Quantitative Structure-Activity Relationship based Virtual screening for Novel Androgen Receptor Antagonists

by

Xin Ren

B.Sc., Department of Pharmacy
Zhejiang University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Experimental Medicine)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2012

Abstract

Androgen receptor (AR) plays a critical role in prostate cancer development and progression. All current therapeutic AR inhibitors modulate the receptor via direct binding to its Hormone Binding Site (HBS). Despite the identification of other small molecule binding areas on the AR surface including Activation Function 2 (AF2), binding function 3 (BF3), and N-terminal domain (NTD), HBS continues to be the major target site for AR antagonists (even though this site is prone to resistant mutations). Thus, there is a high need for the identification and development of novel antagonists targeting HBS of the AR.

In this study, an effective QSAR modeling pipeline was set up and proved to be capable of identifying new AR antagonists from a large ZINC collection of purchasable chemicals. In particular, we have utilized DRAGON, INDUCTIVE and MOE descriptors to create various binary QSAR models of anti-AR activity. When we have applied the developed QSAR solutions to screen more than 2 million chemicals from the ZINC database, we were able to identify 39 potential candidate AR HBS binders. When they were tested in the DHT displacement assay, 9 chemicals demonstrated the corresponding IC_{50} values in efficient low-micromole range. Of those, 9 compounds later exhibited ability to inhibit AR in the eGFP transcriptional assay with the IC_{50} values established at 1.04-16.18 μM level. Notably, 6 discovered chemicals demonstrated concentration-dependent suppression of survival of LNCaP prostate cancer cell lines.

The results of this study set a ground for the development of an entire novel chemical class of AR antagonists that are distinct for the currently marketed drugs such as Nitalutamide, Flutomide, Cassodex, and MDV3100 that all share significant structural similarity.

Preface

In this work, my contributions to the discovery of novel AR antagonists comprised of designing the experiments, collecting the data on published AR binders, building and evaluating QSAR models, screening ZINC database, performing analyses of both computational and biological results. Additionally, I performed a large part of wet-lab measurements including the work on DHT displacement and MTS assays.

Dr. Artem Cherkasov was involved in discussions of all aspects of this study, especially on aspects of QSAR and choice of analysis techniques. Dr. Steve Jones and Dr. Sandra Dunn are my MSc committee members and were also involved in fine suggestions of computational model design and biological testing. Dr. Eric LeBlanc is the research associate at the prostate center and he gave vital guidance on biological testing on in vitro cell lines. Kate Frewin and intern student Jeffrey Leong performed the rest part of the DHT displacement assay and MTS assay, and all the eGFP assay testing.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I owe particular thanks to my supervisor Dr. Artem Cherkasov, for giving me the opportunity to explore the world of cheminformatics. I would like to thank him for his guidance and constant encouragement of my thesis projects. I especially wanted to thank him for his continued support and his tolerance of my research and life pursuits. I thank Dr. Steve Jones and Dr. Sandra Dunn, my supervisory committee members, for lending their expertise in development of this thesis. I offer my gratitude to the faculty, staff and my fellow students at Vancouver Prostate Centre and UBC's program of Experimental Medicine, who have inspired me to continue my work in this field.

I thank the people working at Vancouver Prostate Centre, Peter Arexio, Eric Leblanc, Nathan Lack, Fuqiang Ban, Natalie Kanaan, Huifang Li, Ravi Shashi Nayana, Thomas Klein and Jeffrey Leong, for their support and advice during completion of this thesis.

Most importantly I would like to thank my family. I specially would like to thank my parents Dengjun Ren and Aiping Zhang, for their constant emotional support, for believing in me, encouraging me, and teaching me never give up.

1. Introduction

1.1 Androgen Receptor as Target for Prostate Cancer Chemotherapy Treatment

1.1.1 Prostate Cancer

There are 1 in 7 Canadian men will be diagnosed with prostate cancer, making it the most commonly diagnosed non-skin cancer in men, and one of the leading causes of cancer-death (Society 2012). Approximately 26,500 Canadians will be diagnosed with prostate cancer in 2012. On average, 73 Canadian men are diagnosed with prostate cancer every day and at the same time, on average 11 Canadian men die from the disease every day (Society 2012). While frequently curable in early stage of the disease by surgery or radiation ablation, the first line of treatment for locally advanced, recurrent or metastatic prostate cancer is some form of androgen withdrawal therapy, which is generally designed to block either the production of androgens or their binding to the androgen receptor (AR) (Sharifi, Gulley et al. 2005). Unfortunately, the effectiveness of this type of treatment is usually temporary due to the progression of surviving tumor cells to a castration-resistant state (Gleave, Goldenberg et al. 1998; Albertsen, Hanley et al. 2005), with no curative treatment options and a median life expectancy of around 18 months (Kent and Hussain 2003; Martel, Gumerlock et al. 2003; Antonarakis and Eisenberger 2011). While the molecular mechanisms responsible for progression to the castration-resistant phenotype are largely unknown, typically they do not appear to involve loss of AR expression (Taplin, Rajeshkumar et al. 2003). In over 80% of locally advanced castration-resistant prostate cancers (CRPCs), high levels of nuclear AR have been observed (Tilley, Lim-Tio et al. 1994); and in bone metastases, the amount of AR present is often higher than in primary tumors (Hobisch, Culig et al. 1995).There is evidence that in most cases some form of inappropriate activation of AR is linked to recurrent growth of prostate cancers (Rennie and Nelson 1998).

Since AR is central to the progression to castration resistance, AR knockdown or alternative AR inhibition strategies have been proposed as additional therapy after failure of conventional androgen ablation (Scher, Buchanan et al. 2004).

1.1.2 Androgen Receptor

Androgen receptor (AR) is a member of the steroid and nuclear receptor superfamily, which is regulated by the binding of androgens, mainly testosterone and 5α-dihydrotestosterone (DHT) (Roy, Lavrovsky et al. 1999).

The AR gene is more than 90 kb long and codes for a protein of 919 amino acids that has three major functional domains, as illustrated in Figure 1. The N-terminal domain (NTD), which has a modulatory function, is encoded by exon 1 (1586 bp). The DNA-binding domain (DBD) is encoded by exons 2 and 3 (152 and 117 bp, respectively). The C-terminal ligand binding domain (LBD) which associates with an HSP90 chaperone is encoded by five exons from 131 to 288 bp. There is also a small hinge region between the DNA-binding domain and ligand-binding domain containing a nuclear localization signal (NLS) (Heemers and Tindall 2007; Narayanan, Mohler et al. 2008; Gao 2010).

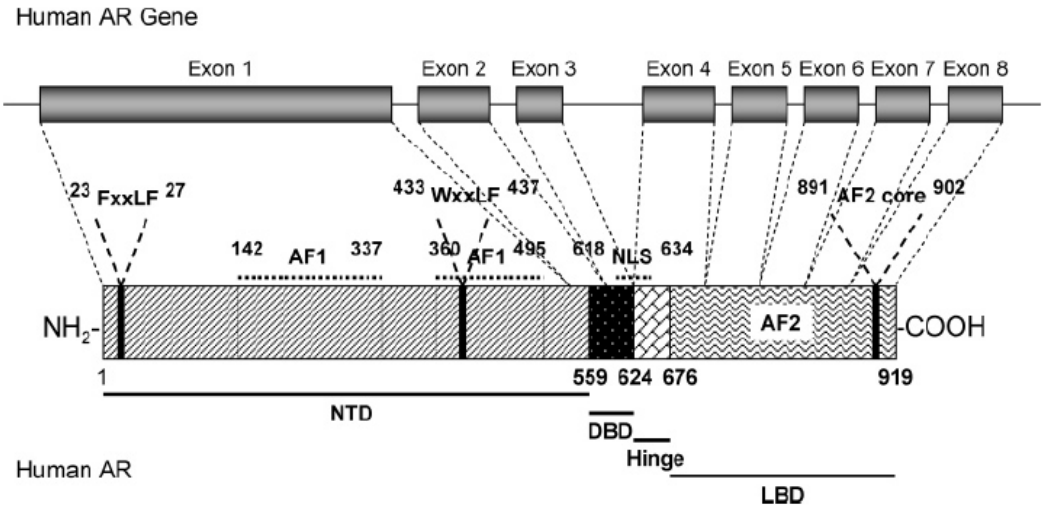


Figure 1 Structural organization of the AR gene and protein.

The LBD is a multifunctional domain, which is very important for ligand recognition and contributes to dimerization and co-regulator interactions. The LBD contains 11 α-helices (H) and two β-turns, arranged in three layers to form an antiparallel "α-helical sandwich". As shown on

Figure 2, there are 11 helices in AR LBD with H2 absent. The central layer was formed by helices H4, H5, H8, H9 and the first β-turn, flanked from one side by H1, H3, and from the other side by H6, H7, H10, H11 and the second β-turn. Hormone Biding Site (HBS) resides in the interior of the LBD underneath the central helical layer, which is intrinsically flexible to support binding of ligands with different sizes (Gao 2010; van de Wijngaart, Molier et al. 2010). Notably, in the active agonist-bound LBD conformation, H12 is placed over the HBS like a lid and spans all the three helical layers. There are other interaction sites available outside of the HBS, such as activation function 2 (AF2) (Axerio-Cilies, Lack et al. 2011) and binding function 3 (BF3)(Lack, Axerio-Cilies et al. 2011).

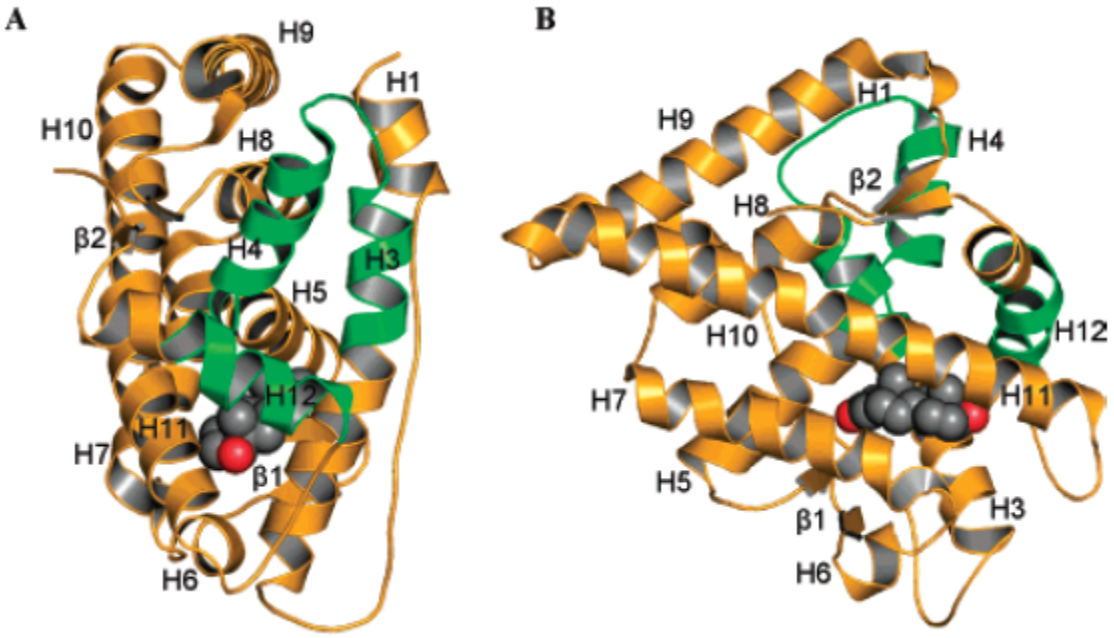


Figure 2 Crystal structures of wild-type AR ligand binding domain bound with DHT

(A) front view; (B) ligand view. Space filled atoms are (black) carbon and (red) oxygen. The activation function 2 region (helices 3, 4, and 12) is highlighted in green.

Binding of agonist (DHT) alters LBD structure (Figure 3) to form the coactivator-binding site, which initially binds an FQNLF peptide in the NTD (represented by triangle). The resulting

intramolecular N-terminal or C-terminal interaction may function to expose the NLS and enhance AR nuclear translocation, chromatin binding, and the initial recruitment of coactivator and chromatin-modifying proteins by the NTD (Masiello, Cheng et al. 2002). Direct interaction between closely positioned DBDs and possibly intermolecular N or C interactions stabilizes the AR homodimer on the DNA. The LBD and NTD then cooperatively recruit transcriptional coactivators such as TRAP220 and steroid receptor coactivators (SRC) that interact with the LBD coactivator-binding site through LxxLL motifs (displacing the NTD FQNLF peptide) (Dubbink, Hersmus et al. 2004). SRC recruitment of the SWI/SNF complex, histone acetyltransferases (CBP/p300), protein methyltransferases (CARM1), and additional factors results in the AR-targeted relaxation of chromatin and propagation of a transcriptionally active gene locus (Janne and Shan 1991; Knee, Froesch et al. 2001; McEwan 2004).

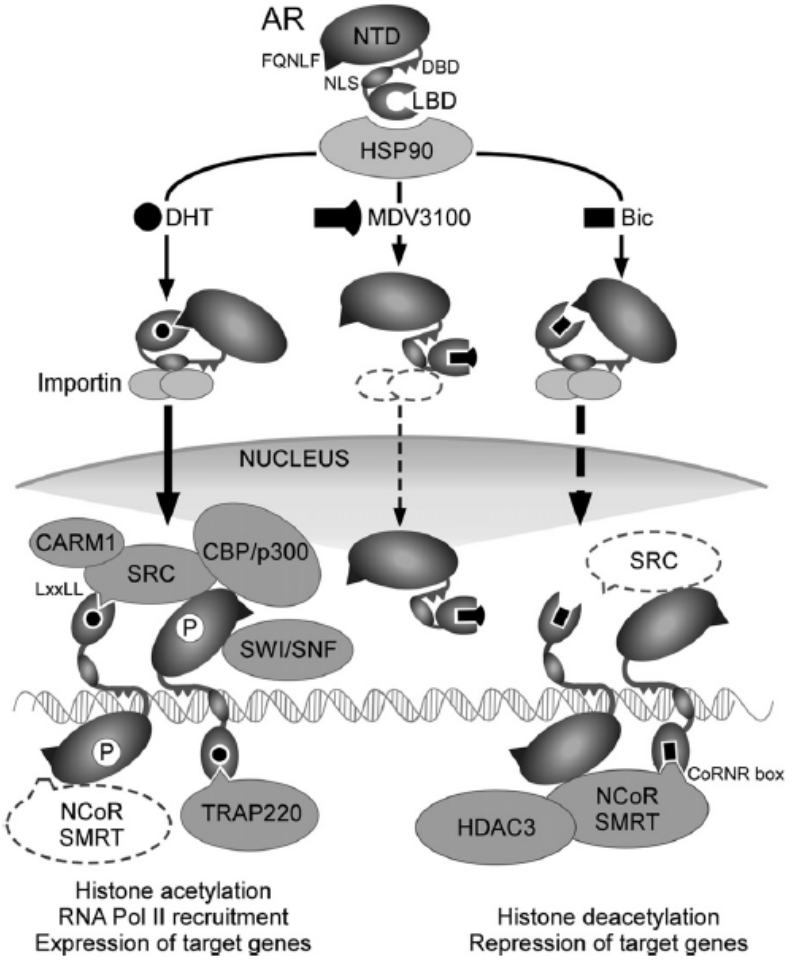


Figure 3 Agonist and Antagonist Modulation of AR Transcriptional Activity

AR antagonists such as Bicalutamide (Bic) still promote nuclear translocation and chromatin binding but fail to induce optimal LBD helix 12 repositioning for generation of the coactivator-binding site. This way AR lacks transcriptional activity due to ineffective coactivator recruitment and enhanced recruitment of corepressors (NCoR and SMRT), which bind weakly to the NTD and are stabilized by binding of extended LxxLL-like motifs (CoRNR boxes) to the LBD. However, in CRPC, high-level AR expression and other mechanisms may increase the recruitment of coactivators versus corepressors, leading to agonist activity. In contrast to Bicalutamide, MDV3100 more effectively impairs nuclear translocation and appears to completely prevent chromatin binding, which may reflect further displacement of helix 12 and abrogation of the FQNLF/LxxLL coactivator-binding site (Saporita, Zhang et al. 2003; Dubbink, Hersmus et al. 2004; Shen and Balk 2009; Foster, Car et al. 2011).

1.1.3 Prostate Cancer Therapeutics

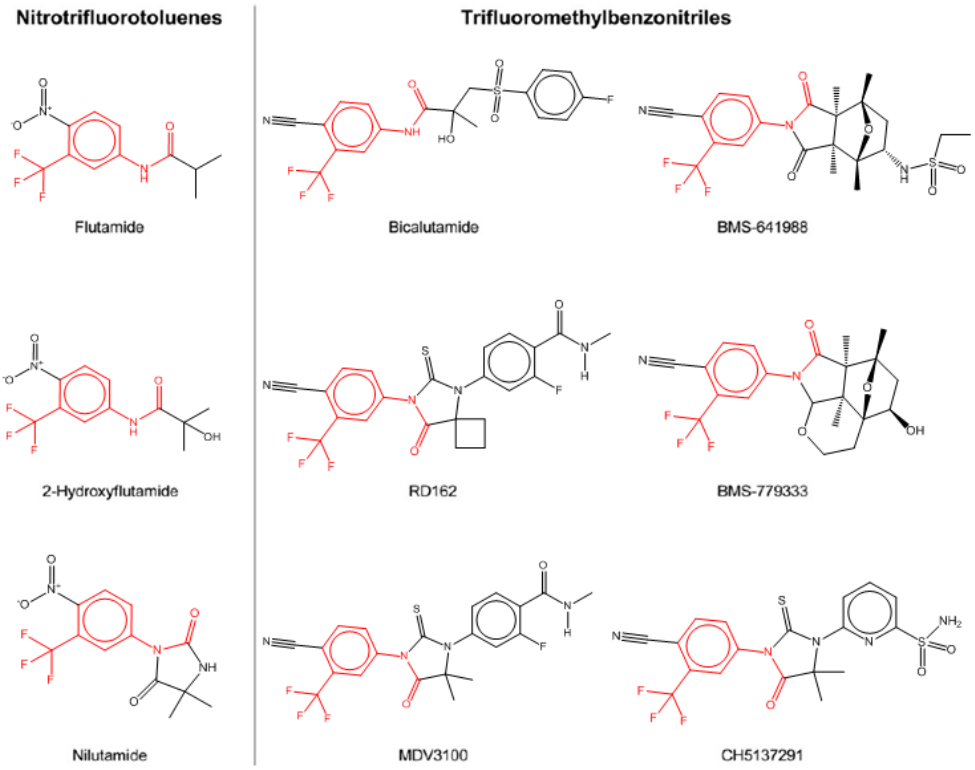


Figure 4 Structures of marketed and developed non-steroid antiandrogens

The first antiandrogen tested in the clinic was Cyproterone acetate in 1966. It functions as a AR antagonist and also decreases the AR levels *in vivo* (Haendler and Cleve 2012).

Flutamide is a non-steroidal chemical with pure antiandrogenic properties (Figure 4). It is derived from acetanilide and it is converted to active form 2-hydroxyflutamide in vivo. It was approved for treatment of locally confined metastatic prostate cancer in combination with a GnRH agonist by FDA in 1989 (Brogden and Chrisp 1991).

Nilutamide is an AR antagonist which is derived from nitrotrifluorotoluene group and is structurally related to Flutamide (Figure 4), but has a longer half-life. It is approved at a 300 mg followed by a 150 mg daily dose in 1996. Its use is limited due to side-effects such as pneumonitis and delayed adaptation to darkness (Dole and Holdsworth 1997).

Bicalutamide (marketed as Casodex, Cosudex, Calutide, Kalumid) is an oral non-steroidal anti-androgen drug for prostate cancer (Figure 4). It was first approved in 1995 as a combination treatment (with surgical or medical castration) for advanced prostate cancer and subsequently launched as monotherapy for the treatment of earlier stages of the cancer (Cockshott 2004). It not only inhibits the activation of the AR but also enhance the degradation of the androgen receptor. The R-enantiomer of this drug is the active conformation (Mukherjee, Kirkovsky et al. 1996). It has a better affinity for the AR, a longer half-life and accumulates in the plasma whereas the S-enantiomer is inactive.

The recently developed RD162 and MDV3100 (Figure 4) were selected based on their high binding affinity for the AR and on their ability to impair the nuclear import of the AR/ligand complex. They are diarylthiohydantoins, derivatives of Nilutamide (Tran, Ouk et al. 2009; Jung, Ouk et al. 2010). They do not enhance translocation of AR to the nucleus. In addition they prevent binding of AR to DNA and AR to coactivator proteins. MDV3100 has about fivefold higher binding affinity for the AR compared to Bicalutamide (Makkonen, Kauhanen et al. 2011).

BMS-641988 is an oxabicyclo-imide-derived AR antagonist (Figure 4) with approximately twenty fold higher AR-binding properties than Bicalutamide, and also three to seven fold higher antagonistic properties in transactivation assays than Bicalutamide (Attar, Jure-Kunkel et al. 2009). BMS-779333 (Figure 4) is a second generation AR antagonist with a comparable anti-tumor profile to BMS-641988, to which it is structurally related (Salvati, Balog et al. 2008).

The dimethylthiohydantoin derivative CH5137291 (Figure 4) is also one among the trifluoromethylbenzonitrile group. It suppresses the growth of CRPC xenograft model LNCaP-BC2 and reduces plasma PSA levels in mice (Yoshino, Sato et al. 2010).

It should be outlined, that all these chemicals share close structural similarity (a common motif is highlighted in red in Figure 4). Therefore, all these chemicals can suffer from the same problems, such as arising resistance driven by mutations in in the AR LBD, which results in a high medical need for the identification and development of alternative, novel antagonists of AR HBS, belonging to a different chemical class.

1.2 Quantitative Structure-Activity Relationships

Quantitative structure–activity relationship (QSAR) approach describes a mathematical relationship between biological activity of a molecular system and its geometric and chemical characteristics. QSAR models can be based on regression, non-linear or classification approximations that could be applied to evaluate activity of new molecules (Dudek, Arodz et al. 2006).

QSAR generates predictive models using statistical tools correlating biological activity (including desirable therapeutic effect and undesirable side effects) of chemicals (drugs/toxicants/environmental pollutants) with descriptors or features representative of molecular structure and properties (Patani and LaVoie 1996). QSAR solutions find applications in many disciplines including risk assessment, toxicity prediction, and regulatory decisions (Tong, Hong et al. 2005) in addition to drug discovery and lead optimization (Dearden 2003). Obtaining a good quality QSAR model is a task depending on many factors, such as quality of biological data, choice of descriptors and statistical methods used (Craig, Hansch et al. 1971; Keiser, Roth et al. 2007). Any QSAR modeling effort should ultimately lead to statistically robust solutions capable of making accurate and reliable predictions of biological activities of yet unstudied chemicals (Keiser, Roth et al. 2007).

The basic assumption of the QSAR is that similar molecules have similar activities. The underlying problem is therefore how to define a small difference on a molecular level, since each kind of activity, e.g. reaction ability, biotransformation ability, solubility, target activity, and so on, might depend on another differences (Patani and LaVoie 1996). In general, researchers are more interested in finding strong trends. QSAR hypotheses usually rely on a limited number of chemical data points. Thus, the further goal is to avoid over fitted hypotheses, and remove over fitted and useless interpretations on structural or molecular data set.

1.2.1 Modeling: Data Mining Approach

QSAR models typically utilize relatively large number of molecular features (descriptors). Since QSAR descriptors often lack structural interpretation ability, the preprocessing steps face a feature selection problem, which can be achieved by visual inspection, by data mining, or by molecule mining. The former typically use such algorithms as support vector machines (Cortes and Vapnik 1995), decision trees (Yuan and Shaw 1995; Ramos-Jimenez, del Campo-Avila et al. 2005) and neural networks (Hopfield 1982) among others to induce a predictive learning model. The latter, molecule mining, is a special example of structured data mining approaches, predicting structure features matrix or performing a fragmentation of a molecules into catalogued substructures (Helma 2005).

1.2.2 Evaluation of Quality of QSAR Models

Validation is the process by which reliability and relevance of a procedure are established for a specific purpose (Roy 2007). For validation of QSAR models usually four strategies are adopted (Hawkins, Basak et al. 2003; Konovalov, Sim et al. 2008): 1. internal validation or cross-validation; 2. validation by dividing the data set into training and test components; 3. true external validation by application of model on external data; 4. data randomization or Y-scrambling (Roy, Paul et al. 2009).

For a binary classification model, its outcomes are labeled either as positive (p) or negative (n). Naturally, there are four possible outcome combinations generated by any binary classifier. If the outcome from a prediction is p and the actual value is also p, then the prediction is called a true positive (TP); however if the actual value is n then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are n, and false negative (FN) is when the prediction outcome is n while the actual value is p. a receiver operating characteristic (ROC) graphically presents the model behavior in a visual way (Li and Gramatica 2010). A ROC curve, which has been proved to be a valuable way to evaluate the quality of a two-class classifier, shows the separation ability of a binary classifier by

iteratively setting the possible classifier threshold. As a result, a plot of the trade-off between the sensitivity (y-axis) and 1-specificity (x-axis) can be obtained. If the plot has Area under the Curve (AUC) of 1, a perfect classifier is found, and if the area equals 0.5, the classifier has no discriminative power at all (Li and Gramatica 2010).

Sensitivity or true positive rate (TPR)

$$TPR = TP/P = TP/(TP + FN)$$

Specificity (SPC) or True Negative Rate

$$SPC = TN/N = TN/(FP + TN) = 1 - FPR$$

Positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

Negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

The success of any QSAR model depends on accurate and clean training data (Roy, Leonard et al. 2008), proper representative descriptors selection methods (Roy and Roy 2008), statistical suitable statistical methods (Roy, Leonard et al. 2008), and most critically both internal and external validation of the built models (Shao 1993; Rao and Wu 2005).

2 Materials and Methods

2.1 Studies Data Sets

Training Set-1 (T-1): This set was taken from the literature and included 625 chemicals known to interact with the AR (Li and Gramatica 2010). The dataset consisted of 394 data points measured by a Danish lab (Li and Gramatica 2010), and 231 activity numbers collected from various open sources. The corresponding experimental values (IC_{25}, μM) stand for the ability of a chemical to inhibit the luminescence response induced by the synthetic androgen R1881. If a given chemical exhibits IC_{25} value at test concentration of <10 μM, then it would be classified as active. If IC_{25} > 10 μM, or if cytotoxicity of a chemical is over IC_{50}=3μM, then the chemical would be classified as inactive. The structures of the studied chemicals were obtained from the supplementary materials of the article (Li and Gramatica 2010) and then were optimized by the MMFF94x forcefield implemented in the Molecular Operating Environment (MOE) program (Boyd 2005).

Training Set-2 (T-2): This data set included 595 activity points collected from public databases. In particular, we investigated such databases as CHEMBL (Gaulton, Bellis et al. 2012), BindingDB (Gilson, Chen et al. 2001), DrugBank (Wishart, Knox et al. 2008) and ChemSpider (Williams and Tkachenko 2010) for IC_{50} parameters of AR inhibition. If IC_{50} value of a given chemical was lower than 20 μM, we have defined it as active.

External Testing Data Set: we have reserved 89 molecules u also used as an external set in the literature (Li and Gramatica 2010) to additionally check the developed QSAR models for their predicting ability.

Combined Set (CS): merging T-1 and T-2 datasets resulted in a Combined Set that has also been used independently for QSAR modeling. All entries from the training set-1, training set 2, and external test set are presented in the Appendices section.

2.2 Database for Virtual Screening

In this study ZINC molecular database (Irwin and Shoichet 2005) was screened by the developed QSAR binary models. This database represents a curated collection of commercially available chemical chemicals prepared especially for virtual screening.

In particular, we have considered a subset of ZINC containing 2 million lead-like chemical structures. The database was firstly processed with MOE 2010 (Molecular Operating Environment) (Boyd 2005) program by applying its 'washing' protocol, followed by the 3D rebuilding of structures and their partial charge assignment, using the default MOE settings.

2.3 Molecular Descriptors

Descriptor is the final result of a logic or mathematical procedure that transforms chemical encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment (Todeschini 2009).

2.3.1 Dragon Descriptors

Dragon is commercial software for calculating molecular descriptors. Dragon descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high throughput screening of molecule databases (Talete 2007). Currently DRAGON collection contains 3214 theoretical descriptors that could be are divided into the following categories: (a) 0D constitutional descriptors, (b) 1D count of functional groups and atom-centered fragments; (c) 2D topological descriptors, walk and path counts, connectivity and information indexes, various autocorrelations from the molecular graph, edge adjacency indices, descriptors of Burden eigenvalues (Burden 1989; Burden 1997), topological charge and eigenvalues-based indices, and 2D binary and frequency fingerprints; (d) 3D Randic molecular profiles, geometrical descriptors, weighted holistic invariant molecular descriptors (WHIMs)

(Todeschini and Gramatica 1997; Gramatica, Navas et al. 1998), and geometry, topology, and atom-weights assembly (GETAWAY) descriptors (Consonni, Todeschini et al. 2002; Consonni, Todeschini et al. 2002), (e) charge descriptors, and (f) molecular properties. The list and meaning of the molecular descriptors is provided by the DRAGON package, and the calculation procedure is explained in detail (Talete 2007).

2.3.2 INDUCTIVE Descriptors

A set of 50 INDUCTIVE descriptors (Cherkasov 2003; Cherkasov and Jankovic 2004; Cherkasov 2005; Cherkasov, Shi et al. 2005) have been calculated initially for all studied chemicals. During the calculation all hydrogen atoms were suppressed and only heavy atoms have been taken into account. The inductive QSAR descriptors were calculated from values of atomic electro-negativities and radii by using custom SVL-scripts downloaded from the SVL exchanger and implemented within the MOE package. To avoid cross correlation among the independent variables, we have computed pairwise correlation among all the 50 QSAR parameters and removed those inductive descriptors which formed any linear dependence with R correlation coefficient ≥ 0.95. As a result of this procedure, only 30 inductive QSAR descriptors have been selected.

2.3.3 MOE Descriptors

MOE descriptors include both 2D and 3D molecular parameters. The 2D molecular descriptors are numerical features derived from the connection table representing a molecule and include physical properties, subdivided surface areas, atom counts, bond counts, Kier and Hall connectivity and Kappa Shape indices, adjacency and distance matrix descriptors containing BCUT and GCUT descriptors, pharmacophore feature descriptors, and partial charge descriptors. 3D molecular descriptors, which are dependent on the conformation of a molecule, include potential energy descriptors, surface area, volume, shape descriptors, and charge descriptors (Boyd 2005). Descriptors are partitioned into *classes*. Each class indicates what is assumed by the descriptor calculators about the molecule presented:

- **2D**. 2D descriptors only use the atoms and connection information of the molecule for the calculation. 3D coordinates and individual conformations are not considered.
- **i3D**. Internal 3D descriptors use 3D coordinate information about each molecule; however, they are invariant to rotations and translations of the conformation.
- **x3D**. External 3D descriptors also use 3D coordinate information but also require an absolute frame of reference (e.g., molecules docked into the same receptor).

2.3.4 Descriptor Selection

The recursive feature elimination (RFE) (Marko Robnik-Sikonja 1997) based on support vector machines (SVM) method was used in this research as the feature selection method for selecting molecular descriptors associated to AR antagonist activity. RFE is an iterative procedure for backward feature elimination, which can be executed in WEKA (Mark Hall 2009), where the descriptors are normalized by default. The first step of SVM-RFE is to train a SVM classifier with all the features. Suppose there is a training set, x= (x_1, x_2, ..., x_i, ..., x_n) and y = (y_1, y_2, ..., y_i, ...,y_n). SVM solves the classification problem by minimizing the following equation:

$$J = (1/2)\sum_{ij} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j + \lambda \delta_{ij}) - \sum_i \alpha_i \quad \text{subject to} \quad 0 \le \alpha_i \le C \quad \text{and} \quad \sum_i \alpha_i y_i = 0$$

Where α is the Langrange coefficient, δ_{ij} is the Kronecker symbol (if $i = j$ and $\delta_{ij} = 1$, otherwise 0), λ and C are two parameters needed to be optimized by SVM. The output of this solution is α_i. The resulting decision function is: $f(x) = w \cdot x + b$, where w is the weight vector calculated as $w = \sum i\ \alpha_i y_i x_i$. The ranking criterion is the square of the weight, calculated as $c_l = (w_l)^2$ for the l th feature. Then the feature with the smallest cl is removed. For computational reasons, it may be more efficient to remove several features in each cycle. At the end, all the features are ranked. The features on the top are the more informative ones.

2.4 Modeling Methods

In this study, 9 classification methods were used, all implemented through the WEKA software. In particular, we have utilized k-Nearest Neighbors (kNN) (Kachigan 1991), Local Lazy method (lazy IB1) (D. Aha 1991), Alternating Decision Tree (ADTree) (Bernhard Pfahringer 2001), Artificial Neural Network (ANN) (Egmont-Petersen 2002), K-star method (John G. Cleary 1995), Bagging method (Breiman 1996), as well as LogitBoost (J. Friedman 2000), Decorate (Breiman 2001), and Random Forest(P. Melville 2003).

2.5 Applicability Domain (AD)

The applicability domain (AD) of a QSAR is the physical or chemical, structural or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new chemicals. To verify the practical applicability of our models to chemicals not used in the model development, the model's AD, which is a theoretical region defined by the used descriptors in modeling, should be quantitatively assessed. To investigate the AD of a training set of chemicals, one can directly analyze properties of the multivariate descriptor space of the training chemicals or more indirectly via range vectors. The values of each descriptor of training set chemicals are checked and the range of the very descriptor is determined. When applying to the test set, if a chemical with all its descriptor values in the range of the values of training set chemicals, then the prediction of this test chemical is in the chemical space of the model and could be reliable (Tropsha, Gramatica et al. 2003; Gramatica 2007).

2.6 Chemical Sources and Conformation

Selected chemicals were purchased from established suppliers, including Asinex, Chembridge, and Sigma. The identity and purity of all chemicals were confirmed by mass spectroscopy (MS) and liquid chromatography–tandem mass spectrometry (LC–MS/MS).

2.7 MTS Assay

The MTS assay is a cell viability assay used to test compounds for their cytotoxicity on cells. In Dr. Paul Rennie's lab, the MTS assay is used to rule out any non-specific inhibition of AR transcriptional activity according to the manufacturer's protocol (CellTiter 961 Aqueous One Solution Reagent, Promega), steps are as following:

1) Seed cells (LNCaP) in RPMI 1640 media supplemented with 10 % CSS at a concentration of 5000 cells per 100 µl in each well. For one compound, seed for 30 wells (3 rows and 10 columns). Seed in rows A to C and columns 2-11. Include a blank control row if needed. Incubate the cells at 37°C for 24 hours.
2) Prepare a 2X concentration of the compound. This can be done in a 24 well plate. First, determine what 2X concentration you will prepare. Prepare the 2X compound solution and 2X DMSO solution as follows:

			2X Concentration
450 µl RPMI 1640 + 10% CSS	+	*450 µl solution 3.)*	*→ 25 µM.*
450 µl RPMI 1640 + 10% CSS	+	*450 µl solution 5.)*	*→ 12.5 µM.*
450 µl RPMI 1640 + 10% CSS	+	*450 µl solution 6.)*	*→ 6.25 µM.*
1000 µl of CSS	+	*2 µl DMSO*	*→ 0.2% DMSO*

3) Prepare Casodex and R1881 as follows:

100 µl RPMI 1640 + 10 % CSS	+	*1 µl (100mM) Casodex*	*→ 1 mM*
1000 µl RPMI 1640 + 10 % CSS	+	*1 µl (10µM) R1881*	*→ 10 nM*

4) Now that all the solutions are prepared, take the cells that were seeded the previous day out of the incubator. Add 100 µl of the 2X compound solution, 2 µl of the casodex solution and 2 µl of the R1881 solution as follows in triplicates. In the row below, add 100 µl of RPMI 1640

+ 10 % CSS and 100 µl of the compound solution as shown. This will be used as the blank. Incubate the cells for 72 hours (or 96 hours depending on your experiment) at 37°C.

	2	3	4	5	6	7
A	100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln -100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln - 2 µl of Casodex Soln - 100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln -100 µl of 6.25 µM Cmpd	- 2µl of R1881 Soln - 100 µl of 12.5 µM Cmpd	- 2µl of R1881 Soln -100 µl of 25 µM Cmpd
B	100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln -100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln - 2 µl of Casodex Soln - 100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln -100 µl of 6.25 µM Cmpd	- 2µl of R1881 Soln - 100 µl of 12.5 µM Cmpd	- 2µl of R1881 Soln -100 µl of 25 µM Cmpd
C	100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln -100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln - 2 µl of Casodex Soln - 100 µl CNTRL DMSO Soln	- 2µl of R1881 Soln -100 µl of 6.25 µM Cmpd	- 2µl of R1881 Soln - 100 µl of 12.5 µM Cmpd	- 2µl of R1881 Soln -100 µl of 25 µM Cmpd
D				-100 µl of RPMI 1640 + 10% CSS -100 µl of 6.25 µM Cmpd	-100 µl of RPMI 1640 + 10% CSS - 100 µl of 12.5 µM Cmpd	-100 µl of RPMI 1640 + 10% CSS -100 µl of 25 µM Cmpd

5) Take the cells out of the incubator and add 20 µl of MTS reagent (found in the fridge in the cell culture room or the -30ºC freezer) into each well. Into any empty well, add 200 µl of RPMI 1640 media + 10%CSS and also add 20 µl of the MTS reagent. Incubate for 1 hour at 37ºC. Read using the TECAN Infinite F500 at 492nm.

2.8 eGFP Cellular Transcription Assay

In the eGFP assay, an LNCaP cell line with a stably transfected androgen responsive probasin-derived promoter upstream of an eGFP reporter gene (ARR2PB-EGFP) is used. Since eGFP expression will be regulated by the AR, an increase in fluorescence will also correlate to an increase in AR induced transcriptional activity. These LN-ARR2PB-EGFP cells are used to screen small molecules against AR transcriptional activity (Tavassoli, Snoek et al. 2007). Briefly, stably transfected eGFP-expressing LNCaP human prostate cancer cells (LN-ARR2PB-eGFP) containing an androgen-responsive probasin-derived promoter (ARR2PB) were grown in phenol-red-free RPMI 1640 supplemented with 5% CSS. After 5 days, the cells were plated into a 96-well plate (35000 cells/well) with 0.1 nM R1881 and increasing concentrations (0–100 µM) of chemical. The cells were incubated for 3 days, and the fluorescence was then measured (excitation, 485 nm; emission, 535 nm). The viability of these cells has been assayed by the MTS cell proliferation assay (CellTiter 961 Aqueous One Solution Reagent, Promega) according to the instructions of the manufacturer.

2.9 Androgen Displacement Assay

Androgen displacement was assessed with the Polar Screen Androgen Receptor Competitor Green Assay Kit as per the instructions of the manufacturer (www.invitrogen.com). Kit contains the reagents necessary to perform a competition assay for identification of Androgen receptor (AR) binding chemicals. The kit uses the rat AR ligand-binding domain tagged with His and GST [AR-LBD(His-GST)]. AR-LBD(His-GST) is added to a fluorescently-tagged androgen ligand (Fluormone™AL Green) in the presence of competitor test chemicals in microwell plates. The presence of effective competitors prevents the formation of a AL Green/ AR-LBD(His-GST) complex resulting in a decrease of the polarization value due to ligand

displacement caused by a competitor. The shift in polarization value in the presence of test chemicals is used to determine relative affinity of test chemicals for AR-LBD(His-GST). 1. Dispense 20 µL 2X Test Chemical in the microwell plate. 2. Add 20 µL 2X AR-LBD (His-GST)/Fluormone™ AL Green Complex to the same plate and mix. 3. Cover the AR Green Assay plates to protect the reagents from light. Incubate the AR Green Assay plates at 20-25°C for 4-8 hours. The polarization values vary less than 10% from maximum values if read within this time period. 4. Measure polarization value of each well.

2.10 Determination of Chemical Purity

Chemical identity and purity of the tested chemicals were confirmed by LC–MS/MS. Briefly, an Acquity ultra-performance liquid chromatograph (UPLC) with a 2.1 × 100 mm BEH, 1.7 µM, C18 column coupled to a photodiode array (PDA) detector in line with a Quattro Premier XE (Waters, Milford, MA) was used with water and acetonitrile containing 0.1% formic acid as mobile phases. A 5–95% acetonitrile gradient from 0.2–10.0 min was used, and 95% was maintained for 2 min followed by re-equilibration to starting conditions for a total run time of 15.0 min. The MS was run at unit resolution with 3 kV capillary, 120 and 300 °C source and desolvation temperatures, 50 and 1000 L/h cone and desolvation N_2 gas flows, and Ar collision gas set to 7.4–3 mbar. On the basis of the full range of the diode array absorbance (210–800 nm), the relative purity [AUCCMPD versus area under the curve (AUC) of all other peaks] was calculated. All chemicals described had a purity of >90–95%.

3. Results

3.1 Development and Validation of the QSAR Models

3.1.1 Development and Cross Validation of QSAR Models

The WEKA software was used to construct QSAR models for screening for potential AR antagonists. WEKA is a collection of machine learning algorithms typically used in data mining studies. In addition, WEKA includes functionalities for data pre-processing, classification, regression, clustering, association and visualization. It is also well-suited for developing new machine learning schemes.

To all studied chemicals of Training Set-2, their anti-AR activity values were transformed to the binary form using IC_{50}=20μM cutoff value. Anti-AR activity of chemicals from Training Set-1 and External Testing Data Set (Li and Gramatica 2010) was already binary and needed no transformation. After that three sets of descriptors including DRAGON, INDUCTIVE and MOE were calculated for all chemicals in the combined set (CS). Consequently, the RFE method (Marko Robnik-Sikonja 1997) was applied to select the most effective descriptors. All descriptors with low variance < 0.1) were removed from the following consideration. Additionally, all correlated descriptors (with the corresponding $R \geq 0.95$) have also been removed.

At the next stage, various machine learning methods were applied here, including k-Nearest Neighbors (kNN) approach, Local Lazy method (lazy IB1), Alternating Decision Tree (ADTree), Artificial Neural Network (ANN), K-star method, Bagging method, as well as LogitBoost, Decorate, and Random Forest to relate the selected QSAR descriptors of compounds to their binary activity parameters.

In order to test the level of performance that we can expect for this data set using all available descriptors, QSAR models were built only based on training set 1 (see Table 1), in order to reproduce similar quality models compared with Li and Gramatica's models (Li and Gramatica 2010). Descriptors used in these models are listed in table 2. In this step, models were

investigated by different number of descriptors, to find an optimized number of descriptors (see Figure 6). To make sure that the QSAR models are rigorous enough, 10 fold cross validation was also used. From the results presented in Table 1 it could be concluded that for each modeling method, descriptors number range from 25 to 30 would give best performance of models with better AUC values of ROC.

After making sure that all data points are reproducible, the training set 1 and training set 2 were combined and modeled again (Table 3). Such undertaking required another description section round, and the corresponding selected descriptors are presented in Table 4. As data in Table 3 indicate, the same conclusion could be achieved - that the optimal number of QSAR descriptor is in the range of 25-30 parameters (Figure 7). The correlation table of all descriptors is also listed here, showing that there are no highly correlated descriptors were included (Table 5). The resulting sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and Area under the Curve (AUC) of receiver operating characteristic (ROC) parameters obtained for the training set 1, training set 2 and combined set are presented in the Appendices section.

Table 1 ROC AUC of different methods based on Training Set 1

Desc. No.	ADTree	ANN	Bagging	Decorate	kNN	Kstar	Lazy-IB1	Logitboost	RandomForest
15	0.765	0.772	0.808	0.806	0.79	0.799	0.691	0.805	0.794
16	0.765	0.754	0.805	0.801	0.781	0.798	0.694	0.798	0.795
17	0.763	0.756	0.809	0.794	0.782	0.793	0.691	0.804	0.806
18	0.781	0.758	0.814	0.798	0.781	0.785	0.691	0.81	0.814
19	0.778	0.74	0.826	0.786	0.782	0.782	0.705	0.799	0.816
20	0.779	0.757	0.825	0.816	0.787	0.787	0.692	0.803	0.819
21	0.789	0.77	0.827	0.815	0.794	0.777	0.691	0.817	0.817
22	0.785	0.762	0.828	0.812	0.796	0.782	0.697	0.83	0.821

Desc. No.	ADTree	ANN	Bagging	Decorate	kNN	Kstar	Lazy-IB1	Logitboost	RandomForest
23	0.785	0.776	0.827	0.81	0.806	0.785	0.698	0.827	0.834
24	0.778	0.771	0.83	0.825	0.81	0.782	0.699	0.822	0.822
25	0.78	0.791	0.829	0.803	0.808	0.778	0.707	0.822	0.84
26	0.771	0.766	0.822	0.819	0.804	0.782	0.709	0.817	0.818
27	0.77	0.783	0.83	0.813	0.783	0.79	0.707	0.817	0.809
28	0.773	0.804	0.828	0.82	0.819	0.811	0.715	0.817	0.816
29	0.768	0.777	0.829	0.811	0.824	0.802	0.726	0.821	0.832
30	0.783	0.825	0.832	0.804	0.817	0.801	0.721	0.818	0.827
best	0.789	0.825	0.832	0.82	0.824	0.811	0.726	0.83	0.84

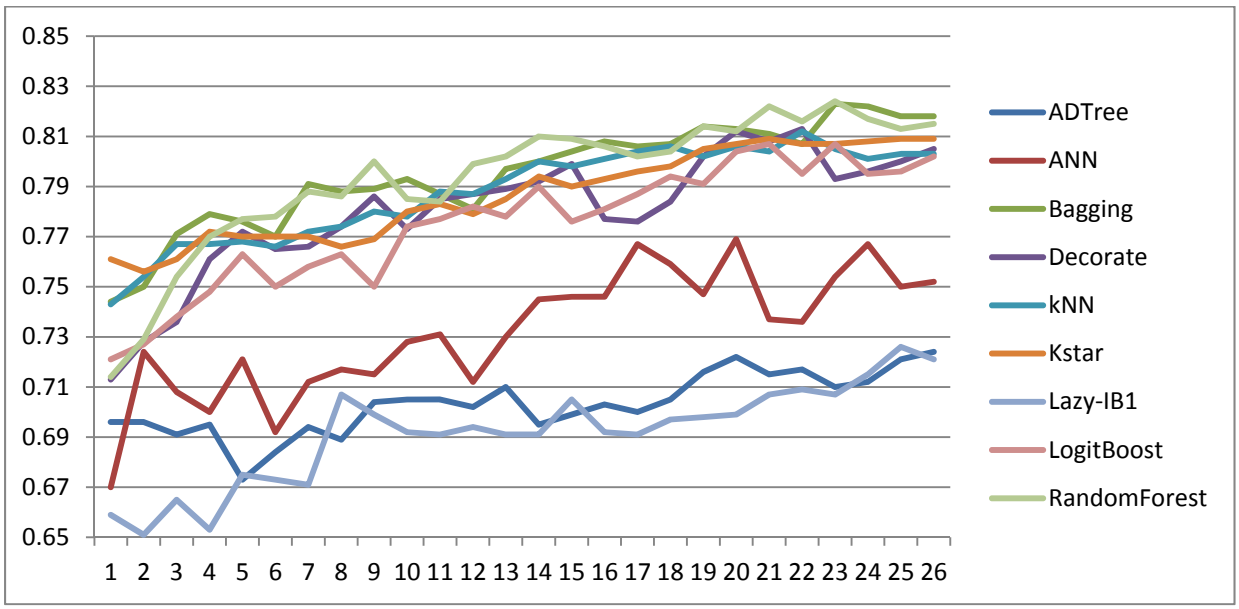


Figure 5 Correlation of AUC and number of descriptors of Training Set 1

Table 2 Descriptors selected basing on Training Set 1 and their corresponding meanings

Symbol	Definition	Class
ARR	aromatic ratio	constitutional descriptors
b_double	Number of double bonds. Aromatic bonds are not considered to be double bonds	2D
B03[O-O]	presence/absence of O - O at topological distance 03	2D binary fingerprints
B05[N-F]	presence/absence of N - F at topological distance 05	2D binary fingerprints
BCUT_SLOGP_2	LogP BCUT (2/3)	2D
BCUT_SLOGP_3	LogP BCUT (3/3)	2D
E_vdw	Van der Waals energy	i3D
Hardness_of_Most_Pos	Atomic hardness of an atom with the most positive charge	IND
IVDE	mean information content on the vertex degree equality	information indices
JGI6	mean topological charge index of order6	topological charge indices
Largest_Neg_Softness	Largest atomic softness among values for positively charged atoms	IND
lip_violation	Lipinski Violation Count	2D
nCb-	number of substituted benzene C(sp2)	functional group counts
nCconj	number of non-aromatic conjugated C(sp2)	functional group counts
nDB	number of double bonds	constitutional descriptors
nN	number of Nitrogen atoms	constitutional descriptors
nRCONHR	number of secondary amides (aliphatic)	functional group counts
PCR	ratio of multiple path count over path count	walk and path counts
PEOE_VSA-5	Total negative 5 vdw surface area	2D
RBF	rotatable bond fraction	constitutional descriptors
rsynth	Synthetic Feasibility	2D

Symbol	Definition	Class
SlogP_VSA6	Bin 6 SlogP (0.20, 0.25]	2D
SlogP_VSA7	Bin 7 SlogP (0.25, 0.30]	2D
std_dim3	Standard dimension 3	i3D
Sum_Pos_Hardness	Sum of hardnesses of atoms with positive partial charge	IND
Total_Pos_Softness	Sum of softnesses of atoms with positive partial charge	IND
vsurf_D8	Hydrophobic volume at -1.6	i3D
vsurf_DW12	vsurf_EWmin1, vsurf_EWmin2 distance	i3D
vsurf_EWmin2	2nd lowest Hydrophobic energy	i3D
vsurf _IW4	Hydrophilic integy moment at -2.0	i3D

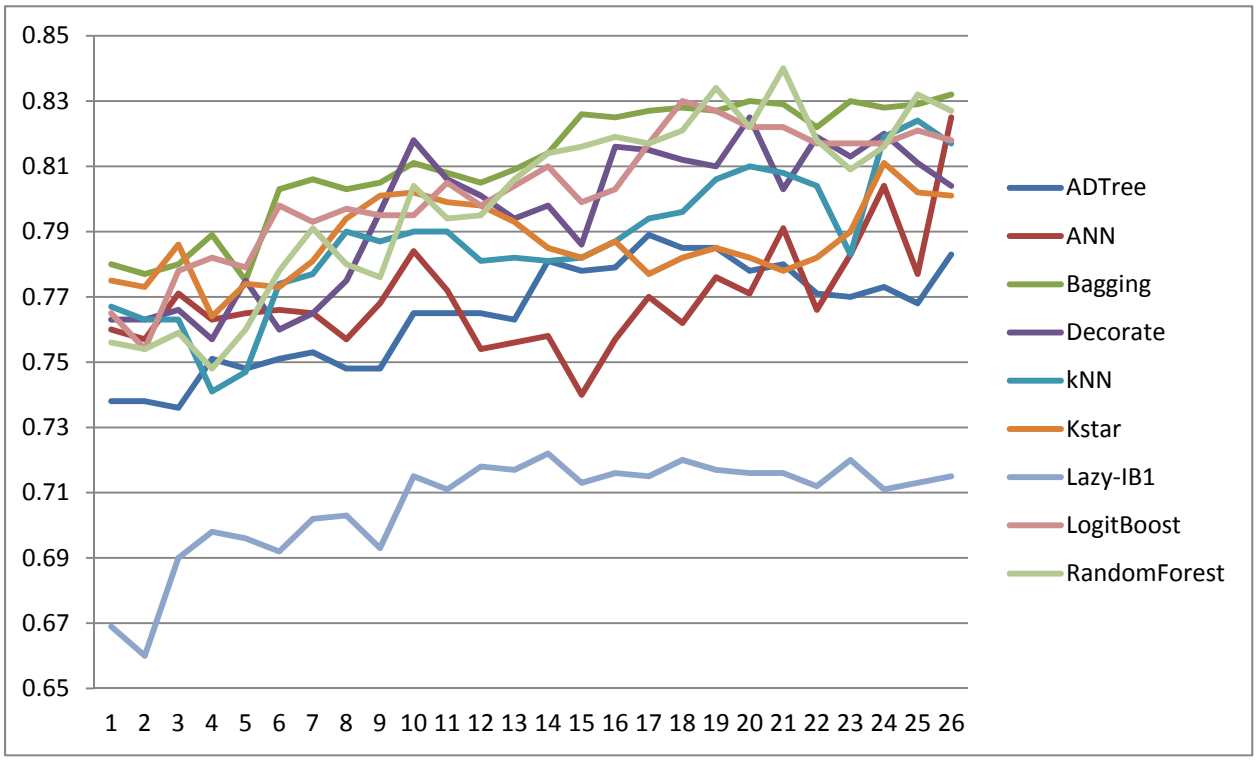


Figure 6 Correlation of AUC and number of descriptors of Training Set 1 and 2

Table 3 ROC AUC of different methods based on Combine Set 1

Desc. No.	ADTree	ANN	bagging	Decorate	kNN	Kstar	Lazy-IB1	Logitboost	RandomForest
15	0.705	0.731	0.787	0.785	0.788	0.783	0.711	0.777	0.784
16	0.702	0.712	0.781	0.787	0.787	0.779	0.718	0.782	0.799
17	0.71	0.73	0.797	0.789	0.793	0.785	0.717	0.778	0.802
18	0.695	0.745	0.8	0.792	0.8	0.794	0.722	0.79	0.81
19	0.699	0.746	0.804	0.799	0.798	0.79	0.713	0.776	0.809
20	0.703	0.746	0.808	0.777	0.801	0.793	0.716	0.781	0.806
21	0.7	0.767	0.806	0.776	0.804	0.796	0.715	0.787	0.802
22	0.705	0.759	0.807	0.784	0.806	0.798	0.72	0.794	0.804
23	0.716	0.747	0.814	0.802	0.802	0.805	0.717	0.791	0.814
24	0.722	0.769	0.813	0.812	0.806	0.807	0.716	0.804	0.812
25	0.715	0.737	0.811	0.808	0.804	0.809	0.716	0.807	0.822
26	0.717	0.736	0.807	0.813	0.812	0.807	0.712	0.795	0.816
27	0.71	0.754	0.823	0.793	0.805	0.807	0.72	0.807	0.824
28	0.712	0.767	0.822	0.796	0.801	0.808	0.711	0.795	0.817
29	0.721	0.75	0.818	0.8	0.803	0.809	0.713	0.796	0.813
30	0.724	0.752	0.818	0.805	0.803	0.809	0.715	0.802	0.815
best	0.724	0.767	0.823	0.813	0.812	0.809	0.722	0.807	0.824

Table 4 Descriptors selected basing on Combine Set 1 and their corresponding meanings

Symbol	**Definition**	**Class**
a_nN	Number of nitrogen atoms	2D
ARR	aromatic ratio	constitutional descriptors

Symbol	Definition	Class
Average_Softness	Arithmetic mean of softnesses of all atoms of a molecule	IND
b_double	Number of double bonds. Aromatic bonds are not considered to be double bonds.	2D
B02[N-O]	presence/absence of N - O at topological distance 02	2D binary fingerprints
B02[O-O]	presence/absence of O - O at topological distance 02	2D binary fingerprints
B09[C-O]	presence/absence of C - O at topological distance 09	2D binary fingerprints
BCUT_SLOGP_0	LogP BCUT (0/3)	2D
BCUT_SMR_0	Molar Refractivity BCUT (0/3)	2D
EEig09r	Eigenvalue 09 from edge adj. matrix weighted by resonance integrals	edge adjacency indices
ESpm08d	Spectral moment 08 from edge adj. matrix weighted by dipole moments	edge adjacency indices
F04[C-C]	frequency of C - C at topological distance 04	2D binary fingerprints
F05[O-F]	frequency of O - F at topological distance 05	2D binary fingerprints
GCUT_SLOGP_3	logP GCUT (3/3)	2D
Kier2	Second kappa shape index	2D
MSD	mean square distance index (Balaban)	topological descriptors
nCb-	number of substituted benzene C(sp2)	functional group counts
nCconj	number of non-aromatic conjugated C(sp2)	functional group counts
nDB	number of double bonds	constitutional descriptors
PCR	ratio of multiple path count over path count	walk and path counts
PEOE_VSA_PNEG	Total polar negative VDW surface area	2D
R7u	R autocorrelation of lag 7 / unweighted	GETAWAY descriptors
SlogP	Log Octanol/Water Partition Coefficient	2D
SMR	Molar Refractivity	2D
SRW05	self-returning walk count of order 05	walk and path counts

Symbol	Definition	Class
std_dim1	Standard dimension 1	i3D
Sum_Hardness	Sum of hardnesses of atoms of a molecule	IND
vsurf_A	Ampiphilic moment	i3D
vsurf_CW1	Capacity factor at -0.2	i3D
vsurf_IW6	Hydrophilic integy moment at -4.0	i3D

Table 5 Correlation of selected descriptors

Descriptors order in the table below, each line from left to right and each column from top to bottom, the name of the descriptors are: a_nN, ESpm08d, SRW05, ARR, nCconj, nDB, PCR, BCUT_SMR_0, B02[O-O], b_double, vsurf_IW6, nCb-, B02[N-O], F04[C-C], GCUT_SLOGP_3, vsurf_A, F05[O-F], MSD, Kier2, vsurf_CW1, std_dim1, BCUT_SLOGP_0, SMR, PEOE_VSA_PNEG, Sum_Hardness, R7s+, SlogP, P_VSA_LogP_8, Average_Softness, B09[C-O].

1	0.16	0.15	0.06	-0.03	0.21	0.16	-0.06	0.08	0.15	0.13	0.05	0.59	-0.07	-0.06	0.02	0.18	0.34	0.3	0.02	0.31	-0.3	0.23	0.28	0.15	0.15	-0.17	-0.12	0.01
0.16	1	0.24	-0.26	0.12	0.33	-0.16	-0.33	0.16	0.32	-0.03	0	0.16	0.5	0.36	-0.05	-0.03	0.55	0.52	-0.15	0.4	-0.36	0.56	0.41	0.56	-0.05	0.15	-0.22	0.35
0.15	0.24	1	-0.33	0.03	0.4	-0.32	-0.37	0.1	0.42	-0.06	-0.27	0.23	0.33	0.47	-0.08	-0.14	0.22	0.07	-0.15	0.2	-0.33	0.34	0.31	0.29	0.09	0	-0.09	0.47
0.06	-0.26	-0.33	1	-0.31	-0.52	0.9	0.74	-0.15	-0.53	-0.08	0.63	-0.19	-0.23	-0.42	-0.11	-0.09	-0.21	-0.17	0.06	-0.14	0.69	-0.17	-0.3	-0.3	-0.04	0.18	0.21	-0.49
-0.03	0.12	0.03	-0.31	1	0.61	-0.05	-0.19	-0.02	0.7	0.09	-0.1	0.08	0.28	0.21	0.09	0.23	0.12	0.1	-0.06	0.13	-0.22	0.18	0.09	0.19	-0.1	-0.01	-0.16	0.16
0.21	0.33	0.4	-0.52	0.61	1	-0.26	-0.35	0.41	0.93	-0.01	-0.2	0.43	0.3	0.32	-0.04	0.15	0.43	0.36	-0.14	0.39	-0.43	0.42	0.65	0.37	0.2	-0.14	-0.12	0.32
0.16	-0.16	-0.32	0.9	-0.05	-0.26	1	0.64	-0.11	-0.27	-0.02	0.76	-0.09	-0.05	-0.3	-0.08	0.03	-0.02	0.01	-0.05	0.05	0.57	0.04	-0.2	-0.14	-0.01	0.26	0.14	-0.38
-0.06	-0.33	-0.37	0.74	-0.19	-0.35	0.64	1	0.01	-0.39	-0.17	0.42	-0.18	-0.58	-0.78	-0.16	-0.06	-0.4	-0.29	0.34	-0.34	0.89	-0.47	-0.23	-0.61	0.08	-0.12	0.39	-0.74
0.08	0.16	0.1	-0.15	-0.02	0.41	-0.11	0.01	1	0.24	0.01	-0.07	0.12	-0.09	-0.08	0.02	0.02	0.26	0.31	-0.05	0.2	-0.06	0.13	0.63	0.11	0.16	-0.18	0.07	0
0.15	0.32	0.42	-0.53	0.7	0.93	-0.27	-0.39	0.24	1	0	-0.22	0.36	0.36	0.36	-0.03	0.14	0.4	0.33	-0.12	0.37	-0.43	0.4	0.5	0.38	0.14	-0.1	-0.17	0.35
0.13	-0.03	-0.06	-0.08	0.09	-0.01	-0.02	-0.17	0.01	0	1	-0.02	0.11	0.05	0.09	0.81	0.21	0.15	0.13	0.03	0.13	-0.15	0.03	-0.03	0.16	-0.09	0.03	-0.27	0.06
0.05	0	-0.27	0.63	-0.1	-0.2	0.76	0.42	-0.07	-0.22	-0.02	1	-0.04	0.12	-0.1	-0.06	0.1	0.17	0.17	-0.2	0.15	0.38	0.27	-0.05	0.02	0.11	0.43	0.33	-0.13
0.59	0.16	0.23	-0.19	0.08	0.43	-0.09	-0.18	0.12	0.36	0.11	-0.04	1	0	0.06	0.05	0.35	0.29	0.22	-0.04	0.29	-0.33	0.23	0.39	0.16	0.24	-0.15	-0.08	0.13
-0.07	0.5	0.33	-0.23	0.28	0.3	-0.05	-0.58	-0.09	0.36	0.05	0.12	0	1	0.89	0.02	-0.04	0.55	0.45	-0.53	0.49	-0.45	0.81	0.15	0.79	-0.18	0.56	-0.27	0.7
-0.06	0.36	0.47	-0.42	0.21	0.32	-0.3	-0.78	-0.08	0.36	0.09	-0.1	0.06	0.89	1	0.06	-0.01	0.5	0.36	-0.55	0.47	-0.59	0.74	0.17	0.75	-0.12	0.5	-0.24	0.84
0.02	-0.05	-0.08	-0.11	0.09	-0.04	-0.08	-0.16	0.02	-0.03	0.81	-0.06	0.05	0.02	0.06	1	0.23	0.07	0.07	0	0.05	-0.15	-0.02	-0.06	0.12	-0.13	0.01	-0.24	0.04
0.18	-0.03	-0.14	-0.09	0.23	0.15	0.03	-0.06	0.02	0.14	0.21	0.1	0.35	-0.04	-0.01	0.23	1	0.07	0.03	0.09	0.1	-0.09	-0.01	0.08	-0.01	0.22	0.01	-0.08	0.05
0.34	0.55	0.22	-0.21	0.12	0.43	-0.02	-0.4	0.26	0.4	0.15	0.17	0.29	0.55	0.5	0.07	0.07	1	0.94	-0.5	0.88	-0.42	0.87	0.53	0.83	0.09	0.37	-0.14	0.49
0.3	0.52	0.07	-0.17	0.1	0.36	0.01	-0.29	0.31	0.33	0.13	0.17	0.22	0.45	0.36	0.07	0.03	0.94	1	-0.52	0.8	-0.33	0.78	0.49	0.79	0.04	0.34	-0.12	0.37
0.02	-0.15	-0.15	0.06	-0.06	-0.14	-0.05	0.34	-0.05	-0.12	0.03	-0.2	-0.04	-0.53	-0.55	0	0.09	-0.5	-0.52	1	-0.48	0.22	-0.67	-0.02	-0.61	0.17	-0.66	-0.07	-0.45
0.31	0.4	0.2	-0.14	0.13	0.39	0.05	-0.34	0.2	0.37	0.13	0.15	0.29	0.49	0.47	0.05	0.1	0.88	0.8	-0.48	1	-0.34	0.75	0.42	0.76	0.07	0.37	-0.16	0.39
-0.3	-0.36	-0.33	0.69	-0.22	-0.43	0.57	0.89	-0.06	-0.43	-0.15	0.38	-0.33	-0.45	-0.59	-0.15	-0.09	-0.42	-0.33	0.22	-0.34	1	-0.43	-0.33	-0.54	0.02	0.1	0.41	-0.6
0.23	0.56	0.34	-0.17	0.18	0.42	0.04	-0.47	0.13	0.4	0.03	0.27	0.23	0.81	0.74	-0.02	-0.01	0.87	0.78	-0.67	0.75	-0.43	1	0.41	0.86	-0.03	0.59	-0.05	0.67
0.28	0.41	0.31	-0.3	0.09	0.65	-0.2	-0.23	0.63	0.5	-0.03	-0.05	0.39	0.15	0.17	-0.06	0.08	0.53	0.49	-0.02	0.42	-0.33	0.41	1	0.37	0.28	-0.26	-0.08	0.26
0.15	0.56	0.29	-0.3	0.19	0.37	-0.14	-0.61	0.11	0.38	0.16	0.02	0.16	0.79	0.75	0.12	-0.01	0.83	0.79	-0.61	0.76	-0.54	0.86	0.37	1	-0.13	0.43	-0.36	0.72
0.15	-0.05	0.09	-0.04	-0.1	0.2	-0.01	0.08	0.16	0.14	-0.09	0.11	0.24	-0.18	-0.12	-0.13	0.22	0.09	0.04	0.17	0.07	0.02	-0.03	0.28	-0.13	1	-0.12	0.19	0.02
-0.17	0.15	0	0.18	-0.01	-0.14	0.26	-0.12	-0.18	-0.1	0.03	0.43	-0.15	0.56	0.5	0.01	0.01	0.37	0.34	-0.66	0.37	0.1	0.59	-0.26	0.43	-0.12	1	0.26	0.38
-0.12	-0.22	-0.09	0.21	-0.16	-0.12	0.14	0.39	0.07	-0.17	-0.27	0.33	-0.08	-0.27	-0.24	-0.24	-0.08	-0.14	-0.12	-0.07	-0.16	0.41	-0.05	-0.08	-0.36	0.19	0.26	1	-0.13
0.01	0.35	0.47	-0.49	0.16	0.32	-0.38	-0.74	0	0.35	0.06	-0.13	0.13	0.7	0.84	0.04	0.05	0.49	0.37	-0.45	0.39	-0.6	0.67	0.26	0.72	0.02	0.38	-0.13	1
0.21	0.41	0.34	-0.37	0.21	0.5	-0.24	-0.5	0.28	0.46	0.14	-0.13	0.33	0.46	0.53	0.09	0.15	0.58	0.5	-0.23	0.56	-0.46	0.55	0.48	0.58	0.04	0.12	-0.2	0.5

3.1.2 Model Validation with External Testing Set

To validate the performance of the developed models we have used 89 chemical structures that have also been previously utilized in the reference paper (Li and Gramatica 2010) as unbiased external set (Table 6). The applicability domain (AD) of the models was validated and not listed here, for the reason that training set 1 and external testing set were both from reference paper (Li and Gramatica 2010). The combined set – based QSAR models were selected for further ZINC database screening, since for each modeling method, the combined set models are slightly better than training set1 models. These models were also performing great on external testing set, and ready for database screening.

Table 6 Validation on external testing set

	ADTREE	**ANN**	**BAGGING**	**Decorate**	**kNN**	**Kstar**	**IB1**	**Logitboost**	**Random Forest**
AUC (ROC)	0.78	0.862	0.817	0.79	0.707	0.707	0.714	0.768	0.732
True Positive Fraction	0.735	1	0.915	0.795	0.807	0.911	0.764	0.765	0.881
True Negative Fraction	0.714	0.793	0.714	0.651	0.714	0.857	0.857	0.714	0.857
Accuracy	0.776	0.809	0.809	0.738	0.708	0.719	0.685	0.764	0.742

3.2 Screening ZINC Database Using Developed QSAR Models

3.2.1 Virtual Screening and Consensus Voting

With the optimized descriptor number and highest ROC AUC values, nine selected models were applied to screen ZINC database. Thus, the ZINC database was processed with 9 pre-trained QSAR models utilizing different mathematical approaches including k-Nearest Neighbors (kNN) approach, Local Lazy method (lazy IB1), Alternating Decision Tree (ADTree), Artificial Neural Network (ANN), K-star method, Bagging method, as well as LogitBoost, Decorate, and Random

Forest. While all those are binary solutions, we have implemented a consensus voting protocol to evaluate the results. In particular, if a given ZINC entry is predicted as active by one model that would give a single vote to the entry. The final cumulative vote (with the maximum possible value of 9) was then used to rank the ZINC database chemicals. On the basis of the cumulative count, 50 most highly voted molecules were selected for future experimental evaluation (Figure 8). After checking the vendor information on these 50 chemicals, we found out that 39 chemicals were commercially available. Thus eventually 39 selected chemicals were tested in the wet lab.

From those 39 chemicals, 9 were found to be active and demonstrated significant ability to displace the fluorescently-tagged androgen ligand Fluormone (www.invitrogen.com) from the target HBS site.

Figure 7 Workflow of screening process

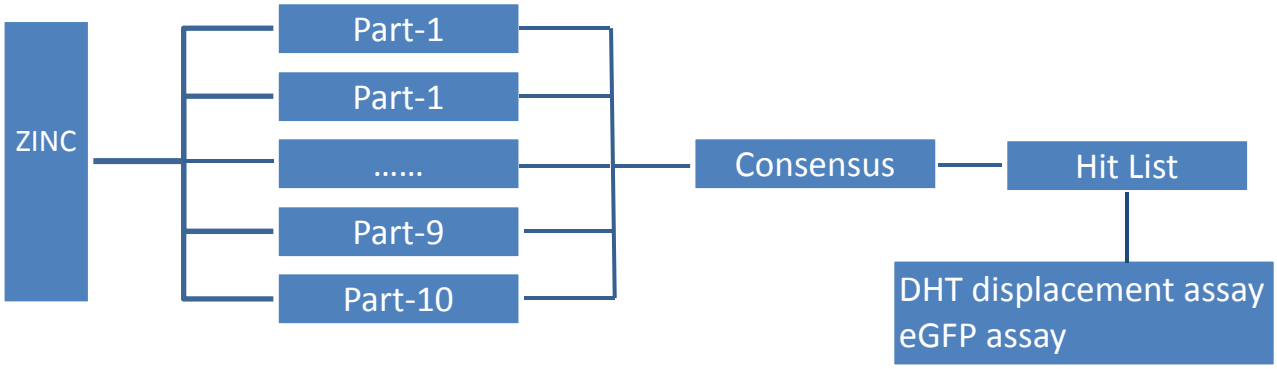


Table 7 Top 39 chemicals from screening

ZINC_ID	INTERNAL NO.	TAG
ZINC00171669	12001	MB00254
ZINC00488165	12002	9280584

ZINC_ID	INTERNAL NO.	TAG
ZINC00564515	12003	STK661827
ZINC04421403	12004	BAS 00121493
ZINC04488996	12005	BAS 12520097
ZINC04600179	12006	F2679-0167
ZINC12394346	12007	STK350797
ZINC12441512	12008	F5027-0039
ZINC58427578	12009	T6905782
ZINC00155064	12010	05038, LT154595
ZINC58282811	12011	T6821353
ZINC11036200	12012	T5921398
ZINC05921978	12014	ST51005340
ZINC00122320	12015	EN300-12312
ZINC04772847	12016	5340274
ZINC27154492	12017	Z396584806
ZINC04517670	12018	ST51043413
ZINC01593346	12051	A38800
ZINC00421637	12052	ASN02562486
ZINC00397781	12053	S696544|ALDRICH
ZINC01577019	12054	S749818|ALDRICH
ZINC01671287	12055	39311
ZINC01681528	12056	49676
ZINC01627374	12057	81910
ZINC01730845	12058	83329
ZINC01716374	12059	127682
ZINC01744320	12060	149906
ZINC01561336	12061	273821
ZINC01571863	12062	317906
ZINC05699716	12063	401483
ZINC01610217	12064	607743
ZINC01629402	12065	645009
ZINC01680651	12066	STK396447
ZINC18173676	12067	STK325933
ZINC13658815	12068	STK553158

ZINC_ID	INTERNAL NO.	TAG
ZINC00196649	12069	STK835521
ZINC00473218	12070	STK842542
ZINC03722619	12071	STK894533
ZINC00055479	12072	ST51005340

3.2.3 Applicability Domain of QSAR models.

The range of descriptors selected in each model were verified and compared to evaluate the applicability domain (AD) of the developed models. This procedure estimates the reliability of predictions for the top-voted 39 chemicals. Values of each descriptor for every top-voted chemical were checked to fit in the training descriptors value range. From this analysis we could verify that all of the selected 39 chemicals fell into suitable AD (Table 8), with the percentage of reliable predictions being 100%.

Table 8 Applicability domain

INTERNAL NO.	a_nN	ESpm08d	SRW05	ARR	nCconj	nDB	PCR	BCUT_SMR_0	B02[O-O]	b_double
12001	1	10.419	0	0.522	1	1	1.439	-2.1933949	0	1
12002	2	10.45	20	0.682	0	0	1.52	-1.9881955	0	0
12003	1	10.434	10	0.727	0	0	1.593	-2.0942636	0	0
12004	1	13.578	0	0.783	0	0	1.58	-2.0333951	0	0
12005	1	11.108	0	0.5	0	1	1.384	-2.2581482	0	1
12006	1	9.402	0	0.895	0	0	1.619	-1.8888065	0	0
12007	2	9.643	10	0.889	0	0	1.576	-1.8972696	0	0
12008	1	11.102	0	0.5	1	1	1.406	-2.4082153	0	1
12009	1	11.143	0	0.429	1	1	1.368	-2.4470885	0	1
12010	1	10.914	0	0.632	1	1	1.509	-2.1816049	0	1
12011	3	11.895	0	0.25	0	1	1.238	-2.489619	0	1
12012	3	12.396	0	0.462	1	1	1.47	-2.31288	0	1
12014	0	10.189	0	0.773	0	0	1.577	-2.2302117	0	0
12015	1	9.384	10	0.889	0	0	1.579	-1.8805335	0	0
12016	2	12.444	10	0.25	0	0	1.219	-2.5238316	0	0
12017	2	10.875	0	0.478	1	1	1.437	-2.5318456	0	1
12018	1	9.565	0	0.941	0	0	1.634	-1.8587706	0	0
12051	1	9.363	0	0.941	0	0	1.636	-1.8908862	0	0
12052	1	10.021	0	0.688	0	0	1.494	-2.0442266	0	0
12053	1	9.503	0	0.706	0	0	1.442	-2.0840466	0	0
12054	1	9.487	0	0.941	0	0	1.635	-1.8915498	0	0
12055	1	10.069	0	0.913	0	0	1.743	-2.1554279	0	0
12056	1	10.598	10	0.706	0	0	1.442	-1.945694	0	0
12057	1	10.247	0	0.913	0	0	1.693	-2.0265048	0	0

INTERNAL NO.	a_nN	ESpm08d	SRW05	ARR	nCconj	nDB	PCR	BCUT_SMR_0	B02[O-O]	b_double
12058	1	10.221	10	0.706	0	0	1.441	-1.9877139	0	0
12059	0	10.436	10	0.727	0	0	1.594	-2.0127611	1	0
12060	1	10.532	20	0.632	0	0	1.386	-2.2237437	0	0
12061	1	10.056	0	0.8	0	0	1.701	-2.0038741	0	0
12062	2	11.318	10	0.462	1	1	1.374	-2.507344	0	1
12063	0	10.042	0	0.647	0	0	1.458	-2.3942502	0	0
12064	1	9.657	10	0.882	0	0	1.614	-1.9878213	0	0
12065	0	11.168	0	0.571	1	1	1.435	-2.3836102	0	1
12066	1	9.517	0	0.941	0	0	1.635	-1.9599934	0	0
12067	1	9.954	20	0.778	0	0	1.607	-2.2086222	0	0
12068	0	11.714	0	0.6	0	0	1.444	-2.4055121	0	0
12069	0	10.373	10	0.762	0	0	1.513	-2.0670538	0	0
12070	0	9.99	0	0.688	0	0	1.54	-2.0473976	0	0
12071	2	10.206	10	0.647	0	0	1.441	-1.959733	0	0
12072	1	10.104	10	0.762	0	0	1.635	-2.0810037	0	0
[min, max]	[0,7.002478]	[4,7]	[0,1]	[7,32]	[2.43, 2.95]	[0.045, 7.56]	[0,4]	[0.20,0.32]	[3.75,8.91]	[2.55, 2.82]
AD range	[0, 10.17758]	[0, 15]	[0, 1]	[0, 96]	[1.53, 3.87]	[0.0062, 11.35]	[0, 6]	[0.14, 0.58]	[1.33, 28.87]	[2.23, 3.35]

INTERNAL NO.	vsurf_IW6	nCb-	B02[N-O]	F04[C-C]	GCUT_SLOGP_3	vsurf_A	F05[O-F]	MSD	Kier2	vsurf_CW1
12001	1.598155	6	0	20	2.661172	1.280883	0	0.231	7.050781	2.584572
12002	2.339491	6	0	16	2.666071	2.076542	0	0.213	6.011719	2.669958
12003	5.130027	4	0	19	2.720508	5.784112	0	0.233	6.405827	2.65524
12004	5.139665	5	0	26	2.722499	5.081868	0	0.214	6.011719	2.592908
12005	5.090312	6	1	23	2.771185	5.779234	0	0.222	7.266436	2.648235
12006	3.461679	4	0	14	2.507102	4.898636	0	0.27	5.325444	2.788729
12007	6.502045	4	0	15	2.532678	7.018967	0	0.262	4.704	2.77437
12008	2.394935	4	1	14	2.623593	2.091413	0	0.242	7.713499	2.629717
12009	2.090391	4	1	22	2.771605	1.453843	0	0.227	8.34714	2.56478
12010	3.892194	4	0	17	2.631717	4.488458	0	0.222	5.325444	2.685056
12011	5.258811	4	1	17	2.754733	4.612169	0	0.26	8.909091	2.696892
12012	3.606653	4	1	18	2.6804	3.225651	3	0.212	7.709141	2.710895
12014	0	4	0	24	2.801295	0.053771	0	0.222	5.080078	2.631952
12015	7.002478	4	0	12	2.509336	7.560006	0	0.273	4.704	2.808223
12016	1.934856	6	0	9	2.795915	2.896147	4	0.206	6.135866	2.738525
12017	3.708907	4	1	18	2.755433	4.197767	0	0.236	7.050781	2.556753
12018	4.208054	5	0	12	2.499531	4.984934	0	0.246	4.107639	2.796486
12051	6.439365	5	0	14	2.576181	6.723872	0	0.254	4.107639	2.823443
12052	1.316714	4	0	13	2.525144	2.662853	0	0.25	4.47259	2.737909
12053	6.022697	4	0	7	2.45886	6.149779	0	0.319	6.07438	2.759996
12054	5.10546	5	0	16	2.592113	5.852851	0	0.242	4.107639	2.811883
12055	0	4	0	29	2.774108	0.105894	0	0.216	5.652893	2.613047
12056	5.923103	5	0	13	2.616901	6.290957	0	0.236	3.7856	2.781679
12057	0	6	0	32	2.808237	0.044572	0	0.21	5.32526	2.623573
12058	6.416921	6	0	13	2.5908	6.402934	0	0.237	3.7856	2.80538
12059	4.661007	7	0	23	2.740037	4.160574	0	0.21	5.080078	2.768719
12060	0	4	0	17	2.739665	6.048038	0	0.234	3.75	2.694052

INTERNAL NO.	vsurf_IW6	nCb-	B02[N-O]	F04[C-C]	GCUT_SLOGP_3	vsurf_A	F05[O-F]	MSD	Kier2	vsurf_CW1
12061	6.855423	6	0	17	2.637182	4.545579	0	0.236	5.551021	2.741952
12062	5.491642	4	1	27	2.951426	6.23444	0	0.2	5.522683	2.628407
12063	3.923757	4	0	16	2.716727	4.96823	0	0.239	4.107639	2.742816
12064	4.671814	4	0	13	2.572924	4.825396	0	0.249	4.107639	2.820383
12065	4.634628	4	0	25	2.871686	5.586563	0	0.204	4.528616	2.682172
12066	5.138324	4	0	16	2.551488	5.395039	0	0.247	4.107639	2.736243
12067	4.910911	4	0	14	2.601564	5.410019	0	0.234	4.349113	2.771729
12068	3.801468	4	0	24	2.862475	4.935515	0	0.201	4.25	2.690836
12069	4.541239	4	0	16	2.615733	5.041495	0	0.248	6.185494	2.686328
12070	5.106693	5	0	12	2.538704	4.993172	0	0.251	4.88843	2.735717
12071	4.527382	4	0	8	2.427496	5.258585	0	0.247	5.104167	2.768139
12072	3.580351	4	0	10	2.555679	4.182735	0	0.269	6.185494	2.666455
[min, max]	[0,7.00]	[4,7]	[0,1]	[7,32]	[2.43, 2.95]	[0.044, 7.56]	[0,4]	[0.2,0.32]	[3.75,8.91]	[2.55, 2.83]
AD range	[0, 10.18]	[0, 15]	[0, 1]	[0, 96]	[1.53, 3.87]	[0.0062, 11.35]	[0, 6]	[0.14, 0.58]	[1.33, 28.87]	[2.23, 3.35]

INTERNAL NO.	std_dim1	BCUT_SLOGP_0	SMR	PEOE_VSA_PNEG	Sum_Hardness	R7u	SlogP	EEig09r	Average_Softness	B09[C-O]
12001	3.064233	-2.46354	8.5208	10.69009	105.2599	1.163	3.75067	1.444	4.843047	1
12002	3.003884	-2.25389	7.33015	19.15416	101.1964	1.216	2.85412	1.693	4.670294	0
12003	2.727879	-2.27818	7.85098	19.67618	105.8795	0.984	4.16222	1.583	4.78199	1
12004	2.458977	-2.28183	8.47717	0.136891	103.7781	1.19	4.839	2	4.608077	0
12005	2.855061	-2.50557	8.60877	18.71133	103.8857	1.193	3.8313	1.348	5.058686	1
12006	3.461143	-2.03936	6.88438	13.45012	91.88747	0.419	3.6074	0.839	4.200919	1
12007	3.091494	-2.32741	6.76874	6.651119	91.94083	0.602	3.2127	0.751	4.254678	0
12008	3.725609	-2.75171	8.25595	27.97931	104.3044	1.038	3.7883	1.972	4.955214	0
12009	3.812982	-2.77883	9.31765	21.07819	123.9041	1.14	3.1555	2.098	5.251559	1
12010	2.463798	-2.39411	6.85598	11.16614	92.53198	0.804	3.01174	1	4.567438	0
12011	3.818302	-2.80121	8.3938	37.38419	119.5301	1.154	2.921284	1.591	5.223744	1
12012	2.403246	-2.66156	8.11652	13.70381	97.39276	1.111	3.782584	1.914	5.109293	0
12014	3.245081	-2.43437	8.3813	0	100.2329	0.755	5.25636	1.833	4.766997	0
12015	3.622174	-2.32512	6.60624	9.154876	90.68082	0.537	3.682	0.786	4.105644	0
12016	2.936185	-2.76849	6.40636	46.14956	89.31811	0.891	3.1834	1.431	5.386538	0
12017	3.325085	-2.79019	8.56715	19.2495	108.9394	0.874	4.1504	1.715	5.16673	0
12018	2.823963	-2.01245	6.09138	13.45012	79.99416	0.477	3.0936	0.551	4.147219	0
12051	3.058632	-2.32297	6.58664	6.651119	86.89626	0.323	3.5752	0.598	4.150276	0
12052	2.825107	-2.1935	6.23548	13.45012	85.04195	0.354	3.26374	0.288	4.401497	0
12053	3.494638	-2.32346	6.70905	7.904431	90.69386	0.608	3.9241	0.518	4.214244	1
12054	2.724902	-2.32277	6.58664	6.651119	85.80129	0.566	3.5752	0.575	4.210214	0
12055	3.073021	-2.38262	8.3992	0	107.3506	1.013	4.29407	1.884	4.72468	0
12056	2.889313	-2.32796	6.71714	6.651119	78.63081	0.226	3.829	0.318	4.394992	0
12057	3.014421	-2.33422	8.4088	0	106.2459	1.197	4.34854	1.827	4.765054	0
12058	2.887466	-2.33152	6.41064	6.651119	78.89175	0.225	3.49337	0.414	4.363979	0
12059	2.543648	-2.29839	7.37048	12.77505	92.14932	0.862	3.4804	1.456	4.647207	0
12060	2.468177	-2.56705	6.58417	0.136891	88.48209	0.518	3.16867	1.116	4.68446	0

INTERNAL NO.	std_dim1	BCUT_SLOGP_0	SMR	PEOE_VSA_PNEG	Sum_Hardness	R7u	SlogP	EEig09r	Average_Softness	B09[C-O]
12061	3.145925	-2.27519	7.2353	10.69009	96.02877	1.001	3.4052	1.131	4.468616	0
12062	2.274837	-2.81283	8.59775	13.56692	103.9817	1.199	3.46157	1.981	5.495072	0
12063	2.622422	-2.52489	6.19568	7.767541	88.43047	0.58	3.30497	0.573	4.647796	0
12064	2.991527	-2.16133	6.19325	7.904431	85.61645	0.514	3.0798	0.505	4.243473	0
12065	2.128557	-2.51618	7.10325	13.56692	90.73332	0.673	3.8922	1.616	5.038194	0
12066	2.969321	-2.11791	6.3986	5.682576	85.96145	0.393	3.69642	0.577	4.229219	0
12067	2.347416	-2.59266	6.94308	7.767541	79.51171	0.98	2.97062	0.703	4.88236	0
12068	2.287024	-2.5298	7.12536	15.53508	98.45079	0.918	3.4052	0.727	4.941827	0
12069	3.138932	-2.24878	7.33808	12.77505	101.0716	0.714	3.6186	1.494	4.517534	1
12070	2.589224	-2.26586	6.43278	10.2713	80.7755	0.576	3.3696	0.47	4.403164	0
12071	2.582908	-2.29618	6.85344	14.83745	69.35173	0.705	3.7077	0.445	4.569382	0
12072	3.827999	-2.47344	7.59258	10.2713	100.9764	0.523	3.636	1.426	4.56817	1
[min, max]	[2.13,3.83]	[-2.81,-2.01]	[6.09, 9.32]	[0, 46.15]	[69.35, 123.91]	[0.22, 1.22]	[2.85, 5.26]	[0.28, 2.10]	[4.10, 5.50]	[0, 1]
AD range	[1.26, 7.35]	[-3.66, -1.79]	[1.28, 27.65]	[0, 132.90]	[0, 295.01]	[0, 2.11]	[-7.57, 11.82]	[-1.1, 3.65]	[0, 7.04]	[0, 1]

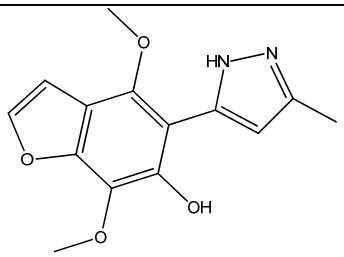
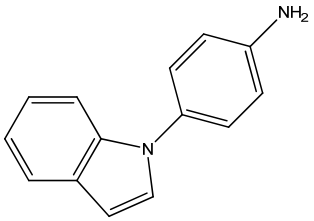
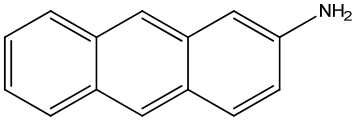
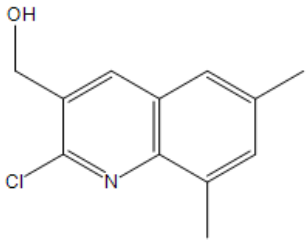
3.3 Cell-based Testing and *in vitro* Biochemical Characterization

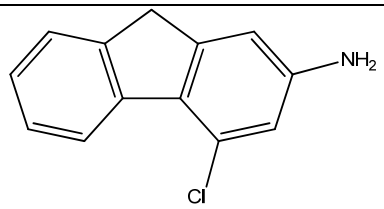
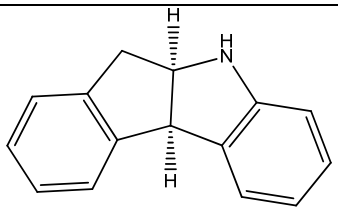
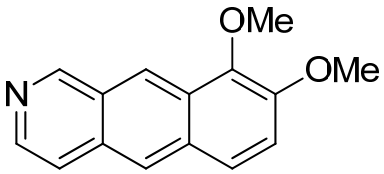
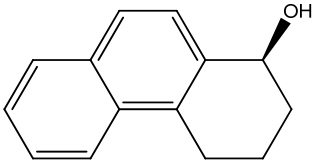
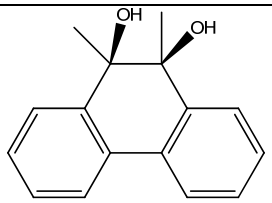
All 39 molecules selected from ZINC database were purchased from their respective vendors. In order to demonstrate whether these selected chemicals interfere with the hormone binding site of the AR they were tested using an androgen displacement assay. After testing, nine chemicals presenting an IC_{50} of DHT displacement below 20 μM have been selected and listed in Table 9.

Next these nine best chemicals went through an additional screening for their ability to inhibit AR transcriptional activity using a non-destructive, cell-based eGFP screening assay. In this assay, the expression of eGFP is under the control of an androgen responsive probasin-derived promoter and thus can correlate with the level of AR activity (Figure 9). Results showed that all nine chemicals selected from the DHT displacement assay exhibited excellent IC_{50} values (Table 9).

After this screen, all chemicals were evaluated for their general cytotoxicity using the MTS assay. Of note compounds VPC-12063 and VPC-12068 were not subjected to MTS assay due to relatively low displacement ability and high GFP IC_{50} values. Five of the remaining seven chemicals demonstrated detectible effect on cell viability and growth when administered at a concentration-dependent way for over 72 h. Compounds VPC-12007 and VPC-12060 decreased cell viability in the MTS assay better compared with other chemicals, identified as the most promising chemicals from all nine hit compounds, as shown in Table 10. In all experimental tests we have used 25μM concentration of Casodex and 10μM concentration of MDV3100 as the controls.

Table 9 Structures and activity profiles of the AR-HBS binders identified from our *in silico* screening

VPC-ID	**Structure**	**eGFP IC_{50}**	**DHT Displacement**
12002		2.014 µM	12.29 µM
12007		2.160 µM	1.84 µM
12051		2.804 µM	1.23 µM
12052		2.838 µM	3.25 µM

VPC-ID	Structure	eGFP IC_{50}	DHT Displacement
12058		4.79 μM	2.024 μM
12060		1.04 μM	3.426 μM
12061		3.26 μM	3.493 μM
12063		9.35 μM	18.63 μM
12068		16.18 μM	5.163 μM

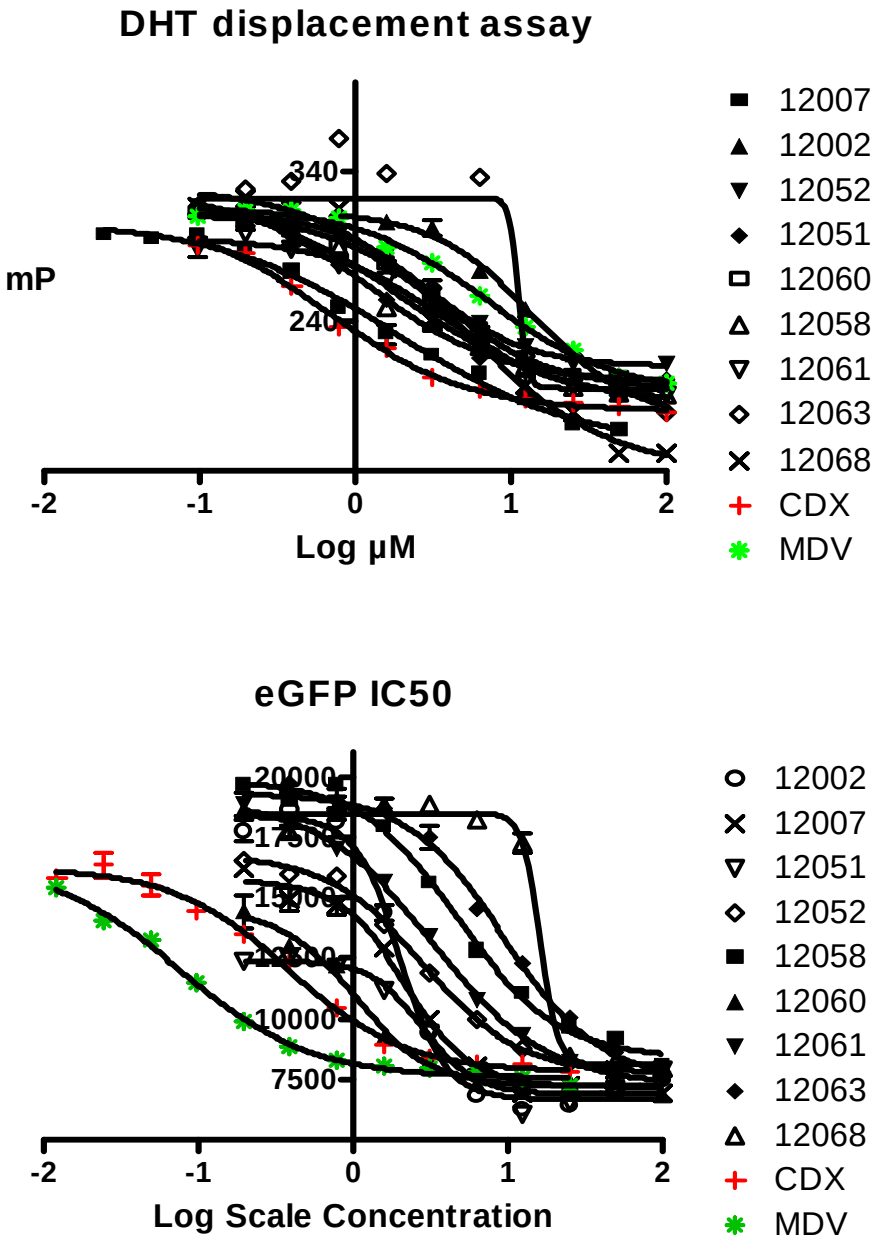
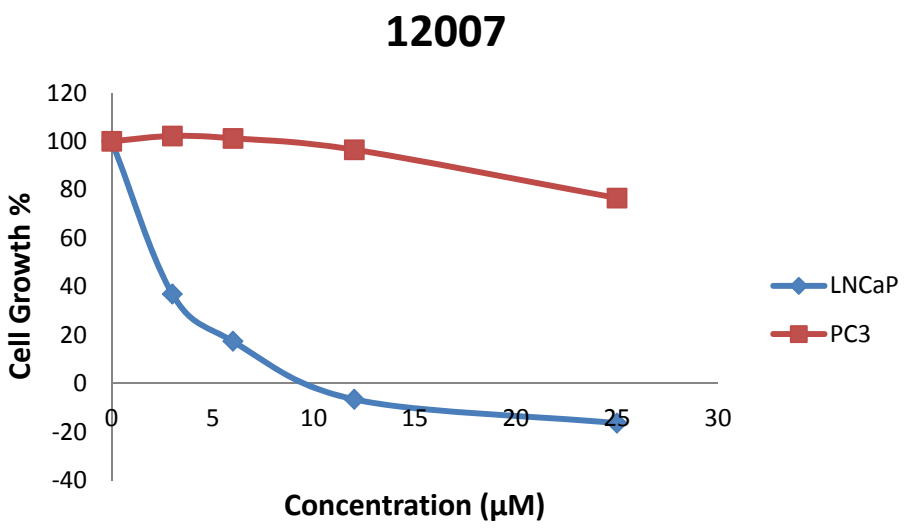
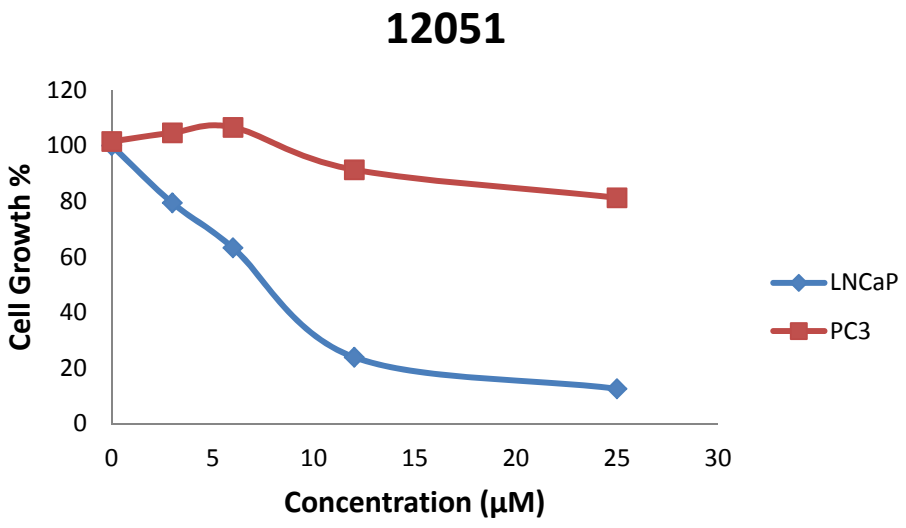
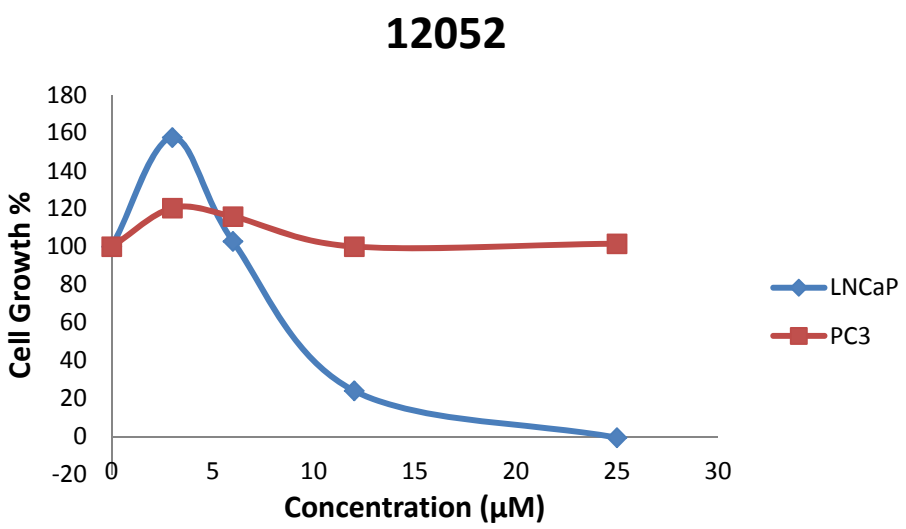
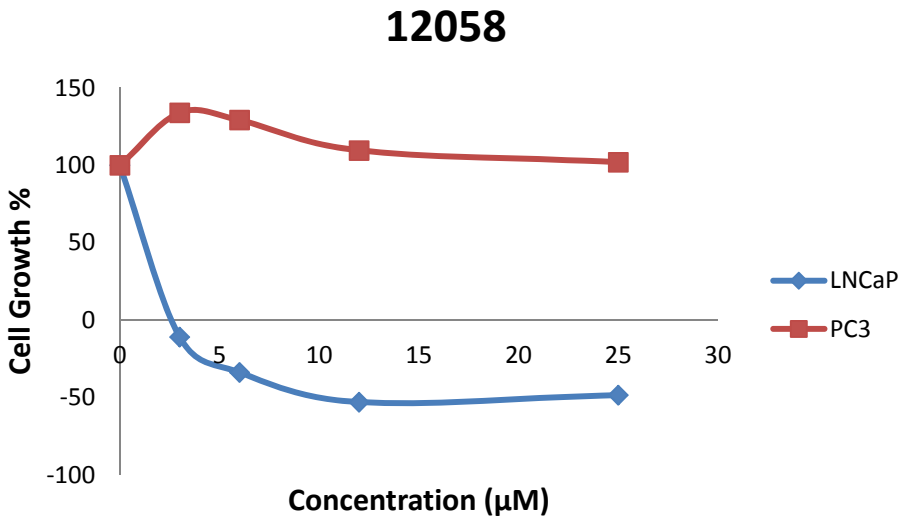
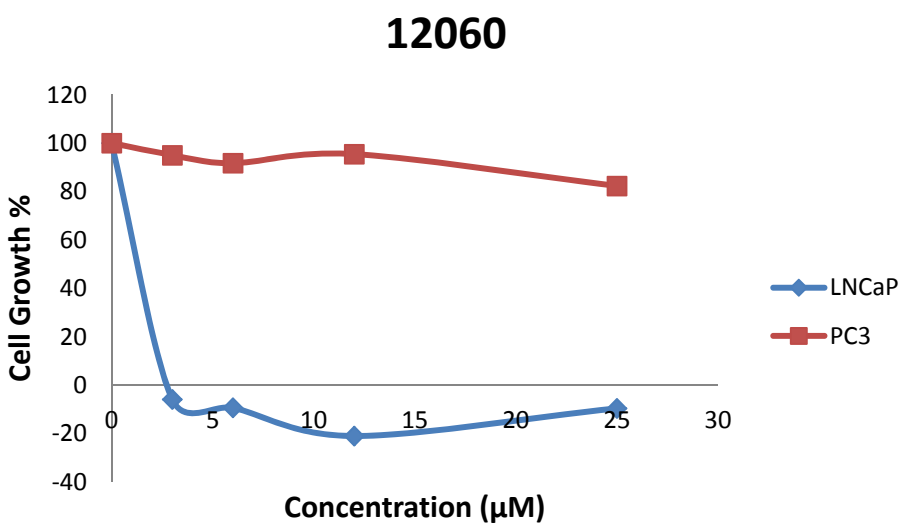
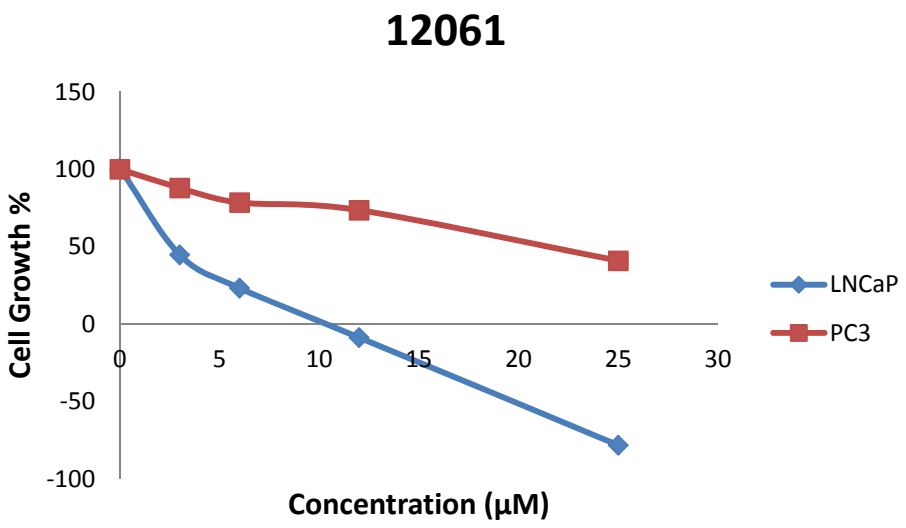


Figure 8 DHT displacement and eGFP assay testing of the identified AR antagonists

Table 10 MTS testing results of best nine chemicals from screening

Compound	MTS
12007	
12051	

Compound	MTS
12052	
12058	

Compound	MTS
12060	
12061	

4. Discussions

4.1 Descriptors

4.1.1 Optimized Number of Descriptors of Combined T-1 and T-2

For the most of the modeling approaches we have utilized, it could be concluded that the optimal number of required QSAR descriptors in in 25-30 range (judged by the corresponding ROC AUC values). It has also been established that ROC AUC values initially increase with inclusion of a larger number of descriptors in the models. Then, after reaching a certain number of utilized descriptors ROC AUC magnitude achieves saturation. Typically, that 'saturation' threshold would represent an optimal number of descriptors required to model a particular dataset.

Our results indicate that Lazy IB1 method reaches the highest ROC AUC magnitude with the use of only 18 QSAR descriptors. Thus, in view of large scale descriptor computation, it is the most optimal method. But the difference between the highest and lowest ROC AUC values in this method is only 1.55%, which means it is relatively stable and less affected by descriptor number.

Decorate and kNN methods both reach their saturation point at descriptor number 26, these two methods are excellent according to the evaluation criteria above except Lazy IB1 method.

Bagging, Logitboost and Random Forest method reach highest ROC AUC at the descriptor number of 27. ANN, Kstar and ADTree reach saturation point with descriptor number 28, 29 and 30 respectively. Comparing difference of highest and lowest ROC AUC of these methods, the values vary from 3% to 8%.

By checking the deviation of the optimized ROC AUC and the worst ROC AUC, by selecting the best descriptor number, the ROC AUC could be increased from 1.5% to 8.0%, which can produce better prediction accuracy.

Except for Lazy IB1 method, all the rest methods are generating optimized models with descriptor number in the range of 25-30. This range is relatively suitable for large scale computation and screening, which lead to the success of next steps of this project.

4.1.2 Contribution of Descriptors

To access the contribution of each descriptor to a given QSAR model, a ranker of principal component analysis was applied (see Table 11). The approach ranks attributes by their individual evaluations, and is used in conjunction with attribute principal component analysis. The principal component analysis (PCA) (Warmuth and Kuzmin 2008) is a mathematical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. A descriptor ranker is a mapping from instances to rankings (total orders) over biological activity. Thus, given any instance as an input, a label ranker produces a prediction in the form of a ranking of the complete set of activities as an output (Hullermeier, Furnkranz et al. 2008).

Table 11 PCA analysis - Eigenvectors

DESCRIPTORS	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**
a_nN	0.071	0.0128	0.319	0.1461	0.1883	0.4734	0.0031	0.1837
ESpm08d	0.1778	-0.0143	0.0789	-0.1147	0.1079	-0.0665	0.1479	0.4489
SRW05	0.1422	0.1341	-0.0455	-0.1942	0.0491	0.2201	-0.6094	0.0162
ARR	-0.1558	-0.3565	0.1394	-0.0115	-0.0141	0.095	-0.209	0.0861
nCconj	0.095	0.1209	0.0515	0.0313	-0.6285	-0.1162	0.1324	0.1948
nDB	0.1867	0.2099	0.2345	-0.1275	-0.3055	-0.0846	-0.0942	-0.0285
PCR	-0.0913	-0.3763	0.2128	0.0207	-0.196	0.0966	-0.163	0.1162
BCUT_SMR_0	-0.2243	-0.1809	0.2405	-0.0907	-0.0855	-0.0881	-0.1135	0.0336
B02[O-O]	0.0621	0.0966	0.2878	-0.1041	0.1798	-0.4331	-0.2074	-0.3836
b_double	0.1851	0.2049	0.1651	-0.1127	-0.3918	-0.0496	-0.0655	0.0918
vsurf_IW6	0.0404	0.0161	0.0038	0.5969	0.0114	-0.1748	-0.2766	0.156
nCb-	-0.0238	-0.3895	0.1594	0.0151	-0.1434	0.0517	-0.0624	-0.0699
B02[N-O]	0.1022	0.1142	0.2922	0.1334	0.0326	0.4443	-0.0039	-0.1784

DESCRIPTORS	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**
F04[C-C]	0.239	-0.1407	-0.1975	-0.0509	-0.1523	0.0138	-0.0839	0.1378
GCUT_SLOGP_3	0.2499	-0.0362	-0.2678	-0.0141	-0.066	0.0819	-0.1334	-0.062
vsurf_A	0.0253	0.0376	-0.0333	0.5928	0.0013	-0.2371	-0.2428	0.0737
F05[O-F]	0.0258	0.0445	0.1538	0.3225	-0.2546	0.1945	0.2843	-0.482
MSD	0.252	-0.1384	0.1734	0.0317	0.1375	-0.0994	0.1521	0.0776
Kier2	0.2226	-0.1542	0.1914	0.0296	0.1663	-0.2058	0.2497	0.0736
vsurf_CW1	-0.1718	0.2264	0.1274	0.0433	0.0023	0.0602	-0.0509	0.3206
std_dim1	0.2248	-0.1462	0.1705	0.0472	0.0889	-0.0568	0.1471	0.0282
BCUT_SLOGP_0	-0.2103	-0.2183	0.0947	-0.1058	-0.1209	-0.1696	-0.1732	-0.086
SMR	0.2699	-0.196	0.0269	-0.061	0.0036	0.0119	-0.0135	0.0214
PEOE_VSA_PNEG	0.1584	0.1317	0.3284	-0.1416	0.1462	-0.1614	-0.1382	-0.1035
Sum_Hardness	0.2735	-0.1143	-0.0447	0.0331	0.0771	-0.085	0.0806	0.1182
R7u	0.2241	-0.0432	-0.1601	0.0403	0.0349	0.0135	-0.085	-0.0752
SlogP	0.1089	-0.3533	-0.1939	0.0166	-0.1148	-0.0233	0.0268	-0.2199
EEig09r	0.2626	-0.1332	0.0404	-0.0143	-0.063	0.1463	-0.1269	-0.0143
Average_Softness	0.2376	0.018	-0.2185	-0.0181	0.0114	0.1026	-0.0985	-0.139
B09[C-O]	0.2171	0.0629	0.0967	0.0346	0.0376	-0.026	-0.075	-0.0737
Ranked attributes:	**0.6395**	**0.5037**	**0.4031**	**0.333**	**0.2775**	**0.231**	**0.1982**	**0.1682**

DESCRIPTORS	V9	V10	V11	V12	V13	V14	V15	V16
a_nN	-0.234	-0.1808	-0.0066	-0.0788	-0.2992	-0.3105	-0.322	0.0254
ESpm08d	0.4617	0.0679	0.5004	-0.2778	-0.0071	-0.2008	0.1023	-0.02
SRW05	-0.0232	0.3873	0.1336	0.1352	-0.2871	-0.0656	0.1295	-0.2514
ARR	0.0351	-0.1283	-0.1534	-0.0966	-0.085	-0.0127	0.2769	0.1485
nCconj	-0.1437	-0.095	-0.0504	-0.0742	-0.1029	-0.1301	-0.082	0.0558
nDB	-0.1233	-0.016	0.0805	0.0239	0.0902	0.0177	-0.0191	0.0079
PCR	-0.0001	-0.1847	-0.1571	0.0344	-0.0361	-0.1227	0.1947	0.0974
BCUT_SMR_0	-0.014	0.1462	0.0759	-0.1734	-0.1315	0.1908	-0.129	-0.0329
B02[O-O]	0.0401	-0.2448	0.0852	-0.0832	-0.1113	-0.3983	0.0189	0.3194
b_double	-0.1416	0.0749	0.0419	0.0569	0.0312	0.056	-0.0286	-0.0278
vsurf_IW6	-0.035	0.0332	0.033	0.0116	0.0366	0.0411	-0.1598	0.1413
nCb-	0.2027	-0.2057	0.1298	0.3567	0.3454	-0.0143	-0.2092	-0.3617
B02[N-O]	-0.1051	0.0695	0.2709	-0.2405	0.4639	0.2786	0.1448	0.2949
F04[C-C]	0.1847	-0.1532	-0.0062	-0.0701	0.0132	0.0512	0.2191	0.1686
GCUT_SLOGP_3	0.0964	-0.0569	-0.1089	0.0255	0.0294	0.0204	0.1212	0.1602
vsurf_A	-0.0146	0.0363	0.1563	0.0399	0.0656	-0.0169	0.1427	-0.1577
F05[O-F]	0.4256	0.104	0.0447	0.0245	-0.4297	0.0044	0.1855	-0.1221
MSD	-0.063	0.167	-0.0487	0.147	-0.0825	0.1003	-0.0784	-0.0433
Kier2	-0.1426	0.0945	0.0207	0.0218	-0.1136	0.1291	-0.1098	-0.1679
vsurf_CW1	0.4576	0.0244	-0.2568	0.1011	-0.0781	0.2214	-0.2689	0.2713
std_dim1	-0.1502	0.3217	-0.2829	0.153	-0.0535	0.0638	0.2463	0.1953
BCUT_SLOGP_0	0.0416	0.2924	0.0275	-0.2942	-0.1259	0.302	-0.1792	0.0895
SMR	-0.0061	-0.0147	0.094	0.082	0.022	-0.0025	-0.0706	0.0076
PEOE_VSA_PNEG	0.1963	-0.2207	-0.0131	0.1887	0.0473	0.3231	0.0038	-0.0545
Sum_Hardness	-0.0258	-0.0211	-0.0614	0.0115	-0.1109	0.1481	0.1698	0.1569
R7u	-0.1088	-0.4365	-0.0047	-0.4555	-0.2512	0.3839	-0.086	-0.2644
SlogP	-0.0233	0.2663	0.2134	-0.1257	0.0528	-0.1665	-0.3356	0.2255
EEig09r	0.1014	0.0135	-0.0851	0.052	0.0011	-0.0609	-0.0612	-0.0679
Average_Softness	0.1519	-0.0564	-0.0258	0.1959	-0.0758	0.0744	-0.4104	0.2922
B09[C-O]	0.227	0.1958	-0.5606	-0.4466	0.3362	-0.2525	-0.1123	-0.2781
Ranked attributes:	**0.1408**	**0.1197**	**0.1015**	**0.0862**	**0.0733**	**0.061**	**0.0518**	**0.0434**

Table 12 Descriptors ranking according to the best eigenvector (EV%=63.95%)

Descriptor	**Contributioin_PCA**	**Categary**
Sum_Hardness	0.2735	IND
SMR	0.2699	2D
EEig09r	0.2626	edge adjacency indices
MSD	0.252	topological descriptors
GCUT_SLOGP_3	0.2499	2D
F04[C-C]	0.239	2D binary fingerprints
Average_Softness	0.2376	IND
std_dim1	0.2248	i3D
R7u	0.2241	GETAWAY descriptors
Kier2	0.2226	2D
B09[C-O]	0.2171	2D binary fingerprints
nDB	0.1867	constitutional descriptors
b_double	0.1851	2D
ESpm08d	0.1778	edge adjacency indices
PEOE_VSA_PNEG	0.1584	2D
SRW05	0.1422	walk and path counts
SlogP	0.1089	2D
B02[N-O]	0.1022	2D binary fingerprints
nCconj	0.095	functional group counts
a_nN	0.071	2D
B02[O-O]	0.0621	2D binary fingerprints
vsurf_IW6	0.0404	i3D
F05[O-F]	0.0258	2D binary fingerprints
vsurf_A	0.0253	i3D
nCb-	-0.0238	functional group counts
PCR	-0.0913	walk and path counts

Descriptor	Contributioin_PCA	Categary
ARR	-0.1558	constitutional descriptors
vsurf_CW1	-0.1718	i3D
BCUT_SLOGP_0	-0.2103	2D
BCUT_SMR_0	-0.2243	2D

In table 12, the Sum_Hardness, SMR and EEig09r descriptors are the most important ones in the training set, with contribution coefficient over 0.26.

INDUCTIVE category: They quantify the inductive effect of the electronegative atoms in a molecule. The inductive effect is the effect of the transmission of charge through a chain of atoms in a molecule by electrostatic induction. Inductive QSAR descriptors derive from free energy equations for inductive and steric substituents, and have the advantage of capturing the electronic properties of different chemicals. Sum_Hardness is Sum of hardnesses of atoms of a molecule, which is a physical measure of how resistant solid matter is to various kinds of permanent shape change when a force is applied. Average_Softness is also an INDUCTIVE descriptor, meaning arithmetic mean of softness of all atoms of a molecule. There are only two INDUCTIVE descriptors selected for the training set, but both of them are contributing greatly to the accuracy of all the models, ranking 1st and 7th respectively in table 12.

Topological descriptors category MSD is the only descriptor of topological descriptors. It is mean square distance index (Balaban), is one of the topological descriptors. The mean square distance index is calculated from the second order distance distribution moment.

Edge adjacency indices category: EEig09r has a high contribution from edge adjacency indices weighted by resonance integrals, illustrating the importance of hydrophobicity. ESpm08d is also an edge adjacency indices descriptor, meaning Spectral moment 08 from edge adj. matrix weighted by dipole moments.

2D category: SMR is contributing the most in 2D category. It is Molar Refractivity, a measure of the total polarizability of a mole of a substance and is dependent on the temperature, the index

of refraction, and the pressure. GCUT_SLOGP_3 is logP GCUT (3/3), Kier2 is Second kappa shape index, b_double is Number of double bonds. Aromatic bonds are not considered to be double bonds. PEOE_VSA_PNEG is Total polar negative VDW surface area. SlogP is Log Octanol/Water Partition Coefficient. a_nN is the number of nitrogen atoms. BCUT_SLOGP_0 is LogP BCUT (0/3). BCUT_SMR_0 is Molar Refractivity BCUT (0/3).

2D binary fingerprints category: F04[C-C] is frequency of C - C at topological distance 04. B09[C-O] is presence/absence of C - O at topological distance 09. B02[N-O] is presence/absence of N - O at topological distance 02. B02[O-O] is presence/absence of O - O at topological distance 02. F05[O-F] is frequency of O - F at topological distance 05.

In the contribution table 12, 2D descriptors rank from top to bottom and scatter randomly, indicating these descriptors are stable and most important category for the training set.

Constitutional descriptors category: nDB is number of double bonds. ARR is aromatic ratio.

Functional group counts category: nCconj is number of non-aromatic conjugated C(sp2). nCb- is number of substituted benzene C(sp2).

GETAWAY descriptors category: R7u is R autocorrelation of lag 7 / unweighted.

i3D category: std_dim1 is Standard dimension 1. vsurf_IW6 is Hydrophilic integy moment at -4.0. vsurf_A is Ampiphilic moment. vsurf_CW1 is Capacity factor at -0.2.

Walk and path counts category: SRW05 ranked the first in association of chemical activities, with the meaning of self-returning walk count. This descriptor can discriminate among isomers and structure situations of atoms. PCR is ratio of multiple path count over path count.

Constitutional, functional group counts, GETAWAY, i3D, walk and path counts descriptors are similar according to model contribution, berried in the middle of contribution table 12.

4.1.3 Descriptor Categories

As shown in Figure 10, 2D descriptors constitute 47% of the whole selected descriptor scale, including 30% 2D descriptors and 17% 2D binary fingerprints. There are in total 14 of 2D descriptors, leading to a relatively faster calculation, for the reason that 2D descriptors need less computation resources than other categories.

i3D category is the third largest of all. 3D molecular descriptors are classified as "i3D" for internal coordinate dependent 3D and "x3D" for external coordinate dependent. The energy descriptors use the MOE potential energy model to calculate energetic quantities from stored 3D conformations. Most of the energy descriptors belong to the i3D class; that is, they depend on internal coordinates alone and not on an external reference frame.

INDUCTIVE, cnstitutional descriptors, edge adjacency indices, functional group counts, walk and path counts are quite similar in quantity, ranging from 6% to 7%.

GETAWAY descriptors and topological descriptors are the least ones in term of percentage. GETAWAY descriptors calculate the leverage matrix obtained by the centered atomic coordinates (molecular influence matrix, MIM), and topological descriptors are calculated based on the molecular graph of a chemical.

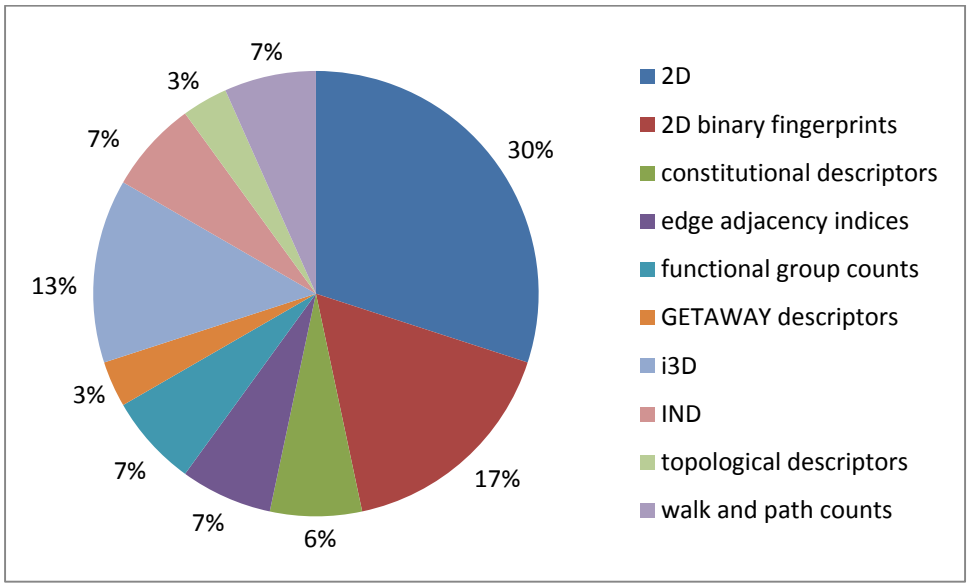


Figure 9 Percentage of each descriptor category

4.2 Modeling Methods

ADTree, ANN, Bagging, Decorate, kNN, Kstar, Lazy-IB1, Logitboost, Random Forest methods were applied here for the virtual screening of ZINC database.

The training results illustrate that Random Forest, bagging and Logitboost are somewhat better than other methods, resulting in higher ROC AUC values, what characterizes them as good binary classifiers. According to the data from Table 13, all classifiers give positive votes for all seven AR antagonists, except for IB1 which votes negative to chemical 12002. Supporting this, IB1 has the lowest internal training ROC AUC as 0.726 (Table 2), which makes it the worst modeling method to generate classifiers.

Table 13 Consensus votes from all nine models

Chemical	**12002**	**12007**	**12051**	**12052**	**12058**	**12060**	**12061**	**12063**	**12065**
kNN	1	1	1	1	1	1	1	1	1
Lazy-IB1	0	1	1	1	1	1	1	1	1

Chemical	**12002**	**12007**	**12051**	**12052**	**12058**	**12060**	**12061**	**12063**	**12065**
ADTree	1	1	1	1	1	1	1	1	1
ANN	1	1	1	1	1	1	1	1	1
Bagging	1	1	1	1	1	1	1	1	1
Decorate	1	1	1	1	1	1	1	1	1
Logitboost	1	1	1	1	1	1	1	1	1
Random forest	1	1	1	1	1	1	1	1	1
Kstar	1	1	1	1	1	1	1	1	1
Consensus votes	**8**	**9**	**9**	**9**	**9**	**9**	**9**	**9**	**9**

4.2.1 Random Forest

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and which outputs the class that is the mode of the classes output by individual trees. It is one of the most accurate learning algorithms currently available. As it is illustrated by Figure 7, this method produces the most highly accurate classifier in this project, and yields ROC AUC value of 0.84 for the internal training.

In our study, we have set the number of trees to 10, and seed was set to 1. Each tree is constructed using the following algorithm: 1. Let the number of training cases be N, and the number of variables in the classifier be M. 2. We are told the number m of input variables to be used to determine the decision at a node of the tree; m should be much less than M. 3. Choose a training set for this tree by choosing n times with replacement from all N available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes. 4. For each node of the tree, randomly choose m variables on which to base the decision at that node. Calculate the best split based on these m variables in the training set. 5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree

classifier). For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the mode vote of all trees is reported as random forest prediction.

4.2.2 Bagging

Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets. Tests on real and simulated data sets using classification and regression trees and subset selection in linear regression show that bagging can give substantial gains in accuracy. The vital element is the instability of the prediction method. If perturbing the learning set can cause significant changes in the predictor constructed, then bagging can improve accuracy.

In this model, the REPTree is used here as the base classification method, its visual tree is displayed below in Figure 11. The number of iterations is set to be 10 and the seed is 1. ESpm08d is used as the root node for the first round classification, and a_nN and b_double are the sub-root nodes. And all the descriptors are shown in this tree as a node or a leaf, but for the ADTree (discussed later) only selected descriptors are used to form the classification tree.

4.2.3 Logitboost

Boosting (Freund & Schapire 1996, Schapire & Singer 1998) is one of the most important recent developments in classification methodology. The performance of many classification algorithms often can be dramatically improved by sequentially applying them to reweighted versions of the input data, and taking a weighted majority vote of the sequence of classifiers thereby produced. For the two-class problem, boosting can be viewed as an approximation to additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion. M5P classifier and 10 iterations were used here.

4.2.4 Artificial Neural Network (ANN)

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. Here the training times is 1000, hidden layers are 10 and the validation threshold is 20. There are 7 Sigmoid Node in total in this model. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

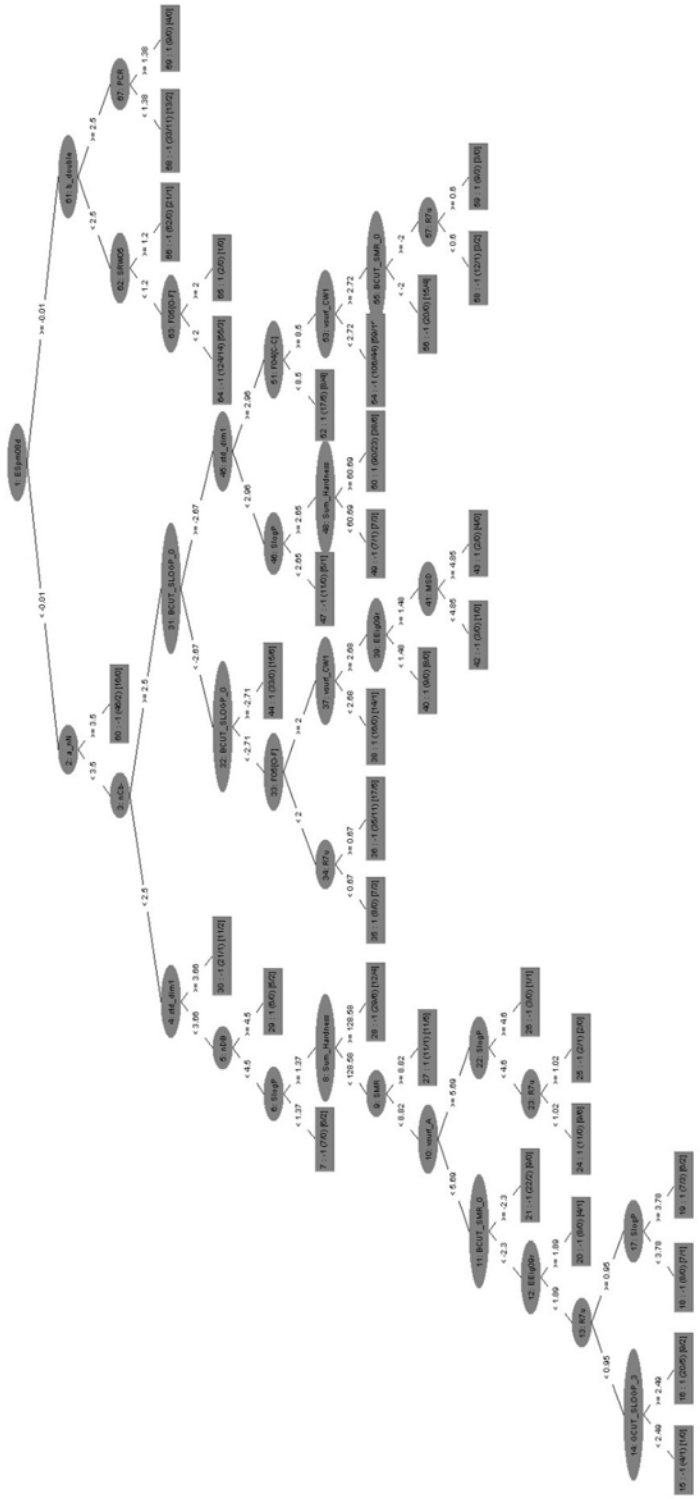


Figure 10 REPTree of bagging method

4.2.5 k-Nearest Neighbors (kNN)

The kNN method is a simple classification method based on local information around each object. kNN is a nonparametric method where the classification of an object depends on the class assignments of its k-nearest neighbors, without making any assumptions about the distribution and the shape of the classes or about the form of class boundaries. The nearness is measured by an appropriate distance metric called Euclidean distance. In this model, the k value is set to be 5, and the linear NN search algorithm is applied. The standard kNN method is implemented simply as follows: (i) calculate distances between each unknown object (u) and all the objects in the training set; (ii) select a range for k; (iii) for each k value, the class to which a majority of the k-nearest training objects belong is assigned to each queryu; (iv) the k value giving the lowest leave-one-out (LOO) cross-validation error rate is the optimal and is used for new object prediction.

4.2.6 Decorate

Decorate (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) is a meta-learner build an effective diverse committee in a simple, straightforward manner (P. Melville 2003). This is accomplished by adding different randomly constructed examples to the training set when building new committee members. These artificially constructed examples are given category labels that disagree with the current decision of the committee, thereby easily and directly increasing diversity when a new classifier is trained on the augmented data and added to the committee. In this model the artificial size is 1.0, the desired number of member classifiers in the Decorate ensemble is 10, and the number of iterations is 10. J48 is used as the classification method.

4.2.7 Kstar

K star is an instance-based classifier, which is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function. It can calculate the probability of an instance a being in category C by summing the probabilities from A to each instance that is A member of C. The probabilities for each category are calculated. The relative probabilities obtained give an estimate of the category distribution at the point of the instance space represented by A. Most other techniques return a single category as the result of classification. For ease of comparison here we choose the category with the highest probability as the classification of the new instance. Alternatives to this include choosing a class at random using the relative probabilities or returning a normalized probability distribution as the answer. The global blend number is 20 in this model.

4.2.8 Alternating Decision Tree (ADTree)

Alternating decision trees (ADTree) is a kind of option tree. Option trees differ from decision trees (such as REPTree mentioned above) in that they contain two types of nodes: a decision node and a prediction node, while decision trees just contain a decision node. When a query reaches a decision node, the sign of this node will be assigned to the query, like in the decision tree. However, when the query reaches a prediction node, it will continue to all the paths of this node. So in an alternating decision tree, the studied chemical could follow different branches (multipath). The sign of the sum of all the prediction nodes that is included in a multipath is the class which the tree associates to the query. One possibility to grow an option tree is incrementally adding nodes to a decision tree. This is commonly done by using the boosting algorithm, and the resulted trees are usually called ADTree instead of option trees. The number of boosting iterations is an important parameter that can be tuned to suit the data set and the desired complexity−accuracy trade-off, which was set as 10 in this model. The default search method of exhaustive search (expands all paths) was used in this research (Figure 12).

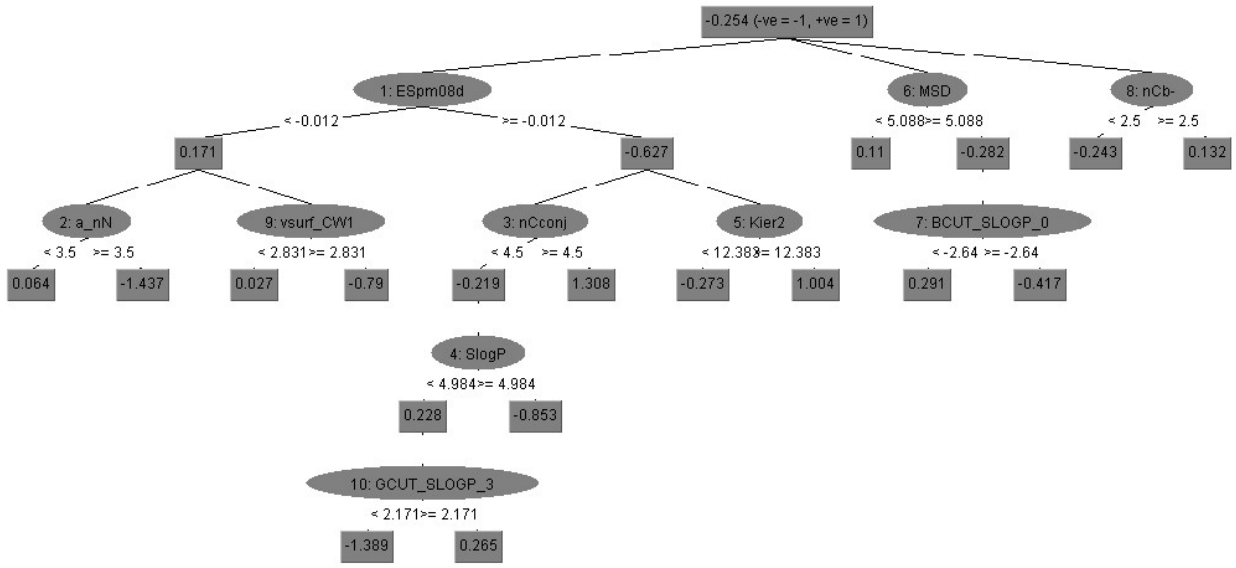


Figure 11 ADTree classification nodes

4.2.9 Local Lazy Method (Lazy IB1)

Lazy learners is a memory-instance-based learning technique, which stores the training objects and does no real work until a prediction is required for an unknown object ***u***. The term lazy arises because the predictions for the test set chemicals are made without producing a model a priori on the whole training set. Considering the close neighborhood of a query point according to a Euclidean distance measure (Farrell, Arnone et al. 2011), the activity of the query is predicted from the activities of the most chemically similar neighbor chemicals in the training set. Once the nearest training sample has been located, IB1 predicts the same class as the training sample for ***u***. If several samples qualify as the closest, then the first one found is used. In this case, it would not give predictions accurate enough. That's why this method is the worst among all nine methods used in this project.

4.3 Identified Chemicals

4.3.1 Cell Line Testing Analysis

DHT displacement assay and eGFP screening assay were used to fast-screening selected chemicals, and then dose-dependent MTS assays were conducted to evaluate cell survival using two prostate cell lines, LNCaP and PC3. Six tested chemicals exhibit profound concentration-dependent suppression of cell survival according to MTS assays (Table 10).

Chemical 12007 is one of the best hits that emerged from the virtual screening. It demonstrates excellent IC_{50} in both DHT displacement assay and eGFP assay, with the corresponding values of 2.16μM and 1.84μM respectively. Chemical 12007 is not a potent suppressor according of the cell growth, and decreases cell survival gently and in a concentration dependent manner. It's a good sign for a potential drug candidate, meaning that it stops cancer cell growth, while exhibiting low general cytotoxicity. .

Chemical 12060 is another promising candidate. Its DHT displacement assay IC_{50} is 1.04 μM which is the lowest among all tested compounds. The IC_{50} established for the compound by the eGFP screen was 3.42 μM which is also very encouraging. The MTS test results indicate that it suppresses LNCaP and MDV3100-resistant LNCaP cells in a similar way as chemical 12007, and demonstrates the overall potency similar to Casodex and MDV3100. Chemical 12060 owns two chiral carbon atoms, and is structurally symmetrical except for one nitrogen atom forming an indole-like structure of one half of the chemical. Derivatives of this compound could be developed by adding various functional groups to the benzyl rings.

Chemicals 12002, 12051 and 12052 demonstrated modest inhibition of cell growth. Besides, their general structure is relatively simple and common compared to with 12007 and 12060, which sets certain limit to their further investigation.

Chemicals 12058 and 12061 showed tremendous suppression effect on PC3 cells, indicating their generic toxicity is very high and is not AR specific. This is possibly due to off-target effects which cannot be ruled out at this stage.

Moreover, chemicals 12007 and 12051 the MTS fluorescence is stronger in 50 μM and 100 μM concentrations testing, for the possible reason that these two chemicals could have auto-fluorescence dependent with chemical concentration.

4.3.2 Docking Poses Analysis

After identifying the best initial hits from the selected chemicals, the corresponding molecules were checked for their possible binding mode to the AR. The AR crystal structure 3L3X (Zhou, Suino-Powell et al. 2010) from the Protein Data Bank was prepared for docking using Maestro suite (Maestro 2011). Docking poses are listed in table 14. Seven of all nine selected chemicals form strong interactions with the pocket amino acid residues of the receptors, especially T877 and N705, which is crucial for the binding of these antagonists according to the docking results we got.

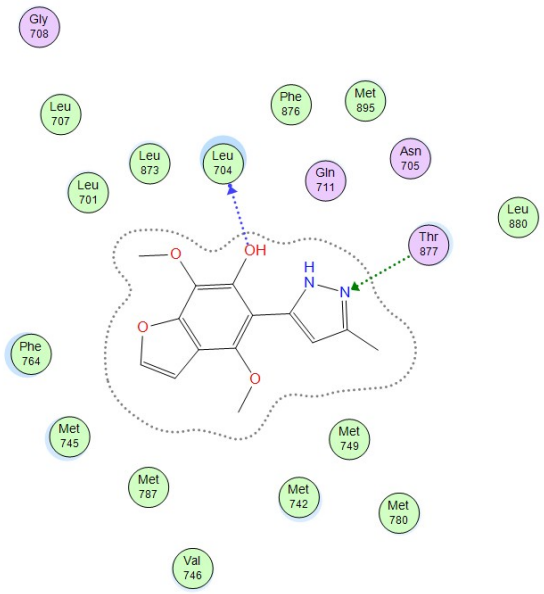
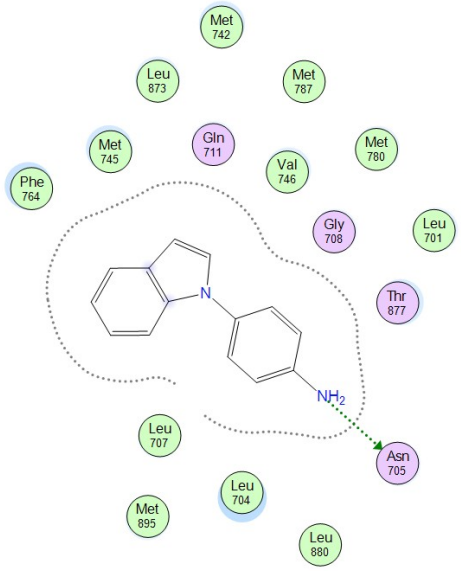
The androgen binding pocket of the AR is mainly composed of hydrophobic residues that can form strong non-polar interactions with androgenic steroids such as testosterone and DHT. The protein-ligand anchoring can be additionally stabilized by a network of hydrogen bonds involving R752, Q711, N705 and T877 polar residues. The interactions between the AR and its steroidal and non-steroidal agonists have been extensively discussed (He, Gampe et al. 2004; De Jesus-Tran, Cote et al. 2006). Combined with the observation that residues forming the AR HBS are remarkably flexible, and can adjust to ligands of various sizes (Bohl, Wu et al. 2008). Previous reports indicated that mutation of W741 to leucine or cysteine will generate additional space in the ABS that allows accommodation of the bulky phenyl ring of bicalutamide and converts its antagonist activity on the AR into an agonist that stimulates transcriptional activity and cancer growth. Similarly, a well-documented agonist-converting T877A mutation was found in the AR present in LNCaP cells(Hara, Miyazaki et al. 2003).

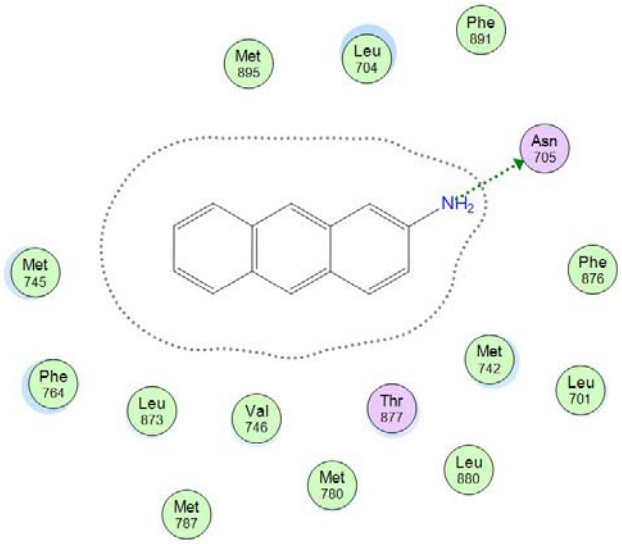
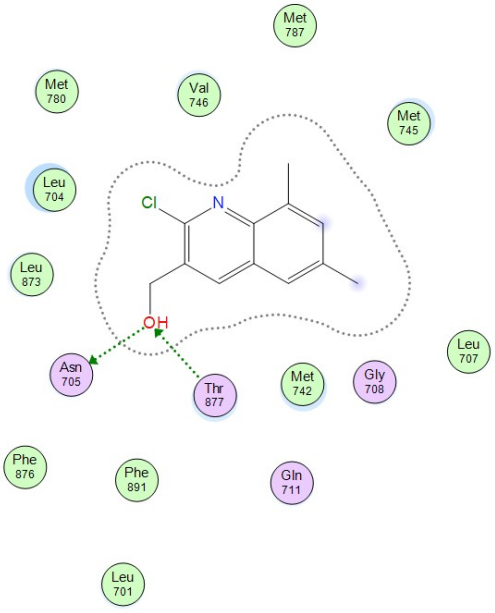
It has been simulated by docking that binding of 12007, 12051, 12052, 12058 and 12063 to the AR LBD occurs at the N705 residue, which is away from the possible mutation residues mentioned above, and therefore a mutation here is not likely to have an effect on the chemical's activity.

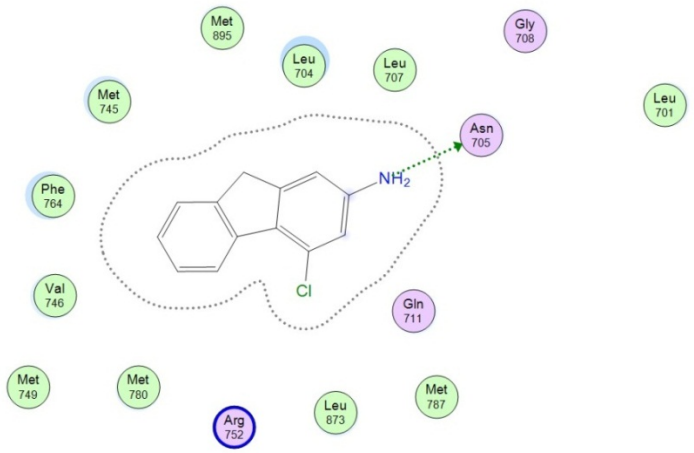
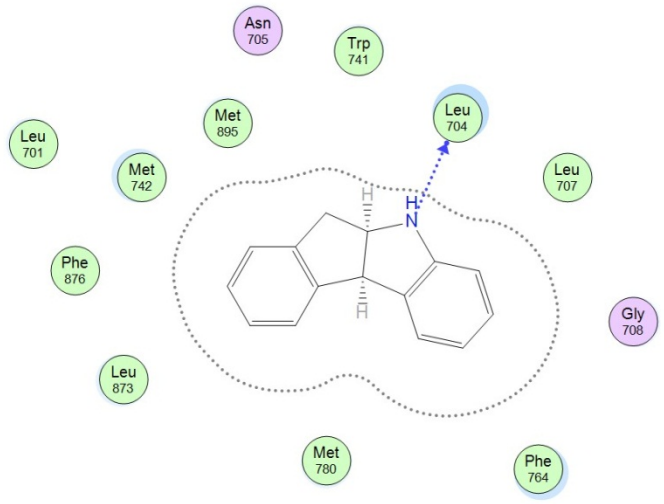
Similarly, a well-documented agonist-converting T877A mutation, as found in the AR present in LNCaP cells, could possibly influence binding of chemical 12002, 12052 and 12063 to this site, as these chemicals form critical contacts with T877 or its mutant(s) as flutamide analogues (Hara, Miyazaki et al. 2003). In support of this, chemicals 12002, 12052 and 12063, didn't demonstrate effective inhibition of the LNCaP cell line (Table 9). eGFP IC_{50} of 12002 and 12063 is more than two times as their DHT displacement IC_{50}.

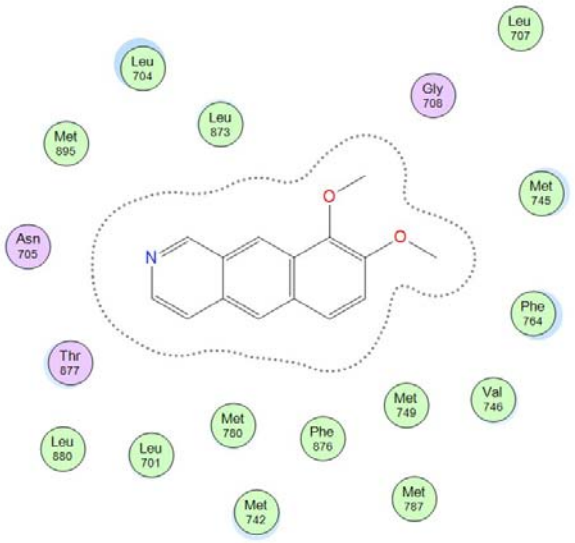
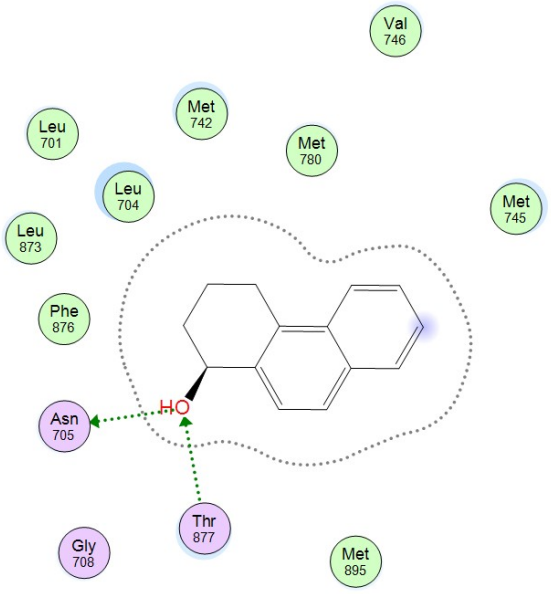
12060 and 12068 form hydrogen bonds at Leu704, which is a residue with almost no report of mutations. Both these two chemicals have chiral carbon atoms, and it's possible that chiral carbons have a favor for Leu704. 12061 form no strong interactions with residues.

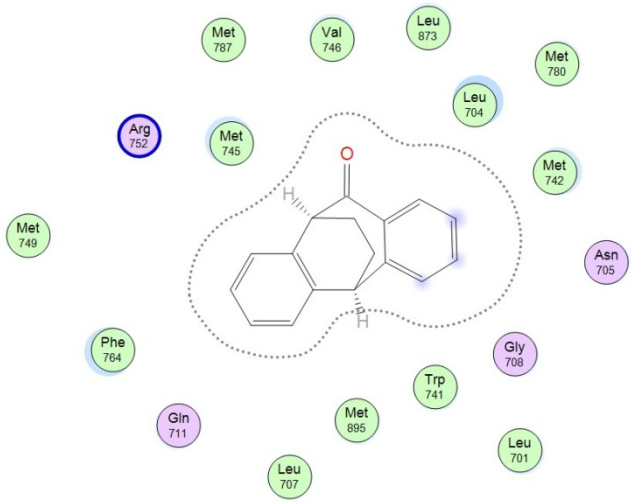
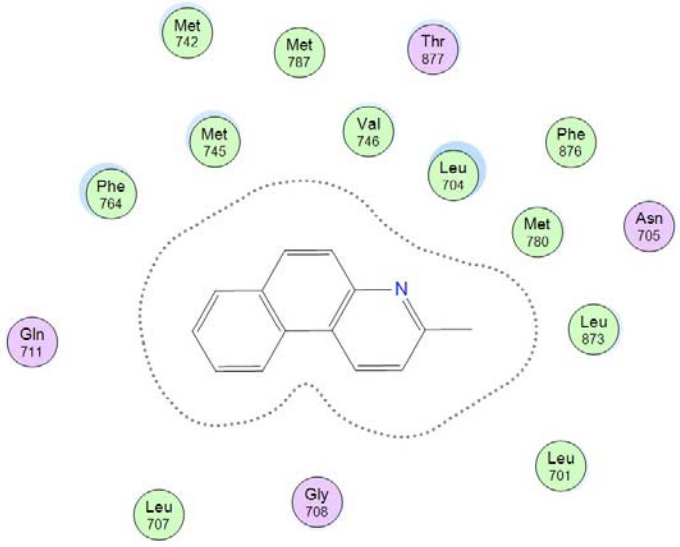
Table 14 Interactions of AR and identified AR antagonists

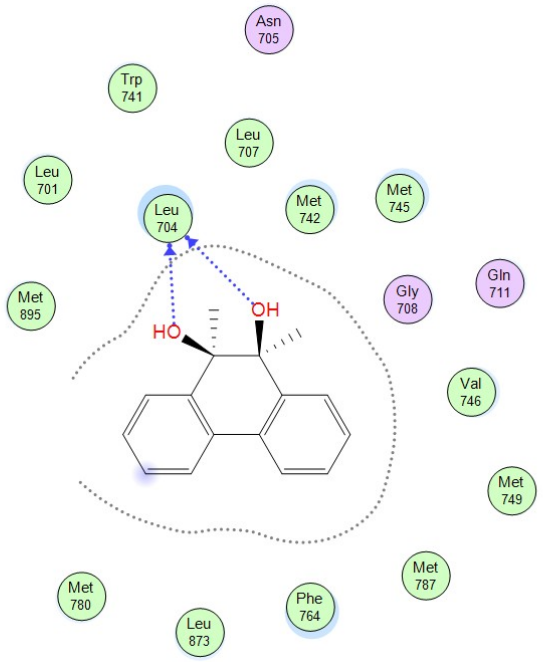
Chemical	Docking result - interactions
12002	
12007	

Chemical	Docking result - interactions
12051	
12052	

Chemical	Docking result - interactions
12058	
12060	

Chemical	Docking result - interactions
12061	
12063	

Chemical	Docking result - interactions
12065	
12066	

Chemical	Docking result - interactions
12068	

4.4. Future Directions

In the future, we will focus on the most promising hit compounds 12007 and 12060 identified. The compounds could be subjected to Med Chem optimization to develop synthetic derivatives with improved bioavailability, half-life time, toxicity and permeability among other properties relevant for a good drug candidate.

Such improved compounds would then undergo experimental evaluation in mice to see whether it can influence tumor growth and survival. If successful, such compounds would represent good candidates for pre-clinical evaluation and eventually clinical trials.

5. Conclusion

In this study, an effective QSAR pipeline was developed and proved to be capable of identifying new AR antagonists from a large public collection of purchasable chemicals. In particular, we have utilized DRAGON, INDUCTIVE and MOE QSAR descriptors to create various binary models on anti-AR activity.

When the developed QSAR solutions were utilized to screen more than 2 M chemicals from the ZINC database, we could identify 39 candidate compounds. When tested within the DHT displacement assay, 9 chemicals demonstrated efficient low-micromole level of activity. Of those, 9 compounds later exhibited ability to inhibit AR in the eGFP transcriptional assay with the corresponding IC_{50} values established in 1.04-16.18μM range. Notably, 5 discovered chemicals demonstrated concentration-dependent suppression of survival of LNCaP prostate cancer cell lines. Chemicals 12007 and 12060 demonstrated the lowest IC_{50} from DHT displacement assay and eGFP assay, and effectively decreased cell viability in the MTS assay, what allowed us to characterize them as our lead compounds.

The results of this study set a ground for the development of an entire novel chemicals class of AR antagonists that are distinct for the currently marketed drugs such as Nitalutamide, Flutomide, Cassodex, and MDV3100 which all share significant structural similarity. The preliminary SAR information obtained around their analogues may serve as a useful basis for the development of an entirely new class of drugs for treating anti-androgen resistant prostate cancer.

Bibliography

Albertsen, P. C., J. A. Hanley, et al. (2005). "20-year outcomes following conservative management of clinically localized prostate cancer." JAMA **293**(17): 2095-101.

Antonarakis, E. S. and M. A. Eisenberger (2011). "Expanding treatment options for metastatic prostate cancer." N Engl J Med **364**(21): 2055-8.

Attar, R. M., M. Jure-Kunkel, et al. (2009). "Discovery of BMS-641988, a novel and potent inhibitor of androgen receptor signaling for the treatment of prostate cancer." Cancer Res **69**(16): 6522-30.

Axerio-Cilies, P., N. A. Lack, et al. (2011). "Inhibitors of androgen receptor activation function-2 (AF2) site identified through virtual screening." J Med Chem **54**(18): 6197-205.

Bernhard Pfahringer, G. H. a. R. K. (2001). "Optimizing the Induction of Alternating Decision Trees. ." Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining: 477-487.

Bohl, C. E., Z. R. Wu, et al. (2008). "Effect of B-ring substitution pattern on binding mode of propionamide selective androgen receptor modulators." Bioorganic & Medicinal Chemistry Letters **18**(20): 5567-5570.

Boyd, S. (2005). "Molecular operating environment." Chemistry World **2**(9): 66-66.

Breiman, L. (1996). "Bagging predictors." Machine Learning **24**(2): 123-140.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Brogden, R. N. and P. Chrisp (1991). "Flutamide. A review of its pharmacodynamic and pharmacokinetic properties, and therapeutic use in advanced prostatic cancer." Drugs Aging **1**(2): 104-15.

Burden, F. R. (1989). "Molecular-Identification Number for Substructure Searches." Journal of Chemical Information and Computer Sciences **29**(3): 225-227.

Burden, F. R. (1997). "A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix." Quantitative Structure-Activity Relationships **16**(4): 309-314.

Cherkasov, A. (2003). "Inductive electronegativity scale. Iterative calculation of inductive partial charges." Journal of Chemical Information and Computer Sciences **43**(6): 2039-2047.

Cherkasov, A. (2005). "'Inductive' Descriptors: 10 Successful Years in QSAR." Current Computer-Aided Drug Design **1**(1): 21-42.

Cherkasov, A. and B. Jankovic (2004). "Application of 'inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides." Molecules **9**(12): 1034-1052.

Cherkasov, A., Z. Shi, et al. (2005). "'Inductive' charges on atoms in proteins: Comparative docking with the extended steroid benchmark set and discovery of a novel SHBG ligand." Journal of Chemical Information and Modeling **45**(6): 1842-1853.

Cockshott, I. D. (2004). "Bicalutamide: clinical pharmacokinetics and metabolism." Clin Pharmacokinet **43**(13): 855-78.

Consonni, V., R. Todeschini, et al. (2002). "Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors." Journal of Chemical Information and Computer Sciences **42**(3): 682-692.

Consonni, V., R. Todeschini, et al. (2002). "Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies." Journal of Chemical Information and Computer Sciences **42**(3): 693-705.

Cortes, C. and V. Vapnik (1995). "Support-Vector Networks." Machine Learning **20**(3): 273-297.

Craig, P. N., C. H. Hansch, et al. (1971). "Minimal Statistical Data for Structure-Function Correlations." Journal of Medicinal Chemistry **14**(5): 447-&.

D. Aha, D. K. (1991). "Instance-based learning algorithms." Machine Learning **6**: 37-66.

De Jesus-Tran, K. P., P. L. Cote, et al. (2006). "Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity." Protein Science **15**(5): 987-999.

Dearden, J. C. (2003). "In silico prediction of drug toxicity." Journal of Computer-Aided Molecular Design **17**(2): 119-127.

Dole, E. J. and M. T. Holdsworth (1997). "Nilutamide: an antiandrogen for the treatment of prostate cancer." Ann Pharmacother **31**(1): 65-75.

Dubbink, H. J., R. Hersmus, et al. (2004). "Distinct recognition modes of FXXLF and LXXLL motifs by the androgen receptor." Mol Endocrinol **18**(9): 2132-50.

Dudek, A. Z., T. Arodz, et al. (2006). "Computational methods in developing quantitative structure-activity relationships (QSAR): A review." Combinatorial Chemistry & High Throughput Screening **9**(3): 213-228.

Egmont-Petersen, M., de Ridder, D., Handels, H. (2002). "Image processing with neural networks - a review." Pattern Recognition **35**(10): 2279–2301.

Farrell, P. G., L. J. Arnone, et al. (2011). "Euclidean distance soft-input soft-output decoding algorithm for low-density parity-check codes." Iet Communications **5**(16): 2364-2370.

Foster, W. R., B. D. Car, et al. (2011). "Drug safety is a barrier to the discovery and development of new androgen receptor antagonists." Prostate **71**(5): 480-8.

Gao, W. (2010). "Peptide antagonist of the androgen receptor." Curr Pharm Des **16**(9): 1106-13.

Gaulton, A., L. J. Bellis, et al. (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic Acids Research **40**(D1): D1100-D1107.

Gilson, M. K., X. Chen, et al. (2001). "BindingDB: An on-line molecular recognition database." Biophysical Journal **80**(1): 33a-33a.

Gleave, M., S. L. Goldenberg, et al. (1998). "Intermittent androgen suppression for prostate cancer: rationale and clinical experience." Prostate Cancer Prostatic Dis **1**(6): 289-296.

Gramatica, P. (2007). "Principles of QSAR models validation: internal and external." Qsar & Combinatorial Science **26**(5): 694-701.

Gramatica, P., N. Navas, et al. (1998). "3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs)." Chemometrics and Intelligent Laboratory Systems **40**(1): 53-63.

Haendler, B. and A. Cleve (2012). "Recent developments in antiandrogens and selective androgen receptor modulators." Mol Cell Endocrinol **352**(1-2): 79-91.

Hara, T., J. Miyazaki, et al. (2003). "Novel mutations of androgen receptor: A possible mechanism of bicalutamide withdrawal syndrome." Cancer Research **63**(1): 149-153.

Hawkins, D. M., S. C. Basak, et al. (2003). "Assessing model fit by cross-validation." Journal of Chemical Information and Computer Sciences **43**(2): 579-586.

He, B., R. T. Gampe, et al. (2004). "Structural basis for androgen receptor interdomain and coactivator interactions suggests a transition in nuclear receptor activation function dominance." Molecular Cell **16**(3): 425-438.

Heemers, H. V. and D. J. Tindall (2007). "Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex." Endocr Rev **28**(7): 778-808.

Helma, C. (2005). "In silico predictive toxicology: The state-of-the-art and strategies to predict human health effects." Current Opinion in Drug Discovery & Development **8**(1): 27-31.

Hobisch, A., Z. Culig, et al. (1995). "Distant metastases from prostatic carcinoma express androgen receptor protein." Cancer Res **55**(14): 3068-72.

Hopfield, J. J. (1982). "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences **79**(8): 2554-2558.

Hullermeier, E., J. Furnkranz, et al. (2008). "Label ranking by learning pairwise preferences." Artificial Intelligence **172**(16-17): 1897-1916.

Irwin, J. J. and B. K. Shoichet (2005). "ZINC - A free database of commercially available compounds for virtual screening." Journal of Chemical Information and Modeling **45**(1): 177-182.

J. Friedman, T. H., R. Tibshirani (2000). "Additive Logistic Regression: a Statistical View of Boosting." Ann. Statist. **28**(2): 337-407.

Janne, O. A. and L. X. Shan (1991). "Structure and function of the androgen receptor." Ann N Y Acad Sci **626**: 81-91.

John G. Cleary, L. E. T. (1995). "K: An Instance-based Learner Using an Entropic Distance Measure." 12th International Conference on Machine Learning: 108-114.

Jung, M. E., S. Ouk, et al. (2010). "Structure-activity relationship for thiohydantoin androgen receptor antagonists for castration-resistant prostate cancer (CRPC)." J Med Chem **53**(7): 2779-96.

Kachigan, S. K. (1991). "MultiVariate Statistical Analysis: A Conceptual Introduction." Radius Press: New York.

Keiser, M. J., B. L. Roth, et al. (2007). "Relating protein pharmacology by ligand chemistry." Nat Biotechnol **25**(2): 197-206.

Kent, E. C. and M. H. Hussain (2003). "The patient with hormone-refractory prostate cancer: determining who, when, and how to treat." Urology **62 Suppl 1**: 134-40.

Knee, D. A., B. A. Froesch, et al. (2001). "Structure-function analysis of Bag1 proteins. Effects on androgen receptor transcriptional activity." J Biol Chem **276**(16): 12718-24.

Konovalov, D. A., N. Sim, et al. (2008). "Statistical confidence for variable selection in QSAR models via Monte Carlo cross-validation." Journal of Chemical Information and Modeling **48**(2): 370-383.

Lack, N. A., P. Axerio-Cilies, et al. (2011). "Targeting the binding function 3 (BF3) site of the human androgen receptor through virtual screening." J Med Chem **54**(24): 8563-73.

Li, J. Z. and P. Gramatica (2010). "Classification and Virtual Screening of Androgen Receptor Antagonists." Journal of Chemical Information and Modeling **50**(5): 861-874.

Maestro (2011). "Schrödinger, LLC, www.schrodinger.com."

Makkonen, H., M. Kauhanen, et al. (2011). "Androgen receptor amplification is reflected in the transcriptional responses of Vertebral-Cancer of the Prostate cells." Mol Cell Endocrinol **331**(1): 57-65.

Mark Hall, E. F., Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009). "The WEKA Data Mining Software: An Update." SIGKDD Explorations **11**(1).

Marko Robnik-Sikonja, I. K. (1997). "An adaptation of Relief for attribute estimation in regression." Fourteenth International Conference on Machine Learning: 296-304.

Martel, C. L., P. H. Gumerlock, et al. (2003). "Current strategies in the management of hormone refractory prostate cancer." Cancer Treat Rev **29**(3): 171-87.

Masiello, D., S. Cheng, et al. (2002). "Bicalutamide functions as an androgen receptor antagonist by assembly of a transcriptionally inactive receptor." J Biol Chem **277**(29): 26321-6.

McEwan, I. J. (2004). "Molecular mechanisms of androgen receptor-mediated gene regulation: structure-function analysis of the AF-1 domain." Endocr Relat Cancer **11**(2): 281-93.

Mukherjee, A., L. Kirkovsky, et al. (1996). "Enantioselective binding of Casodex to the androgen receptor." Xenobiotica **26**(2): 117-22.

Narayanan, R., M. L. Mohler, et al. (2008). "Selective androgen receptor modulators in preclinical and clinical development." Nucl Recept Signal **6**: e010.

P. Melville, R. J. M. (2003). "Constructing Diverse Classifier Ensembles Using Artificial Training Examples." Eighteenth International Joint Conference on Artificial Intelligence: 505-510.

Patani, G. A. and E. J. LaVoie (1996). "Bioisosterism: A Rational Approach in Drug Design." Chem Rev **96**(8): 3147-3176.

Ramos-Jimenez, G., J. del Campo-Avila, et al. (2005). "Induction of decision trees using an internal control of induction." Computational Intelligence and Bioinspired Systems, Proceedings **3512**: 795-803.

Rao, C. R. and Y. Wu (2005). "Linear model selection by cross-validation." Journal of Statistical Planning and Inference **128**(1): 231-240.

Rennie, P. S. and C. C. Nelson (1998). "Epigenetic mechanisms for progression of prostate cancer." Cancer Metastasis Rev **17**(4): 401-9.

Roy, A. K., Y. Lavrovsky, et al. (1999). "Regulation of androgen action." Vitam Horm **55**: 309-52.

Roy, K. (2007). "On some aspects of validation of predictive quantitative structure-activity relationship models." Expert Opinion on Drug Discovery **2**(12): 1567-1577.

Roy, P. P., J. T. Leonard, et al. (2008). "Exploring the impact of size of training sets for the development of predictive QSAR models." Chemometrics and Intelligent Laboratory Systems **90**(1): 31-42.

Roy, P. P., S. Paul, et al. (2009). "On Two Novel Parameters for Validation of Predictive QSAR Models." Molecules **14**(5): 1660-1701.

Roy, P. P. and K. Roy (2008). "On some aspects of variable selection for partial least squares regression models." Qsar & Combinatorial Science **27**(3): 302-313.

Salvati, M. E., A. Balog, et al. (2008). "Identification and optimization of a novel series of [2.2.1]-oxabicyclo imide-based androgen receptor antagonists." Bioorg Med Chem Lett **18**(6): 1910-5.

Saporita, A. J., Q. Zhang, et al. (2003). "Identification and characterization of a ligand-regulated nuclear export signal in androgen receptor." J Biol Chem **278**(43): 41998-2005.

Scher, H. I., G. Buchanan, et al. (2004). "Targeting the androgen receptor: improving outcomes for castration-resistant prostate cancer." Endocr Relat Cancer **11**(3): 459-76.

Shao, J. (1993). "Linear-Model Selection by Cross-Validation." Journal of the American Statistical Association **88**(422): 486-494.

Sharifi, N., J. L. Gulley, et al. (2005). "Androgen deprivation therapy for prostate cancer." Jama-Journal of the American Medical Association **294**(2): 238-244.

Shen, H. C. and S. P. Balk (2009). "Development of androgen receptor antagonists with promising activity in castration-resistant prostate cancer." Cancer Cell **15**(6): 461-3.

Society, C. C. (2012). "“General Statistics” media backgrounder."

Talete, s. (2007). "DRAGON for Window (Software for Molecular Descriptor Calculations). Version 5.5 - 2007 - http://www.talete.mi.it."

Taplin, M. E., B. Rajeshkumar, et al. (2003). "Androgen receptor mutations in androgen-independent prostate cancer: Cancer and Leukemia Group B Study 9663." J Clin Oncol **21**(14): 2673-8.

Tavassoli, P., R. Snoek, et al. (2007). "Rapid, non-destructive, cell-based screening assays for agents that modulate growth, death, and androgen receptor activation in prostate cancer cells." Prostate **67**(4): 416-426.

Tilley, W. D., S. S. Lim-Tio, et al. (1994). "Detection of discrete androgen receptor epitopes in prostate cancer by immunostaining: measurement by color video image analysis." Cancer Res **54**(15): 4096-102.

Todeschini, R. and P. Gramatica (1997). "3D-modelling and prediction by WHIM descriptors .6. Application of WHIM descriptors in QSAR studies." Quantitative Structure-Activity Relationships **16**(2): 120-125.

Todeschini, R. C., V. (2009). "Handbook of Molecular Descriptors." Wiley-VCH: Weinheim, Germany,.

Tong, W. D., H. X. Hong, et al. (2005). "Assessing QSAR Limitations - A Regulatory Perspective." Current Computer-Aided Drug Design **1**(2): 195-205.

Tran, C., S. Ouk, et al. (2009). "Development of a second-generation antiandrogen for treatment of advanced prostate cancer." Science **324**(5928): 787-90.

Tropsha, A., P. Gramatica, et al. (2003). "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models." Qsar & Combinatorial Science **22**(1): 69-77.

van de Wijngaart, D. J., M. Molier, et al. (2010). "Systematic structure-function analysis of androgen receptor Leu701 mutants explains the properties of the prostate cancer mutant L701H." J Biol Chem **285**(7): 5097-105.

Warmuth, M. K. and D. Kuzmin (2008). "Randomized Online PCA Algorithms with Regret Bounds that are Logarithmic in the Dimension." Journal of Machine Learning Research **9**: 2287-2320.

Williams, A. J. and V. Tkachenko (2010). "ChemSpider: How an online resource of chemical compounds, reaction syntheses, and property data can support green chemistry." Abstracts of Papers of the American Chemical Society **239**.

Wishart, D. S., C. Knox, et al. (2008). "DrugBank: a knowledgebase for drugs, drug actions and drug targets." Nucleic Acids Research **36**: D901-D906.

Yoshino, H., H. Sato, et al. (2010). "Design and synthesis of an androgen receptor pure antagonist (CH5137291) for the treatment of castration-resistant prostate cancer." Bioorg Med Chem **18**(23): 8150-7.

Yuan, Y. F. and M. J. Shaw (1995). "Induction of Fuzzy Decision Trees." Fuzzy Sets and Systems **69**(2): 125-139.

Zhou, X. E., K. M. Suino-Powell, et al. (2010). "Identification of SRC3/AIB1 as a preferred coactivator for hormone-activated androgen receptor." J Biol Chem **285**(12): 9161-71.

Appendices

1. Screening top-scored 39 chemicals

2. Training set chemicals

3. External Testing Set Chemicals

4. Positive predictive value (PPV), Negative Predictive Value (NPV), sensitivity, specificity, concordance, ROC AUC of all training sets

4.1 Training Set 1 (T1)

4.2 Training Set 2 (T2)

4.3 Training Set 1 and Set 2 (T1+T2)

1. Screening Top-scored 39 Chemicals

INTERNAL ID	ZINC_ID	SMILES
12001	ZINC00171669	COc1cc2c(cc1OC)C(=NCC2)c3cccc(c3)Cl
12002	ZINC00488165	Cc1cc([nH]n1)c2c(c(c3c(c2OC)cco3)OC)O
12003	ZINC00564515	Cc1c(c2ccccc2n1Cc3ccccc3F)CO
12004	ZINC04421403	c1ccc(cc1)[S+]2c3ccccc3Nc4c2cccc4
12005	ZINC04488996	COc1cc2c(c(c1)OC)[C@H](CC(=O)N2)c3ccc(cc3)Cl
12006	ZINC04600179	c1ccc2c(c1)ccc(n2)c3ccc(cc3)O
12007	ZINC12394346	c1ccc2c(c1)ccn2c3ccc(cc3)N
12008	ZINC12441512	c1ccc(cc1)[C@@H]2CN(CCO2)C(=O)c3c(cccc3Cl)F
12009	ZINC58427578	C[C@H]1CN(C[C@@H](O1)c2ccccc2)C(=O)c3cccc4c3OCCO4
12010	ZINC00155064	c1ccc2c(c1)CCc3ccccc3C2=NO
12011	ZINC58282811	CCN(CC)C(=O)C1CCN(CC1)c2c(cc(cc2F)C#N)F
12012	ZINC11036200	c1ccc2c(c1)C(=O)N([C@H](N2)c3cccc(c3)C#N)CC(F)(F)F
12014	ZINC05921978	Cc1c2ccccc2[s+]c-3c1CCc4c3cccc4
12015	ZINC00122320	c1ccc2c(c1)cc(o2)c3ccc(cc3)N
12016	ZINC04772847	C1CC[C@H]2[C@H](C1)N(C(N2O)c3c(c(c(c(c3F)F)F)F)F)O
12017	ZINC27154492	CC[C@@H]1CCCN(C1)C(=O)c2ccc(c3c2nccc3)Cl
12018	ZINC04517670	c1ccc2c(c1)cc3cccc(c3n2)O
12051	ZINC01593346	c1ccc2cc3cc(ccc3cc2c1)N
12052	ZINC00421637	Cc1cc(c2c(c1)cc(c(n2)Cl)CO)C
12053	ZINC00397781	c1cc(ccc1CNc2ccc(cc2)Cl)O
12054	ZINC01577019	c1ccc2c(c1)ccc3c2cc(cc3)N
12055	ZINC01671287	CCc1c2ccc3ccccc3c2c[n+]4c1cccc4
12056	ZINC01681528	c1ccc2c(c1)-c3ccc(cc3[C@@H]2Br)N
12057	ZINC01627374	Cc1c2ccccc2c(c3c1cc[n+]4c3cccc4)C
12058	ZINC01730845	c1ccc-2c(c1)Cc3c2c(cc(c3)N)Cl
12059	ZINC01716374	c1ccc2c(c1)cc(c3c2c4c(cc3)OCO4)CO
12060	ZINC01744320	c1ccc2c(c1)C[C@@H]3[C@H]2c4ccccc4N3
12061	ZINC01561336	COc1ccc2cc3ccncc3cc2c1OC
12062	ZINC01571863	c1ccc2c(c1)CC[C@H]3N2[C@@H]4c5ccccc5C(=O)N4CC3
12063	ZINC05699716	c1ccc2c(c1)ccc3c2CCC[C@@H]3O
12064	ZINC01610217	c1ccc2c(c1)ccc3c2[nH]c(c3)CO

INTERNAL ID	ZINC_ID	SMILES
12065	ZINC01629402	c1ccc2c(c1)[C@H]3CC[C@@H]2C(=O)c4c3cccc4
12066	ZINC01680651	Cc1ccc2c3ccccc3ccc2n1
12067	ZINC18173676	Cc1[n+](c2c3ccsc3ccc2s1)CCO
12068	ZINC13658815	C[C@]1(c2ccccc2-c3ccccc3[C@]1(C)O)O
12069	ZINC00196649	COc1ccc2c(c1)c(co2)[C@@H](c3ccccc3)O
12070	ZINC00473218	COc1ccc2cc(ccc2c1CO)Br
12071	ZINC03722619	COc1c(cc(cc1Cl)Cl)c2csc(n2)N
12072	ZINC00055479	C[n+]1c2ccccc2sc1COc3ccccc3O

2. Training Set 1 and Training Set 2 Chemicals

	Trainingset-1 Smiles	activity		Training set-2 Smiles	activity
1	FC(F)(F)c1cc(Nc2ccccc2O)ccc1N(O)O	1	1	O=C1CCC2=C3C(C4CC[C@H](O)[C@]4(C[C@@H]3CCCCC)C)CCC2=C1	1
2	O(C)c1cc(ccc1OC)\C=C\C(=O)\C=C(/O)\C=C\c1cc(OC)c(OC)cc1	1	2	O=C1CCC2=C3C(C4CC[C@H](O)[C@]4(C[C@@H]3CCCCCCCCCCCC)C)CCC2=C1	1
3	O(C)c1cc(ccc1OC)\C=C\C(=O)CC(=O)\C=C\c1cc(OC)c(OC)cc1	1	3	O=C1CCC=2C(=C1)CCC1C3CC[C@H](O)[C@]3(CCC1=2)C	1
4	O(CC(OC)=O)c1cc(ccc1OC)\C=C\C(\O)=C\C(=O)\C=C\c1cc(OC)c(OCC(OC)=C)cc1	1	4	FC(F)(F)c1cc(ncc1C#N)N(CC1CC1)CCC	1
5	O(CC(OC)=O)c1cc(ccc1OC)\C=C\C(=O)CC(=O)\C=C\c1cc(OC)c(OCC(OC)=C)cc1	1	5	O=C1CCC2=C3C(C4CC[C@H](O)[C@]4(C[C@@H]3CCCCCCCC)C)CCC2=C1	1
6	O(C)c1cc(ccc1O)\C=C\C(=O)\C(=C(/O)\C=C\c1cc(OC)c(O)cc1)\CCC(O)=O	1	6	FC(F)(F)C1=CC(Oc2c1cc1CCC(Nc1c2)(C)C)=O	1
7	O(C)c1cc(ccc1O)\C=C\C(=O)C(CCC(O)=O)C(=O)\C=C\c1cc(OC)c(O)cc1	1	7	Clc1cc2nc([nH]c2cc1Cl)C(O)(C(C)C)C(F)(F)F	1
8	O(C)c1cc(ccc1O)\C=C\C(=O)\C(=C(/O)\C=C\c1cc(O)c(O)cc1)\CCC(OCC)=O	1	8	O=C1CCC2=C3C(C4CC[C@H](O)[C@]4(C[C@@H]3CC)C)CCC2=C1	1
9	O(C)c1cc(ccc1O)\C=C\C(=O)C(CCC(OCC)=O)C(=O)\C=C\c1cc(OC)c(O)cc1	1	9	FC(F)(F)c1cc(ncc1C#N)N(CCCC)C	1
10	O(C)c1cc(ccc1OC)C(=O)\C=C(/O)\c1cc(OC)c(OC)cc1	1	10	O=C1CC[C@]2([C@@H]3[C@@H]([C@H]4CC[C@@H](O)[C@@]4(CC3)C)CCC2=C1)C	1
11	O1c2cc(CCCC(=O)CCCc3cc1c(O)cc3)ccc2O	1	11	ClC=1C2=CC(=O)[C@H]3C(C3)[C@@]2(C2C(C=1)C1CC[C@](OC(=O)C)(C(=O)C)[C@]1(CC2)C)C	1
12	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(cc1)C#N)C)c1ccc(N)cc1	1	12	O=C1CCC2=C3[C@@H]([C@@H]4CC[C@](O)(C#CC)[C@]4(C[C@@H]3c3ccc(N(C)C)cc3)C)CCC2=C1	1
13	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(N)cc1	1	13	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)[C@@H]2[C@H]1C1CC2C=C1	1
14	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(NS(=O)(=O)C)cc1	1	14	FC(F)(F)c1cc(N2C(=O)C(NC2=O)(C)C)ccc1C#N	1
15	S(=O)(=O)(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(NS(=O)(=O)C)cc1	1	15	O=C1CCC2=C3[C@H]([C@@H]4CC[C@@](OC)(COC)[C@]4(C[C@@H]3c3ccc(cc3)\C=N\O)C)CCC2=C1	1

	Trainingset-1 Smiles	activity		Training set-2 Smiles	activity
16	S(=O)(=O)(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(NC(C(F)(F)F)=C)cc1	1	16	O=C1CC[C@@]2([C@@H]3C([C@@H]4CC[C@@](C(=O)C)(CC(=O)C)[C@]4(CC3)C)C[C@@H](C2=C1)C)C	1
17	FC(F)(F)c1ccc(NC(=O)C(O)(C)C)cc1	0	17	O=C1CC[C@@]2([C@@H]3[C@H]([C@@H]4CC[C@](OC(=O)C)(C(=O)C)[C@]4(CC3)C)C[C@@H](C2=C1)C)C	0
18	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(cc1)C#N)C)c1ccc(NC(=O)C)cc1	1	18	Clc1cc2nc([nH]c2cc1Cl)C(O)(C=C)C(F)(F)F	1
19	S(=O)(=O)(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(cc1)C#N)C)c1ccc(NC(=O)C)cc1	1	19	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)ccc1C#N	1
20	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(cc1)C#N)C)c1ccc(NC(=O)CC)cc1	1	20	FC(F)(F)c1cc(ncc1C#N)N(C(CC)C)C	1
21	ClCC(=O)Nc1ccc(S(=O)(=O)CC(O)(C(=O)Nc2cc(C(F)(F)F)c(cc2)C#N)C)cc1	1	21	Ic1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)ccc1C#N	1
22	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(NC(=O)C)cc1	1	22	O=C1CCC2=C3C(C4CC[C@](OC#CC)(O)[C@]4(C[C@@H]3c3ccc(N(C)C)cc3)C)CCC2=C1	1
23	S(=O)(=O)(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(NC(=O)C)cc1	1	23	O=C1CC[C@@]2(C3C(C4CC[C@H](O)[C@]4(CC3)C)CCC2=C1)C=C	1
24	ClCC(=O)Nc1ccc(SCC(O)(C(=O)Nc2cc(C(F)(F)F)c(N(O)O)cc2)C)cc1	1	24	O=C1CC[C@@]2([C@@H]3[C@H]([C@@H]4CC[C@H](O)[C@]4(CC3)C)CCC2=C1)C=C	1
25	ClCC(=O)Nc1ccc(S(=O)(=O)CC(O)(C(=O)Nc2cc(C(F)(F)F)c(N(O)O)cc2)C)cc1	1	25	FC(F)(F)c1cc(N2C(=O)C(NC2=O)(C)C)ccc1[N+](=O)[O-]	1
26	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C(F)(F)F)c1ccc(NC(=O)C)cc1	1	26	FC(F)(F)c1cc(nc2c1cc1CCC(Nc1c2)(C)C)C#N	1
27	ClCC(=O)Nc1ccc(SCC(O)(C(=O)Nc2cc(C(F)(F)F)c(N(O)O)cc2)C(F)(F)F)cc1	1	27	Clc1ccccc1-c1cc(C(F)(F)F)c(cc1)C#N	1
28	ClCC(=O)Nc1ccc(S(=O)(=O)CC(O)(C(=O)Nc2cc(C(F)(F)F)c(N(O)O)cc2)C(F)(F)F)cc1	1	28	FC(F)(F)c1cc(Oc2c(cccc2C)C)ccc1C#N	1
29	S(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(N(O)O)cc1)C)c1ccc(NC(=O)C(F)(F)F)cc1	1	29	Ic1cc(N2C(=O)[C@@H]3[C@@H](C4(OC3(C)[C@@H](O)C4)C)C2=O)ccc1C#N	1
30	S(CC(O)(C(Nc1cc2C=CC(Oc2cc1)=O)=C)C)c1ccc(F)cc1	0	30	O=C1N(c2cc(C)c(cc2)C)C(=O)[C@@H]2[C@H]1C1CC2C=C1	0
31	S(CC(O)(C(Nc1cc2OC(=O)C=C(c2cc1)C(F)(F)F)=C)C)c1ccc(NC(=O)C)cc1	0	31	FC(F)(F)c1ccccc1-c1cc(C(F)(F)F)c(cc1)C#N	0
32	BrCC(O)(C(Nc1cc2C=CC(Oc2cc1)=O)=C)C	0	32	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)ccc1[N+](=O)[O-]	0
33	S=C1NC(C)C(=O)N1c1cc2C=CC(Oc2cc1)=O	0	33	FC(F)(F)c1cc(O)nc2c1cc1CC[C@H](Nc1c2)CC	0
34	S=C1NC(C)C(=O)N1c1cc2OC(=O)C=C(c2cc1)C(F)(F)F	0	34	F\C=C/[C@@]12C3C(C4CC[C@H](O)[C@]4(CC3)C)CC=C1C[C@@H](O)CC2	0
35	S=C1NC(C)(C)C(=O)N1c1cc2OC(=O)C=C(c2cc1)C(F)(F)F	0	35	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4CC3CC4)C2=O)ccc1[N+](=O)[O-]	0
36	Ic1c2OC(=O)C=C(c2ccc1N1C(=O)C(NC1=S)(C)C)C(F)(F)F	0	36	FC(F)(F)C1=CC(=O)NC2C1C=C1CCC(NC1=C2)CC	0
37	C1CC(c2c(cccc2)C1(C)C)(C)C	0	37	FC(F)(F)c1cc(ccc1C#N)-c1c(OC)cccc1OC	0
38	O=C(C)c1cc2c(cc1)C(CCC2(C)C)(C)C	1	38	FC(F)(F)c1c(O)c(NC(=O)C(C)C)ccc1[N+](=O)[O-]	1
39	O=C(C)c1cc2c(cc1CC)C(CCC2(C)C)(C)C	1	39	FC(F)(F)c1cc(NC(=O)C(O)(C)C)ccc1[N+](=O)[O-]	1
40	O=C1C(C)C(c2c(cccc2)C1(C)C)(C)C	0	40	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(=CC3)[C@@H](CC=C)[C@@H](O)CC1)C=C	0
41	Oc1cc2c(cc1)C(CCC2(C)C)(C)C	1	41	S=C1N(C(=O)C(N1C)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
42	O(C(=O)C)c1cc2c(cc1)C(CCC2(C)C)(C)C	1	42	FC(F)(F)c1cc(N2C(=O)C(N(CCCCO)C2=O)(C)C)ccc1C#N	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
43	Oc1cc2c(c(C)c1C)C(CCC2(C)C)(C)C	1	43	Clc1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)cc(Cl)c1	1
44	O(C)c1cc(c2c(c1)C(CCC2(C)C)(C)C)C(=O)C	1	44	[SnH2]1N=C2C(=CC=C(N3C(=O)[C@H]4[C@H](C5(OC4(CC5)C)C)C3=O)[C-]12)C#N	1
45	Oc1cc2c(cc1C)C(C)(C)C(CC2(C)C)C	1	45	Clc1cc2nc([nH]c2cc1Cl)C(O)(C(F)(F)F)CSCC	1
46	Oc1cc2c(cc1C)C(CCC2(C)C)(C)C	1	46	FC(F)(F)C1=CC(=O)N(c2c1cc1CCC(N(c1c2)C)(C)C)C	1
47	Oc1cc(c2c(c1)C(CCC2(C)C)(C)C)C	1	47	Ic1cc(N2C(=O)[C@H]3[C@H](C4(OC3(C)[C@H](O)C4)C)C2=O)ccc1C#N	1
48	OC1Cc2c(cccc2)C(C)(C)C1C	1	48	S(C)c1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)ccc1C#N	1
49	O(C(=O)C)c1cc2c(cc1C)C(CCC2(C)C)(C)C	1	49	Oc1cc2CC[C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@@H]3c2cc1)C	1
50	O(CC(C)c1cc2c(cc1)C(C)(C)C(C)C2(C)C)C	1	50	Br\C=C/[C@@]12C3C(C4CC[C@H](O)[C@]4(CC3)C)CC=C1C[C@@H](O)CC2	1
51	O=C1CC(c2c1cc1c(c2)C(C)(C)C(C)C1(C)C)(C)C	1	51	FC(F)(F)C1=CC(=O)N(c2c1cc1c(NCCC1CC)c2)C	1
52	O=Cc1cc2c(cc1)C(C)(C)C(C)C2(C)C	1	52	FC(F)(F)c1cc(Oc2ccccc2CC)ccc1C#N	1
53	O=Cc1cc2c(cc1C)C(C)(C)C(C)C2(C)C	1	53	FC(F)(F)c1cc(ncc1C#N)N([C@H](C)c1ccccc1)C	1
54	O=C(C)c1cc2c(cc1C)C(C)(C)C(C)C2(C)C	1	54	O1C2([C@@H]3[C@H](C1(CC2)C)C(=O)N(c1cc(OC)c(cc1)-c1ocnc1)C3=O)C	1
55	O=C1c2c(c3c(cc2C)C(C)(C)C(C)C3(C)C)CC1	1	55	S(C(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
56	O=C1CCc2c1cc1c(c2)C(C)(C)C(C)C1(C)C	1	56	S(c1ccccc1OC)c1cc(OC)c(cc1)C#N	1
57	O=C1c2c(CC1Cc1ccccc1)cccc2	1	57	FC(F)(F)c1cc(O)nc2c1cc1c(NC(CC1C)(C)C)c2C	1
58	O=C1c2c(CC1C(C)C)cccc2	1	58	Fc1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	1
59	O=C1c2c(cccc2)C(C)(C)C1C	1	59	ClC=1C2=CC(=O)[C@H]3[C@H](C3)[C@@]2(C2C(C=1)C1CC[C@](OC(=O)C)(C(=O)C)[C@]1(CC2)C)C	1
60	O1CC(c2c(C1)cc1c(c2)C(C)(C)C(C)C1(C)C)C	1	60	FC(F)(F)c1cc(O)nc2c1cc1CCC(Nc1c2)(CC)C	1
61	O1CC(c2c(C1)cc1c(c2)C(C)(C)C(C)C1(C)C)CC	1	61	FC(F)(F)c1cc(O)nc2c1cc1c(NC(C=C1C)(C)C)c2C	1
62	O1CC(c2c(c3c(cc2C)C(CC3(C)C)(C)C)C1)C	1	62	S(CCC)c1cc(C(F)(F)F)c(cc1)C#N	1
63	O1CC(c2c(C1)cc1c(c2)C(CC1(C)C)(C(C)C)C)C	0	63	FC(F)(F)C1=CC(=O)N(c2c1cc1CCC(N(c1c2)C)(CC)C)C	0
64	O1CC(c2c(cc3c(c2)C(C)(C)C(C)C3(C)C)C1C)C	1	64	FC(F)(F)c1cc(O)nc2c1cc1CC(C)(C)[C@H](Nc1c2)C	1
65	O1Cc2c(CC1C)cc1c(c2)C(C)(C)C(C)C1(C)C	1	65	O(c1ccccc1C)c1cc(OC)c(cc1)C#N	1
66	FC(F)(F)C1=CC(=O)Nc2c1cc1CCC(Nc1c2)CC	1	66	FC(F)(F)c1cc(O)nc2c1cc1CC[C@H](Nc1c2)C(C)C	1
67	FC(F)(F)C1=CC(=O)Nc2c1cc1CCC(Nc1c2)C(C)C	1	67	FC(F)(F)C1=CC(=O)NC2C1C=C(N)C=C2	1
68	FC(F)(F)C1=CC(=O)Nc2c1cc1CC(C)C(Nc1c2)C	1	68	FC(F)(F)c1cc(O)nc2c1cc(N)cc2	1
69	FC(F)(F)C1=CC(=O)Nc2c1cc1CC(C)(C)C(Nc1c2)C	1	69	FC(F)(F)c1cc(O)nc2c1cc1[C@@H]3[C@H](Nc1c2)C(CCC3)(C)C	1

	Trainingset-1 Smiles	activity		Training set-2 Smiles	activity
70	FC(F)(F)C1=CC(=O)Nc2c1cc1CC(C)(C)C(Nc1c2)CC	1	70	FC(F)(F)c1cc(O)nc2c1cc1c(NC(CC1C)(C)C)c2	1
71	FC(F)(F)C1=CC(=O)Nc2c1cc1CC(CC)(CC)C(Nc1c2)CC	1	71	FC(F)(F)c1cc(ccc1C#N)C1CCCCC1O	1
72	FC(F)(F)C1=CC(=O)Nc2c1cc1CC(C(C)C)C(Nc1c2)C(C)(C)C	1	72	FC(F)(F)C1=CC(=O)N(c2c1cc1CCC3(N(CCC3)c1c2)C)C	1
73	FC(F)(F)C1=CC(=O)Nc2c1cc1CC(C(C)C)C(Nc1c2C)C(C)(C)C	1	73	FC(F)(F)C1=CC(=O)N(c2c1cc1[C@@H]3[C@H](N(c1c2)C)C(CCC3)(C)C)C	1
74	FC(F)(F)C1=CC(=O)Nc2c1cc1CCC(Nc1c2)(C)C	1	74	FC(F)(F)C1=CC(=O)NC2C1C=C(NCC(C)(C)C)C=C2	1
75	FC(F)(F)C1=CC(=O)Nc2c1cc1CCC(Nc1c2)(CC)C	1	75	FC(F)(F)C1=CC(=O)NC2C1C=C1C(NC(C=C1)(C)C)=C2	1
76	c12c(cccc1)c(c1c(cccc1)c2C#Cc1ccccc1)C#Cc1ccccc1	0	76	FC(F)(F)c1cc(O)nc2c1cc1CCC(Nc1c2)(C)C	0
77	Clc1ccccc1Nc1nc(Cl)nc(Cl)n1	0	77	FC(F)(F)c1cc(O)nc2c1cc1CCC(Nc1c2)(CC)CC	0
78	Clc1cc(NC(OC(C)C)=O)ccc1	0	78	FC(F)(F)c1cc(O)nc2c1cc1c(NC(CC1CC)(C)C)c2	0
79	Clc1cc(NC(OCC#CCCl)=O)ccc1	0	79	FC(F)(F)c1cc(O)nc2c1cc1c(NC(C=C1C)(C)C)c2	0
80	S(C)c1nc(nc(n1)NCC)NCC	0	80	S(C(C)C)c1cc(OC)c(cc1)C#N	0
81	OC1CCC2C3C(=C4C(=CC(=O)CC4)CC3)C=CC12C	1	81	Clc1cc(Oc2ccccc2C)ccc1C#N	1
82	OCCN(CCCC)CCCC	1	82	Clc1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	1
83	Clc1ccc(cc1)\C(=C\Cl)\c1ccc(Cl)cc1	0	83	Clc1nc2c(cc3c(NC(C=C3C)(C)C)c2)c(c1)C(F)(F)F	0
84	ClC12C3C(C4OC4C3Cl)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	1	84	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC)=C2	1
85	S(P(OC)(=O)N)C	0	85	FC(F)(F)c1cc(O)nc2c1cc1CC(C)C(Nc1c2)(C)C	0
86	Clc1cc(ccc1NC(C(C)C)C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O)C(F)(F)F	0	86	O=C1N(c2ccc([N+](=O)[O-])cc2)C(=O)[C@@H]2[C@H]1C1CC2C=C1	0
87	O(C(=O)CCCCC(OCC(CCCC)CC)=O)CC(CCCC)CC	0	87	FC(F)(F)C1=CC(=O)N(c2c1cc1c(NC(CC1C)(C)C)c2C)C	0
88	ClC12C3C(COS(OC3)(=O)=O)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	0	88	Fc1nc2c(cc3c(NC(C=C3C)(C)C)c2)c(c1)C(F)(F)F	0
89	O=C1C2(CCC(C2(C)C)C1=O)C	0	89	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCCC)=C2	0
90	O(C)c1ccc(cc1)CCC(=O)C	0	90	FC(F)(F)c1cc(O)nc2c1cc1CC[C@H](Nc1c2)CCC	0
91	Oc1ccc(cc1)CCCCCCCCC	0	91	FC(F)(F)c1cc(O)nc2c1cc1C[C@H](C)[C@H](Nc1c2)C	0
92	n1cc(ccc1C)CC	0	92	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C=C	0
93	o1cc(cc1Cc1ccccc1)COC(=O)C1C(C)(C)C1\C=C(\C)/C	0	93	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)CC	0
94	O(C(=O)CCCCCCCCCCC(OCC)=O)CC	0	94	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C=C	0
95	ClC=1C2=CC(OCC2(C2C(C3CCC(OC(=O)C)(C(=O)C)C3(CC2)C)C=1)C)=O	1	95	FC(F)(F)C1=CC(=O)N(c2c1cc1c(NC(C=C1C)(C)C)c2)C	1
96	n1c2ncc(cc2n(C)c1N)-c1ccccc1	0	96	FC(F)(F)c1cc(O)nc2c1cc(N(C)C)cc2	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
97	Sc1ccc(cc1)C	1	97	FC(F)(F)c1cc(Oc2ccccc2O)ccc1C#N	1
98	OC1CCN(CC1)C	0	98	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCC)=C2	0
99	Clc1ccc(S)cc1	0	99	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)ccc1	0
100	O(C(=O)Nc1[nH]c2c(n1)cccc2)C	0	100	FC(F)(F)c1cc(O)nc2c1cc1CCC3(Nc1c2)CCCCC3	0
101	Clc1ccc(cc1)CCC(O)(C(C)(C)C)Cn1ncnc1	1	101	S(CC1CC1)c1cc(C(F)(F)F)c(cc1)C#N	1
102	Oc1ccc(O)cc1-c1ccccc1	1	102	FC(F)(F)C1=CC(=O)N(c2c1cc1CCC(Nc1c2)(C)C)C	1
103	N(C(C)C)C(C)C	0	103	FC(F)(F)C1=CC(=O)N(c2c1cc1[C@@H]3[C@H](N(c1c2)C)CCCC3)C	0
104	O1C2CC3C4C(CCC3(C)C12C(=O)C)C1(C(=CC(=O)CC1)CC4)C	1	104	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)ccc1	1
105	O(C(=O)CCCCCCCCC)CC	0	105	Oc1nc2c(cc3c(NC(C=C3C)(C)C)c2)c(c1)C	0
106	O=Cc1nc(ccc1)C	0	106	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1OC	0
107	Oc1cc2CCCCc2cc1	1	107	FC(F)(F)c1cc(ccc1C#N)[C@@H]1CCCC[C@@H]1O	1
108	O(C(OC)c1ccccc1)C	1	108	O(C)c1cc(ccc1C#N)-c1ccccc1O	1
109	Clc1cc(cc(Cl)c1O)-c1ccccc1	1	109	O=C1C=C2CC[C@H]3[C@@H]4CC[C@H](C(=O)C)[C@]4(CC[C@@H]3[C@]2(C=C1)C)C	1
110	Clc1ccc(cc1)C(O)(C(Cl)(Cl)Cl)c1ccc(Cl)cc1	1	110	Clc1cc(N2C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C2=O)cnc1C#N	1
111	Oc1c(O)cc(cc1O)C(OCCCCCCCCCCCC)=O	0	111	S(c1c(cc(cc1C)C)C)c1cc(OC)c(cc1)C#N	0
112	Clc1c(Cl)c(Cl)c2c(c1Cl)C(OC2=O)=O	0	112	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1O	0
113	ClC1=C(Cl)C(=O)c2c(cccc2)C1=O	1	113	Brc1c2c(ccc1)c(N1C(=O)[C@H]3[C@H](C4(OC3(CC4)C)C)C1=O)ccc2[N+](=O)[O-]	1
114	O(C(=O)c1ccccc1C(OCC(CCCC)CC)=O)CC(CCCC)CC	0	114	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(=CC(CC1)=C)CC3)C	0
115	OC1C2C(C3CCC(C(=O)COC(=O)C)C3(C1)C)CCC1=CC(=O)CCC12C	1	115	S(CCC)c1cc(OC)c(cc1)C#N	1
116	O(C(=O)c1ccccc1O)C1CC(CC(C1)C)(C)C	1	116	Clc1cc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)cc(Cl)c1	1
117	Clc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	0	117	FC(F)(F)C1=CC(=O)NC2C1C=C(NC1CCCC1)C=C2	0
118	S=P(OC1=NN(C(=O)C=C1)c1ccccc1)(OCC)OCC	0	118	FC(F)(F)C1=CC(=O)NC2C1C=CC(N(CC(F)(F)F)CC)=C2	0
119	O(C)c1ccc(cc1)C(O)C(=O)c1ccc(OC)cc1	0	119	FC(F)(F)c1cc(O)nc2c1cc1CC(C)(C)[C@H](Nc1c2)CC	0
120	C1CCc2c(C1)cccc2	0	120	FC(F)(F)c1cc(O)nc2c1cc1CCC(Nc1c2)(CCC)C	0
121	O(C)c1cc(ccc1N)-c1cc(OC)c(N)cc1	1	121	BrC1CC(CCC1Br)C(Br)CBr	1
122	Clc1cc(NC(OC(C=C)(C(O)=O)C)=O)cc(Cl)c1	1	122	Clc1cc(ccc1Sc1ccccc1C)C#N	1
123	Clc1cccc(Cl)c1C#N	0	123	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC1CCC1)=C2	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
124	O=C1Nc2c(N=C1)cccc2	0	124	FC(F)(F)c1cc(O[C@H]2CCCC[C@@H]2C#N)ccc1C#N	0
125	c12c(cc3c(c1)cccc3)cccc2	1	125	S(c1c(cccc1C)C)c1cc(OC)c(cc1)C#N	1
126	Oc1ccc(cc1)C(OCC)=O	0	126	FC(F)(F)C1=CC(=O)Nc2c1cc1N(CC(F)(F)F)[C@H](COc1c2)c1ccc(cc1)C	0
127	Clc1cc(Cl)ccc1Cl	0	127	FC(F)(F)C1=CC(=O)Nc2c1c1OC[C@H](N(c1cc2)CC1CC1)c1ccccc1	0
128	O(C(=O)C1C(C)(C)C1\C=C(/C(OC)=O)\C)C1CC(=O)C(C\C=C\C=C)=C1C	0	128	FC(F)(F)c1cc(nc2c1cc1CCC(N(c1c2)C)(C)C)C#N	0
129	S=P(Oc1cc(C)c(N(O)O)cc1)(OC)OC	1	129	s1c(-c2cc3c(NC(CC3C)(C)C)cc2)c(cc1C#N)C	1
130	Clc1nc(nc(n1)NCC)NCC	0	130	FC(F)(F)C1=CC(=O)N(c2c1cc1c(N(C)C(C=C1C)(C)C)c2)C	0
131	O(C(=O)CCCCCCCCC(OCC(CCCC)CC)=O)CC(CCCC)CC	0	131	FC(F)(F)c1cc(O)nc2c1cc1CC(CC)C(Nc1c2)(C)C	0
132	O(C(=O)C)c1ccccc1	0	132	FC(F)(F)C1=CC(=O)N(c2c1cc1CCC(Nc1c2)(CC)C)C	0
133	NCCCCCCCCCCCC	0	133	FC(F)(F)C1=CC(=O)N(c2c1cc1[C@@H]3[C@H](Nc1c2)CC[C@@H](C3)C)C	0
134	OC1(C(=O)CO)C2(CC(=O)C3C(C2CC1C)CCC1=CC(=O)C=CC13C)C	1	134	FC(F)(F)c1cc(O)nc2c1cc1CC(CC)(CC)[C@H](Nc1c2)CC	1
135	S(=O)(=O)(c1c(cc(cc1C)C)C)C1=N[CH-][NH+](N1)C(=O)N(CC)CC	0	135	O1c2c(cc(OC)cc2)-c2c(c3c(NC(=O)C=C3C)cc2)C1=O	0
136	OC1CC2CCC=3C4CCC(C(CC\C=C(\C)/C)C)C4(CCC=3C2(CC1)C)C	0	136	FC(F)(F)c1cc(N2C(=O)C(N(CC#CCOCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	0
137	Oc1c(cc(cc1C(C)(C)C)C)C(C)(C)C	0	137	FC(F)(F)c1cc(N2C(=O)C(N(OCC#CCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	0
138	c12c3c4ccc1cccc2ccc3ccc4	1	138	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)[C@@H]2[C@H]1C1CCCC2C=C1	1
139	S(=O)(CCCC(F)(F)C(F)(F)F)CCCCCCCCCC1C2C3CCCC(O)C3(CCC2c2c(C1)cc(O)cc2)C	0	139	S(CCCC)c1cc(C(F)(F)F)c(cc1)C#N	0
140	Clc1ccccc1-c1ccccc1Cl	1	140	FC(F)(F)C1=CC(OC2C1C=CC(NCC(F)(F)F)=C2)=O	1
141	O(C)c1cc(O)c(cc1)C(=O)c1ccccc1	1	141	FC(F)(F)C1=CC(=O)Nc2c1cc1N(CC(F)(F)F)[C@@H](COc1c2)Cc1ccccc1	1
142	S(P(SCCC)(OCC)=O)CCC	0	142	FC(F)(F)c1cc(O)nc2c1cc1c(N[C@H]3[C@@]1(CCCC3(C)C)C)c2	0
143	O(C(c1ccccc1)c1ccccc1)C1CCN(CC1)C	0	143	FC(F)(F)C1=CC(=O)N(c2c1cc1c(N([C@@H]3CCCC[C@]13C)C)c2)C	0
144	ClC(Cl)(Cl)SN1C(=O)C2C(CC=CC2)C1=O	0	144	FC(F)(C(F)(F)F)[C@]1(O)CC[C@H]2[C@H]3C(=C4C(=CC(=O)CC4)CC3)[C@H](C[C@]12C)c1ccc(cc1)C(=O)C	0
145	ClC(Cl)(Cl)SN1C(=O)c2c(cccc2)C1=O	0	145	FC(F)(F)C1=CC(=O)N(c2c1cc1[C@@H]3[C@H](N(c1c2)C)CC[C@@H](C3)C)C	0
146	FC(F)(F)c1cc(NC(=O)C(C)C)ccc1N(O)O	1	146	FC(F)(F)C1=CC(=O)N(c2c1cc1CCNc1c2)C	1
147	Clc1cc(-n2nc(-c3ccc(F)cc3F)c(C#N)c2N)ccc1	0	147	FC(F)(F)C1=CC(=O)N(c2c1cc1c(N(C)[C@H](C)C1(C)C)c2)C	0
148	Oc1cc2c(cc1)cccc2	0	148	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC(C)C)=C2	0
149	c12c(cc3c(c1)cc1c(c3)cccc1)cc1c(c2)cccc1	0	149	S(c1ccccc1OC)c1cc(C(F)(F)F)c(cc1)C#N	0
150	N(=C(\N)/N)/CCCCCCCCNCCCCCCCC\N=C(\N)/N	0	150	FC(F)(F)c1cc(ccc1C#N)-c1ccc(OC)cc1	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
151	Clc1ccccc1C1OC1(n1ncnc1)c1ccc(F)cc1	1	151	FC(F)(F)c1cc(O)nc2c1cc1c(NC(C)(C)C(C)(C)C1C)c2	1
152	S=P(Oc1nc2c(nc1)cccc2)(OCC)OCC	1	152	O(c1cc(OC)c(cc1)C#N)c1cc(ccc1)C	1
153	Oc1cc(O)ccc1CCCCCC	1	153	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCC(F)(F)C(F)(F)F)=C2	1
154	O(C(=O)Nc1cc(ccc1)C)c1cc(NC(OC)=O)ccc1	0	154	FC(F)(F)c1cc(O)nc2c1cc1C=CC(Nc1c2)(C)C	0
155	ClC=1C2=CC(=O)C=CC2(C2C(C3CCC(OC(=O)C)(C(=O)C)C3(CC2)C)C=1)C	1	155	FC(F)(F)c1cc(OC2CCCCC2C#N)ccc1C#N	1
156	OCC(CC)C	0	156	FC(F)(F)c1cc(O[C@@H]2CCCC[C@@H]2C#N)ccc1C#N	0
157	Clc1ncc(cc1)CN/1CCN\C\1=N/N(O)O	0	157	Fc1cc(cc-2c1OCc1c3c(NC(C=C3C)(C)C)ccc1-2)C#N	0
158	O(C(=O)C)C1C(C2CC1(CC2)C)(C)C	0	158	O1Cc2c3c(NC(C=C3C)(C)C)ccc2-c2cc([N+](=O)[O-])ccc12	0
159	S=C1NC(=O)N(C=C1)C1OC(CO)C(O)C1O	0	159	O=C1N(c2ccc([N+](=O)[O-])cc2)C(=O)N2[C@@H]1[C@H]1N(C[C@@H]2C1)C(=O)c1ccc(cc1)	0
160	O(C)c1ccc(cc1)CC=C	0	160	FC(F)(F)c1cc(ccc1C#N)-c1ccc(O)cc1	0
161	O(C(=O)CCC(OCCCC)=O)CCCC	0	161	O1c2c(-c3c(c4c(NC(=O)C=C4C)cc3)C1C)c(OC)ccc2	0
162	Oc1cc(O)c(O)cc1C(=O)CCC	0	162	FC(F)(F)c1cc(N2C(=O)C(N(CC#CCOCc3cc(C(=O)NCN[C@H]4CC(O[C@@H](C)[C@H]4O)OC4c5c(C[C@](O)(C4)C(=O)CO)c(O)c4c(C(=O)c	0
163	OC(=O)CCCCCCCCCCC	0	163	O1Cc2c3c(NC(CC3C)(C)C)ccc2-c2cc([N+](=O)[O-])ccc12	0
164	ClC12C3(Cl)C4(Cl)C5(Cl)C(Cl)(C1(Cl)C4=O)C2(Cl)C(Cl)(Cl)C35Cl	0	164	S(=O)(C)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	0
165	OC(C(CO)(C)C)C(C)C	0	165	S(CC1CCC1)c1cc(C(F)(F)F)c(cc1)C#N	0
166	s1c(nnc1NS(=O)(=O)c1ccc(N)cc1)C	0	166	Clc1cc(ccc1Oc1ccccc1C)C#N	0
167	s1cc(nc1)-c1[nH]c2c(n1)cccc2	0	167	FC(F)(F)C1=CC(=O)Nc2c1c1OC[C@H](N(c1cc2)CC(F)(F)F)c1ccccc1	0
168	Fc1ccccc1N(O)O	0	168	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCCCC)=C2	0
169	c12c(cccc1)c(c1c(cccc1)c2-c1ccccc1)-c1ccccc1	0	169	Brc1ccc(N2C(=O)[C@H]3[C@H](C4CCC3C=C4)C2=O)cc1C	0
170	OC(=O)c1ccc(N)cc1	0	170	FC(F)(F)c1cc(O)nc2c1cc1c(NC(C)(C)C(C)=C1C)c2	0
171	O=C1/C(/C2CCC1(C)C2(C)C)=C/c1ccccc1	0	171	O[C@H]1CC[C@H]2[C@H]3[C@@H]([C@@]45[C@H]([C@@H](CCC4)[C@@H](O)CC5)CC3)CC[C@]12C	0
172	n1ccccc1CNCc1ncccc1	0	172	FC(F)(F)c1cc(Oc2ccc(cc2OC)C)ccc1C#N	0
173	Clc1nc(nc(N)c1)N	0	173	O1C2([C@@H]3[C@H](C1(CC2)C)C(=O)N(C=1n2c(ncc2)C=CC=1N)C3=O)C	0
174	O1c2c(CC1(C)C)cccc2OC(=O)NC	0	174	Brc1nc2c(cc3CCC(N(c3c2)C)(C)C)c(c1)C(F)(F)F	0
175	ClCC(=O)N(COC)c1c(cccc1CC)CC	1	175	Brc1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	1
176	Nc1c2c3c4c(cc2)cccc4ccc3cc1	1	176	FC(F)(F)C1=CC(=O)Nc2c1c1OC[C@H](Nc1cc2)C	1
177	O(C)c1nc(nc(n1)NC(C)C)NC(C)C	0	177	FC(F)(F)C1=CC(=O)N(c2c1cc1c(NC(CC1C)(C)C)c2)C	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
178	O=C1CCCCC1C(OCC)=O	0	178	FC(F)(F)c1cc(O)nc2c1cc1c(NCCC1(CC)CC)c2	0
179	Clc1c(cccc1Cl)-c1ccccc1	1	179	FC(F)(F)c1cc(ncc1C#N)N1CCC(CC1)C	1
180	OC1(CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C)C(O)C	1	180	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCC(C)C)=C2	1
181	O1CC1COc1ccc(cc1)C(C)(C)c1ccc(OCC2OC2)cc1	0	181	S(C)c1ccccc1Oc1cc(C(F)(F)F)c(cc1)C#N	0
182	S(\C(=N\OC(=O)NC)\C)C	0	182	FC(F)(F)C1=CC(=O)NC2C1C=C(C)C(N)=C2	0
183	ClC=1C(=O)N(N=CC=1N)c1ccccc1	0	183	FC(F)(F)c1cc(N2C(=O)C(N(CCOCCOCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	0
184	S(P(Sc1ccccc1)(OCC)=O)c1ccccc1	0	184	FC(F)(F)c1cc(N2C(=O)C(N(OCCOCCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	0
185	Clc1cc(cc(Cl)c1O)C(OCC)=O	0	185	FC(F)(F)C1=CC(=O)Nc2c1c1OCc3c(-c1cc2)c(OC)ccc3	0
186	Clc1cc2Oc3cc(Cl)c(Cl)cc3Oc2cc1Cl	0	186	FC(F)(F)c1cc(ncc1C#N)N(CCC)CCC	0
187	O(C(=O)Nc1nc2c(n1C(=O)NCCCC)cccc2)C	0	187	FC(F)(F)c1cc(O)nc2c1cc1CC(C)(C)C(Nc1c2)(C)C	0
188	O1C(CCCC(=O)CCC\C=C/c2c(C1=O)c(O)cc(O)c2)C	1	188	FC(F)(F)c1cc(Oc2ccccc2OC)ccc1C#N	1
189	Oc1ccccc1-c1ccccc1O	1	189	FC(F)(F)C1=CC(=O)Nc2c1c1OC[C@H](Nc1cc2)CC	1
190	Brc1ccc(cc1)C(O)(C(OC(C)C)=O)c1ccc(Br)cc1	1	190	O=C1N(c2cc3c(cc2)cccc3)C(=O)[C@@H]2[C@H]1C1CC2CC1	1
191	O=N\C(=C\1/NC=CC=C/1)\c1ccccc1	0	191	S(CCCC)c1cc(OC)c(cc1)C#N	0
192	Clc1cc(Cl)ccc1Oc1ccc(N(O)O)cc1	1	192	F\C=C\[C@@]12C3C(C4CC[C@H](O)[C@]4(CC3)C)CC=C1C[C@@H](O)CC2	1
193	Clc1cc(Cl)cc(Cl)c1Oc1ccc(N(O)O)cc1	1	193	O=C1CCC2=C3C(C4CC[C@H](O)[C@]4(C[C@@H]3CCCCCCCCCC)C)CCC2=C1	1
194	Oc1cc(C)c(cc1C(C)(C)C)C(CC(C)c1cc(C(C)(C)C)c(O)cc1C)c1cc(C(C)(C)C)c(O)cc1C	0	194	O[C@H]1CC[C@H]2[C@H]3[C@@H]([C@]45CC[C@H](O)[C@H](CC=C4)C5=CC3)CC[C@]12C	0
195	S=P(Oc1noc(c1)-c1ccccc1)(OCC)OCC	0	195	O1C2([C@@H]3[C@H](C1(CC2)C)C(=O)N(C=1c2n(ncc2)C(=CC=1)C#N)C3=O)C	0
196	Brc1c(Oc2ccc(Br)cc2Br)c(Br)c(Br)c(Br)c1Br	1	196	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)CC	1
197	Brc1c(Oc2cc(Br)c(Br)cc2)c(Br)c(Br)c(Br)c1Br	0	197	Brc1cc(F)c2OCc3c4c(NC(C=C4C)(C)C)ccc3-c2c1	0
198	Clc1c(C#N)c(Cl)c(Cl)c(Cl)c1C#N	0	198	FC(F)(F)c1cc(N2C(=O)C(N(CC#CCOCc3cc(C(=O)NCNC4CC(O[C@@H](C)[C@H]4O)OC4c5c(C[C@](O)(C4)C(=O)CO)c(O)c4c(C(=O)c6c(c	0
199	c12c3c4c5c6c1c(ccc2ccc3ccc4ccc5)ccc6	0	199	FC(F)(F)c1cc(N2C(=O)C(N(CC#CCOCc3cc(C(=O)NCN[C@H]4CC(O[C@@H](C)[C@H]4O)OC4c5c(C[C@](O)(C4)C(=O)CO)c(O)c4c(C(=O)c	0
200	Clc1nc(nc(n1)NCC)NC(C)C	0	200	O(CCCCC(O)=O)c1ccc(cc1)CCCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
201	Oc1ccc(O)cc1C(C)(C)C	1	201	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCC(C)(C)C)=C2	1
202	Clc1ccc(Oc2ccc(NC(=O)N(C)C)cc2)cc1	0	202	FC(F)(F)c1cc(O[C@@H]2C(CN(Cc3ccccc3)C2=O)(C)C)ccc1C#N	0
203	C1c2c3c4c1cccc4ccc3ccc2	1	203	FC(F)(F)c1cc(Oc2cc(ccc2OC)C)ccc1C#N	1
204	S(C)c1c(cc(OC(=O)NC)cc1C)C	1	204	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(nc2)C(F)(F)F)C)ccc1C#N	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
205	c-12c(-c3c4c-1cccc4ccc3)ccc1c2cccc1	1	205	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCC(=O)N(C)C)C1)C	1
206	c-12c(-c3c4c-1cc1c(c4ccc3)cccc1)cccc2	0	206	S(=O)(CCCC(F)(F)C(F)(F)F)CCCCCCCCCC[C@H]1[C@H]2[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]2[C@@]2(C(CC(=O)CC2)C1)C)C	0
207	Clc1cc(ccc1)-c1cc(Cl)ccc1	1	207	FC(F)(F)C1=CC(=O)NC2C1C=C(NCC)C=C2	1
208	Clc1ccc(cc1)-c1ccc(Cl)cc1	0	208	FC(F)(F)C1=CC(=O)Nc2c1c1OC[C@H](Nc1cc2)CCC	0
209	Clc1c(-c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl)c(Cl)c(Cl)c(Cl)c1Cl	0	209	O(c1ccccc1C)c1ccc(cc1OC)C#N	0
210	Clc1ccccc1-c1ccccc1	1	210	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)ccc1[N+](=O)[O-]	1
211	Clc1ccc(cc1)-c1ccccc1	0	211	Br\C=C\[C@@]12C3C(C4CC[C@H](O)[C@]4(CC3)C)CC=C1C[C@@H](O)CC2	0
212	c-12c(-c3c4c-1cccc4ccc3)cccc2	1	212	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cccnc2)C)ccc1C#N	1
213	c-12c(-c3c4c-1cccc4ccc3)cc1c(c2)cccc1	1	213	FC(F)(F)c1cc(Oc2ccccc2OCC)ccc1C#N	1
214	S=P(Oc1ccc(N(O)O)cc1)(OCC)c1ccccc1	1	214	O(CCCCCC(O)=O)c1cc(ccc1)CCCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	1
215	Brc1cc(Cl)c(OP(=S)(OC)OC)cc1Cl	0	215	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(CC)CC)C1)C	0
216	S(C)C1=NN=C(C(C)(C)C)C(=O)N1N	0	216	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N2CCCCC2)C1)C	0
217	O=C(C)c1cc2c(cc1C)C(C)(C)C(CC2(C)C)C	1	217	s1c2c(nc1)ccc(N)c2N1C(=O)[C@H]2[C@H](C3(OC2(CC3)C)C)C1=O	1
218	n1ccc(cc1)Cc1ccccc1	1	218	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC(C=C)(C)C)=C2	1
219	Clc1c(-c2c(Cl)c(Cl)cc(Cl)c2Cl)c(Cl)c(Cl)cc1Cl	1	219	Clc1ccccc1Oc1cc(C(F)(F)F)c(cc1)C#N	1
220	P(OC)(OC)(O\C(=C\C(=O)NC)\C)=O	0	220	FC(F)(F)c1cc(ncc1C#N)N(C(C)c1ccccc1)C	0
221	Brc1cc(Cl)c(OP(=S)(OC)c2ccccc2)cc1Cl	1	221	FC(F)(F)c1cc(Oc2cc(ccc2O)C)ccc1C#N	1
222	c12c(c3c(c4c1cccc4)cccc3)cccc2	0	222	O1C(c2cc(ccc2NC1=O)-c1cc(oc1)C#N)(C)C	0
223	c12c(c3c(cc1)cccc3)ccc1c2cccc1	1	223	FC(F)(F)c1cc(N2C(=O)[C@@H]3N(C4CC3CC4)C2=O)ccc1	1
224	S(CC)C(=O)N1CCCCCC1	0	224	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(nc2)C)C)ccc1C#N	0
225	P(Oc1cc(C)c(N(O)O)cc1)(OC)(OC)=O	0	225	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccncc2)C)ccc1C#N	0
226	S(C(C(=O)NC)C)CCSP(OC)(OC)=O	0	226	O1c2c(-c3c(c4c(NC(=O)C=C4C)cc3)C1CC=C)c(OC)ccc2	0
227	O1c2c(OC1(C)C)cccc2OC(=O)NC	0	227	Brc1ccc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)cc1C(F)(F)F	0
228	Clc1cc2OC(=O)N(c2cc1)CSP(=S)(OCC)OCC	1	228	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1C	1
229	O(C(=O)N(C)C)c1nc(nc(C)c1C)N(C)C	0	229	Fc1cc(F)cc-2c1OCc1c3c(NC(CC3C)(C)C)ccc1-2	0
230	S(\C(=N\OC(=O)NC)\C(=O)N(C)C)C	0	230	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC1CCCC1)=C2	0
231	c1cc(ccc1-c1ccccc1)C=C	0	231	Clc1cc(N2C(=O)[C@H]3[C@H](C4CCC3C=C4)C2=O)ccc1Cl	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
232	S=C(Nc1ccccc1NC(=S)NC(OC)=O)NC(OC)=O	0	232	s1c(-c2cc3c(NC(C)(C)C(=O)C3(C)C)cc2)c(cc1C#N)C	0
233	Clc1cc(cc(Cl)c1)C(=O)NC(C#C)(C)C	0	233	FC(F)(F)c1cc(NC(=O)C(C)C)ccc1[N+](=O)[O-]	0
234	S(P(=S)(OCCC)OCCC)CC(=O)N1CCCCC1C	1	234	O=C(Nc1cc(C#N)c([N+](=O)[O-])cc1)C(C)C	1
235	O1C(CO)C(O)C(O)C1O	0	235	FC(F)(F)c1cc(O)nc2c1cc1c(N(C)C(C=C1C)(C)C)c2	0
236	OC(C)c1cc(N)ccc1	0	236	FC(F)(F)c1cc(Oc2c(OC)cccc2OC)ccc1C#N	0
237	OC1CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)C=C3)C	0	237	Fc1cccc(OC)c1Oc1cc(C(F)(F)F)c(cc1)C#N	0
238	O(C(=O)C=C)CCCCCC	0	238	FC(F)(F)c1cc(O)nc2c1cc1c(NC(CC1C(C)C)(C)C)c2	0
239	O(C)c1cccc(CCCC)c1O	0	239	S=C1N(C(=O)C(N1CCCS(=O)(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
240	S1(=O)(=O)Nc2c(cccc2)C(=O)N1C(C)C	0	240	FC(F)(F)c1cc(O)nc2c1cc1c(NC(C)(C)[C@@H](C)[C@H]1C)c2	0
241	S=P(Oc1ccccc1C(OC(C)C)=O)(OCC)NC(C)C	1	241	FC(F)(F)c1cc(ncc1C#N)N1CC2C(CCCC2)CC1	1
242	O=C1NC(=O)c2c1cc1c(c2)C(=O)NC1=O	0	242	O[C@]1(CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C=C)C	0
243	ClCC1OCCO1	0	243	S(CC1CC1)c1cc(OC)c(cc1)C#N	0
244	ClC(Cl)(Cl)c1nc(sn1)OCC	0	244	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(OC)nc2)C)ccc1C#N	0
245	OC(C(O)C(O)C=O)C(O)CO	0	245	O[C@]1(CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C=C)C	0
246	S(P(OC(C)C)(OC(C)C)=O)Cc1ccccc1	0	246	S(=O)(=O)(Nc1cccc(N(Cc2ccccc2)Cc2ccccc2)c1C)C	0
247	[nH]1c2c(c3ccc(nc13)N)cccc2	0	247	Cl\C=C/[C@@]12C3C(C4CC[C@H](O)[C@]4(CC3)C)CC=C1C[C@@H](O)CC2	0
248	Clc1cc(Cl)cc(Cl)c1Oc1ccc(N)cc1	1	248	Fc1nc2c(cc3CCC(N(c3c2)C)(C)C)c(c1)C(F)(F)F	1
249	O(C(=O)NC)c1ccccc1C(C)C	0	249	Fc1ccccc1CN1CC(C)(C)[C@@H](Oc2cc(C(F)(F)F)c(cc2)C#N)C1=O	0
250	S=P(Oc1ccc(cc1)C#N)(OC)OC	1	250	S1CCCSC1=C1Oc2c(cc(F)cc2)-c2c1c1c(NC(C=C1C)(C)C)cc2	1
251	C(CCCCC)(C)C	0	251	S(c1ccccc1F)c1cc(C(F)(F)F)c(cc1)C#N	0
252	ClC(Cl)(Cl)C(NC=O)N1CCN(CC1)C(NC=O)C(Cl)(Cl)Cl	0	252	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cccnc2OC)C)ccc1C#N	0
253	Oc1ccc(cc1)C(CC(C)(C)C)(C)C	1	253	S=C1N(C(=O)C(N1CCCS(=O)(=O)NC)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
254	Clc1c(Cl)c(Cl)c2c(c1Cl)C(OC2)=O	0	254	S(C1CCC1)c1cc(C(F)(F)F)c(cc1)C#N	0
255	S1(=O)(=O)N=C(OCC=C)c2c1cccc2	0	255	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4CCC3C=C4)C2=O)ccc1	0
256	S(O)(=O)(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	0	256	Fc1cc(cc2c1NC(=O)C2(C)C)-c1n(C)c(cc1)C#N	0
257	Clc1ccc(cc1)CSC(=O)N(CC)CC	1	257	FC(F)(F)C1=CC(=O)N(c2c1cc1CCN(c1c2)C)C	1
258	O(C(=O)c1ccccc1C(OCCCCCCC(C)C)=O)CCCCCCC(C)C	0	258	FC(F)(F)c1cc(Oc2c(cc(cc2C)C)C)ccc1C#N	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
259	Clc1ccc(OC)c(/C(/OC(=O)c2ccccc2)=N/OCC)c1OC	0	259	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(nc2)C#N)C)ccc1C#N	0
260	Clc1cc(Cl)c(Cl)nc1OP(=S)(OCC)OCC	0	260	FC(F)(F)c1cc(ccc1C#N)-c1ccc(cc1)C	0
261	S=P(Oc1nc(nc(c1)C)N(CC)CC)(OC)OC	0	261	Fc1cc(F)ccc1NC(=O)N1C[C@@H](N(C[C@H]1C)c1cc(C(F)(F)F)c(cc1)C#N)C	0
262	S=P(Oc1ccc(N(O)O)cc1)(OC)OC	1	262	Fc1ccc(OC(=O)N2[C@@H]3[C@H]4N([C@@H](C3)C2)C(=O)N(c2cc(C(F)(F)F)c(cc2)C#N)C4=O)cc1	1
263	S(CC)CSP(=S)(OCC)OCC	0	263	O(c1cc(OC)c(cc1)C#N)c1ccc(cc1)C	0
264	S(CCSP(=S)(OCC)OCC)CC	0	264	s1c(-c2cc3c(NC(C=C3C)(C)C)cc2)c(cc1C#N)C	0
265	Clc1cc(Cl)c(Cl)cc1OP(=S)(OC)OC	1	265	FC(F)(F)c1cc(Oc2ccccc2CCC)ccc1C#N	1
266	S(Cc1ccccc1OC(=O)NC)CC	0	266	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C\c1ncccc1C	0
267	O(C(=O)C)C1(CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C)C(=O)C	0	267	Clc1cc(ccc1F)-c1cc2c(NC(OC2(C)C)=O)cc1	0
268	S(P(OC)(=O)NC(=O)C)C	0	268	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC(CC)(C)C)=C2	0
269	ClC12C3C(C4CC3C=C4)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	0	269	FC(F)(F)c1cc(ncc1C#N)N1CCCCC1	0
270	OC1CC2=CCC3C4CCC(C(CC\C=C(\C)/C)C)C4(CCC3C2(CC1)C)C	0	270	FC(F)(F)c1cc(N2CC(N(CC2C)C(=O)Nc2ccc(nc2)C(F)(F)F)C)ccc1C#N	0
271	Fc1cc(F)cc(F)c1N(O)O	0	271	Fc1cc(cc(c1)C#N)-c1cc2c(NC(C=C2C)(C)C)cc1	0
272	Clc1cc(Cl)c(Cl)cc1-c1cc(Cl)c(Cl)cc1	1	272	FC(F)(F)c1c2c(ncc1)cc1NC(CCc1c2)(C)C	1
273	Fc1ccccc1OC	0	273	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C	0
274	Fc1ccccc1-c1ccccc1	1	274	O1CC(=O)Nc2c(cc(cc2)-c2n(C)c(cc2)C#N)C1(C)c1occc1	1
275	O(C(=O)C(C)=C)c1ccc(cc1)C(C)(C)c1ccc(OC(=O)C(C)=C)cc1	1	275	s1cc(cc1)\C=C/1\Oc2c(cc(F)cc2)-c2c\1c1c(NC(C=C1C)(C)C)cc2	1
276	Brc1cc(Br)c(Br)cc1Oc1ccc(Br)cc1Br	1	276	FC(F)(F)C1=CC(=O)N(c2c1cc1c(N(CCC1CC)C)c2)C	1
277	Clc1cc(Cl)ccc1-c1cc(Cl)c(Cl)cc1	1	277	FC(F)(F)C1=CC(=O)NC2C1C=C(CC)C(NCC(F)(F)F)=C2	1
278	Clc1cc(Cl)cc(Cl)c1-c1ccc(Cl)cc1	1	278	FC(F)(F)c1cc(Oc2cc(ccc2OC)CO)ccc1C#N	1
279	Clc1c(ccc(Cl)c1Cl)-c1cc(Cl)c(Cl)cc1	1	279	Brc1ccc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)cc1C	1
280	O=C(C)c1c2c3c4c(cc2)cccc4ccc3cc1	1	280	O=C/1CC2C3C(CC[C@]2(C)\C\1=C/C)[C@@]1(C(=CC(=O)CC1)CC3)C	1
281	Clc1cc(Cl)c(Cl)cc1-c1ccc(Cl)cc1	1	281	S(C)c1ncc(NC(=O)N2C[C@@H](N(C[C@H]2C)c2cc(C(F)(F)F)c(cc2)C#N)C)cc1	1
282	Clc1c(Cl)cc(cc1Cl)-c1cc(Cl)c(Cl)c(Cl)c1	0	282	FC(F)(F)C1=CC(=O)Nc2c1cc1N(CCOc1c2)Cc1ccccc1	0
283	Brc1cc(cc(c1)C(F)(F)F)C(F)(F)F	0	283	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)[C@@H]2[C@H]1C1CCC2C=C1	0
284	Clc1ccc(cc1C#N)C(F)(F)F	0	284	S(c1ccccc1C)c1cc(C(F)(F)F)c(cc1)C#N	0
285	Clc1cc(N2C(=O)C3(CC3(C)C2=O)C)cc(Cl)c1	1	285	FC(F)(F)c1cc(ccc1C#N)-c1cc(O)ccc1	1

	Trainingset-1 Smiles	activity		Training set-2 Smiles	activity
286	Clc1cc(Cl)ccc1Oc1cc(OC)c(N(O)O)cc1	1	286	S(CC)c1ccccc1Oc1cc(C(F)(F)F)c(cc1)C#N	1
287	Clc1cc(NC(=O)N(C)C)ccc1Cl	1	287	FC(F)(F)C1=CC(=O)N(c2c1cc1[C@@H]3[C@H](Nc1c2)CCCC3)CC	1
288	Clc1cc(NC(=O)N(OC)C)ccc1Cl	1	288	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C	1
289	Clc1cccc(Cl)c1-c1ccccc1	1	289	Brc1cc(O)c(cc1OC)C[C@H]1C(C)(C)[C@H](Br)CCC1=C	1
290	ClC12C3C(COS(OC3)=O)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	1	290	O=C1CC[C@@]2(C3C(C4CC=C(n5c6c(nc5)cccc6)[C@]4(CC3)C)CCC2=C1)C	1
291	Clc1c(-c2ccccc2)c(Cl)c(Cl)cc1Cl	1	291	FC(F)(F)c1c2c(ncc1)c(c1NC(CC(c1c2)C)(C)C)C	1
292	S=P(Oc1nc(nc(c1)C)C(C)C)(OCC)OCC	0	292	Clc1cc(F)ccc1[C@@H]1C(C#N)=C(NC(C)=C1C(OC(C)(C)C)=O)C	0
293	S(=O)(=O)(NC(OC)=O)c1ccc(N)cc1	0	293	Br\C=C\1/Oc2c(cc(F)cc2)-c2c/1c1c(NC(C=C1C)(C)C)cc2	0
294	FC(F)(C(F)(F)C(F)(F)C(F)(F)C(F)(F)F)C(F)(F)C(F)(F)C(O)=O	0	294	Clc1cc(N(Cc2ccccc2C(F)(F)F)[C@H]2CCN(C2)C)ccc1C#N	0
295	Clc1cc(Cl)ccc1Oc1ccc(Cl)cc1O	1	295	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cc(cnc2)C)C)ccc1C#N	1
296	Clc1cc(Cl)cc(Cl)c1-c1c(Cl)cc(Cl)cc1Cl	0	296	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C(\CC)/C	0
297	C(CCC)(C=C)C	1	297	O=C1N(c2ccc([N+](=O)[O-])cc2)C(=O)N2[C@H]1[C@@H]1N(C[C@H]2C1)C(OC(C)(C)C)=O	1
298	O=C(Nc1ccc(cc1)C(C)C)N(C)C	0	298	O=[N+]([O-])c1ccc(cc1)CC1C(CC(=CC1=C)C)(C)C	0
299	Clc1ccccc1\C(=C(/Cl)\Cl)\c1ccc(Cl)cc1	1	299	S1c2c(N(c3c1cccc3)CC(=O)c1c3c(ccc1)cccc3)cc(cc2)C(F)(F)F	1
300	OC(=O)CCCc1c2c3c4c(cc2)cccc4ccc3cc1	0	300	Clc1nc2c(cc3CCC(N(c3c2)C)(C)C)c(c1)C(F)(F)F	0
301	S(O)(=O)(=O)c1c(cc(cc1C)C)C	0	301	FC(F)(F)c1cc(cc(c1)C(F)(F)F)C1CCNc2c1cc1c(nc(O)cc1C(F)(F)F)c2	0
302	Fc1cc(ccc1O)C(C(C)c1cc(F)c(O)cc1)CC	1	302	Clc1cc(ccc1)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
303	Clc1cc(Cl)ccc1OP(SCCC)(=S)OCC	1	303	FC(F)(F)c1cc(Oc2ccc(cc2OCC)C)ccc1C#N	1
304	Clc1cc(cc(Cl)c1)-c1ccccc1	1	304	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C\c1occc1	1
305	Clc1ccccc1-c1ccc(Cl)cc1	1	305	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(C)C)C1)C	1
306	Fc1ccc(cc1)C(C)=C	1	306	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N)C1)C	1
307	Clc1cc(Cl)c(Cl)cc1-c1cc(Cl)c(Cl)cc1Cl	0	307	O=C1CC[C@]2([C@H](C1)C[C@H]([C@H]1[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]12)C)CCCCCCCC(=O)N(C)C)C	0
308	Clc1c(ccc(Cl)c1Cl)-c1cc(Cl)c(Cl)cc1Cl	1	308	O=C1CCC2=C3C(C4CC[C@](O)(CCCO)[C@]4(CC3c3ccc(N(C)C)cc3)C)CCC2=C1	1
309	Clc1c(cc(Cl)c(Cl)c1Cl)-c1cc(Cl)c(Cl)cc1Cl	1	309	O=C1CCC2=C3[C@H]([C@@H]4CC[C@](O)(CCCO)[C@@]4(C[C@@H]3c3ccc(cc3)C(=O)C)C)CCC2=C1	1
310	Clc1ccc(NC(=O)NC(=O)c2c(F)cccc2F)cc1	0	310	O=C1CC2C[C@H]([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@@H]3[C@]2(CC1)C)C)CCCCCCCC(=O)N(CCCC)C	0
311	Clc1cc(Cl)ccc1C(OCC=C)Cn1ccnc1	1	311	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(CCCC)C)C1)C	1
312	Clc1cc(Cl)cc(Cl)c1-c1ccccc1	1	312	s1cc(cc1C#N)-c1cc2c(NC(OC2(C)C)=O)cc1	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
313	Clc1ccc(Cl)cc1-c1cc(Cl)ccc1Cl	0	313	O[C@]1(CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C=C)C#C	0
314	OC(C(O)C(=O)CO)C(O)CO	0	314	O[C@]1(CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C=C)C#C	0
315	S=P(Oc1cc(ccc1N(O)O)C)(OCC)NC(CC)C	1	315	O=C1N([C@@H]2CC[C@@H]3[C@@H]4C\C(=C\c5ncccc5)\[C@@H](O)[C@@]4(CC[C@H]3[C@@]2(C=C1)C)C)C	1
316	S1CCSC1C(O)C(O)C(O)C(O)CO	0	316	FC(F)(F)C1=CC(=O)NC2C1C=CC(NCc1ccccc1)=C2	0
317	n1ccccc1-c1ncccc1	0	317	FC(F)(F)C1=CC(=O)Nc2c1cc1NCCOc1c2	0
318	O1c2c(OCC1CO)cccc2	1	318	Clc1cc(F)c(cc1)CN(Cc1ccccc1)c1cccc(NS(=O)(=O)C)c1C	1
319	Clc1cc(N2C(=O)CN(C(=O)NC(C)C)C2=O)cc(Cl)c1	0	319	FC(F)(F)c1cc(N2C(=O)[C@@H]3N([C@H]4C[C@@H]3N(C4)C(=O)c3ccccc3)C2=O)ccc1C#N	0
320	O=C1/C(/C2CCC1(C)C2(C)C)=C/c1ccc(cc1)C	0	320	O(CCC(O)=O)c1ccc(cc1)CCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
321	O1C(=O)C(C)C(OC1(C)C)=O	0	321	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(CCC)C)C1)C	0
322	Clc1nc(nc(Cl)c1)-c1ccccc1	0	322	S=C1N(C(=O)C(N1CCCS(=O)(=O)N)(C)C)c1cc(C)c(cc1)C#N	0
323	O(C(=O)NC)c1ccccc1C(CC)C	0	323	FC(F)(F)C1=CC(=O)NC2C1C=C(C)C(NCC(C)C)=C2	0
324	Clc1ccccc1-c1cc(Cl)cc(Cl)c1	1	324	Clc1cc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)ccc1Cl	1
325	Clc1cc(Cl)c(Cl)cc1-c1cc(Cl)ccc1Cl	0	325	FC(F)(F)c1cc(ccc1)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
326	Clc1c(cc(Cl)c(Cl)c1Cl)-c1cc(Cl)c(Cl)cc1	1	326	S(C(C)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
327	Clc1cccc(Cl)c1-c1ccccc1Cl	1	327	S(C)c1ccc(cc1)CN1CC(C)(C)[C@@H](Oc2cc(C(F)(F)F)c(cc2)C#N)C1=O	1
328	Clc1c(cccc1Cl)-c1ccc(Cl)cc1	1	328	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4CC3CC4)C2=O)ccc1	1
329	Clc1cc(cc(Cl)c1)-c1ccc(Cl)cc1	1	329	FC(F)(F)c1cc(O)nc2c1cc1CCC3(N(CCC3)c1c2)C	1
330	Clc1c(cccc1Cl)-c1cccc(Cl)c1Cl	0	330	s1cc(cc1)C1(OCC(=O)Nc2c1cc(cc2)-c1n(C)c(cc1)C#N)C	0
331	Clc1cc(Cl)ccc1OP(SCCC)(OCC)=O	0	331	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cnccc2C)C)ccc1C#N	0
332	Clc1ccc(cc1)CP(OCC)(OCC)=O	0	332	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCC(=O)N(CC)CC)C1)C	0
333	Brc1c(Br)c(Br)ccc1Oc1ccc(Br)cc1	1	333	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(=CC(CC1)=C)CC3)C=C	1
334	Clc1c(Cl)c2Oc3cc(Cl)c(Cl)cc3Oc2cc1Cl	0	334	FC(F)(F)c1cc(N2C(=O)C(N(CCOCCOCCOCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	0
335	ON(O)c1c(C)c(cc(N(O)O)c1NC(CC)CC)C	1	335	FC(F)(F)c1cc(N2C(=O)C(N(OCCOCCOCCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	1
336	ClC(c1ccc(OC)cc1)(c1ccc(OC)cc1)c1ccccc1	1	336	S(=O)(CCCC)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	1
337	ClCCCC(=O)c1ccc(OC)cc1	0	337	Clc1nc2c(cc3c(N(C)C(C=C3C)(C)C)c2)c(c1)C(F)(F)F	0
338	Clc1c2c(ccc1)c(c1c(cccc1)c2C#Cc1ccccc1)C#Cc1ccccc1	0	338	S(=O)(=O)(C[C@](O)(C(=O)Nc1cc(C(F)(F)F)c(cc1)C#N)C)c1ccc(F)cc1	0
339	O(C(=O)CCCCCCCCCCCCCC)CC	0	339	FC(F)(F)C1=CC(=O)Nc2c1cc1N(CC(F)(F)F)[C@H](COc1c2)c1ccccc1	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
340	Brc1cc(Cl)c(OP(SCCC)(OCC)=O)cc1	0	340	Fc1ccc(OC(=O)N2[C@@H]3[C@H]4N([C@@H](C3)C2)C(=O)N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C4=O)cc1	0
341	Clc1cc(Cl)ccc1-c1cc(Cl)ccc1Cl	1	341	FC(F)(F)c1cc(N2C(=O)[C@@H]3[C@@H](C4OC3(CC4)C)C2=O)ccc1	1
342	Clc1c(-c2ccccc2Cl)c(Cl)ccc1Cl	1	342	FC(F)(F)c1cc(N2C(=O)C(N(CCOCCN3CCN(CC3)CCOCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	1
343	Clc1c(cccc1Cl)-c1cc(Cl)cc(Cl)c1	0	343	Ic1ccc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)cc1Cl	0
344	Clc1cc(Cl)ccc1Oc1cc(C(OC)=O)c(N(O)O)cc1	1	344	O(CCCCC(O)=O)c1cc(ccc1)CCCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	1
345	O=C(Nc1ccc(cc1)C)NC(C)(C)c1ccccc1	0	345	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N2CCCC2)C1)C	0
346	ClC=1C2=CC(=O)C3C(C3)C2(C2C(C3CCC(OC(=O)C)(C(=O)C)C3(CC2)C)C=1)C	1	346	O=C1CC[C@]2([C@H](C1)C[C@H]([C@H]1[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]12)C)CCCCCCCCCC(O)=O)C	1
347	Clc1cc(ccc1Oc1cc(OCC)c(N(O)O)cc1)C(F)(F)F	1	347	O(c1ccc(cc1OC)C#N)c1cc(ccc1)C	1
348	Clc1ccc(OC(n2ncnc2)C(=O)C(C)(C)C)cc1	0	348	FC(F)(F)c1cc(Oc2ccc(cc2O)C)ccc1C#N	0
349	OC1CC2CCC3C4CCC(C(CCC(O)=O)C)C4(CCC3C2(CC1)C)C	0	349	O[C@@H]1CC2=CCC3C4CC=C([C@]4(CCC3[C@]2(CC1)C)C)c1nccnc1	0
350	O(C(=O)c1ccccc1C(O)=O)CC(CCCC)CC	0	350	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(nc2)N)C)ccc1C#N	0
351	n1c(cccc1C)-c1nc(ccc1)C	0	351	Brc1cc(O)c(cc1OC)C[C@@H]1C(CCCC1=C)(C)C	0
352	O1C=C(C(=O)c2c1cc(O)cc2O)c1ccc(O)cc1	0	352	O[C@@H]1CC2=CCC3C4CC=C(n5c6c(nc5)cccc6)[C@]4(CCC3[C@]2(CC1)C)C	0
353	FC(F)(F)c1ccccc1C#N	0	353	Brc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)COc2cc1	0
354	O(C)c1cc(ccc1O)\C=C\C(=O)CC(=O)\C=C\c1cc(OC)c(O)cc1	0	354	Clc1ccc(cc1)C\1=NOC(=O)/C/1=C\c1ccc(N2CCCC2)cc1	0
355	OC(=O)CCCCCCC\C=C\C\C=C\C\C=C\CC	0	355	S=C1N(C(=O)C(N1CCCC(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
356	OC(C(O)(c1ccccc1)c1ccccc1)(c1ccccc1)c1ccccc1	0	356	Brc1cc(ccc1)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
357	O=C1CC(C)C2(CC(CCC2=C1)C(C)=C)C	1	357	Clc1cc(N2C(=O)C(N(CCCS(=O)(=O)N)C2=S)(C)C)ccc1C#N	1
358	[NH+]1(C=CC(=C[CH-]1)C1=C[CH-][NH+](C=C1)C)C	0	358	FC(F)(CCCC(CCCCCCCC[C@H]1[C@H]2[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]2[C@@]2([C@H](CC(=O)CC2)C1)C)C)C(O)=O)C(F)(F)	0
359	O(C(=O)CCCCCCCCC)C=C	0	359	S(c1c(cccc1C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
360	OC(CCCCCCCCCCCC)C	0	360	Oc1c2ncccc2c(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)cc1	0
361	OC1CCC2(C(CCC3(C2CCC2C4C(CCC23C)(CCC4C(C)=C)CO)C)C1(C)C)C	0	361	O(CCCCCC(O)=O)c1ccc(cc1)CCCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
362	OC1C2C3CCC(C(CCC(O)=O)C)C3(CCC2C2(C(C1)CC(O)CC2)C)C	0	362	Oc1cc2c(cc1)[C@]1([C@@H](C[C@](O)(CC1)C#CC)CC2)CC	0
363	OC1CC2C(C3CCC(C(CCC(=O)NCC(O)=O)C)C13C)C(O)CC1CC(O)CCC12C	0	363	Clc1cc(F)c2OCc3c4c(NC(C=C4C)(C)C)ccc3-c2c1	0
364	O1c2cc(O)ccc2-c2oc3cc(O)ccc3c2C1=O	0	364	FC(F)(F)c1cc(Oc2ccccc2-c2ccccc2)ccc1C#N	0
365	Oc1cc(O)cc(O)c1C(=O)C	0	365	O(CCCC(O)=O)c1ccc(cc1)CCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
366	OC1CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C	1	366	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(C(C)C)C)C1)C	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
367	Brc1cc(Cl)c(OP(=S)(OCC)OCC)cc1Cl	1	367	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cc(OC)cnc2)C)ccc1C#N	1
368	c12c(c3c(cc(cc3)C(C)C)cc1)cccc2C	1	368	O=C(N1CCCC1)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	1
369	ClCc1c(C)c(C)c(C)c(C)c1C	0	369	S(CC1CCC1)c1cc(OC)c(cc1)C#N	0
370	O1C=C(C(=O)c2c1cc(O)cc2)c1ccc(O)cc1	0	370	Brc1cc(O)c(cc1OC)C[C@@H]1C(C)(C)[C@H](Br)CCC1=C	0
371	O1C=C(C(=O)c2c1cc(O)cc2O)c1ccc(OC)cc1	0	371	FC(F)(F)c1cc(nc2c1cc1CCC(Nc1c2)(C)C)C(F)(F)F	0
372	O(C)c1cc(OC)ccc1N(O)O	1	372	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(nc2)C(OC)=O)C)ccc1C#N	1
373	O1CCc2c(C1)cccc2	0	373	S(=O)(=O)(C)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	0
374	Brc1ccccc1Oc1cc(Br)c(Br)cc1	1	374	Clc1cc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)ccc1Cl	1
375	FC12C(C3CC(C)C(O)(C(=O)CO)C3(CC1O)C)CCC1=CC(=O)C=CC12C	1	375	Brc1ccc(cc1)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
376	OC1(CCC2C3C(C4(C(=CC(=O)CC4)CC3)C)C(=O)CC12C)C(=O)COC(=O)C	0	376	Fc1cc(F)ccc1Oc1cc(C(F)(F)F)c(cc1)C#N	0
377	OC1C2C(C3CCC(C(=O)CO)C3(C1)C)CCC1=CC(=O)CCC12C	1	377	s1cc(cc1C#N)-c1cc2c(NC(C=C2C)(C)C)cc1	1
378	OC1C2(C(CC1O)C1C(CC2)c2c(cc(O)cc2)CC1)C	1	378	Clc1ccc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)cc1C	1
379	OC1CCC2C3C(CCC12C)c1c(cc(O)cc1)CC3	0	379	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ccc(nc2)C(=O)C)C)ccc1C#N	0
380	Clc1ccc(cc1)C(C(Cl)(Cl)Cl)c1ccc(Cl)cc1	1	380	O=C1N(c2ccc(cc2)C#N)C(=O)N2[C@H]1[C@H]1N(C[C@@H]2C1)C(OC(C)(C)C)=O	1
381	Clc1cccc(Cl)c1C(O)=O	0	381	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C=C(CC1)CC)=CC3)C=C	0
382	c12c3c4ccc1c1c(cc2ccc3ccc4)cccc1	1	382	S(=O)(=O)(Nc1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F)C	1
383	O(C(=O)c1ccccc1)c1cc2CCC3C4CCC(O)C4(CCC3c2cc1)C	0	383	FC(F)(F)C1=CC(=O)NC2C1C=C(CC)C(NC)=C2	0
384	OC(C(O)C(O)C=O)C(O)CO	0	384	FC(F)(F)C1=CC(=O)NC2C1C=CC(NC1CCCCC1)=C2	0
385	Oc1cc(ccc1O)CC(C(Cc1cc(O)c(O)cc1)C)C	0	385	Clc1cc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)ccc1F	0
386	OCCCO	0	386	S(CCC(C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
387	Clc1cc(N2C(=O)C(OC2=O)(C=C)C)cc(Cl)c1	1	387	O[C@]1(CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C=C(CC1)C)=CC3)C=C)C	1
388	S1CCSC1=C(C(OC(C)C)=O)C(OC(C)C)=O	0	388	FC(F)(F)c1cc(Oc2ccccc2C)ccc1C#N	0
389	Clc1cc(ccc1Oc1cc(C(O)=O)c(N(O)O)cc1)C(F)(F)F	0	389	FC(F)(F)C1=CC(=O)NC2C1C=CC(N(CCC)CC(F)(F)F)=C2	0
390	Clc1cc(ccc1Oc1cc(C(OC)=O)c(N(O)O)cc1)C(F)(F)F	1	390	O=C1CC2C[C@H]([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@@H]3[C@]2(CC1)C)C)CCCCCCCCC(=O)N(CCCC)C	1
391	OC(=O)CCCC\C=C\C\C=C\C\C=C\CCCCC	1	391	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C=C(CC1)C)=CC3)C=C	1
392	OC(=O)CCCCCCCCCCCCCCCCCC	0	392	O=C1CC2C[C@H]([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@@H]3[C@]2(CC1)C)C)CCCCCCCCCC(=O)N(CCCC)C	0
393	OC1CCC2(C(CCC3(C2CCC2C4CC(CCC4(CCC23C)CO)(C)C)C)C1(C)C)C	0	393	FC(F)(F)C1=CC(=O)Nc2c1cc1N(CC(F)(F)F)[C@H](COc1c2)c1cc(ccc1)C	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
394	FC(F)(C(O)=O)C(F)(F)F	0	394	FC(F)(F)C1=CC(=O)NC2C1C=CC(N(CC)CC)=C2	0
395	Clc1ccc(cc1)C(O)(C(OCC)=O)c1ccc(Cl)cc1	1	395	O1N=C(C)\C(=C\c2ccc(N(CC)CC)cc2)\C1=O	1
396	O1C(OC2OC(COC3OC(CO)C(O)C(O)C3O)C(O)C(O)C2O)(CO)C(O)C(O)C(O)C1CO	0	396	O=C1CC[C@@]2([C@@H]3[C@H]([C@@H]4CC[C@@](O)(C)[C@]4(CC3)C)[C@@H](CC2=C1)CCCCCCCCCC(O)=O)C	0
397	ClCC(=O)N(C(COC)C)c1c(cccc1C)CC	0	397	O=C(N1CCCC1)c1n(cc(N(Cc2ccccc2)c2ccc(cc2)C#N)c1)C	0
398	ClCC(=O)N(CCOCCC)c1c(cccc1CC)CC	0	398	S(=O)(=O)(N(S(=O)(=O)C)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F)C	0
399	O1C2(C(OC1CCC)CC1C3C(C4(C(=CC(=O)C=C4)CC3)C)C(O)CC12C)C(=O)CO	1	399	FC(F)(F)c1cc(N2C(=O)[C@H]3N([C@H]4C[C@@H]3N(C4)C(OC(C)(C)C)=O)C2=O)ccc1C#N	1
400	Clc1cc(Cl)ccc1Oc1ccc(OC(C(OC)=O)C)cc1	0	400	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C/c1ccccc1	0
401	S(Cc1nc[nH]c1C)CCN\C(=N/C)\NC#N	0	401	Fc1cc(F)cc-2c1OCc1c3c(NC(C=C3C)(C)C)ccc1-2	0
402	Clc1ccc(cc1)C(C(C)C)C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O	0	402	O\1c2c(-c3c(c4c(NC(C=C4C)(C)C)cc3)/C/1=C/c1ccccc1)cccc2	0
403	c12c(c(c3c(cccc3)c1-c1ccccc1)-c1ccccc1)c(c1c(cccc1)c2-c1ccccc1)-c1ccccc1	0	403	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4CCC3CCC4)C2=O)ccc1	0
404	Clc1ccc(cc1)CS(=O)C(=O)N(CC)CC	0	404	O(CCCCC(O)=O)c1ccc(cc1)CCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
405	S(C(=O)C)C1C2C3CCC4(OC(=O)CC4)C3(CCC2C2(C(C1)=CC(=O)CC2)C)C	1	405	O=C/1CC2C3C(CC[C@]2(C)\C\1=C\C)[C@@]1(C(=CC(=O)CC1)CC3)C	1
406	ClC(Cl)(Cl)C(P(OC)(OC)=O)O	0	406	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(CC)C)C1)C	0
407	Clc1ccc(cc1)C1(O)CCN(CC1)CCCC(=O)c1ccc(F)cc1	0	407	O=C1CC[C@]2([C@H](C1)C[C@H]([C@H]1[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]12)C)CCCCCCCC(O)=O)C	0
408	OC1(CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)C(C3)C)C)C(=O)C	0	408	Brc1c2c(cccc2)c(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)cc1	0
409	Brc1c2c(cccc2)c(Br)c2c1cccc2	0	409	O(CCCC(O)=O)c1cc(ccc1)CCCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
410	Cl\C(\Cl)=C\C1C(C)(C)C1C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O	0	410	FC(F)(F)C1=CC(=O)NC2C1C=C(C)C(NCCC)=C2	0
411	Cl\C(\Cl)=C\C1C(C)(C)C1C(OCc1cc(Oc2ccccc2)ccc1)=O	0	411	Clc1nc2c(cc3c(NC(CC3C)(C)C)c2C)c(c1)C(F)(F)F	0
412	Clc1c(-c2ccc(Cl)cc2)c(Cl)ccc1Cl	1	412	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cccnc2C)C)ccc1C#N	1
413	Clc1c(Cl)cc(cc1Cl)-c1cc(Cl)c(Cl)cc1Cl	1	413	Fc1cc(cc(c1)C#N)-c1cc2c(NC(=O)COC2(CC)c2occc2)cc1	1
414	Clc1c(N)c(Cl)c(Cl)c(Cl)c1Cl	1	414	FC(F)(F)C1=CC(=O)N(c2c1cc1c(NC(C=C1C)(C)C)c2)CC	1
415	O(C)c1cc(ccc1OC)C(=O)CC(=O)c1cc(OC)c(OC)cc1	1	415	O1[C@@H]2CC=3[C@@]([C@@H]4[C@H]([C@@H]5CC[C@H](O)[C@]5(CC4)C)CC=3)(CC2)CC1	1
416	FC(F)(F)c1cc(NC(=O)C(O)(C)C)ccc1N(O)O	1	416	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)NC)C1)C	1
417	N#Cc1ccccc1C	0	417	O=C1N([C@@H]2CC[C@@H]3[C@H]4C\C(=C\c5cnc(nc5)-c5ncccc5)\[C@@H](O)[C@@]4(CC[C@H]3[C@@]2(C=C1)C)C)C	0
418	Br\C(\Br)=C\C1C(C)(C)C1C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O	0	418	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)\C=C/C	0
419	Oc1cc2CCC3C4CCC(=O)C4(CCC3c2cc1)C	0	419	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)\C=C/C	0
420	c12c3c(ccc1cc1c4c(ccc1c2)cccc4)cccc3	0	420	FC(F)(F)c1c2c(ncc1)cc1N(C)C(CCc1c2)(C)C	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
421	O1CC(Cc2c1cc(O)cc2)c1ccc(O)cc1	0	421	Clc1ncc(NC(=O)N2C[C@@H](N(C[C@H]2C)c2cc(C(F)(F)F)c(cc2)C#N)C)cc1	0
422	Oc1c2c3c4c(cc2)cccc4ccc3cc1	1	422	O(C(=O)C=1[C@@H](C(C#N)=C(NC=1CCC)C)c1c2c(ncc1)cccc2)CC	1
423	c1ccccc1CC(C)C	0	423	O1CC(=O)Nc2c(cc(cc2)-c2n(C)c(cc2)C#N)C1(c1occc1)c1occc1	0
424	O=C(C(C)(C)c1cccnc1)c1cccnc1	0	424	O1c2c(-c3c(c4c(NC(C=C4C)(C)C)cc3)C1c1ccccc1)cccc2	0
425	N(C)(C)c1cc(C)c(N=Nc2ccccc2)cc1	0	425	S=C1N(C(=O)C(N1CCS(=O)(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
426	Clc1cc(Cl)ccc1	0	426	S(C)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	0
427	n1c(nc(nc1N)N)C	0	427	FC(F)(F)c1cc(N2C(=O)[C@@H]3N(C4CC3N(C4)C(OC(C)(C)C)=O)C2=O)ccc1C#N	0
428	S(C)c1ccc(OP(=S)(OC)OC)cc1C	1	428	s1cccc1C1(OCC(=O)Nc2c1cc(cc2)-c1n(C)c(cc1)C#N)c1sccc1	1
429	S(OCCCCOS(=O)(=O)C)(=O)(=O)C	1	429	S=C1N(C(=O)C(N1CCCNS(=O)(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
430	ON(O)c1n(C)c(nc1)C	0	430	Brc1nc2c(cc3c(NC(CC3C)(C)C)c2C)c(c1)C(F)(F)F	0
431	O(C(n1ncnc1)C(O)C(C)(C)C)c1ccc(cc1)-c1ccccc1	1	431	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)[C@@H]2[C@H]1C1[C@H]3[C@@H](C2C=C1)C3	1
432	Clc1ccc(OC(n2ncnc2)C(O)C(C)(C)C)cc1	0	432	Clc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)COc2cc1	0
433	ON(O)c1c2c3c4c(cc2)cccc4ccc3cc1	1	433	Fc1cc(ccc1)\C=C\1/Oc2c(-c3c/1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
434	Clc1c(-c2ccccc2)c(Cl)ccc1Cl	1	434	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C/c1ccccc1C	1
435	Clc1cc(Cl)c(Cl)nc1OP(=S)(OC)OC	0	435	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)N(Cc2ccccc2)C)C1)C	0
436	OC(=O)CCCN	0	436	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@@]12C)[C@@]1(C(C[C@@](O)(CC1)C)=CC3)C=C	0
437	Oc1ccc(cc1)/C(=C(\CC)/c1ccc(O)cc1)/CC	1	437	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)\C=C\C	1
438	c12c3c(ccc1cc1c(c2)cccc1)cccc3	1	438	O(CCCC(O)=O)c1ccc(cc1)CCCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	1
439	OC(CO)CO	0	439	O(Cc1cccnc1C)c1cc2c(cc1)[C@]1([C@@H](C[C@](O)(CC1)C#CC)CC2)CC	0
440	OC(=O)CCCc1ccc(N(O)O)cc1	0	440	O=C(N1CCCC1)c1n(cc(N(CC2CCCCC2)c2ccc([N+](=O)[O-])cc2)c1)C	0
441	S(P(=S)(OCC)OCC)CSP(=S)(OCC)OCC	1	441	S(CC(F)(F)F)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	1
442	OC1(CCC2C3C(CCC12C)c1c(cc(O)cc1)CC3)C#C	1	442	Fc1c2O\C(\c3c4c(NC(C=C4C)(C)C)ccc3-c2ccc1)=C/c1ccccc1	1
443	ClC12C3C(CC(Cl)C3Cl)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	1	443	Fc1cc(ccc1F)\C=C\1/Oc2c(-c3c/1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
444	O=C1CCC2(C3C(C4CCC(C(=O)C)C4(CC3)C)CCC2=C1)C	1	444	Clc1cc(ccc1Cl)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
445	O(C(=O)CC)C1CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C	0	445	O(C(=O)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C)CC	0
446	OC1CC2=CCC3C4CCC(C(CCCC(C)C)C)C4(CCC3C2(CC1)C)C	0	446	FC(F)(F)c1cc(ccc1C#N)-c1cc(OC)ccc1	0
447	c12c(ccc3c1cccc3)c(c1c(cccc1)c2C)C	1	447	S(=O)(CCCC(F)(F)C(F)(F)F)CCCCCCC[C@H]1[C@H]2[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]2[C@@]2(C(CC(=O)CC2)C1)C)C	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
448	Clc1cc(cc(Cl)c1OP(=S)(OC)OC)C	1	448	FC(F)(F)c1cc(ccc1C#N)-c1c(O)cccc1O	1
449	Clc1c(Cl)c2oc3cc(Cl)c(Cl)cc3c2cc1Cl	0	449	O=C(N(CC)CC)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
450	O(CCCC)c1ccc(cc1)C(=O)C	0	450	FC(F)(F)c1cc(N2C(=O)C(N(CCCCN3CCN(CC3)CCOCc3cc(C(=O)N)c(O)cc3)C2=O)(C)C)ccc1C#N	0
451	Clc1c(Cl)cc(cc1Cl)-c1cc(Cl)c(Cl)cc1	1	451	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cncnc2)C)ccc1C#N	1
452	Oc1c2cc(O)ccc2ccc1	0	452	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C/C=C(C)C	0
453	c12c(cc(cc1)C)cccc2C	0	453	O=C(N1CCc2c1cccc2)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
454	Clc1c(Cl)c2Oc3c(Oc2cc1Cl)c(Cl)c(Cl)c(Cl)c3	0	454	O=C1N(c2cc3ncccc3cc2)C(=O)[C@@H]2[C@H]1C1CC2C=C1	0
455	Clc1ccc(cc1)-c1c(nc(nc1N)N)CC	0	455	O\1c2c(-c3c(c4c(NC(C=C4C)(C)C)cc3)/C/1=C/c1ccccc1C)cccc2	0
456	OC1(CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C)C	0	456	Clc1cc(cc(Cl)c1)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
457	O1C(CO)C(O)C(O)C1n1c2NC=NC(=O)c2nc1	0	457	S=C1N(C(=O)C(N1CCCS(=O)(=O)N)(C)C)c1ccc(C#N)c(C(F)(F)F)c1C	0
458	P(OCC1OC(n2c3ncnc(N)c3nc2)C(O)C1O)(OP(O)(O)=O)(O)=O	0	458	s1cc(cc1C#N)-c1cc2c(NC(C)(C)C(=O)C2(C)C)cc1	0
459	ClC1C(Cl)C(Cl)C(Cl)C(Cl)C1Cl	0	459	s1c(ccc1C#N)-c1cc2c(NC(C=C2C)(C)C)cc1	0
460	Clc1cc2NCNS(=O)(=O)c2cc1S(=O)(=O)N	0	460	O[C@H]1CC[C@H]2[C@H]3[C@@H]([C@]45CC[C@@H](O)[C@H](CC4)C5=CC3)CC[C@]12C	0
461	Clc1cc(Cl)ccc1C(=O)c1c(nn(C)c1OS(=O)(=O)c1ccc(cc1)C)C	0	461	Clc1ccc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)cc1[N+](=O)[O-]	0
462	Clc1ccc(cc1)C(O)(C(OC(C)C)=O)c1ccc(Cl)cc1	1	462	FC(F)(F)c1cc(nc2c1cc1CCC(Nc1c2)(C)C)N	1
463	O(C(=O)C1C(C)(C)C1\C=C(\C)/C)C1CC(=O)C(CC=C)=C1C	0	463	FC(F)(F)C1=CC(=O)NC2C1C=CC(N(C)C)=C2	0
464	N(CCc1ccccc1)C	0	464	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2cnccc2OC)C)ccc1C#N	0
465	Nc1ccc(cc1)CC	0	465	O=C1N(c2cc3c(cc2)cccc3)C(=O)[C@@H]2[C@H]1C1CC2C=C1	0
466	OC1CCC(CC1)C	0	466	Fc1cc(ccc1C)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
467	O=C1Nc2c(C=C1)cccc2	0	467	Brc1cc(cc(c1)C)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
468	Clc1ccc(O)cc1C	1	468	O1c2c(-c3c(c4c(NC(C=C4C)(C)C)cc3)C1c1ccc(cc1)C)cccc2	1
469	Clc1nc(nc(n1)NCC)NC(C)(C)C	0	469	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)COc2cc1	0
470	ClC12C3C(C4C5OC5C3C4)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	1	470	O(C(=O)C=1[C@@H](C(C(OC)=O)=C(NC=1C)C)c1cc([N+](=O)[O-])ccc1)C(C)(C)C	1
471	Oc1cc(O)cc(O)c1C(=O)CCc1ccc(O)cc1	0	471	Fc1cc(cc(c1)C#N)-c1cc2c(NC(OC2(C)C)=O)cc1	0
472	Clc1ccccc1C(O)(c1ccc(Cl)cc1)c1cncnc1	1	472	FC(F)(F)c1cc(N2C[C@@H](N(C[C@H]2C)C(=O)Nc2ccc(nc2)C(F)(F)F)C)ccc1C#N	1
473	Clc1cc(Cl)ccc1C1(OC(CO1)CCC)Cn1ncnc1	1	473	O1CCN(CC1)C(=O)CCCCCCCC1[C@H]2[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]2[C@@]2([C@@H](C1)CC(=O)CC2)C)C	1
474	N(C)(C)c1ccc(cc1)C(c1ccc(N(C)C)cc1)c1ccc(N(C)C)cc1	0	474	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1OCC(OC)=O	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
475	Clc1c2c(cccc2)c(Cl)c2c1cccc2	1	475	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)N2[C@H]1[C@H]1N(C[C@@H]2C1)C(OC(C)(C)C)=O	1
476	n1c2c(ccc1)c(N)ccc2	0	476	S(c1c(cc(cc1C)C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
477	Nc1cc2c(cc1)cc1c(c2)cccc1	1	477	O[C@H]1CC[C@H]2[C@H]3[C@@H]([C@]45CC[C@H](O)[C@H](CC4)C5=CC3)CC[C@]12C	1
478	Clc1cc(N(O)O)cc(Cl)c1O	0	478	S=C1N(C(=O)C(N1CCCCS(=O)(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
479	Cl\C(\Cl)=C\OP(OC)(OC)=O	0	479	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)\C=C/OC	0
480	Oc1ccc(cc1)Cc1ccc(O)cc1	1	480	Clc1ccc(cc1C)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
481	S=C(Nc1ccc(cc1)C)Nc1ccc(cc1)C	0	481	Cl\C=C\[C@@]12C3C(C4CC[C@H](O)[C@]4(CC3)C)CC=C1C[C@@H](O)CC2	0
482	O=C(N)\C=C\c1ccccc1	0	482	Fc1nc2c(cc3c(NC(CC3C)(C)C)c2C)c(c1)C(F)(F)F	0
483	Oc1ccc(cc1)CO	0	483	s1c(ccc1[N+](=O)[O-])-c1cc2c(NC(C=C2C)(C)C)cc1	0
484	Clc1cc(F)ccc1	0	484	FC(F)(F)C1=CC(=O)Nc2c1c1OC[C@H](Nc1cc2)c1ccccc1	0
485	Brc1cccnc1	0	485	FC(F)(F)C1=CC(=O)Nc2c1c1OCCNc1cc2	0
486	O(C(=O)C)CCCCC	0	486	Fc1cc2-c3c(c4c(NC(C=C4C)(C)C)cc3)/C(/Oc2cc1)=C/c1ccccc1N(C)C	0
487	ICCCCCI	0	487	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)NCC)C1)C	0
488	C(CCCCCCCCCCCC)CCCCCCCCCCCC	0	488	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1OCCO	0
489	O=C1CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C	0	489	Clc1ccc(cc1)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
490	O(C(=O)NC)c1c2c(ccc1)cccc2	0	490	FC(F)(F)c1cc(N(C(C(=O)N(C)C)C)CC(F)(F)F)ccc1C#N	0
491	O1C(CO)C(O)C(O)C(O)C1OC1C(O)C(O)C(OC1CO)O	0	491	Fc1ccc(F)cc1\C=C/1\Oc2c(-c3c\1c1c(NC(C=C1C)(C)C)cc3)cccc2	0
492	c1ccccc1C(=C(c1ccccc1)c1ccccc1)c1ccccc1	0	492	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCCC(=O)NCCC)C1)C	0
493	S(=O)(=O)(N)c1ccc(N(O)O)cc1	0	493	S(=O)(=O)(Nc1cccc(N(Cc2ccc(Oc3cc(OCCCC(=O)NCCCC(O)=O)ccc3)cc2)Cc2ccccc2)c1C)C	0
494	FC(F)(F)c1cc(N2C(=O)C(NC2=O)(C)C)ccc1N(O)O	1	494	S=C1N(C(=O)C(N1CCCNC(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
495	O=C1CCC(=O)CC1	0	495	O[C@]1(CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C=C(CC1)C)=CC3)C=C)C#C	0
496	OCC	0	496	FC(F)(F)c1cc(Oc2cc(ccc2)C)ccc1C#N	0
497	S(CCSP(=S)(OC)OC)CC	0	497	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4[C@@H]5[C@H](C3C=C4)C5)C2=O)ccc1	0
498	Clc1ccc(N(C(C)C)C(=O)CSP(=S)(OC)OC)cc1	1	498	FC(F)(F)c1cc(ncc1C#N)N([C@@H](C)c1ccccc1)C	1
499	Clc1nc(nc2c1cccc2)-c1ccccc1	0	499	FC(F)(F)c1ccc(N(Cc2ccccc2)c2cc(n(c2)C)C(=O)N2CCCC2)cc1	0
500	Clc1cc(Cl)ccc1C1(OC(CO1)COc1ccc(N2CCN(CC2)C(=O)C)cc1)Cn1ccnc1	1	500	O=C(N1CCCCC1)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	1
501	O=C1NC(=O)NC=C1	0	501	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)CC#C	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
502	Clc1ccc(cc1)CN(C(=O)Nc1ccccc1)C1CCCC1	1	502	S(CCC(C)C)c1cc(OC)c(cc1)C#N	1
503	n1c(nc(nc1N)N)NC1CC1	0	503	FC(F)(F)c1cc(cc(N2C(=O)[C@H]3[C@H](C4CC3C=C4)C2=O)c1)C(F)(F)F	0
504	BrC(C(Br)(Br)Br)C1C(C)(C)C1C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O	0	504	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C#C	0
505	O=C(C)C	0	505	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)CC=C	0
506	S(=O)(C)C	0	506	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)CCC	0
507	n1c2n3c(nc2ccc1N)C=CC=C3	0	507	S(CCCC)c1nc2c(cc3CCC(Nc3c2)(C)C)c(c1)C(F)(F)F	0
508	n1c2n3c(nc2ccc1N)C(=CC=C3)C	0	508	Fc1ccc(OC(=O)N2[C@@H]3[C@H]4N([C@@H](C3)C2)C(=O)N(c2c3c(cccc3)c(cc2)C#N)C4=O)cc1	0
509	Clc1cc(Cl)cc(Cl)c1OCCN(CCC)C(=O)n1ccnc1	1	509	FC(F)(F)c1cc(ccc1C#N)-c1cc(ccc1)C	1
510	O=CN(C)C	0	510	O(C)c1cc(ccc1O)CC=C	0
511	[nH]1c2c(c3cc(C)c(nc13)N)cccc2	0	511	O=C1CC2C[C@H]([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@@H]3[C@]2(CC1)C)C)CCCCCCCCCCC(=O)N(CCCC)C	0
512	Cl\C(\Cl)=C\C1C(C)(C)C1C(OC(C#N)c1cc(Oc2ccccc2)c(F)cc1)=O	0	512	S=C1N(C(=O)C(N1CCCS(=O)(=O)N)(C)C)c1cc(OC)c(cc1)C#N	0
513	O(CC)c1c2c(cccc2)c(OCC)c2c1cccc2	0	513	O=C(N)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
514	O=C1NC(=O)Nc2nc[nH]c12	0	514	O=C(C)c1ccc(N(Cc2ccccc2)c2cc(n(c2)C)C(=O)N2CCCC2)cc1	0
515	Clc1ccc(cc1)/C(/Cl)=C/C1C(C)(C)C1C(OC(C#N)c1cc(Oc2ccccc2)c(F)cc1)=O	0	515	S=C1N(C(=O)C(N1CCCC(O)=O)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	0
516	Clc1c(ccc(Cl)c1Cl)-c1cc(Cl)c(Cl)c(Cl)c1	1	516	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)NCCCCCC)C1)C	1
517	FC(F)(F)c1ccc(Oc2ccc(OC(C(OCCCC)=O)C)cc2)nc1	0	517	OCc1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
518	Clc1cc(Cl)ccc1-c1ccc(Cl)cc1	1	518	Clc1ccc(cc1F)C1Oc2c(-c3c1c1c(NC(C=C1C)(C)C)cc3)cccc2	1
519	FC(F)(F)Oc1ccc(cc1)C(C(C)C)C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O	1	519	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1OCCOC	1
520	O=Cc1c2c(cccc2)c(c2c1cccc2)C=O	0	520	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C=C(CC1)C)=CC3)C	0
521	Fc1cc(ccc1)C(=O)C(F)(F)F	0	521	FC(F)(F)c1cc(Oc2ccccc2C(C)C)ccc1C#N	0
522	Clc1cc(NC(=O)CC)ccc1Cl	1	522	Clc1cc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)ccc1C	1
523	O(C(=O)C)C1(CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)C(C3)C)C)C(=O)C	0	523	Clc1cc(F)ccc1Cn1nc2c(cccc2C(F)(F)F)c1-c1ccc(F)cc1	0
524	Clc1cc(Cl)ccc1C(=O)c1c(nn(C)c1OCC(=O)c1ccccc1)C	1	524	FC(F)(F)C1=CC(=O)N(c2c1cc1c(NC(C=C1C)(C)C)c2C)C	1
525	s1ccnc1NS(=O)(=O)c1ccc(N)cc1	0	525	FC(F)(F)c1cc(N(Cc2ccccc2)c2cc(n(c2)C)C(=O)N2CCCC2)ccc1[N+](=O)[O-]	0
526	ClC(Cl)(Cl)C(c1ccc(OC)cc1)c1ccc(OC)cc1	1	526	O=C1N(c2ccc([N+](=O)[O-])cc2)C(=O)N2[C@@H]1[C@H]1N(C[C@@H]2C1)C(=O)c1ccccc1	1
527	Clc1ccc(cc1)C(C(Cl)Cl)c1ccc(Cl)cc1	1	527	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C[C@](O)(CC1)C)=CC3)C=C	1
528	Clc1ccc(cc1)C(=C(Cl)Cl)c1ccc(Cl)cc1	1	528	Clc1cc(Oc2cc(C(F)(F)F)c(cc2)C#N)ccc1	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
529	S(=O)(=O)(Nc1noc(c1)C)c1ccc(N)cc1	0	529	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)ccc1C#N	0
530	S(C)c1nc(nc(n1)NC(C)C)NC(C)C	0	530	Fc1cc(cc(c1)C#N)-c1cc2c(NC(=O)COC2(C)C)cc1	0
531	C(=C\C)/CC=C	0	531	Fc1cc(cc(c1)C#N)-c1cc2c(NC(=O)COC2(c2occc2)c2occc2)cc1	0
532	S(P(=S)(OC)OC)CN1C(=O)c2c(cccc2)C1=O	0	532	O[C@H]1CC[C@H]2[C@H]3[C@H](CC[C@]12C)[C@@]1(C(C=C(CC1)c1ccccc1)=CC3)C=C	0
533	s1c2c(nc1OCC(=O)N(C)c1ccccc1)cccc2	1	533	FC(F)(F)c1cc(nc2c1cc1CCC(Nc1c2)(C)C)NCc1ccc(OC)cc1	1
534	N(c1ccc(Nc2ccccc2)cc1)c1ccccc1	0	534	O1c2c(cc(O)cc2)C(=O)CC1c1ccccc1	0
535	S(C(CC1CC(=O)C(/C(=N/OCC)/CCC)C(=O)C1)C)CC	0	535	Clc1ccc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)cc1C(F)(F)F	0
536	N(c1ccc(Nc2ccccc2)cc1)c1ccccc1	1	536	Clc1ccc(Oc2ccc(cc2)CN(Cc2ccc(F)cc2F)c2cccc(NS(=O)(=O)C)c2C)cc1OCC(O)=O	1
537	ClC12C3C(C=CC3Cl)C(Cl)(C1(Cl)Cl)C(Cl)=C2Cl	1	537	O=C(N1CCCC1)c1ccc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)cc1	1
538	ClC(c1ccccc1)(c1ccccc1)c1ccccc1	1	538	S=C1N(C(=O)C(N1CCCS(=O)(=O)N(C)C)(C)C)c1cc(C(F)(F)F)c(cc1)C#N	1
539	OC(c1ccccc1)(c1ccccc1)c1ccccc1	0	539	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)C#CC	0
540	OC(C(O)=O)(c1ccccc1)c1ccccc1	0	540	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)\C=C/CC	0
541	C(CC)(C=C)C	0	541	Fc1ccc(cc1)CN(c1ccc([N+](=O)[O-])cc1)c1cc(n(c1)C)C(=O)N1CCCC1	0
542	n1c2c3c(nccc3)ccc2n(C)c1N	0	542	s1cc(cc1)[C@](O)(C)[C@H]1CCCC2=Cc3n(ncc3C[C@]12C)-c1ccc(F)cc1	0
543	n1c2c(n(C)c1N)c(cc1ncccc12)C	0	543	O=C1N(C2[C@H]3[C@@H](C1C=C2)C(=O)N(c1c2c(cccc2)c([N+](=O)[O-])cc1)C3=O)C	0
544	n1c2c3nc(n(c3ccc2ncc1C)C)N	0	544	FC(F)(F)c1cc(N2C(=O)[C@@H]3N([C@H]4C[C@@H]3N(C4)C(OC(C)(C)C)=O)C2=O)ccc1C#N	0
545	c12c(cc3c(cccc3)c1C)cccc2	1	545	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)N2[C@@H]1[C@H]1N(C[C@@H]2C1)C(OC(C)(C)C)=O	1
546	c12c(cccc1)c(c1c(cccc1)c2C)C	1	546	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ncncc2)C)ccc1C#N	1
547	Clc1ccccc1C(C(Cl)(Cl)Cl)c1ccc(Cl)cc1	1	547	O[C@H]1CCC2C3C(CC[C@]12C)[C@@]1(C(C[C@@H](O)CC1)=CC3)\C=C\OC	1
548	O=C(N)C=C	0	548	O=C(N1CCCC1)c1n(cc(N(Cc2ccc(cc2)C)c2ccc([N+](=O)[O-])cc2)c1)C	0
549	OC1CCC2(C(CCC3=C2CCC2(C)C3(CCC2C(CC\C=C(\C)/C)C)C)C1(C)C)C	0	549	Oc1ccc(cc1)CN(c1ccc([N+](=O)[O-])cc1)c1cc(n(c1)C)C(=O)N1CCCC1	0
550	Brc1cc(cc(Br)c1O)C(C)(C)c1cc(Br)c(O)c(Br)c1	1	550	O=C(N1CCCC1)c1cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)ccc1	1
551	Cl\C(=C\C1C(C)(C)C1C(OCc1c(F)c(F)c(C)c(F)c1F)=O)\C(F)(F)F	0	551	FC(F)(F)c1cc(ncc1C#N)N[C@H](C)c1ccccc1	0
552	Clc1c(N(O)O)c(Nc2ncc(cc2Cl)C(F)(F)F)c(N(O)O)cc1C(F)(F)F	0	552	O=C1N(c2ccc(cc2)C#N)C(=O)N2[C@@H]1[C@H]1N(C[C@@H]2C1)C(OC(C)(C)C)=O	0
553	Oc1ccc(cc1)C(C)(C)c1ccc(O)cc1	1	553	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H]([C@H]4CC[C@@H]3C4)C2=O)ccc1[N+](=O)[O-]	1
554	OC1C2C(C3CCC(C(=O)C)C3(C1)C)CCC1=CC(=O)CCC12C	1	554	O=C1N(c2c3c(cccc3)c([N+](=O)[O-])cc2)C(=O)N2[C@@H]1[C@H]1CC[C@@H]2C1	1
555	ClC12C(CCl)(CCl)C(=C)C(Cl)(C1(Cl)Cl)C(Cl)C2Cl	1	555	S=C1N(C(=O)C(N1CCCS(=O)(=O)N)(C)C)c1cc(C(F)(F)F)c(cc1C)C#N	1

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
556	Clc1ccc(cc1)C1(CCCC1)C(O)=O	0	556	O(c1ccc(cc1OC)C#N)c1ccc(cc1)C	0
557	O(c1cc(ccc1)COCC(C)(C)c1ccc(OCC)cc1)c1ccccc1	1	557	Brc1ccc(N2C(=O)[C@H]3[C@H](C4C=CC3[C@@H](O)[C@H]4O)C2=O)cc1C	1
558	S(O)(=O)(=O)CCNC(=O)CCC(C)C1CCC2C3C(CC(O)C12C)C1(C(CC3O)CC(O)CC1)C	0	558	FC(F)(F)c1cc(Oc2ccc(cc2)C)ccc1C#N	0
559	OC1CC2C(C3CCC(C(CCC(O)=O)C)C13C)C(O)CC1CC(O)CCC12C	0	559	OCc1ccc(cc1)CN(c1ccc([N+](=O)[O-])cc1)c1cc(n(c1)C)C(=O)N1CCCC1	0
560	O1C(=O)c2c3c(ccc4c3c(cc2)C(OC4=O)=O)C1=O	0	560	O(CCCC(O)=O)c1cc(ccc1)CCC[C@H]1[C@H]2[C@@H]3CC[C@@](O)(C)[C@]3(CC[C@@H]2[C@@]2(C(=CC(=O)CC2)C1)C)C	0
561	O=C1NC(=O)c2c3c1cccc3ccc2	0	561	OCc1ccc(cc1)CN(c1cc(cc(c1)C)C)c1cc(n(c1)C)C(=O)N1CCCC1	0
562	Clc1c(N(O)O)c(Cl)c(Cl)c(Cl)c1Cl	0	562	O(C(=O)c1ccc(N(Cc2ccccc2)c2cc(n(c2)C)C(=O)N2CCCC2)cc1)C	0
563	OC1CC2=CCC3C4CCC(C(CCC(C(C)C)CC)C)C4(CCC3C2(CC1)C)C	0	563	O=C(N(CCCC)CCCC)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
564	S(=O)(=O)(NC(=O)Nc1nc(OC)cc(OC)n1)Cc1ccccc1C(OC)=O	0	564	OC(=O)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
565	O(C(=O)C)C1CCC2C3C(CCC12C)=C(CC)C(=O)CC3	1	565	O=C1CC[C@]2([C@@H](CC([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@H]23)C)CCCCCCCC(=O)NCCCC)C1)C	1
566	Clc1cc(NC(=O)C(O)(C=C)C)cc(Cl)c1	1	566	O=C1CC[C@]2(C(=C1)CC[C@H]1[C@@H]3CC[C@@](O)(C#CC)[C@]3(CC=C12)C)Cc1ccc(cc1)C	1
567	O(C)c1ccc(cc1)\C=C\C(OCC(CCCC)CC)=O	0	567	S(=O)(=O)(Nc1cc2c(-c3c(OC2c2ccccc2)cccc3OC)cc1)C	0
568	O=C1c2c(cccc2)C(=O)c2c1cccc2	0	568	Oc1ccc(cc1)CN(c1cc(cc(c1)C)C)c1cc(n(c1)C)C(=O)N1CCCC1	0
569	O(C(=O)c1ccccc1C(OCC)=O)CC	0	569	Clc1ccc(Oc2cc(C(F)(F)F)c(cc2)C#N)cc1	0
570	O(C(=O)c1ccccc1C(OCCCC)=O)CCCC	0	570	O=C1N(c2ccc(cc2)C#N)C(=O)N2[C@@H]1[C@H]1N(C[C@@H]2C1)C(=O)c1ccc(cc1)CCCC	0
571	Clc1cc(N2C(=O)C(OC2=O)(C(OCC)=O)C)cc(Cl)c1	1	571	S(=O)(CCCC(F)(F)C(F)(F)F)CCCCC[C@H]1[C@H]2[C@@H]3CC[C@H](O)[C@]3(CC[C@@H]2[C@@]2(C(CC(=O)CC2)C1)C)C	1
572	OC1(CCC2C3C(=C4C(=CC(=O)CC4)CC3)C(CC12C)c1ccc(N(C)C)cc1)C#CC	1	572	FC(F)(F)c1cc(O)nc2c1cc1c(NC(C)(C)[C@@H](C)[C@@H]1C)c2	1
573	c12c3c(ccc1cccc2)cccc3	0	573	Brc1cc(cc(Br)c1OCc1cc(Br)ccc1)CCC(O)=O	0
574	Clc1cc(C(=O)c2ccccc2)c(O)cc1	1	574	FC(F)(F)c1cc(N2C[C@H](N(C[C@@H]2C)C(=O)Nc2ncccc2)C)ccc1C#N	1
575	O(C(=O)c1ccccc1C(OCCCC)=O)Cc1ccccc1	0	575	FC(F)(F)c1cc(ccc1C#N)-c1ccccc1OCCOCCOC	0
576	C1c2c(-c3c1cccc3)cccc2	0	576	S(=O)(=O)(Nc1cc2c(-c3c(OC2c2ccccc2)cccc3OC(F)F)cc1)C	0
577	Clc1c(O)c(Cl)c(Cl)c(Cl)c1Cl	0	577	O=C1CC2C[C@H]([C@H]3[C@@H]4CC[C@H](O)[C@]4(CC[C@@H]3[C@]2(CC1)C)C)CCCCCCCCCCCC(=O)N(CCCC)C	0
578	Clc1c(O)c(Cl)c(Cl)c(O)c1Cl	1	578	O=C(N1CCCC1)c1n(cc(N(C(C)C)c2cc(cc(c2)C)C)c1)C	1
579	Nc1ccc(cc1)C#N	0	579	S(=O)(=O)(Nc1cc2c(-c3c(OC2c2ccccc2)cccc3OCC)cc1)C	0
580	Clc1cc(Cl)cc(Cl)c1O	0	580	FC(F)(F)c1cc(N2C(=O)[C@@]3(N([C@H]4C[C@@H]3N(C4)C(OC(C)(C)C)=O)C2=O)C)ccc1C#N	0
581	Oc1c(cc(cc1C(C)(C)C)CO)C(C)(C)C	1	581	FC(F)(F)c1cc(N2C(=O)[C@H]3[C@H](C4OC3CC4)C2=O)ccc1	1
582	Oc1cc(C(C)(C)C)c(O)cc1C(C)(C)C	0	582	S(=O)(=O)(Nc1cccc(N(Cc2ccc(Oc3ccc(OCC(O)=O)cc3)cc2)Cc2ccccc2)c1C)C	0

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
583	Clc1c(N(O)O)cc(N(O)O)cc1N(O)O	0	583	O[C@H]1CN(C[C@@H]1O)C(=O)c1n(cc(N(Cc2ccccc2)c2ccc([N+](=O)[O-])cc2)c1)C	0
584	S=C(Oc1cc(ccc1)C(C)(C)C)N(C)c1nc(OC)ccc1	0	584	O=C(N1CCCC1)c1n(cc(N(C(C)C)c2ccc([N+](=O)[O-])cc2)c1)C	0
585	n1c(cc(nc1N\N=C(\C)/c1ccccc1C)C)C	0	585	O=C(N1CCCC1)c1n(cc(Nc2ccc([N+](=O)[O-])cc2)c1)C	0
586	O=C1CC2C(C1)CCC2CCCCC(OC)CC	1	586	O=C(N1CCCC1)c1n(cc(Nc2cc(cc(c2)C)C)c1)C	1
587	Nc1ccccc1-c1ccccc1	1	587	O=C(N1CCCC1)c1n(cc(N(CC)c2ccc([N+](=O)[O-])cc2)c1)C	1
588	Oc1ccccc1-c1ccccc1	1	588	FC(F)(F)c1cc(N2C(=O)[C@@H]3N([C@H]4C[C@@H]3N(C4)C(=O)c3ccc(cc3)CCCC)C2=O)ccc1C#N	1
589	S(=O)(=O)(CC(O)(C(=O)Nc1cc(C(F)(F)F)c(cc1)C#N)C)c1ccc(F)cc1	1	589	O=C(N1CCCC1)c1n(cc(N(CC)c2cc(cc(c2)C)C)c1)C	1
590	N#Cc1ccccc1C#N	0	590	O1c2c(cccc2)C(=O)C(O)=C1c1ccccc1	0
591	c12c(cccc1)cccc2	0	591	O=C(N1CCCC1)c1n(cc(N(C)c2cc(cc(c2)C)C)c1)C	0
592	O(CC)c1cc2c(NC(C=C2C)(C)C)cc1	0	592	O=C(N1CCCC1)c1n(cc(N(C)c2ccc([N+](=O)[O-])cc2)c1)C	0
593	N(CC)(CC)C1CCCCC1	1	593	O(C)c1cc(ccc1O)CCC(=O)C	1
594	Oc1cc(N(CC)CC)ccc1	0	594	Oc1ccc(cc1)CCC(=O)C	0
595	Clc1cc(ccc1N)-c1cc(Cl)c(N)cc1	1	595	Oc1ccc(cc1)CC(=O)C	1
596	Cl\C(=C\C1C(C)(C)C1C(OC(C#N)c1cc(Oc2ccccc2)ccc1)=O)\C(F)(F)F	0			
597	OC(=O)c1ccc(cc1)-c1ccccc1	0			
598	O(C)c1cc(ccc1OC)CC=C	0			
599	Clc1cc(Cl)c(Cl)cc1OCC(O)=O	0			
600	ClC1CCCC1	0			
601	Sc1cc(ccc1)C(F)(F)F	0			
602	Oc1ccc(cc1)C(OCCC)=O	0			
603	Oc1ccc(cc1)C(OCCCC)=O	0			
604	Clc1cc(C)c(OCC(O)=O)cc1	0			
605	Clc1cc(Cl)ccc1OCC(O)=O	0			
606	OC(C(CC)CO)CCC	0			
607	O=C(C)c1c2c(ccc1)cccc2	1			
608	Clc1ccc(cc1)C(O)(C(C)C1CC1)Cn1ncnc1	0			
609	Clc1ccccc1N	0			

Trainingset-1 Smiles		activity	Training set-2 Smiles		activity
610	Clc1cc(Cl)c(Cl)cc1Cl	0			
611	Clc1cc(Cl)c(Cl)cc1O	0			
612	S1C(=O)N(N=C1OC)CSP(=S)(OC)OC	0			
613	n1c2c3nc(n(c3c(cc2ncc1C)C)C)N	0			
614	ClCC(O)CO	0			
615	Clc1cc(Cl)c(Cl)cc1/C(/OP(OC)(OC)=O)=C/Cl	0			
616	s1ccc(OC)c1CN(C(=O)C)c1c(cccc1C)C	1			
617	Clc1cc(Cl)ccc1OP(=S)(OCC)OCC	1			
618	O(C(=O)C(O)C)CC	0			
619	OC(C(CC1C(CCCC1C)(C)C)C)C	0			
620	S(O)(=O)(=O)c1ccccc1	0			
621	O=C1NC2CCC3C4CCC(C(=O)NC(C)(C)C)C4(CCC3C2(C=C1)C)C	1			
622	Oc1ccc(cc1)C(OC)=O	0			
623	Oc1ccc(cc1)C(O)=O	0			
624	Clc1cc(C(F)(F)F)c(\N=C(\n2ccnc2)/COCCC)cc1	1			
625	ClC[CH-][N+](C)(C)C	0			

3. External Testing Set Chemicals

ID	smiles	activity
1	c1(c(c(c2c(c1)C(CCC2(C)C)(C)C)C)O)C	1
2	c1ccc2c(c1)C[C@@H](C2=O)C	0
3	c1(ccc2c(c1)C[C@@H](C2=O)C)C	0
4	c1c(cc2c(c1)[C@@](CC2(C)C)(C)CC)C(=O)C	0
5	c1(c(cc2c(c1)C(CC2(C)C)(C)C)COC)C	0
6	c1ccc2c(c1)C(=O)[C@@H](C2)CC	0
7	C(c1ccc(N(C)C)cc1)c1ccc(N(C)C)cc1	0
8	c1(ccccc1)CCCC	0
9	OP(=O)(O)CNCC(=O)O	0
10	c1(ccccc1)O	0
11	O[C@H](CCCCCC)CO	0
12	c1c(OP(=S)(OCC)OCC)ccc(c1)[S@](=O)C	0
13	O(C(=O)c1c(O)cccc1)C	0
14	Oc1c(cc(C=O)cc1)OC	0
15	COP(=S)(S[C@@H](C(=O)OCC)CC(=O)OCC)OC	0
16	N(CCCCN)CCCN	0
17	S=P(Oc1ccc(cc1)C#N)(c1ccccc1)OCC	0
18	S(CSP(=S)(OCC)OCC)C(C)(C)C	0
19	C(=O)(C)c1c2CCC(c2cc(c1)C(C)(C)C)(C)C	0
20	C(=O)(C(C)C)Nc1cc(c(cc1)N(=O)=O)C(F)(F)F	1
21	Clc1cc([C@@]2(OC2)C[C@]2(CC)C(=O)c3c(C2=O)cccc3)ccc1	0
22	c1(ccccc1)[C@H](CC)C	0
23	N(C)(C)C(=S)SSC(=S)N(C)C	0
24	C(=O)(C=Cc1ccccc1)O	0
25	O=S1(=O)O[C@](c2ccccc12)(c1ccc(O)cc1)c1ccc(O)cc1	0
26	O[C@@H]1CC2=CC[C@H]3[C@H]4[C@](CC[C@@H]3[C@]2(CC1)C)([C@H](CC4)C(=O)C)C	0
27	O(C)c1cc2c(c3CC[C@](C)([C@@H](CC)c3cc2)C(=O)O)cc1	0
28	c1c(cc(c(c1N(O)O)N(CCC)CCC)N(O)O)C(F)(F)F	0
29	[Si](O[Si](Cc1ccccc1)(C)C)(Cc1ccccc1)(C)C	0
30	c12c(CCC1)c(=O)n(c(=O)[nH]2)C1CCCCC1	0
31	CC(C)NP(=O)(OCC)Oc1cc(c(cc1)SC)C	0
32	P(=O)(OC(=CCl)c1c(cc(c(c1)Cl)Cl)Cl)(OC)OC	0
33	ClC1(Cl)[C@]2(Cl)[C@@]3(Cl)[C@]4(Cl)C(Cl)(Cl)[C@]5(Cl)[C@@]3(Cl)[C@@]1(Cl)[C@]5(Cl)[C@]24Cl	0
34	O=c1sc2nc3cc(C)ccc3nc2s1	0
35	S(P(=S)(OCC)OCC)CC(=O)N(C)C(=O)OCC	0
36	S=P(S[C@@H](c1ccccc1)C(=O)OCC)(OC)OC	0
37	[C@@H]1([C@@H]([C@@H]([C@H]([C@H]([C@@H]1Cl)Cl)Cl)Cl)Cl)Cl	0
38	[C@@H]1([C@@H]([C@H]([C@@H]([C@H]([C@@H]1Cl)Cl)Cl)Cl)Cl)Cl	0
39	N(C)(C=Nc1c(cc(cc1)C)C)C=Nc1c(cc(cc1)C)C	0

ID	smiles	activity
40	O[C@@H](CCCCC)C=C	0
41	c1cccc2c1OP(=S)(OC2)OC	0
42	Ic1c(OC(=O)CCCCCCC)c(I)cc(c1)C#N	0
43	[C@@H]1([C@@H]([C@H]([C@H]([C@H]([C@@H]1Cl)Cl)Cl)Cl)Cl)Cl	0
44	c1ccc(c2c1n1c(s2)nnc1)C	0
45	c1cccc2c1CCC2	0
46	O[C@]1([C@@]2([C@H]([C@H]3[C@H]([C@H](C2)O)[C@@]2(C(=CC(=O)C=C2)CC3)C)CC1)C)C(=O)CO	0
47	C(=O)(CCCCCCC(=O)O)O	0
48	Cl[C@H]1C[C@@H]2[C@@]3(C(=C([C@]([C@@H]2[C@H]1Cl)(C3(Cl)Cl)Cl)Cl)Cl)Cl	1
49	C1(=C([C@@]2([C@H]3[C@@H]([C@]1(C2(Cl)Cl)Cl)C[C@H]([C@@H]3Cl)Cl)Cl)Cl)Cl	1
50	C1(=O)CC[C@]2(C(=C1)CC[C@H]1[C@@H]3CC[C@H](C(=O)CO)[C@@]3(C=O)C[C@H](O)[C@H]21)C	0
51	S1C(=S)N(CN(C1)C)C	0
52	O(c1ccc(NC(=O)[C@@H](O)C)cc1)CC	0
53	C(=O)(c1c(cccc1)C)Nc1cc(ccc1)OC(C)C	0
54	P(=S)(Oc1ccc(N(=O)=O)cc1)(OCC)OCC	1
55	OC(=O)CCCCCCCCCCCCCCCC	0
56	O=C1N2c3c(CC2)cccc3CC1	0
57	c1ccc(c(c1C)N([C@@H](C)C(=O)OC)C(=O)COC)C	0
58	O=c1n(C)c2c(n(C)cn2)c(=O)n1C	0
59	Clc1c(=O)n(ncc1N)c1ccccc1	0
60	C(=O)(CC[C@H](C(=O)O)NC(=O)c1ccc(cc1)NCc1nc2c(=O)[nH]c(N)nc2nc1)O	0
61	O=C(CSP(=S)(OC)OC)NC	0
62	S(C(=O)N1CCCCC1)C(c1ccccc1)(C)C	0
63	C(CCCCCCCO)O	0
64	Oc1c(cccc1)C(=O)N	0
65	O(C(C)C)c1cc(NC(=O)c2c(cccc2)C(F)(F)F)ccc1	0
66	c1(cc(=O)c2ccc(cc2o1)O)c1ccccc1	0
67	ClC(=C[C@H]1C([C@H]1C(=O)O[C@@H](c1cc(Oc2ccccc2)ccc1)C#N)(C)C)C(F)(F)F	0
68	S(=O)(=O)(Oc1cc2C(COc2cc1)(C)C)CC	0
69	COP(=O)(OC)OC(=CC(=O)NC)C	0
70	C(=O)([C@H](Nc1c(cc(cc1)C(F)(F)F)Cl)C(C)C)O[C@H](c1cc(ccc1)Oc1ccccc1)C#N	0
71	Cl[C@@]12C(=C(Cl)[C@@](C1(Cl)Cl)([C@@H]1[C@H]2[C@@H]2C[C@H]1[C@@H]1O[C@H]21)Cl)Cl	0
72	S(c1ccc(OP(=O)(OCCC)OCCC)cc1)C	0
73	N(C(=O)C)CCc1c[nH]c2ccc(OC)cc12	0
74	O=C(O)c1c(cccc1)[C@@H](c1ccc(O)cc1)c1ccc(O)cc1	0
75	O=C1c2c(C(=O)c3c1ccc(O)c3)ccc(O)c2	0
76	S(C(=O)N([C@@H](C(C)C)C)CC)Cc1ccccc1	0
77	O(C(C)C)C(=O)Nc1cc(OCC)c(OCC)cc1	0
78	c1cccc(c1)C=CC=Cc1ccccc1	0
79	c1c(cccc1)c1ccccc1	0
80	[C@H]([C@@H](c1ccccc1)C)(c1ccccc1)c1ccccc1	0

ID	smiles	activity
81	c1(c(cccc1)Cl)O	0
82	C1[C@@H]2[C@H](CO[S@@](=O)O1)[C@]1(C(=C([C@@]2(C1(Cl)Cl)Cl)Cl)Cl)Cl	1
83	c1(ccc(cc1)C(CCCCC)(C)C)O	1
84	[C@@H]12C=C(CO)C[C@]3([C@H]([C@]1([C@@H]([C@H]([C@@]1([C@H]2C1)OC(=O)C)OC(=O)CCCCCCCC	0
85	[C@@H]1(C(=O)C[C@]2([C@H](C1)CC[C@H]1[C@@H]3CCC[C@]3(CC[C@H]21)C)C)O	0
86	C1(=O)CC[C@]2(C(=C1)CC[C@H]1[C@@H]3CC[C@H](O)[C@]3(CC[C@H]21)C)C	0
87	C(COc1ccc(cc1)C(=C(CC)c1ccccc1)c1ccc(cc1)O)N(C)C	0
88	C1(=O)CC[C@]2(C(=C1)CC[C@H]1[C@@H]3CC[C@]([C@]3(C[C@@H]([C@]21F)O)C)(C)O)C	0
89	C1CC(=O)C=C2C1=C1[C@@H](CC2)[C@H]2[C@](C=C1)([C@](CC2)(O)C)C	0

4. PPV, NPV, Sensitivity, Specificity, Concordance, ROC AUC of Different Combination of Training Sets

4.1 Training Set 1 (T1)

T1-PPV

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.60	0.60	0.60	0.58	0.58	0.61	0.59	0.59	0.61	0.61	0.61	0.61	0.60	0.62	0.63	0.63	0.64	0.64	0.64	0.64	0.64	0.66	0.63	0.63	0.64	0.64
ADTree	0.58	0.58	0.58	0.60	0.61	0.60	0.60	0.61	0.60	0.65	0.65	0.65	0.61	0.61	0.63	0.63	0.65	0.65	0.65	0.64	0.65	0.63	0.63	0.63	0.65	0.63
ANN	0.59	0.60	0.65	0.62	0.64	0.60	0.61	0.63	0.67	0.66	0.65	0.61	0.63	0.65	0.61	0.63	0.63	0.63	0.65	0.65	0.66	0.63	0.67	0.67	0.66	0.68
BAGGING	0.64	0.63	0.64	0.66	0.65	0.68	0.71	0.68	0.68	0.69	0.69	0.68	0.70	0.71	0.74	0.73	0.71	0.73	0.71	0.72	0.70	0.71	0.73	0.71	0.73	0.71
BayesNet	0.62	0.62	0.62	0.62	0.63	0.63	0.61	0.61	0.62	0.62	0.63	0.64	0.65	0.64	0.65	0.64	0.66	0.66	0.66	0.66	0.66	0.64	0.65	0.65	0.67	0.67
Classification ViaRegression	0.60	0.60	0.61	0.63	0.63	0.65	0.63	0.63	0.61	0.62	0.64	0.64	0.60	0.64	0.65	0.63	0.64	0.64	0.66	0.66	0.65	0.65	0.66	0.65	0.65	0.66
DAGGING	0.80	0.80	0.77	0.72	0.68	0.64	0.66	0.65	0.65	0.68	0.69	0.68	0.68	0.68	0.67	0.65	0.67	0.66	0.68	0.67	0.67	0.71	0.70	0.68	0.72	0.69
Decision Table	0.61	0.61	0.62	0.60	0.61	0.61	0.61	0.61	0.65	0.65	0.65	0.66	0.68	0.69	0.68	0.68	0.68	0.67	0.67	0.67	0.67	0.66	0.62	0.66	0.67	0.67
Decorate	0.67	0.69	0.67	0.62	0.64	0.63	0.65	0.66	0.68	0.71	0.70	0.70	0.68	0.66	0.68	0.71	0.70	0.69	0.70	0.71	0.70	0.72	0.71	0.70	0.70	0.68
DNTB	0.61	0.61	0.62	0.61	0.61	0.61	0.63	0.63	0.66	0.69	0.69	0.66	0.65	0.66	0.67	0.67	0.64	0.61	0.61	0.62	0.62	0.64	0.62	0.62	0.63	0.63
END	0.71	0.71	0.65	0.62	0.62	0.62	0.64	0.66	0.67	0.70	0.70	0.68	0.68	0.66	0.62	0.63	0.64	0.65	0.66	0.66	0.66	0.68	0.68	0.69	0.67	0.65
IB1	0.67	0.67	0.69	0.61	0.64	0.64	0.70	0.71	0.69	0.68	0.68	0.68	0.69	0.71	0.72	0.70	0.70	0.70	0.71	0.71	0.70	0.72	0.63	0.72	0.74	0.74
KNN	0.64	0.64	0.66	0.64	0.63	0.64	0.63	0.64	0.61	0.62	0.63	0.65	0.65	0.64	0.64	0.66	0.64	0.64	0.65	0.65	0.65	0.67	0.67	0.70	0.67	0.67
Kstar	0.58	0.57	0.60	0.59	0.61	0.60	0.61	0.64	0.64	0.64	0.63	0.63	0.63	0.63	0.64	0.64	0.63	0.64	0.64	0.64	0.65	0.64	0.63	0.64	0.65	0.66
LogicBoost	0.62	0.60	0.63	0.63	0.64	0.68	0.67	0.67	0.69	0.69	0.70	0.69	0.70	0.71	0.68	0.69	0.70	0.70	0.70	0.71	0.71	0.73	0.69	0.71	0.71	0.73
Logistic	0.62	0.62	0.63	0.64	0.64	0.65	0.65	0.66	0.66	0.67	0.66	0.67	0.65	0.65	0.64	0.65	0.66	0.67	0.67	0.67	0.68	0.64	0.64	0.63	0.64	0.65
LWL-random	0.66	0.64	0.66	0.63	0.64	0.68	0.69	0.68	0.72	0.70	0.72	0.68	0.69	0.73	0.71	0.76	0.74	0.79	0.75	0.77	0.73	0.75	0.77	0.76	0.75	0.74
NBTree	0.62	0.62	0.61	0.64	0.60	0.63	0.62	0.63	0.64	0.64	0.66	0.63	0.62	0.62	0.64	0.67	0.67	0.66	0.62	0.65	0.63	0.67	0.67	0.64	0.68	0.67
PART	0.63	0.63	0.65	0.66	0.64	0.68	0.66	0.70	0.69	0.71	0.71	0.70	0.71	0.68	0.72	0.72	0.73	0.76	0.77	0.73	0.71	0.74	0.75	0.77	0.76	0.77
Random Committee	0.62	0.62	0.66	0.64	0.64	0.68	0.72	0.68	0.70	0.70	0.69	0.70	0.72	0.72	0.72	0.74	0.76	0.73	0.77	0.74	0.77	0.74	0.71	0.73	0.74	0.75
Random Forest	0.64	0.63	0.64	0.65	0.66	0.72	0.69	0.69	0.67	0.66	0.70	0.70	0.72	0.72	0.71	0.72	0.71	0.72	0.70	0.71	0.70	0.72	0.71	0.72	0.73	0.72
Random Sub Space	0.63	0.62	0.67	0.64	0.61	0.64	0.68	0.66	0.66	0.65	0.66	0.65	0.65	0.66	0.66	0.68	0.66	0.66	0.65	0.66	0.66	0.63	0.65	0.62	0.62	0.63
REPTree	0.67	0.62	0.65	0.67	0.65	0.67	0.66	0.67	0.65	0.69	0.69	0.68	0.69	0.73	0.73	0.72	0.73	0.72	0.73	0.72	0.72	0.75	0.72	0.74	0.74	0.76
Rotationforest	0.63	0.63	0.65	0.63	0.64	0.63	0.62	0.62	0.64	0.67	0.66	0.66	0.66	0.66	0.66	0.66	0.64	0.67	0.67	0.68	0.68	0.64	0.65	0.66	0.63	0.67
SPegasos	0.68	0.65	0.69	0.70	0.70	0.69	0.67	0.68	0.67	0.68	0.69	0.69	0.71	0.71	0.72	0.71	0.72	0.71	0.71	0.72	0.72	0.72	0.71	0.72	0.75	0.75
SVM	0.60	0.60	0.60	0.58	0.58	0.61	0.59	0.59	0.61	0.61	0.61	0.61	0.60	0.62	0.63	0.63	0.64	0.64	0.64	0.64	0.64	0.66	0.63	0.63	0.64	0.64

T1-NPV

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.77	0.77	0.77	0.76	0.76	0.77	0.78	0.78	0.79	0.77	0.77	0.77	0.79	0.77	0.76	0.76	0.74	0.74	0.74	0.74	0.74	0.77	0.77	0.77	0.75	0.74
ADTree	0.76	0.76	0.76	0.78	0.81	0.77	0.77	0.77	0.75	0.75	0.75	0.75	0.76	0.78	0.76	0.74	0.76	0.76	0.76	0.76	0.76	0.77	0.78	0.78	0.74	0.75
ANN	0.77	0.75	0.75	0.76	0.78	0.77	0.77	0.74	0.75	0.75	0.76	0.74	0.76	0.76	0.74	0.74	0.76	0.75	0.78	0.77	0.76	0.76	0.78	0.79	0.77	0.79
BAGGING	0.75	0.75	0.74	0.77	0.76	0.78	0.78	0.77	0.77	0.79	0.78	0.77	0.78	0.77	0.80	0.79	0.78	0.79	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.77
BayesNet	0.75	0.76	0.76	0.76	0.75	0.75	0.78	0.78	0.80	0.80	0.81	0.82	0.82	0.82	0.83	0.83	0.84	0.86	0.86	0.86	0.86	0.83	0.83	0.83	0.85	0.85
Classification ViaRegression	0.71	0.71	0.71	0.72	0.75	0.77	0.75	0.74	0.74	0.76	0.79	0.80	0.76	0.80	0.78	0.77	0.78	0.76	0.76	0.76	0.76	0.76	0.77	0.75	0.76	0.77
DAGGING	0.97	0.97	0.96	0.94	0.91	0.89	0.90	0.89	0.89	0.86	0.86	0.85	0.85	0.85	0.84	0.84	0.85	0.84	0.84	0.84	0.84	0.87	0.86	0.85	0.86	0.85
DecisionTable	0.78	0.78	0.78	0.78	0.77	0.77	0.76	0.76	0.78	0.77	0.77	0.77	0.78	0.78	0.76	0.76	0.76	0.75	0.75	0.75	0.75	0.76	0.75	0.75	0.75	0.75
Decorate	0.82	0.84	0.80	0.75	0.74	0.75	0.78	0.78	0.78	0.81	0.80	0.80	0.79	0.78	0.80	0.82	0.81	0.80	0.81	0.81	0.79	0.82	0.81	0.80	0.81	0.80
DNTB	0.78	0.78	0.78	0.77	0.78	0.78	0.77	0.77	0.77	0.78	0.79	0.79	0.79	0.80	0.78	0.78	0.75	0.74	0.74	0.75	0.75	0.76	0.75	0.75	0.76	0.76
END	0.86	0.86	0.78	0.76	0.75	0.76	0.78	0.78	0.78	0.80	0.80	0.78	0.78	0.76	0.73	0.74	0.75	0.75	0.77	0.76	0.76	0.78	0.79	0.80	0.78	0.77
IB1	0.73	0.72	0.73	0.72	0.74	0.74	0.73	0.77	0.76	0.75	0.75	0.75	0.75	0.75	0.77	0.75	0.75	0.75	0.75	0.76	0.76	0.77	0.77	0.78	0.79	0.78
KNN	0.73	0.73	0.74	0.73	0.76	0.77	0.75	0.76	0.74	0.75	0.74	0.73	0.73	0.74	0.74	0.74	0.75	0.75	0.75	0.76	0.75	0.76	0.77	0.77	0.77	0.77
Kstar	0.76	0.75	0.76	0.75	0.75	0.76	0.76	0.76	0.76	0.76	0.76	0.77	0.76	0.76	0.76	0.77	0.76	0.76	0.76	0.76	0.76	0.78	0.79	0.80	0.79	0.78
LogicBoost	0.73	0.71	0.73	0.73	0.75	0.78	0.78	0.78	0.80	0.80	0.81	0.81	0.81	0.82	0.79	0.79	0.79	0.79	0.80	0.80	0.80	0.83	0.80	0.81	0.80	0.81
Logistic	0.67	0.67	0.67	0.69	0.69	0.69	0.70	0.70	0.70	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.72	0.74	0.74	0.73	0.74	0.71	0.72	0.71	0.73	0.75
LWL-random	0.73	0.72	0.73	0.71	0.71	0.72	0.74	0.74	0.74	0.75	0.75	0.75	0.74	0.75	0.74	0.76	0.77	0.78	0.76	0.78	0.75	0.76	0.77	0.76	0.77	0.76
NBTree	0.71	0.71	0.73	0.73	0.71	0.76	0.69	0.69	0.71	0.74	0.73	0.73	0.71	0.74	0.71	0.74	0.75	0.74	0.73	0.73	0.73	0.76	0.72	0.73	0.76	0.76
PART	0.72	0.72	0.71	0.79	0.72	0.76	0.71	0.76	0.75	0.73	0.75	0.71	0.68	0.70	0.73	0.78	0.78	0.76	0.68	0.76	0.71	0.78	0.78	0.74	0.78	0.75
Random Committee	0.75	0.76	0.78	0.79	0.79	0.82	0.80	0.83	0.83	0.84	0.83	0.83	0.83	0.82	0.84	0.83	0.84	0.87	0.87	0.84	0.83	0.85	0.86	0.87	0.88	0.87
Random Forest	0.77	0.77	0.79	0.79	0.79	0.82	0.85	0.81	0.84	0.83	0.83	0.82	0.84	0.85	0.84	0.85	0.87	0.84	0.87	0.85	0.88	0.85	0.84	0.85	0.86	0.86
RandomSubSpace	0.69	0.70	0.71	0.70	0.73	0.74	0.73	0.72	0.73	0.74	0.76	0.75	0.76	0.76	0.78	0.76	0.76	0.77	0.75	0.76	0.76	0.78	0.77	0.76	0.76	0.76
REPTree	0.76	0.77	0.76	0.76	0.75	0.73	0.74	0.72	0.75	0.75	0.75	0.74	0.76	0.74	0.74	0.73	0.72	0.72	0.72	0.72	0.72	0.72	0.73	0.74	0.74	0.71
Rotationforest	0.77	0.75	0.77	0.80	0.76	0.79	0.78	0.78	0.77	0.79	0.79	0.79	0.80	0.83	0.83	0.83	0.83	0.82	0.84	0.82	0.82	0.84	0.82	0.83	0.84	0.85
SPegasos	0.83	0.84	0.85	0.83	0.84	0.80	0.80	0.79	0.81	0.82	0.82	0.82	0.81	0.81	0.82	0.82	0.78	0.80	0.80	0.82	0.82	0.80	0.81	0.81	0.79	0.79
SVM	0.78	0.75	0.79	0.82	0.81	0.81	0.79	0.78	0.78	0.79	0.79	0.80	0.80	0.81	0.82	0.80	0.81	0.81	0.81	0.81	0.81	0.82	0.81	0.82	0.84	0.83

T1-Sensitivity

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.70	0.70	0.70	0.69	0.69	0.71	0.73	0.73	0.74	0.71	0.71	0.71	0.74	0.69	0.66	0.66	0.61	0.61	0.61	0.61	0.61	0.67	0.69	0.69	0.63	0.61
ADTree	0.71	0.71	0.70	0.73	0.78	0.72	0.71	0.71	0.67	0.64	0.64	0.64	0.68	0.73	0.67	0.62	0.66	0.65	0.65	0.66	0.66	0.69	0.70	0.70	0.62	0.63
ANN	0.72	0.67	0.64	0.67	0.71	0.71	0.71	0.63	0.62	0.63	0.66	0.64	0.67	0.65	0.65	0.62	0.68	0.63	0.69	0.68	0.66	0.66	0.70	0.71	0.66	0.70
BAGGING	0.65	0.63	0.63	0.68	0.66	0.68	0.66	0.65	0.67	0.70	0.68	0.66	0.67	0.66	0.69	0.69	0.67	0.69	0.67	0.68	0.67	0.66	0.66	0.67	0.67	0.65
BayesNet	0.66	0.67	0.67	0.67	0.65	0.65	0.72	0.72	0.75	0.76	0.77	0.78	0.78	0.79	0.80	0.79	0.81	0.84	0.84	0.84	0.84	0.79	0.80	0.80	0.82	0.82
Classification ViaRegression	0.60	0.60	0.61	0.63	0.63	0.65	0.63	0.63	0.61	0.62	0.64	0.64	0.60	0.64	0.65	0.63	0.64	0.64	0.66	0.66	0.65	0.65	0.66	0.65	0.65	0.66
DAGGING	0.80	0.80	0.77	0.72	0.68	0.64	0.66	0.65	0.65	0.68	0.69	0.68	0.68	0.68	0.67	0.65	0.67	0.66	0.68	0.67	0.67	0.71	0.70	0.68	0.72	0.69
DecisionTable	0.72	0.72	0.71	0.72	0.70	0.70	0.69	0.69	0.70	0.69	0.68	0.67	0.68	0.68	0.65	0.65	0.63	0.63	0.63	0.63	0.63	0.64	0.65	0.63	0.63	0.63
Decorate	0.67	0.69	0.67	0.62	0.64	0.63	0.65	0.66	0.68	0.71	0.70	0.70	0.68	0.66	0.68	0.71	0.70	0.69	0.70	0.71	0.70	0.72	0.71	0.70	0.70	0.68
DNTB	0.72	0.72	0.71	0.71	0.72	0.72	0.70	0.70	0.67	0.70	0.70	0.72	0.72	0.73	0.69	0.68	0.65	0.63	0.63	0.65	0.65	0.66	0.65	0.65	0.67	0.67
END	0.71	0.71	0.65	0.62	0.62	0.62	0.64	0.66	0.67	0.70	0.70	0.68	0.68	0.66	0.62	0.63	0.64	0.65	0.66	0.66	0.66	0.68	0.68	0.69	0.67	0.65
IB1	0.63	0.62	0.63	0.59	0.63	0.65	0.62	0.67	0.66	0.63	0.64	0.65	0.65	0.66	0.68	0.64	0.65	0.65	0.65	0.66	0.67	0.68	0.69	0.70	0.72	0.70
KNN	0.57	0.56	0.57	0.62	0.65	0.68	0.60	0.62	0.59	0.60	0.59	0.57	0.57	0.57	0.58	0.58	0.61	0.59	0.59	0.62	0.61	0.61	0.69	0.63	0.64	0.64
Kstar	0.66	0.65	0.66	0.65	0.65	0.66	0.66	0.67	0.68	0.67	0.66	0.67	0.66	0.67	0.66	0.68	0.67	0.65	0.66	0.66	0.65	0.68	0.71	0.71	0.70	0.69
LogicBoost	0.62	0.60	0.63	0.63	0.64	0.68	0.67	0.67	0.69	0.69	0.70	0.69	0.70	0.71	0.68	0.69	0.70	0.70	0.70	0.71	0.71	0.73	0.69	0.71	0.71	0.73
Logistic	0.43	0.42	0.41	0.48	0.46	0.48	0.48	0.50	0.50	0.52	0.52	0.52	0.52	0.53	0.52	0.52	0.55	0.60	0.60	0.58	0.59	0.54	0.56	0.55	0.57	0.63
LWL-random	0.59	0.57	0.57	0.55	0.54	0.55	0.59	0.59	0.58	0.60	0.59	0.63	0.59	0.60	0.57	0.60	0.65	0.65	0.60	0.65	0.60	0.61	0.63	0.62	0.63	0.62
NBTree	0.62	0.62	0.64	0.64	0.61	0.65	0.60	0.60	0.63	0.66	0.64	0.64	0.62	0.65	0.64	0.66	0.66	0.66	0.65	0.66	0.65	0.66	0.64	0.65	0.67	0.66
PART	0.62	0.62	0.61	0.64	0.60	0.63	0.62	0.63	0.64	0.64	0.66	0.63	0.62	0.62	0.64	0.67	0.67	0.66	0.62	0.65	0.63	0.67	0.67	0.64	0.68	0.67
Random Committee	0.63	0.63	0.65	0.66	0.64	0.68	0.66	0.70	0.69	0.71	0.71	0.70	0.71	0.68	0.72	0.72	0.73	0.76	0.77	0.73	0.71	0.74	0.75	0.77	0.76	0.77
Random Forest	0.62	0.62	0.66	0.64	0.64	0.68	0.72	0.68	0.70	0.70	0.69	0.70	0.72	0.72	0.72	0.74	0.76	0.73	0.77	0.74	0.77	0.74	0.71	0.73	0.74	0.75
RandomSubSpace	0.48	0.50	0.53	0.51	0.57	0.58	0.57	0.54	0.58	0.61	0.63	0.61	0.63	0.63	0.67	0.61	0.63	0.63	0.61	0.63	0.63	0.66	0.65	0.62	0.62	0.62
REPTree	0.67	0.69	0.64	0.66	0.66	0.58	0.58	0.56	0.62	0.64	0.64	0.60	0.65	0.60	0.60	0.57	0.55	0.56	0.56	0.56	0.56	0.58	0.59	0.63	0.63	0.55
Rotationforest	0.67	0.62	0.65	0.67	0.65	0.67	0.66	0.67	0.65	0.69	0.69	0.68	0.69	0.73	0.73	0.72	0.73	0.72	0.73	0.72	0.72	0.75	0.72	0.74	0.74	0.76
SPegasos	0.63	0.63	0.65	0.63	0.64	0.63	0.62	0.62	0.64	0.67	0.66	0.66	0.66	0.66	0.66	0.66	0.64	0.67	0.67	0.68	0.68	0.64	0.65	0.66	0.63	0.67
SVM	0.68	0.65	0.69	0.70	0.70	0.69	0.67	0.68	0.67	0.68	0.69	0.69	0.71	0.71	0.72	0.71	0.72	0.71	0.71	0.72	0.72	0.72	0.71	0.72	0.75	0.75

T1-Specificity

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.67	0.67	0.67	0.66	0.66	0.68	0.65	0.65	0.66	0.69	0.69	0.69	0.66	0.71	0.73	0.73	0.76	0.76	0.76	0.76	0.76	0.76	0.72	0.72	0.76	0.76
ADTree	0.65	0.65	0.65	0.66	0.66	0.68	0.68	0.68	0.69	0.76	0.76	0.76	0.70	0.67	0.73	0.75	0.75	0.76	0.76	0.75	0.75	0.72	0.72	0.72	0.77	0.75
ANN	0.66	0.69	0.76	0.71	0.73	0.67	0.69	0.75	0.79	0.78	0.75	0.72	0.73	0.76	0.72	0.74	0.72	0.75	0.75	0.74	0.77	0.73	0.76	0.76	0.76	0.78
BAGGING	0.75	0.75	0.75	0.76	0.75	0.78	0.82	0.78	0.78	0.78	0.79	0.78	0.80	0.81	0.83	0.83	0.81	0.82	0.81	0.82	0.81	0.82	0.83	0.81	0.83	0.82
BayesNet	0.72	0.71	0.71	0.71	0.73	0.73	0.68	0.68	0.68	0.68	0.68	0.70	0.71	0.69	0.70	0.70	0.71	0.71	0.71	0.71	0.71	0.69	0.70	0.70	0.71	0.71
DecisionTable	0.68	0.68	0.70	0.67	0.69	0.69	0.69	0.69	0.74	0.74	0.74	0.76	0.78	0.79	0.78	0.78	0.79	0.79	0.79	0.79	0.79	0.77	0.73	0.78	0.79	0.79
DNTB	0.68	0.68	0.69	0.68	0.68	0.68	0.71	0.71	0.76	0.76	0.78	0.74	0.74	0.74	0.76	0.77	0.75	0.72	0.72	0.72	0.72	0.75	0.73	0.73	0.73	0.73
IB1	0.69	0.68	0.71	0.72	0.72	0.70	0.72	0.74	0.74	0.75	0.75	0.74	0.74	0.73	0.73	0.75	0.73	0.75	0.75	0.74	0.75	0.73	0.72	0.73	0.74	0.75
KNN	0.81	0.81	0.82	0.73	0.74	0.74	0.82	0.82	0.82	0.81	0.81	0.81	0.82	0.84	0.84	0.83	0.82	0.83	0.84	0.82	0.82	0.84	0.72	0.83	0.84	0.85
Kstar	0.74	0.74	0.76	0.75	0.74	0.74	0.72	0.73	0.70	0.72	0.73	0.75	0.75	0.73	0.74	0.76	0.74	0.75	0.75	0.76	0.75	0.77	0.76	0.79	0.76	0.77
Logistic	0.82	0.82	0.83	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.81	0.81	0.80	0.81	0.81	0.79	0.79	0.80	0.81	0.79	0.79	0.78	0.78	0.77
LWL-random	0.79	0.78	0.79	0.77	0.79	0.82	0.81	0.81	0.84	0.82	0.84	0.80	0.82	0.84	0.84	0.87	0.84	0.88	0.86	0.86	0.85	0.86	0.87	0.86	0.85	0.85
RandomSubSpace	0.81	0.80	0.80	0.81	0.80	0.84	0.82	0.83	0.80	0.79	0.82	0.82	0.83	0.83	0.81	0.84	0.82	0.83	0.81	0.82	0.82	0.82	0.82	0.83	0.84	0.83
REPTree	0.73	0.71	0.78	0.74	0.71	0.77	0.81	0.80	0.78	0.76	0.77	0.77	0.76	0.79	0.79	0.81	0.81	0.80	0.79	0.80	0.80	0.76	0.78	0.73	0.73	0.78

T1-Concordance

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.69	0.69	0.69	0.67	0.67	0.69	0.68	0.68	0.70	0.70	0.70	0.70	0.69	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.72	0.71	0.71	0.70	0.70
ADTree	0.67	0.67	0.67	0.69	0.71	0.69	0.69	0.69	0.68	0.71	0.71	0.71	0.69	0.70	0.70	0.69	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.70
ANN	0.68	0.68	0.71	0.70	0.72	0.69	0.69	0.70	0.72	0.72	0.72	0.69	0.70	0.71	0.69	0.69	0.70	0.70	0.72	0.72	0.72	0.70	0.73	0.74	0.72	0.74
BAGGING	0.71	0.70	0.70	0.73	0.71	0.74	0.75	0.73	0.73	0.75	0.74	0.73	0.75	0.75	0.77	0.77	0.75	0.77	0.75	0.76	0.75	0.75	0.76	0.75	0.76	0.75
BayesNet	0.69	0.69	0.69	0.69	0.70	0.70	0.69	0.69	0.71	0.71	0.72	0.73	0.73	0.73	0.74	0.74	0.75	0.76	0.76	0.76	0.76	0.73	0.74	0.74	0.76	0.76
DecisionTable	0.70	0.70	0.70	0.69	0.70	0.70	0.69	0.69	0.72	0.72	0.72	0.72	0.74	0.74	0.73	0.73	0.73	0.72	0.72	0.72	0.72	0.72	0.69	0.72	0.72	0.72
DNTB	0.70	0.70	0.70	0.69	0.69	0.69	0.70	0.70	0.72	0.73	0.75	0.73	0.73	0.73	0.73	0.73	0.71	0.68	0.68	0.69	0.69	0.71	0.69	0.69	0.70	0.70
IB1	0.66	0.66	0.67	0.66	0.68	0.68	0.68	0.71	0.71	0.70	0.70	0.70	0.70	0.70	0.71	0.70	0.70	0.71	0.71	0.71	0.71	0.71	0.71	0.72	0.73	0.73
KNN	0.71	0.71	0.72	0.68	0.71	0.71	0.73	0.74	0.72	0.72	0.72	0.71	0.72	0.73	0.74	0.73	0.73	0.73	0.74	0.74	0.73	0.74	0.71	0.75	0.76	0.76
Kstar	0.71	0.70	0.72	0.71	0.70	0.71	0.70	0.71	0.69	0.70	0.70	0.72	0.71	0.71	0.71	0.73	0.71	0.71	0.71	0.72	0.71	0.73	0.74	0.76	0.74	0.74
Logistic	0.66	0.66	0.66	0.68	0.67	0.68	0.68	0.69	0.69	0.70	0.69	0.70	0.69	0.69	0.69	0.69	0.70	0.71	0.71	0.71	0.72	0.69	0.69	0.69	0.69	0.71
LWL-random	0.71	0.69	0.70	0.68	0.69	0.71	0.72	0.72	0.73	0.73	0.74	0.73	0.73	0.74	0.73	0.76	0.76	0.79	0.76	0.77	0.75	0.76	0.77	0.76	0.76	0.76
RandomSubSpace	0.68	0.68	0.69	0.69	0.70	0.73	0.72	0.71	0.71	0.71	0.74	0.73	0.75	0.75	0.75	0.74	0.74	0.75	0.73	0.74	0.74	0.76	0.75	0.74	0.75	0.74
REPTree	0.70	0.70	0.72	0.71	0.69	0.69	0.72	0.70	0.71	0.71	0.72	0.70	0.72	0.71	0.71	0.71	0.70	0.70	0.70	0.70	0.70	0.69	0.70	0.69	0.69	0.68

4.2 Training Set 2 (T2)

T2-PPV

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.52	0.57	0.59	0.59	0.58	0.6	0.6	0.62	0.62	0.62	0.62	0.63	0.63	0.62	0.62	0.62	0.62	0.62	0.62	0.61	0.61	0.61	0.58	0.58	0.58	0.58
ADTree	0.59	0.59	0.56	0.55	0.56	0.58	0.62	0.59	0.6	0.59	0.59	0.63	0.63	0.65	0.65	0.65	0.63	0.62	0.61	0.58	0.61	0.61	0.61	0.61	0.61	0.61
ANN	0.56	0.6	0.59	0.62	0.63	0.59	0.62	0.67	0.63	0.61	0.62	0.61	0.63	0.63	0.64	0.6	0.66	0.67	0.64	0.62	0.63	0.61	0.61	0.66	0.6	0.66
BAGGING	0.63	0.63	0.62	0.62	0.62	0.64	0.64	0.65	0.64	0.65	0.64	0.65	0.69	0.65	0.67	0.68	0.65	0.68	0.66	0.67	0.64	0.65	0.66	0.67	0.66	0.67
BayesNet	0.51	0.58	0.58	0.58	0.58	0.54	0.52	0.52	0.53	0.54	0.54	0.55	0.54	0.54	0.54	0.55	0.56	0.56	0.56	0.56	0.55	0.54	0.53	0.54	0.54	0.55
Classification ViaRegression	0.52	0.63	0.58	0.62	0.62	0.63	0.63	0.62	0.61	0.65	0.65	0.62	0.62	0.58	0.59	0.59	0.6	0.59	0.64	0.61	0.63	0.61	0.61	0.62	0.61	0.6
DAGGING	0	0.51	0.48	0.56	0.61	0.6	0.59	0.6	0.59	0.63	0.64	0.68	0.63	0.63	0.68	0.62	0.62	0.63	0.63	0.63	0.67	0.62	0.64	0.63	0.65	0.66
Decision Table	0.56	0.55	0.55	0.55	0.53	0.61	0.61	0.61	0.61	0.6	0.61	0.61	0.61	0.62	0.62	0.62	0.6	0.58	0.58	0.58	0.58	0.59	0.6	0.53	0.51	0.52
Decorate	0.68	0.63	0.64	0.65	0.65	0.69	0.67	0.63	0.64	0.64	0.62	0.63	0.66	0.69	0.65	0.64	0.68	0.65	0.67	0.65	0.67	0.64	0.69	0.66	0.66	0.67
DNTB	0.53	0.56	0.57	0.57	0.55	0.6	0.62	0.61	0.59	0.59	0.59	0.58	0.6	0.6	0.59	0.58	0.56	0.57	0.6	0.6	0.55	0.55	0.59	0.55	0.59	0.6
END	0.64	0.62	0.62	0.61	0.63	0.66	0.64	0.62	0.62	0.62	0.62	0.58	0.62	0.63	0.64	0.62	0.63	0.62	0.65	0.66	0.65	0.65	0.65	0.65	0.65	0.64
IB1	0.58	0.64	0.64	0.62	0.62	0.63	0.62	0.63	0.64	0.66	0.66	0.64	0.65	0.66	0.67	0.67	0.67	0.67	0.67	0.67	0.68	0.65	0.65	0.67	0.68	0.68
KNN	0.6	0.61	0.59	0.62	0.6	0.63	0.62	0.63	0.64	0.66	0.63	0.62	0.62	0.62	0.64	0.65	0.65	0.66	0.65	0.64	0.63	0.63	0.63	0.64	0.64	0.65
Kstar	0.66	0.67	0.71	0.71	0.69	0.64	0.66	0.66	0.66	0.65	0.66	0.66	0.65	0.67	0.67	0.68	0.67	0.69	0.66	0.66	0.67	0.66	0.67	0.66	0.65	0.66
LogicBoost	0.6	0.64	0.63	0.58	0.59	0.64	0.65	0.63	0.66	0.59	0.63	0.67	0.62	0.65	0.68	0.71	0.66	0.69	0.66	0.69	0.65	0.7	0.65	0.7	0.71	0.68
Logistic	0.59	0.62	0.59	0.61	0.63	0.63	0.63	0.63	0.65	0.62	0.62	0.63	0.61	0.62	0.62	0.61	0.61	0.6	0.61	0.61	0.61	0.61	0.6	0.61	0.62	0.64
LWL-random	0.63	0.63	0.63	0.68	0.63	0.64	0.64	0.66	0.68	0.66	0.67	0.64	0.67	0.65	0.67	0.69	0.7	0.67	0.66	0.7	0.65	0.68	0.7	0.67	0.68	0.66
NBTree	0.62	0.59	0.61	0.58	0.58	0.6	0.66	0.67	0.66	0.65	0.66	0.65	0.64	0.63	0.64	0.61	0.58	0.6	0.58	0.6	0.63	0.65	0.62	0.62	0.62	0.61
PART	0.59	0.62	0.61	0.61	0.58	0.64	0.6	0.65	0.6	0.64	0.65	0.62	0.6	0.63	0.65	0.64	0.64	0.62	0.59	0.59	0.63	0.62	0.6	0.62	0.63	0.62
Random Committee	0.58	0.63	0.64	0.65	0.63	0.64	0.65	0.65	0.64	0.66	0.65	0.64	0.67	0.63	0.66	0.68	0.68	0.67	0.67	0.68	0.68	0.66	0.68	0.7	0.68	0.68
Random Forest	0.6	0.63	0.64	0.66	0.65	0.62	0.66	0.64	0.65	0.7	0.69	0.64	0.65	0.63	0.64	0.67	0.66	0.67	0.63	0.65	0.67	0.67	0.66	0.66	0.66	0.69
Random Sub Space	0.66	0.65	0.67	0.67	0.67	0.67	0.69	0.68	0.67	0.68	0.71	0.66	0.67	0.66	0.64	0.7	0.66	0.68	0.63	0.68	0.67	0.69	0.65	0.7	0.66	0.66
REPTree	0.6	0.59	0.62	0.59	0.53	0.53	0.55	0.55	0.55	0.56	0.57	0.56	0.54	0.54	0.54	0.54	0.54	0.59	0.58	0.58	0.61	0.57	0.58	0.56	0.56	0.58
Rotationforest	0.64	0.67	0.67	0.68	0.7	0.7	0.68	0.7	0.73	0.69	0.66	0.65	0.69	0.68	0.67	0.69	0.68	0.65	0.66	0.7	0.67	0.69	0.7	0.69	0.71	0.7
SPegasos	0.6	0.59	0.58	0.6	0.61	0.62	0.61	0.61	0.62	0.61	0.62	0.61	0.6	0.61	0.6	0.6	0.6	0.6	0.61	0.6	0.61	0.61	0.65	0.65	0.63	0.64
SVM	0.62	0.63	0.61	0.62	0.65	0.69	0.68	0.68	0.7	0.68	0.68	0.68	0.68	0.68	0.68	0.69	0.69	0.69	0.69	0.7	0.7	0.69	0.7	0.7	0.71	0.7

T2-NPV

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.73	0.74	0.74	0.74	0.75	0.75	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.75	0.75	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.76	0.76	0.76
ADTree	0.76	0.8	0.78	0.78	0.78	0.74	0.74	0.74	0.73	0.74	0.74	0.75	0.74	0.76	0.78	0.78	0.78	0.79	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79
ANN	0.73	0.78	0.78	0.79	0.81	0.81	0.78	0.81	0.81	0.79	0.81	0.81	0.82	0.8	0.8	0.79	0.81	0.81	0.82	0.79	0.81	0.79	0.79	0.81	0.81	0.84
BAGGING	0.75	0.77	0.77	0.78	0.77	0.77	0.78	0.78	0.78	0.79	0.78	0.79	0.8	0.79	0.79	0.81	0.8	0.8	0.79	0.79	0.79	0.79	0.81	0.8	0.79	0.8
BayesNet	0.77	0.77	0.77	0.77	0.77	0.76	0.82	0.82	0.78	0.79	0.79	0.79	0.81	0.82	0.82	0.84	0.85	0.86	0.86	0.86	0.86	0.85	0.86	0.85	0.85	0.85
Classification ViaRegression	0.7	0.76	0.74	0.75	0.76	0.76	0.77	0.77	0.76	0.77	0.77	0.78	0.76	0.77	0.77	0.77	0.77	0.76	0.78	0.78	0.77	0.79	0.78	0.78	0.77	0.78
DAGGING	0.66	0.67	0.67	0.68	0.68	0.69	0.7	0.7	0.71	0.74	0.72	0.76	0.74	0.74	0.76	0.75	0.75	0.75	0.74	0.75	0.77	0.75	0.76	0.76	0.76	0.77
Decision Table	0.73	0.75	0.75	0.75	0.73	0.74	0.74	0.74	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.76	0.77	0.77	0.76	0.77	0.77	0.77	0.76	0.76
Decorate	0.76	0.79	0.79	0.81	0.79	0.81	0.8	0.78	0.8	0.81	0.79	0.79	0.8	0.82	0.8	0.81	0.82	0.83	0.82	0.82	0.82	0.81	0.83	0.82	0.81	0.82
DNTB	0.73	0.75	0.76	0.76	0.75	0.75	0.76	0.76	0.76	0.79	0.79	0.79	0.8	0.8	0.81	0.83	0.81	0.78	0.79	0.79	0.77	0.78	0.79	0.79	0.8	0.78
END	0.75	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.8	0.79	0.79	0.78	0.79	0.79	0.8	0.8	0.8	0.8	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
IB1	0.8	0.82	0.82	0.8	0.81	0.82	0.81	0.82	0.83	0.83	0.82	0.82	0.83	0.83	0.83	0.83	0.83	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.84
KNN	0.77	0.78	0.79	0.79	0.8	0.82	0.81	0.81	0.81	0.81	0.8	0.81	0.8	0.81	0.81	0.81	0.81	0.82	0.81	0.81	0.81	0.82	0.82	0.82	0.82	0.81
Kstar	0.77	0.8	0.81	0.83	0.82	0.82	0.82	0.82	0.83	0.82	0.82	0.83	0.82	0.83	0.83	0.83	0.82	0.83	0.83	0.82	0.82	0.82	0.83	0.83	0.82	0.82
LogicBoost	0.75	0.77	0.79	0.77	0.77	0.78	0.8	0.79	0.8	0.76	0.8	0.81	0.8	0.8	0.82	0.84	0.81	0.82	0.82	0.83	0.82	0.83	0.82	0.84	0.83	0.84
Logistic	0.7	0.76	0.75	0.77	0.77	0.77	0.77	0.78	0.78	0.77	0.78	0.78	0.77	0.77	0.77	0.77	0.76	0.77	0.77	0.77	0.77	0.77	0.78	0.78	0.79	0.8
LWL-random	0.8	0.81	0.82	0.84	0.84	0.83	0.83	0.85	0.85	0.84	0.84	0.84	0.84	0.84	0.85	0.86	0.86	0.85	0.84	0.86	0.84	0.85	0.87	0.85	0.85	0.84
NBTree	0.78	0.78	0.78	0.78	0.79	0.77	0.79	0.78	0.8	0.8	0.81	0.81	0.8	0.81	0.81	0.82	0.79	0.81	0.79	0.8	0.8	0.82	0.81	0.81	0.82	0.81
PART	0.76	0.78	0.78	0.78	0.8	0.8	0.78	0.79	0.81	0.8	0.78	0.79	0.79	0.82	0.82	0.82	0.82	0.82	0.81	0.78	0.81	0.81	0.8	0.79	0.79	0.79
Random Committee	0.8	0.83	0.83	0.84	0.83	0.83	0.84	0.84	0.83	0.84	0.83	0.84	0.85	0.84	0.85	0.85	0.84	0.85	0.85	0.85	0.85	0.84	0.84	0.86	0.86	0.84
Random Forest	0.8	0.82	0.82	0.85	0.84	0.83	0.83	0.84	0.83	0.84	0.83	0.85	0.85	0.83	0.83	0.85	0.85	0.85	0.83	0.86	0.85	0.84	0.86	0.85	0.84	0.84
Random Sub Space	0.72	0.73	0.76	0.77	0.78	0.77	0.78	0.78	0.78	0.77	0.8	0.77	0.79	0.78	0.78	0.81	0.79	0.8	0.78	0.79	0.79	0.78	0.79	0.8	0.8	0.8
REPTree	0.75	0.76	0.77	0.77	0.75	0.75	0.74	0.75	0.74	0.75	0.75	0.75	0.74	0.77	0.76	0.76	0.76	0.76	0.77	0.75	0.78	0.77	0.77	0.77	0.77	0.77
Rotationforest	0.75	0.78	0.79	0.8	0.8	0.79	0.8	0.81	0.82	0.81	0.81	0.8	0.82	0.82	0.82	0.82	0.83	0.81	0.81	0.84	0.8	0.82	0.82	0.83	0.84	0.81
SPegasos	0.71	0.75	0.75	0.76	0.76	0.76	0.76	0.77	0.77	0.76	0.77	0.76	0.76	0.76	0.76	0.76	0.76	0.77	0.77	0.77	0.77	0.77	0.78	0.78	0.78	0.79
SVM	0.75	0.78	0.78	0.79	0.81	0.82	0.81	0.81	0.81	0.82	0.83	0.83	0.82	0.83	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.83

T2-Sensitivity

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.42	0.44	0.43	0.43	0.48	0.46	0.47	0.47	0.47	0.47	0.47	0.46	0.46	0.45	0.45	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.5	0.5	0.5	0.5
ADTree	0.5	0.62	0.58	0.58	0.58	0.43	0.42	0.41	0.38	0.4	0.4	0.42	0.41	0.47	0.54	0.52	0.55	0.56	0.54	0.61	0.57	0.57	0.57	0.57	0.57	0.57
ANN	0.41	0.56	0.56	0.56	0.64	0.66	0.55	0.62	0.63	0.59	0.63	0.64	0.66	0.6	0.6	0.57	0.61	0.6	0.65	0.56	0.62	0.59	0.59	0.61	0.63	0.69
BAGGING	0.44	0.5	0.51	0.53	0.52	0.52	0.53	0.55	0.55	0.55	0.55	0.56	0.58	0.57	0.57	0.6	0.58	0.59	0.56	0.55	0.56	0.55	0.6	0.57	0.56	0.57
BayesNet	0.59	0.55	0.55	0.55	0.55	0.53	0.72	0.72	0.6	0.63	0.63	0.63	0.67	0.71	0.71	0.74	0.77	0.79	0.78	0.78	0.78	0.77	0.8	0.78	0.77	0.77
Classification ViaRegression	0.25	0.48	0.43	0.45	0.47	0.48	0.52	0.51	0.47	0.48	0.48	0.54	0.5	0.53	0.51	0.54	0.53	0.51	0.55	0.55	0.52	0.58	0.55	0.55	0.53	0.55
DAGGING	0	0.09	0.06	0.12	0.13	0.19	0.21	0.24	0.28	0.38	0.31	0.45	0.39	0.41	0.44	0.43	0.43	0.42	0.4	0.42	0.48	0.46	0.47	0.47	0.47	0.48
Decision Table	0.38	0.5	0.5	0.5	0.42	0.42	0.42	0.42	0.51	0.53	0.53	0.53	0.51	0.51	0.51	0.5	0.5	0.52	0.53	0.53	0.51	0.53	0.53	0.55	0.54	0.53
Decorate	0.46	0.56	0.55	0.62	0.57	0.59	0.58	0.55	0.6	0.61	0.57	0.56	0.58	0.63	0.58	0.63	0.64	0.67	0.63	0.65	0.64	0.63	0.66	0.65	0.63	0.65
DNTB	0.4	0.5	0.49	0.5	0.48	0.45	0.5	0.48	0.51	0.59	0.58	0.59	0.6	0.61	0.65	0.7	0.66	0.58	0.57	0.57	0.56	0.59	0.58	0.6	0.63	0.56
END	0.45	0.59	0.58	0.59	0.58	0.57	0.57	0.56	0.6	0.57	0.56	0.57	0.58	0.57	0.59	0.6	0.61	0.59	0.64	0.66	0.65	0.65	0.65	0.64	0.65	0.65
IB1	0.62	0.65	0.65	0.62	0.64	0.64	0.64	0.65	0.67	0.66	0.65	0.66	0.67	0.67	0.67	0.68	0.68	0.68	0.68	0.67	0.67	0.67	0.67	0.68	0.69	0.7
KNN	0.52	0.56	0.58	0.58	0.61	0.66	0.62	0.63	0.63	0.61	0.61	0.63	0.62	0.62	0.64	0.63	0.63	0.64	0.63	0.62	0.63	0.65	0.65	0.66	0.65	0.63
Kstar	0.49	0.58	0.6	0.65	0.65	0.65	0.66	0.66	0.67	0.65	0.65	0.67	0.65	0.66	0.66	0.66	0.65	0.67	0.66	0.66	0.65	0.65	0.66	0.66	0.65	0.66
LogicBoost	0.44	0.52	0.56	0.53	0.54	0.53	0.59	0.57	0.58	0.5	0.59	0.62	0.59	0.59	0.63	0.67	0.63	0.64	0.66	0.65	0.65	0.67	0.64	0.68	0.66	0.69
Logistic	0.23	0.48	0.48	0.51	0.52	0.52	0.51	0.53	0.53	0.52	0.54	0.53	0.52	0.52	0.52	0.51	0.5	0.51	0.5	0.52	0.52	0.52	0.56	0.55	0.57	0.59
LWL-random	0.61	0.64	0.65	0.7	0.7	0.69	0.69	0.72	0.71	0.69	0.7	0.7	0.7	0.71	0.71	0.74	0.73	0.72	0.71	0.74	0.71	0.71	0.75	0.71	0.72	0.71
NBTree	0.53	0.56	0.56	0.56	0.58	0.52	0.55	0.52	0.57	0.6	0.61	0.63	0.59	0.64	0.62	0.67	0.6	0.63	0.61	0.62	0.61	0.64	0.64	0.63	0.65	0.64
PART	0.5	0.55	0.54	0.54	0.63	0.59	0.56	0.56	0.64	0.6	0.53	0.56	0.58	0.64	0.65	0.66	0.65	0.66	0.65	0.56	0.63	0.63	0.61	0.58	0.57	0.58
Random Committee	0.64	0.67	0.67	0.7	0.68	0.67	0.69	0.69	0.68	0.71	0.68	0.7	0.71	0.71	0.72	0.72	0.7	0.71	0.71	0.71	0.71	0.69	0.7	0.72	0.73	0.69
Random Forest	0.62	0.65	0.66	0.72	0.69	0.69	0.68	0.7	0.67	0.7	0.67	0.73	0.73	0.68	0.69	0.71	0.71	0.71	0.68	0.74	0.71	0.7	0.75	0.72	0.7	0.68
Random Sub Space	0.3	0.34	0.47	0.49	0.53	0.5	0.53	0.51	0.51	0.5	0.57	0.5	0.56	0.54	0.55	0.61	0.57	0.57	0.54	0.55	0.55	0.52	0.57	0.59	0.57	0.59
REPTree	0.46	0.5	0.52	0.52	0.49	0.51	0.45	0.47	0.46	0.48	0.49	0.47	0.47	0.57	0.54	0.52	0.51	0.49	0.54	0.48	0.54	0.53	0.54	0.56	0.55	0.53
Rotationforest	0.44	0.53	0.57	0.59	0.57	0.56	0.57	0.61	0.62	0.61	0.61	0.6	0.63	0.64	0.63	0.64	0.65	0.62	0.62	0.69	0.6	0.62	0.64	0.66	0.67	0.59
SPegasos	0.28	0.46	0.46	0.51	0.49	0.5	0.49	0.5	0.51	0.48	0.5	0.5	0.49	0.5	0.5	0.5	0.5	0.51	0.51	0.51	0.52	0.52	0.53	0.53	0.54	0.55
SVM	0.45	0.55	0.56	0.57	0.61	0.62	0.62	0.61	0.6	0.62	0.65	0.65	0.64	0.65	0.64	0.66	0.66	0.66	0.65	0.66	0.67	0.66	0.67	0.66	0.67	0.67

T2-Specificity

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.8	0.83	0.85	0.85	0.82	0.84	0.84	0.85	0.85	0.85	0.85	0.86	0.86	0.86	0.86	0.84	0.84	0.84	0.84	0.84	0.84	0.83	0.82	0.82	0.82	0.82
ADTree	0.82	0.77	0.76	0.76	0.76	0.84	0.87	0.86	0.87	0.86	0.86	0.88	0.88	0.87	0.85	0.86	0.84	0.82	0.82	0.77	0.81	0.81	0.81	0.81	0.81	0.81
ANN	0.84	0.8	0.8	0.82	0.8	0.77	0.82	0.84	0.81	0.8	0.8	0.79	0.8	0.82	0.82	0.81	0.84	0.85	0.81	0.83	0.81	0.8	0.8	0.84	0.79	0.82
BAGGING	0.87	0.85	0.84	0.83	0.84	0.85	0.85	0.85	0.84	0.85	0.85	0.85	0.86	0.84	0.86	0.85	0.84	0.86	0.85	0.86	0.84	0.85	0.84	0.86	0.85	0.86
BayesNet	0.71	0.8	0.8	0.8	0.8	0.77	0.65	0.65	0.72	0.73	0.73	0.73	0.71	0.69	0.69	0.69	0.69	0.68	0.68	0.68	0.68	0.66	0.64	0.66	0.66	0.67
Classification ViaRegression	0.88	0.86	0.84	0.86	0.85	0.86	0.84	0.84	0.85	0.87	0.87	0.83	0.84	0.8	0.82	0.81	0.82	0.82	0.84	0.82	0.84	0.81	0.82	0.82	0.83	0.81
DAGGING	1	0.96	0.96	0.95	0.96	0.93	0.92	0.92	0.9	0.89	0.91	0.89	0.88	0.88	0.9	0.87	0.87	0.87	0.88	0.87	0.88	0.86	0.87	0.86	0.87	0.87
Decision Table	0.85	0.79	0.79	0.79	0.81	0.86	0.86	0.86	0.84	0.82	0.82	0.82	0.83	0.84	0.84	0.85	0.83	0.81	0.81	0.81	0.81	0.81	0.82	0.75	0.74	0.75
Decorate	0.89	0.83	0.84	0.83	0.85	0.87	0.85	0.83	0.83	0.82	0.82	0.83	0.85	0.85	0.84	0.81	0.85	0.81	0.84	0.82	0.84	0.81	0.85	0.82	0.84	0.83
DNTB	0.82	0.8	0.81	0.8	0.8	0.85	0.84	0.85	0.82	0.79	0.79	0.78	0.79	0.79	0.76	0.74	0.73	0.77	0.8	0.8	0.77	0.75	0.79	0.74	0.77	0.81
END	0.87	0.81	0.82	0.8	0.83	0.85	0.83	0.82	0.81	0.82	0.82	0.79	0.81	0.83	0.83	0.81	0.81	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.81
IB1	0.77	0.81	0.81	0.8	0.8	0.81	0.8	0.81	0.81	0.82	0.83	0.81	0.81	0.82	0.83	0.82	0.82	0.83	0.83	0.83	0.84	0.81	0.82	0.83	0.83	0.83
KNN	0.82	0.82	0.8	0.81	0.79	0.8	0.8	0.81	0.82	0.84	0.81	0.8	0.8	0.81	0.81	0.82	0.82	0.83	0.82	0.82	0.81	0.8	0.8	0.81	0.81	0.83
Kstar	0.87	0.86	0.88	0.86	0.85	0.81	0.82	0.82	0.82	0.82	0.83	0.82	0.82	0.83	0.83	0.84	0.84	0.85	0.83	0.82	0.84	0.83	0.83	0.83	0.82	0.82
LogicBoost	0.85	0.85	0.83	0.8	0.81	0.85	0.84	0.83	0.85	0.82	0.82	0.84	0.81	0.84	0.85	0.86	0.83	0.85	0.82	0.85	0.82	0.85	0.82	0.85	0.86	0.83
Logistic	0.92	0.85	0.83	0.84	0.85	0.85	0.85	0.84	0.85	0.84	0.83	0.84	0.83	0.84	0.84	0.84	0.83	0.83	0.84	0.83	0.83	0.83	0.81	0.81	0.82	0.83
LWL-random	0.81	0.8	0.8	0.83	0.79	0.8	0.8	0.8	0.83	0.81	0.82	0.8	0.82	0.8	0.82	0.82	0.84	0.81	0.81	0.84	0.81	0.82	0.84	0.82	0.83	0.81
NBTree	0.84	0.8	0.81	0.79	0.78	0.82	0.85	0.87	0.85	0.84	0.84	0.82	0.83	0.8	0.82	0.78	0.77	0.78	0.77	0.79	0.81	0.82	0.8	0.8	0.8	0.79
PART	0.82	0.82	0.83	0.82	0.76	0.83	0.81	0.85	0.78	0.82	0.86	0.82	0.8	0.81	0.82	0.81	0.81	0.79	0.77	0.8	0.81	0.8	0.79	0.82	0.83	0.81
Random Committee	0.76	0.8	0.8	0.81	0.79	0.81	0.81	0.81	0.8	0.81	0.81	0.8	0.82	0.78	0.81	0.82	0.83	0.82	0.82	0.82	0.83	0.82	0.84	0.84	0.82	0.83
Random Forest	0.79	0.8	0.81	0.81	0.81	0.78	0.82	0.79	0.81	0.84	0.84	0.79	0.79	0.79	0.8	0.82	0.81	0.82	0.79	0.79	0.82	0.82	0.8	0.81	0.82	0.84
Random Sub Space	0.92	0.91	0.88	0.88	0.87	0.87	0.88	0.88	0.87	0.88	0.88	0.87	0.86	0.86	0.84	0.87	0.85	0.86	0.84	0.87	0.86	0.88	0.84	0.87	0.85	0.85
REPTree	0.85	0.82	0.84	0.82	0.78	0.77	0.81	0.8	0.8	0.8	0.81	0.81	0.79	0.75	0.77	0.78	0.78	0.83	0.8	0.82	0.82	0.8	0.8	0.78	0.78	0.8
Rotationforest	0.88	0.87	0.86	0.86	0.87	0.88	0.86	0.86	0.88	0.86	0.84	0.84	0.85	0.84	0.84	0.86	0.84	0.83	0.84	0.85	0.85	0.85	0.86	0.85	0.86	0.87
SPegasos	0.9	0.84	0.83	0.82	0.84	0.84	0.84	0.84	0.84	0.84	0.85	0.84	0.83	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.85	0.85	0.84	0.84
SVM	0.86	0.83	0.81	0.82	0.84	0.85	0.85	0.85	0.87	0.85	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.86	0.85	0.85	0.86	0.86	0.85

T2-Concordance

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.67	0.7	0.7	0.7	0.7	0.71	0.71	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.73	0.73	0.73	0.73	0.72	0.72	0.72	0.71	0.71	0.71	0.71
ADTree	0.71	0.72	0.7	0.7	0.7	0.7	0.71	0.7	0.7	0.7	0.7	0.72	0.72	0.73	0.75	0.74	0.74	0.74	0.73	0.72	0.73	0.73	0.73	0.73	0.73	0.73
ANN	0.69	0.72	0.72	0.73	0.75	0.73	0.73	0.77	0.75	0.73	0.75	0.74	0.75	0.75	0.75	0.73	0.76	0.76	0.76	0.74	0.75	0.73	0.73	0.76	0.73	0.77
BAGGING	0.72	0.73	0.73	0.73	0.73	0.74	0.74	0.75	0.74	0.75	0.74	0.75	0.77	0.75	0.76	0.77	0.75	0.77	0.75	0.76	0.75	0.75	0.76	0.76	0.76	0.76
BayesNet	0.67	0.71	0.71	0.71	0.71	0.68	0.68	0.68	0.68	0.69	0.69	0.7	0.7	0.7	0.7	0.71	0.71	0.72	0.72	0.72	0.71	0.7	0.69	0.7	0.7	0.7
Classification ViaRegression	0.67	0.73	0.7	0.72	0.72	0.73	0.73	0.73	0.72	0.74	0.74	0.73	0.72	0.71	0.71	0.71	0.72	0.71	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.72
DAGGING	0.66	0.66	0.66	0.67	0.68	0.68	0.68	0.69	0.69	0.71	0.71	0.74	0.71	0.72	0.74	0.72	0.72	0.72	0.72	0.72	0.74	0.72	0.73	0.72	0.73	0.74
Decision Table	0.69	0.69	0.69	0.69	0.68	0.71	0.71	0.71	0.72	0.72	0.72	0.72	0.72	0.73	0.73	0.73	0.72	0.71	0.71	0.71	0.71	0.71	0.72	0.68	0.67	0.67
Decorate	0.74	0.74	0.74	0.76	0.75	0.77	0.76	0.74	0.75	0.75	0.74	0.74	0.76	0.78	0.75	0.75	0.78	0.77	0.77	0.76	0.77	0.75	0.79	0.77	0.77	0.77
DNTB	0.68	0.69	0.7	0.7	0.69	0.71	0.72	0.72	0.71	0.72	0.72	0.71	0.73	0.73	0.72	0.73	0.71	0.71	0.72	0.72	0.7	0.7	0.72	0.7	0.72	0.72
END	0.73	0.74	0.74	0.73	0.74	0.75	0.75	0.74	0.74	0.74	0.73	0.72	0.73	0.74	0.75	0.74	0.75	0.74	0.76	0.77	0.76	0.76	0.76	0.76	0.76	0.76
IB1	0.72	0.76	0.76	0.74	0.75	0.75	0.75	0.75	0.76	0.77	0.77	0.76	0.77	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.77	0.78	0.78	0.79
KNN	0.72	0.73	0.72	0.73	0.73	0.76	0.74	0.75	0.75	0.76	0.75	0.74	0.74	0.75	0.75	0.76	0.76	0.77	0.76	0.75	0.75	0.75	0.75	0.76	0.76	0.76
Kstar	0.74	0.76	0.78	0.79	0.78	0.76	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.77	0.77	0.78	0.77	0.79	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.77
LogicBoost	0.71	0.74	0.74	0.71	0.71	0.74	0.75	0.74	0.76	0.71	0.75	0.77	0.74	0.76	0.77	0.8	0.76	0.78	0.77	0.78	0.77	0.79	0.76	0.79	0.79	0.78
Logistic	0.68	0.72	0.71	0.72	0.73	0.73	0.73	0.74	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.73	0.73	0.73	0.75
LWL-random	0.75	0.75	0.75	0.78	0.76	0.76	0.76	0.78	0.79	0.77	0.78	0.77	0.78	0.77	0.78	0.8	0.8	0.78	0.78	0.8	0.77	0.79	0.81	0.78	0.79	0.78
NBTree	0.73	0.72	0.73	0.71	0.72	0.72	0.75	0.75	0.76	0.76	0.76	0.76	0.75	0.75	0.75	0.74	0.71	0.73	0.72	0.73	0.75	0.76	0.74	0.74	0.75	0.74
PART	0.71	0.73	0.73	0.72	0.72	0.75	0.72	0.75	0.73	0.75	0.75	0.74	0.73	0.75	0.76	0.76	0.76	0.75	0.73	0.72	0.75	0.75	0.73	0.74	0.74	0.73
Random Committee	0.72	0.76	0.76	0.77	0.76	0.76	0.77	0.77	0.76	0.78	0.77	0.77	0.79	0.76	0.78	0.79	0.79	0.78	0.78	0.79	0.79	0.78	0.79	0.8	0.79	0.78
Random Forest	0.73	0.75	0.76	0.78	0.77	0.75	0.77	0.76	0.76	0.79	0.78	0.77	0.77	0.75	0.76	0.78	0.78	0.78	0.76	0.78	0.78	0.78	0.79	0.78	0.78	0.79
Random Sub Space	0.71	0.71	0.74	0.75	0.75	0.75	0.76	0.75	0.75	0.75	0.78	0.74	0.76	0.75	0.74	0.78	0.76	0.76	0.73	0.76	0.75	0.76	0.75	0.77	0.76	0.76
REPTree	0.71	0.71	0.73	0.72	0.68	0.68	0.69	0.69	0.69	0.69	0.7	0.69	0.68	0.69	0.69	0.69	0.69	0.71	0.71	0.7	0.72	0.71	0.71	0.7	0.7	0.71
Rotationforest	0.73	0.75	0.76	0.77	0.77	0.77	0.76	0.78	0.79	0.78	0.76	0.76	0.78	0.78	0.77	0.78	0.78	0.76	0.76	0.79	0.76	0.78	0.78	0.79	0.79	0.78
SPegasos	0.69	0.71	0.7	0.72	0.72	0.72	0.72	0.72	0.73	0.72	0.73	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.74	0.74	0.74	0.75
SVM	0.72	0.74	0.73	0.73	0.76	0.78	0.77	0.77	0.78	0.77	0.78	0.78	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79

T2-ROC AUC

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.65	0.77	0.77	0.77	0.76	0.76	0.76	0.76	0.76	0.75	0.75	0.77	0.77	0.77	0.77	0.78	0.78	0.78	0.78	0.77	0.77	0.77	0.76	0.76	0.76	0.76
ADTree	0.71	0.77	0.77	0.76	0.77	0.76	0.77	0.76	0.75	0.75	0.75	0.76	0.76	0.77	0.79	0.82	0.8	0.8	0.8	0.81	0.81	0.81	0.81	0.81	0.81	0.81
ANN	0.67	0.77	0.78	0.79	0.81	0.8	0.79	0.82	0.8	0.79	0.82	0.8	0.8	0.8	0.81	0.78	0.81	0.82	0.83	0.79	0.8	0.79	0.79	0.82	0.79	0.82
Bagging	0.74	0.79	0.8	0.8	0.8	0.8	0.8	0.8	0.79	0.8	0.8	0.8	0.81	0.81	0.82	0.82	0.82	0.83	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
BayNet	0.61	0.72	0.72	0.72	0.72	0.73	0.74	0.74	0.75	0.75	0.75	0.75	0.76	0.76	0.76	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
ClassificationViaRegression	0.65	0.77	0.75	0.77	0.77	0.78	0.78	0.77	0.77	0.79	0.79	0.78	0.78	0.78	0.77	0.77	0.79	0.78	0.79	0.79	0.77	0.79	0.79	0.78	0.77	0.79
Dagging	0.57	0.71	0.71	0.71	0.75	0.74	0.73	0.73	0.75	0.74	0.73	0.76	0.76	0.76	0.75	0.74	0.76	0.76	0.75	0.75	0.75	0.74	0.76	0.77	0.77	0.77
DecisionTable	0.69	0.73	0.73	0.73	0.7	0.71	0.72	0.72	0.76	0.75	0.76	0.76	0.77	0.76	0.76	0.77	0.76	0.74	0.74	0.74	0.75	0.75	0.77	0.74	0.72	0.73
Decorate	0.69	0.8	0.8	0.81	0.81	0.82	0.81	0.8	0.81	0.83	0.81	0.81	0.82	0.84	0.83	0.82	0.83	0.81	0.83	0.83	0.83	0.83	0.82	0.81	0.82	0.83
DNTB	0.66	0.74	0.74	0.74	0.74	0.75	0.77	0.77	0.77	0.78	0.78	0.77	0.77	0.78	0.78	0.79	0.78	0.76	0.76	0.76	0.75	0.74	0.77	0.75	0.74	0.75
END	0.7	0.77	0.76	0.77	0.77	0.77	0.75	0.75	0.75	0.73	0.74	0.72	0.73	0.75	0.75	0.73	0.75	0.75	0.77	0.78	0.77	0.76	0.76	0.74	0.75	0.74
IB1	0.7	0.73	0.73	0.71	0.72	0.73	0.72	0.73	0.74	0.74	0.74	0.73	0.74	0.75	0.75	0.75	0.75	0.76	0.75	0.75	0.75	0.74	0.75	0.75	0.76	0.77
KNN	0.75	0.79	0.79	0.79	0.79	0.81	0.8	0.8	0.8	0.81	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.81	0.81	0.8	0.8	0.8	0.8	0.8	0.8	0.81
Kstar	0.77	0.84	0.84	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.84	0.85	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.83	0.83	0.83
Logistic	0.68	0.77	0.77	0.78	0.78	0.78	0.77	0.77	0.78	0.78	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.79	0.78	0.78	0.79
LogitBoost	0.69	0.79	0.81	0.79	0.78	0.81	0.83	0.81	0.81	0.77	0.81	0.82	0.81	0.82	0.82	0.85	0.82	0.83	0.83	0.84	0.82	0.84	0.84	0.84	0.84	0.84
LWL	0.76	0.79	0.81	0.82	0.81	0.82	0.81	0.84	0.83	0.83	0.84	0.81	0.84	0.82	0.82	0.85	0.84	0.84	0.83	0.84	0.82	0.84	0.85	0.83	0.83	0.84
NBTree	0.68	0.74	0.75	0.75	0.74	0.75	0.78	0.76	0.79	0.8	0.8	0.8	0.77	0.79	0.78	0.8	0.77	0.77	0.76	0.79	0.79	0.8	0.78	0.78	0.79	0.78
PART	0.71	0.78	0.76	0.76	0.77	0.78	0.76	0.77	0.78	0.77	0.76	0.76	0.73	0.75	0.76	0.77	0.76	0.76	0.75	0.71	0.75	0.75	0.74	0.73	0.74	0.7
RandomCommittee	0.75	0.8	0.8	0.81	0.81	0.82	0.83	0.83	0.83	0.82	0.83	0.83	0.82	0.82	0.84	0.85	0.83	0.84	0.85	0.85	0.85	0.84	0.82	0.85	0.84	0.83
RandomForest	0.76	0.81	0.82	0.83	0.82	0.82	0.84	0.82	0.83	0.84	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.84	0.82	0.84	0.84	0.85	0.84	0.84	0.83	0.84
RandomSubSpace	0.72	0.79	0.81	0.82	0.82	0.81	0.83	0.81	0.82	0.81	0.82	0.81	0.82	0.82	0.8	0.83	0.82	0.83	0.8	0.82	0.82	0.82	0.83	0.83	0.82	0.83
RatationForest	0.74	0.82	0.82	0.82	0.84	0.83	0.83	0.83	0.85	0.84	0.83	0.83	0.83	0.84	0.84	0.84	0.84	0.83	0.83	0.85	0.84	0.84	0.85	0.84	0.85	0.84
RepTree	0.7	0.74	0.77	0.76	0.72	0.73	0.72	0.73	0.72	0.72	0.72	0.73	0.71	0.72	0.72	0.71	0.73	0.74	0.74	0.73	0.74	0.73	0.75	0.74	0.74	0.75
Spegasos	0.67	0.77	0.76	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.76	0.76	0.76	0.76	0.77	0.76	0.78	0.78	0.78	0.78
SVM	0.66	0.69	0.69	0.69	0.72	0.74	0.73	0.73	0.74	0.74	0.75	0.75	0.74	0.75	0.74	0.75	0.75	0.75	0.75	0.76	0.76	0.76	0.76	0.76	0.76	0.76

4.3 Training Set 1 and Set 2 (T1+T2)

T1+T2-PPV

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.52	0.57	0.54	0.56	0.56	0.56	0.59	0.57	0.58	0.58	0.58	0.59	0.59	0.59	0.59	0.61	0.61	0.59	0.59	0.57	0.58	0.60	0.58	0.58	0.58	0.59
ADTree	0.63	0.63	0.65	0.66	0.65	0.65	0.67	0.67	0.68	0.67	0.67	0.65	0.67	0.67	0.68	0.68	0.67	0.67	0.70	0.70	0.71	0.71	0.71	0.71	0.70	0.71
ANN	0.59	0.57	0.56	0.53	0.54	0.55	0.56	0.61	0.60	0.60	0.59	0.58	0.63	0.59	0.61	0.62	0.60	0.60	0.61	0.62	0.63	0.61	0.63	0.60	0.63	0.64
BAGGING	0.57	0.58	0.61	0.62	0.62	0.63	0.65	0.67	0.65	0.65	0.64	0.64	0.67	0.66	0.65	0.66	0.66	0.66	0.69	0.68	0.69	0.68	0.68	0.66	0.58	0.67
BayesNet	0.61	0.60	0.62	0.64	0.64	0.62	0.63	0.63	0.63	0.65	0.67	0.65	0.65	0.68	0.64	0.66	0.66	0.67	0.66	0.67	0.68	0.68	0.69	0.66	0.67	0.69
Classification ViaRegression	0.56	0.56	0.61	0.60	0.63	0.62	0.63	0.65	0.63	0.66	0.66	0.63	0.65	0.65	0.66	0.67	0.65	0.68	0.68	0.66	0.68	0.66	0.66	0.66	0.67	0.67
DAGGING	0.51	0.54	0.53	0.57	0.54	0.54	0.57	0.58	0.56	0.57	0.57	0.54	0.55	0.55	0.55	0.56	0.57	0.57	0.58	0.57	0.57	0.59	0.59	0.60	0.60	0.60
Decision Table	0.58	0.58	0.60	0.61	0.61	0.60	0.59	0.58	0.62	0.61	0.62	0.64	0.59	0.61	0.60	0.60	0.60	0.60	0.60	0.63	0.60	0.62	0.59	0.59	0.62	0.63
Decorate	0.62	0.58	0.60	0.56	0.57	0.57	0.59	0.58	0.55	0.59	0.59	0.59	0.58	0.61	0.61	0.64	0.63	0.63	0.61	0.64	0.61	0.60	0.62	0.61	0.63	0.61
DNTB	0.46	0.46	0.47	0.49	0.48	0.50	0.51	0.51	0.50	0.49	0.49	0.50	0.52	0.53	0.53	0.53	0.54	0.54	0.56	0.56	0.54	0.53	0.54	0.54	0.54	0.54
END	1.00	1.00	0.67	0.50	0.80	0.75	0.60	0.60	0.66	0.66	0.62	0.64	0.66	0.64	0.66	0.59	0.63	0.64	0.62	0.66	0.62	0.61	0.64	0.64	0.64	0.64
IB1	0.59	0.61	0.60	0.58	0.60	0.55	0.62	0.60	0.58	0.56	0.56	0.56	0.61	0.57	0.59	0.59	0.61	0.60	0.59	0.60	0.60	0.61	0.59	0.59	0.57	0.61
KNN	0.66	0.64	0.66	0.63	0.68	0.72	0.68	0.70	0.67	0.67	0.69	0.68	0.70	0.70	0.68	0.69	0.71	0.70	0.69	0.68	0.69	0.70	0.71	0.71	0.70	0.69
Kstar	0.65	0.61	0.65	0.67	0.65	0.66	0.66	0.64	0.65	0.68	0.68	0.66	0.65	0.65	0.68	0.67	0.67	0.68	0.69	0.68	0.68	0.67	0.69	0.68	0.69	0.69
LogicBoost	0.58	0.57	0.59	0.61	0.61	0.63	0.63	0.64	0.65	0.64	0.63	0.64	0.66	0.65	0.66	0.65	0.65	0.64	0.64	0.65	0.68	0.67	0.70	0.65	0.68	0.66
Logistic	0.58	0.57	0.61	0.61	0.61	0.60	0.61	0.61	0.60	0.63	0.63	0.64	0.64	0.64	0.63	0.64	0.63	0.64	0.64	0.64	0.64	0.63	0.64	0.63	0.64	0.64
LWL-random	0.51	0.55	0.59	0.62	0.63	0.62	0.60	0.61	0.59	0.58	0.59	0.61	0.62	0.64	0.62	0.63	0.62	0.60	0.62	0.63	0.62	0.63	0.63	0.61	0.61	0.64
NBTree	0.64	0.63	0.62	0.66	0.64	0.63	0.63	0.63	0.62	0.64	0.63	0.63	0.64	0.65	0.64	0.65	0.65	0.64	0.66	0.67	0.67	0.67	0.66	0.66	0.67	0.67
PART	0.64	0.65	0.64	0.65	0.65	0.66	0.67	0.67	0.66	0.68	0.69	0.68	0.69	0.69	0.69	0.70	0.69	0.69	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
Random Committee	0.62	0.61	0.60	0.62	0.62	0.62	0.62	0.62	0.62	0.63	0.63	0.63	0.64	0.65	0.63	0.64	0.65	0.64	0.66	0.66	0.66	0.67	0.66	0.65	0.66	0.66
Random Forest	0.57	0.58	0.58	0.57	0.58	0.57	0.58	0.56	0.59	0.59	0.61	0.61	0.58	0.60	0.59	0.59	0.59	0.59	0.59	0.58	0.58	0.61	0.60	0.60	0.58	0.58
Random Sub Space	0.55	0.57	0.59	0.58	0.59	0.60	0.60	0.59	0.54	0.61	0.58	0.58	0.62	0.60	0.59	0.62	0.57	0.63	0.61	0.60	0.60	0.60	0.62	0.62	0.59	0.63
REPTree	0.61	0.60	0.64	0.64	0.64	0.66	0.65	0.65	0.65	0.64	0.66	0.67	0.66	0.65	0.65	0.65	0.61	0.64	0.66	0.66	0.67	0.69	0.66	0.66	0.68	0.66
Rotationforest	0.54	0.55	0.57	0.59	0.57	0.58	0.54	0.54	0.58	0.58	0.58	0.57	0.54	0.54	0.56	0.54	0.53	0.55	0.57	0.57	0.58	0.59	0.58	0.58	0.58	0.58
SPegasos	0.58	0.58	0.58	0.55	0.55	0.55	0.59	0.59	0.58	0.58	0.58	0.58	0.60	0.58	0.59	0.59	0.57	0.59	0.59	0.59	0.62	0.61	0.60	0.60	0.60	0.60
SVM	0.48	0.47	0.38	0.42	0.42	0.49	0.56	0.55	0.54	0.55	0.55	0.55	0.55	0.54	0.54	0.54	0.54	0.55	0.61	0.61	0.59	0.55	0.55	0.55	0.51	0.51

T1+T2-NPV

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.67	0.69	0.68	0.69	0.69	0.70	0.73	0.72	0.73	0.72	0.72	0.74	0.76	0.75	0.76	0.77	0.77	0.76	0.75	0.76	0.76	0.76	0.75	0.74	0.75	0.74
ADTree	0.73	0.73	0.74	0.75	0.75	0.75	0.75	0.75	0.75	0.76	0.75	0.75	0.76	0.77	0.76	0.76	0.77	0.76	0.77	0.77	0.77	0.77	0.78	0.77	0.77	0.77
ANN	0.65	0.65	0.64	0.64	0.65	0.65	0.67	0.68	0.68	0.68	0.68	0.68	0.69	0.68	0.68	0.69	0.69	0.69	0.69	0.69	0.70	0.70	0.71	0.71	0.71	0.71
BAGGING	0.74	0.75	0.78	0.78	0.78	0.79	0.80	0.81	0.79	0.80	0.80	0.80	0.81	0.82	0.81	0.81	0.80	0.81	0.82	0.81	0.81	0.81	0.82	0.81	0.76	0.81
BayesNet	0.72	0.72	0.74	0.74	0.74	0.73	0.73	0.74	0.74	0.76	0.76	0.76	0.76	0.77	0.76	0.77	0.77	0.78	0.77	0.78	0.77	0.77	0.78	0.77	0.77	0.78
Classification ViaRegression	0.74	0.74	0.77	0.78	0.79	0.78	0.79	0.81	0.79	0.81	0.81	0.78	0.80	0.80	0.81	0.81	0.81	0.82	0.83	0.81	0.81	0.81	0.81	0.81	0.82	0.82
DAGGING	0.66	0.67	0.67	0.69	0.68	0.68	0.71	0.71	0.71	0.70	0.71	0.69	0.72	0.72	0.72	0.72	0.74	0.74	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75
Decision Table	0.70	0.72	0.73	0.75	0.76	0.76	0.75	0.75	0.76	0.76	0.76	0.77	0.74	0.75	0.75	0.75	0.76	0.75	0.75	0.77	0.75	0.76	0.74	0.75	0.75	0.76
Decorate	0.70	0.73	0.72	0.71	0.73	0.72	0.72	0.73	0.74	0.75	0.75	0.76	0.75	0.76	0.77	0.76	0.78	0.77	0.76	0.78	0.75	0.76	0.76	0.76	0.78	0.75
DNTB	0.64	0.64	0.64	0.66	0.66	0.72	0.71	0.71	0.72	0.72	0.72	0.74	0.75	0.76	0.76	0.76	0.77	0.77	0.78	0.76	0.76	0.76	0.77	0.77	0.77	0.77
END	0.63	0.63	0.63	0.63	0.63	0.63	0.64	0.64	0.65	0.65	0.65	0.66	0.66	0.65	0.65	0.65	0.66	0.66	0.66	0.67	0.66	0.67	0.68	0.68	0.68	0.68
IB1	0.69	0.70	0.69	0.71	0.69	0.73	0.72	0.72	0.74	0.72	0.72	0.72	0.74	0.76	0.75	0.71	0.75	0.73	0.77	0.76	0.77	0.75	0.75	0.77	0.75	0.73
KNN	0.71	0.72	0.73	0.72	0.74	0.76	0.74	0.75	0.75	0.77	0.75	0.76	0.77	0.77	0.77	0.77	0.78	0.77	0.78	0.78	0.78	0.77	0.79	0.79	0.79	0.78
Kstar	0.69	0.67	0.72	0.72	0.71	0.73	0.73	0.73	0.73	0.75	0.76	0.74	0.74	0.74	0.75	0.74	0.74	0.76	0.76	0.75	0.75	0.75	0.75	0.76	0.75	0.76
LogicBoost	0.75	0.75	0.76	0.78	0.78	0.78	0.79	0.79	0.80	0.79	0.79	0.80	0.82	0.80	0.80	0.81	0.79	0.80	0.80	0.79	0.82	0.81	0.82	0.82	0.81	0.82
Logistic	0.75	0.75	0.77	0.78	0.78	0.77	0.78	0.78	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.80	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
LWL-random	0.68	0.71	0.71	0.72	0.73	0.72	0.68	0.68	0.69	0.72	0.72	0.72	0.75	0.74	0.73	0.75	0.74	0.74	0.75	0.75	0.74	0.75	0.76	0.75	0.76	0.71
NBTree	0.75	0.75	0.76	0.78	0.78	0.77	0.78	0.78	0.78	0.79	0.79	0.79	0.80	0.81	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.81	0.81
PART	0.72	0.75	0.74	0.75	0.75	0.76	0.77	0.77	0.77	0.78	0.78	0.78	0.79	0.78	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.80	0.80	0.79	0.80
Random Committee	0.75	0.76	0.76	0.77	0.77	0.77	0.77	0.78	0.78	0.79	0.77	0.78	0.78	0.79	0.78	0.77	0.78	0.78	0.78	0.79	0.79	0.80	0.79	0.79	0.79	0.79
Random Forest	0.71	0.72	0.72	0.72	0.72	0.71	0.73	0.72	0.73	0.73	0.73	0.74	0.73	0.75	0.74	0.75	0.74	0.73	0.73	0.73	0.73	0.74	0.74	0.74	0.74	0.74
Random Sub Space	0.68	0.69	0.71	0.72	0.72	0.75	0.74	0.75	0.73	0.77	0.73	0.74	0.76	0.74	0.75	0.76	0.75	0.78	0.74	0.75	0.76	0.76	0.76	0.79	0.75	0.77
REPTree	0.71	0.73	0.74	0.75	0.74	0.77	0.77	0.76	0.77	0.77	0.77	0.79	0.78	0.77	0.79	0.77	0.78	0.77	0.79	0.78	0.79	0.78	0.78	0.77	0.79	0.78
Rotationforest	0.72	0.72	0.71	0.72	0.70	0.70	0.71	0.71	0.71	0.72	0.72	0.71	0.72	0.71	0.70	0.72	0.72	0.73	0.73	0.74	0.72	0.72	0.71	0.71	0.72	0.72
SPegasos	0.65	0.65	0.65	0.65	0.66	0.66	0.68	0.68	0.68	0.69	0.69	0.69	0.70	0.70	0.70	0.70	0.69	0.70	0.70	0.70	0.72	0.72	0.72	0.72	0.72	0.72
SVM	0.64	0.64	0.63	0.63	0.63	0.63	0.68	0.68	0.68	0.68	0.68	0.68	0.67	0.67	0.67	0.67	0.67	0.68	0.68	0.68	0.68	0.68	0.69	0.69	0.68	0.68

T1+T2-Sensitivity

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.31	0.36	0.36	0.40	0.39	0.45	0.53	0.50	0.53	0.50	0.50	0.55	0.60	0.59	0.61	0.61	0.62	0.62	0.58	0.61	0.62	0.60	0.58	0.57	0.58	0.57
ADTree	0.50	0.50	0.51	0.55	0.53	0.54	0.54	0.53	0.54	0.57	0.54	0.54	0.57	0.58	0.56	0.57	0.57	0.55	0.59	0.57	0.57	0.58	0.59	0.59	0.57	0.58
ANN	0.14	0.14	0.14	0.14	0.20	0.20	0.28	0.33	0.30	0.31	0.31	0.31	0.34	0.34	0.33	0.35	0.35	0.35	0.37	0.37	0.39	0.39	0.43	0.45	0.43	0.43
BAGGING	0.58	0.59	0.63	0.64	0.65	0.65	0.68	0.68	0.66	0.68	0.69	0.68	0.69	0.71	0.70	0.69	0.68	0.69	0.70	0.69	0.69	0.70	0.72	0.70	0.61	0.70
BayesNet	0.46	0.49	0.53	0.51	0.52	0.48	0.51	0.53	0.53	0.56	0.57	0.57	0.57	0.58	0.58	0.58	0.58	0.61	0.59	0.62	0.59	0.58	0.61	0.60	0.59	0.61
Classification ViaRegression	0.58	0.57	0.62	0.65	0.65	0.64	0.66	0.69	0.66	0.69	0.69	0.64	0.67	0.67	0.69	0.70	0.69	0.71	0.73	0.69	0.68	0.69	0.70	0.69	0.71	0.71
DAGGING	0.30	0.33	0.33	0.37	0.36	0.38	0.47	0.47	0.46	0.45	0.45	0.43	0.52	0.52	0.53	0.53	0.57	0.55	0.56	0.56	0.56	0.57	0.57	0.58	0.57	0.57
Decision Table	0.44	0.51	0.49	0.57	0.59	0.61	0.57	0.57	0.59	0.59	0.58	0.59	0.56	0.57	0.57	0.58	0.61	0.59	0.58	0.60	0.57	0.59	0.55	0.57	0.57	0.58
Decorate	0.39	0.51	0.47	0.47	0.53	0.51	0.50	0.53	0.57	0.58	0.58	0.61	0.58	0.61	0.62	0.57	0.63	0.61	0.58	0.62	0.57	0.59	0.60	0.61	0.63	0.58
DNTB	0.17	0.17	0.21	0.33	0.34	0.57	0.53	0.53	0.58	0.59	0.59	0.64	0.66	0.66	0.67	0.67	0.67	0.67	0.67	0.64	0.65	0.66	0.67	0.67	0.68	0.68
END	0.01	0.00	0.00	0.01	0.03	0.02	0.11	0.13	0.13	0.16	0.15	0.18	0.17	0.17	0.16	0.18	0.20	0.21	0.21	0.23	0.21	0.25	0.28	0.31	0.29	0.28
IB1	0.38	0.40	0.37	0.45	0.37	0.56	0.48	0.49	0.55	0.51	0.51	0.52	0.54	0.62	0.59	0.45	0.57	0.52	0.63	0.60	0.62	0.56	0.57	0.64	0.59	0.52
KNN	0.42	0.45	0.47	0.45	0.48	0.53	0.50	0.53	0.53	0.58	0.54	0.56	0.58	0.57	0.59	0.59	0.59	0.59	0.61	0.61	0.61	0.59	0.63	0.63	0.64	0.61
Kstar	0.34	0.28	0.45	0.44	0.43	0.46	0.48	0.48	0.49	0.52	0.55	0.52	0.50	0.52	0.53	0.50	0.51	0.55	0.54	0.53	0.54	0.52	0.52	0.57	0.54	0.54
LogicBoost	0.59	0.59	0.62	0.64	0.65	0.65	0.65	0.65	0.66	0.66	0.66	0.68	0.71	0.69	0.66	0.70	0.65	0.67	0.67	0.65	0.70	0.69	0.70	0.73	0.68	0.72
Logistic	0.60	0.59	0.62	0.64	0.64	0.64	0.65	0.66	0.65	0.67	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.67	0.65	0.65	0.65	0.65	0.66	0.65	0.65	0.65
LWL-random	0.39	0.47	0.44	0.47	0.48	0.47	0.32	0.33	0.36	0.48	0.47	0.47	0.56	0.53	0.50	0.54	0.54	0.54	0.57	0.56	0.53	0.56	0.57	0.57	0.60	0.43
NBTree	0.55	0.56	0.58	0.62	0.62	0.62	0.63	0.64	0.65	0.65	0.66	0.66	0.68	0.69	0.67	0.67	0.68	0.68	0.66	0.67	0.68	0.68	0.68	0.67	0.69	0.68
PART	0.45	0.53	0.53	0.55	0.55	0.56	0.57	0.58	0.59	0.60	0.61	0.61	0.62	0.62	0.63	0.63	0.62	0.63	0.63	0.63	0.63	0.63	0.64	0.64	0.64	0.64
Random Committee	0.55	0.59	0.61	0.61	0.61	0.61	0.62	0.64	0.63	0.65	0.62	0.63	0.64	0.64	0.62	0.62	0.63	0.62	0.63	0.64	0.65	0.66	0.66	0.66	0.66	0.65
Random Forest	0.46	0.47	0.49	0.49	0.48	0.48	0.53	0.50	0.50	0.52	0.51	0.53	0.52	0.57	0.56	0.57	0.54	0.53	0.54	0.53	0.52	0.53	0.55	0.54	0.55	0.55
Random Sub Space	0.36	0.39	0.45	0.50	0.50	0.57	0.54	0.57	0.55	0.61	0.52	0.55	0.60	0.56	0.57	0.60	0.60	0.63	0.54	0.59	0.59	0.60	0.59	0.65	0.58	0.61
REPTree	0.43	0.51	0.51	0.54	0.53	0.59	0.60	0.58	0.59	0.59	0.59	0.63	0.61	0.60	0.64	0.60	0.63	0.61	0.64	0.61	0.64	0.60	0.62	0.61	0.63	0.63
Rotationforest	0.54	0.51	0.45	0.47	0.44	0.40	0.50	0.50	0.44	0.51	0.51	0.46	0.52	0.48	0.45	0.53	0.52	0.54	0.53	0.56	0.48	0.49	0.46	0.47	0.50	0.50
SPegasos	0.16	0.16	0.16	0.17	0.22	0.23	0.32	0.33	0.35	0.37	0.37	0.38	0.40	0.41	0.40	0.40	0.39	0.42	0.41	0.41	0.46	0.46	0.48	0.48	0.48	0.48
SVM	0.14	0.13	0.03	0.07	0.07	0.10	0.35	0.35	0.35	0.35	0.35	0.35	0.31	0.32	0.32	0.32	0.32	0.35	0.32	0.32	0.34	0.37	0.40	0.40	0.41	0.41

T1+T2-Specificity

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.83	0.83	0.81	0.82	0.81	0.79	0.78	0.77	0.77	0.78	0.78	0.77	0.75	0.75	0.75	0.77	0.76	0.74	0.76	0.73	0.74	0.76	0.75	0.75	0.75	0.76
ADTree	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.84	0.85	0.83	0.84	0.83	0.83	0.83	0.84	0.84	0.83	0.83	0.85	0.85	0.86	0.86	0.86	0.85	0.85	0.86
ANN	0.94	0.93	0.93	0.92	0.90	0.90	0.87	0.87	0.88	0.88	0.87	0.87	0.88	0.86	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.85	0.85	0.82	0.85	0.85
BAGGING	0.74	0.74	0.76	0.77	0.76	0.77	0.79	0.80	0.79	0.78	0.77	0.77	0.79	0.78	0.78	0.78	0.79	0.79	0.81	0.80	0.81	0.80	0.79	0.79	0.73	0.79
BayesNet	0.83	0.81	0.80	0.83	0.82	0.83	0.82	0.81	0.81	0.82	0.83	0.82	0.81	0.84	0.80	0.82	0.82	0.82	0.82	0.81	0.83	0.84	0.84	0.82	0.82	0.84
Classification ViaRegression	0.73	0.74	0.76	0.74	0.77	0.77	0.76	0.78	0.77	0.79	0.78	0.77	0.79	0.79	0.79	0.79	0.78	0.80	0.79	0.79	0.80	0.79	0.78	0.79	0.79	0.79
DAGGING	0.83	0.84	0.83	0.83	0.82	0.80	0.79	0.80	0.78	0.79	0.79	0.78	0.75	0.74	0.74	0.75	0.74	0.75	0.76	0.75	0.75	0.76	0.76	0.76	0.77	0.77
Decision Table	0.81	0.78	0.80	0.78	0.77	0.75	0.76	0.75	0.78	0.78	0.79	0.80	0.77	0.78	0.77	0.77	0.75	0.76	0.76	0.78	0.77	0.78	0.77	0.76	0.79	0.80
Decorate	0.86	0.78	0.81	0.78	0.76	0.77	0.79	0.77	0.72	0.75	0.76	0.75	0.75	0.76	0.76	0.81	0.77	0.79	0.78	0.79	0.78	0.76	0.78	0.76	0.78	0.78
DNTB	0.88	0.88	0.85	0.79	0.78	0.66	0.70	0.69	0.66	0.63	0.63	0.61	0.63	0.64	0.64	0.64	0.66	0.66	0.68	0.69	0.66	0.65	0.65	0.65	0.65	0.65
END	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.95	0.96	0.95	0.95	0.94	0.95	0.94	0.95	0.93	0.93	0.93	0.92	0.93	0.93	0.91	0.90	0.90	0.90	0.90
IB1	0.84	0.85	0.86	0.80	0.85	0.72	0.82	0.80	0.76	0.76	0.76	0.76	0.79	0.72	0.76	0.81	0.77	0.80	0.73	0.76	0.75	0.78	0.76	0.73	0.74	0.80
KNN	0.87	0.85	0.85	0.84	0.86	0.87	0.86	0.86	0.85	0.83	0.85	0.85	0.85	0.85	0.83	0.84	0.86	0.85	0.84	0.83	0.84	0.85	0.85	0.84	0.84	0.83
Kstar	0.89	0.89	0.86	0.87	0.86	0.86	0.85	0.84	0.85	0.86	0.84	0.84	0.84	0.83	0.85	0.85	0.85	0.84	0.85	0.85	0.85	0.85	0.86	0.84	0.86	0.85
LogicBoost	0.74	0.73	0.74	0.76	0.76	0.77	0.77	0.78	0.79	0.77	0.77	0.77	0.78	0.77	0.80	0.77	0.79	0.78	0.78	0.79	0.80	0.80	0.82	0.77	0.80	0.77
Logistic	0.74	0.73	0.76	0.75	0.75	0.75	0.76	0.75	0.73	0.76	0.77	0.77	0.77	0.78	0.77	0.77	0.77	0.77	0.78	0.78	0.78	0.77	0.78	0.78	0.78	0.78
LWL-random	0.78	0.77	0.82	0.83	0.83	0.82	0.87	0.87	0.85	0.79	0.80	0.82	0.80	0.82	0.82	0.81	0.80	0.78	0.79	0.80	0.80	0.80	0.80	0.78	0.77	0.85
NBTree	0.82	0.80	0.79	0.81	0.79	0.78	0.78	0.77	0.76	0.78	0.77	0.77	0.77	0.78	0.77	0.78	0.78	0.78	0.79	0.80	0.80	0.80	0.79	0.79	0.80	0.79
PART	0.84	0.83	0.82	0.82	0.82	0.82	0.83	0.83	0.81	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.83	0.83	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.84
Random Committee	0.79	0.78	0.75	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.78	0.78	0.78	0.79	0.78	0.79	0.79	0.79	0.80	0.80	0.80	0.81	0.79	0.79	0.79	0.79
Random Forest	0.79	0.79	0.78	0.78	0.80	0.78	0.77	0.76	0.79	0.79	0.80	0.80	0.77	0.77	0.77	0.76	0.77	0.78	0.77	0.77	0.78	0.79	0.78	0.78	0.76	0.76
Random Sub Space	0.83	0.82	0.81	0.79	0.79	0.77	0.78	0.76	0.72	0.76	0.78	0.76	0.78	0.77	0.76	0.78	0.73	0.78	0.79	0.76	0.76	0.76	0.78	0.76	0.75	0.79
REPTree	0.83	0.80	0.82	0.82	0.82	0.82	0.80	0.82	0.81	0.80	0.82	0.81	0.81	0.80	0.80	0.81	0.76	0.80	0.81	0.81	0.82	0.84	0.80	0.81	0.82	0.80
Rotationforest	0.72	0.75	0.80	0.80	0.80	0.82	0.74	0.75	0.80	0.78	0.78	0.79	0.74	0.75	0.79	0.73	0.72	0.73	0.76	0.75	0.79	0.80	0.79	0.79	0.78	0.78
SPegasos	0.93	0.93	0.93	0.92	0.89	0.89	0.87	0.86	0.85	0.84	0.84	0.83	0.84	0.82	0.83	0.83	0.82	0.83	0.83	0.83	0.83	0.82	0.80	0.81	0.81	0.81
SVM	0.90	0.91	0.97	0.94	0.94	0.93	0.83	0.83	0.82	0.83	0.83	0.83	0.85	0.84	0.84	0.84	0.84	0.83	0.87	0.87	0.85	0.82	0.80	0.80	0.77	0.77

T1+T2-Concordance

Descriptor Number	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
AdaBoostM1	0.63	0.66	0.64	0.66	0.66	0.66	0.68	0.67	0.68	0.67	0.67	0.69	0.69	0.69	0.70	0.71	0.71	0.69	0.69	0.68	0.69	0.70	0.68	0.68	0.69	0.69
ADTree	0.70	0.70	0.71	0.72	0.72	0.72	0.73	0.72	0.73	0.73	0.73	0.72	0.73	0.73	0.73	0.74	0.74	0.73	0.75	0.75	0.75	0.75	0.76	0.75	0.75	0.76
ANN	0.64	0.64	0.64	0.63	0.64	0.64	0.65	0.67	0.66	0.66	0.66	0.66	0.68	0.66	0.67	0.67	0.67	0.67	0.67	0.68	0.68	0.68	0.69	0.68	0.69	0.69
BAGGING	0.68	0.68	0.71	0.72	0.72	0.73	0.74	0.75	0.74	0.74	0.74	0.74	0.75	0.76	0.75	0.75	0.75	0.75	0.77	0.76	0.77	0.76	0.77	0.75	0.69	0.76
BayesNet	0.69	0.69	0.70	0.71	0.71	0.70	0.70	0.70	0.71	0.72	0.73	0.73	0.72	0.74	0.72	0.73	0.73	0.74	0.73	0.74	0.74	0.74	0.75	0.74	0.73	0.75
Classification ViaRegression	0.67	0.67	0.71	0.71	0.73	0.72	0.72	0.74	0.73	0.75	0.75	0.73	0.74	0.74	0.75	0.76	0.74	0.76	0.77	0.75	0.76	0.75	0.75	0.75	0.76	0.76
DAGGING	0.63	0.64	0.64	0.66	0.65	0.64	0.67	0.67	0.66	0.66	0.66	0.65	0.66	0.66	0.66	0.67	0.68	0.68	0.68	0.68	0.68	0.69	0.69	0.70	0.70	0.70
Decision Table	0.67	0.68	0.69	0.70	0.71	0.70	0.69	0.69	0.71	0.71	0.71	0.72	0.69	0.70	0.69	0.70	0.70	0.70	0.69	0.72	0.70	0.71	0.69	0.69	0.71	0.72
Decorate	0.68	0.68	0.69	0.67	0.67	0.67	0.68	0.68	0.66	0.69	0.69	0.69	0.68	0.70	0.71	0.72	0.72	0.72	0.71	0.73	0.70	0.70	0.71	0.71	0.72	0.70
DNTB	0.61	0.61	0.61	0.62	0.62	0.63	0.63	0.63	0.63	0.61	0.61	0.62	0.64	0.65	0.65	0.65	0.66	0.66	0.68	0.67	0.66	0.65	0.66	0.66	0.66	0.66
END	0.63	0.63	0.63	0.63	0.63	0.63	0.64	0.64	0.65	0.65	0.65	0.65	0.66	0.65	0.65	0.65	0.66	0.66	0.66	0.67	0.66	0.66	0.67	0.68	0.67	0.67
IB1	0.67	0.68	0.67	0.67	0.67	0.66	0.69	0.68	0.68	0.67	0.67	0.67	0.70	0.68	0.69	0.68	0.70	0.69	0.69	0.70	0.70	0.70	0.69	0.70	0.68	0.69
KNN	0.70	0.70	0.71	0.70	0.72	0.74	0.72	0.74	0.73	0.73	0.73	0.74	0.75	0.75	0.74	0.74	0.76	0.75	0.75	0.75	0.75	0.75	0.77	0.76	0.76	0.75
Kstar	0.68	0.66	0.70	0.71	0.70	0.71	0.71	0.71	0.71	0.73	0.73	0.72	0.71	0.72	0.73	0.72	0.72	0.73	0.74	0.73	0.73	0.73	0.73	0.74	0.74	0.73
LogicBoost	0.68	0.68	0.70	0.71	0.72	0.72	0.73	0.73	0.74	0.73	0.73	0.74	0.75	0.74	0.75	0.75	0.74	0.74	0.74	0.74	0.76	0.76	0.77	0.75	0.76	0.75
Logistic	0.69	0.68	0.71	0.71	0.71	0.71	0.72	0.72	0.70	0.73	0.73	0.73	0.73	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.74	0.73	0.73	0.73
LWL-random	0.63	0.65	0.68	0.69	0.70	0.69	0.66	0.67	0.67	0.67	0.68	0.69	0.71	0.71	0.70	0.71	0.70	0.69	0.71	0.71	0.70	0.71	0.71	0.70	0.71	0.69
NBTree	0.72	0.71	0.71	0.74	0.73	0.72	0.73	0.72	0.72	0.73	0.73	0.73	0.74	0.74	0.74	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75	0.75	0.76	0.75
PART	0.70	0.72	0.71	0.72	0.72	0.72	0.74	0.73	0.73	0.74	0.75	0.74	0.75	0.75	0.75	0.76	0.76	0.75	0.76	0.76	0.76	0.76	0.77	0.76	0.76	0.76
Random Committee	0.70	0.71	0.70	0.71	0.71	0.71	0.71	0.72	0.72	0.73	0.72	0.72	0.73	0.73	0.72	0.73	0.73	0.73	0.74	0.74	0.74	0.75	0.74	0.74	0.74	0.74
Random Forest	0.67	0.67	0.67	0.67	0.68	0.67	0.68	0.67	0.68	0.69	0.69	0.70	0.68	0.69	0.69	0.69	0.69	0.69	0.68	0.68	0.68	0.69	0.69	0.69	0.68	0.68
Random Sub Space	0.65	0.66	0.68	0.68	0.68	0.69	0.69	0.69	0.66	0.71	0.68	0.68	0.71	0.69	0.69	0.71	0.68	0.72	0.70	0.70	0.70	0.70	0.71	0.72	0.69	0.72
REPTree	0.68	0.69	0.71	0.72	0.71	0.73	0.73	0.73	0.73	0.72	0.73	0.75	0.74	0.73	0.74	0.73	0.71	0.73	0.74	0.73	0.75	0.75	0.74	0.73	0.75	0.74
Rotationforest	0.65	0.66	0.67	0.68	0.66	0.66	0.65	0.65	0.67	0.68	0.68	0.67	0.66	0.65	0.66	0.65	0.65	0.66	0.67	0.68	0.67	0.68	0.67	0.67	0.67	0.68
SPegasos	0.64	0.64	0.64	0.64	0.64	0.64	0.66	0.66	0.66	0.66	0.66	0.66	0.68	0.67	0.67	0.67	0.66	0.67	0.67	0.67	0.69	0.69	0.68	0.68	0.68	0.68
SVM	0.62	0.62	0.62	0.61	0.61	0.62	0.65	0.65	0.65	0.65	0.65	0.65	0.64	0.64	0.64	0.64	0.64	0.65	0.67	0.67	0.66	0.65	0.65	0.65	0.63	0.63