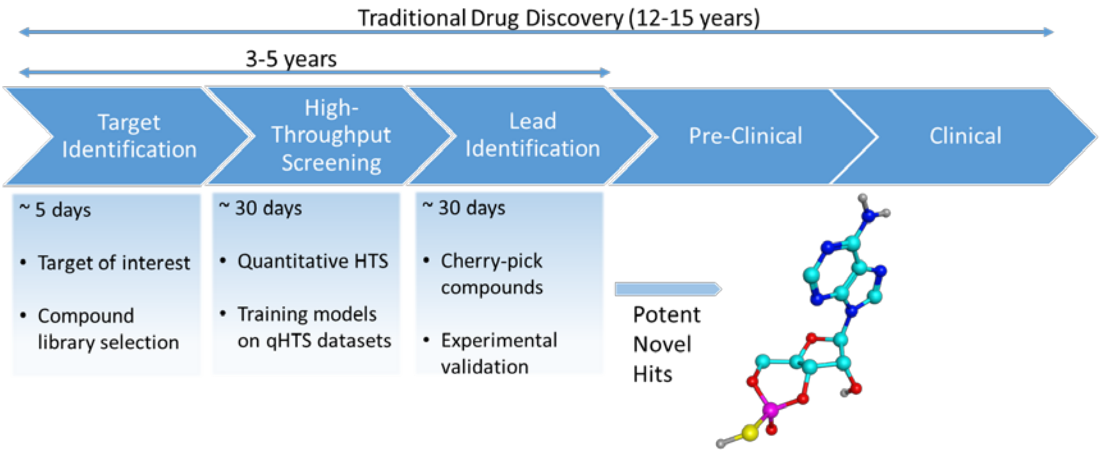


Figure 16. Traditional drug discovery process and with comparison of our current study

Our current study which is mainly focused on a target-based approach aids in drug repurposing. But the drug discovery has been changing its focus for the last several years, by not only including target-based ways, but by giving ways to a system-based through omics studies [103]. So the individual targets are replaced by molecular pathways with cell-based or phenotypic responses. Such advancements in phenotypic screening offers several advantages over the original target-based approach such as providing a biological response that is physiologically more relevant. Using qHTS data the computational models for androgen receptor (AR) pathway have been developed and validated to predict the chemical activity against AR pathway [44]. Also binding of these chemicals to more than one target has been studied too, which refers to polypharmacology [104]. Unintended off-target binding results in adverse effects, whereas desirable off-target binding can lead to drug repurposing approach.

It is our expectation that the current studies outlined in this dissertation will help to predict the effect of environmental chemicals/drugs on various new targets and/or cellular and mechanistic pathways and aid in drug repurposing. Based on these predictions the chemicals can be further prioritized for human health and treatment as *in vitro* (HTS) assays are time-consuming, expensive and limited.

REFERENCES

1. Broach, J.R. and J. Thorner, *High-throughput screening for drug discovery.* Nature, 1996. **384**(6604): p. 14-16.

2. Entzeroth, M., H. Flotow, and P. Condron, *Overview of high-throughput screening.* Current protocols in pharmacology, 2009. **44**(1): p. 9.4. 1-9.4. 27.

3. Inglese, J., et al., *Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries.* Proceedings of the National Academy of Sciences, 2006. **103**(31): p. 11473-11478.

4. Huang, R., *A Quantitative High-Throughput Screening Data Analysis Pipeline for Activity Profiling.* Methods Mol Biol, 2016. **1473**: p. 111-22.

5. Attene-Ramos, M.S., et al., *The Tox21 robotic platform for the assessment of environmental chemicals--from vision to reality.* Drug Discov Today, 2013. **18**(15-16): p. 716-23.

6. Tice, R.R., et al., *Improving the human hazard characterization of chemicals: a Tox21 update.* Environ Health Perspect, 2013. **121**(7): p. 756-65.

7. Shukla, S.J., et al., *The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform.* Drug Discov Today, 2010. **15**(23-24): p. 997-1007.

8. Huang, R., et al., *Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization.* Nat Commun, 2016. **7**: p. 10425.

9. Huang, R., et al., *Chemical genomics profiling of environmental chemical modulation of human nuclear receptors.* Environ Health Perspect, 2011. **119**(8): p. 1142-8.

10. Jin, G. and S.T. Wong, *Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines.* Drug discovery today, 2014. **19**(5): p. 637-644.

11. Hao, M., S.H. Bryant, and Y. Wang, *Cheminformatics analysis of the AR agonist and antagonist datasets in PubChem.* Journal of cheminformatics, 2016. **8**(1): p. 37.

12. Norinder, U. and S. Boyer, *Conformal prediction classification of a large data set of environmental chemicals from ToxCast and Tox21 estrogen receptor assays.* Chemical research in toxicology, 2016. **29**(6): p. 1003-1010.

13. Huang, R., et al., *The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics.* Science translational medicine, 2011. **3**(80): p. 80ps16-80ps16.

14. Ashburn, T.T. and K.B. Thor, *Drug repositioning: identifying and developing new uses for existing drugs.* Nature reviews Drug discovery, 2004. **3**(8): p. 673.

15. Wishart, D.S., et al., *DrugBank: a knowledgebase for drugs, drug actions and drug targets.* Nucleic Acids Res, 2008. **36**(Database issue): p. D901-6.

16. Zanger, U.M. and M. Schwab, *Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation.* Pharmacology & therapeutics, 2013. **138**(1): p. 103-141.

17. Ince, I., et al., *Developmental changes in the expression and function of cytochrome P450 3A isoforms: evidence from in vitro and in vivo investigations.* Clinical pharmacokinetics, 2013. **52**(5): p. 333-345.

18. Sommer, S. and S.A. Fuqua, *Estrogen receptor and breast cancer.* Semin Cancer Biol, 2001. **11**(5): p. 339-52.

19. Lefkowitz, R. and M. Caron, *Adrenergic receptors. Models for the study of receptors coupled to guanine nucleotide regulatory proteins.* Journal of Biological Chemistry, 1988. **263**(11): p. 4993-4996.

20. Chang, D.J., et al., *Molecular cloning, genomic characterization and expression of novel human α1A-adrenoceptor isoforms.* FEBS letters, 1998. **422**(2): p. 279-283.

21. Minneman, K.P. and T.A. Esbenshade, *Alpha 1-adrenergic receptor subtypes.* Annu Rev Pharmacol Toxicol, 1994. **34**: p. 117-33.

22. Mauri, A., et al., *Dragon software: An easy approach to molecular descriptor calculations.* Match, 2006. **56**(2): p. 237-248.

23. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* Journal of chemical information and computer sciences, 2003. **43**(6): p. 1947-1958.

24. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.

25. Abdo, A., et al., *Ligand-based virtual screening using bayesian networks.* Journal of chemical information and modeling, 2010. **50**(6): p. 1012-1020.

26. Sheridan, R.P., et al., *Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships.* J Chem Inf Model, 2016. **56**(12): p. 2353-2360.

27. Miao, J. and L. Niu, *A survey on feature selection.* Procedia Computer Science, 2016. **91**: p. 919-926.

28. Kursa, M.B., A. Jankowski, and W.R. Rudnicki, *Boruta–a system for feature selection.* Fundamenta Informaticae, 2010. **101**(4): p. 271-285.

29. Sokolova, M. and G. Lapalme, *A systematic analysis of performance measures for classification tasks.* Information Processing & Management, 2009. **45**(4): p. 427-437.

30. Zweig, M.H. and G. Campbell, *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.* Clinical chemistry, 1993. **39**(4): p. 561-577.

31. Hsu, C.-W., et al., *Advances in high-throughput screening technology for toxicology.* International Journal of Risk Assessment and Management, 2017. **20**(1-3): p. 109-135.

32. Lavecchia, A., *Machine-learning approaches in drug discovery: methods and applications.* Drug discovery today, 2015. **20**(3): p. 318-331.

33. Lo, Y.-C., et al., *Machine learning in chemoinformatics and drug discovery.* Drug discovery today, 2018.

34. Huang, R., et al., *Expanding biological space coverage enhances the prediction of drug adverse effects in human using in vitro activity profiles.* Sci Rep, 2018. **8**(1): p. 3783.

35. Sun, H., et al., *Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data.* Journal of chemical information and modeling, 2011. **51**(10): p. 2474-2481.

36. Sun, H., et al., *Prediction of cytochrome P450 profiles of environmental chemicals with QSAR models built from drug-like molecules.* Molecular informatics, 2012. **31**(11-12): p. 783-792.

37. He, S.B., et al., *Construction of Metabolism Prediction Models for CYP450 3A4, 2D6, and 2C9 Based on Microsomal Metabolic Reaction System.* Int J Mol Sci, 2016. **17**(10).

38. Hukkanen, J., *Induction of cytochrome P450 enzymes: a view on human in vivo findings.* Expert review of clinical pharmacology, 2012. **5**(5): p. 569-585.

39. Chen, Y., et al., *Computational models to predict endocrine-disrupting chemical binding with androgen or oestrogen receptors.* Ecotoxicology and environmental safety, 2014. **110**: p. 280-287.

40. Heo, S., U. Safder, and C. Yoo, *Deep learning driven QSAR model for environmental toxicology: effects of endocrine disrupting chemicals on human health.* Environmental Pollution, 2019. **253**: p. 29-38.

41. Yang, Y., et al., *MP44-05 EARLY-ONSET SYMPTOMATIC BPH TISSUES OF MEN LESS THAN OR EQUAL TO 50 YEARS OLD APPEARED TO BE INCREASED STROMAL COMPONENTS AND VASCULARITY.* The Journal of Urology, 2016. **195**(4S): p. e600-e600.

42. Kuo, G.-H., et al., *Design, Synthesis, and Structure− Activity Relationships of Phthalimide-Phenylpiperazines: A Novel Series of Potent and Selective α1a-Adrenergic Receptor Antagonists.* Journal of medicinal chemistry, 2000. **43**(11): p. 2183-2195.

43. Ghanemi, A., *Targeting G protein coupled receptor-related pathways as emerging molecular therapies.* Saudi Pharmaceutical Journal, 2015. **23**(2): p. 115-129.

44. Kleinstreuer, N.C., et al., *Development and Validation of a Computational Model for Androgen Receptor Activity.* Chem Res Toxicol, 2017. **30**(4): p. 946-964.

45. Dhawan, B.N., et al., *International Union of Pharmacology. XII. Classification of opioid receptors.* Pharmacol Rev, 1996. **48**(4): p. 567-92.

46. McDonald, J. and D.G. Lambert, *Opioid receptors.* Continuing Education in Anaesthesia Critical Care & Pain, 2005. **5**(1): p. 22-25.

47. Vallejo, R., R.L. Barkin, and V.C. Wang, *Pharmacology of opioids in the treatment of chronic pain syndromes.* Pain Physician, 2011. **14**(4): p. E343-60.

48. Williams, J., *Basic Opioid Pharmacology.* Rev Pain, 2008. **1**(2): p. 2-5.

49. Waldhoer, M., S.E. Bartlett, and J.L. Whistler, *Opioid receptors.* Annu Rev Biochem, 2004. **73**: p. 953-90.

50. Hedegaard, H., A.M. Minino, and M. Warner, *Drug Overdose Deaths in the United States, 1999-2017.* NCHS Data Brief, 2018(329): p. 1-8.

51. Bruchas, M.R. and B.L. Roth, *New Technologies for Elucidating Opioid Receptor Function.* Trends Pharmacol Sci, 2016. **37**(4): p. 279-289.

52. Huang, R., et al., *Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs.* Frontiers in Environmental Science, 2016. **3**: p. 85.

53. Russo, D.P., et al., *Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction.* Mol Pharm, 2018. **15**(10): p. 4361-4370.

54. Du, H., et al., *In Silico Prediction of Chemicals Binding to Aromatase with Machine Learning Methods.* Chem Res Toxicol, 2017. **30**(5): p. 1209-1218.

55. Drwal, M.N., et al., *Molecular similarity-based predictions of the Tox21 screening outcome.* Frontiers in Environmental Science, 2015. **3**.

56. Manglik, A., et al., *Structure-based discovery of opioid analgesics with reduced side effects.* Nature, 2016. **537**(7619): p. 185-190.

57. Negri, A., et al., *Discovery of a novel selective kappa-opioid receptor agonist using crystal structure-based virtual screening.* J Chem Inf Model, 2013. **53**(3): p. 521-6.

58. Huang, R., et al., *The NCATS Pharmaceutical Collection: a 10-year update*. Drug Discov Today, 2019. **24**(12): p. 2341-2349.

59. Beisken, S., et al., *KNIME-CDK: Workflow-driven cheminformatics.* BMC Bioinformatics, 2013. **14**: p. 257.

60. Venables, W.N.a.R., B. D., *Neural Networks*, in *Modern Applied Statistics with S*. 2002, Springer. p. 243-250.

61. Kuhn, M., *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 2008. **28**(5).

62. Lunardon, N.M., G.; Torelli, N., *ROSE: A Package for Binary Imbalanced Learning.* The R Journal, 2014. **6**(1).

63. Fernanda, B.M.S., L. A.; Stead, L. G. , *Understanding statistical tests in the medical literature: which test should I use?* Int J Emerg Med, 2008. **1**: p. 197-199.

64. McHugh, M.L., *The Chi-square test of independence.* Biochemia Medica, 2013. **23**(2): p. 143-9.

65. Sokolova, M., N. Japkowicz, and S. Szpakowicz. *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation*. in *Australasian joint conference on artificial intelligence*. 2006. Springer.

66. Tharwat, A., *Classification assessment methods.* Applied Computing and Informatics, 2018.

67. Huang, Z., *A fast clustering algorithm to cluster very large categorical data sets in data mining.* DMKD, 1997. **3**(8): p. 34-39.

68. Willett, P., J.M. Barnard, and G.M. Downs, *Chemical similarity searching.* Journal of chemical information and computer sciences, 1998. **38**(6): p. 983-996.

69. Todeschini, R., et al., *Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets.* Journal of chemical information and modeling, 2012. **52**(11): p. 2884-2901.

70. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.

71. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* J Comput Chem, 2009. **30**(16): p. 2785-91.

72. Huang, W., et al., *Structural insights into micro-opioid receptor activation.* Nature, 2015. **524**(7565): p. 315-21.

73. Manglik, A., et al., *Crystal structure of the micro-opioid receptor bound to a morphinan antagonist.* Nature, 2012. **485**(7398): p. 321-6.

74. Che, T., et al., *Structure of the Nanobody-Stabilized Active State of the Kappa Opioid Receptor.* Cell, 2018. **172**(1-2): p. 55-67 e15.

75. Wu, H., et al., *Structure of the human kappa-opioid receptor in complex with JDTic.* Nature, 2012. **485**(7398): p. 327-32.

76. Claff, T., et al., *Elucidating the active delta-opioid receptor crystal structure with peptide and small-molecule agonists.* Sci Adv, 2019. **5**(11): p. eaax9115.

77. Granier, S., et al., *Structure of the delta-opioid receptor bound to naltrindole.* Nature, 2012. **485**(7398): p. 400-4.

78. Dimitrov, S., et al., *A stepwise approach for defining the applicability domain of SAR and QSAR models.* J Chem Inf Model, 2005. **45**(4): p. 839-49.

79. Sheridan, R.P.F., B. P.; Maiorov V. N.; Kearsley S. K., *Similarity to molecules in the training set Is a good discriminator for prediction accuracy in QSAR.* J Chem Inf Comput Sci, 2004. **44**: p. 1912-1928.

80. Bajusz, D.R., A.; Heberger, K., *Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?* Journal of Cheminformatics, 2015. **7**(20).

81. Trescot, A.M., et al., *Opioid pharmacology.* Pain Physician, 2008. **11**(2 Suppl): p. S133-53.

82. Matthes, H.W., et al., *Loss of morphine-induced analgesia, reward effect and withdrawal symptoms in mice lacking the mu-opioid-receptor gene.* Nature, 1996. **383**(6603): p. 819-23.

83. Nguyen, H., et al., *LLY-507, a Cell-active, Potent, and Selective Inhibitor of Protein-lysine Methyltransferase SMYD2.* J Biol Chem, 2015. **290**(22): p. 13641-53.

84. Rasmussen, K., et al., *The novel 5-Hydroxytryptamine(1A) antagonist LY426965: effects on nicotine withdrawal and interactions with fluoxetine.* J Pharmacol Exp Ther, 2000. **294**(2): p. 688-700.

85. Casarosa, P., et al., *Identification of the first nonpeptidergic inverse agonist for a constitutively active viral-encoded G protein-coupled receptor.* J Biol Chem, 2003. **278**(7): p. 5172-8.

86. Su, W., et al., *5-Fluoro-2-indolyl des-chlorohalopemide (FIPI), a phospholipase D pharmacological inhibitor that alters cell spreading and inhibits chemotaxis.* Mol Pharmacol, 2009. **75**(3): p. 437-46.

87. Zamponi, G.W., et al., *Scaffold-based design and synthesis of potent N-type calcium channel blockers.* Bioorg Med Chem Lett, 2009. **19**(22): p. 6467-72.

88. Rivera, V.M., et al., *Ridaforolimus (AP23573; MK-8669), a potent mTOR inhibitor, has broad antitumor activity and can be optimally administered using intermittent dosing regimens.* Mol Cancer Ther, 2011. **10**(6): p. 1059-71.

89. Choi, Y.S. and J.A. Billings, *Opioid antagonists: a review of their role in palliative care, focusing on use in opioid-related constipation.* J Pain Symptom Manage, 2002. **24**(1): p. 71-90.

90. Prachayasittikul, V., et al., *8-Hydroxyquinolines: a review of their metal chelating properties and medicinal applications.* Drug Des Devel Ther, 2013. **7**: p. 1157-78.

91. Moon, J.A., et al., *IN-1130, a novel transforming growth factor-beta type I receptor kinase (ALK5) inhibitor, suppresses renal fibrosis in obstructive nephropathy.* Kidney Int, 2006. **70**(7): p. 1234-43.

92. Kist, R. and R.A. Caceres, *New potential inhibitors of mTOR: a computational investigation integrating molecular docking, virtual screening and molecular dynamics simulation.* J Biomol Struct Dyn, 2017. **35**(16): p. 3555-3568.

93. Dancey, J., *mTOR signaling and drug development in cancer.* Nat Rev Clin Oncol, 2010. **7**(4): p. 209-19.

94. Liu, J.J., et al., *Phosphoproteomic approach for agonist-specific signaling in mouse brains: mTOR pathway is involved in kappa opioid aversion.* Neuropsychopharmacology, 2019. **44**(5): p. 939-949.

95. Gentilucci, L., et al., *Molecular docking of opiates and opioid peptides, a tool for the design of selective agonists and antagonists, and for the investigation of atypical ligand-receptor interactions.* Curr Med Chem, 2012. **19**(11): p. 1587-601.

96. Kaserer, T., et al., *mu Opioid receptor: novel antagonists and structural modeling.* Sci Rep, 2016. **6**: p. 21548.

97. Ruiu, S., et al., *Methoxyflavones from Stachys glutinosa with binding affinity to opioid receptors: in silico, in vitro, and in vivo studies.* J Nat Prod, 2015. **78**(1): p. 69-76.

98. Floresta, G., A. Rescifina, and V. Abbate, *Structure-Based Approach for the Prediction of Mu-opioid Binding Affinity of Unclassified Designer Fentanyl-Like Molecules.* Int J Mol Sci, 2019. **20**(9).

99. Ellis, C.R., et al., *Predicting opioid receptor binding affinity of pharmacologically unclassified designer substances using molecular docking.* PLoS One, 2018. **13**(5): p. e0197734.

100. Dumitrascuta, M., et al., *Synthesis, Pharmacology, and Molecular Docking Studies on 6-Desoxo-N-methylmorphinans as Potent mu-Opioid Receptor Agonists.* J Med Chem, 2017. **60**(22): p. 9407-9412.

101. Kormos, C.M., et al., *Design, synthesis, and biological evaluation of (3R)-1,2,3,4-tetrahydro-7-hydroxy-N-[(1S)-1-[[(3R,4R)-4-(3-hydroxyphenyl)-3,4-dim ethyl-1-piperidinyl]methyl]-2-methylpropyl]-3-isoquinolinecarboxamide (JDTic) analogues: in vitro pharmacology and ADME profile.* J Med Chem, 2014. **57**(17): p. 7367-81.

102. Hughes, J.P., et al., *Principles of early drug discovery.* British journal of pharmacology, 2011. **162**(6): p. 1239-1249.

103. Moffat, J.G., et al., *Opportunities and challenges in phenotypic drug discovery: an industry perspective.* Nat Rev Drug Discov, 2017. **16**(8): p. 531-543.

104. Anighoro, A., J. Bajorath, and G. Rastelli, *Polypharmacology: challenges and opportunities in drug discovery.* J Med Chem, 2014. **57**(19): p. 7874-87.

BIOGRAPHY

Srilatha Sakamuru received her Master of Science in Biotechnology from Johns Hopkins University, Baltimore, MD, in 2008. She has been working as a research biologist at National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH) since 2008 (then NIH Chemical Genomics Center). She enrolled for Ph.D. Bioinformatics and Computational Biology program at GMU in fall 2013.