Development and Extension of Cheminformatics Techniques for Integration of Diverse Data to Enhance Drug Discovery

Christopher M. Grulke

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Pharmacy (Division of Medicinal Chemistry and Natural Products).

Chapel Hill
2011

Approved by:

Dr. Alexander Tropsha
Dr. J. S. Marron
Dr. Diane Pozefsky
Dr. Bryan Roth
Dr. Christopher Waller
Dr. Qisheng Zhang

ABSTRACT

Christopher M. Grulke

Development and Extension of Cheminformatics Techniques for Integration of Diverse Data to Enhance Drug Discovery

The scientific community has fallen headlong into the age of data. With the available crop of information available to scientists growing at an exponential pace, tools to harvest this data and process it into knowledge are needed. This blanket statement is nowhere more true than in drug discovery today.

The increasing quantities of bioactivity and protein crystallographic data provide key information capable of improving the state of virtual screening. The CoLiBRI methodology attempts to learn from the large knowledge base of protein-ligand interactions to discover a comprehensive model capable of filtering large libraries very quickly using only a protein structure. This modeling procedure has been greatly expanded to encompass a wide range of descriptor techniques and to use advanced statistical methods of multidimensional mapping.

The growth of virtual screening methods (including CoLiBRI) has provided a plethora of options to cheminformaticians with little guidance on their strengths and weaknesses. This oversight in methodology benchmarking should be addressed to reduce the time and effort wasted applying subpar screening protocols. To attend to this issue, we developed a benchmark dataset that will enable a flood of methodology experimentation and validation.

The recent generation of gene expression data and cancer cell growth inhibition data enable identification of signatures of cellular resistance. These signatures can be used as validated prognostic markers to guide patient management thereby fueling the personalization of cancer treatment. From the available data, we have derived hypothetical biomarkers of multidrug resistance and a flood of links between gene expression and chemical specific resistance that require experimental validation.

The increasing capabilities of cheminformatics techniques require dissemination to the public to produce the greatest impact. We have therefore developed a web portal providing cheminformatics software and models to fuel public drug discovery efforts.

Table of Contents

List of Tables

List of Figures

Abbreviations

MLI, Molecular Libraries Initiative

HGP, Human Genome Project

PDSP, Psychoactive Drug Screening Program

PDB, Protein DataBank

SCOP, Structural Classification of Proteins

NCBI, National Center of Biotechnology Information

KEGG, Kyoto Encyclopedia of Genes and Genomes

QSAR, Quantitative Structure Activity Relationship

HIV, Human Immunodeficiency Virus

SNP, Single Nucleotide Polymorphism

CoLiBRI, Complementary Ligand Bind Receptor Interactions

CADD, Computer Aided Drug Discovery

GPCR, G-Protein Coupled Receptor

SCREEN, Surface Cavity REcognition and EvaluatioN

TAE, Transferrable Atom Equivalents

GSK, GlaxoSmithKline

RDF, Radial Distribution Function

kNN, k Nearest Neighbor

SA, Simulated Annealing

PMR, Predicted Mean Rank

LOO, Leave One Out

CV, Cross Validation

CCA, Canonical Correlation Analysis

UNC, University of North Carolina

WDI, World Drug Index

PCA, Principle Component Analysis

PMRR, Predicted Mean Rank of Receptors

FKBP, FK506 binding protein

MDDR, MDL Drug Data Report

FCFP, Functional Connectivity Fingerprint

ACK1, Activated Cdc42-associated Kinase

ACHE, Acetylcholinesterase

AR, Androgen Receptor

B2AR, Beta-2 Adrenergic Receptor

CA2, Carbonic Anhydrase II

CDK2, Cyclin Dependent Kinase 2

COX2, Cyclooxygenase 2

DHFR, Dihydrofolate Reductase

ESR1, Estrogen Receptor Alpha

ESR2, Estrogen Receptor Beta

F10, Coagulation Factor X

GR, Glucocorticoid Receptor

HIV-Int, HIV Integrase

HIV-Pr, HIV Protease

HIV-RT, HIV Reverse Transcriptase

PARP1, Poly [ADP-ribose] Polymerase-1

PDE5, Phosphodiesterase 5A

PNP, Purine Nucleoside Phosphorylase

PPARG, Peroxisome Proliferator-Activated Receptor Gamma

REN, Renin

SRC, Tyrosine Protein Kinase SRC

F2, Thrombin

ROC, Receiver Operating Characteristic

CCR, Correct Classification Rate

NCI, National Cancer Institute

IVCLSP, In Vitro Cell Line Screening Project

SVD, Singular Value Decomposition

SAM, Significance Analysis of Microarrays

IPA, Ingenuity Pathway Analysis

FDR, False Discovery Rate

MYC, c-Myc

DFFB, DNA Fragmentation Factor Beta

APC, Adenomatous Polyposis Coli

JSP, JavaServer Page

SVM, Support Vector Machines

GA, Genetic Algorithm

ADME, Absorption, Distribution, Excretion, Metabolism

Chapter 1: Background and Significance

1.1. Data and Drug Discovery

Publicly available data in all forms of science is increasing at an exponential pace.[1] This increase in data is nowhere more evident than in the field of drug discovery. High-throughput technologies (e.g., parallel synthesis and high throughput screening) have become commonplace and bold publicly funded projects (e.g., the Molecular Libraries Initiative (MLI)[2] and the Human Genome Project(HGP)) harness these technologies to create large amounts of publicly accessible data. As a result, these large publicly available databases are ready to accelerate chemical biology and drug discovery.

As an example, the availability of data linked to chemical compounds in the public domain has exploded. Several databases have sprung up offering structural, biochemical, and phenotypic data in a chemocentric way. The *PubChem* database[3] (http://pubchem.ncbi.nlm.nih.gov/), developed as the central repository for chemical structure-activity data, is just a single instance of such databases. In the short time since its introduction in 2005, PubChem has grown to contain nearly 31 million chemical compound records; over 1.5 million of these chemicals have been "tested" in an assay with more than 300 thousand appearing active at least once. Many similarly structured databases have emerged recently as well (e.g., ChemSpider,[4] ChEMBL,[5] PDSP Ki,[6] and others (cf. this recent review[7])).

Additional resources are available with protein centric data (PDB, SCOP, UniProt), gene centric data (NCBI Genome, UniGene), and pathway/protein interaction centric data (KEGG, BioCyc, GeneNet). With the vast increase in the data related to the function of our bodies, one would expect that the rate of drug discovery would also have increased.

Unfortunately, while available data to fuel drug discovery has drastically expanded, the number of new drugs introduced into the market has, in fact, remained stagnant.[8, 9] With the increased output of "me-too" drugs[10], one could argue that the rate of drug discovery has decreased even as our knowledge has increased. Also, the attrition of drug candidates entering clinical trials remains high.[11, 12] The cost of drug discovery and development is continuing to grow,[13] and the time to develop a drug remains roughly the same as it was 30 years ago.[12]

The slow rate of drug discovery in the midst of an explosion of biomedical data is a conundrum that can be addressed by developing methods and studies that utilize the expanse of data to inform decisions related to the discovery of drugs. In the last few years, the use of *in silico* methods to leverage data for drug discovery has become much more common.[14] However, because the amount and breadth of the available data is constantly increasing, there is an abundance of unaddressed areas that require attention.

To leverage the enormous amount and types of available data to address effectively the variety of questions in the field of drug discovery, specialized techniques are needed. For many years, our group has been engaged in the development and application of innovative methodologies and approaches in the field of QSAR modeling. This focus on conversion of statistical techniques to enable cheminformatics research, has given us a unique ability to select,

modify, and apply methods for analysis of data involving chemical structure. The studies contained in this dissertation address only a few situations where developments have been made.

1.2. Virtual Screening

The contributions of virtual screening to drug discovery are many:[15] cheminformatics techniques have aided in the discovery of such drugs as Dorxolamide for glaucoma[16], Zanamavir for influenza[17], and Raltegravir for HIV infection.[18] The advancement of virtual screening could provide a steady stream of new hits to the drug discovery process, but such advancement requires both novel techniques and detailed comparisons of available tools.

'Virtual screening' has typically implied the use of protein structure to identify subsets of molecules in large chemical databases or virtual chemical libraries that are likely to bind to the target protein with appreciable affinity and specificity. Structure-based virtual screening has become a fundamental part of modern computer-aided drug design[19, 20]. It requires the posing and scoring of libraries of small molecules to find compounds that fit into the binding site and bind tightly to the receptor. Since the seminal publication by the Kuntz group in 1982[21], this approach has been used successfully in numerous studies (such as that of HIV protease inhibitors) resulting in the design of approved drugs[22]. Numerous algorithms and programs have been introduced. (For reviews, see Wong and McCammon[23], Taylor et al.[24], and Muegge[25].) Examples of widely used docking programs include Autodock[26], FlexE[27], and Gold[28].

While the implication has been that virtual screening and structure-based virtual screening are synonymous, there has been an increase in the use of ligand-based techniques to identify hits from large chemical databases[29-32]. Numerous algorithms have been introduced (cf. the recent reviews[33-35]). Most recently, reviews[36, 37] in the area of virtual screening have begun noting

studies comparing structure-based techniques to ligand-based techniques. Use of cheminformatics techniques to cull large chemical databases to a size more conducive to application of slower docking methods is discussed frequently. Novel scoring functions generated using methods typically applied in QSAR modeling[38] or directly integrating ligand similarity[39] have been reported. This view of virtual screening as a field in which both ligand and structure-based techniques are applied to yield optimal results is good for drug discovery as a whole.

Two chapters of this dissertation are focused on advancing the field of virtual screening. Chapter 2 details the efforts of the author to advance a novel method of structure-based virtual screening that uses techniques commonly applied in ligand-based virtual screening. Chapter 3 documents the creation of benchmark dataset intended to thoroughly assess various virtual screening methods and preliminary studies to verify its usefulness.

1.3. Chemotherapeutic Resistance

The resistance of cancer cells to chemotherapeutic treatment has been of interest for more than half a century.[40] With chemotherapy being the preferred method of tumor treatment, the understanding of drug resistance is vital to provide quality care to cancer patients. With the dawn of the genomic age, the investigation of chemotherapy resistance turned to analysis of genomic data to determine underlying factors and markers indicative of a tumor's resistance to chemical treatment. Even with the many new discoveries, our ability to accurately predict the response of a patient to a chemotherapeutic is limited.[41]

The goal of personalizing treatments for patients to yield better clinical outcomes has been marked by both successes and disappointments.[42, 43] It is hypothesized that a portion of the

difficulty in predicting patient outcomes is due to the large variety of mechanisms for drug resistance, while the remainder can be attributed to the lack of a single measurement type capable of capturing all types of resistance. The use of gene expression signature to predict outcome only captures a portion of the potential causes of therapy failure.

While the study described in Chapter 4 focused purely on the use of gene expression profiles to develop a series of markers of chemotherapeutic resistance, we recognize that to fully address the problems in personalization of medicine we need to include other data types such as SNP variations. However, we believe that gene-expression profiles provide a great deal of insight in cellular resistance to drug agents and that treatment of this data (and others) to identify generic or multidrug biomarkers followed by analysis of single compound outcome biomarkers is most rational.

1.4. Dissemination of Tools and Results

A large portion of the results obtained by analysis of data is published, but not easily searchable, accessible, or usable. The recently increased public availability of experimental data highlights the lack of a public repository to store tools to examine such data and the hypotheses generated by such examinations. This deficiency is most evident in the field of cheminformatics.

The field of bioinformatics may be considered the most closely related discipline to cheminformatics. However, when we compare the two fields, the lack of publicly available cheminformatics tools is underscored. In bioinformatics, tools are widely available to accomplish gene and protein sequence alignments[44, 45] and classifications.[46-48] Web interfaces are provided for several protein pocket identification schemes.[49-52] The analysis of gene expression can be complete using software available through the web.[53] This availability of tools

in the field of bioinformatics aids in the advancement of their field. Access to some of this bioinformatics software was vital to complete portions of the studies discussed in this dissertation.

Because the availability of tools enables research, the development of a web portal for cheminformatics investigation of data and dispersement of cheminformatics techniques was undertaken. Chapter 5 provides information regarding the completion and impact of this portal, which we call Chembench.

Chapter 2: Complementary Ligand Binding Receptor Interactions (CoLiBRI)

2.1. Introduction

Computer Aided Drug Design (CADD) can be defined as any method that uses computational power to analyze input information in order to enhance the drug discovery process. Traditionally, these methods have been subcategorized into two classes based on the input that they require. Structure based methods rely on the three-dimensional structure of the macromolecular target for which drugs are being designed while ligand-based methods analyze the chemical structures of compounds with known activity. As such, each class of methodologies has its own domain of applicability and its own limitations.

Structure based methods are often used to screen chemical databases for potential compound leads based on steric and electronic complementarity to a macromolecular target's binding pocket. Several successes have been reported using a variety of popular software; however, accurate scoring and ranking of chemicals using structure-based methods is still difficult[54] and being thoroughly researched[38, 55, 56]. Additionally, since the docking technique relies on accurate 3-D macromolecular structure, it is difficult to apply to several potential targets for which structures are rarely available (notably G-Protein Coupled Receptors (GPCRs) and ion channels). Finally, because of the complexity of conformational sampling and posing of chemicals, even the fastest methods take several seconds to screen a single compound.

Ligand-based methods, likely due to a longer history, cover a broader range of techniques; however, the most common approach is to represent compounds which have known target activity using chemical descriptors and subsequently apply statistical tools to discover correlations between the calculated descriptors and the target activity. This activity need not be related to interaction with a single known macromolecular target. This type of approach has been used successfully to screen chemical libraries to find new chemical leads[29, 57, 58]. While screening with these methods is typically very fast, a drawback to this traditional Quantitative Structure-Activity Relationship (QSAR) approach is that it may be less likely to find active chemicals of different structural class from the set used to discern the SAR (though this is a point of contention among computational scientists). Additionally, it requires a certain amount of bioactivity information that is often lacking in the early discovery process.

In a recent publication from our lab, a traditional ligand-based method (SA-kNN) trained using publicly available 3-D macromolecular data was shown to be fast and effective for screening a large number of protein targets[59]. This novel computational drug discovery strategy outlined in Figure 1 combines the strengths of both *structure-based* and *ligand-based* approaches while attempting to surpass their individual shortcomings. The training of CoLiBRI models starts from a dataset of protein-ligand complexes. From this dataset, the binding pocket of each protein is identified. While the task of pocket selection has been well studied[60], it is one that still lacks a complete solution. Both ligands and the identified pockets are then transformed into multidimensional descriptors. While description of ligands is often done in QSAR studies, there is little precedent for description of the chemical fragments that comprise a binding pocket. Based on best practices in analogous ensemble QSAR modeling workflows, modeling sets are separated into training and test sets. Based on the hypothesis that the relative location of a novel

binding site with respect to other binding sites in multidimensional chemistry space could be used to predict the location of the ligand(s) complementary to this site in the ligand chemistry space, models that map the two multidimensional spaces are developed using the training sets. These models are used to rank the test set ligands within a large chemical library of putative inactives. Models that appear to be predictive are then applied to the binding pocket of a protein of interest to generate a virtual ligand point that is used as a query in chemical similarity searches to identify putative ligands of the protein in available chemical databases.

In the published approach testing of the CoLiBRI workflow was completed using 800 diverse protein-ligand complexes comprising the PDBBind dataset[61]. The authors extracted the binding pocket from the protein using protein-ligand tessellation and then represented both the receptor

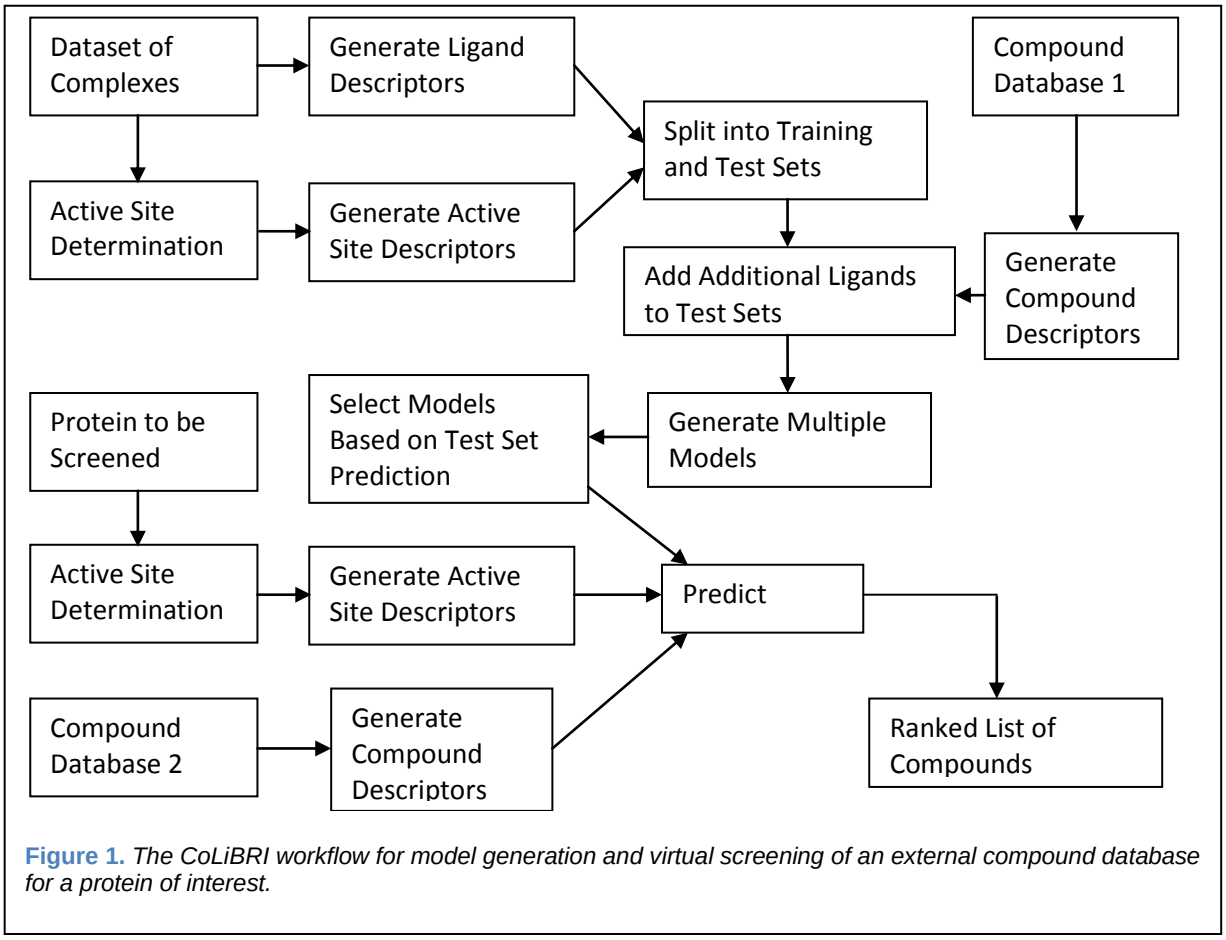


Figure 1. *The CoLiBRI workflow for model generation and virtual screening of an external compound database for a protein of interest.*

active site and its corresponding ligand in the same universal, multidimensional chemical descriptor space (note that in principle, the descriptors used for receptors and ligands do not have to be the same, and we explored this aspect in the current project). The authors reasoned that mapping of both binding pockets and corresponding ligands onto the same multidimensional chemistry space would preserve the complementary relationships between binding sites and their respective ligands. Thus, it is expected that ligands binding to similar active sites are also similar. Using a k nearest neighbor (kNN) pattern recognition approach and variable selection, it has been shown that knowledge of the binding pocket structure affords identification of its complimentary ligand among the top 1% of a large chemical database in over 90% of all test binding sites when a binding pocket of the same protein family was present in the training set. However, in a more realistic case where test receptors are highly dissimilar and not present among the receptor families in the training set, the prediction accuracy is decreased; still, CoLiBRI was able to quickly eliminate 75% of the chemical database as improbable ligands. The authors also showed that the method was highly computationally efficient allowing a user to process ca. 30K compounds per minute on a single Pentium 4 CPU [59].

Unfortunately, the seminar work on CoLiBRI was relatively limited. Herein, we document our attempts to improve upon this method by examining and enhancing its key components: active site determination, active site and ligand descriptor generation, and model generation.

2.2. Materials and Methods

2.2.1. Dataset Preparation

Coordinates for the protein-ligand complexes were obtained from multiple versions of the PDBBind Database[61]. The PDBBind database provides an organized repository of protein ligand

complexes extracted from the PDB and annotated with binding constants extracted from literature. From this compendium of protein-ligand complexes with affinities, a "refined" set of complexes meeting the following criteria is pulled. Complexes must have a resolution of greater than 2.5 angstrom; not contain covalent bonds between the protein and ligand; contain a ligand consisting only of C, N, O, S, P, H, and halogens with a molecular weight less than 1000; and have no unnatural amino acids in the binding pocket. The "refined" set is clustered using BLAST and a threshold of 90% similarity. For each cluster containing 4 or more complexes, 3 representatives are chosen—the one with the highest binding affinity, the one with the lowest binding affinity, and a one with the medium binding affinity—to form the "core" set.

In all cases, Sybyl[62] was used to preprocess the proteins including the removal of crystallographic water, elimination of salts and metals, and addition of hydrogen atoms. Ligands were "washed" using the Wash Molecules application in MOE[63]. This application normalizes chemical structures by carrying out a number of operations including 2D depiction layout, hydrogen correction, salt and solvent removal, chirality and bond type normalization, adjustment and enumeration of protonation states, and expansion of fragment abbreviations.

2.2.2. Active Site Determination

The identification of the binding pocket is a crucial part of the CoLiBRI workflow. In this study three methods of active site determination were investigated: protein-ligand tessellation, CastP, and SCREEN.

2.2.2.1. Protein-Ligand Tessellation

To appropriately calculate binding pocket descriptors, we are first required to identify individual atoms or amino acid fragments that are the pocket. The first method we applied to

complete this task utilized a computational geometry technique known as Delaunay tessellation to isolate the protein atoms that made contacts with bound ligands. Applied to a collection of randomly distributed points, Delaunay tessellation partitions the space occupied by these points into an aggregate of space filling, irregular triangles (in 2D) or tetrahedra (in 3D) with the original points as vertices. Thus, this approach effectively identifies all nearest neighbor triplets (or quadruplets) of vertices. An example of Delaunay tessellation in two dimensions is illustrated in Figure 2.

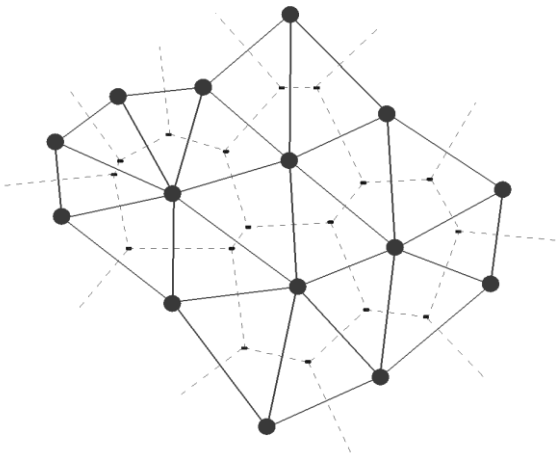


Figure 2. *Illustration of Voronoi/Delaunay tessellation in 2D space (Voronoi polyhedra are represented by dashed line, and Delaunay simplices by solid line). For the collection of points with 3D coordinates, such as atoms of the protein-ligand complex, Delaunay simplices are tetrahedra whose vertices correspond to the atoms.*

Protein-ligand complexes are represented by the coordinates of their heavy atoms (i.e., in a hydrogen-depleted form). Delaunay tessellation of this representation uniquely defines all sets of nearest neighbor atom quadruplets, including three types of interfacial quadruplets: three receptor atoms and one ligand atom; two receptor and two ligand atoms; and one receptor and three ligand atoms. Thus, Delaunay tessellation affords an easy way of detecting all receptor atoms that directly contact the ligand. These are then specified as the binding site.

2.2.2.2. CastP

The CastP method[64] of identifying pockets also relies on the use of tessellation; however, this tessellation does not involve the bound ligand. Instead, the protein is tessellated with all small molecules removed and cavities are detected using alpha shape theory. Figure 3 and Figure 4

taken from Liang et al illustrate how the alpha shape theory can be applied in 2D space to identify protein pockets. First the protein is tessellated (triangulated) and the Voronoi diagram is determined. All Voronoi edges fully external to the protein are omitted. Delaunay tetrahedra (triangles) that have edges crossing these fully external Voronoi edges are considered "empty". Empty tetrahedra (triangles) are merged together as long as they share a triangle (edge) into potential pockets. Provided one of the tetrahedra (triangles) contained in a potential pocket is acute, the pocket is designated a protein pocket. Pockets were identified in this manner for all proteins in our datasets using the CastP webserver[50, 65] at http://sts.bioengr.uic.edu/castp/. The binding

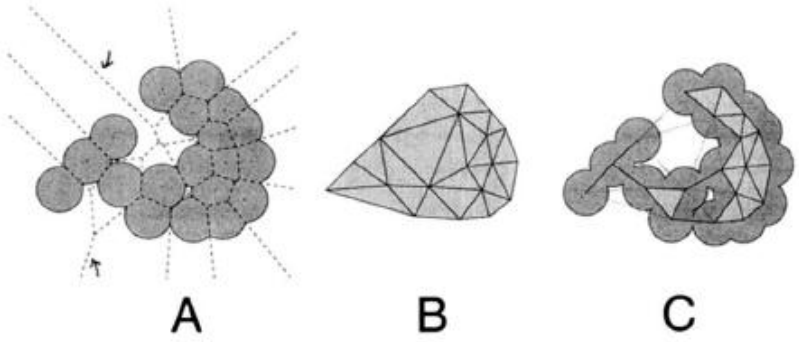


Figure 3. *Illustration of concepts in alpha theory (from Liang et al.[63]).* **A***: A two-dimensional molecule consisting of disks of uniform radii. The dashed lines show the Voronoi diagram of the molecule. Arrows indicate 2 of the 10 Voronoi edges that are completely outside the molecule.* **B***: The convex hull of the atom centers in Figure IA (all shaded area) with Delaunay triangulation (triangles defined by dark lines).* **C***: The alpha shape of the molecule in* A. *The alpha shape. or dual complex, consists of the light-shaded triangles. the dark line segments, and the atom centers. There are 10 shaded line segments corresponding to the 10 Voronoi edges that are completely outside the molecule. Any triangle with one or more shaded edges is an "empty triangle." A void formed by three empty triangles can be seen at the bottom center. It encloses a molecular cavity.*

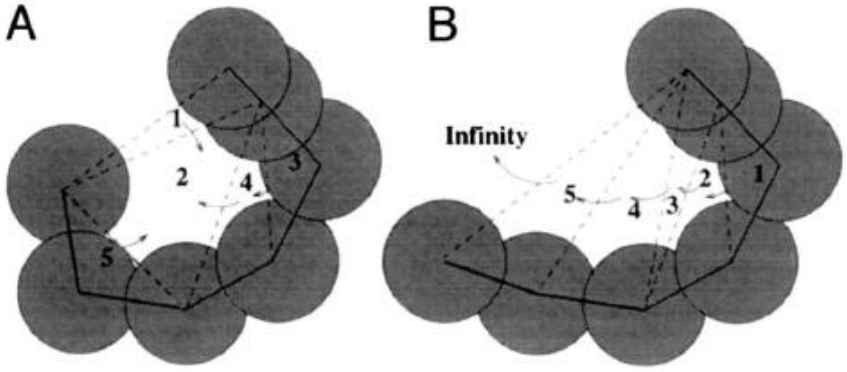


Figure 4. *Illustration of discrete flow for two-dimensional pockets (from Liang et al [63]).* **A***: Discrete flow of a pocket. Obtuse empty triangles (I , 3. 4. and 5) flow to the acute triangle (2). Collectively, they form a pocket of the dual complex, which can he mapped to the molecular pocket.* **B***: A depression for which obtuse triangles sequentially flow to the outside (to infinity). Depressions of this type are not identified as pockets.*

pocket was identified by visual inspection of the identified pockets overlaid with the bound ligand.

2.2.2.3. SCREEN

The Surface Cavity REcognition and EvaluatioN (SCREEN) method[49] of pocket detection identifies the gap between a protein's molecular surface and a surface generated by rolling a intermediately sized sphere over the molecular surface. Cavities were identified in this manner for all proteins in our datasets using the SCREEN2 webserver at http://luna.bioc.columbia.edu/honiglab/screen2/cgi-bin/screen2.cgi. The binding pocket was identified by visual inspection of the identified cavities overlaid with the bound ligand.

2.2.3. Active Site Descriptor Calculation

Descriptor generation for the set of chemical fragments is a significant difficulty in the CoLiBRI process. Most molecular descriptors cannot be generated for chemical fragments. As such, two newly developed methods of protein pocket description (feature point pairs and RDF) were added to the previously published TAE/RECON technique.

2.2.3.1. TAE/RECON

The generation of TAE/RECON descriptors relies on the concepts of Transferable Atom Equivalents (TAE) developed by Breneman and co-workers[66-68]. The major advantage of these descriptors over other descriptor types is that they are derived from the electronic and shape properties of isolated atoms or chemical groups. The additivity principle is used to calculate molecular descriptors by summing up the individual descriptor type values for all atoms in the molecule, using the RECON method. In the case of ligands, this leads to the generation of molecular descriptors, similar to other approaches. The same additivity principle can also be

used to derive pseudo-molecular descriptors for any group of atoms, e.g., binding site fragments, making the TAE descriptors exceptionally well suited for our approach.

2.2.3.2. Feature Point Pairs

While the application of atom pairs as a description of binding pockets is straightforward, the use of feature points overlaid on chemical structure rather than specific atom points provides an abstraction that could prove more biologically relevant. Through collaboration with computational scientists in GlaxoSmithKline (GSK), a set of feature point representations of amino acids were implemented and used to transform binding pockets to a feature space. This feature space first described by Yang[69] provides a simple representation of amino acids based on their physicochemical properties. Counts of feature pairs occurring within respective distance bins were used as a set of quantitative descriptors of the 3D characteristics of the binding pocket. The table of amino acid atom to feature transformations is contained in Appendix I.

2.2.3.3. Radial Distribution Function

Radial Distribution Function (RDF) descriptors were developed in 1999 by Hemmer, et al.[70] to better describe the three dimensional characteristics of small molecules. Because other implementations of RDF descriptor generation could not be used to describe the disconnected chemical fragments that comprise our binding pockets, we implemented our own version of these descriptors.

To start, a peptide containing each of the 20 standard amino acids bordered on its N- and C-termini by glycine was treated as a small molecule within the PETRA software from Molecular Networks[71] to generate a table of atomic properties (including partial charge, electronegativity, and polarizability) for each atom type contained within proteins. The methods of property

calculation within the PETRA program have been shown to be quite accurate[72-76]. This table of properties along with the coordinates of the chemical fragments comprising the pocket was then processed using Equation 1. In Equation 1, prop($atom_i$) is a predefined property of atom i; $Dist_{ij}$ is distance in 3D coordinate space between the atoms measured in Angstroms; and D and damp are bin and damping parameters. RDF descriptors were calculated for binding pockets using values of D ranging from 0 to 20 with a step size of 0.2.

(1)

2.2.4. Ligand Descriptor Calculation

Ligand descriptors were generated using a variety of commercially available techniques. Specifically, TAE/RECON, Dragon, MOE2D, MolconnZ, and autocorrelation descriptors were all generated in the course of our study of the CoLiBRI methodology.

2.2.5. Model Generation

2.2.5.1. Simulated Annealing k-Nearest Neighbors

While kNN is an excellent pattern recognition technique, it requires that the similarities to which it is applied be related to the property being modeled. Because some descriptors generated for a ligand/binding pocket may be irrelevant to the binding interaction, these descriptors generate a level of inaccuracy within the resulting compound rankings. Variable selection—in particular simulated annealing (SA)—is a technique that has been successfully applied with the kNN principle to generate more robust and predictive models for traditional QSAR datasets[58, 77, 78].

The first version of the CoLiBRI methodology is a SA-kNN model generation tool which optimizes the predicted mean rank (PMR) of the true ligand of a binding pocket using Leave-One-Out(LOO) cross validation (outlined in Figure 5) where PMR is calculated by averaging the ranks at which a pocket's true bind ligand is retrieved across all pockets. This generates a model that in theory should be more accurate in virtual screening than the kNN principle applied to distances calculated in the whole descriptor space. Additional details of this method are described in the original CoLiBRI publication[59].

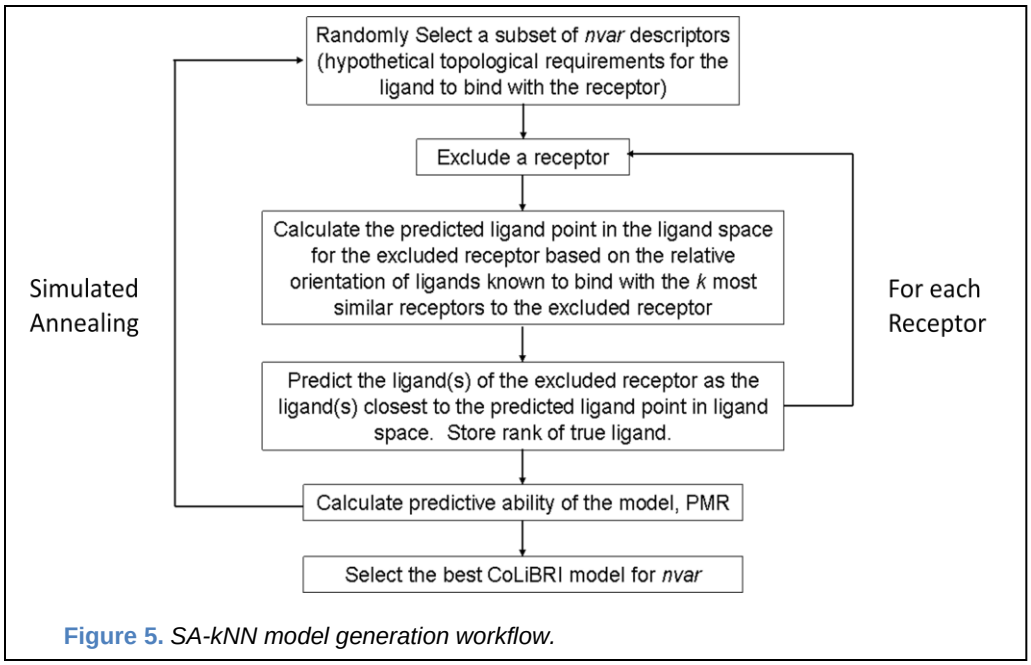


Figure 5. *SA-kNN model generation workflow.*

2.2.5.2. CCA and kCCA

The SA-kNN method attempts to select a descriptor subspace where similar proteins bind similar chemicals; however, when dealing with two multi-dimensional spaces the optimization becomes more complex. Fortunately, Canonical Correlation Analysis (CCA) originally developed by Hotelling[79] is specifically formulated to correlate multidimensional spaces. Therefore, its application in this situation is ideal.

Considering two multidimensional spaces X and Y, if we limit ourselves to bilinear mapping of the multidimensional spaces, the optimization problem can be written as Equation 2 where w_x

and w_y are the corresponding mapping matrices. The problem defined in Equation 2 is that of the well-known canonical correlation analysis that can be rearranged into a generalized eigen problem and subsequently solved. This provides a mapping of the two multidimensional spaces such that corresponding proteins and chemicals should be located near each other in their projected spaces.

(2)

Additionally, CCA can be extended using kernel methods. Although there are a multitude of potential kernels which could be applied, in this study we applied a newly developed spectral kernel[80]. Because the datasets are diverse and the similarity principle is only applicable in a local sense, the spectral kernel defined in Equation 3 and 4 provides a logical extension to the CCA method for this application.

(3)

(4)

where: x_i = descriptor vector for observation i

$N(x_i)$ = the k nearest neighbors of observation i

n = number of observations

Once CCA is completed, the similarity in the projected spaces allows prediction of the point in chemical space using a more advanced method than weighted averaging of the ligand descriptors for the neighboring proteins. Ridge regression[81] is used to build two models. Both models are generated using the binding site descriptors of the active site being predicted as the independent variables; however, one uses ligand descriptors of the k nearest neighbors as independent variables and the other uses the binding site descriptors of the k nearest neighbors as independent variables. The weights generated by this modeling are averaged and then applied to the ligand descriptors of the k nearest neighbors to predict the ligand point in the projected chemical space. This ligand point is then used to rank the chemical library.

2.3. Results and Discussion

2.3.1. External Validation of the CoLiBRI Workflow

The previously published work by our lab indicated that the CoLiBRI methodology may have potential as a fast and accurate structure–based virtual screening methodology; however, the experiment in the published work focus on extraction from a large chemical database of a single co-crystallized binding partner. While similar, this

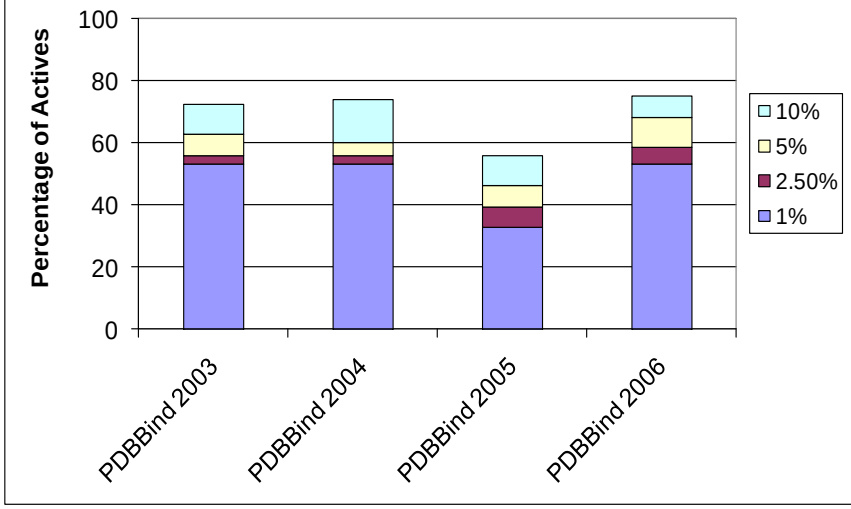


Figure 6. *Recall of HIV protease ligands dissolved in World Drug Index using consensus prediction by CoLiBRI models.*

is not the same experiment as attempting to find all known chemicals that bind to a protein. Additionally, the accuracies reported were for the test set, not a fully external validation set.

Since, we have expanded the training set using more recent versions of the PDBBind dataset. We have also applied the models generated from these sets to fully external test cases in order to more accurately determine the validity of this modeling technique. Figure 6 contains the results of a sample virtual screen for HIV protease inhibitors dissolved in the World Drug Index (WDI) using SA-kNN CoLiBRI models. Models trained using three versions of the PDBBind database with all HIV protease complexes removed recalled more than half the active ligands in 1% of the database. These results are comparable to those found in literature for virtual screening by conventional 3-D docking methods[82]. However, while docking typically takes over second for each screened ligand, the entire library of over 50,000 compounds were screened in less than 100 seconds The success of this pilot study indicates that the CoLiBRI method can filter large chemical databases to recall cognate ligands much more quickly than traditional methods.

2.3.2. Preliminary CCA Testing

The limitations of using only a single method of multidimensional optimization to build CoLiBRI models led to the desire to integrate additional methods. In particular, research into CCA and kCCA being conducted in the Department of Statistics at UNC provided access to a deterministic method of multidimensional optimization that is significantly faster than the stochastic SA-kNN method applied previously. To ascertain the capabilities of this technique, CCA was applied to the 800 protein-ligand complexes contained in PDBBind. Figure 7 demonstrates that when the multidimensional binding pocket and ligand points are projected onto their respective CCA vectors generated from combined analysis of the respective TAE

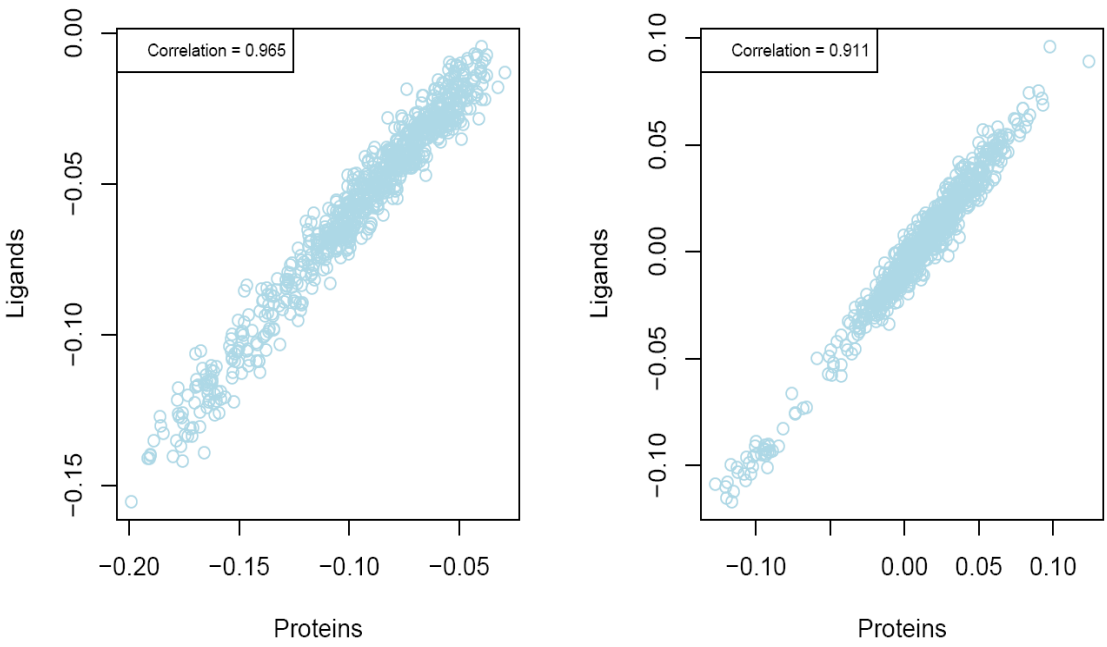


Figure 7. *Correlation of the (a) first and (b) second canonical variates from ligands and proteins*

multidimensional spaces, they correlate very well. Figure 8 is a scatter plot of projection on these same variates of the 800 binding sites (in red) and ligands (in blue) represented using their internal ids. The subfigure contains a magnified view of a portion of the project space. Visual inspection indicates that although the overlay is not perfect, it can be noted that ligands from a complex are near to their corresponding binding site and the neighborhood distributions are quite similar. For example, when inspecting the region surrounding pocket 666, it is clear that ligand 666 is located near it in space. Additionally, the pocket neighbors of pocket 666 (pockets 96, 135, 657, 658, 659, and 686) match the ligand neighbors of ligand 666.

2.3.3. Integration of CCA in the CoLiBRI Workflow

Because CCA provides a telling visual correlation between these two spaces, we believed that applying this method during the model development process could greatly improve our prediction accuracy. Therefore, we initiated a direct study comparing the SA-kNN, linear CCA, and kCCA. This study relied on the 1300 complexes of the refined set of the 2007 version of

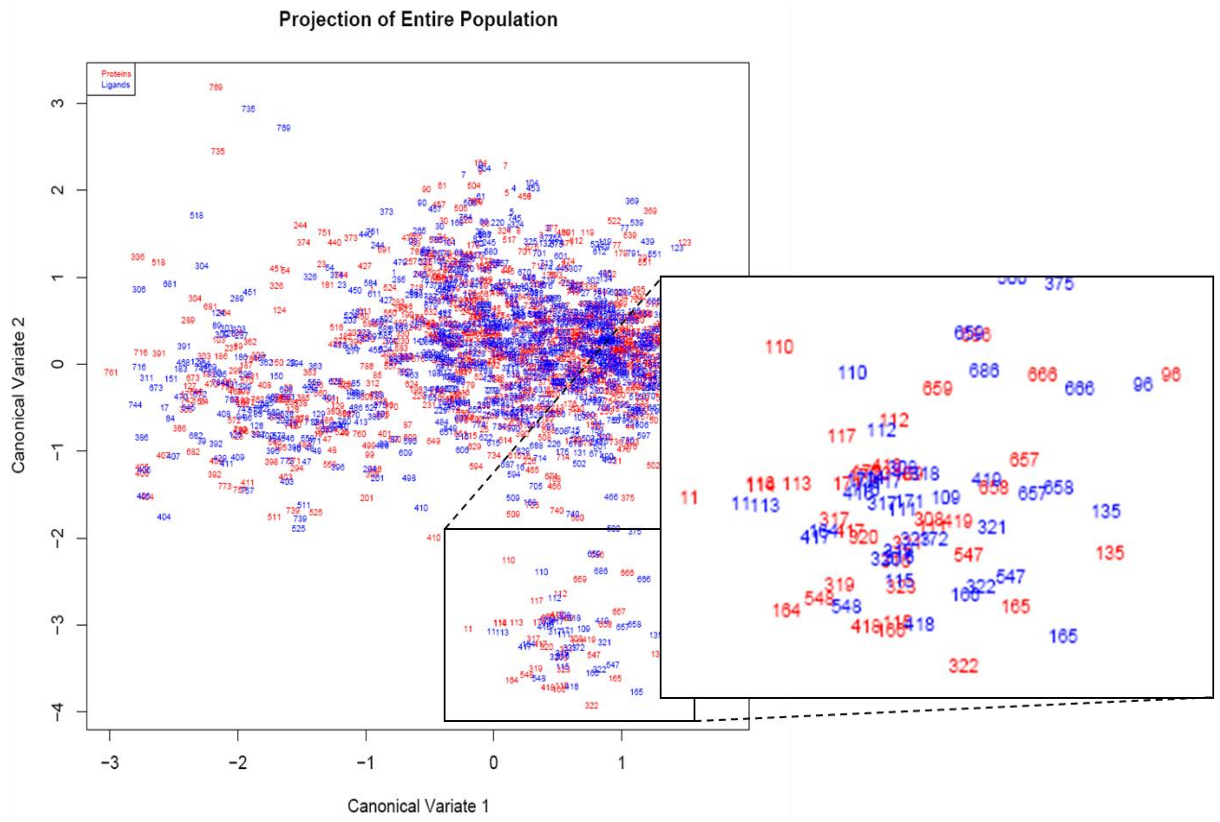


Figure 8. *Projection of binding sites and their cognate ligands from PDBBind V2003 onto the first two canonical vectors*

PDBBind. In addition, 210 of these complexes belonged to the core set, which is selected based on clustering of the 1300 complexes using protein sequence similarity and retaining only 3 complexes from each cluster.

From this data, we selected three different separations of the data by applying different methods for extracting the external validation set. For the first separation, an external validation set of 135 complexes was selected randomly from the 1300 complex refined set. For the second separation, an external validation set of 4 proteins (132 complexes) was extracted from the 1300 complex refined set based on the protein names stored in PDBBind for these complexes. For the third separation, an external validation set of 7 clusters (21 complexes) was taken randomly from the 70 cluster (210 complexes) core set. The remaining complexes which were not to be

used for external validation were then split using the Sphere Exclusion Algorithm[83] yielding training and test sets of size 966 and 169 complexes for the first separation, 1006 and 162 complexes for the second separation, 153 and 36 complexes for the third separation.

The first separation was referred to as the "standard set" since it closely mimics the normal method of 3-way data splitting applied by our lab to generate traditional QSAR models[84]. The second separation is referred to as the "name set" intended to have completely virgin proteins in the external set, and therefore, provide a more robust test of the modeling methods. The third separation is referred to as the "cluster set" and while containing the least amount of data, guarantees that the external set proteins are not exceedingly similar to proteins used during model development.

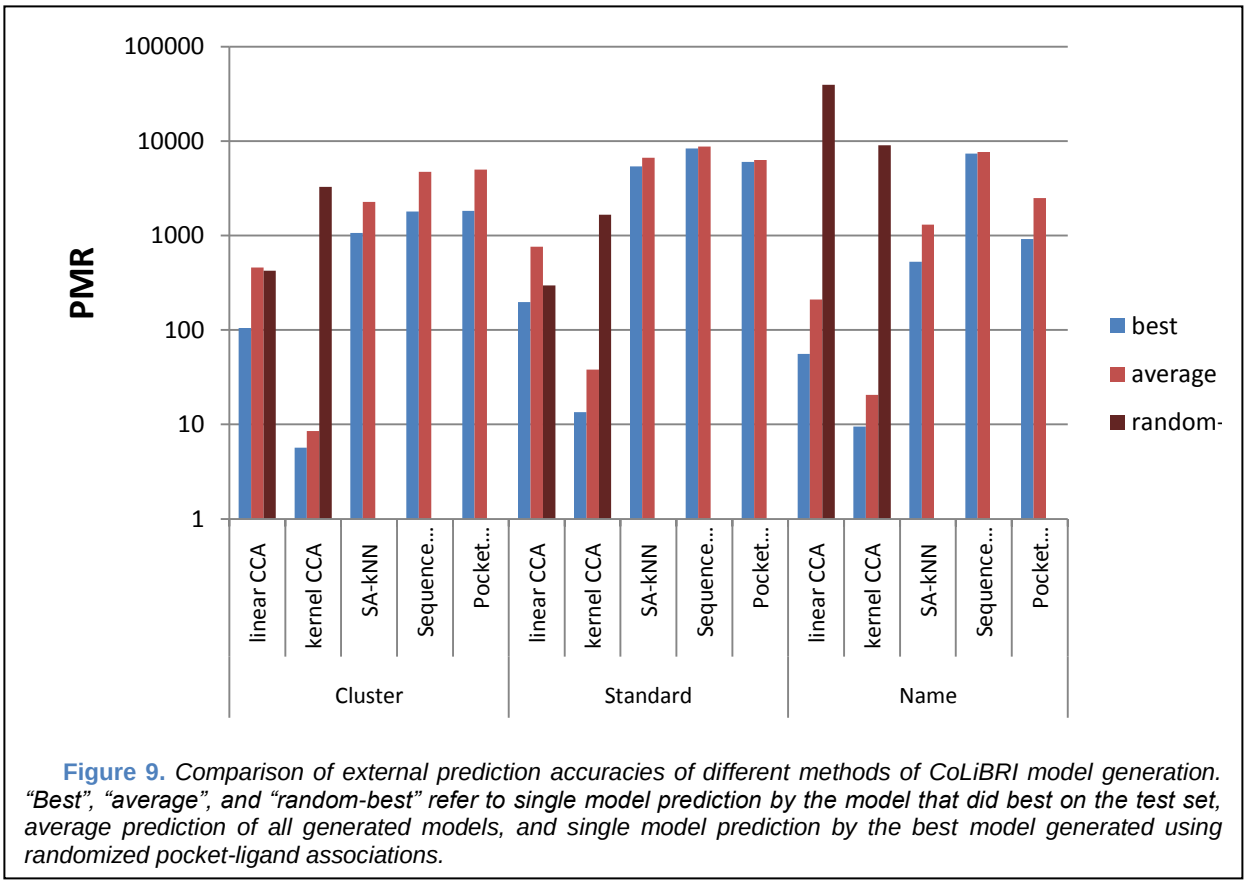


Figure 9. *Comparison of external prediction accuracies of different methods of CoLiBRI model generation. "Best", "average", and "random-best" refer to single model prediction by the model that did best on the test set, average prediction of all generated models, and single model prediction by the best model generated using randomized pocket-ligand associations.*

Both training and test sets were used to optimize the models for prediction of the external set though only the training set is used as a knowledge-base for the kNN predictions. In order to more closely replicate the act of virtual screening, the World Drug Index (WDI) compound database[85] was added to the test and external sets. To verify accuracies were not based on chance correlations, an additional set of models were generated where the training set ligand-pocket associations were randomly shuffled.

To assess the necessity of optimization procedures, both sequence similarity based (as determined using ClustalW[45]) and pocket similarity based kNN methods were also applied to all data splits with varied k values. TAE/RECON descriptors were used for both binding pockets (identified using protein-ligand tessellation) and ligands.

The PMRs for prediction of the external sets are shown in Figure 9. While SA-kNN provided only a very minor improvement over non-variable selected kNN techniques, CCA and kCCA performance was clearly superior. However, the results of linear CCA appear to be for both the cluster and standard sets indistinguishable from the results with randomized pocket-ligand associations. On the other hand, kCCA provided the best predictions for every external set and for what could be considered the most difficult case (the cluster set) predicted the true ligand on average in the top 10 compounds of the nearly 54000 contained in the screening database. This results indicates that CoLiBRI is capable of re-identifying the "true ligand" for a pocket in less than 0.1% of the database.

2.3.4. RDF descriptors

Through collaboration with Molecular Networks[71], software capable of developing RDF descriptors for binding pockets was developed. The effect of this method of descriptor

calculation was compared to the TAE/RECON method of binding pocket description on modeling of a regeneration of the cluster set described above from the 2008 version of PDBBind. Figure 10 displays the distribution of retrieved ranks for true ligands when screened with CoLiBRI for both the test and external sets. Based on the test set results, it appears as though neither TAE nor RDF descriptors of active sites provide superior predictive power. However, when applying the best model for each descriptor type, RDF descriptors and autocorrelation descriptors for pockets and ligands respectively showed clearly improved prediction over using TAE descriptors. This indicates that similar to traditional QSAR modeling, a "combi" approach may lead to more predictive results.

2.3.5. True Ligand Identification or Virtual Screening

Generally speaking, CoLiBRI models have been developed and validated using the retrieval of the "true ligand" for a binding pocket. While similar, this is not the goal of virtual screening which attempts to identify all (or at least most) of the ligands that will bind to a pocket. This distinction required a reprocessing of the PDBBind core set in order to properly test it.

In the 2009 version of PDBBind, the core set consists of 219 protein-ligand complexes organized into 73 clusters. For each cluster, its 3 members were aligned using ClustalX[45] to determine whether the proteins contained therein were actually the same protein. A protein was considered to be the same as another if there was no more than 1 point mutation or insertion in the body of the protein. 5% of the protein's residues at the head and tail of the protein were omitted from consideration when examining the protein sequence since alterations at the head or tail are common to aid protein purification and crystallization. 49 of the 73 clusters proved to meet the above criteria and the three complexes' ligands for each cluster were considered to be "true

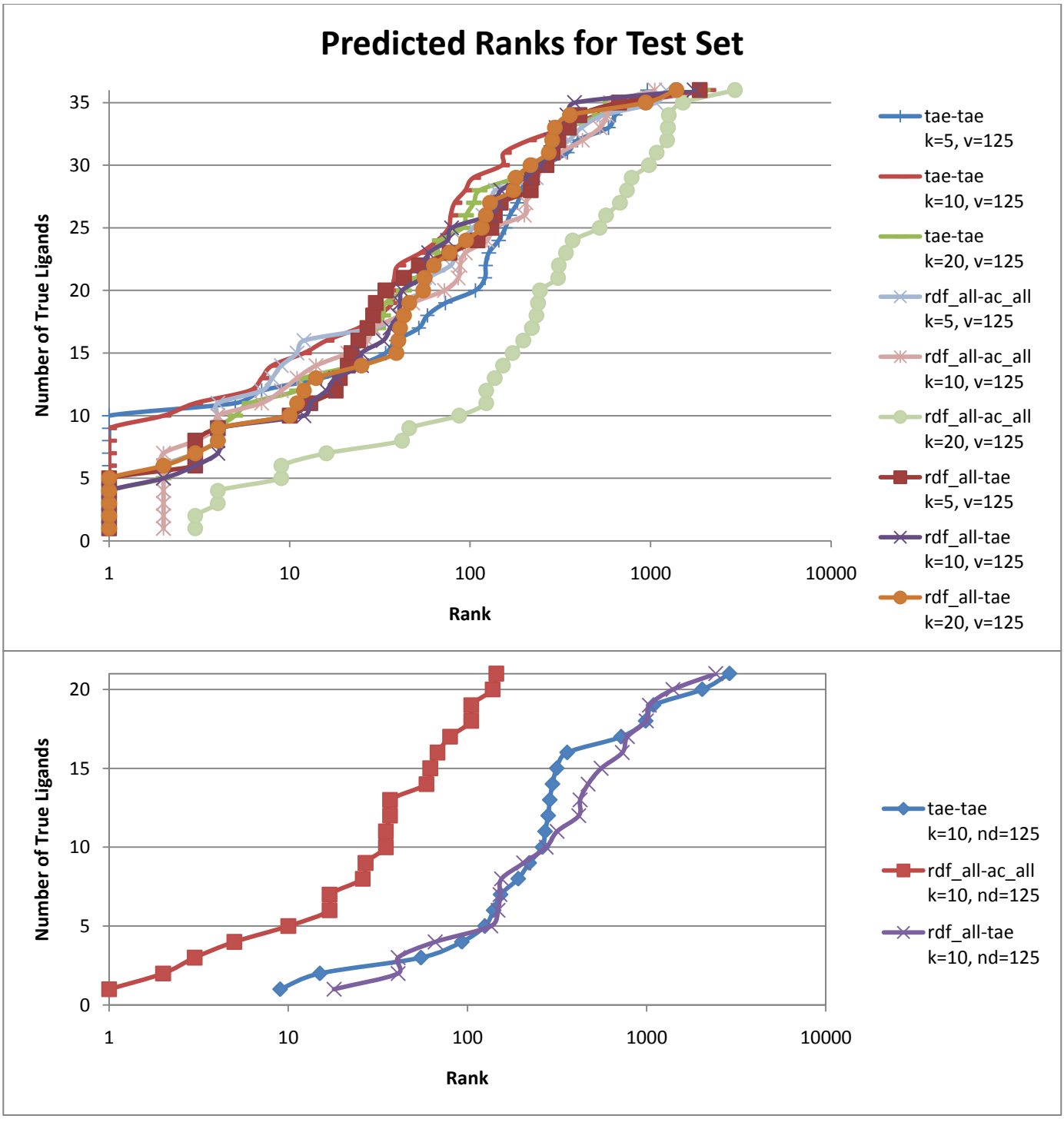


Figure 10. *Comparison of (a) test and (b) external predictive power for different methods of CoLiBRI binding pocket/ligand description.*

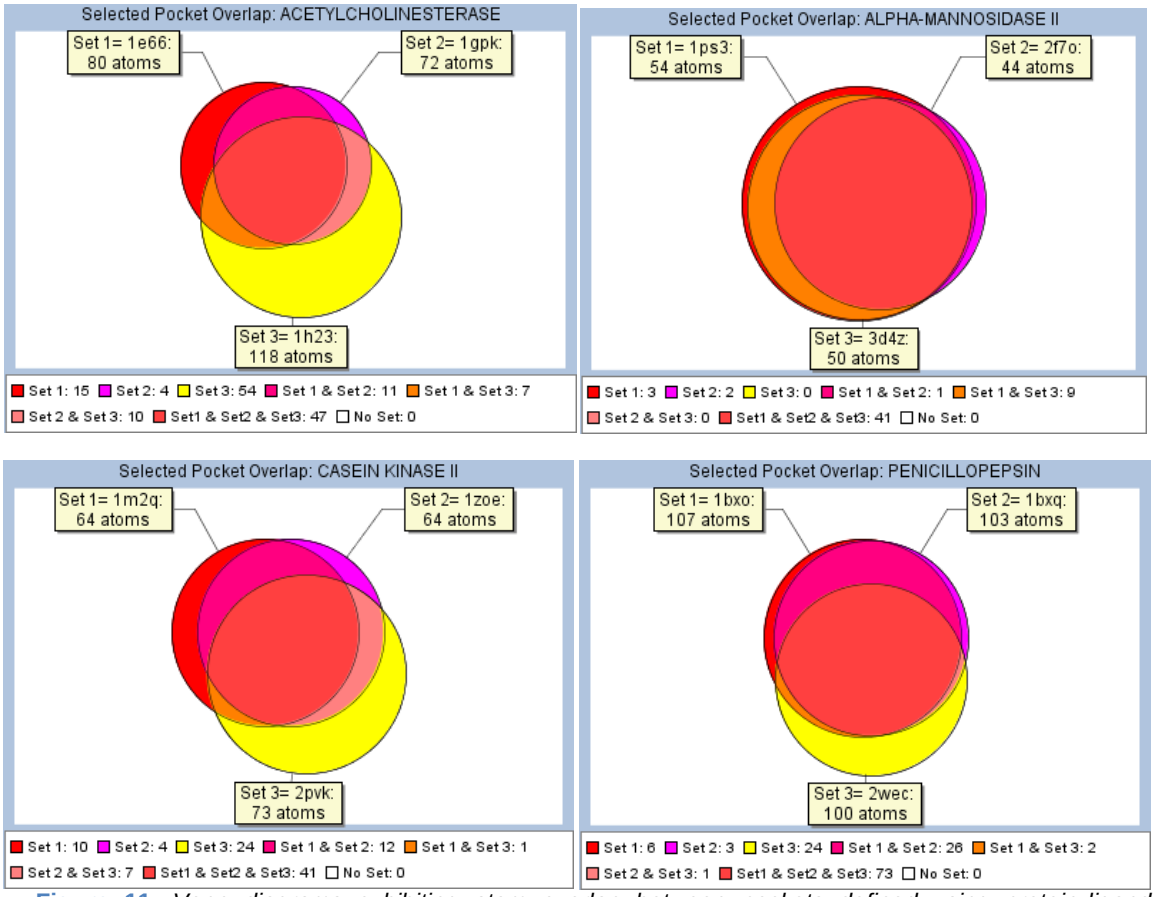


Figure 11. *Venn diagrams exhibiting atom overlap between pockets defined using protein-ligand tessellation.*

ligands" for that protein target. This set was used for all further analysis of the CoLiBRI technique and its members are recorded in Appendix II.

2.3.5.1. Pocket Consistency

A key difficulty that must be addressed when examining the three protein-ligand complexes as a whole is the definition of a pocket. The protein-ligand tessellation method of pocket identification provides a unique pocket for each protein-ligand complex. These pocket definitions can have a wide variability in their level of overlap. Figure 11 contains example Venn diagrams of the atoms selected as pocket members for different PDB entries for the same protein. (Additional diagrams are provided in Appendix III.) While there is a large degree of overlap in pockets, the difference of on average more than 15% of a pocket's atoms is alarming.

The level of uncertainty in descriptor values caused by this difference in a pocket's constitution for the same protein often leads to being unable to determine if two proteins are the same based on their pocket representations in multidimensional space. Figure 12 shows the distributions of distances in TAE/RECON space between pockets selected using protein-ligand tessellation. While distances between representations of the same protein are skewed toward zero, nearly 50% of the distances between the same protein are larger than the smallest of distances between different proteins. We hypothesized that a portion of this difference in pocket definition could be rectified by training CCA with connections between all three of a protein's pockets and all three ligands rather than a single connection between each complex's pocket and ligand. To obtain a better grasp of the feasibility of this approach, PCA and CCA were applied to the dataset and the co-localization of pockets and ligands of the same protein were visually inspected. An example of this analysis with representatives of acetyl-cholinesterase marked in red is displayed in Figures 13-15. It was visually apparent that while correspondence between pockets and ligands in their respective spaces is high after CCA analysis, the multiple representatives for a single protein still have other pockets interspersed. Thus, modeling alone was insufficient for dealing with pocket differences.

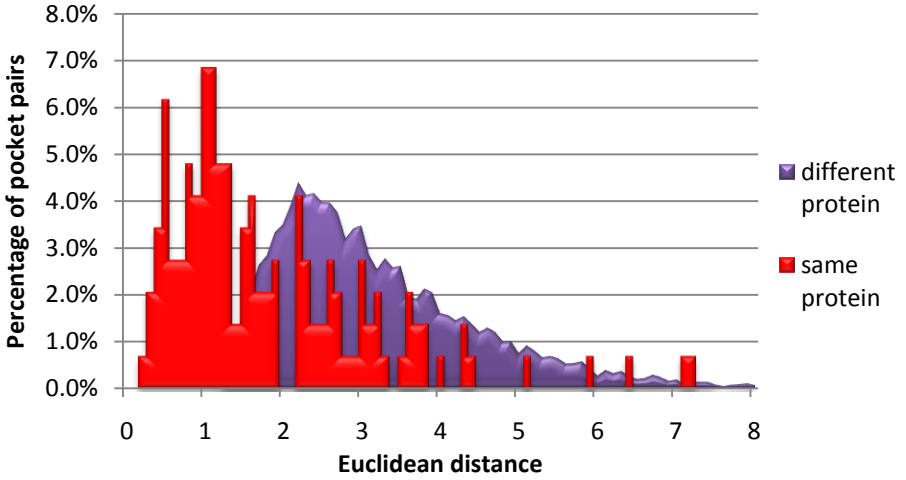


Figure 12. *Histogram of distances between pockets selected with protein-ligand tessellation in TAE/RECON space*

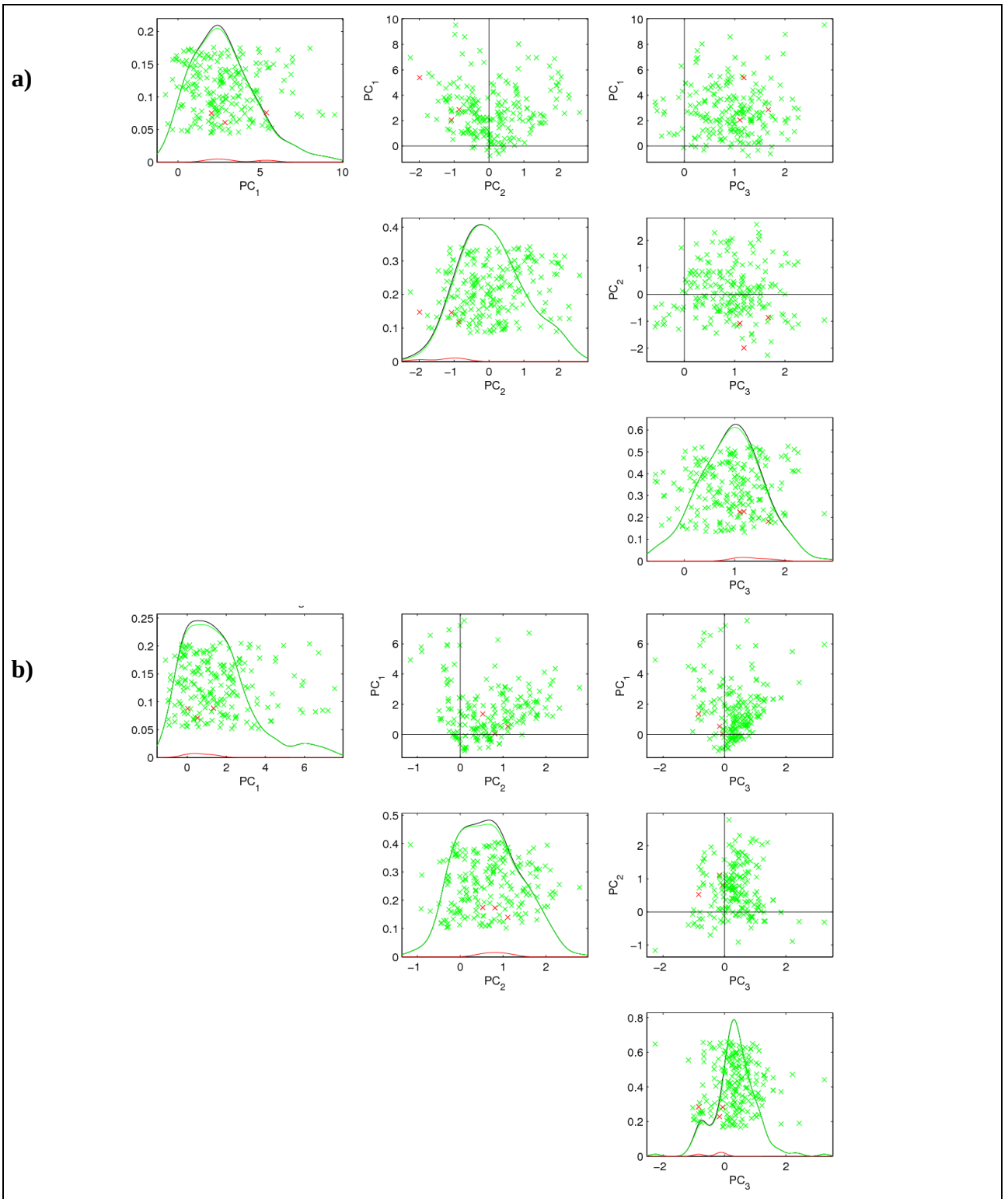


Figure 13. *PCA projections of (a) protein pockets and (b) ligands in TAE/RECON space.*

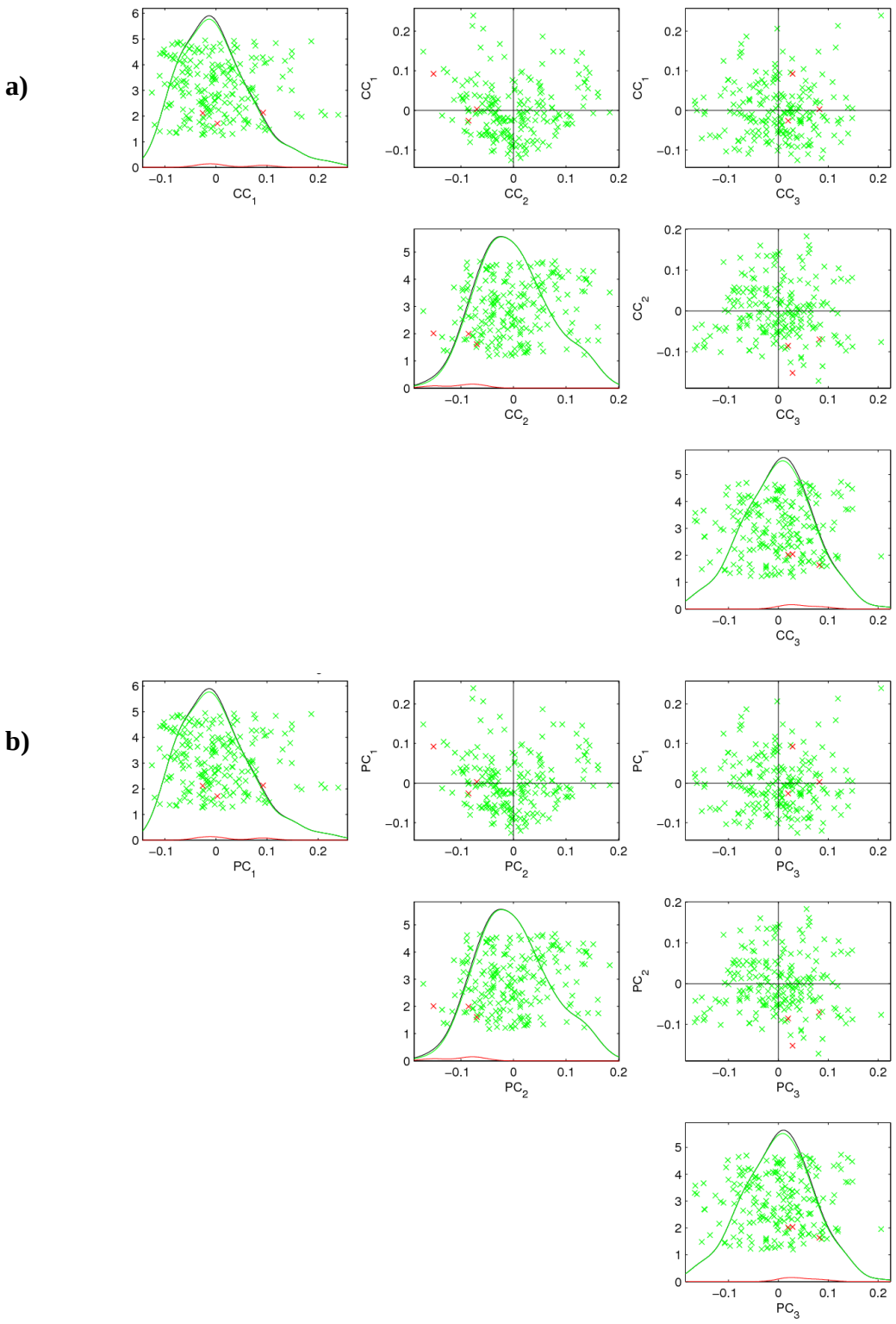


Figure 14. *CCA projections of (a) protein pockets and (b) ligands in TAE/RECON space when only a single connection between complex's pocket and ligand was modeled.*

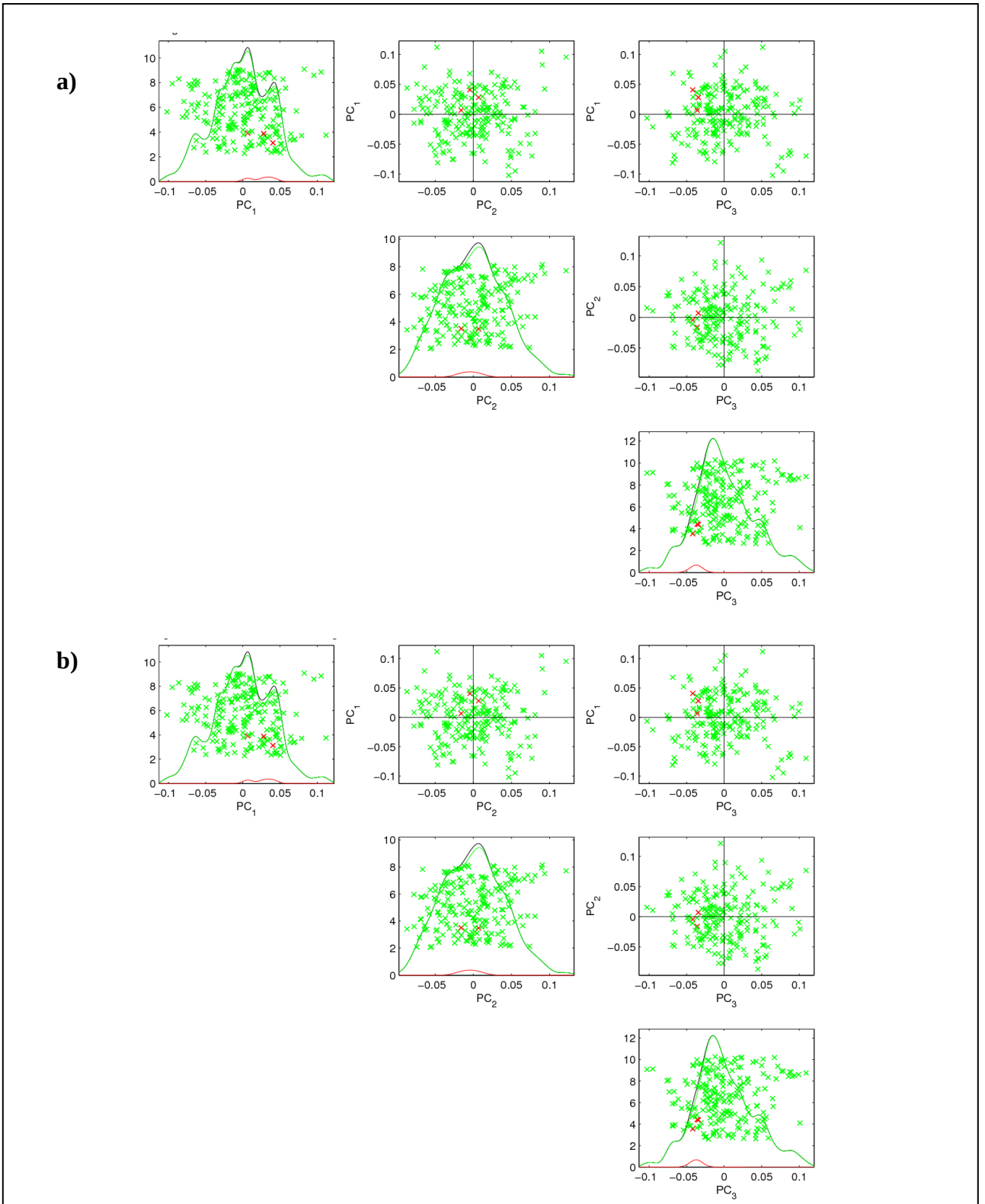


Figure 15. *CCA projections of (a) protein pockets and (b) ligands in TAE/RECON space when connection between all three representatives of a protein pocket and all three ligands were modeled.*

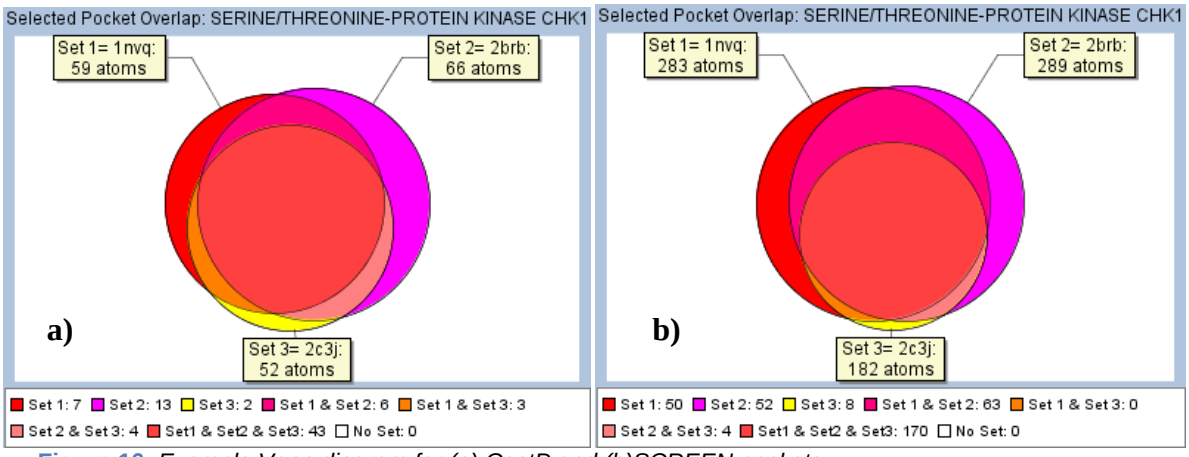


Figure 16. *Example Venn diagram for (a) CastP and (b)SCREEN pockets.*

Protein-ligand tessellation to determine the pocket for each protein was convenient; however, some (including me) would claim that use of the crystallized ligand to map the specific atoms of a protein pocket might imprint ligand information not inherent to protein onto the defined pocket. Therefore we had great interest in converting to a method of pocket identification that was ligand insensitive. We applied 2 protein-only methods of pocket detection: CastP and SCREEN. However, analysis of the pockets identified with both methods indicated that the consistency of pockets identified with these methods was no better than that of pockets identified with protein-ligand tessellation.

Figure 16 displays exemplar Venn diagrams for pockets defined using CastP and SCREEN (with additional examples in Appendix III). It is important to note that for many protein-ligand complexes, CastP and SCREEN were unable to identify the pocket of interest. While many more methods of pocket identification exist, a broad survey of such methods was outside the scope of this study. To discover the achievability of identifying multiple ligands that bind to a single protein, a union of the 3 representative pockets defined using protein-ligand tessellation was considered as the "true" pocket that would be defined by an accurate and consistent pocket identification scheme.

2.3.5.2. Combi-CoLiBRI Modeling

While temporarily setting aside the issue of pocket consistency, generation of CoLiBRI models was realized for the union pockets defined using protein-ligand tessellation. Models were generated with CCA and kCCA modeling methodology using ProFeat[86] descriptors of proteins; TAE/RECON, Feature Pairs, and 3 variants of RDF descriptors for pockets; and Dragon with hydrogens, Dragon without hydrogens, MACCS keys, and MOE2D descriptors for ligands. 5-fold external validation was used to ensure statistical robustness. External folds were dissolved into DrugBank[87] rather than WDI to reduce computational cost. Figure 17 reports the calculated predictive capability using PMRR. PMRR is the average across the 49 proteins of the PMR for the ligands of that protein. Even in the contrived case of "true" pocket definition, prediction accuracy was mediocre. Being that the DrugBank database contains only includes roughly 4500 compounds, the retrieval rates were rough on the order of 10% of the database. Surprisingly, linear CCA performed better than kCCA on prediction of the external sets which directly contradicts the results obtained previously. In addition, it is unexpected that MOE2D descriptors would perform better than dragon descriptors, which are more comprehensive. The ranks of ligands that bind to a pocket (shown in

Figure 18 for one descriptor type and method) by applying kCCA to RDF pocket descriptors calculated using partial charge and polarizability and MOE2D ligand descriptors shows that prediction accuracy for individual proteins covers a braod range. To extract at least one active for 98% of external set proteins at least 250 compounds (around 5% of the database) would have to be screened. If retrieval of all three ligands for a protein was desired, for 80% of the proteins in the database 500 compounds (roughly 10% of the database) would have to be screened.

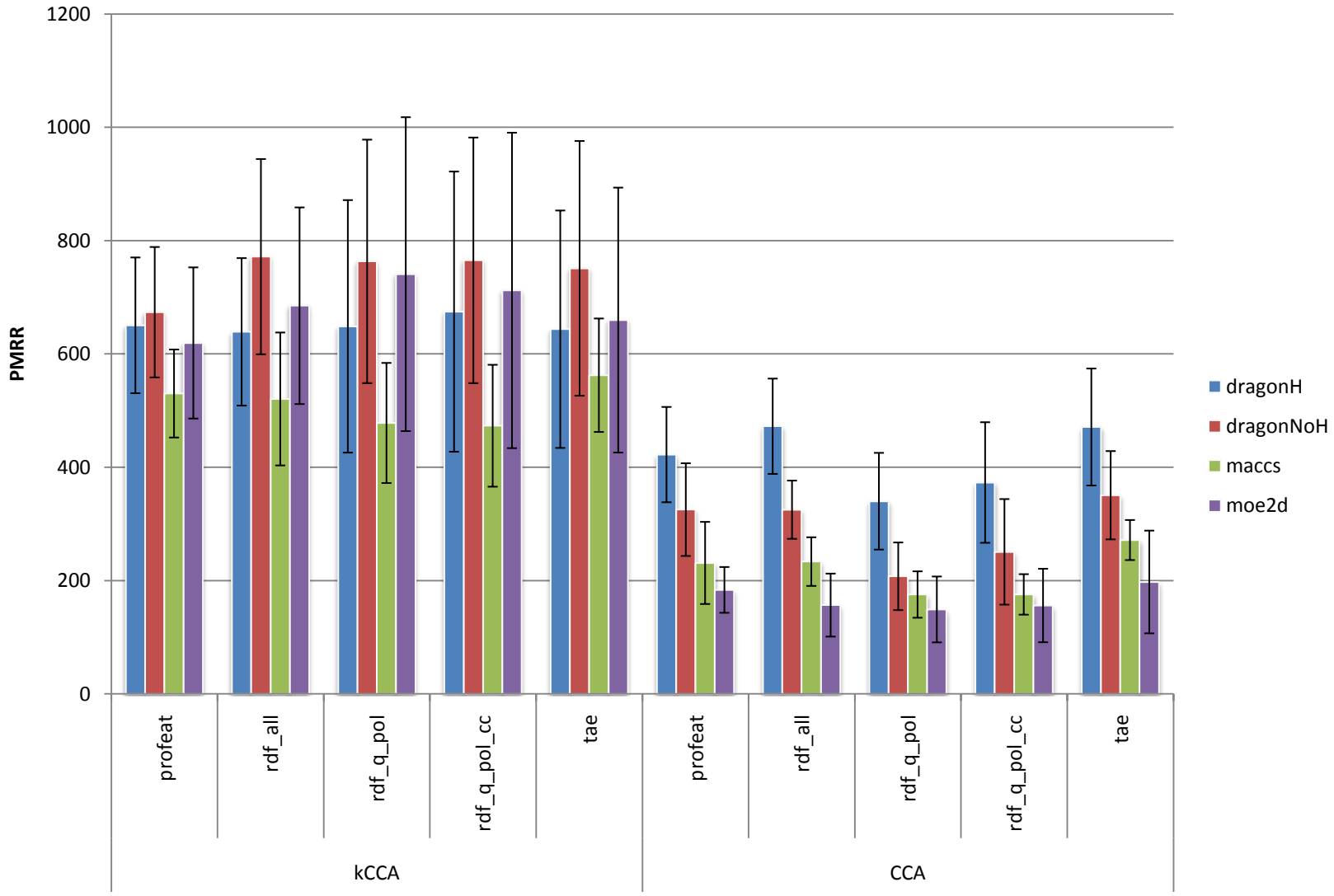


Figure 17. *PMRRs for "Combi-CoLiBRI" analysis*

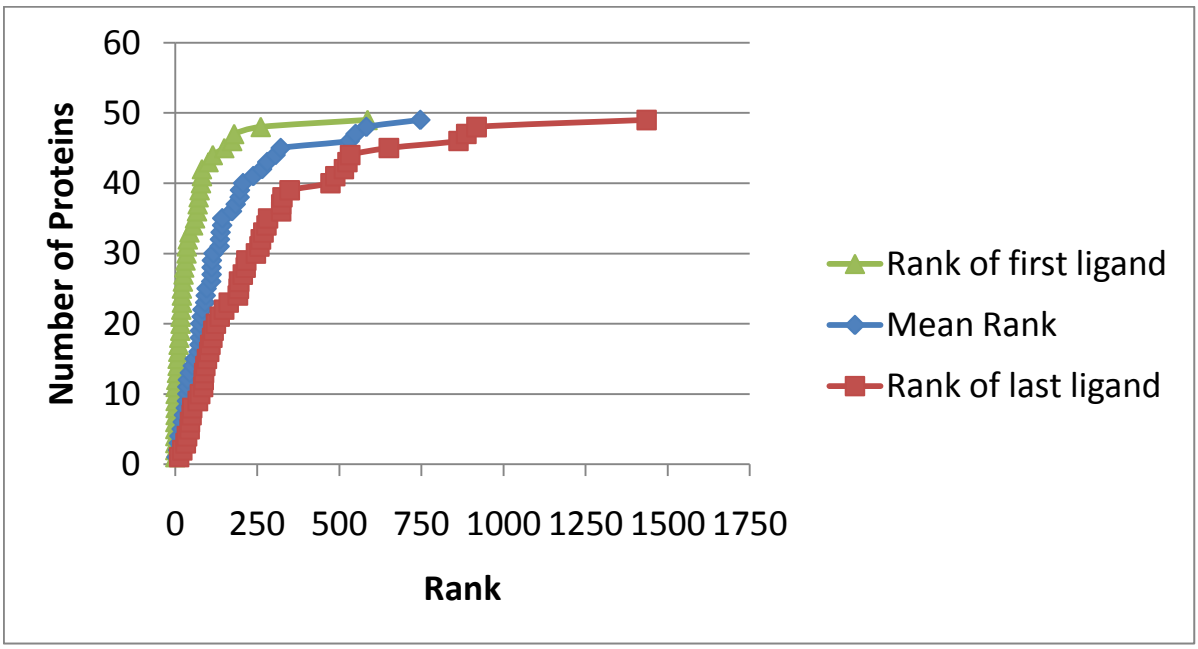


Figure 18. *Retrieval ranks for ligands when modeled by CCA CoLiBRI using rdf_q_pol and MOE2D descriptors.*

2.3.5.3. Ligand Consistency

An additional difficulty unconsidered prior to modeling is that of ligand consistency. Figure 19 displays a PCA projection of ligands in Dragon space with those of the same protein connected by a line. This plot realized with the ADDRAGRA software written in our lab shows the variation in ligand structure that is inherent in our dataset.

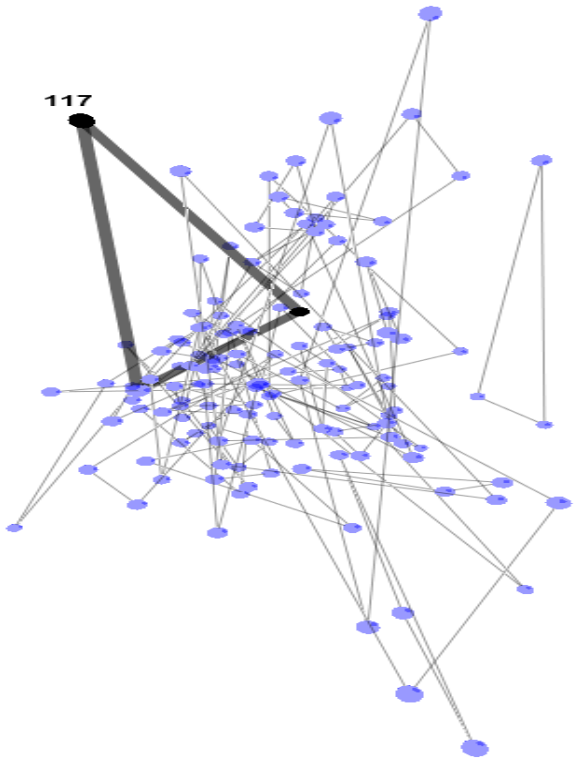


Figure 19. *ADDAGRA plot of ligands in Dragon space with ligands of the same protein connected.*

This variation of ligand structure is likely directly related to the method by which representative ligands were chosen in PDBBind. For each cluster, the complexes in which the ligands had the highest, lowest and median activities were chosen. This means that in several cases the difference between the measured binding affinities for the representative ligands of a protein can be quite large (average of 2.76 log units, maximum of 8.57 log units). This corresponds well to the large differences in ligand structure for the same protein and may explain a portion of the difficulty in prediction. Figure 20 shows the binding modes and affinities for two different ligands of FBKP.

2.4. Conclusions and Future Work

During the course of research into the CoLiBRI workflow for virtual screening of large compound libraries, I have carried out the following tasks:

1. Performed external validation of the original CoLiBRI methodology in screening HIV-protease (Section 2.3.1)
2. Integrated into the CoLiBRI workflow a novel method for optimizing multiple multidimensional spaces (Sections 2.3.2 and 2.3.3)
3. Assessed the capabilities of two additional techniques for protein pocket designation (Sections 2.3.5.1)
4. Implemented two new methods of protein pocket description (Sections 2.2.3.2 and 2.2.3.3)
5. Performed “Combi-CoLiBRI” using available methods of pocket and ligand description (Section 2.3.5.2)

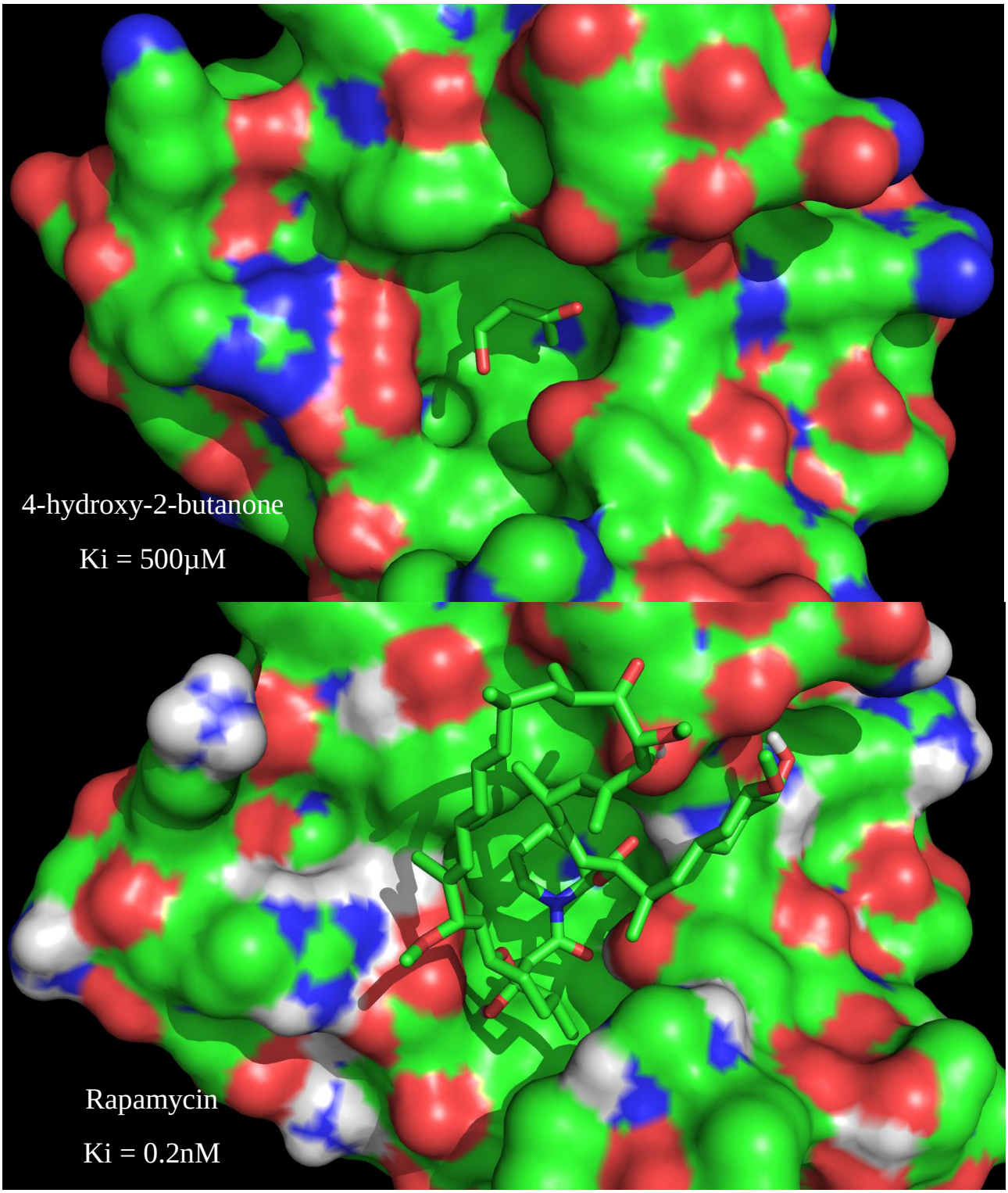


Figure 20. *Binding modes and affinities of 2 ligands of FKBP*

The CoLiBRI workflow exhibits excellent results when "re-docking" ligands with the protein pocket extracted from the same complex. Additionally, when applied in a situation more akin to virtual screening (i.e. the extraction of multiple ligands using a single pocket), CoLiBRI allows the elimination of a large portion of the chemical library and does so with a rate of screening several orders of magnitude faster than typical structure-based methods. Pockets in the database for which this wasn't the case typically had diverse ligands with a broad range of binding affinities. The primary limitation to the use of CoLiBRI in general structure-based studies is that the identification of a protein's pocket reproducibly across the PDB entries of a single protein was unattainable by the tested pocket definition software. This causes an unacceptable level of uncertainty in the description of the protein pocket and subsequently in ligand ranking. For this reason, a more extensive analysis of the consistency of pocket prediction should be completed; perhaps even the development of a technique for pocket detection that will provide consistently defined boundaries should be included.

There is additional refinement required in the analysis of CoLiBRI as a virtual screening technique. While the retrieval of all three representative ligands for a protein is more similar to virtual screening than the identification of just one, a thorough benchmarking of CoLiBRI using the benchmark developed in Chapter 2 is required. Also, the variation in prediction accuracy for different protein members should be examined in greater detail to determine if there is a rational way to form an applicability domain for CoLiBRI models. Finally, CoLiBRI should be built and using a more extensive set of protein crystal structures and known binders.

Chapter 3: Benchmarking of Virtual Screening Techniques

3.1. Introduction

Virtual screening methodologies all have unique advantages and disadvantages. As such, it is generally accepted that no method is unilaterally better than every other method of virtual screening. While typically new methods are tested and shown to be useful on a small number of well documented sets[88, 89], this type of investigation provides little statistical validation of the usefulness of the tool and no understanding of the proper situation for application of the technique.

While comparison within the fields of structure-based and ligand-based techniques are often undertaken[54, 82, 90, 91], there has only been a minimal amount of study across the two fields. Theoretically, if enough active compounds are known for a particular target, ligand-based methods should provide better predictions than structure-based methods; however, the amount of binding data required to a make the application of a ligand-based technique more advantageous is still unknown. Therefore, a thorough comparison between structure-based and ligand-based methods for virtual screening must be carried out on several targets that have a sufficient amount of known binding data.

Though there are already benchmarks for docking (Directory of Useful Decoys (DUD)[92]) and QSAR (Mittal et al.[93]), there is not a benchmark intended to be utilized by both methods. The importance of such a benchmark can be noted by the attempts of some to benchmark ligand-based tools[94, 95] with the DUD database even though it was designed so that decoys could easily be separated from binders with topological indices. This study's intent was to define a set of targets with available binding data to be used as a benchmark for virtual screening in the public domain. After generation of this set, preliminary testing of QSAR methods, similarity searching, and docking were carried out to demonstrate the utility of such a set.

3.2. Materials and Methods

3.2.1. Databases

The benchmark datasets were drawn from extensive databases containing large amounts of biological activity data. The databases (ChEMBL, WOMBAT, and MDDR) used in this study are described *infra*.

3.2.1.1. ChEMBL

The ChEMBL database[5] is a publicly available repository of "drug-like" small molecules linked with biological assay data. This biological assay data is

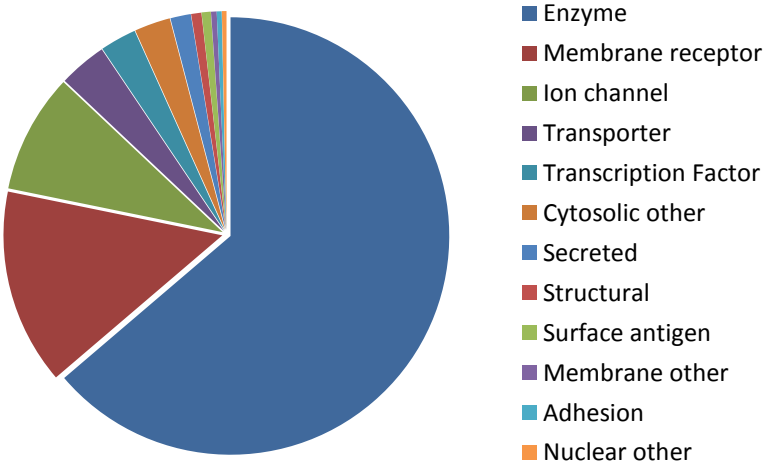


Figure 21. *Distribution of ChEMBL targets.*

extracted from peer reviewed scientific literature and curated by members of EMBL-EBI. The ChEMBL database contains nearly 600K compounds, 450K assays, and 7.5k targets. The distribution of ChEMBL's targets over the proteome can be seen in Figure 21.

3.2.1.2. WOMBAT

WOMBAT[96] is a commercial product of Sunset Molecular. The database is also populated with data from scientific literature, but the data is specifically taken from selected articles within medicinal chemistry journals. In total, the wombat team has indexed more than 15,000 articles and annoted over 300,000 entries of biological activity. The distribution of activities in WOMBAT is shown in Figure 22.

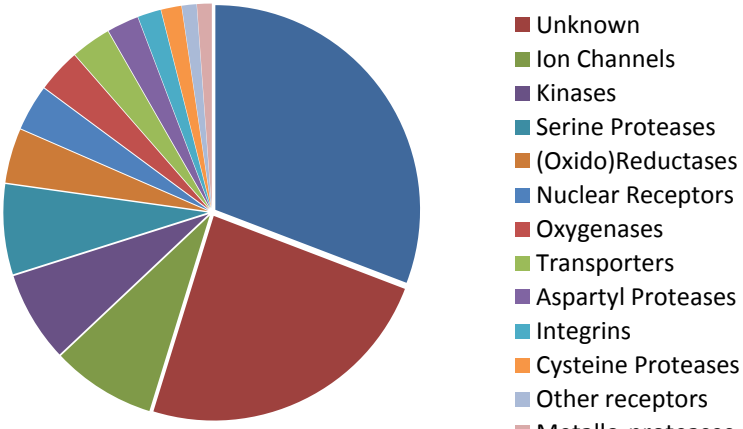


Figure 22. *Distribution of WOMBAT activities.*

3.2.1.3. MDDR

The MDL Drug Data Report[97] has long been an industry standard database covering patent literature and journal submissions. The database was jointly produced by Symyx and Prous Science and is currently being marketed by Accelrys. The database contains over 150K biologically relevant compounds with biological activities classified using the Prous classification system. The compounds are also annotated with trade names, company codes, generic names, originating company, and its current phase of development.

3.2.2. Dataset Extraction

Based on preliminary searching of the available data, a subset of biological targets was selected to form the benchmark set. Compounds and their associated activities were compiled to create datasets for both modeling and validation from the ChEMBL database with the one exception being the Ack1 dataset, which was compiled from 3 patents. For each target, an additional search of WOMBAT and MDDR was completed to extract fully external sets.

Each compound set was cleaned thoroughly. All molecules were processed with Pipeline Pilot[98] to remove salts and solvents, normalize protonation states, standardize chiral definitions, and aromatize the molecules. Activities for the ligands of each target were categorized as either active or inactive using an upper and lower threshold that provided roughly balanced sets for each target and eliminated compounds with uncertain activities. Subsequently, duplicate structures were identified and an inspection of the activities of duplicates was carried out. Duplicates for which binned activity disagreed were removed while duplicates for which binned activity agreed had a single representative retained. A detailed description of processing of ligands for each target is contained in Appendix IV.

3.2.3. Dataset Splitting and Screening

To properly assess the effect of modeling set size on modeling statistics and virtual screening the data splitting scheme show in Figure 23 was applied for each target.

In all cases, the dataset of chemicals for a target drawn from ChEMBL was split into modeling and validation sets using the 5-fold method. While the definition of a modeling set is unnecessary in the case of docking, defining external sets that are the same across all methods is

ideal for comparison purposes. The external sets were then dissolved into the compendium of ligands from other targets to provide a larger number of decoys in our screening sets.

Subsets were generated from each modeling set to preserve the integrity of the validation sets. For each modeling set, five subsets were selected of size 26, 50, 100, 250, and 500 for analysis with QSAR and similarity searching yielding 25 total subsets. These subsets were randomly selected with an equal number of representatives from each class. For some datasets, subsets of size 250 and 500 were omitted due to a lack of data. An additional five subsets were selected of size 1 and 5 from each modeling set yielding 10 more sets for similarity searching. The smaller sets selected for similarity searching were drawn only from the active class. In total, at most 130 ensemble QSAR models were developed (25 subsets * 5 modeling sets + 5 modeling sets). Each model was used to predict the appropriate screening library. A total of 180 similarity searches using probes drawn from the modeling sets were completed on the appropriate screening library. In addition, each screening library was ranked by docking and

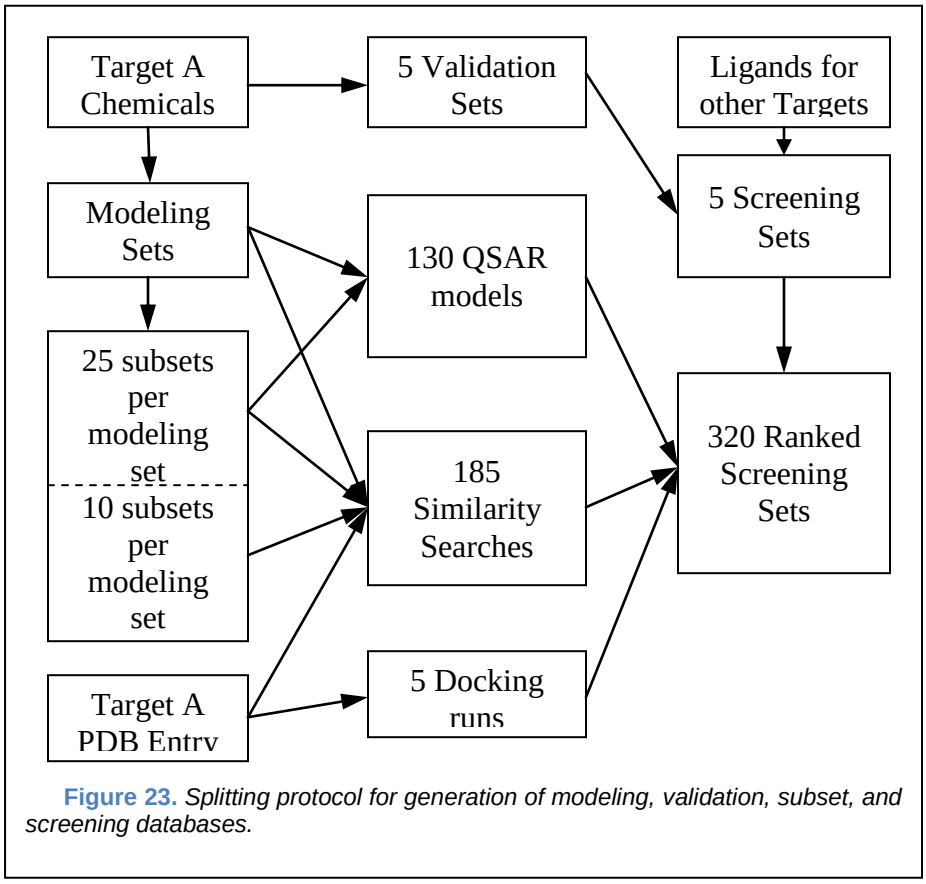


Figure 23. *Splitting protocol for generation of modeling, validation, subset, and screening databases.*

by a similarity search with the ligand contained in the PDB entry. Thus at most 320 differently ranked screening libraries were generated for each target.

3.2.4. Docking

To prepare the protein structures for docking, all the water molecules and ions associated with the structure were removed. eHiTS was used to preprocess the protein by extracting the ligand from the PDB complex files and generating a native eHiTS file format. A radius of 7.5Å to 10Å was used to define the active site and calculate steric grids and feature descriptions.

The ligand database for docking was prepared using LigPrep Module as implemented in Schrodinger 9.2. LigPrep provides an efficient way to prepare all-atom 3D structures, starting from 2D or 3D structures. Low energy 3D structures of ligands were generated from canonical SMILES strings using OPLS 2005 force field. Calculation of possible states at pH 7±2 resulted in generation of the correct ionization state. Specified chiralities were retained from the canonical SMILES and the lowest energy conformation of rings was retained for each ligand. Hydrogen atoms were added to complete valences as necessary. Ligands just comprising ions or molecule fragments having 4 atoms or less were removed. Structures that caused processing failures in the energy minimization of the structures were also removed. In the end one unique conformation per ligand was retained for docking.

eHiTS v2009.1 (www.symbiosys.com), an automated docking software, was used for virtual screening. The eHiTS software package[99] is a flexible ligand docking program that utilizes exhaustive fragment based search algorithm to dock and then energetically optimizes the 3D coordinates of docked poses within the active site of target. One of the critical steps in a successful docking approach is to correctly position each ligand in the binding site based on the

defined constraints. This step involves exploration of the configurational and conformational space for the interaction between target and ligand. This step attempts to correctly identify the most favorable binding mode of the ligand in the target active site.

The eHiTS docking algorithm docks rigid fragments generated from a ligand independently within a binding pocket. The binding pocket is represented using Geometric Shape and Chemical Feature graph (GSCF), where nodes of the GSCF graph represent a rigid shape by a simplified geometric hull generated from regular polyhedra where each vertex of a polyhedron is encoded with its chemical properties. The ligand is broken down into rigid fragments and flexible chain/linker atoms/fragments. Each fragment is also represented using a GSCF graph made up of regular polyhedra with chemical properties associated with each vertex of the polyhedron. Each rigid ligand fragment is docked in each cavity polyhedron during the rigid docking phase by matching and exploring each cavity-ligand fragment orientation. Thousands to millions of fragment poses are generated within the binding cavity depending on the size and fragmentation pattern of original ligand.

Poses are then selected using a fast graph matching algorithm and rigid fragments are reconnected through their flexible linker atoms that comprise the matching pose set. Flexible chains are tweaked and optimized such that its end matches the rigid fragment precisely without violating any energetic and steric constraints. The final binding poses are refined by a local energy minimization in the active site of the receptor, driven by eHiTS scoring function. The binding energy of each pose is calculated and reported as eHiTS score.

The eHiTS scoring function is based on a combination of novel scoring term (local surface point contact evaluation) plus a hybrid scoring term based on traditional empirical and

knowledge based scoring functions. The interaction score between fragment surface points and receptor surface points are computed from the interaction statistics collected separately for distinct types of surface point pairs. Surface points are classified into 23 types and interactions between ligand and receptor surface points are recorded. The random probability of interaction is used to convert to interactions into an energy term using energy scaling factors. Besides this energy term the final scoring also includes terms for steric clash, depth value, conformational strain energy of the ligand, entropy, intra-molecular interaction, receptor surface coverage, and family coverage. The terms of the scoring function are combined using adjustable weights for each protein family. To train these weights, interaction statistics were collected for all pairs of atoms within 5.6 Å of each other for a set of ~1420 high resolution protein-ligand complex. The complexes were clustered into 71 clusters or families and family specific weight sets were generated. In addition to the family optimized scoring function, eHiTS allows new scoring weight sets to be generated by training the scoring function with addition protein-ligand complex data or with known active and inactive ligands.

While the scoring function of eHiTS program can be trained using known actives and inactives to bias the function toward finding ligands that are more similar to known actives. However, in our benchmarking study we carried out an unbiased docking based on default eHiTS parameters. Compounds were ranked based on the returned eHiTS score.

3.2.5. Similarity Searching

Similarity searching is the simplest form of ligand-based virtual screening. The method typically involves generating a set of multidimensional descriptors for both the known ligand(s) and the chemical database, then ranking all compounds in the chemical database based on their similarity to the ligand. There are many different types of descriptors that can be applied and

several different ways to assess similarity. In this study, similarity searching was carried out using both the ligand contained within the protein-ligand complex obtained from the PDB and the actives in the modeling sets defined *supra*. Similarity was assessed using the Tanimoto coefficient and FCFP4 from Pipeline Pilot. Compounds were ranked based on their similarity to the nearest probe.

3.2.6. QSAR

The generation of all QSAR models was accomplished through use of the Chembench web portal. Only random forest modeling was applied as the number of modeling set to be analyzed was large. In all cases, the random forest modeling procedure Chembench was applied to Dragon descriptors[100] of chemical structure with the following selections: range scaling of descriptors and elimination of descriptors with perfect correlation, 50 random divisions of training/test set containing between 20% and 30% of the dataset, and 50 trees generated for each split using 50 descriptors. Further discussion of the random forest procedure implemented in Chembench is contained in section 5.2.2. During screening, each compound was scored and ranked using the percentage of models within the ensemble that predicted it to be active.

3.3. Results and Discussion

3.3.1. Preliminary Error Analysis

Prior to defining datasets for each target, it was necessary to assess the level of noise that appears in the numeric activity values contained in the bioactivity databases. Since these databases are extracted from peer reviewed literature, we expected the data to be of high quality. To verify this hypothesis, activity values were extracted for all 22 targets of interest from the ChEMBL database. Of the 43319 entries returned, 6917 were identified as duplicates (the same

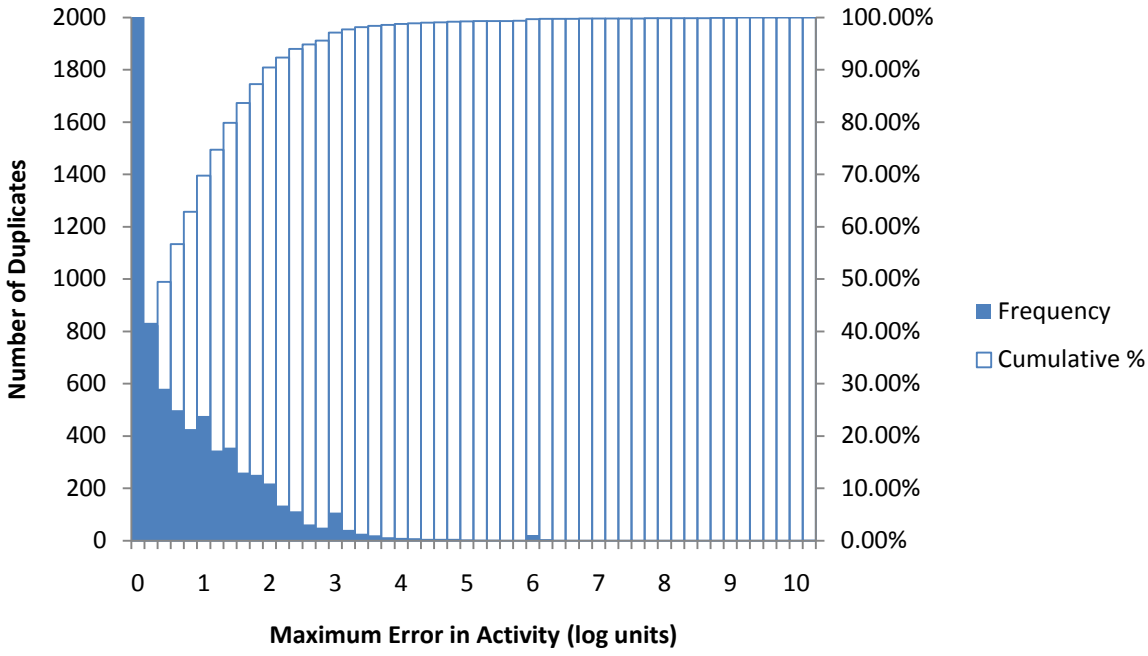


Figure 24. *Maximal difference in activity reported for duplicates in ChEMBL.*

ChEMBL ID had multiple listed activity values for a target and an activity type). Figure 24 contains the histogram of the maximal difference in activity values for each duplicate. The cumulative histogram indicates that the majority (70%) of duplicates have a maximal difference between reported activities of less than one log unit. 90% of duplicates have a maximal difference of less than two log units. Based on this information, we decided that the inconsistency of reported values within ChEMBL made modeling them problematic. We decided to categorize the continuous values into active and inactive classes. To reduce the error in labeling, the threshold to be considered active was two log units greater than the threshold to be considered inactive.

An interesting side note concerning maximal error measurements is that there are distinctive increases in the number of duplicates with errors of three and six in the histogram. These increases are likely due to errors in interpretation of units when data is being extracted from literature sources.

3.3.2. Selected Datasets

In total, datasets were extracted for 22 targets. These targets cover multiple protein classes whose activities can be modulated by small compounds. This set includes GPCRs, nuclear hormone receptors, and several enzyme families such as kinases and proteases. High resolution protein structures for all targets were identified within the PDB and ligands for each target were extracted from the three bioactivity databases: ChEMBL, WOMBAT, and MDDR. Information regarding the data extracted for each target is contained in

Table 1.

3.3.3. Ranking with Docking

The ranking of screening sets was completed for all targets with eHiTS, a commonly used fast flexible docking solution. For each of the resultant ordered screening sets two Receiver Operating Characteristic (ROC) curves were generated. The first ROC curve is generated considering decoy compounds in the screening set coming from the other targets to be inactives. The second is generated considering only the compounds in the screening set belonging to that target's dataset. Figure 25 contains examples of the former while Figure 26 contains examples of the latter. All ROC curves are available in Appendix V.

While docking did an excellent job in selecting true actives from the full screening sets in the majority of cases, in some cases it was indistinguishable from random prediction. These cases correspond to the lack of a family based scoring function of certain proteins. In order for optimal performance, eHiTS requires that a family be known for a protein. This limitation makes identification of binders of proteins under-populated in the PDB difficult.

Table 1. *Summary of results for benchmark dataset generation*

Abbreviation	Target	PDB ID	Active Threshhold (nM)	Inactive Threshold (nM)	# Compounds in modeling/ validation set	# Compounds in WOMBAT external set	# Compounds in MDDR external set
ACK1	Activated Cdc42-associated Kinase	3EQR	≤ 100	≥ 10000	172	NA	NA
ACHE	Acetylcholinesterase	1EVE	≤ 100	≥ 10000	887	652	860
AR	Androgen Receptor	2AM9	≤ 100	≥ 10000	422	258	NA
B2AR	Beta-2 Adrenergic Receptor	2RH1	≤ 100	≥ 10000	248	137	238
CA2	Carbonic Anhydrase II	3K34	≤ 10	≥ 1000	1073	778	267
CDK2	Cyclin Dependent Kinase 2	2R3I	≤ 100	≥ 10000	1360	756	NA
COX2	Cyclooxygenase 2	3PGH	≤ 100	≥ 10000	1429	811	1168
DHFR	Dihydrofolate Reductase	2W3A	≤ 100	≥ 10000	463	240	250
ESR1	Estrogen Receptor Alpha	2OUZ	≤ 10	≥ 1000	878	799	311
ESR2	Estrogen Receptor Beta	2NV7	≤ 10	≥ 1000	703	681	NA
F10	Coagulation Factor X	2XBV	≤ 10	≥ 1000	999	2050	1648
GR	Glucocorticoid Receptor	3K22	≤ 10	≥ 1000	385	387	NA
HIV-Int	HIV Integrase	1QS4	≤ 1000	≥ 50000	749	954	475
HIV-Pr	HIV Protease	1G35	≤ 10	≥ 1000	1526	2691	1140
HIV-RT	HIV Reverse Transcriptase	2ZD1, 3KK1	≤ 100	≥ 10000	1133	1411	NA
PARP1	Poly [ADP-ribose] Polymerase-1	3GJW	≤ 10	≥ 1000	299	293	377
PDE5	Phosphodiesterase 5A	1TBF	≤ 10	≥ 1000	687	499	660
PNP	Purine Nucleoside Phosphorylase	1VHW	≤ 10	≥ 1000	173	81	82
PPARG	Peroxisome Proliferator-Activated Receptor Gamma	3ET3	≤ 100	≥ 10000	376	340	NA
REN	Renin	3K1W	≤ 10	≥ 1000	1235	536	1529
SRC	Tyrosine Protein Kinase SRC	3G5D	≤ 100	≥ 10000	1443	689	NA
F2	Thrombin	2BVR	≤ 100	≥ 10000	1150	1933	1373

The usefulness of ranking when only known active and inactive compounds were considered was much lower. The ability to select true binders from known non-binders using only eHiTS scoring appears to be quite limited.

3.3.4. Ranking with Similarity Searches

Similarity searching using Tanimoto score and FCFP-4 fingerprints was completed using Pipeline Pilot. Results were obtained using either the ligand from the PDB entry or the actives from a selected modeling set as probes. ROC curves are provided only for the cases of the PDB ligand and full modeling sets. (ROCS using probes extracted from subsets were not generated.) Example ROC curves of type I are displayed in Figure 27 and Figure 28 while those of type II are shown in Figure 29 and Figure 30. All ROC curves are available in Appendix V.

3.3.5. Ranking with QSAR models

The ranking of all screening sets was completed using random forest models developed on Chembench. Models were obtained for all 130 modeling sets (including subsets). The predictive power of all these sets was assessed using their validation sets. Graphics exemplifying the effects of modeling set size on the predictive power of the resultant models are shown in Figure 31. The stability of the resultant models (measured with the standard deviation in CCR) is displayed in Figure 32 for select targets. Appendix VI holds additional examples of these plots.

Examination of Figure 31 confirms that as a modeling set size increases, its predictive power increases. Additionally, Figure 32 shows that generally the stability of a model increased as more compounds are modeled. These results completely agree with what would be expected as set sampling increases.

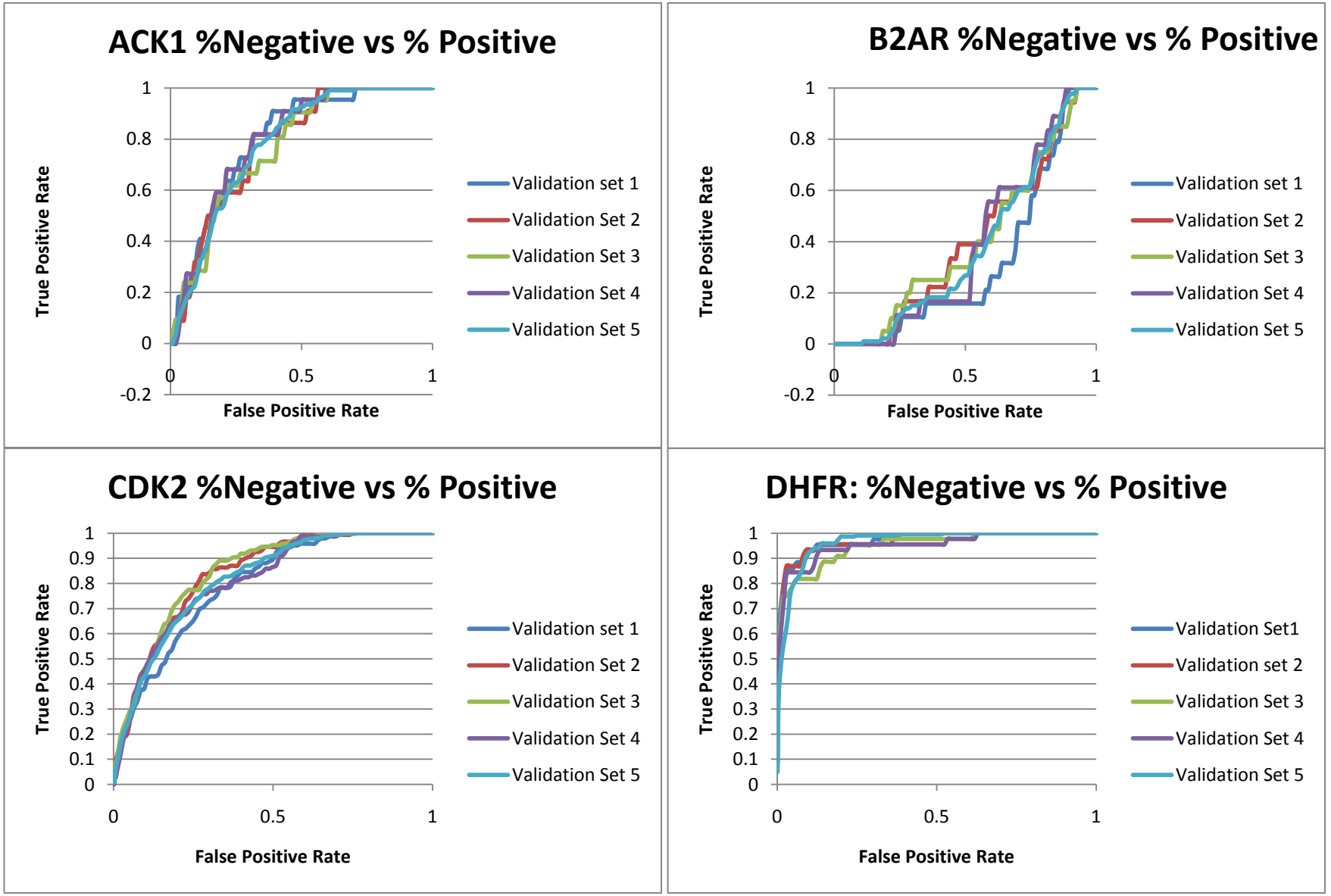


Figure 25. *Example ROC curves resulting from the docking the full screening library using eHiTS.*

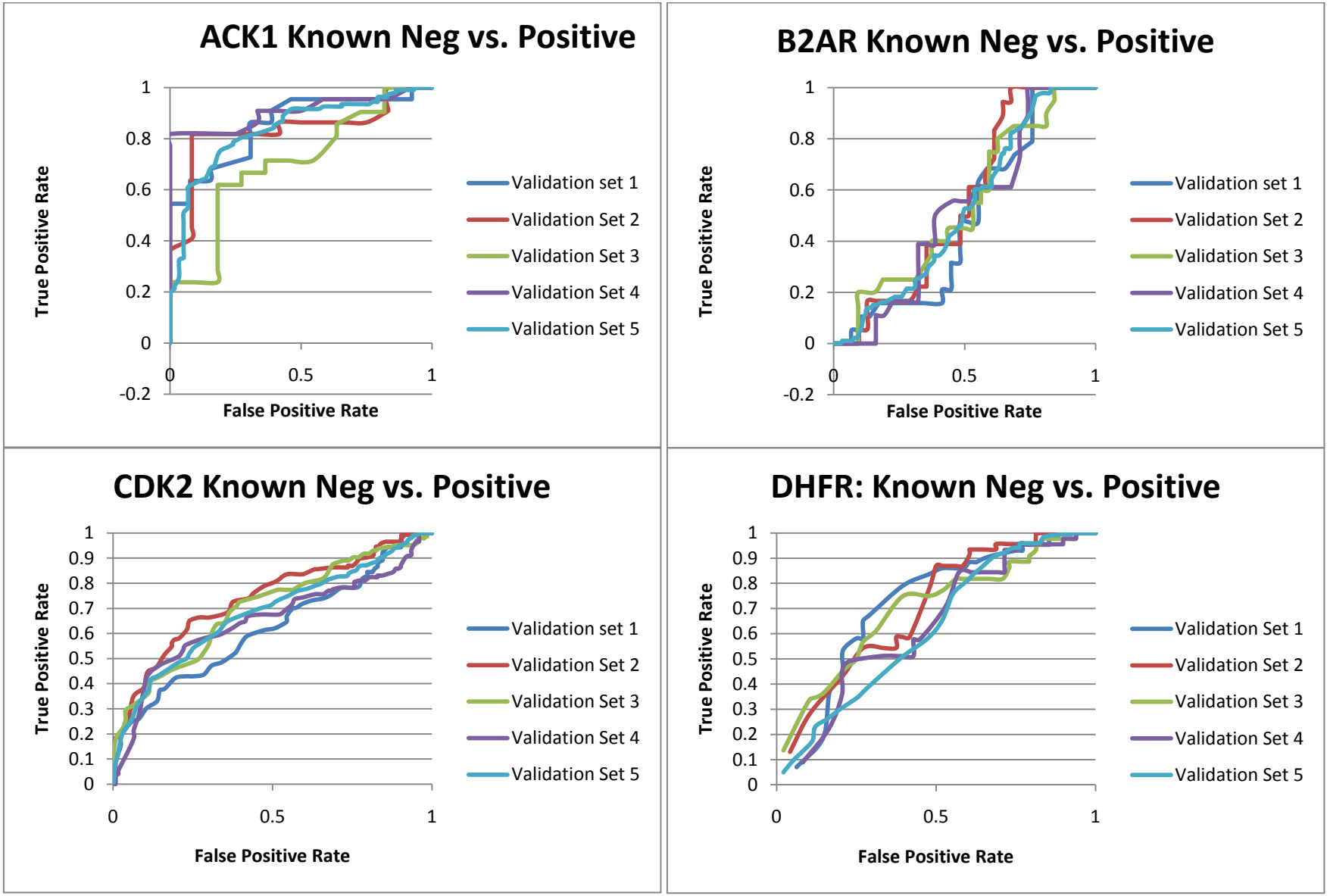


Figure 26. *Example ROC curves resulting from the docking of compounds with known activity using eHiTS.*

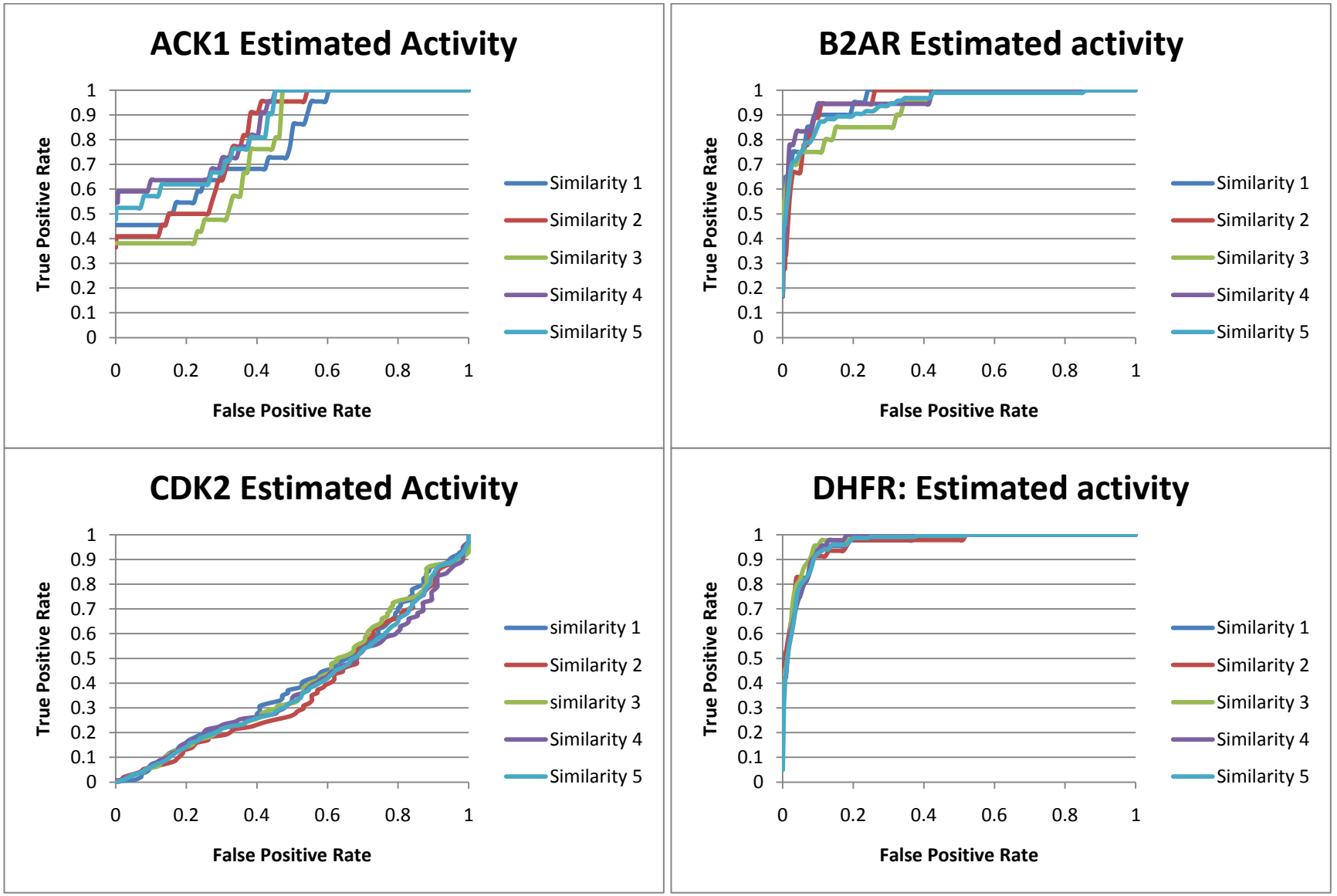


Figure 27. *Example ROC curves resulting from searching the full screening library using the PDB ligand and Tanimoto similarity with FCFP-4.*

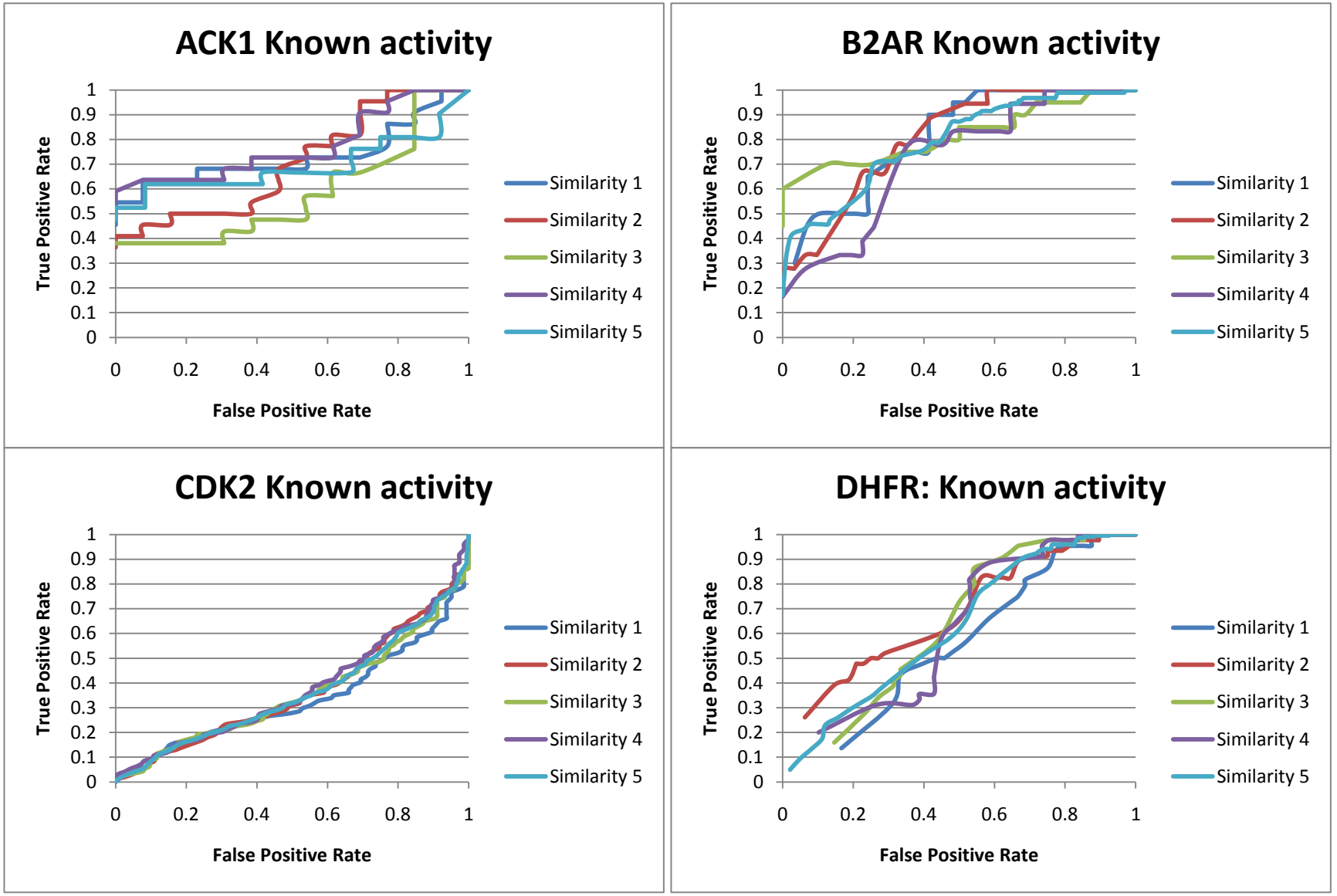


Figure 28. *Example ROC curves resulting from searching compound with known activity using the PDB ligand and Tanimoto similarity with FCFP-4.*

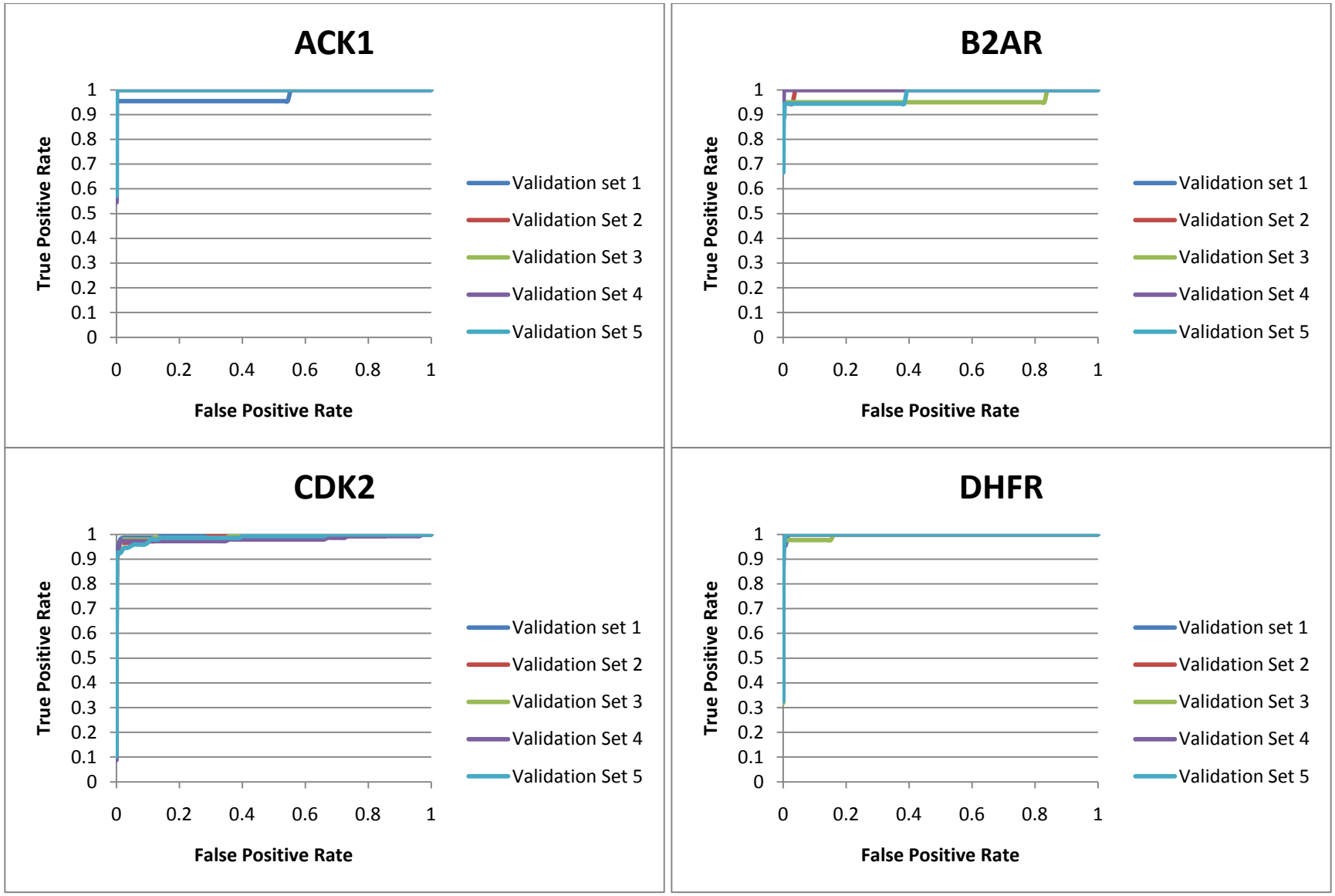


Figure 29. *Example ROC curves resulting from searching the screening library using the modeling set actives and Tanimoto similarity with FCFP-4.*

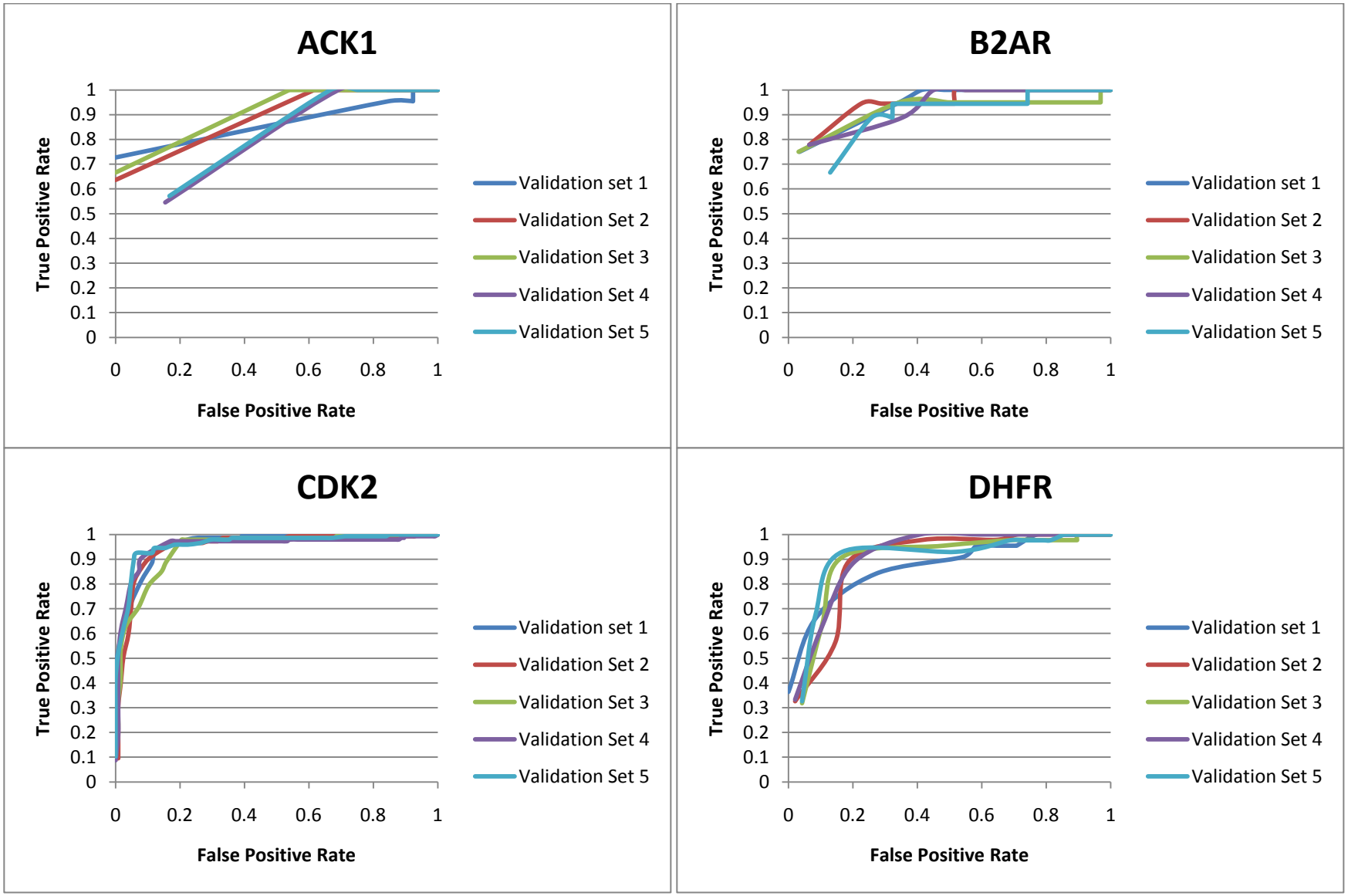


Figure 30. *Example ROC curves resulting from searching compounds with known activity using the modeling set actives and Tanimoto similarity with FCFP-4.*

ROC curves were generated only for models developed using the entire modeling set (not subsets). ROC curves for the entire screening set are available in Figure 33 while those for compounds with known target activity are displayed in Figure 34.

QSAR models were able to effectively rank both the screening library and the compounds with known activities. Being that QSAR models are specifically trained to make separations between actives and inactives in the modeling set, they can be insensitive when predicting compounds not similar to the modeling set; however this is not readily apparent in ROC curves generated when using the entire modeling set. In typical applications of QSAR for virtual screening, a global applicability domain filter is used to guarantee that selected compounds are similar to the modeling set. However, an applicability domain filter was not applied in this case as its most tradition implementation is a similarity search using all modeling set members as probes. The use of multiple modeling methods in concert to obtain superior predictive power was beyond the scope of this study's focus of ascertaining the usefulness of the benchmark set.

3.3.6. Method Comparison

Being that different virtual screening methods require different inputs, it is hard to compare them in an unbiased way. QSAR modeling using 1000 modeling compounds cannot be fairly compared to docking results that rely on single protein structure. However, there are two fair comparisons that can be made. Similarity searching using the ligand contained within the PDB entry as a probe and docking both use only a single protein-ligand complex to rank a chemical database. Also similarity searching using all actives from a modeling set as probes uses the base of knowledge as a QSAR model.

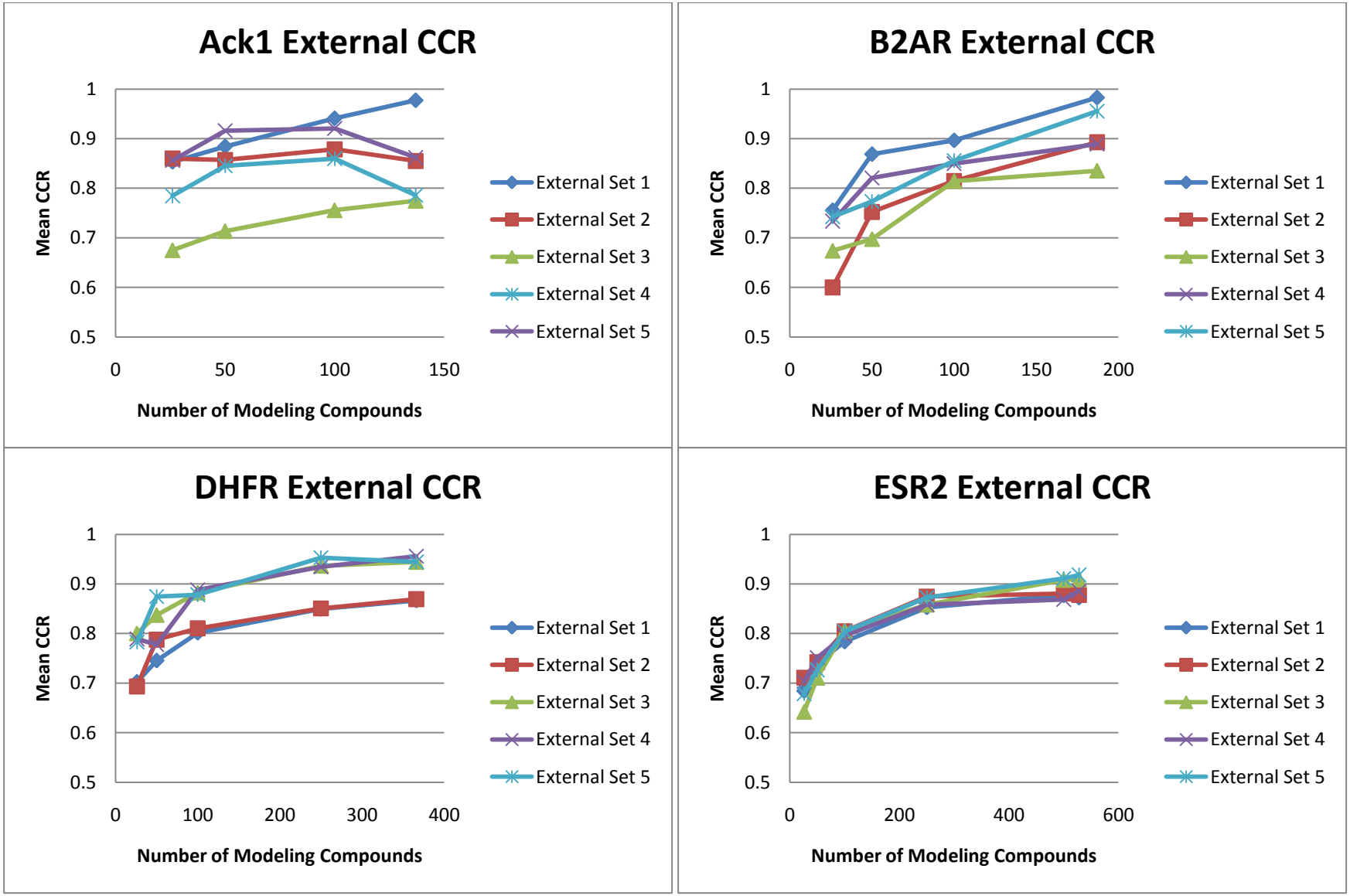


Figure 31. *Prediction accuracy as measured using mean CCR for selected targets.*

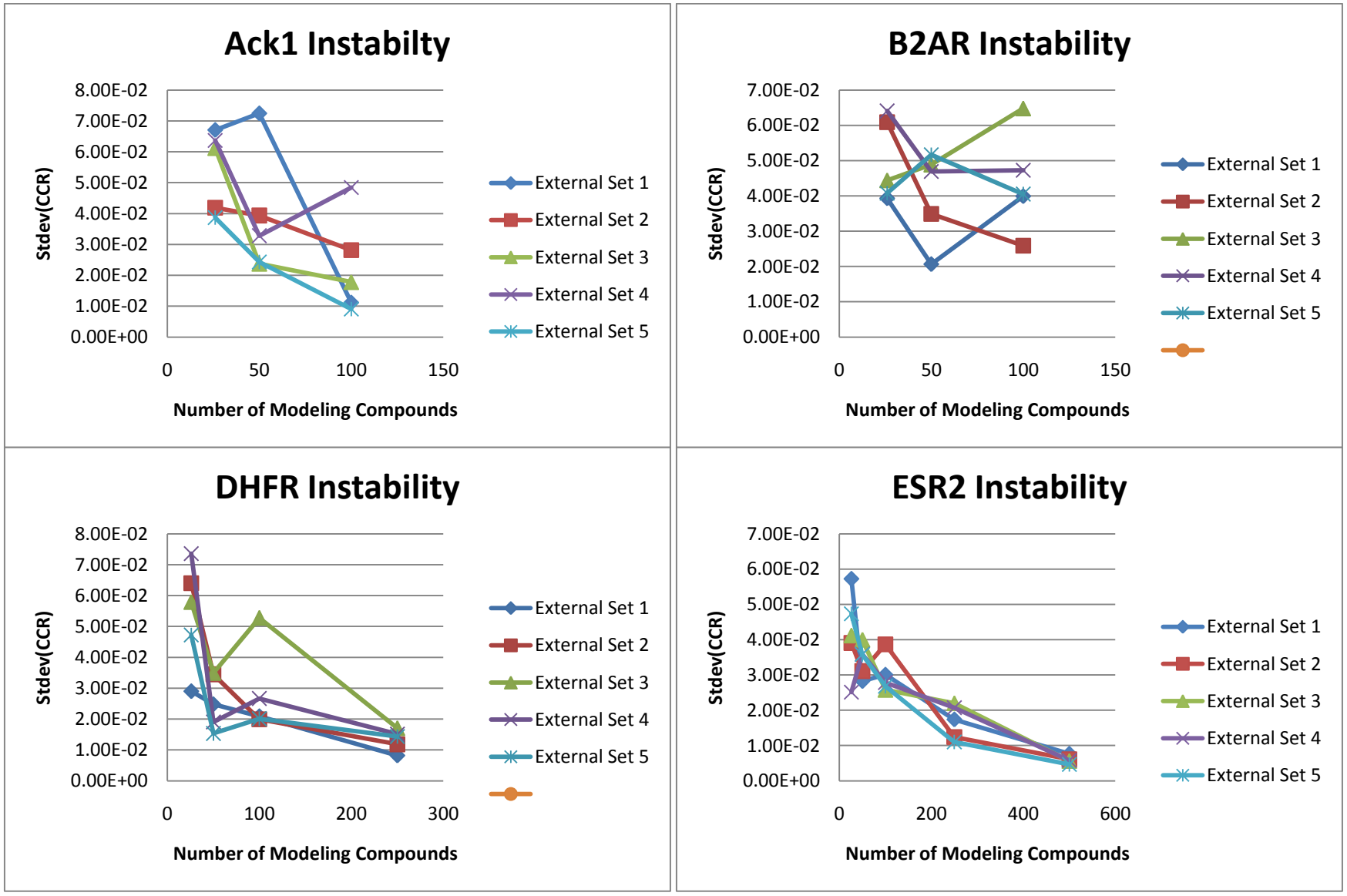


Figure 32. *Prediction stability as measured using the standard deviation in validation set CCR for selected targets.*

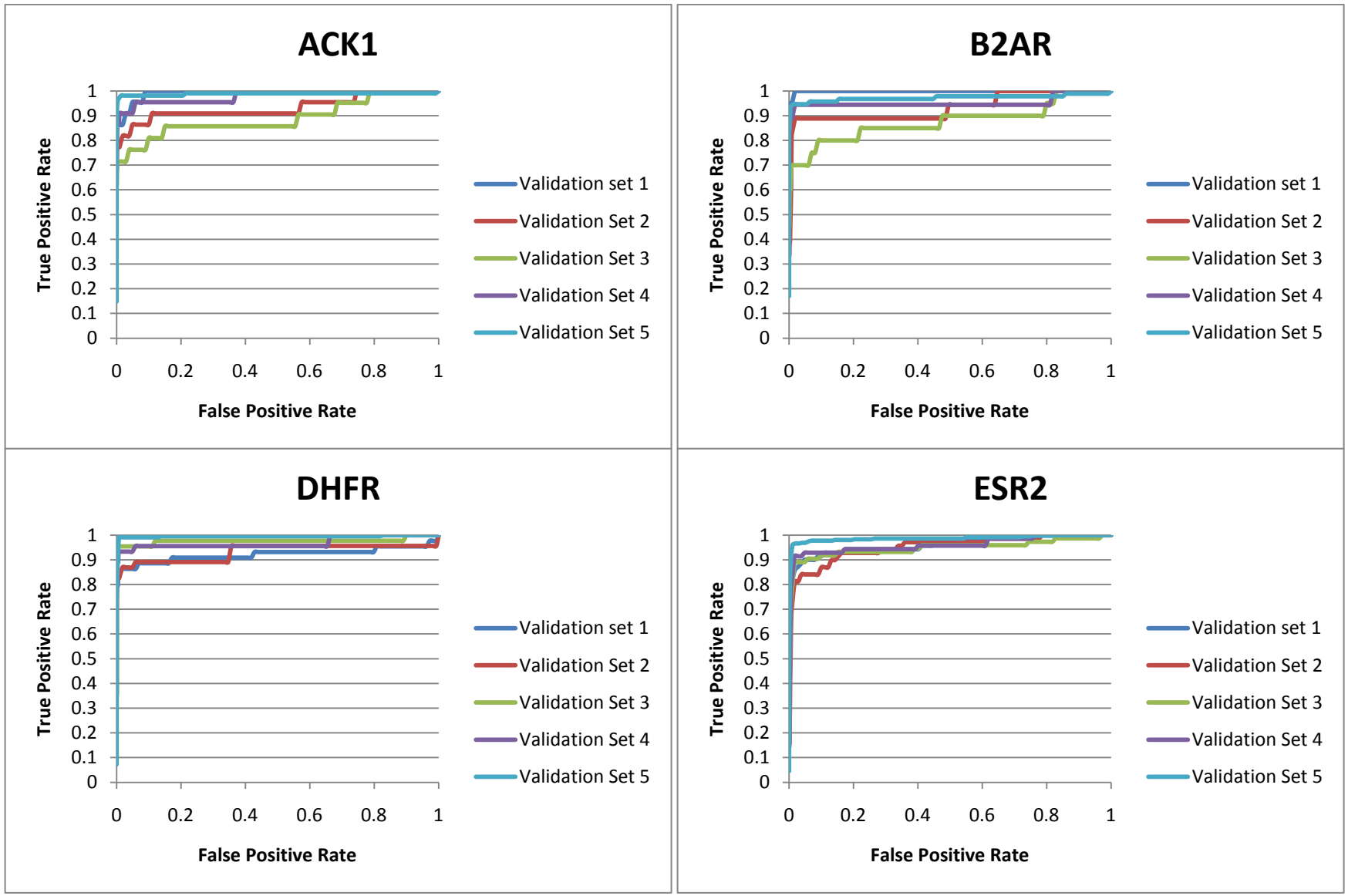


Figure 33. *Example ROC curves resulting from prediction of the screening library using QSAR models.*

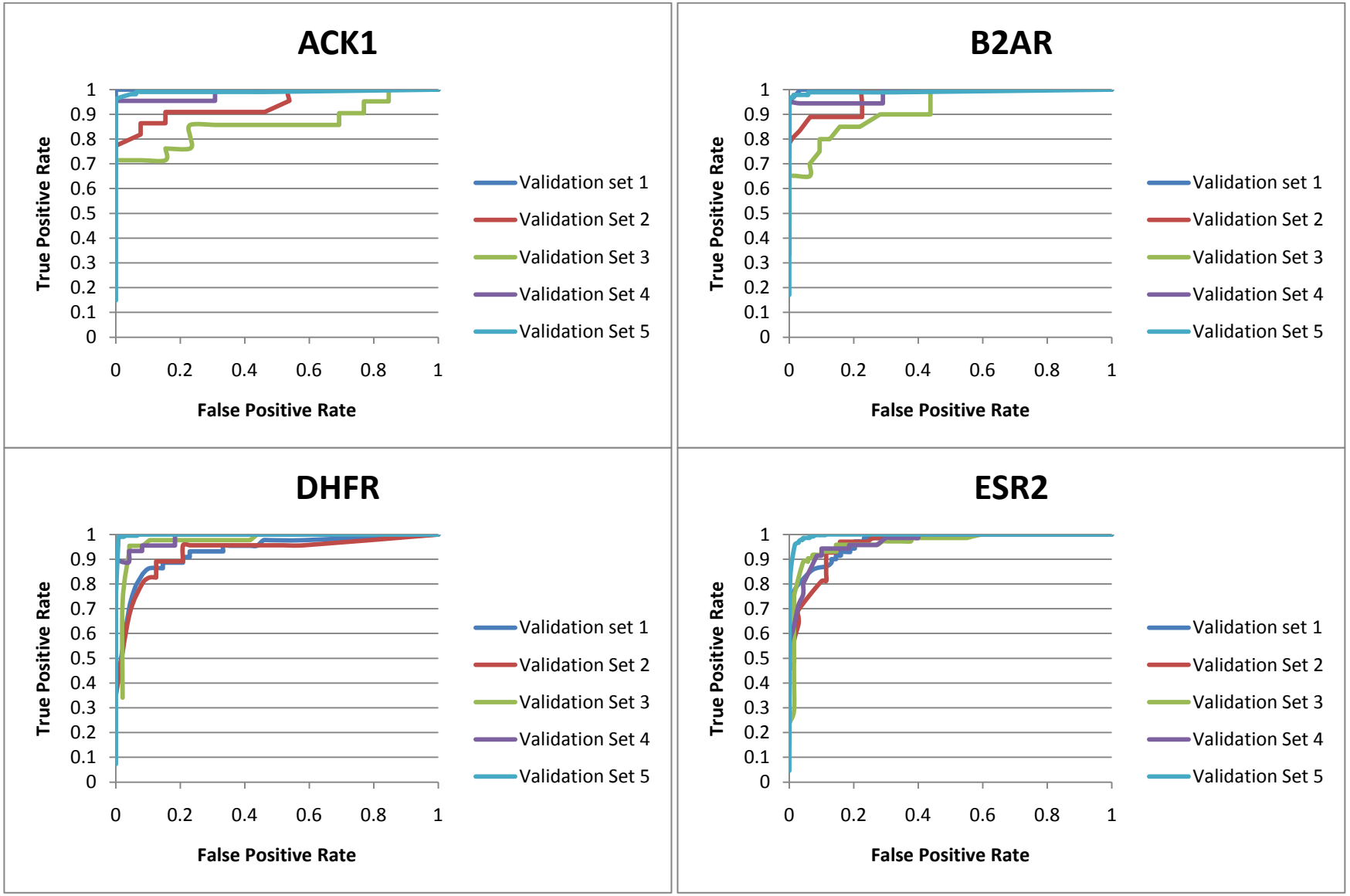


Figure 34. *Example ROC curves resulting from prediction of compounds with known activity using QSAR models.*

Figure 35 demonstrates that in there is no clear pattern in whether similarity searching or docking will provide better ranking of a screening library. While docking certainly fails in some cases, there are also examples of similarity searching proving mediocre in recall of actives. When examining the capacity of the two methods to properly classify only compounds with known activities, it is apparent (see Figure 36) that often neither method proves successful.

Figure 37 and Figure 38 display the ROC curves for similarity searching and QSAR modeling using each of the modeling sets. While similarity searching appears to be much better at extracting active compounds from a large chemical library, QSAR does a superior job of separating the known actives from the known inactives. This sensitivity of QSAR to fine differences in chemical structure while similarity searching provides coarse separation of actives from a large set of putative inactives speaks to the complementary of the two methods in virtual screening.

While different methods use different sets of knowledge to rank chemical libraries, all methods are united in that their goal is enrichment of known actives in a subset of a database. Therefore, it is reasonable to compare all methods and their respective knowledge bases on the criteria of enrichment. For each target, enrichments were calculated at 0.5%, 1%, 5%, and 10% of the database. An example of the resulting enrichment comparison is contained in Figure 39. Enrichment comparisons for additional targets are available in Appendix VII.

Based on the generated figures, enrichment appears to usually increase as ligand information is added to the model system. In terms of raw enrichment of active compounds, similarity searching appears to be the best method for utilizing this information.

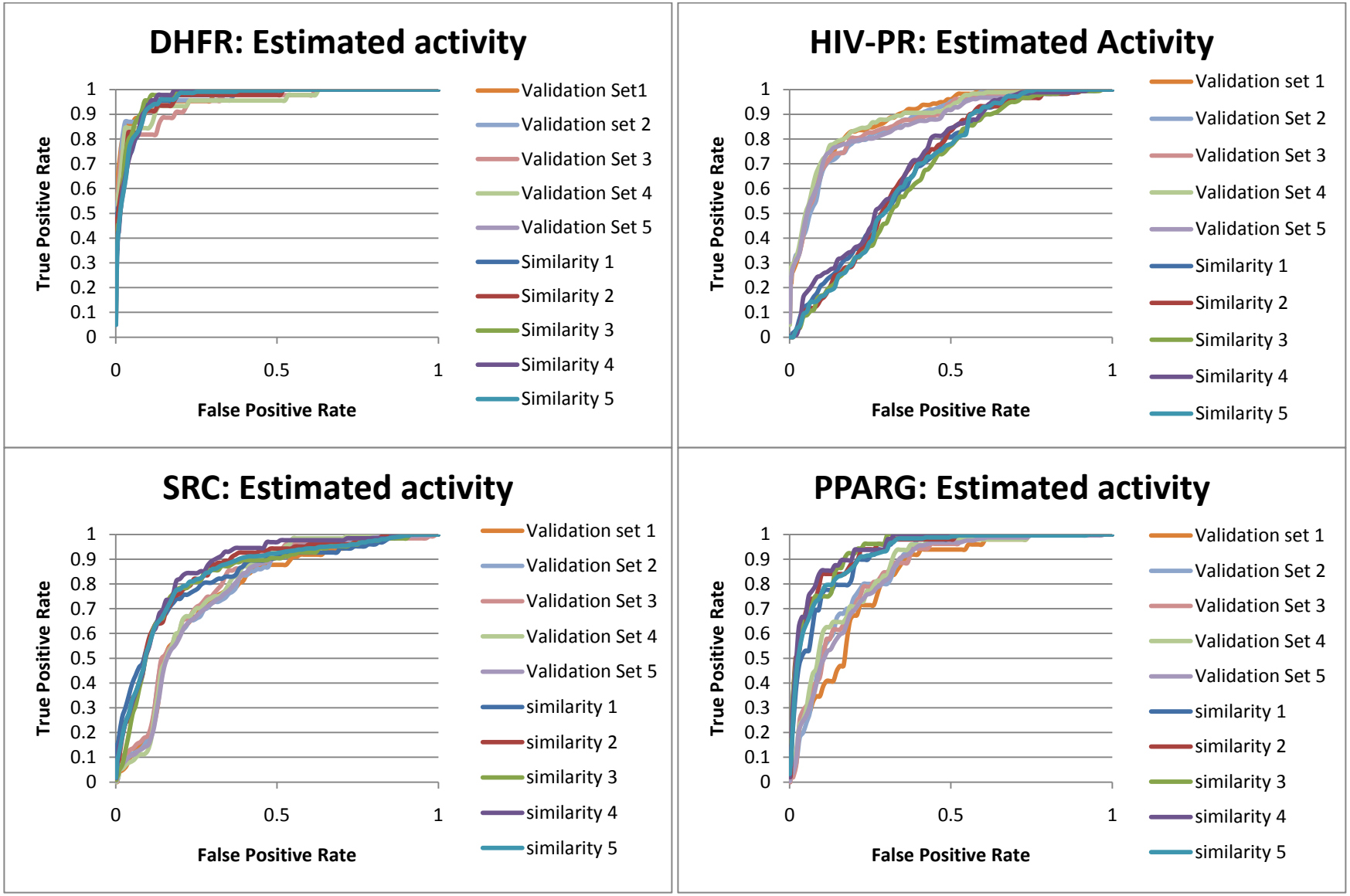


Figure 35. *Comparison of ROC curves between docking and similarity searching on the full screening library.*

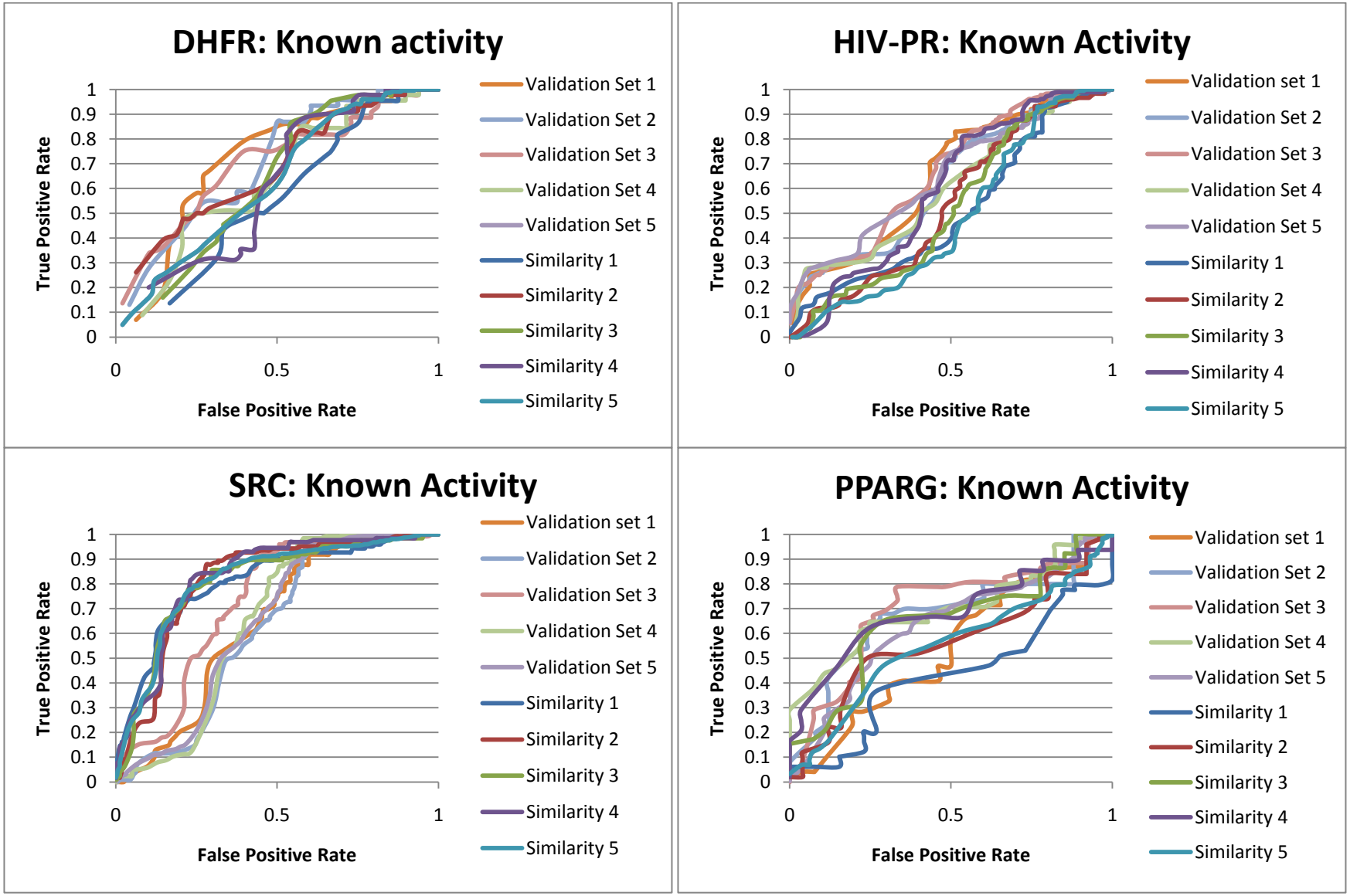


Figure 36. *Comparison of ROC curves between docking and similarity searching on compounds with known activity.*

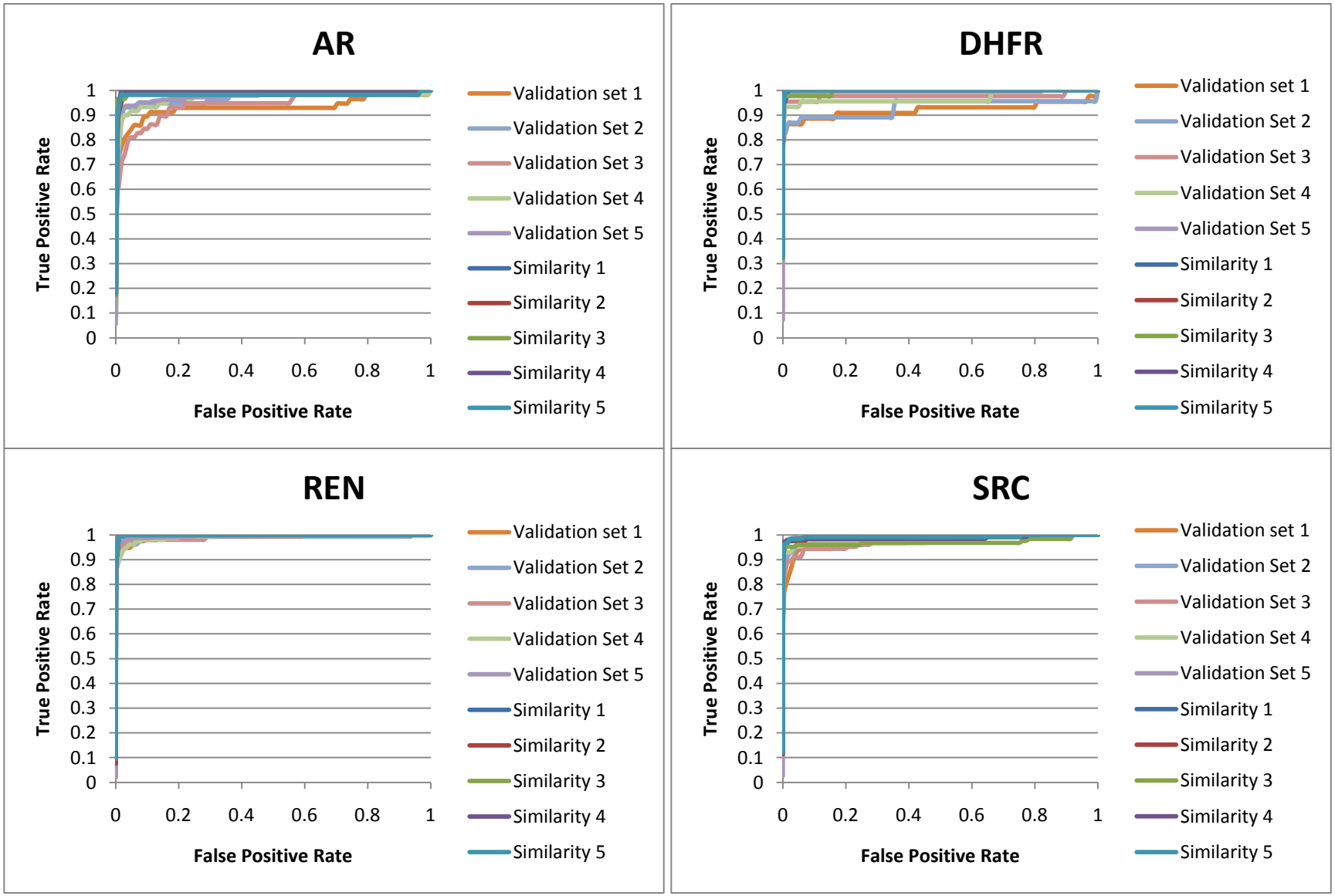


Figure 37. *Comparison of ROC Curves from QSAR modeling and similarity searching using full modeling sets on screening library.*

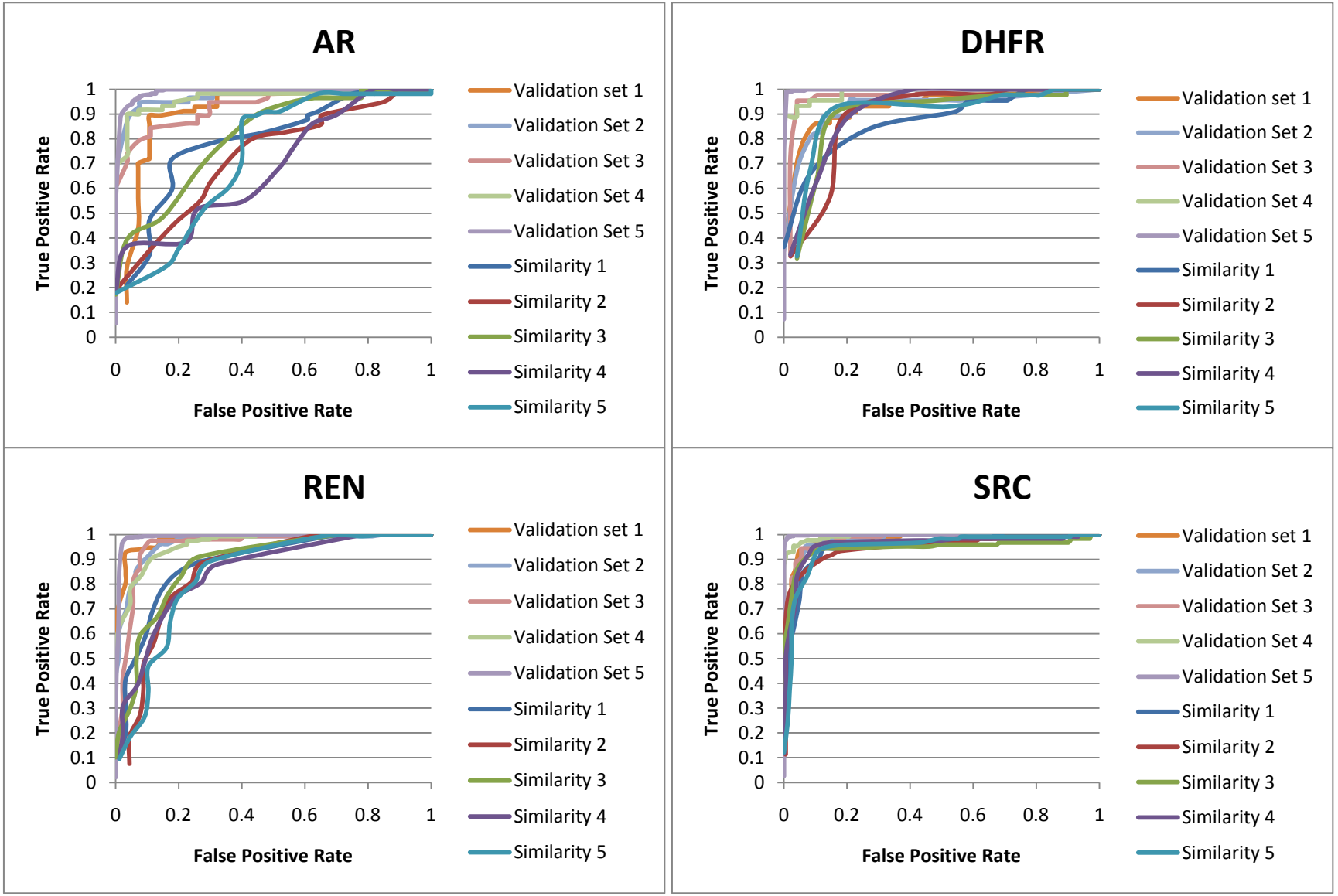


Figure 38. *Comparison of ROC Curves from QSAR modeling and similarity searching using full modeling sets on compounds with known activity.*

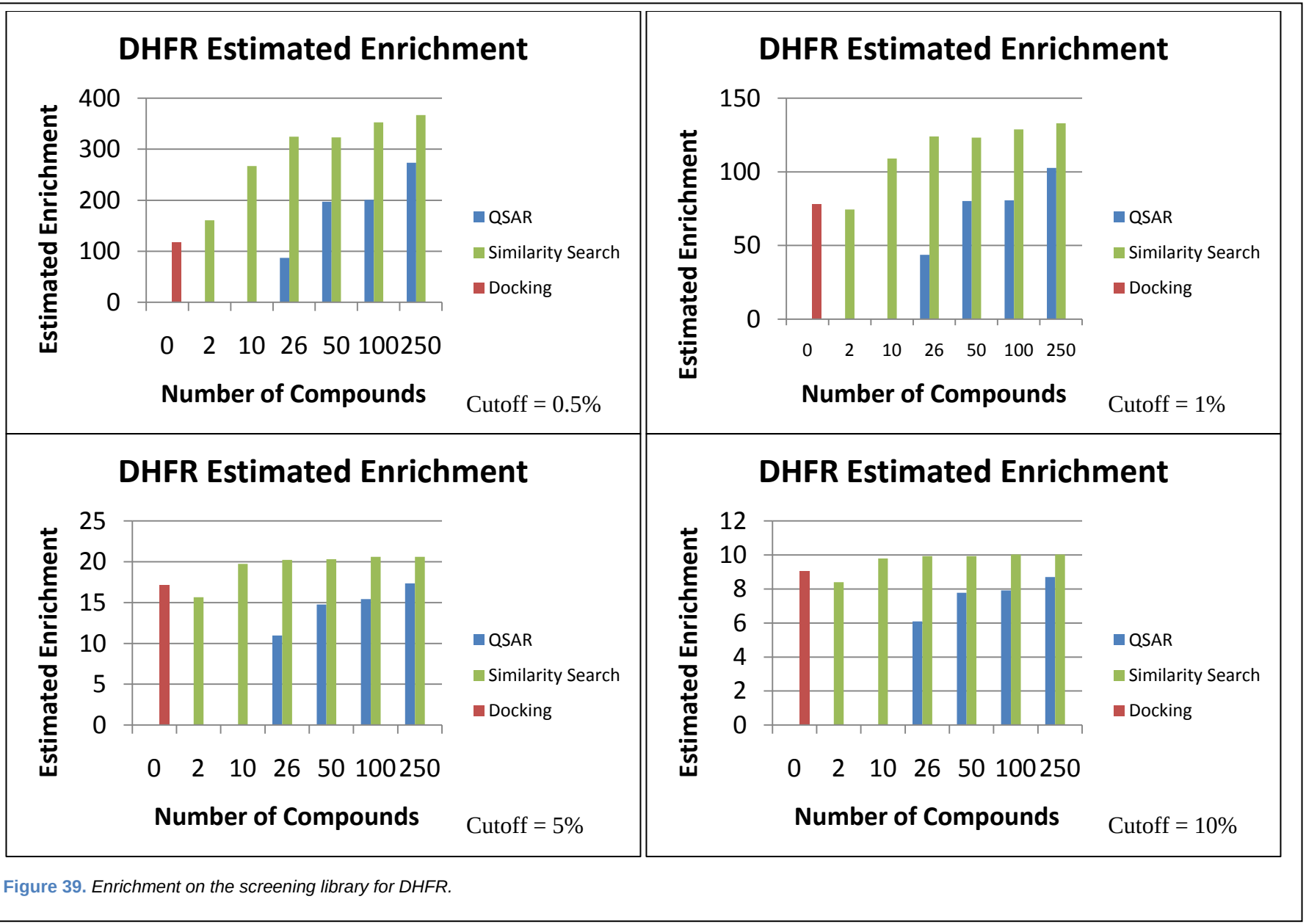


Figure 39. *Enrichment on the screening library for DHFR.*

3.3.7. CCR or Enrichment for Model Characterization

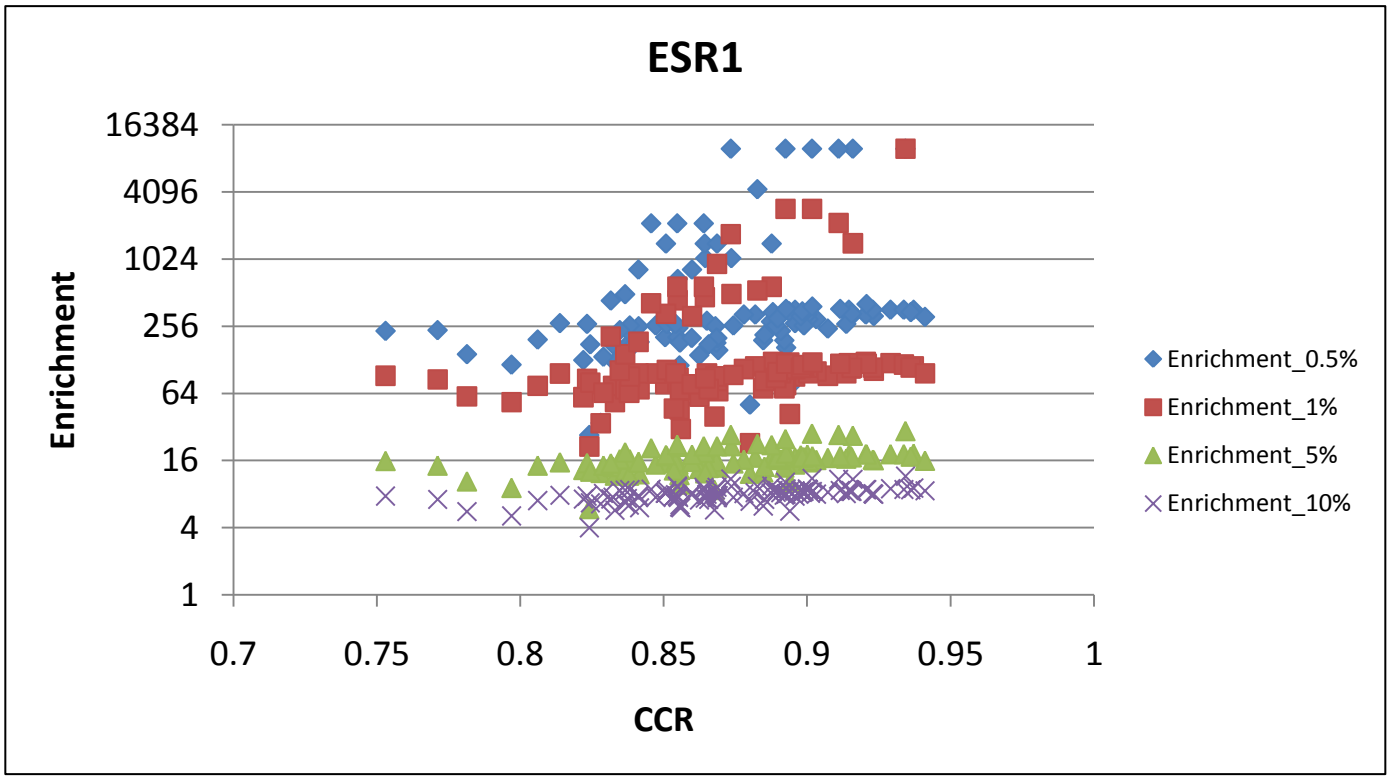
Questions often arise regarding how to select QSAR models that will yield superior virtual screening results. While our lab primarily relies on CCR as a measurement of a model's usefulness, for the goal of virtual screening one would expect that metrics more commonly applied within the fields such as enrichment would provide a better assessment of a model's capabilities. While the experimental design of this study was specifically focused on the evaluation of the effect of modeling set size on QSAR in relation to both docking and similarity searching, the abundance of derived data allows us to examine the relationship between CCR and enrichment.

Figure 40 contains a scatter plot of CCR vs. enrichment for two selected targets. Additional figures of this type are contained in Appendix VIII. While the relationship between CCR and enrichment does appear to have slight correlation, that correlation appears to be inconsistent and in many cases weak.

3.4. Conclusions and Future Directions

During the generation and assessment of a benchmark dataset for assessment of virtual screening techniques, the following goals have been achieved

1. Extraction and curation of ligand datasets for 22 targets from three (one public and two commercial) bioactivity databases (Section 3.3.2)
2. Docking of a library of nearly 17,000 compounds to 22 different protein targets (Section 3.3.3
3. Similarity searching using nearly 4000 different probe sets (Section 3.3.4)



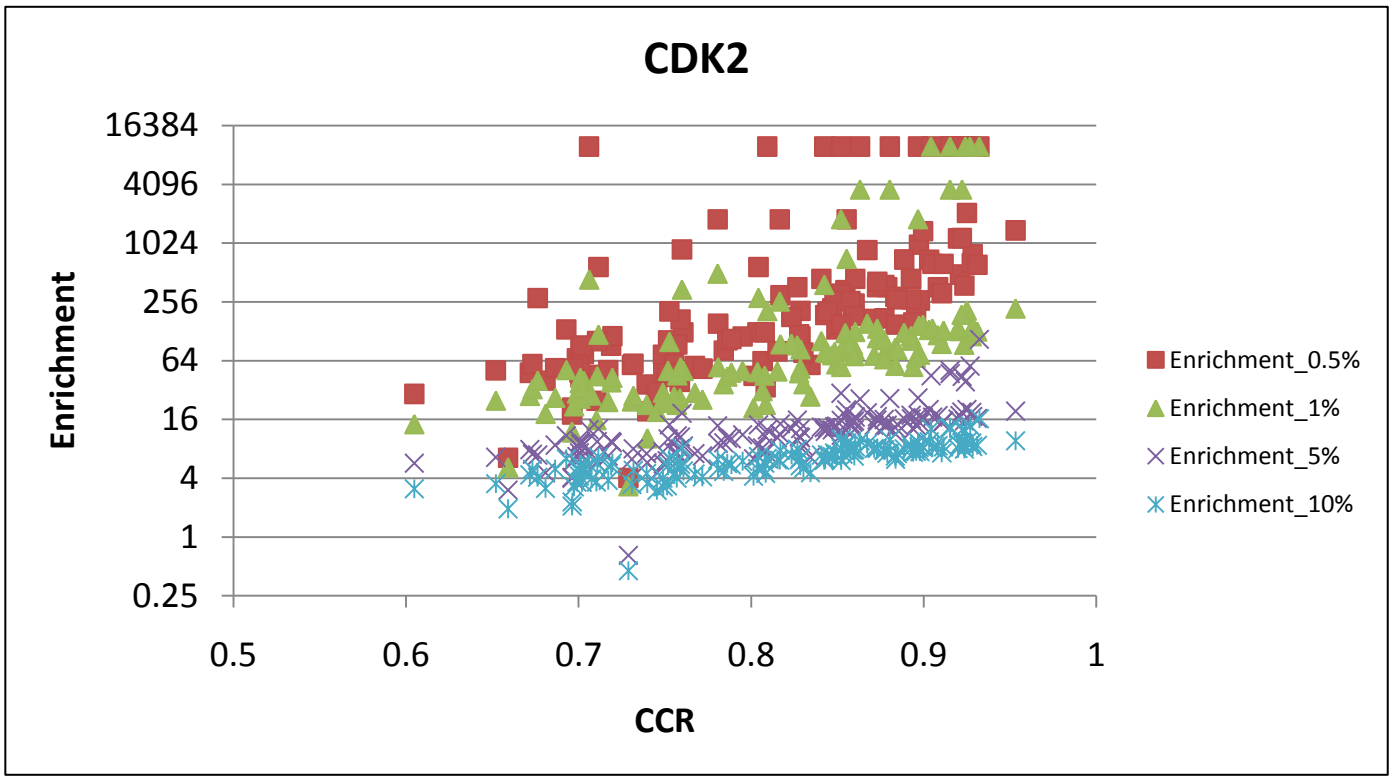


Figure 40. *The lack of correlation between CCR and enrichment.*

4. Generation of more than 2500 ensemble QSAR models including 22 externally validated predictors of biological activity (Section 3.3.5)

Using the data generated from assessment of the benchmark datasets, we have determined that the selected docking and similarity searching protocols perform very poorly in separating tested actives from tested inactives. We have validated the importance of being able to classify a target in the family based scoring scheme promoted by eHiTS. We have determined that in terms of ability to identify hits from a chemical library, similarity searching and docking perform nearly equivalently.

While assessing the selected QSAR method, it was apparent that QSAR alone is poor in comparison to similarity searching at enriching a large chemical library; however, QSAR models significantly outperform docking and similarity searching in their ability to separate the known actives from the known inactives. The inability of QSAR models to effectively separate the most interesting compounds from a chemical database is easily rectified with the use of a global applicability domain (an assessment of whether a compound in the chemical library is similar enough to members of modeling set to make a prediction). The results of this study show that the use of a global applicability domain as is often done when performing a virtual screen with QSAR is vital to achieve optimal selection of hypothetical binders.

The performance of QSAR models in classification and virtual screening as measured using CCR and enrichment respectively often correlate, but are not equivalent. Optimizing enrichment rather than CCR could generate models better suited to virtual screening of large chemical libraries.

Clearly this study is limited in the number of methods applied to analyze these datasets. To gain a better understanding of the capability of cheminformatics in the task of virtual screening, a larger study involving more cheminformatics specialists must be initiated. By encouraging a collaborative study, a better assessment of cheminformatics tools will be obtained since experts will use the tools with which they are most familiar and comfortable. This will lead to comparison of tools when applied in the best manner.

The metrics of virtual screening success should be improved. Rather than assessing the number of compounds returned, a better measure of success is the number of new chemical classes identified. Clustering the dataseta then manually defining the boundaries between the different classes of actives could achieve this goal. Then the recall of active classes could be measured when virtually screening the library.

The above consideration highlights a limitation in our strategy for determining the effects of knowledge base size on similarity searching and QSAR. In realistic applications, the knowledge base often contains only a subset of the known active classes for a target whereas with random sampling no attempt to control the diversity of modeling set was made. It is expected that if a compound's target class were considered when selecting compounds for modeling sets, a more distinct drop would be seen in predictive power as modeling set size was decreased. This hypothesis surely bears testing as the usefulness of ligand-based methods should be assessed in the most realistic manner so the method comparison can inform application scientists.

Finally, while sets were generated from both WOMBAT and MDDR, they have not been utilized in benchmarking screening tools. The commercial restrictions on the extracted sets are

surely a strike against them, but studies should be completed verifying that virtual screening on these sets and the ChEMBL set are similar.

Chapter 4: Chemical Sensitivity of Cancer Cell Lines

4.1. Introduction

Over the past decade there has been increased interest in shifting the treatment of cancer from a tissue or organ specific approach to a more personalized approach[43]. Personalized medicine relies on the measurement of biomarkers that indicate how an individual will respond to a particular treatment. However, a comprehensive set of biomarkers is still unavailable. This is disappointing as there has been a decided increase in our capacity for genetic screening.

Biomarkers can be defined using a variety of techniques in the fields of genomics, proteomics, or metabolomics. Herein, we focus on the use of gene expression profiles to predict the resistance or sensitivity of a cell line to a chemotherapeutic or several chemotherapeutics. The NCI-60 dataset provides an excellent resource to mine to identify gene expression biomarkers as it provides a measure of drug-induced cytotoxicity for a large number of chemicals in a panel of 60 cell lines. These cell lines also have their gene expressions profiled.

While many have mined this data to identify biomarkers of resistance, most works focus on analysis of single compounds at a time[101]. At most a small set of compounds are examined[102]. This lack of comprehensive analysis of the NCI-60 dataset likely obscures markers that are relevant to large set of compounds (i.e. multidrug resistance genes). Therefore, we have completed a study of the entirety of the NCI-60 dataset looking to identify both multidrug resistance biomarkers and drug specific resistance biomarkers.

4.2. Materials and Methods

4.2.1. NCI-60 dataset

The In Vitro Cell Line Screening Project (IVCLSP) has been fully operational since April of 1990. This project, tasked with the direct support of the Development Therapeutics Program (DTP) anticancer drug discovery effort, is designed to screen up to 3,000 compounds per year for growth inhibition of 60 different human tumor cell lines representing a variety of tissue types. Portions of the results of this screening are made available to the public. Our data was taken from the following locations: GI_{50} values were taken from the archive file available from http://dtp.nci.nih.gov/docs/cancer/cancer_data.html, chemical data was drawn from both the structural file contained within the bioactivity data archive file and the 2D structural file available at http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html, and Affymetrix HG-U133(A-B) raw data[103] were extracted with use of Cellminer[104].

4.2.2. Dataset Curation

Being that the dataset contains chemical, screening, and gene expression data, the first step of curation was to ensure consistency of representatives across data types. When examining the chemical and screening data, we determined that 585 identifiers in the screening data had no stored chemical data. Of the 60 cell lines commonly screened in the IVCLSP, only 59 had recorded gene expression data. The 585 identifiers that did not have chemical data and the cell line without gene expression data were eliminated from further analysis.

The chemical structures for the remaining 47039 compounds were then standardized and compared using Pipeline Pilot to determine if duplicates were present. 532 duplicate structures

were identified linked to 1114 nsc_ids. The screening results for duplicate structures were treated as having been submitted with the same identifier (see process of curation *infra*).

The screening data provided via download often contained multiple pGI_{50} values for the same identifier-cell line pair. Additionally, the data was occasionally reported in more than one type of unit. To deal with this multiplicity of values, unless the pGI_{50} was equal to maximum concentration tested, we weighted the pGI_{50} measurements (in M units) by the number of tests from which that measurement was obtained and then averaged them. When the reported pGI_{50} was equal to the maximum concentration tested, it was only included in the averaging if it was less than minimum pGI_{50} reported for other instances of the identifier-cell line pair. While inclusion of any data where the reported pGI_{50} is equal to the maximum concentration tested may be considered questionable, elimination of all such instances significantly reduces the amount of available data and obscures the chemical sensitivity trends across cell lines. All data not reported in M units was ignored.

After coalescing duplicative pGI_{50} values, only the 4614 compounds for which all 59 cell lines had pGI_{50} values were retained. Additionally, compounds that did not have a difference of greater than one order of magnitude between their most active GI_{50} and their least active GI_{50} were removed leaving 3555 compounds.

In order to apply QSAR techniques, additional curation was required prior to generation of chemical descriptors. As the descriptor techniques being employed were insensitive to chirality, all chirality was removed using Pipeline Pilot prior to QSAR modeling and duplicates were again analyzed leaving 3524 compounds. Additionally, the Dragon descriptor generation software was unable to process chemicals that contained certain atoms eliminating another 11 compounds.

4.2.3. Computation Study Design

The simultaneous analysis of chemical, bioactivity, and gene expression data is quite difficult. Therefore, we decided to progressively segment the data to analyze individual portions at a time. In short, we hypothesize that the GI_{50} values contained within our dataset can be estimated by adding a wholly chemical component and a wholly cellular component to an interaction component as described in Equation 5.

(5)

This description of activity values allows us to eliminate the wholly chemical component by normalizing each compound's GI_{50} values using the average and standard deviation in the activity of that compound across the cell lines. This normalized GI_{50} value becomes our measure of a cell's resistance or sensitivity to the drug (see Equation 6).

(6)

Our separation of the data into parts leads to a cellular resistance that while certainly a function of both cellular composition and chemical structure can be analyzed as a multidrug resistance (an estimate of the hardiness of a cell when treated by a spectrum of chemicals) and specific cellular resistance (the specific interaction between a cell and chemical that is separate from the mechanisms for generic resistance).

4.2.4. Multidrug Resistance

Multidrug resistance can be described as the hardiness of a cell line against a broad spectrum of chemical stimuli. The resistance of a cell line to a chemical probe is only apparent in relation to the effects of the same stimulus on other cell lines. This being the case, the pGI_{50} values across the cell lines for each compound were centered and scaled using the mean and standard

deviation for that compound as an estimate of each cell line's resistance to that compound. The resulting matrix of 59 cell lines with resistance estimates for 3555 compounds was then subjected to Singular Value Decomposition (SVD)[105] to select a single vector that represented the general resistance (or multidrug resistance) of the cell lines.

4.2.5. Gene Identification

After definition of a response variable (either generic cellular resistance as above or a particular compound's GI_{50} spectrum as below), selection of significant genes was carried out using Significance Analysis of Microarrays (SAM)[106]. Specifically, we applied the SAMR package available from http://www-stat.stanford.edu/~tibs/SAM/ using 1000 permutations. The delta parameter was altered to obtain an appropriate level of significance based on each case.

4.2.6. Pathway Analysis

Analysis of the networks and pathways populated and formed by the identified genes was accomplished using Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com). In our case, a core analysis was conducted using a maximum network size of 35 members. The core analysis includes network analysis, functional analysis, and canonical pathway analysis.

Network analysis was carried out by first mapping each identifier to its corresponding object in Ingenuity's Knowledge Base. These molecules, called Network Eligible molecules, were overlaid onto a global molecular network developed from information contained in Ingenuity's Knowledge Base. Networks of Network Eligible Molecules were then algorithmically generated based on their connectivity.

The Functional Analysis identified the biological functions and/or diseases that were most significant to the set of genes. The identified markers associated with biological functions and/or

diseases in Ingenuity's Knowledge Base were considered for the analysis. Right-tailed Fisher's exact test was used to calculate a p-value determining the probability that each biological function and/or disease assigned to that data set is due to chance alone.

Canonical pathways analysis identified the pathways from the Ingenuity Pathways Analysis library of canonical pathways that were most represented by the identified genes. The significance of the association between the data set and the canonical pathway was measured in 2 ways: 1) A ratio of the number of molecules from the data set that map to the pathway divided by the total number of molecules that map to the canonical pathway is displayed. 2) Fisher's exact test was used to calculate the probability that the association between the genes in the dataset and the canonical pathway is explained by chance alone.

4.2.7. QSAR modeling of expected/aberrant behavior

The apparent of nature of compounds belonging to one of two classes based on the hierarchical clustering of correlation value (see Section 4.3.3) was used as a response variable to build a QSAR model. Compounds were loaded into Chembench and standardized. Five-fold external validation was used to ensure model robustness. The random forest procedure implemented in Chembench was applied to Dragon descriptors[100] of chemical structure with the following selections: range scaling of descriptors and elimination of descriptors with perfect correlation, 50 random divisions of training/test set containing between 20% and 30% of the dataset, and 50 trees generated for each split using 50 descriptors. Further discussion of the random forest procedure implemented in Chembench is contained in section 5.2.2.

4.2.8. Nearest Neighbor Analysis

The aberrant compounds identified via hierarchical clustering were treated to individual SAM analysis to identify the genes most significantly related to their pGI_{50} profile. For each compound, the nearest neighbor compound in FCFP4 space was identified and the overlap of significant genes between neighbors was assessed.

4.3. Results and Discussion

4.3.1. Gene Expression Markers of Multidrug Resistance

To visualize the amount of multidrug resistance evident within the cells contained within the NCI-60 panel, the centered and scaled pGI_{50} values indicative of the level of resistance were separated into resistant, sensitive, and neutral groupings where any normalized $pGI_{50} < -1$ was considered resistant, any normalized $pGI_{50} > 1$ was considered sensitive, and the remainder were considered neutral. Figure 41 contains a bar graph of the number of chemicals to which a cell line was sensitive or resistant. These results indicate that not only are some cell lines resistant to multiple drugs, but some cell lines are sensitive to multiple drugs.

With the knowledge that a large portion of the measured cellular resistance and sensitivity appears to be caused by multidrug effects, we quantized the multidrug resistance of a cell using SVD projection of the normalized pGI_{50} matrix into a single vector. The application of SAM to this quantized multidrug resistance identified 361 genes (121 linked to sensitivity and 240 linked to resistance) with less than a 0.1% 90^{th} percentile FDR. A listing of these hypothetical markers of multidrug resistance is contained in Appendix IX.

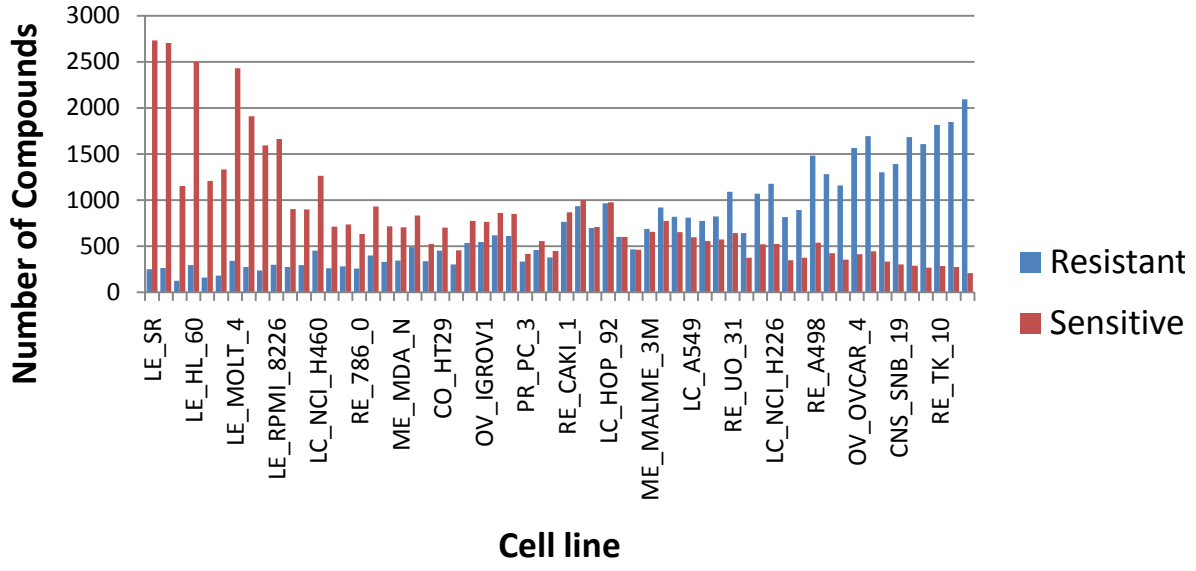


Figure 41. *Multidrug resistance profile of cell lines.*

4.3.2. Pathways of Multidrug Resistance

Following identification with SAM of markers of multidrug resistance, the 361 markers were subjected to Ingenuity Pathway Analysis. When loaded, a total of 11 probe set ids failed to map to genes. The resulting gene list was subjected to IPA core analysis. This analysis resulted in the identification of several protein networks that have a high degree of connection amongst the identified markers. One such network is shown in Figure 42. Additional networks and table of the networks and their linked functions are displayed in Appendix X.

When examining the network, it is interesting to note that a large number of the markers identified have previously been linked to cancer. c-Myc (MYC) is a transcription factor that has been identified in several cases to be linked to cancer and which is currently being investigated as a cancer target. c-Myc has been previously linked to the sensitization of melanoma cells to radiotherapy[107]. DNA Fragmentation Factor Beta (DFFB) is a protein that when activated initiates DNA fragmentation and chromosome condensation[108]. Lowered DFFB expression has

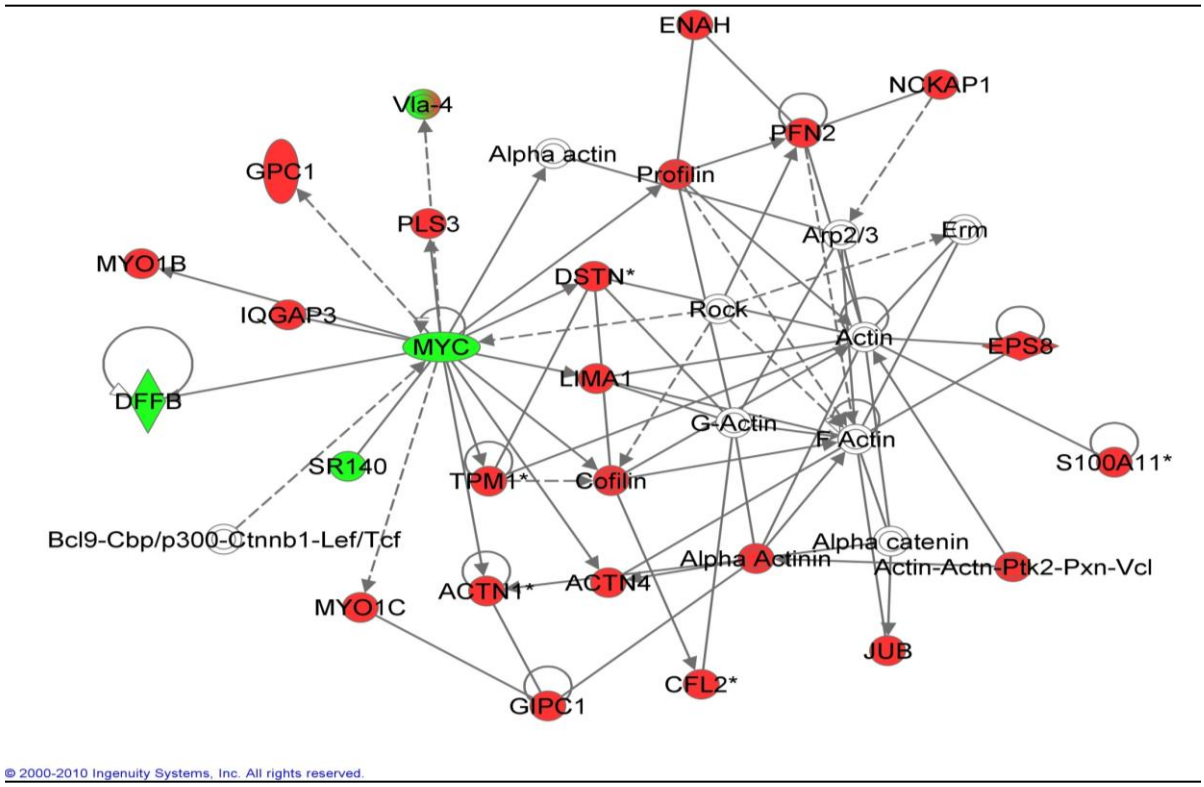


Figure 42. *Network of identified gene expression markers of multidrug resistance. (Red nodes are resistance genes and Green nodes are Sensitizing genes.*

been linked with Oligodendrogliomas.[109] Increased expression of alpha-actinins (including ACTN1 and ACTN4) has been identified in hepatocellular carcinomas.[110] Profilin is an actin binding protein that has previously been shown to decrease cancer cell motility[111] and suppress tumors.[112] Ajuba (JUB) is a protein that is known to be essential to enter into mitosis.[113] It has been found to interact with protein 14-3-3σ, a protein commonly silenced in cancers[114]. These are just a subset of links that can be made between this network (which contains a large number of motility effecting genes) and cancer.

Additionally, IPA detected both the canonical pathways and cellular functions that were highly represented by the hypothetic markers. These pathways and function are documented in Figure 43. Several of these pathways and functions are linked to cancer. With respect to pathways, the similarity of leukocyte extravasation to tumor cell extravasation has been

previously noted and reviewed.[115] Agrin interaction and neuromuscular signaling have been shown to be affected by the mouse tumor suppressor protein Adenomatous Polyposis Coli (APC).[116] Integrin signaling is known to be required for development and metastasis of cancer.[117] With respect to the functions, cellular movement is needed for cancer spread, and cellular assembly and organization is required for any proliferating cell line.

The high degree of linkage between the identified markers, their pathways, and their

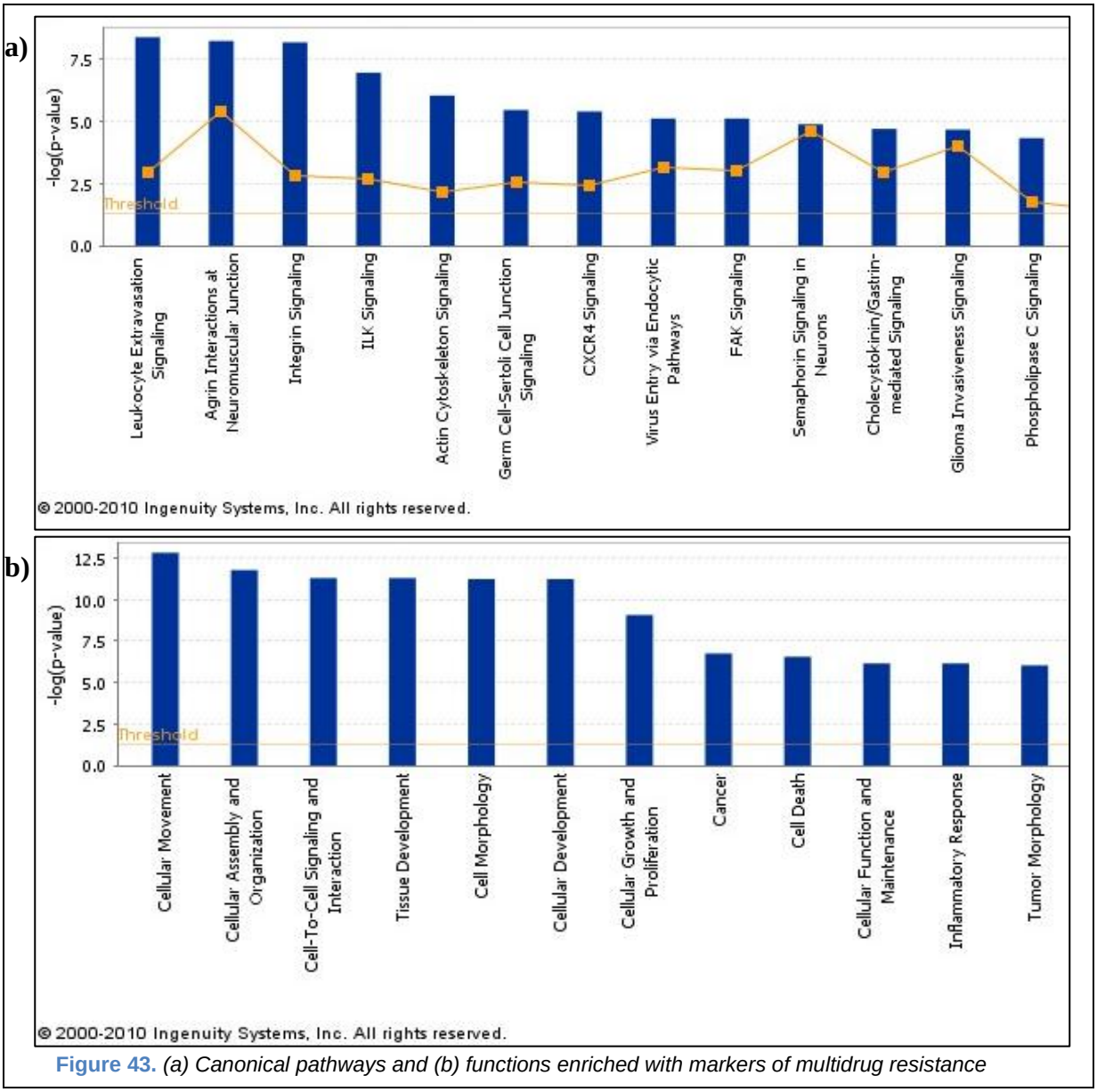


Figure 43. *(a) Canonical pathways and (b) functions enriched with markers of multidrug resistance*

functions and cancer lend credence to the hypothesis that the identified genes are in fact related to multidrug resistance.

4.3.3. Correlation of GI_{50}s and Marker Expression

Based on our separation of resistance into a multidrug component and a drug specific component, we expected that several chemicals contained within the dataset would behave differently than projected by multidrug markers. To detect these compounds, we correlated the gene expression for the selected multidrug resistance and sensitizing genes to the pGI_{50} values for compounds across the 59 cell lines. Using these correlation values, the compounds were clustered using Partek Genomics Suite's[118] hierarchical clustering with Euclidean distance and average linkage. (See Figure 44a.)

Chemicals are clearly segregated into two clusters: one comprised of 2933 compounds whose pGI_{50} values correlate as expected with the expression of selected generic resistance and sensitivity genes and one comprised of 622 compounds whose pGI_{50} values in general do not correlate as expected. While these two classes were apparent when clustering was done on the correlation values, they were not evident when the chemicals were clustered using the normalized pGI_{50} values (Figure 44b) or two sets of chemical descriptors (Figure 44c,d).

While the two noted classes are the most glaring result of the clustering, the heatmap also indicated that there is still variation within the 2933 compounds that generally have expected behavior. These variations could also be related to gene expression effects. Unfortunately, these deviations were not addressed in this study.

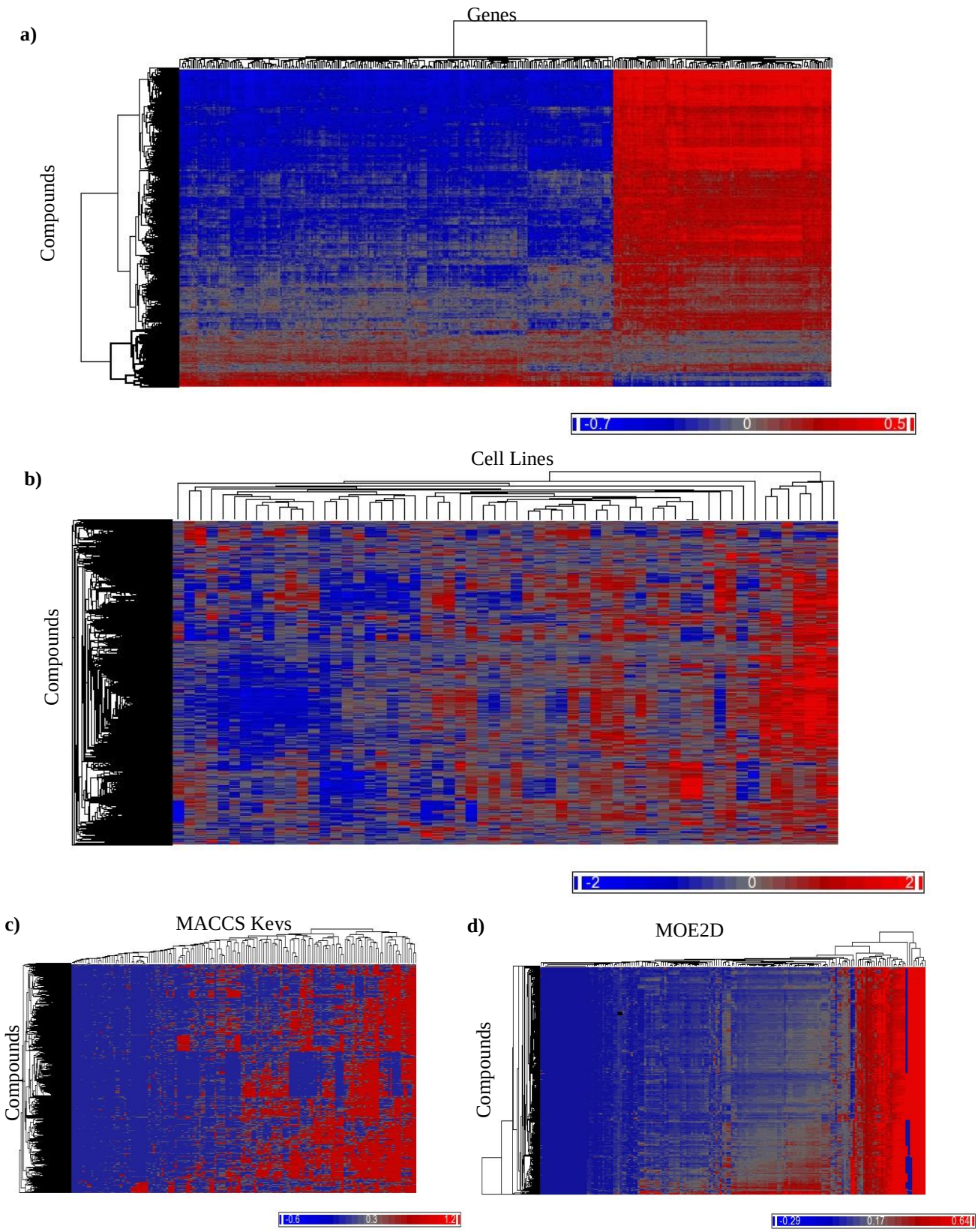


Figure 44. *Clustering results of the NCI-60 compounds based on (a) correlation of pGI_{50} values and gene expression of multidrug resistance markers, (b) Normalized pGI_{50} values, (c) MACCS keys, and (d) MOE2D descriptors*

4.3.4. Prediction of Aberrant Behavior

While it can be expected that the resistance of a cell line to a chemical can usually be understood by an estimate of a cell line's generic hardiness, it is not surprising that some compounds may have specific interactions that allow them to exhibit cellular growth inhibition profiles that are uncommon. As global similarity in chemical descriptor spaces appeared to be insufficient for predicting which compounds would elicit abnormal growth inhibition profiles, we built a QSAR to aid in this task.

The imbalance in the dataset provided a significant complication to the modeling. To address this imbalance, three methods for down-sampling the overrepresented class patterned after those used in a recent unpublished study of anti-malarial compounds were applied: random selection of five folds of the overrepresented class, selection of compounds from the overrepresented class most similar to underrepresented class, and selection of two neighbors from the overrepresented class for each member of the underrepresented class. These down sampling techniques were applied after five-fold extraction of validation sets to ensure accuracies would be comparable.

Since we were applying Random Forests, our QSAR modeling was consensus in nature. This being the case, each compound was predicted with a numeric value between 0.0 and 1.0 with 0.0 representing high consensus that a compound was normal, 1.0 representing high consensus that a compound was aberrant, and 0.5 indicating that there was no consensus amongst models. While typically a threshold at 0.5 is used to separate active and inactive predictions, there are times where compounds with numeric values near 0.5 are thrown out. To define the level of agreement required, an agreement threshold was defined such that if the agreement threshold were 0.1, only compounds with numeric values above 0.6 would be considered

aberrant while compounds with numeric value below 0.4 would be considered normal. Compounds with predicted numeric value between 0.4 and 0.6 are eliminated.

Figure 45 presents the prediction accuracy in terms of CCR and the coverage when predicting the validation sets as a function of the agreement threshold. As expected the prediction accuracy increases as the compounds with lower levels of consensus are removed. No method of down-sampling appears to be definitively superior to another. Additionally, only a small number of the compounds (roughly 10%) can be accurately identified as being aberrant using only information from their chemical structure. This is understandable considering there are likely a large number of ways for a compound to elicit a resistance profile that cannot be predicted based on the selected markers of multidrug resistance. As such, it is likely that the 622

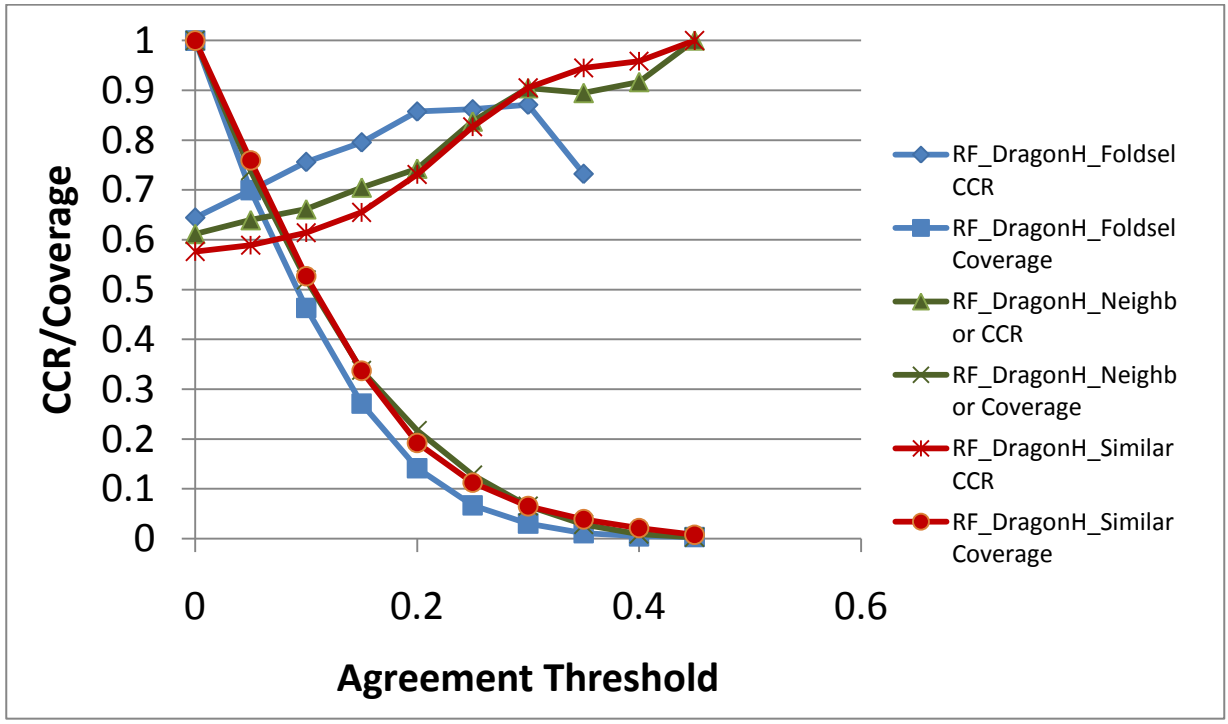


Figure 45. *Validation set prediction accuracy and coverage for QSAR prediction of aberrant compounds.*

compounds identified as having aberrant behavior in this study only sparsely populate the pathways that modulate resistance in a chemical specific manner.

4.3.5. Genetic Markers of Aberrant Compounds

As the causes of compound specific resistance remain unknown, SAM analysis was applied to the resistance profile of each compound in the aberrant set. Genes with less than a 1% median FDR were identified as potential biomarkers for a chemical's specific resistance profile. While no genes were found for nearly half of the chemicals at this cutoff, 58000 gene-compound pairs were identified. The amount of overlap of genes between chemicals was measured. Figure 46 shows the distribution of overlap in gene expression markers for chemicals.

Figure 46 clearly shows that very little overlap occurs between potential biomarkers of specific chemical resistance within the aberrant set and this plot omits the nearly 130 thousand

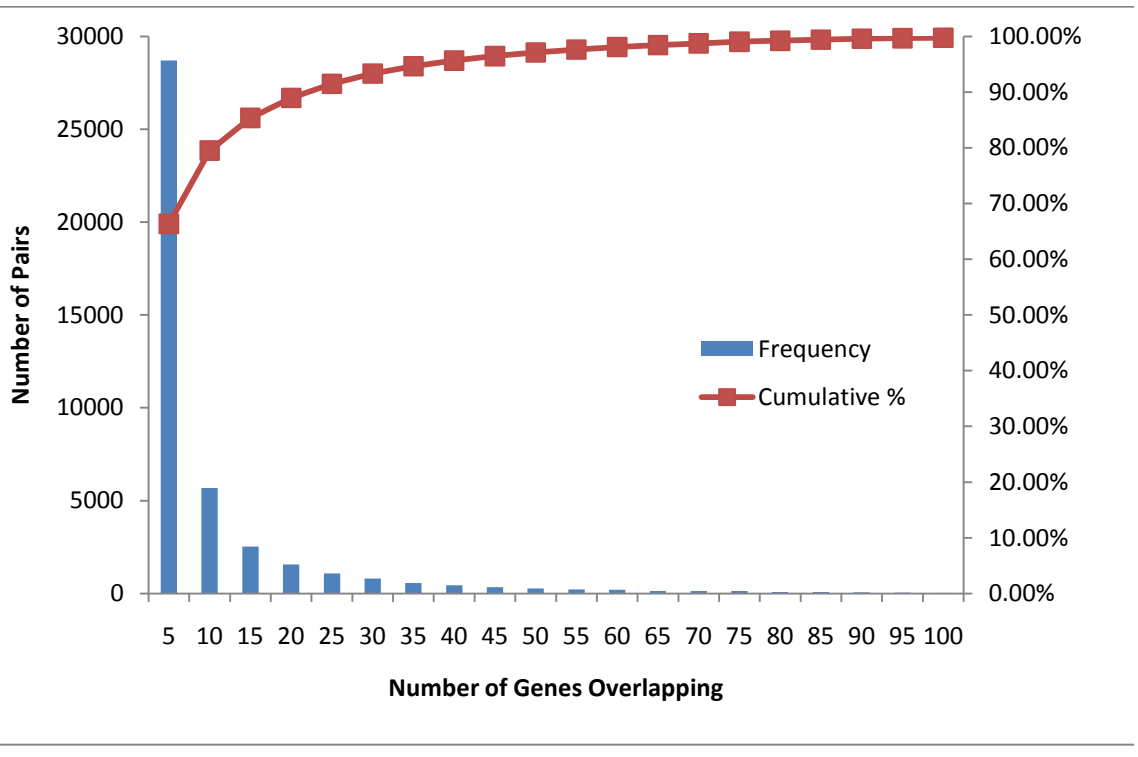


Figure 46. *Infrequency of overlap in resistance genes of aberrant compounds.*

pairs for which there were zero gene overlaps. To determine if chemical similarity would be capable of predicting which chemicals would overlap in the resistance genes, we plotted for the three nearest neighbors of each compound the Tanimoto similarity of compound pairs in FCFP-4 space against the count of their overlapping resistance genes. This plot (displayed in Figure 47) clearly shows that the correlation between similarity and overlap is very weak.

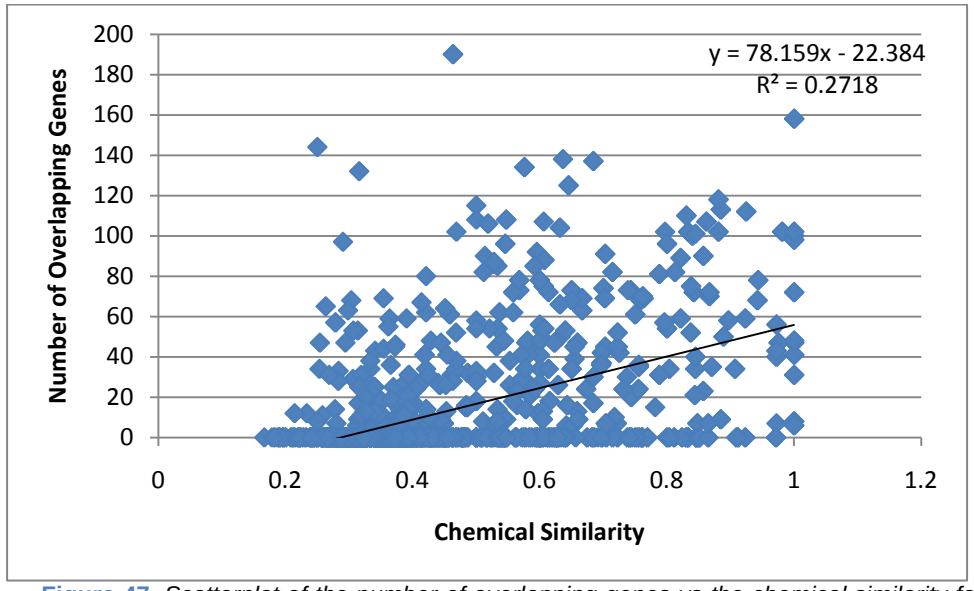


Figure 47. *Scatterplot of the number of overlapping genes vs the chemical similarity for pairs of neighbor compounds*

4.4. Conclusions and Future Directions

During analysis of the NCI-60 dataset to identify gene expression markers of resistance, multiple tasks were completed.

1. An unbiased method was defined for quantifying the multidrug resistance potential of a cell line. (Section 4.3.1)
2. SAM was used to identify 361 genes whose expression appears linked to multidrug resistance (Section 4.3.1)

3. These potential biomarkers were analyzed for their biological significance and connections to one another thereby implicating several functions in the conference of resistance. (Section 4.3.2)
4. Compounds having drastically different responses than expected based on expression of multidrug resistance markers were identified. (Section 4.3.3)
5. QSAR analysis was completed indicating that only a small portion of aberrantly behaving compounds can be predicted based on structure. (Section 4.3.4)
6. 58000 chemical-genes resistances were hypothesized. (Section 4.3.5)

These results provide the basis for a great deal of experimental validation. In addition to experimental validation of the hypothetical genes that were identified in this study, there are computational studies that could be carried out to refine the selection of markers. In particular, the treatment of genes as individual entities during identification of markers ignored knowledge of how the genes work as a network. While the identified genes appear to be highly connected in networks, by quantifying the differences in these networks between cells rather than the expression of individual genes, greater insight may be possible. Instead of selecting genes that appear linked to resistance phenomenon and building the networks with these blocks, it would be more logical to directly link alterations in networks to resistance. Difficulties in appropriately quantifying the fluctuations in a network prevented us from carrying out this study.

Ability to predict compounds for which there would be specific resistance effects eluded us for a large portion of our dataset. This appears to be a limitation in either the dataset or in the methods applied to it. It may be that a local approach to modeling of this data would lead to more predictive models since we consider the causes of specific chemical resistance to be very

local in nature. Inclusion of compounds eliminated because they were not tested against every cell line could expand the dataset and increase understanding of different resistance profiles.

Chapter 5: Chembench

5.1. Introduction

Thanks in large part to publicly funded efforts, there has been an accumulation of bioactivity data in the public domain. The size and complexity of databases containing this data rivals that of the large biological datasets that established the need for bioinformatics. However, the rapidly growing data about interactions of small-molecule probes with biological systems remain largely underexplored because of the absence of appropriate public domain tools for their analysis. This is particularly distressful given the significance of chemical biology for understanding the functions of living organisms.

Within the last decade, cheminformatics has emerged as a burgeoning discipline combining computational, statistical, and informational methodologies with some of the key concepts in chemistry and biology.[119-121] We describe modern cheminformatics broadly as a chemocentric scientific discipline encompassing the creation, retrieval, management, visualization, modeling, discovery, and dissemination of chemical knowledge. Cheminformatics plays a critical role in understanding the fundamental problem of structure-property relationships and therefore applies to almost any area of chemical and biological research. Similar to the role that bioinformatics has played in transforming modern biomedical research, cheminformatics is poised to revolutionize all areas of research in chemical genomics and drug discovery.

While cheminformatics has been recognized as a distinct, impactful scientific discipline, there is a painful absence of cheminformatics tools in the public domain. While some advancement was stimulated by the NIH cheminformatics planning grants awarded to six research groups nationally in 2006, the majority of attainable cheminformatics tools (see Table 2) can perform only rudimentary functions; even the most advanced of the accessible tools lack thorough validation protocols, are poorly integrated with each other, or require specialized knowledge to apply them. Therefore, we chose to develop Chembench, a web portal providing access to several techniques used within the field of cheminformatics.

Table 2. *Limited cheminformatics resources available online or for download (mostly free to academia)*

Repository	Website	Cheminformatics Capabilities
RECCR	http://reccr.chem.rpi.edu	Multiple Modeling Methods Descriptor Generation (paid)
PowerMV	http://nisla05.niss.org/PowerMV/	Multiple Modeling Methods Descriptor Generation Calculation of Drug-like Properties
Cheminformatics.org	http://www.cheminformatics.org/	Similarity Search Diversity Estimation
Molinspiration	http://molinspiration.com/	Calculation of Drug-like Properties Prediction of Drug Class
Indiana	http://sites.google.com/site/davidjwild/home	Similarity Search/ Data Extraction
PubChem	http://pubchem.ncbi.nlm.nih.gov/	Heatmap Generation Similarity Search/Clustering
ChemSpider	http://www.chemspider.com/	Prediction of Properties (ACD/Labs) Similarity Search
VCCLab	http://www.vcclab.org/	Prediction of a Property Descriptor Calculation Multiple Modeling Methods
Laboratoire d'Infochimie	http://infochim.u-strasbg.fr/recherche/Download/Download.php	Fragment Generation MLR modeling Prediction of Biological Activity
SEA	http://sea.bkslab.org/	Prediction of Biological Activity
Mold2	http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/default.htm	Descriptor Generation
Chemistry Development Kit (CDK)	http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page	Descriptor Generation Multiple Modeling Methods
QSAR application Toolbox	http://www.oecd.org/document/23/0,3343,en_2649_34379_33957015_1_1_1_1,00.html	Prediction of Biological Activity Similarity Search Data-Gap Filling
Chemaxon	http://www.chemaxon.com/free-software/	Calculation of Drug-like Properties Similarity Searching (Free) Clustering (paid)
Pipeline Pilot Student	http://accelrys.com/solutions/industry/academic/student-edition.html	Calculation of Drug-like Properties Clustering Fingerprint Generation Multiple Modeling Methods

5.2. Materials and Methods

5.2.1. Chembench Architecture

The Chembench system is quite complicated. Figure 48 contains a simplified representation of the Chembench system detailing general structure and component interaction. A brief summary follows.

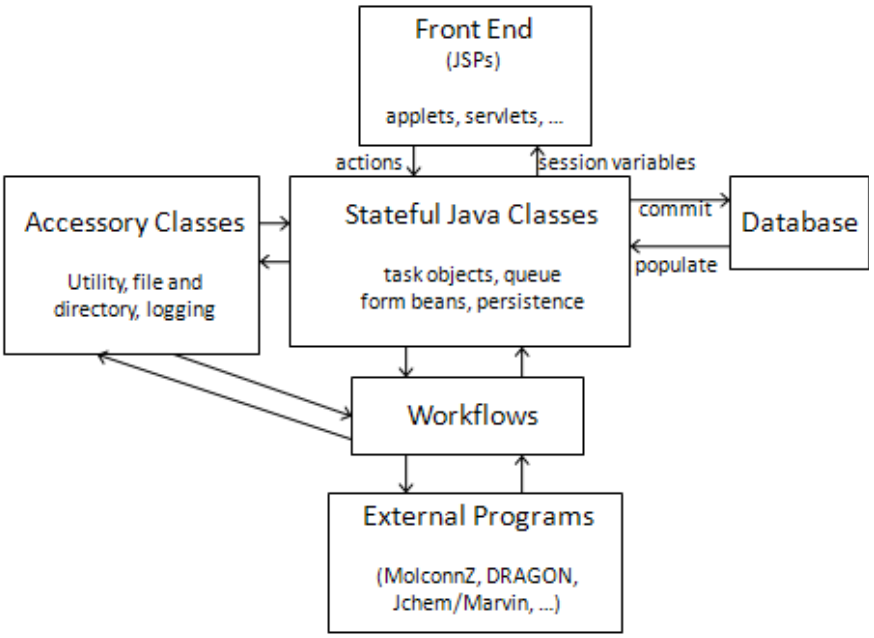


Figure 48. *Overview of Chembench architecture.*

The front end is comprised primarily of JavaServer Pages (JSPs) with the occasional inclusion of an embedded java applet. Information displayed by the JSPs typically is provided via session variables set by the stateful java classes. User-provided input is processed via servlets and passed to stateful java classes when an action within the JSP is executed.

Stateful java classes hold all the data with which a user interacts. The majority of logic within Chembench is carried out within this part of the system. Contained within is the job queuing system.

Accessory classes manage the mundane tasks of the Chembench system. The classes control all global constant definition, I/O operations, and error logging.

Workflows are a set of java functions that interface between the Chembench system and external programs. Several external programs are needed to properly carry out cheminformatics analysis. The workflows portion of code also contains standalone functions that carry out necessary functions of cheminformatics analysis that are not handled by external programs such as data format transformation.

External programs perform the primary actions of cheminformatics analysis of data. Many of these programs are commercial and provided through the generous support of software contributors. External software generates chemical images, calculates descriptors, splits datasets, and develops models.

Information about the users, datasets, models, predictors, predictions, and tasks are all stored within a MySQL database. Stateful java classes access this database is accessed through Hibernate.

5.2.2. Integrated Methods

A large number of external programs have been integrated into the Chembench system. This allows users to perform a series of cheminformatics analyses. The general workflow of data analysis implemented within the system can be seen in Figure 49 taken from Tropsha's recent review of QSAR best practices. [84]

The key steps of the QSAR modeling process are outlined in Figure 49. In Chembench, we have integrated software to standardize structures, split datasets, calculate descriptors, perform y-randomization, build models, and enforce applicability domains. Generally, these tasks are implemented in a modular manner, allowing users to mix and match techniques in each category with members of other categories. As such, Chembench users can undertake a large number of

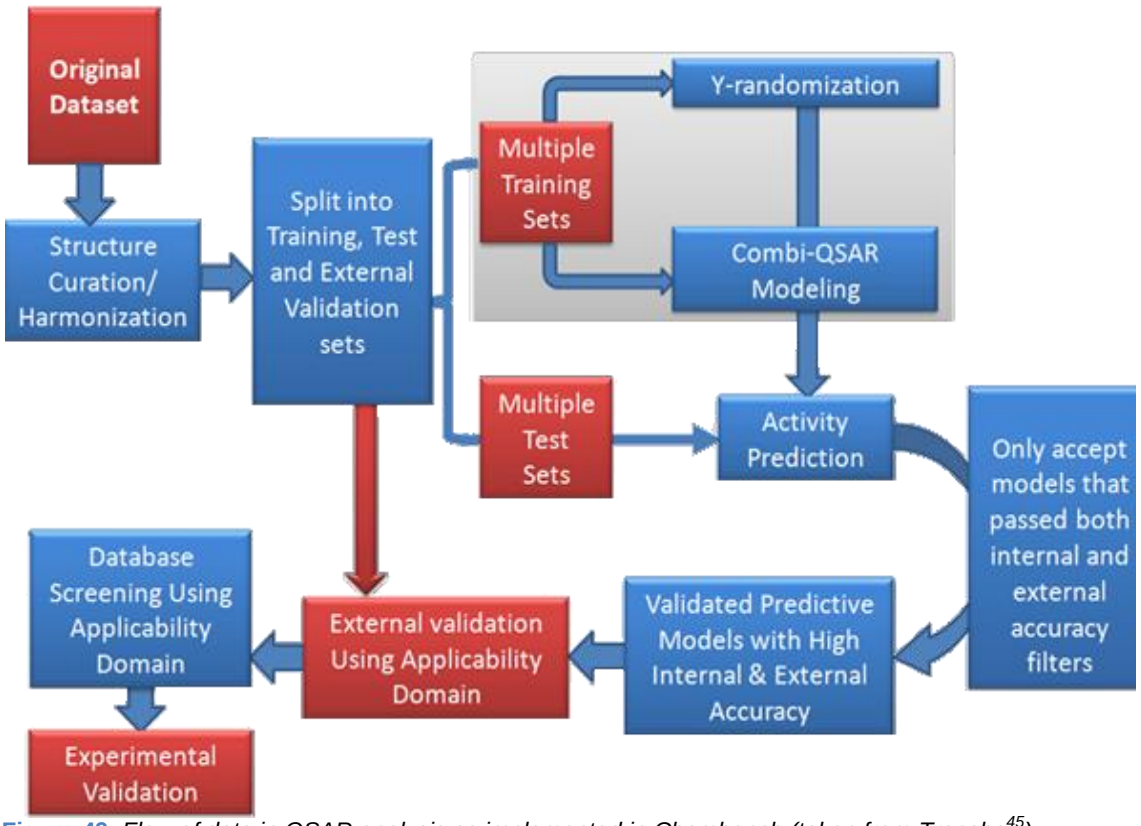


Figure 49. *Flow of data in QSAR analysis as implemented in Chembench (taken from Tropsha[45])*

varying QSAR analyses, and variables in the process can be analyzed individually for their effect on modeling accuracy. Below is described in more detail the techniques that have been integrated into the system.

5.2.2.1. Structure Curation

The importance of structure curation and harmonization has been recently documented[122]. The accuracy of chemical structure representation may have a profound effect on the outcome of cheminformatics studies. Therefore, we have devised a standardized chemical data curation strategy that should be followed at the onset of any molecular modeling investigation. Figure 50 illustrates major steps of this strategy enabled by several publicly available and free-for–academic-use tools. The simple, but important, steps for cleaning chemical records in a database include the removal of a fraction of the data that cannot be appropriately handled by

conventional cheminformatics techniques, e.g., inorganic and organometallic compounds, counterions, salts and mixtures; structure validation; ring aromatization; normalization of specific chemotypes; curation of tautomeric forms; and the deletion of duplicates. It is also critical to visualize and manually inspect at least a fraction of chemical data that go into model development.

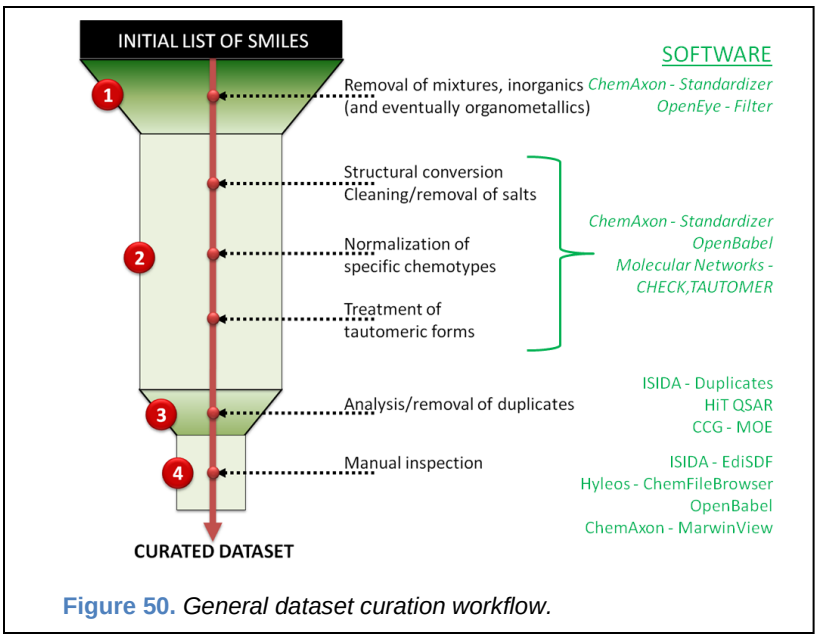


Figure 50. *General dataset curation workflow.*

The current version of Chembench does not have a fully integrated data curation procedure; however, portions have been integrated. The Standardizer component from ChemAxon's Suite[123] of cheminformatics products is used to perform normalization of chemical structures upon user request. Structures can then be manually inspected once a dataset is uploaded.

5.2.2.2. Data Splitting and Validation

As detailed in the dataflow overview in Figure 49, the Chembench website relies on the three way split of datasets into training, test, and external sets. Training sets are used for model generation. Test sets are used for model analysis and selection. External sets are used to validate the predictive power of the ensemble models.

Currently, there are two methods of dataset splitting available in Chembench. The most intuitive is the random split technique that randomly divides the dataset into two subsets whose

proportions are determined by the user. This random split technique can be tempered by use of activity binning to ensure that both subsets have similar activity profiles.

The second technique is that of Sphere Exclusion[83] originated in our lab. This algorithm considers each compound as a point in the multidimensional descriptor space. The procedure starts with the calculation of the distance matrix **D** between representative points in the descriptor space. Let D_{min} and D_{max} be the minimum and maximum elements of **D**, respectively. *N* sphere radii are defined by the following formulas, $R_{min}=R_1=D_{min}$, $R_{max}=R_N=D_{max}/4$, $R_i=R_1+(i-1)*(R_N-R_1)/(N-1)$, where i=2,…,N-1. Each sphere radius corresponds to one division of the set in training and test set. A sphere-exclusion algorithm consists of the following steps.

1. Select randomly a compound.
2. Include it in the training set.
3. Construct a sphere around this compound.
4. Select compounds from this sphere and include them alternatively into test and training sets.
5. Exclude all compounds from within this sphere for further consideration.
6. If no more compounds left, stop. Otherwise let *m* be the number of spheres constructed and *n* be the number of remaining compounds. Let d_{ij} (i=1,…,m; j=1,…,n) be the distances between the remaining compounds and sphere centers. Select a compound corresponding to a user defined rule.

To properly assess the robustness of generated models, models are also always generated for y-randomized data. Statistics of y-randomized models can then be directly compared to those generated on the true data and the significance of generated models can be determined.

5.2.2.3. Descriptor Generation

The generation of Combi-QSAR models requires the calculation of multiple descriptor types in addition to multiple modeling methods. Chembench provides the methods of descriptor generation detailed below. After generation, the descriptors can be normalized either by range-scaling (so that their values are distributed within the interval 0-1) or auto-scaling (subtraction of the mean and then division by the standard deviation). Additionally highly correlated descriptors can be removed.

*DRAGON Descriptors***.** The DragonX software[124] is used to calculate all 2D Dragon descriptors. These included topological descriptors, constitutional descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, functional group counts, atom-centered fragments and molecular properties. DragonX can calculate descriptors for either hydrogen depleted or hydrogen containing representations of a compound.

MolconnZ Descriptors. The MolconnZ software[125] available from EduSoft affords the computation of a wide range of topological indices of molecular structure. These indices include, but are not limited to, the following descriptors: valence, path, cluster, path/cluster and chain molecular connectivity indices[126-128], kappa molecular shape indices[129, 130], topological[131] and electrotopological state indices[132-135], differential connectivity indices[126, 136], graph's radius and diameter[137], Wiener[138] and Platt[139] indices, Shannon [140] and Bonchev-Trinajstić[141] information indices, counts of different vertices, counts of paths and edges between different types of vertices (http://www.edusoft-lc.com/molconn/manuals/400).

MOE2D Descriptors. MOE software[142] is used to generate MOE2D descriptors. These included physical properties, subdivided surface areas, atom and bond counts, Kier and Hall connectivity[126-128] and kappa shape indices[130, 143], adjacency and distance matrix descriptors[138, 144-146], pharmacophore feature descriptors, and partial charge descriptors[147].

MACCS keys. MOE software[148] is used to generate MACCS keys. MACCS keys consist of a set of 166 chemical rules commonly associated with biological activity. This fingerprint was first developed by MDL.

5.2.2.4. Model Development

Three methods of model generation are currently available in Chembench. Of these three techniques, one has been fully developed and its effectiveness has been validated in our lab. Two of these techniques are modifications upon techniques developed elsewhere and modified to enable integration into the Chembench system.

Variable Selected kNN. The first method implemented in Chembench was the variable selected kNN procedure first introduced in the field of cheminformatics in 2000 .[149] This method has been applied in many situations to develop predictive models.

The first version of kNN implemented in the system employed the leave-one-out (LOO) cross-validation (CV) procedure and a simulated-annealing algorithm[150, 151] to optimize variable selection. The procedure starts with the random selection of a predefined number of descriptors from all descriptors. If the number of nearest neighbors k is higher than one, the estimated activities $\hat{y}_i$ of compounds excluded by the LOO procedure are calculated using the following formula:

$$\hat{y}_i = \frac{\sum_{j=1}^{k} y_j w_{ij}}{\sum_{j=1}^{k} w_{ij}} \quad (7)$$

where y_j is the activity of the *j*-th compound. Weights w_{ij} are defined as:

$$w_{ij} = \left(1 + \frac{d_{ij}}{\sum_{j'=1}^{k} d_{ij'}} \right)^{-1} \quad (8)$$

and d_{ij} is Euclidean distances between compound *i* and its *j*-th nearest neighbor. However, if the number of nearest neighbors *k* is equal to one, then the estimated activity $\hat{y}_i$ of the compound will be equal to the activity of this one nearest neighbor.

For classification *k*NN, the predicted $\hat{y}_i$ values (see Equation 7) are rounded to the closest whole numbers (which are, in fact, the class numbers), and the prediction accuracy (correct classification rate, CCR_{train}) is calculated as follows:

$$CCR = 0.5 \left(\frac{N_1^{corr}}{N_1^{total}} + \frac{N_2^{corr}}{N_2^{total}} \right) \quad (9)$$

where N_j^{corr} and N_j^{total} are the number of correctly classified and total number of compounds of class *j* (*j*=1, 2). Then, a predefined small number of descriptors are randomly replaced by other descriptors from the original pool, and the new value of CCR_{train} is obtained. If CCR_{train} (new) > CCR_{train} (old), the new set of descriptors is accepted. If CCR_{train} (new) ≤ CCR_{train} (old), the new set of descriptors is accepted with probability p = exp (CCR (new) - CCR (old))/T, or rejected with probability (1-p), where T is a simulated annealing (SA) "temperature" parameter.

During this process, T is decreasing until a predefined threshold. Thus, the optimal (highest) CCR_{train} is achieved. For the prediction, the final set of selected descriptors is used, and Equation 7 and 8 are applied to predict activities of compounds of the test sets. Then the activities are rounded to the closest whole numbers, and the correct classification rate for the test set is calculated using Equation 9.

For continuous kNN, this procedure is maintained, but the optimization function is changed from CCR to q^2. q^2 is calculated according to Equation 10.

——— (10)

In addition to the simulated annealing procedure for variable selected, we have recently added the genetic algorithm (GA)[152] method of optimization. Rather than starting with a single randomly selected set of descriptors, the genetic algorithm is initiated with a population of different randomly selected descriptors. Similar to SA, the fitness of member of the population is calculated using Equation 9 or 10 for classification or continuous modeling respectively based on predicted values determined using a LOO-CV procedure and Equations 7 and 8. A second generation of the population is spawned through breeding (crossover) of parents selected based on their fitness. Generation will continue to be spawned until a predefined number of generations have been created or none of the members of the population have become more fit in a set number of generations.

Support Vector Machine. A common learning technique applied in the field of data classification is that of Support Vector Machines (SVM). SVM was developed by Vapnik[153] as a general data modeling methodology where both the training set error and the model complexity are incorporated into a special loss function that is minimized during model development. SVM

has been extended to afford the development of SVM regression models for datasets with continuous activities. It has been used in several QSAR applications.[154, 155]

To provide access to SVM learning, we have integrated the LIBSVM package[156] in Chembench. LIBSVM provides several SVM variants including traditional SVM (C-SVC), Regression SVM (epsilon-SVR), and nu-SVM implementation for both classification and regression. LIBSVM also provides several kernels for transformation of the descriptor space. Our own grid modeling technique was implemented on top of LIBSVM to generate ensembles of SVM models.

Random Forest. Random forest is a technique developed by Breiman[157] that builds series of decision trees based on a dataset and then uses them as an ensemble predictor. Typically, the optimal decision tree is generated for a randomly selected subset of a dataset and a randomly select subset of descriptors. This typical implementation of random forests is unfortunately not modular as it requires a specific implementation of dataset splitting. Therefore, a variant of random forests with alterations to internal training and test set selection was done to maintain the modular nature of modeling within Chembench so splitting techniques can be altered without variation of learning methods.

The modified random forest procedure in Chembench is quite similar to the traditional application of random forest but varies in the way that modeling set selection is done. Rather than a new training set being selected for each new tree grown, a manageable number of internal training sets are defined and then multiple trees (a grove) grown for each of these sets. Additionally, these sets are selected without replacement. The generation of groves is done using the randomForest package for R available from http://stat-

www.berkeley.edu/users/breiman/RandomForests. The original implementation of random forests can be mimicked by performing a large number of data splits and generating only a single tree for each split.

5.3. Results and Discussion

The Chembench web portal was officially released to the public in April of 2010 at http://chembench.mml.unc.edu. Users upon entering the site can choose to either register or to use the system as a guest. Guest users have all the capabilities of registered users, but their data objects are subject to periodic deletion and their data is accessible by any other guest user.

The capabilities of the website consist of functions organized around 3 key components of the QSAR workflow: Datasets, Modeling, and Prediction. These three components become the objects generated within the portal upon use of three tabs containing forms controlling their generation. An additional tab (My Bench) allows management and further analysis of these three types of objects.

5.3.1. Datasets

Datasets are generally the entrance point for users to a cheminformatics analysis. A dataset is required for a user to develop a model or make predictions (though the inclusion of public datasets and models allows Chembench users to bypass this step).

5.3.1.1. Dataset Creation

The Chembench interface for dataset uploading allows many options for users inputting their data. The primary option is the type of dataset a user would like to upload. Users can choose to upload modeling and prediction sets either with or without pre-calculated descriptors. Modeling sets require the inclusion of an activity value for each compound, allowing the generation of

models. Users are required to designate whether activity values are continuous or categorical in nature. Prediction sets can consist of purely structural data meant to be annotated by previously generated predictors. The inclusion of pre-calculated descriptors allows users to compare their own descriptor generation packages to those integrated in the Chembench framework. If for confidentiality reason, a user wishes to use the site without uploading chemical structures, they can upload a set with no structure file, but pre-calculated descriptors. However, this precludes the use of the integrated descriptor generation techniques and the use of the system for commercial calculations is prohibited. Upload format standards are defined in the help documentation.

When uploading the dataset, users are expected to define an external set. This external set will be extracted from the uploaded set by random selection. For users that have already defined an external set outside the Chembench site, input of a list of identifiers is provided to ensure comparability of Chembench results to those of nonintegrated methods.

Once a dataset is named and the form is submitted, a series of data checks are done to ensure that formatting of the uploaded data files is correct. Additionally, identifiers are checked for uniqueness and are matched across all uploaded files to verify their capability to be used as a key. Once data compatibility with the Chembench system has been validated, the dataset is created; the external set is defined; descriptors are calculated; and 2D chemical images are generated.

5.3.1.2. Dataset Inspection

Once a dataset has been created in Chembench, it can be accessed via the My Bench page. Also, it will be populated in lists of datasets on the Modeling and Prediction pages where relevant.

Selection of a dataset on the My Bench page will allow user to inspect several aspects of their dataset. All compounds are contained in a table so users can manually ascertain the correctness of interpretation of their upload structures. The selected external set can be viewed. A histogram of the activity values uploaded for the dataset is provided. A heatmap of Mahalanobis distance or Tanimoto similarity between compounds in MACCS key space is available. Finally, any warnings or errors in descriptor generation are provided to the user for consideration prior to modeling.

5.3.2. Modeling

The modeling step is considered the primary contribution of Chembench to the public. Modeling is complex and option rich. It depends on the consistency and accuracy of the uploaded dataset and is required for identification of compounds of interest from chemical library. To better distinguish the difference between individual models generated and the ensemble models (i.e. the consensus of individual models), with in Chembench the latter is referred to as a "predictor".

5.3.2.1. Model Generation

The initiation of model generation depends on the selection of a dataset. Modeling datasets are segregated into two groups, continuous sets and category sets, because the applicability of modeling techniques and parameters of the modeling techniques are dependent on that

designation. All datasets available for modeling are provided in a drop-down list for selection with user uploaded datasets being listed first.

Once a user has selected a dataset, they can choose the descriptor type that they would like to use for model generation. Descriptor types for which at least one compound in the dataset could not be generated are grayed out and not selectable. If a user desires to apply a grayed out descriptor type, they must address the issues identified in the "Descriptor Warnings" section when they view that dataset. The descriptors selected for model development can be additionally processed by eliminating highly correlated descriptors and by normalizing descriptors using either range-scaling or autoscaling.

The internal splitting of the dataset can be accomplished using either random splitting or sphere exclusion. Both methods allow the user to specify the number of splits to generate and the approximate size of the test sets. Additional parameters of the sphere exclusion method are made available to users on its tab.

The modeling method section of the page allows users to select from the currently supported methods of model generation. Each method has its own tab which when selected provides access to the many parameters necessary to control the model development algorithm. Default values for all these parameters are provided based on the modeling experience of the site developers and parameter limitations are enforced to prevent improper parameter inputs.

5.3.2.2. Predictor Review

Once a predictor name is defined and the job is submitted, it will be sent to the queue and modeled will be completed either locally or on the emerald computing cluster depending on the modeling type. Job progress can be tracked on the My Bench page. Upon job completion, it can

be accessed through the My Bench page and users will be notified via email (if requested on submission). It will also be made available to that user as a predictor on the Prediction page.

Accessing a predictor from My Bench allows users to see several aspects of the modeling results. Of most interest is the prediction accuracy of the predictor on the external set. A table is provided containing the predicted value, actual value, residual, and number of models applicable for each compound of the external set. A summary of this information is contained either in a confusion matrix for categorical modeling or a plot of predicted vs. actual values for continuous modeling. The correct statistic (either CCR or R^2) is calculated to provide a quantitative assessment of the predictor's accuracy. In addition to external results, internal modeling information is recorded in the models or trees tab. Herein, the statistics of the individual models that compose the ensemble are provided. Also, the results of model generation using Y-randomized activities are provided so the user can validate that the models generated using the correct data are significantly better than those using y-randomized activities. Finally, the model generation parameters are displayed to remind the user of the protocol applied.

5.3.3. Prediction

While modeling is the expertise of the authors of the Chembench web portal, the most publicly beneficial portion of the site may be the Prediction tab. Here, users can quickly and easily identify compounds that are expected to have properties of interest.

Prediction is a two-step process. First a user must select the predictors they would like to use. These predictors are separated into private (the predictors that user has generated) and public (the predictors provided by the authors). Public predictors are categorized by the type of activity (specific target interaction, toxicity, or ADME related properties) that they predict.

Multiple predictors can be selected for a single prediction allowing users the ability to see a spectrum of predicted activities if they desire. Once a user has selected the predictor(s) that they wish to apply, they then are given the opportunity to predict either a previously uploaded dataset or a single compound defined by a SMILES string or drawn in MarvinSketch applet. All prediction jobs are submitted to the queue and can be accessed on completion from the My Bench page. Prediction results are paginated and can be sorted on any of the predicted values.

5.3.4. Additional Features

The Chembench web portal has several components and aspects that are vital for its function but not directly related to cheminformatics analysis of data.

One of the most important aspects of the website is that it is user specific. User sessions are created upon login. All objects within Chembench are linked to a user. This allows the website to protect the private data of individual users. It also enables the customization of interfaces for users depending on their level of expertise. Several parameters available for tuning of model building are only of interest to experts in the field of cheminformatics. Display of these parameters can be turned on and off under a user's profile. The amount of public data a user wishes to access can also be modified. Also, the definition of users allows the ability to provide special access to data for some. In particular, the ability to download descriptors can be enabled for users with the appropriate software licenses. The user system also provides interface for administrative actions within Chembench. Users defined as administrators can view and control many aspects of the system including canceling of other user's jobs. However, the most important aspect of the user oriented aspect of Chembench is that it allows users to submit jobs and easily retrieve them at a later time.

The queue component of the system enables the efficient use of computational power. Chembench is hosted on an 8 processor system managed by ITS research computing at UNC. It is linked to the Emerald computing cluster, which has more than 800 processors. While the available computing power is large, shorter jobs are much more efficiently handled by the local system whereas larger jobs typically are better treated on the cluster. As such, the queue in Chembench has been designed to handle different types of jobs in different ways. It has also been structured to allow the easy addition of other computational resources. The design of the queue provides users with fast and efficient generation of their models and prevents jobs from overrunning the host server and causing portal usability issues.

The most important piece of the Chembench portal is that it has developed a user base. There are now over 200 registered users of the Chembench site and frequently multiple users are logged on at once. In total the site has run over 9000 jobs and provided nearly 11.5 years of compute time to the public.

5.3.5. Public Datasets and Models

Chembench was originally intended as a way to provide access to the results of work within the Tropsha lab to the public. As such, the site is populated with many datasets generated or used within our lab as well as several validated and published predictors. Table 3 and Table 4 list the datasets and predictors currently available via Chembench.

The lack of availability of the datasets and predictors generated as part of the development of a benchmark for virtual screening detailed in Chapter 3 is an omission caused by the fact that we are in the process of upgrading the site to handle datasets with multifold external sets. As such, these datasets and predictors will be available shortly. Also, the addition of the random forest

Table 3. *Selected datasets made available through Chembench*

Dataset Name	Number of Compounds	Reference	Description
HDAC_59	59	J Chem Inf Model. 2009 Feb;49(2):461-76.	A set of 59 hdac inhibitors used to generate models as discussed in the above referenced article.
Ames_Mutagenicity	6542	Pending	A set of 6452 compounds with a binary assessment of the mutagenic liability.
T.Pyriformis_Mod	983	http://dx.doi.org/10.1021/ci700443v	The set of 983 compound with measured values of toxicity against T.Pyriformis used for modeling in the above reference.
T.Pyriformis_Ext2	110	http://dx.doi.org/10.1021/ci700443v	The set of 110 compounds with measured values of toxicity against T.Pyriformis used as a second validation set in the above reference.
Drugbank	4494	http://www.drugbank.ca	A set of 4494 compounds retrived from Drugbank after standardization, cleaning, and de-duplication.
P-Glycoprotein	195	http://dx.doi.org/10.1021/ci0504317	A set of 195 substrate/non-substrates for P-Glycoprotein used as a modeling set in the linked reference.

and SVM modeling algorithms to the site has been recent. Several predictors reliant on these methods are currently waiting reformatting for input into the Chembench framework.

5.4. Conclusions and Future Directions

We have completed the following key steps in the generation of web portal to allow the application of cheminformatics techniques by worldwide users.

1. Integration of cheminformatics software for structure standardization, descriptor generation, model development, and prediction (Sections 5.3.1-5.3.3)
2. Development of a queuing system to manage cluster and local job (Section 5.3.4)
3. Creation of an easy-to-use interface allowing experts and non-experts in cheminformatics access to needed tools (Section 5.3.4)
4. Publication through the portal of more than 50 datasets and 7 validated predictors (Section 5.3.5)

Table 4. *Predictors made available through Chembench*

Predictor Name	Modeling Method	Descriptor Type	Predictor Class	Description
48_ ANTICONV	KNN	MOLCONNZ	DrugDiscovery	This predictor is a regeneration of the SA-kNN models developed by M Shen; et al in http://dx.doi.org/10.1021/jm030584q. These models built using 48 Functionalized Amino Acids (FAAs) predict the log(ED50 Âµmol/kg) of chemicals in the mice Maximal Electroshock Seizure (MES) test.
T.Pyriformis	KNN	MOLCONNZ	Toxicity	This predictor contains the kNN-MolconnZ models generated by H Zhu; et al in http://dx.doi.org/10.1021/ci700443v. These models built using 983 compounds (644 training/339 external test) predict aquatic toxicity (pIGC50) against Tetrahymena Pyriformis.
P-Glycoprotein _DragonkNN	KNN	DRAGONH	ADME	This predictor is the regeneration of models developed by P de Cerqueira Lima; et al in http://dx.doi.org/10.1021/ci0504317 using DRAGON descriptors with SA-kNN . These binary models built using 195 compounds predict whether a compound will be a substrate for P-Glycoprotein (1) or will be a non-substrate (0).
Blood_Brain_Barrier_MZkNN	KNN	MOLCONNZ	ADME	This predictor contains the kNN-MolconnZ models generated by L Zhang; et al in http://dx.doi.org/10.1007/s11095-008-9609-0. These models built using 159 compounds (144 training/15 external test) predict the log(BB) in rats. .
Anti-Malarial_DragonkNN	KNN	DRAGONH	DrugDiscovery	This predictor is a collection of models generated in the Tropsha lab on a set of 3133 compounds screened for their antimalarial activities in St. Jude Children's Research Hospital. These binary models predict whether a compound will inhibit growth of the P. falciparum 3D7 strain (1) or not (0).
5HT2B_Binder_ DragonkNN	KNN	DRAGONH	Toxicity	This predictor contains models generated using Dragon and kNN by R Hajjo; etal in http://dx.doi.org/10.1021/jm100600y. These models built and validated using 304 compounds with binder/non-binder classification defined based on functional assays.
RAT-ACUTE-LD50_DragonkNN	KNN	DRAGONH	Toxicity	This predictor contains models generated using Dragon and kNN by H Zhu; etal in http://dx.doi.org/10.1021/tx900189p. These models built and validated using 3472 compounds predict Acute Toxicity (pLD50(mol/kg)) in Rats.

This site provides access to methods commonly used in the field of cheminformatics through a simple user interface that can be tailored to allow more advanced usage. Additionally, the site contains several predictors of biological properties that could be used by non-experts in the field of cheminformatics to assess compounds of interest prior to synthesis or experimental testing.

The creation of Chembench was completed in a multidisciplinary team headed by Dr. Diane Pozefsky. The writing of the software was primarily completed by hiring of developers with a computer science background. As the scientific lead on the team my primary contribution was in communication of the workflows used by cheminformaticians, training developers in cheminformatics software, and definition of user interface requirements. In addition, I wrote the original version of the underlying MySQL database and was tasked with collection of public datasets and predictors.

Development of the site is an ongoing project. There are additional methods and techniques to be added to the site, in particular the integration of molecular descriptors that are not bound by license. While we are grateful to software contributors for providing their tools for descriptor use within the site, allowing users to download descriptors would increase the usefulness of the web portal within the cheminformatics community.

The integration of the website with repositories for biological data is undergoing development. Creation of web service protocols allowing efficient transfer of data between ChemSpider and Chembench has been completed, but integration of the protocols into the user interface is still ongoing. Completion of integration with ChemSpider will provide a proof of concept to aid the integration of Chembench with PubChem and other public databasing efforts.

Appendix I: Amino Acid to Feature Transformations

Contained in this appendix is a table of the transformations used to generate features from amino acids. For each amino acid fragment, the atoms selected as part of the binding pocket were transformed to features. Seeing as some features contain more than one atom from an amino acid, as long as a portion of the atoms of that feature were contained in the defined pocket, the feature was included. The location of the feature was calculated as the average of the atomic coordinates of atoms defined as being a part of the binding pocket which comprise that feature.

Feature ID	Residue	Pharmacophore Feature	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6
1	GLU	-	CD	OE1	OE2			
2	ALA	A	OXT					
3	ALA	H	CB					
4	ALA	D	N					
5	ALA	A	O					
6	ARG	A	OXT					
7	ASN	A	OXT					
8	ARG	H	CB	CG				
9	ASP	A	OXT					
10	CYS	A	OXT					
11	ARG	+	CZ	NE	NH1	NH2		
12	ARG	D	N					
13	ARG	D	NE					
14	ARG	D	NH1					
15	ARG	D	NH2					
16	ARG	A	O					
17	GLN	A	OXT					
18	HIS	A	OXT					
19	ASN	H	CB					
20	LYS	A	OXT					
21	ASN	D	N					
22	ASN	A	ND2					
23	ASN	A	O					
24	ASN	A	OD1					

Feature ID	Residue	Pharmacophore Feature	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6
25	MET	A	OXT					
26	PRO	A	OXT					
27	ASP	H	CB					
28	ASP	-	CG	OD1	OD2			
29	ASP	D	N					
30	ASP	A	O					
31	ASP	A	OD1					
32	ASP	A	OD2					
33	SER	A	OXT					
34	TRP	A	OXT					
35	CYS	H	CB					
36	CYS	D	N					
37	CYS	A	O					
38	CYS	A	SG					
39	TYR	A	OXT					
40	ALA	-	C	O	OXT			
41	GLN	H	CB	CG				
42	ARG	-	C	O	OXT			
43	ASN	-	C	O	OXT			
44	GLN	D	N					
45	GLN	A	NE2					
46	GLN	A	O					
47	GLN	A	OE1					
48	ASP	-	C	O	OXT			
49	CYS	-	C	O	OXT			
50	GLU	H	CB	CG				
51	GLN	-	C	O	OXT			
52	GLU	-	C	O	OXT			
53	GLU	D	N					
54	GLU	A	O					
55	GLU	A	OE1					
56	GLU	A	OE2					
57	GLU	A	OXT					
58	GLY	-	C	O	OXT			
59	HIS	-	C	O	OXT			
60	GLY	D	N					
61	GLY	A	O					
62	GLY	A	OXT					
63	ILE	-	C	O	OXT			

Feature ID	Residue	Pharmacophore Feature	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6
64	LEU	-	C	O	OXT			
65	HIS	H	CB					
66	LYS	-	C	O	OXT			
67	MET	-	C	O	OXT			
68	HIS	R	CG	CD2	NE2	CE1	ND1	
69	HIS	D	N					
70	HIS	+	ND1					
71	HIS	+	NE2					
72	HIS	A	O					
73	PHE	-	C	O	OXT			
74	PRO	-	C	O	OXT			
75	ILE	H	CB	CG1	CD1	CG2		
76	SER	-	C	O	OXT			
77	THR	-	C	O	OXT			
78	TRP	-	C	O	OXT			
79	ILE	D	N					
80	ILE	A	O					
81	ILE	A	OXT					
82	TYR	-	C	O	OXT			
83	VAL	-	C	O	OXT			
84	LEU	H	CB	CG	CD1	CD2		
85	ASN	D	ND2					
86	ASN	D	OD1					
87	CYS	D	SG					
88	LEU	D	N					
89	LEU	A	O					
90	LEU	A	OXT					
91	GLN	D	NE2					
92	GLN	D	OE1					
93	LYS	H	CB	CG	CD			
94	HIS	D	ND1					
95	HIS	D	NE2					
96	LYS	D	NZ					
97	LYS	D	N					
98	LYS	+	NZ					
99	LYS	A	O					
100	SER	D	OG					
101	THR	D	OG1					
102	MET	H	CB	CG				

Feature ID	Residue	Pharmacophore Feature	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5	Atom 6
103	MET	H	CE	SD				
104	TYR	D	OH					
105	MET	D	N					
106	MET	A	O					
107	PHE	H	CB					
108	PHE	H	CG	CD1	CE1	CZ	CDE2	CD2
109	PHE	R	CG	CD1	CE1	CZ	CE2	CD2
110	PHE	D	N					
111	PHE	A	O					
112	PHE	A	OXT					
113	PRO	H	CB	CG				
114	PRO	A	O					
115	SER	D	N					
116	SER	A	O					
117	SER	A	OG					
118	THR	H	CG2					
119	THR	D	N					
120	THR	A	O					
121	THR	A	OG1					
122	THR	A	OXT					
123	TRP	H	CB	CG				
124	TRP	R	CG	CD1	NE1	CE2	CD2	
125	TRP	R	CD2	CE2	CZ2	CH2	CZ3	CE3
126	TRP	H	CD2	CE2	CZ2	CH2	CZ3	CE3
127	TRP	D	N					
128	TRP	D	NE1					
129	TRP	A	O					
130	TYR	H	CB					
131	TYR	R	CG	CD1	CE1	CE2	CD2	CZ
132	TYR	H	CG	CD1	CE1	CE2	CD2	
133	TYR	D	N					
134	TYR	A	O					
135	TYR	A	OH					
136	VAL	H	CB	CG1	CG2			
137	VAL	D	N					
138	VAL	A	O					
139	VAL	A	OXT					

Appendix II: Selected Clusters from the PDBBind Core Set

Contained in this appendix is the list of PDBBind cluseters defined as containing a single protein using the criteria described in Section 2.3.5. These proteins formed the dataset for a more accurate test of CoLiBRI's virtual screening capabilities.

CLUSTER_ID	PDB_ID	NAME
1	1ps3 3d4z 2f7o	ALPHA-MANNOSIDASE II
2	1amw 1bgq 2iwx	HEAT SHOCK PROTEIN 90 HEAT SHOCK PROTEIN 82
5	3cj2 1nhu 2d3u	RNA-DEPENDENT RNA POLYMERASE
7	1ajp 1ai5 1ajq	PENICILLIN AMIDOHYDROLASE
8	1gpk 1h23 1e66	ACETYLCHOLINESTERASE
9	2rkm 1b9j 1b7h	OLIGO-PEPTIDE BINDING PROTEIN
10	2qv4 1u33 1xd1	ALPHA-AMYLASE
11	1uwt 2ceq 2cer	BETA-GALACTOSIDASE
12	2qwb 2qwd 2qwe	NEURAMINIDASE
13	2j77 2j78 2cet	BETA-GLUCOSIDASE A
14	3ccw 3cdb 3cd5	3-HYDROXY-3-METHYLGLUTARYL-COENZYME A REDUCTASE
15	3bra 3ckp 2g94	BETA-SECRETASE 1
16	2qfu 1x8r 2pq9	3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE
18	1n2v 1k4g 1s39	QUEUINE TRNA-RIBOSYLTRANSFERASE TRNA GUANINE TRANSGLYCOSYLASE
19	1kv1 2bak 3e93	MITOGEN-ACTIVATED PROTEIN KINASE P38 MITOGEN-ACTIVATED PROTEIN KINASE 14

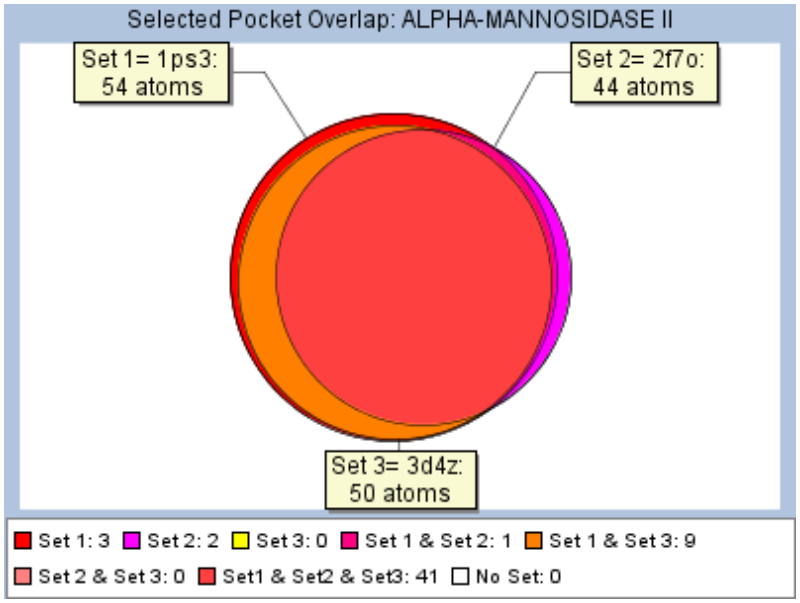
CLUSTER_ID	PDB_ID	NAME
20	2hu6 3f19 3f17	MACROPHAGE METALLOELASTASE (MMP-12)
21	1ndw 1ndy 1ndz	ADENOSINE DEAMINASE
22	1m2q 1zoe 2pvk	CASEIN KINASE II
24	2v00 5er2 4er2	ENDOTHIAPEPSIN
25	2qbp 1nl9 2azr	PROTEIN-TYROSINE PHOSPHATASE PROTEIN-TYROSINE PHOSPHATASE 1B
26	2wec 1bxq 1bxo	PENICILLOPEPSIN
27	2brb 2c3j 1nvq	SERINE/THREONINE-PROTEIN KINASE CHK1
28	4tln 1tmn 4tmn	THERMOLYSIN
31	2exm 1b38 1pxo	CELL DIVISION PROTEIN KINASE 2
33	1qi0 1w3k 1w3l	ENDOGLUCANASE B ENDOGLUCANASE 5A
34	1bcu 1c1v 1sl3	THROMBIN
37	1jqd 1jqe 2aou	HISTAMINE N-METHYLTRANSFERASE
38	1y1z 1pb8 1pbq	N-METHYL-D-ASPARTATE RECEPTOR SUBUNIT 1
39	2obf 1hnn 2g71	PHENYLETHANOLAMINE N- METHYLTRANSFERASE
41	1p1q 1syh 1ftm	GLUTAMATE RECEPTOR 2
43	1fcx 1fd0 1fcz	RETINOIC ACID RECEPTOR GAMMA-1
44	1f4e 1f4f 1f4g	THYMIDYLATE SYNTHASE
45	1yc1 3ekr 2uwd	HEAT SHOCK PROTEIN HSP90-ALPHA
46	2osf 2pow 1if7	CARBONIC ANHYDRASE II
47	2bok 1mq6 1nfy	COAGULATION FACTOR X COAGULATION FACTOR XA
48	2usn 2d1o 1hfs	STROMELYSIN-1

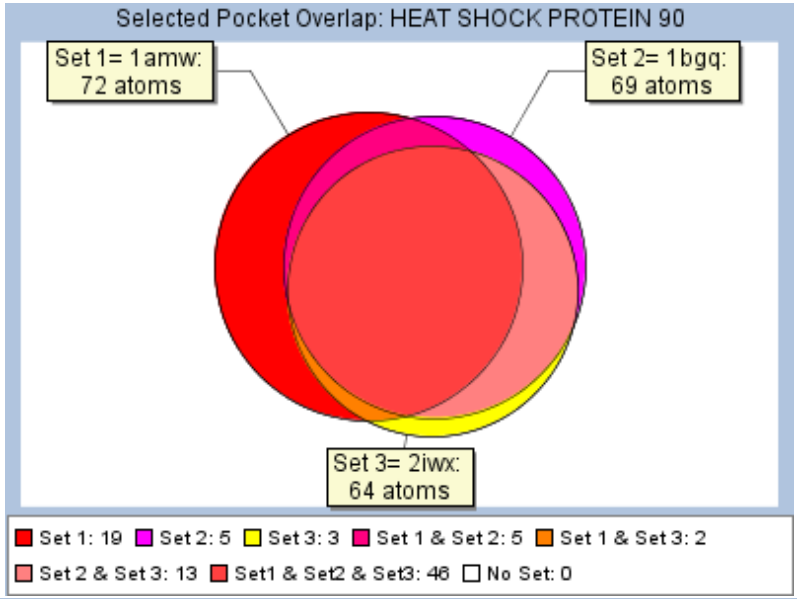
CLUSTER_ID	PDB_ID	NAME
49	2flr 2b7d 2bz6	COAGULATION FACTOR VII COAGULATION FACTOR VIIA
50	1loq 1lol 1x1z	OROTIDINE 5'-MONOPHOSPHATE DECARBOXYLASE
52	1uto 1g3e 1o3f	TRYPSINOGEN TRYPSIN BETA
53	1jys 1nc1 1y6q	MTA/SAH NUCLEOSIDASE
54	1bma 1ela 1elb	ELASTASE
55	1pr5 1a69 1k9s	PURINE NUCLEOSIDE PHOSPHORYLASE
56	3pce 3pcn 3pcj	PROTOCATECHUATE 3
57	2pgz 3c84 2bys	ACETYLCHOLINE-BINDING PROTEIN
61	6std 2std 3std	SCYTALONE DEHYDRATASE
62	1jaq 1zs0 1zvx	NEUTROPHIL COLLAGENASE (MMP-8)
64	2g8r 1o0h 1u1b	RIBONUCLEASE PANCREATIC
65	1sv3 1jq8 2arm	PHOSPHOLIPASE A2
68	1d7j 1fki 1fkb	FK506 BINDING PROTEIN (FKBP)

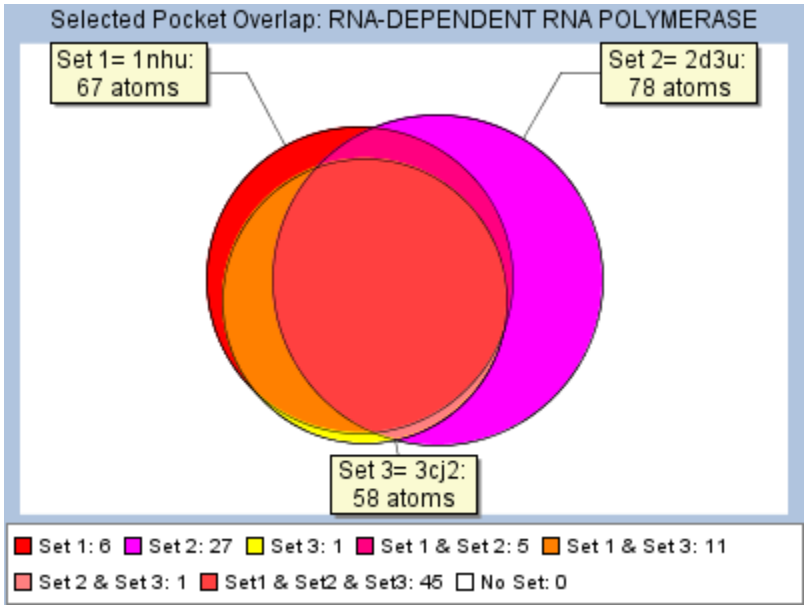
Appendix III:Venn Diagrams of Pocket Overlap

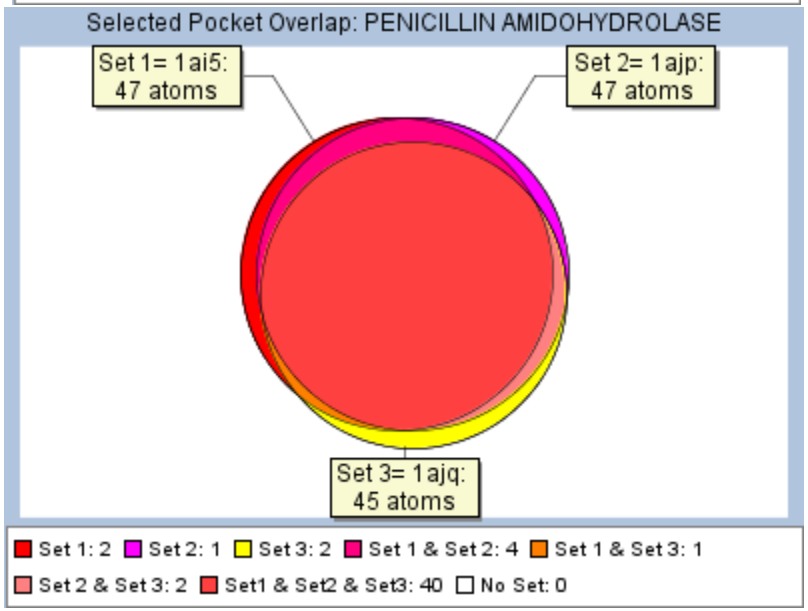
Contained in this appendix is the compendium of venn diagrams (as exemplified in Figure 11 and 16) generated while assessing the consistency of pockets defined for different protein-ligand complexes of the same protein. The figures are separated based on the technique used to identify the pocket. There are 49 diagrams for protein-ligand tessellated pockets, 11 for CastP pockets, and 24 for SCREEN. The reduced number of examples for the two latter methods is due to those methods not identifying the binding pocket for at least one of the protein-ligand complexes for a protein. Overall the venn diagrams display that pocket detection with these methods is inadequate for consistent identification of the same protein pocket for the multiple representatives of a protein.

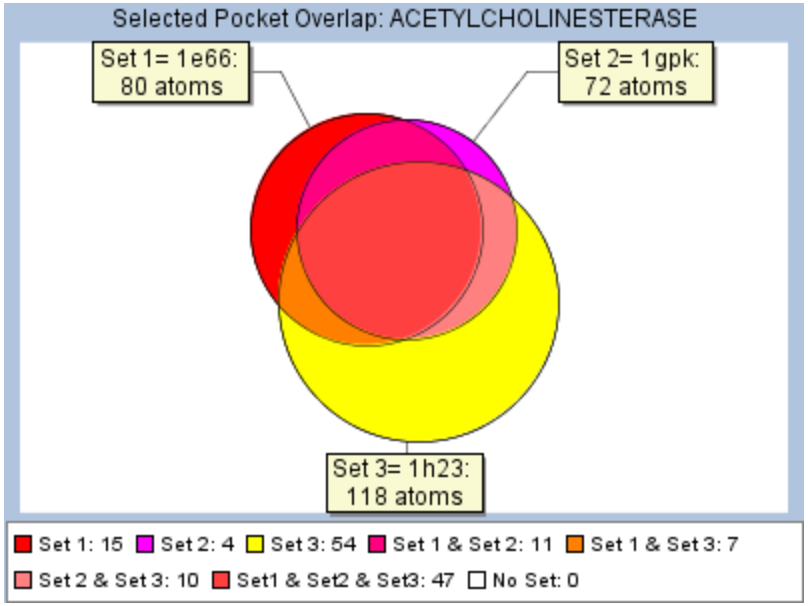
Protein-Ligand Tessellation

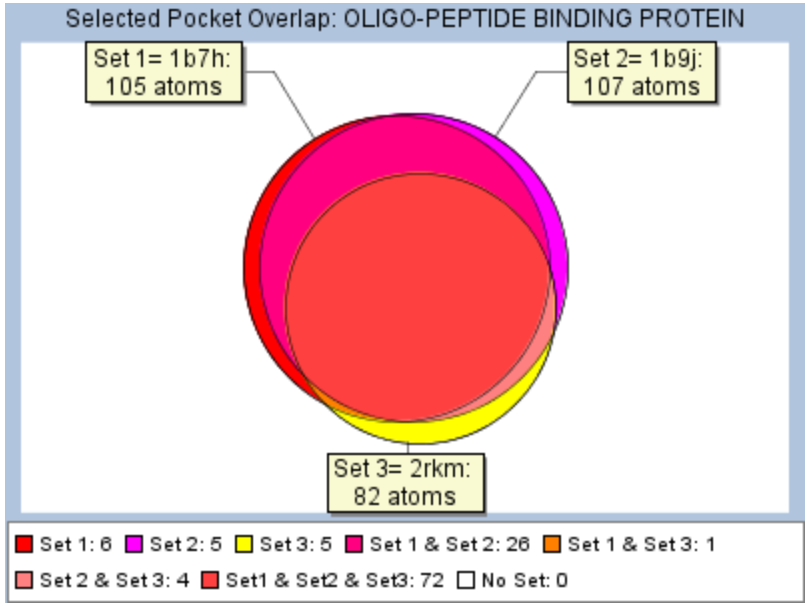

Selected Pocket Overlap: ALPHA-MANNOSIDASE II
Set 1= 1ps3: 54 atoms
Set 2= 2f7o: 44 atoms
Set 3= 3d4z: 50 atoms
Set 1: 3
Set 2: 2
Set 3: 0
Set 1 & Set 2: 1
Set 1 & Set 3: 9
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 41
No Set: 0

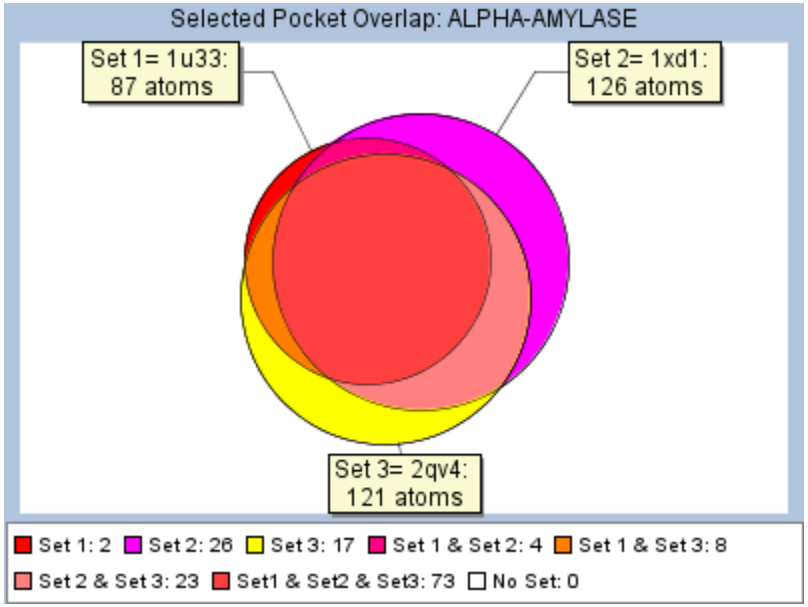

Selected Pocket Overlap: HEAT SHOCK PROTEIN 90
Set 1= 1amw: 72 atoms
Set 2= 1bgq: 69 atoms
Set 3= 2iwx: 64 atoms
Set 1: 19
Set 2: 5
Set 3: 3
Set 1 & Set 2: 5
Set 1 & Set 3: 2
Set 2 & Set 3: 13
Set1 & Set2 & Set3: 46
No Set: 0

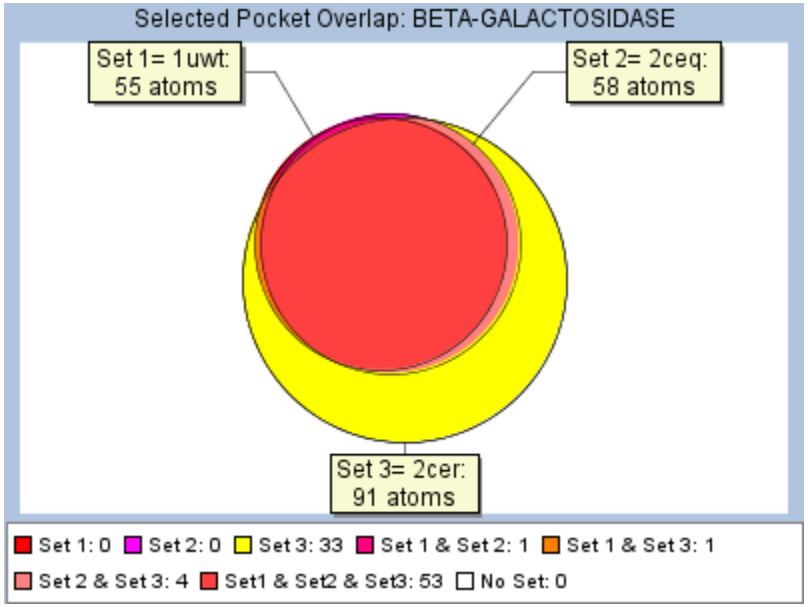

Selected Pocket Overlap: RNA-DEPENDENT RNA POLYMERASE
Set 1= 1nhu: 67 atoms
Set 2= 2d3u: 78 atoms
Set 3= 3cj2: 58 atoms
Set 1: 6
Set 2: 27
Set 3: 1
Set 1 & Set 2: 5
Set 1 & Set 3: 11
Set 2 & Set 3: 1
Set1 & Set2 & Set3: 45
No Set: 0

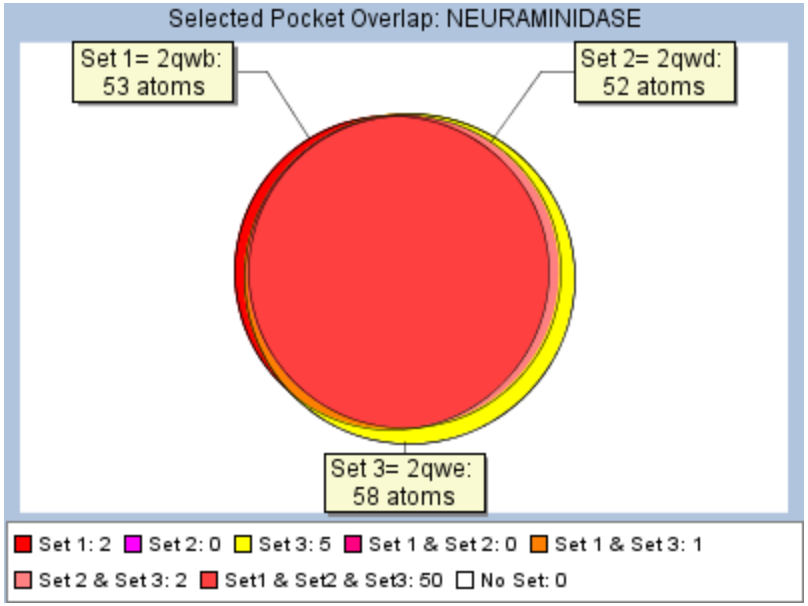

Selected Pocket Overlap: PENICILLIN AMIDOHYDROLASE
Set 1= 1ai5: 47 atoms
Set 2= 1ajp: 47 atoms
Set 3= 1ajq: 45 atoms
Set 1: 2
Set 2: 1
Set 3: 2
Set 1 & Set 2: 4
Set 1 & Set 3: 1
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 40
No Set: 0

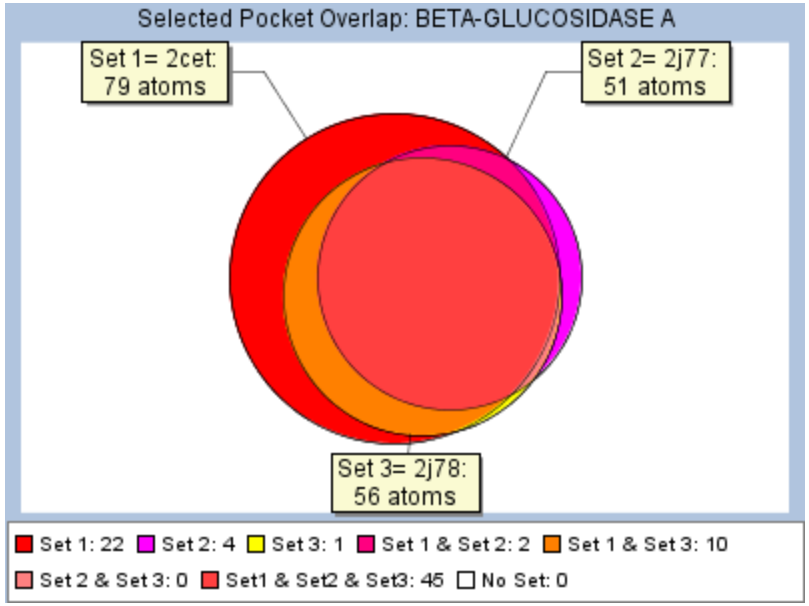

Selected Pocket Overlap: ACETYLCHOLINESTERASE
Set 1= 1e66:
80 atoms
Set 2= 1gpk:
72 atoms
Set 3= 1h23:
118 atoms
Set 1: 15
Set 2: 4
Set 3: 54
Set 1 & Set 2: 11
Set 1 & Set 3: 7
Set 2 & Set 3: 10
Set1 & Set2 & Set3: 47
No Set: 0

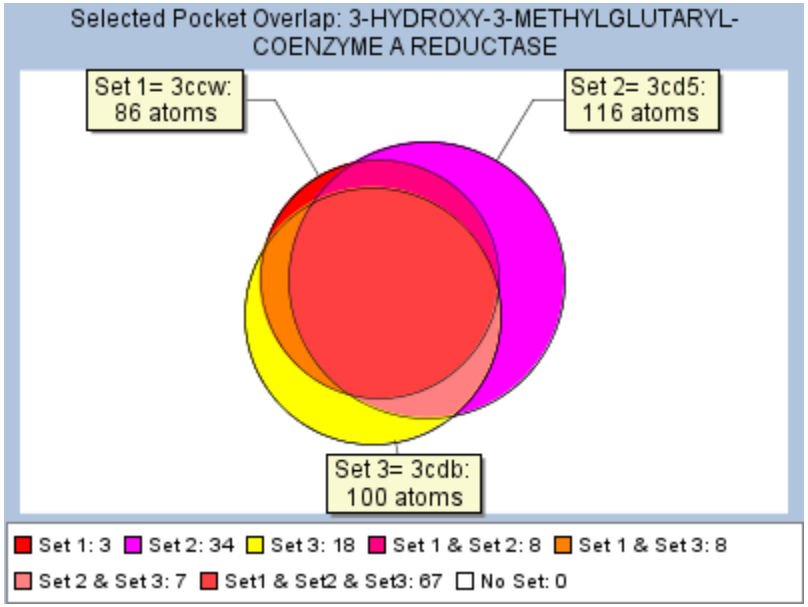

Selected Pocket Overlap: OLIGO-PEPTIDE BINDING PROTEIN
Set 1= 1b7h:
105 atoms
Set 2= 1b9j:
107 atoms
Set 3= 2rkm:
82 atoms
Set 1: 6
Set 2: 5
Set 3: 5
Set 1 & Set 2: 26
Set 1 & Set 3: 1
Set 2 & Set 3: 4
Set1 & Set2 & Set3: 72
No Set: 0

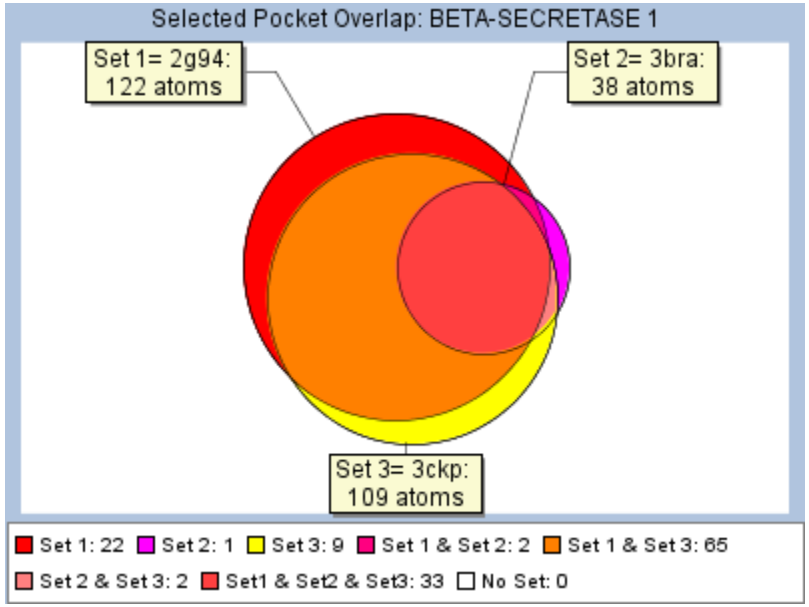

Selected Pocket Overlap: ALPHA-AMYLASE
Set 1= 1u33:
87 atoms
Set 2= 1xd1:
126 atoms
Set 3= 2qv4:
121 atoms
Set 1: 2
Set 2: 26
Set 3: 17
Set 1 & Set 2: 4
Set 1 & Set 3: 8
Set 2 & Set 3: 23
Set1 & Set2 & Set3: 73
No Set: 0

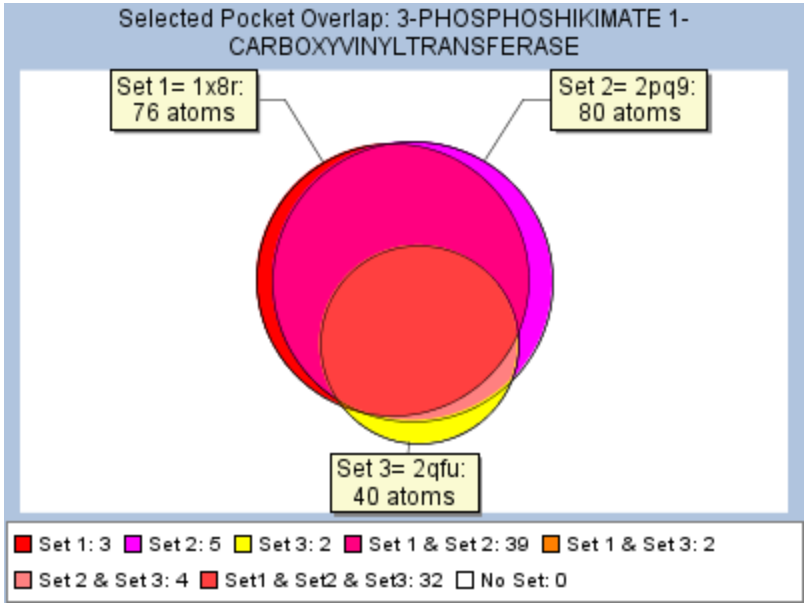

Selected Pocket Overlap: BETA-GALACTOSIDASE
Set 1= 1uwt:
55 atoms
Set 2= 2ceq:
58 atoms
Set 3= 2cer:
91 atoms
Set 1: 0
Set 2: 0
Set 3: 33
Set 1 & Set 2: 1
Set 1 & Set 3: 1
Set 2 & Set 3: 4
Set1 & Set2 & Set3: 53
No Set: 0

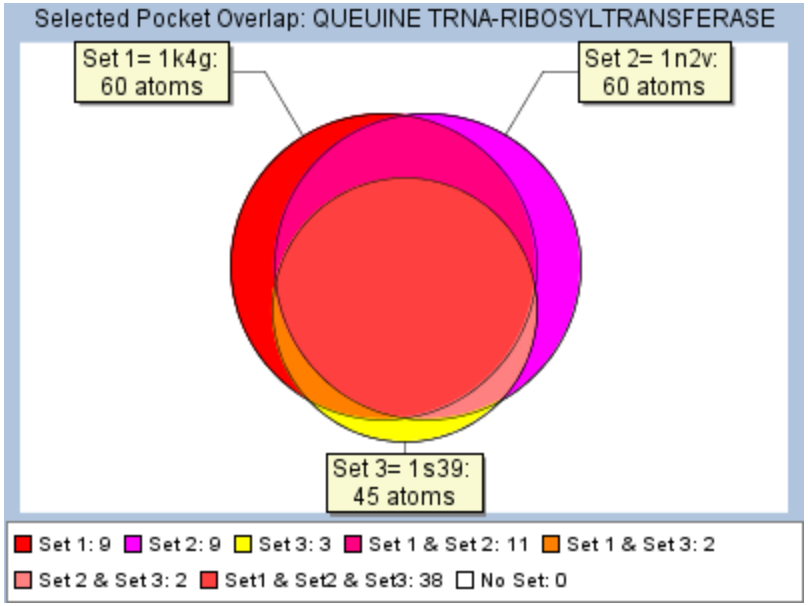

Selected Pocket Overlap: NEURAMINIDASE
Set 1= 2qwb: 53 atoms
Set 2= 2qwd: 52 atoms
Set 3= 2qwe: 58 atoms
Set 1: 2
Set 2: 0
Set 3: 5
Set 1 & Set 2: 0
Set 1 & Set 3: 1
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 50
No Set: 0

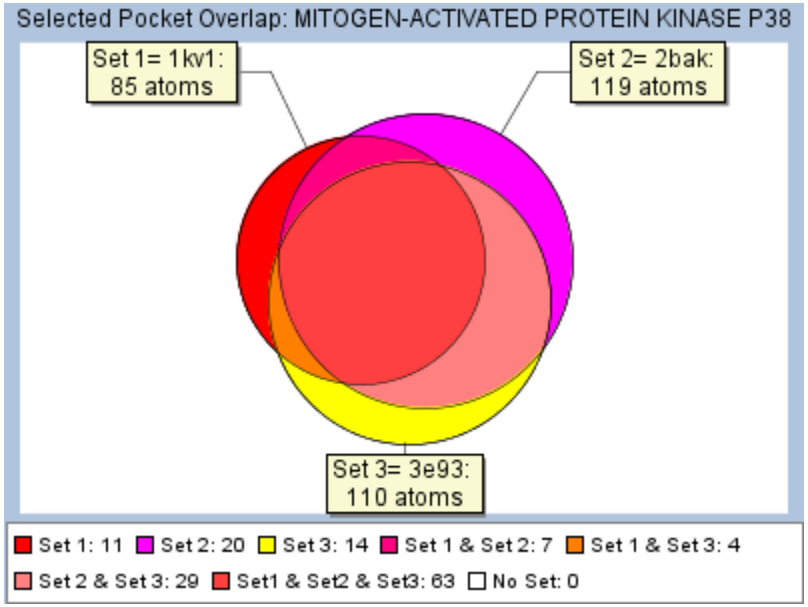

Selected Pocket Overlap: BETA-GLUCOSIDASE A
Set 1= 2cet: 79 atoms
Set 2= 2j77: 51 atoms
Set 3= 2j78: 56 atoms
Set 1: 22
Set 2: 4
Set 3: 1
Set 1 & Set 2: 2
Set 1 & Set 3: 10
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 45
No Set: 0

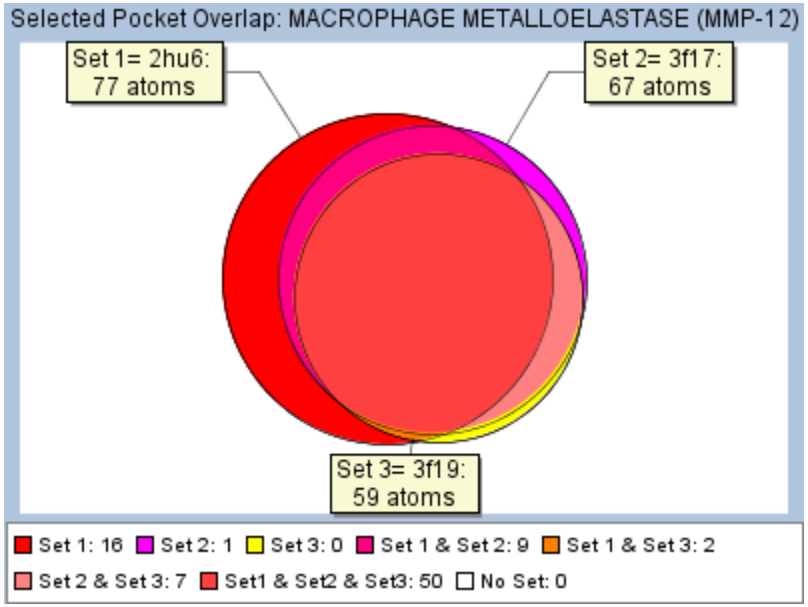

Selected Pocket Overlap: 3-HYDROXY-3-METHYLGLUTARYL-COENZYME A REDUCTASE
Set 1= 3ccw: 86 atoms
Set 2= 3cd5: 116 atoms
Set 3= 3cdb: 100 atoms
Set 1: 3
Set 2: 34
Set 3: 18
Set 1 & Set 2: 8
Set 1 & Set 3: 8
Set 2 & Set 3: 7
Set1 & Set2 & Set3: 67
No Set: 0

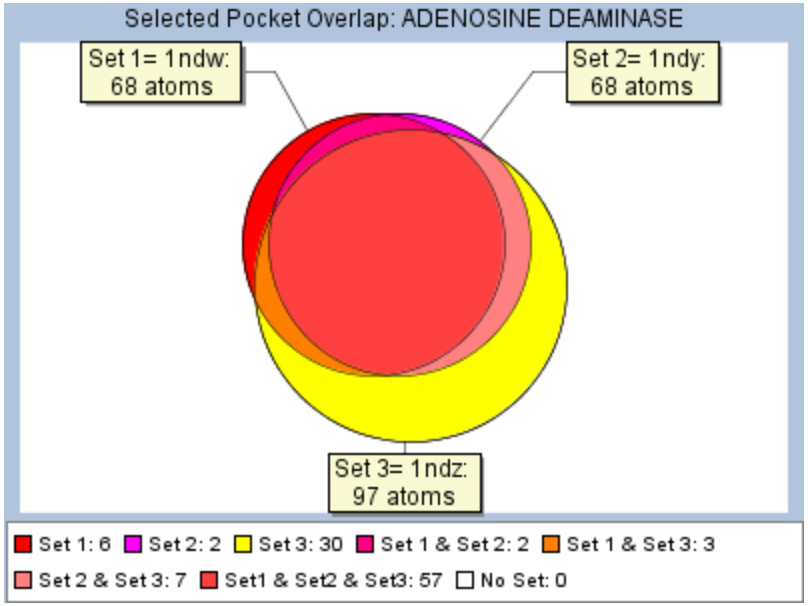

Selected Pocket Overlap: BETA-SECRETASE 1
Set 1= 2g94: 122 atoms
Set 2= 3bra: 38 atoms
Set 3= 3ckp: 109 atoms
Set 1: 22
Set 2: 1
Set 3: 9
Set 1 & Set 2: 2
Set 1 & Set 3: 65
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 33
No Set: 0

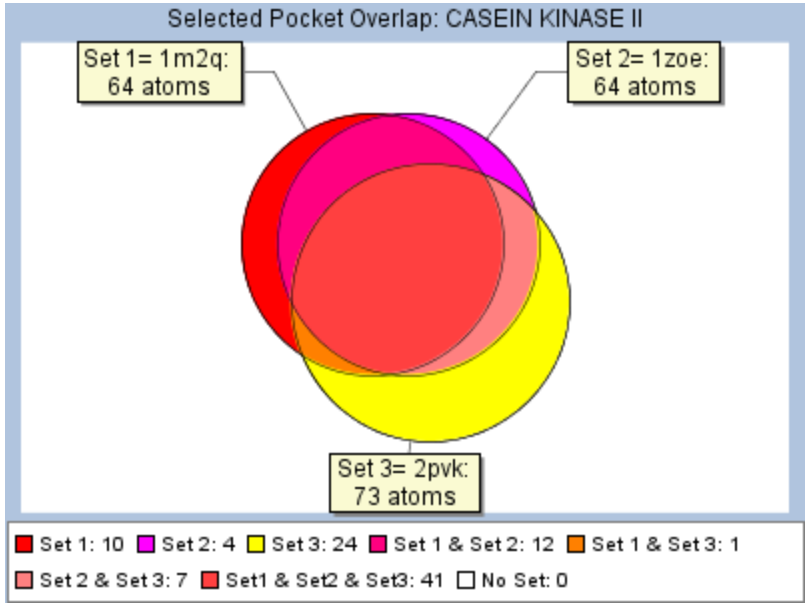

Selected Pocket Overlap: 3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE
Set 1= 1x8r: 76 atoms
Set 2= 2pq9: 80 atoms
Set 3= 2qfu: 40 atoms
Set 1: 3
Set 2: 5
Set 3: 2
Set 1 & Set 2: 39
Set 1 & Set 3: 2
Set 2 & Set 3: 4
Set1 & Set2 & Set3: 32
No Set: 0

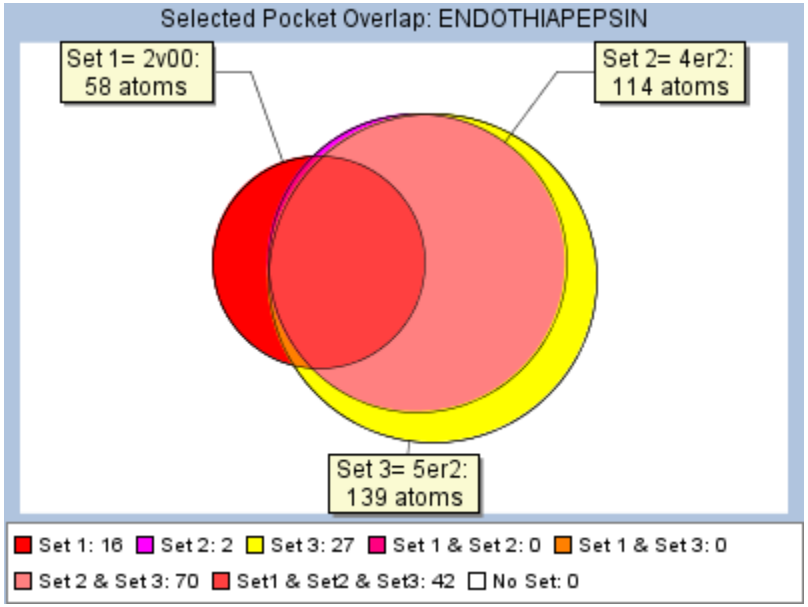

Selected Pocket Overlap: QUEUINE TRNA-RIBOSYLTRANSFERASE
Set 1= 1k4g: 60 atoms
Set 2= 1n2v: 60 atoms
Set 3= 1s39: 45 atoms
Set 1: 9
Set 2: 9
Set 3: 3
Set 1 & Set 2: 11
Set 1 & Set 3: 2
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 38
No Set: 0

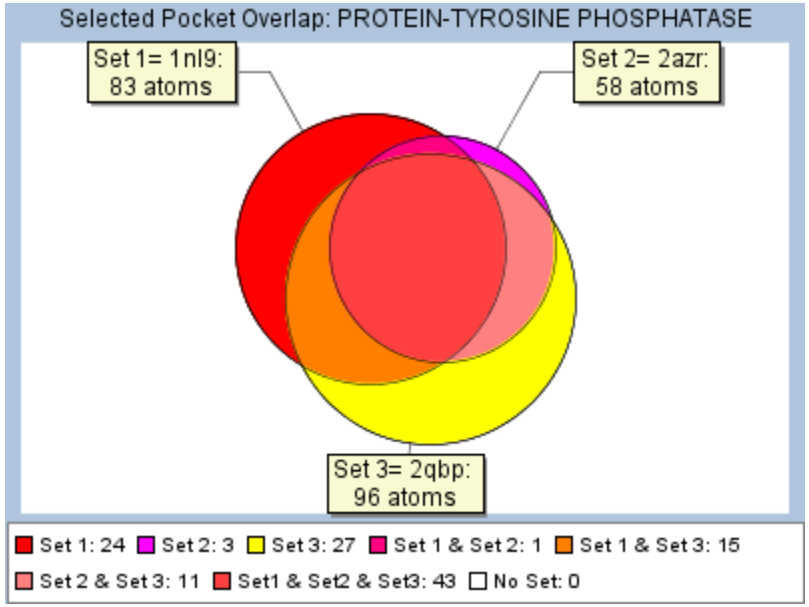

Selected Pocket Overlap: MITOGEN-ACTIVATED PROTEIN KINASE P38
Set 1= 1kv1: 85 atoms
Set 2= 2bak: 119 atoms
Set 3= 3e93: 110 atoms
Set 1: 11
Set 2: 20
Set 3: 14
Set 1 & Set 2: 7
Set 1 & Set 3: 4
Set 2 & Set 3: 29
Set1 & Set2 & Set3: 63
No Set: 0

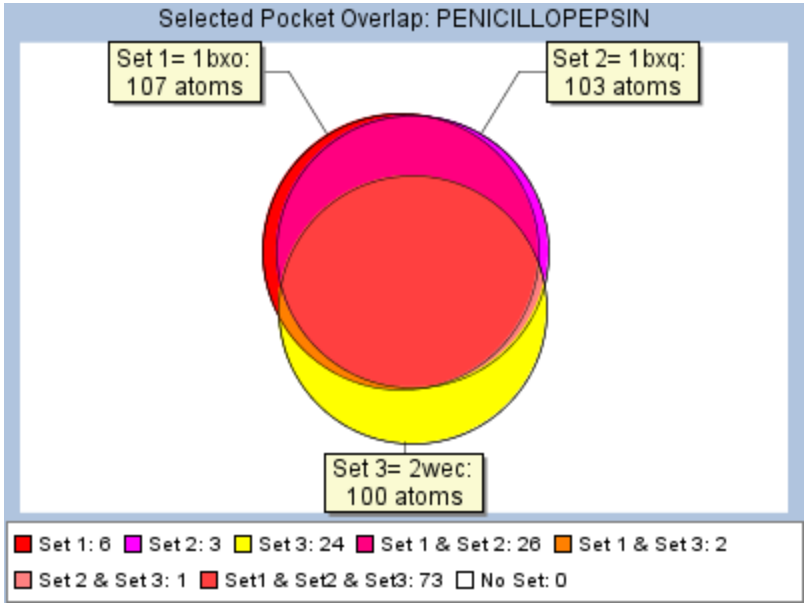

Selected Pocket Overlap: MACROPHAGE METALLOELASTASE (MMP-12)
Set 1= 2hu6: 77 atoms
Set 2= 3f17: 67 atoms
Set 3= 3f19: 59 atoms
Set 1: 16
Set 2: 1
Set 3: 0
Set 1 & Set 2: 9
Set 1 & Set 3: 2
Set 2 & Set 3: 7
Set1 & Set2 & Set3: 50
No Set: 0

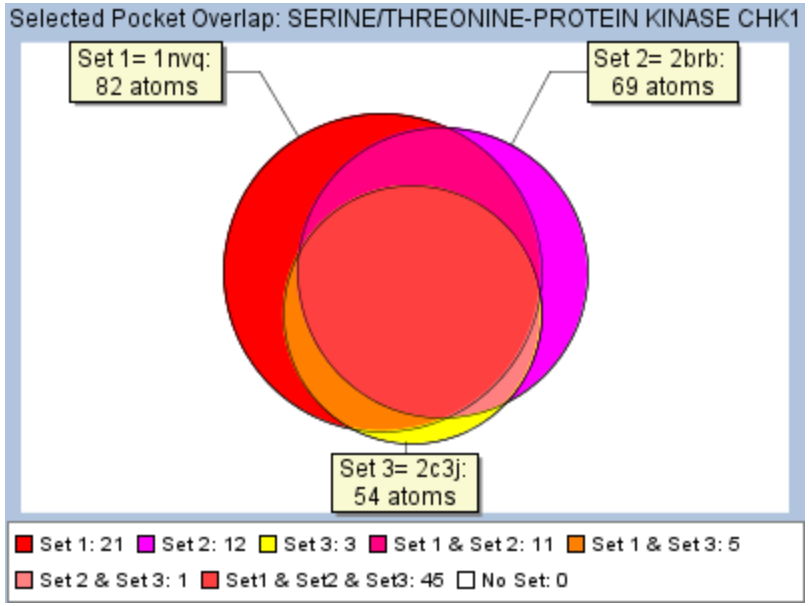

Selected Pocket Overlap: ADENOSINE DEAMINASE
Set 1= 1ndw: 68 atoms
Set 2= 1ndy: 68 atoms
Set 3= 1ndz: 97 atoms
Set 1: 6
Set 2: 2
Set 3: 30
Set 1 & Set 2: 2
Set 1 & Set 3: 3
Set 2 & Set 3: 7
Set1 & Set2 & Set3: 57
No Set: 0

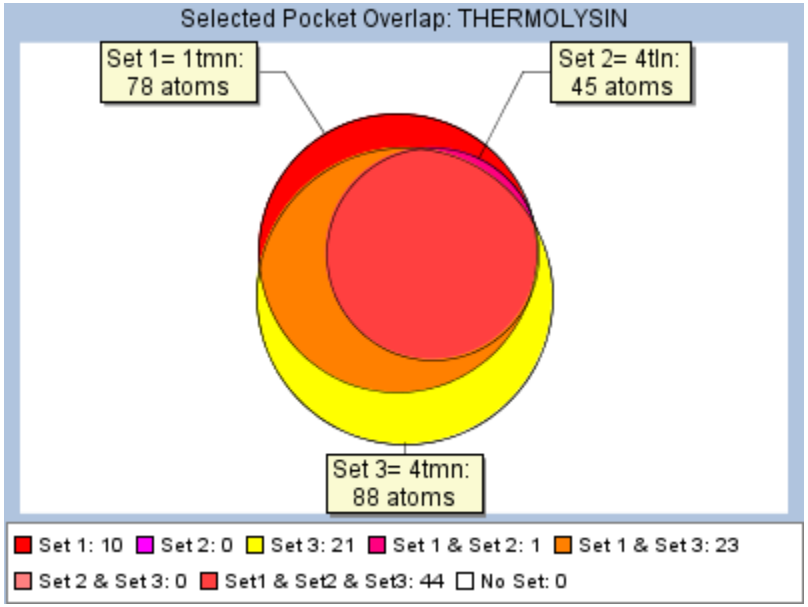

Selected Pocket Overlap: CASEIN KINASE II
Set 1= 1m2q: 64 atoms
Set 2= 1zoe: 64 atoms
Set 3= 2pvk: 73 atoms
Set 1: 10
Set 2: 4
Set 3: 24
Set 1 & Set 2: 12
Set 1 & Set 3: 1
Set 2 & Set 3: 7
Set1 & Set2 & Set3: 41
No Set: 0

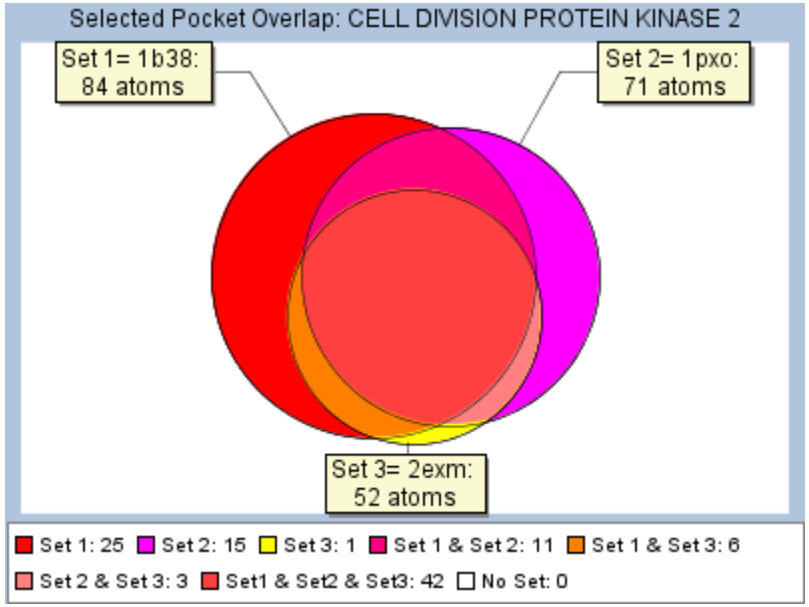

Selected Pocket Overlap: ENDOTHIAPEPSIN
Set 1= 2v00: 58 atoms
Set 2= 4er2: 114 atoms
Set 3= 5er2: 139 atoms
Set 1: 16
Set 2: 2
Set 3: 27
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 70
Set1 & Set2 & Set3: 42
No Set: 0

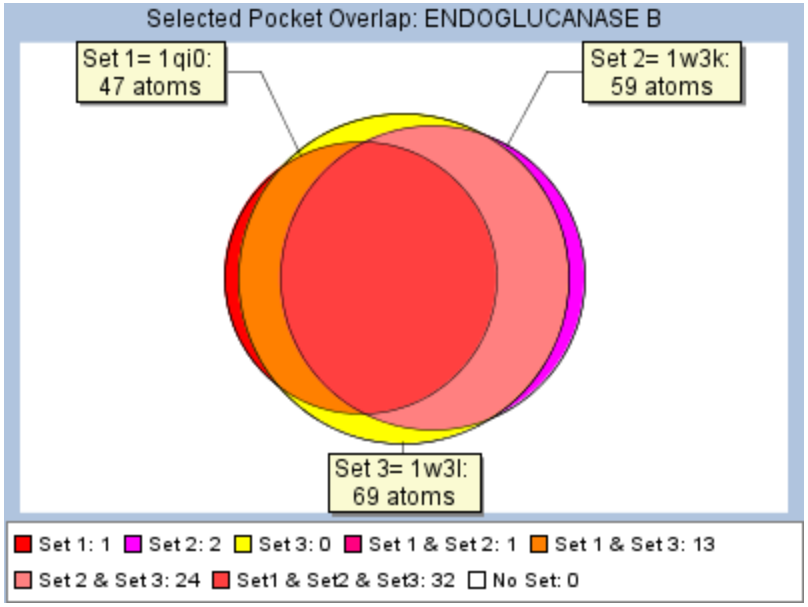

Selected Pocket Overlap: PROTEIN-TYROSINE PHOSPHATASE
Set 1= 1nl9: 83 atoms
Set 2= 2azr: 58 atoms
Set 3= 2qbp: 96 atoms
Set 1: 24
Set 2: 3
Set 3: 27
Set 1 & Set 2: 1
Set 1 & Set 3: 15
Set 2 & Set 3: 11
Set1 & Set2 & Set3: 43
No Set: 0

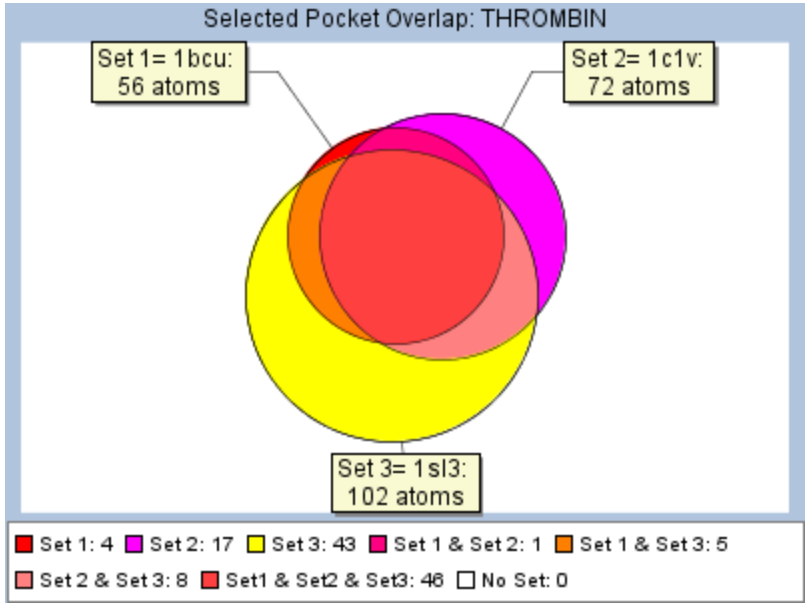

Selected Pocket Overlap: PENICILLOPEPSIN
Set 1= 1bxo:
107 atoms
Set 2= 1bxq:
103 atoms
Set 3= 2wec:
100 atoms
Set 1: 6
Set 2: 3
Set 3: 24
Set 1 & Set 2: 26
Set 1 & Set 3: 2
Set 2 & Set 3: 1
Set1 & Set2 & Set3: 73
No Set: 0

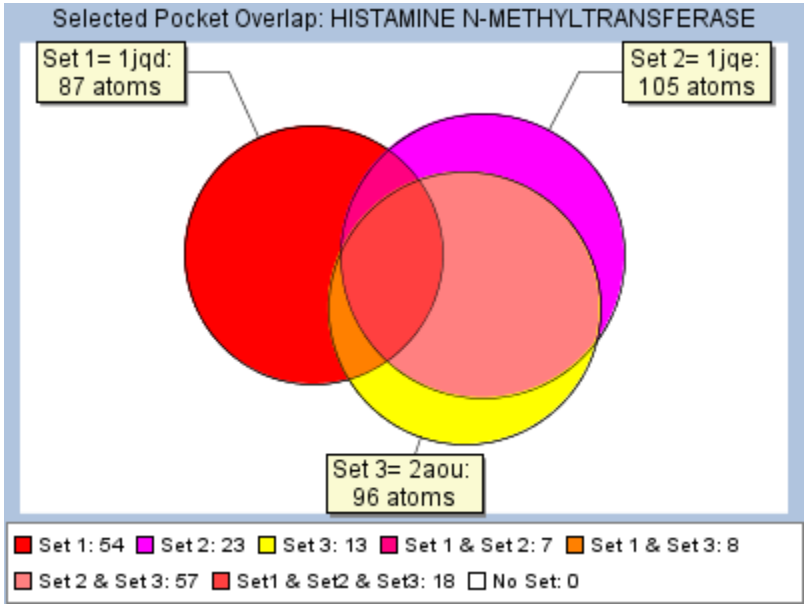

Selected Pocket Overlap: SERINE/THREONINE-PROTEIN KINASE CHK1
Set 1= 1nvq:
82 atoms
Set 2= 2brb:
69 atoms
Set 3= 2c3j:
54 atoms
Set 1: 21
Set 2: 12
Set 3: 3
Set 1 & Set 2: 11
Set 1 & Set 3: 5
Set 2 & Set 3: 1
Set1 & Set2 & Set3: 45
No Set: 0

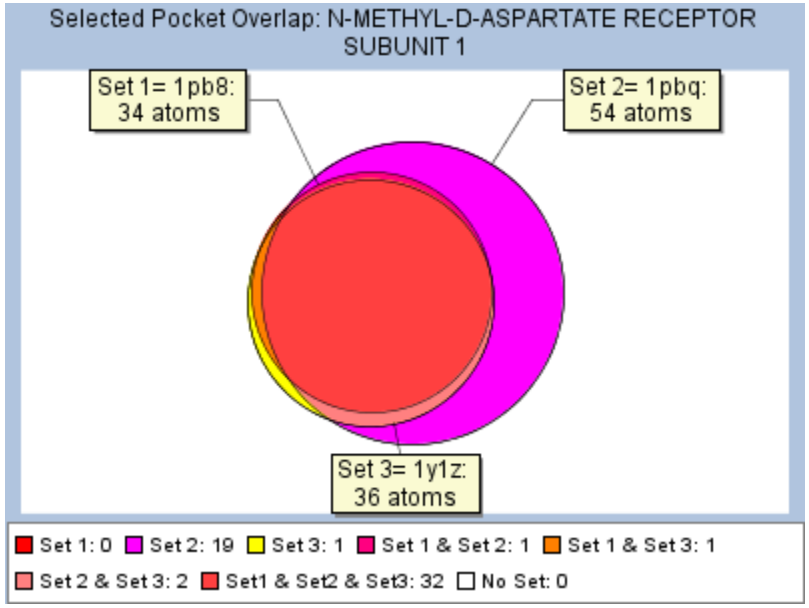

Selected Pocket Overlap: THERMOLYSIN
Set 1= 1tmn:
78 atoms
Set 2= 4tln:
45 atoms
Set 3= 4tmn:
88 atoms
Set 1: 10
Set 2: 0
Set 3: 21
Set 1 & Set 2: 1
Set 1 & Set 3: 23
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 44
No Set: 0

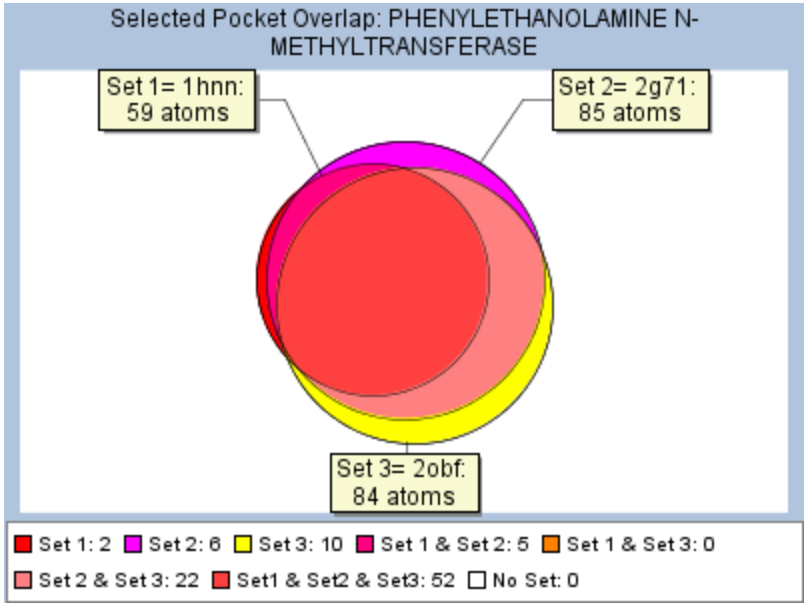

Selected Pocket Overlap: CELL DIVISION PROTEIN KINASE 2
Set 1= 1b38:
84 atoms
Set 2= 1pxo:
71 atoms
Set 3= 2exm:
52 atoms
Set 1: 25
Set 2: 15
Set 3: 1
Set 1 & Set 2: 11
Set 1 & Set 3: 6
Set 2 & Set 3: 3
Set1 & Set2 & Set3: 42
No Set: 0

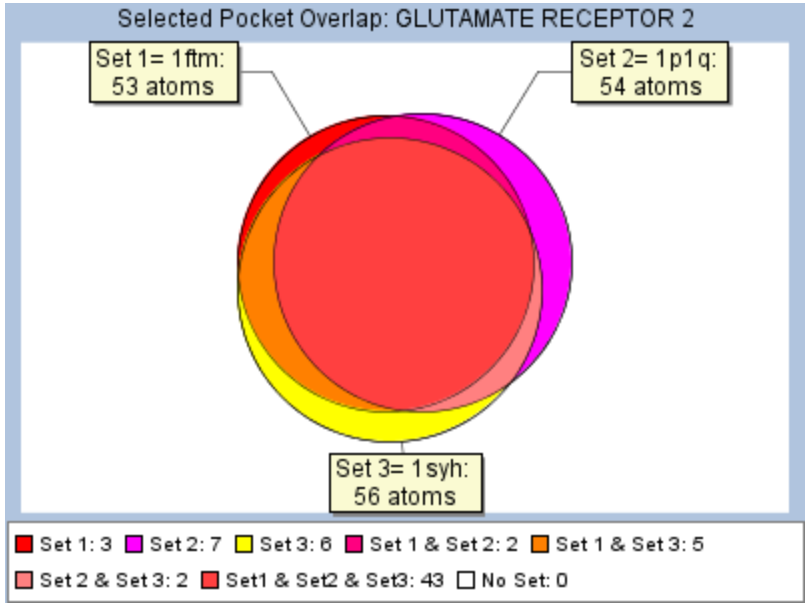

Selected Pocket Overlap: ENDOGLUCANASE B
Set 1= 1qi0:
47 atoms
Set 2= 1w3k:
59 atoms
Set 3= 1w3l:
69 atoms
Set 1: 1
Set 2: 2
Set 3: 0
Set 1 & Set 2: 1
Set 1 & Set 3: 13
Set 2 & Set 3: 24
Set1 & Set2 & Set3: 32
No Set: 0

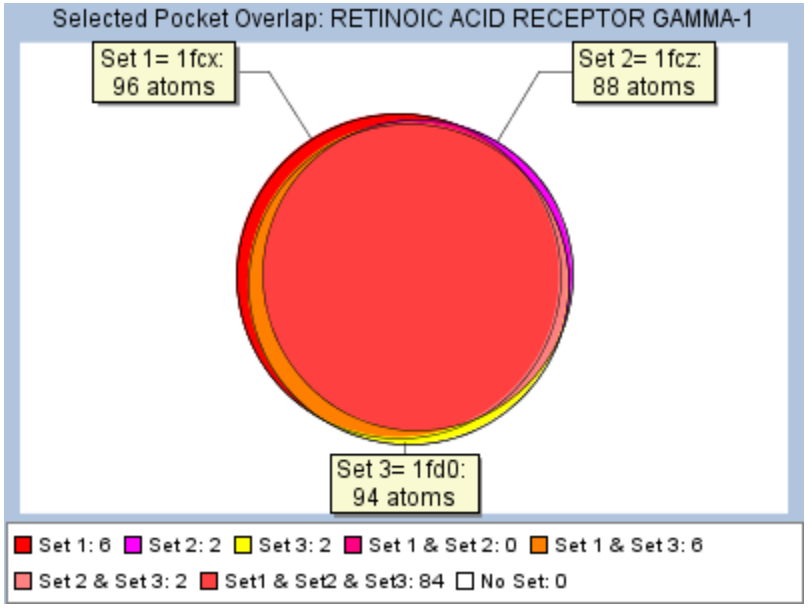

Selected Pocket Overlap: THROMBIN
Set 1= 1bcu:
56 atoms
Set 2= 1c1v:
72 atoms
Set 3= 1sl3:
102 atoms
Set 1: 4
Set 2: 17
Set 3: 43
Set 1 & Set 2: 1
Set 1 & Set 3: 5
Set 2 & Set 3: 8
Set1 & Set2 & Set3: 46
No Set: 0

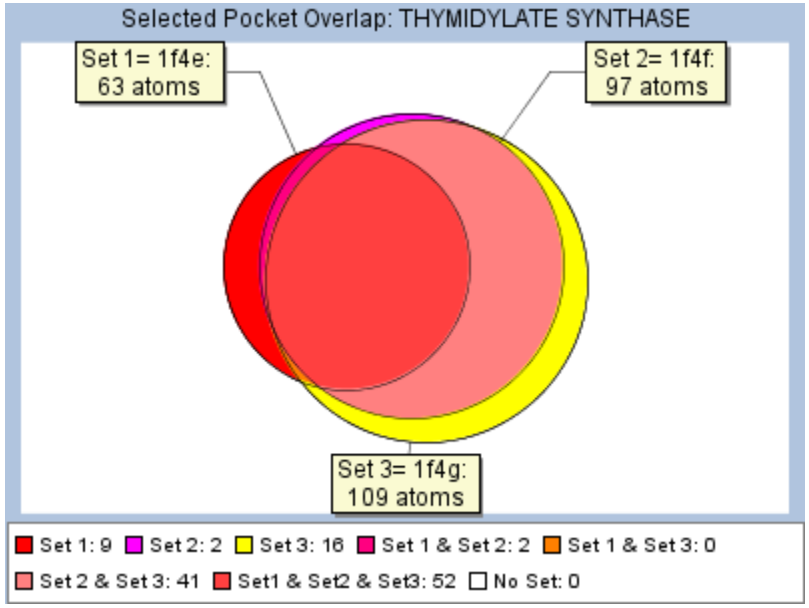

Selected Pocket Overlap: HISTAMINE N-METHYLTRANSFERASE
Set 1= 1jqd:
87 atoms
Set 2= 1jqe:
105 atoms
Set 3= 2aou:
96 atoms
Set 1: 54
Set 2: 23
Set 3: 13
Set 1 & Set 2: 7
Set 1 & Set 3: 8
Set 2 & Set 3: 57
Set1 & Set2 & Set3: 18
No Set: 0

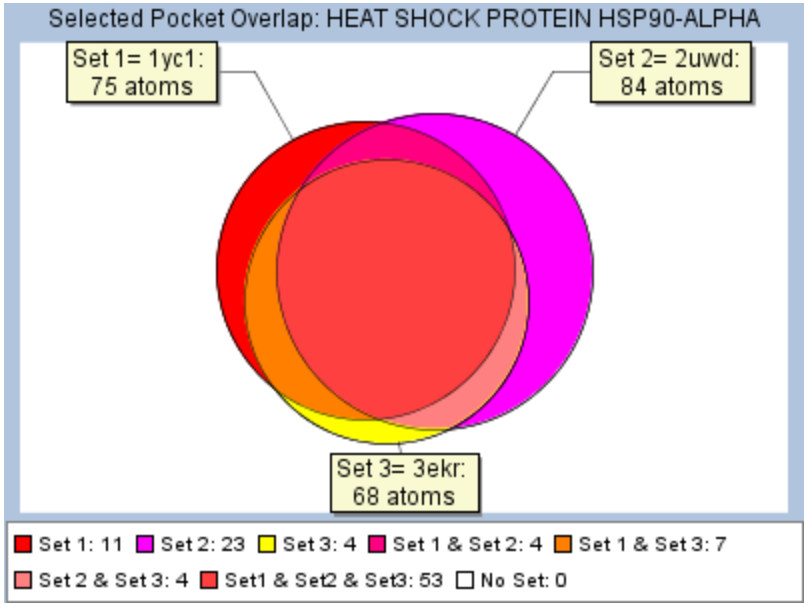

Selected Pocket Overlap: N-METHYL-D-ASPARTATE RECEPTOR SUBUNIT 1
Set 1= 1pb8:
34 atoms
Set 2= 1pbq:
54 atoms
Set 3= 1y1z:
36 atoms
Set 1: 0
Set 2: 19
Set 3: 1
Set 1 & Set 2: 1
Set 1 & Set 3: 1
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 32
No Set: 0

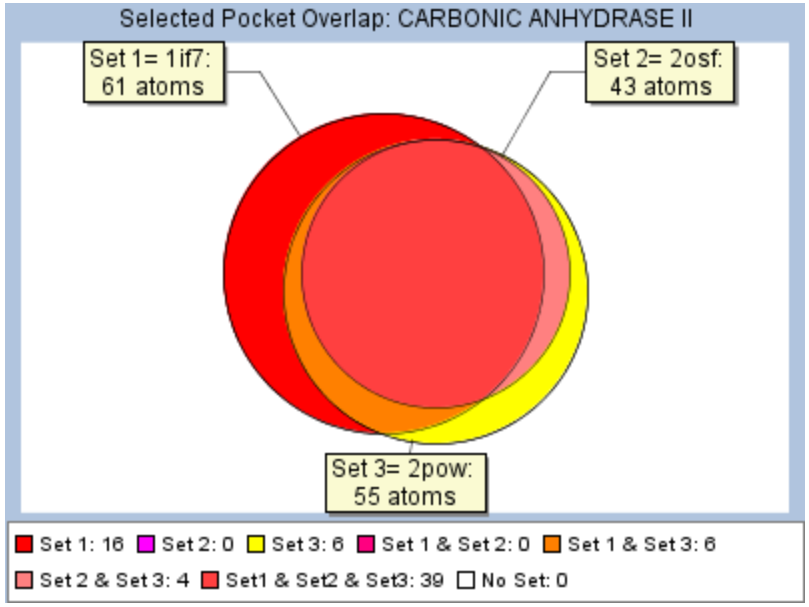

Selected Pocket Overlap: PHENYLETHANOLAMINE N-METHYLTRANSFERASE
Set 1= 1hnn: 59 atoms
Set 2= 2g71: 85 atoms
Set 3= 2obf: 84 atoms
Set 1: 2
Set 2: 6
Set 3: 10
Set 1 & Set 2: 5
Set 1 & Set 3: 0
Set 2 & Set 3: 22
Set1 & Set2 & Set3: 52
No Set: 0

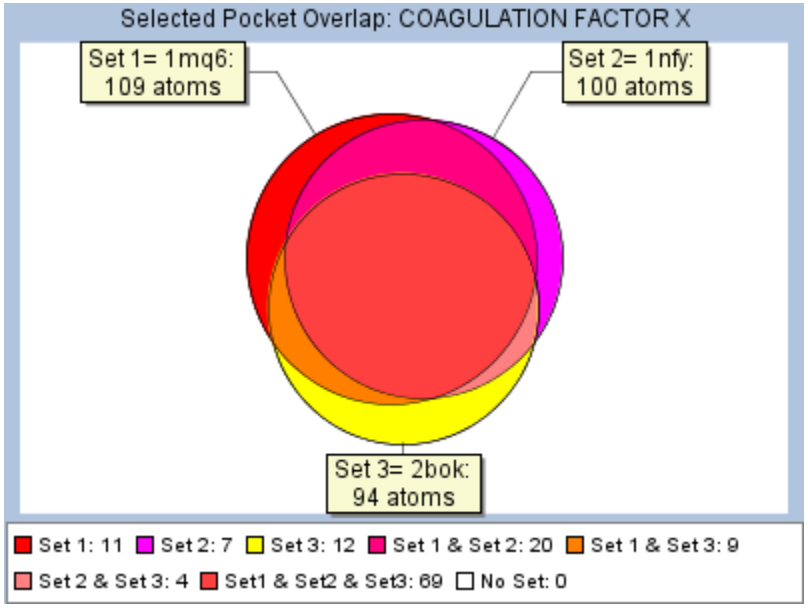

Selected Pocket Overlap: GLUTAMATE RECEPTOR 2
Set 1= 1ftm: 53 atoms
Set 2= 1p1q: 54 atoms
Set 3= 1syh: 56 atoms
Set 1: 3
Set 2: 7
Set 3: 6
Set 1 & Set 2: 2
Set 1 & Set 3: 5
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 43
No Set: 0

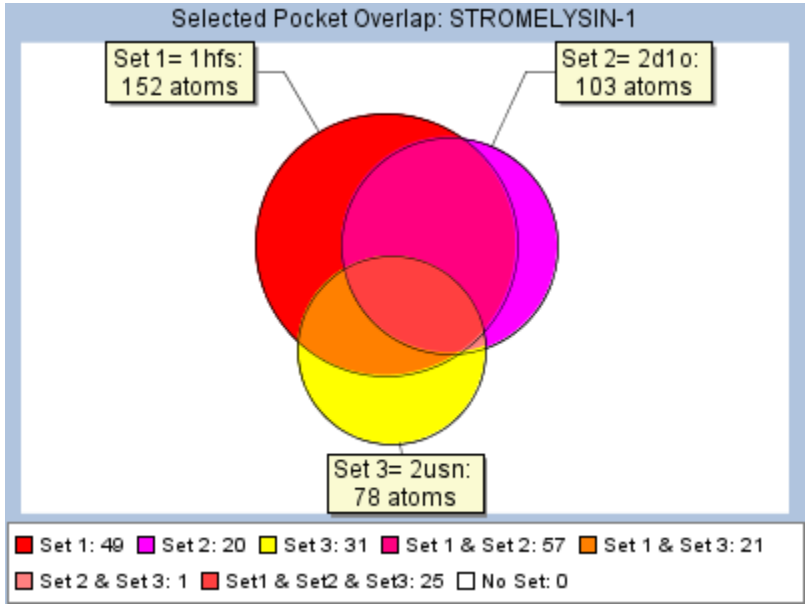

Selected Pocket Overlap: RETINOIC ACID RECEPTOR GAMMA-1
Set 1= 1fcx: 96 atoms
Set 2= 1fcz: 88 atoms
Set 3= 1fd0: 94 atoms
Set 1: 6
Set 2: 2
Set 3: 2
Set 1 & Set 2: 0
Set 1 & Set 3: 6
Set 2 & Set 3: 2
Set1 & Set2 & Set3: 84
No Set: 0

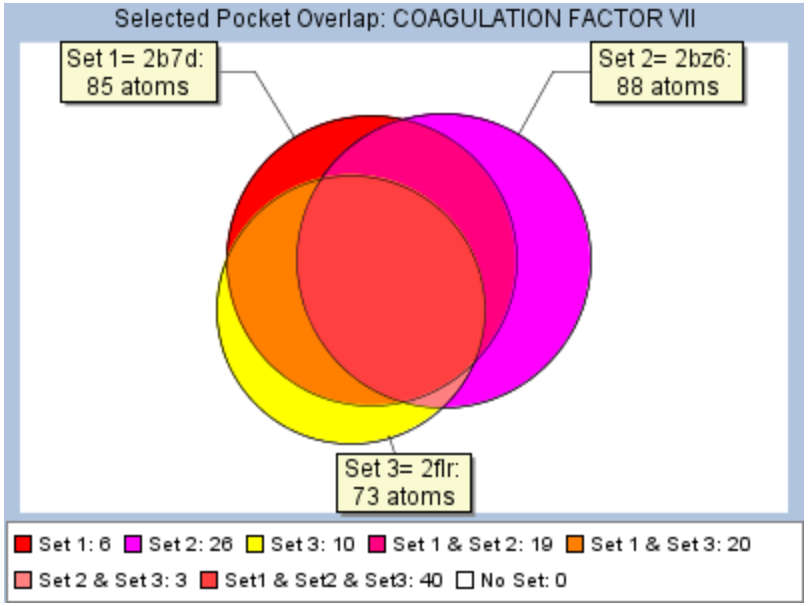

Selected Pocket Overlap: THYMIDYLATE SYNTHASE
Set 1= 1f4e: 63 atoms
Set 2= 1f4f: 97 atoms
Set 3= 1f4g: 109 atoms
Set 1: 9
Set 2: 2
Set 3: 16
Set 1 & Set 2: 2
Set 1 & Set 3: 0
Set 2 & Set 3: 41
Set1 & Set2 & Set3: 52
No Set: 0

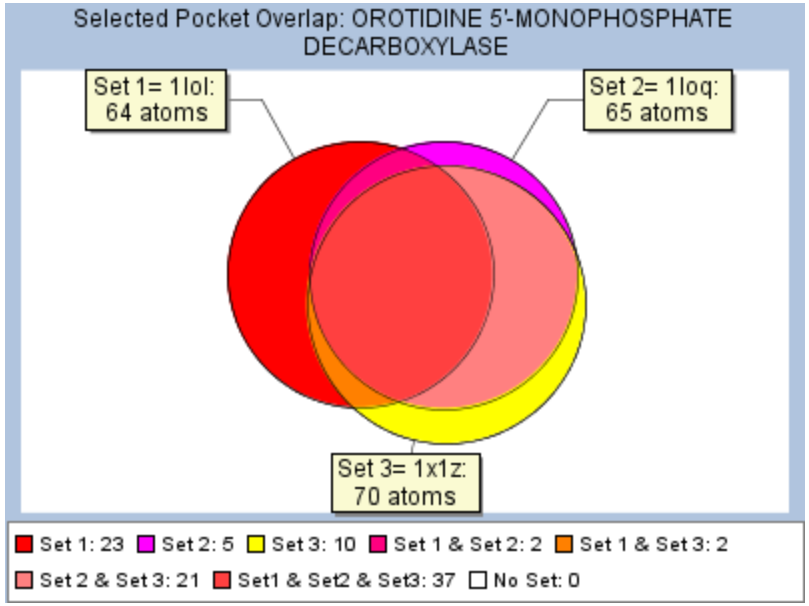

Selected Pocket Overlap: HEAT SHOCK PROTEIN HSP90-ALPHA
Set 1= 1yc1: 75 atoms
Set 2= 2uwd: 84 atoms
Set 3= 3ekr: 68 atoms
Set 1: 11 Set 2: 23 Set 3: 4 Set 1 & Set 2: 4 Set 1 & Set 3: 7
Set 2 & Set 3: 4 Set1 & Set2 & Set3: 53 No Set: 0

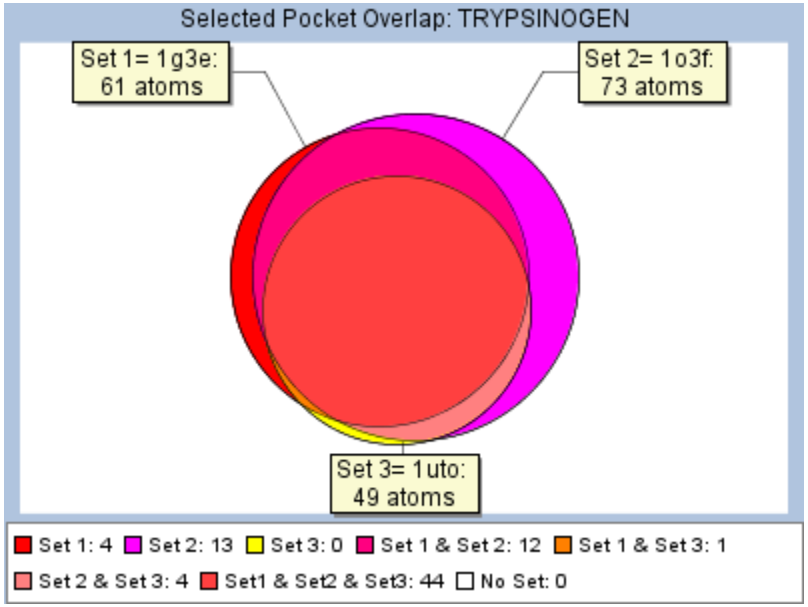

Selected Pocket Overlap: CARBONIC ANHYDRASE II
Set 1= 1if7: 61 atoms
Set 2= 2osf: 43 atoms
Set 3= 2pow: 55 atoms
Set 1: 16 Set 2: 0 Set 3: 6 Set 1 & Set 2: 0 Set 1 & Set 3: 6
Set 2 & Set 3: 4 Set1 & Set2 & Set3: 39 No Set: 0

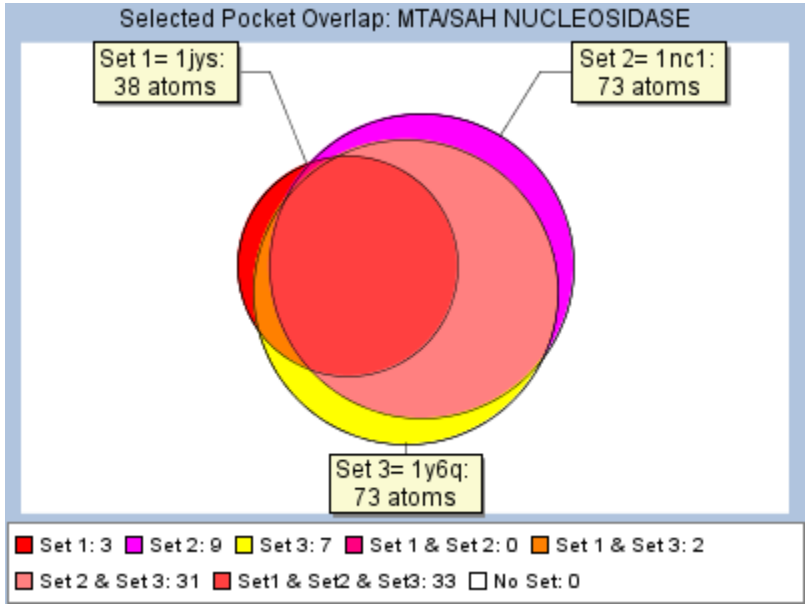

Selected Pocket Overlap: COAGULATION FACTOR X
Set 1= 1mq6: 109 atoms
Set 2= 1nfy: 100 atoms
Set 3= 2bok: 94 atoms
Set 1: 11 Set 2: 7 Set 3: 12 Set 1 & Set 2: 20 Set 1 & Set 3: 9
Set 2 & Set 3: 4 Set1 & Set2 & Set3: 69 No Set: 0

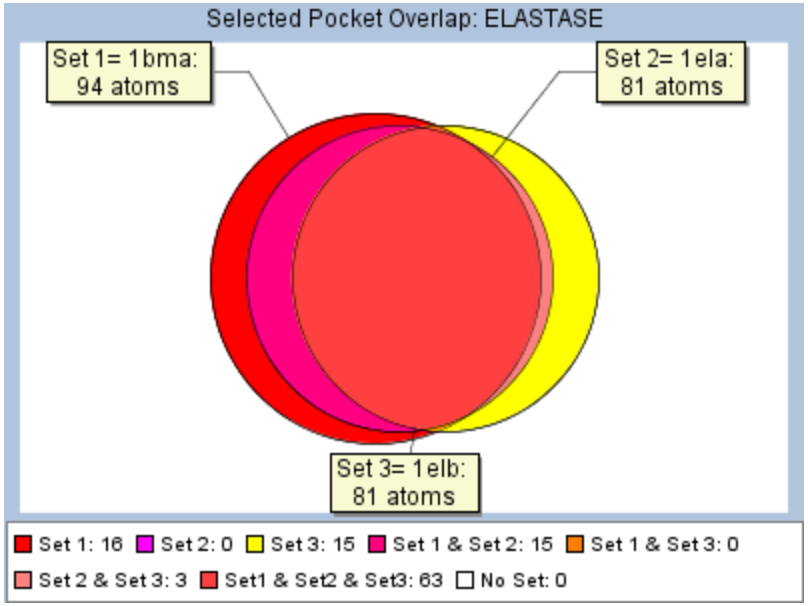

Selected Pocket Overlap: STROMELYSIN-1
Set 1= 1hfs: 152 atoms
Set 2= 2d1o: 103 atoms
Set 3= 2usn: 78 atoms
Set 1: 49 Set 2: 20 Set 3: 31 Set 1 & Set 2: 57 Set 1 & Set 3: 21
Set 2 & Set 3: 1 Set1 & Set2 & Set3: 25 No Set: 0

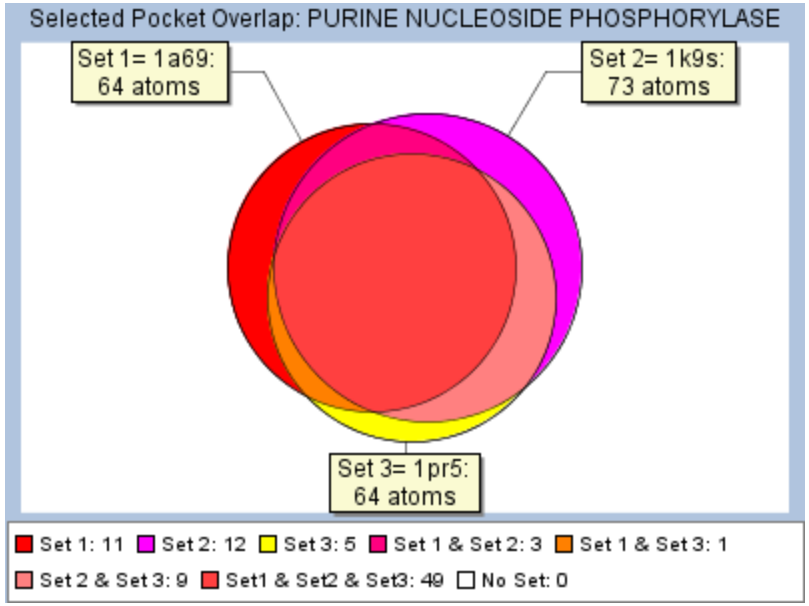

Selected Pocket Overlap: COAGULATION FACTOR VII
Set 1= 2b7d: 85 atoms
Set 2= 2bz6: 88 atoms
Set 3= 2flr: 73 atoms
Set 1: 6 Set 2: 26 Set 3: 10 Set 1 & Set 2: 19 Set 1 & Set 3: 20
Set 2 & Set 3: 3 Set1 & Set2 & Set3: 40 No Set: 0

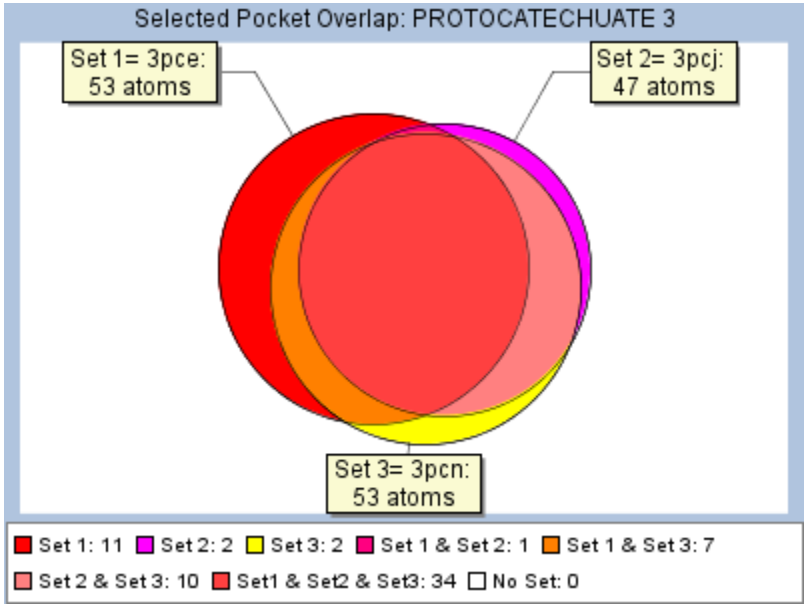

Selected Pocket Overlap: OROTIDINE 5'-MONOPHOSPHATE DECARBOXYLASE
Set 1= 1lol: 64 atoms
Set 2= 1loq: 65 atoms
Set 3= 1x1z: 70 atoms
Set 1: 23 Set 2: 5 Set 3: 10 Set 1 & Set 2: 2 Set 1 & Set 3: 2
Set 2 & Set 3: 21 Set1 & Set2 & Set3: 37 No Set: 0

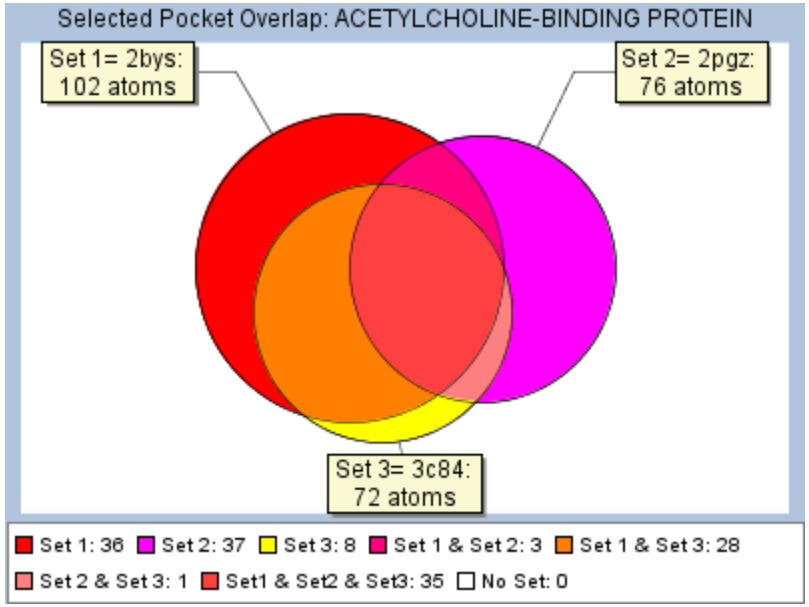

Selected Pocket Overlap: TRYPSINOGEN
Set 1= 1g3e: 61 atoms
Set 2= 1o3f: 73 atoms
Set 3= 1uto: 49 atoms
Set 1: 4 Set 2: 13 Set 3: 0 Set 1 & Set 2: 12 Set 1 & Set 3: 1
Set 2 & Set 3: 4 Set1 & Set2 & Set3: 44 No Set: 0

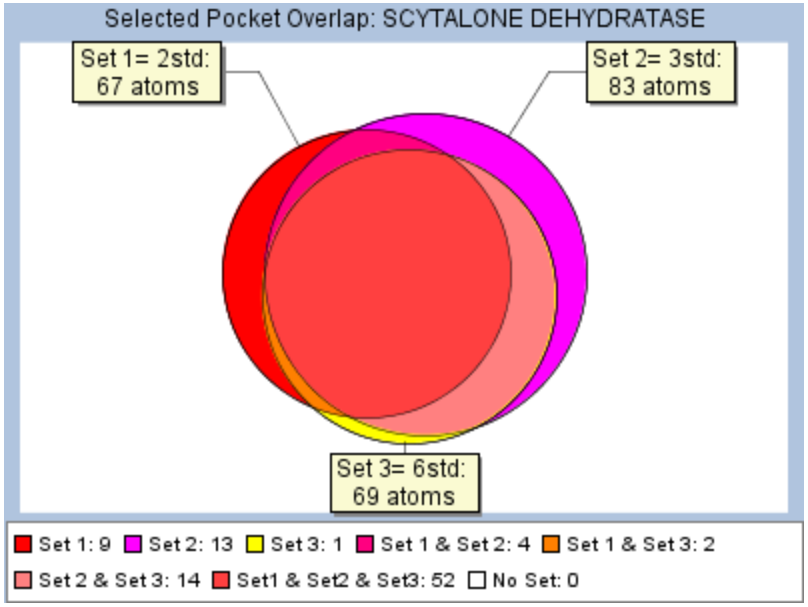

Selected Pocket Overlap: MTA/SAH NUCLEOSIDASE
Set 1= 1jys: 38 atoms
Set 2= 1nc1: 73 atoms
Set 3= 1y6q: 73 atoms
Set 1: 3 Set 2: 9 Set 3: 7 Set 1 & Set 2: 0 Set 1 & Set 3: 2
Set 2 & Set 3: 31 Set1 & Set2 & Set3: 33 No Set: 0

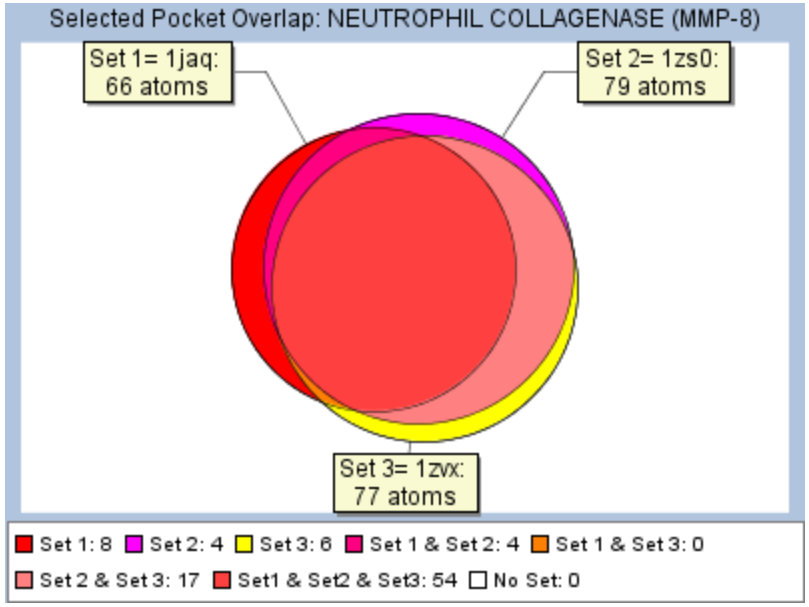

Selected Pocket Overlap: ELASTASE
Set 1= 1bma: 94 atoms
Set 2= 1ela: 81 atoms
Set 3= 1elb: 81 atoms
Set 1: 16 Set 2: 0 Set 3: 15 Set 1 & Set 2: 15 Set 1 & Set 3: 0
Set 2 & Set 3: 3 Set1 & Set2 & Set3: 63 No Set: 0

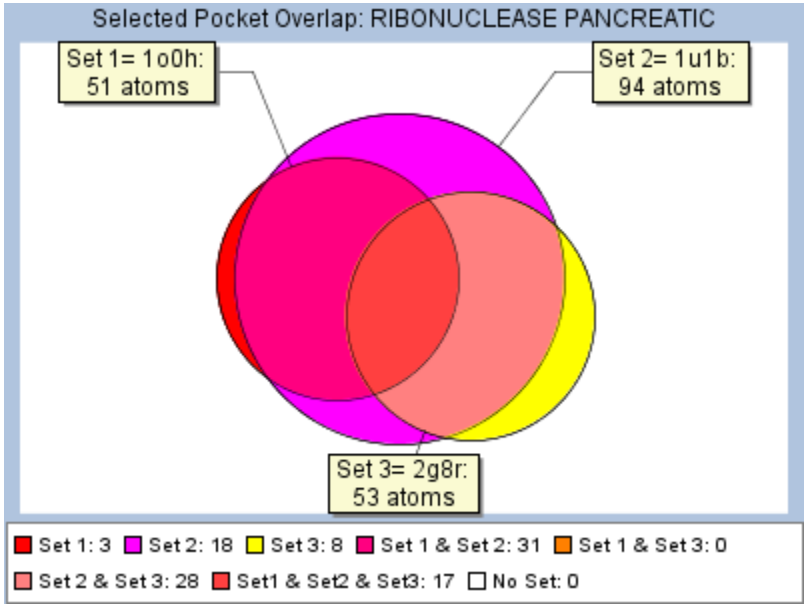

Selected Pocket Overlap: PURINE NUCLEOSIDE PHOSPHORYLASE
Set 1= 1a69: 64 atoms
Set 2= 1k9s: 73 atoms
Set 3= 1pr5: 64 atoms
Set 1: 11 Set 2: 12 Set 3: 5 Set 1 & Set 2: 3 Set 1 & Set 3: 1
Set 2 & Set 3: 9 Set1 & Set2 & Set3: 49 No Set: 0

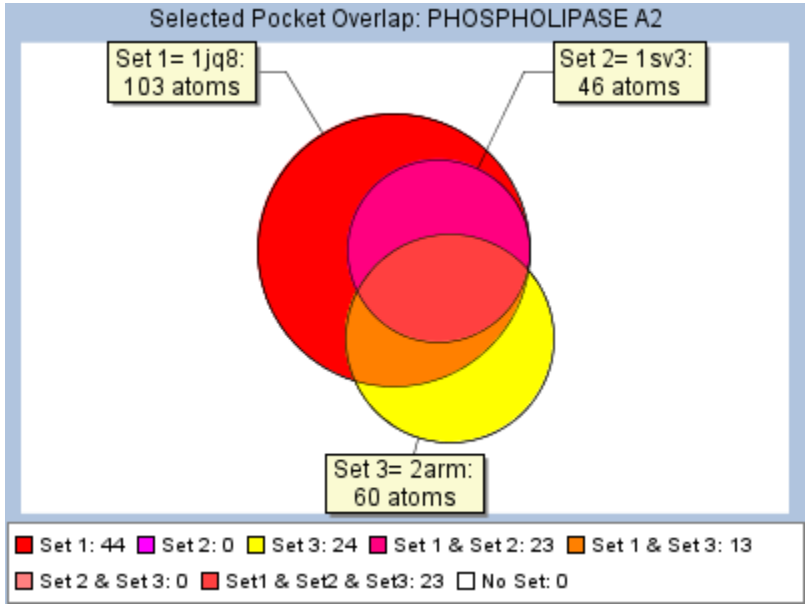

Selected Pocket Overlap: PROTOCATECHUATE 3
Set 1= 3pce: 53 atoms
Set 2= 3pcj: 47 atoms
Set 3= 3pcn: 53 atoms
Set 1: 11 Set 2: 2 Set 3: 2 Set 1 & Set 2: 1 Set 1 & Set 3: 7
Set 2 & Set 3: 10 Set1 & Set2 & Set3: 34 No Set: 0

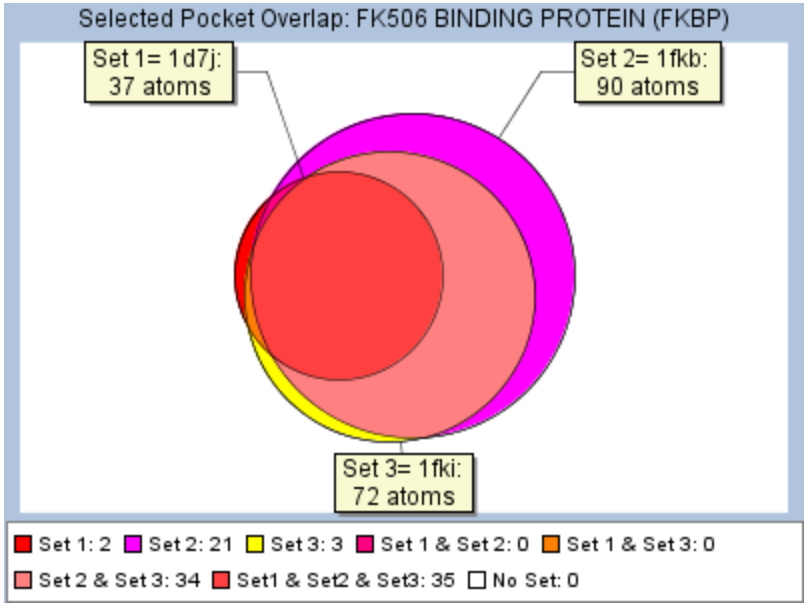

Selected Pocket Overlap: ACETYLCHOLINE-BINDING PROTEIN
Set 1= 2bys: 102 atoms
Set 2= 2pgz: 76 atoms
Set 3= 3c84: 72 atoms
Set 1: 36 Set 2: 37 Set 3: 8 Set 1 & Set 2: 3 Set 1 & Set 3: 28
Set 2 & Set 3: 1 Set1 & Set2 & Set3: 35 No Set: 0


Selected Pocket Overlap: SCYTALONE DEHYDRATASE
Set 1= 2std: 67 atoms
Set 2= 3std: 83 atoms
Set 3= 6std: 69 atoms
Set 1: 9 Set 2: 13 Set 3: 1 Set 1 & Set 2: 4 Set 1 & Set 3: 2
Set 2 & Set 3: 14 Set1 & Set2 & Set3: 52 No Set: 0

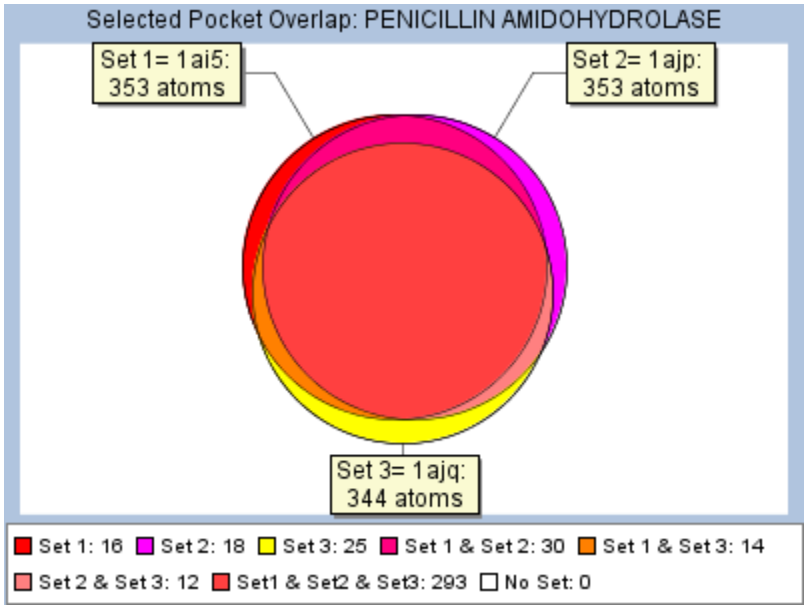

Selected Pocket Overlap: NEUTROPHIL COLLAGENASE (MMP-8)
Set 1= 1jaq: 66 atoms
Set 2= 1zs0: 79 atoms
Set 3= 1zvx: 77 atoms
Set 1: 8 Set 2: 4 Set 3: 6 Set 1 & Set 2: 4 Set 1 & Set 3: 0
Set 2 & Set 3: 17 Set1 & Set2 & Set3: 54 No Set: 0

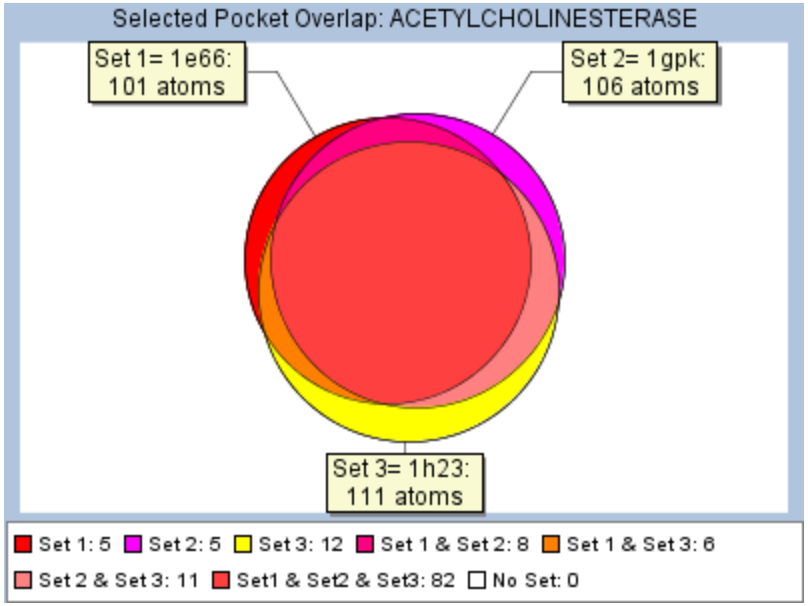

Selected Pocket Overlap: RIBONUCLEASE PANCREATIC
Set 1= 1o0h: 51 atoms
Set 2= 1u1b: 94 atoms
Set 3= 2g8r: 53 atoms
Set 1: 3 Set 2: 18 Set 3: 8 Set 1 & Set 2: 31 Set 1 & Set 3: 0
Set 2 & Set 3: 28 Set1 & Set2 & Set3: 17 No Set: 0

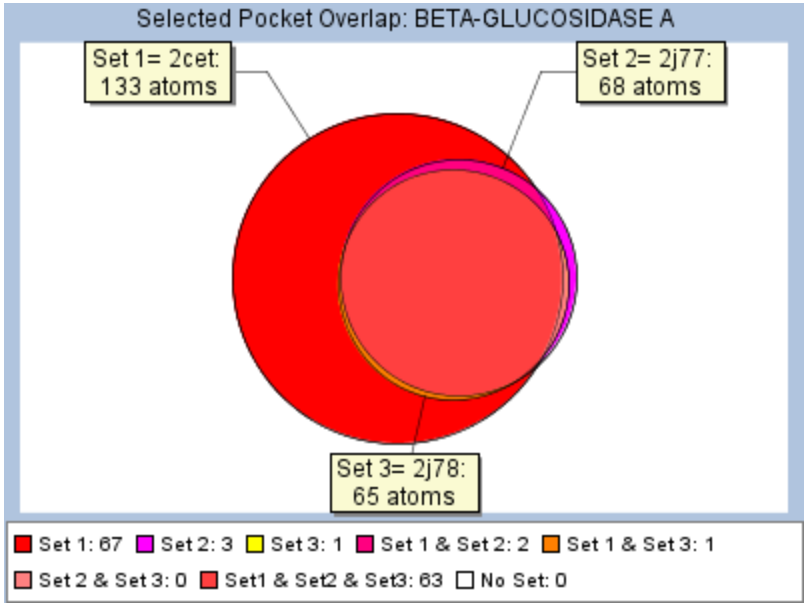

Selected Pocket Overlap: PHOSPHOLIPASE A2
Set 1= 1jq8: 103 atoms
Set 2= 1sv3: 46 atoms
Set 3= 2arm: 60 atoms
Set 1: 44 Set 2: 0 Set 3: 24 Set 1 & Set 2: 23 Set 1 & Set 3: 13
Set 2 & Set 3: 0 Set1 & Set2 & Set3: 23 No Set: 0

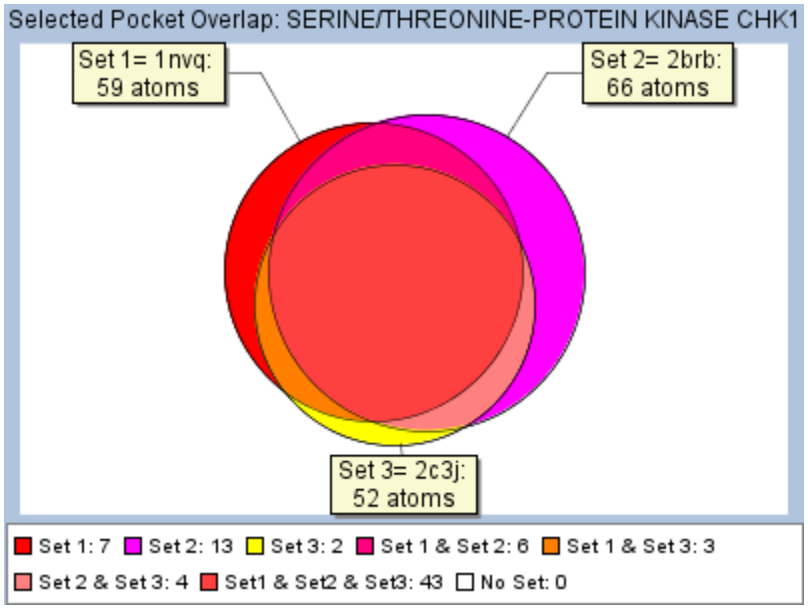

Selected Pocket Overlap: FK506 BINDING PROTEIN (FKBP)
Set 1= 1d7j:
37 atoms
Set 2= 1fkb:
90 atoms
Set 3= 1fki:
72 atoms
Set 1: 2
Set 2: 21
Set 3: 3
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 34
Set1 & Set2 & Set3: 35
No Set: 0

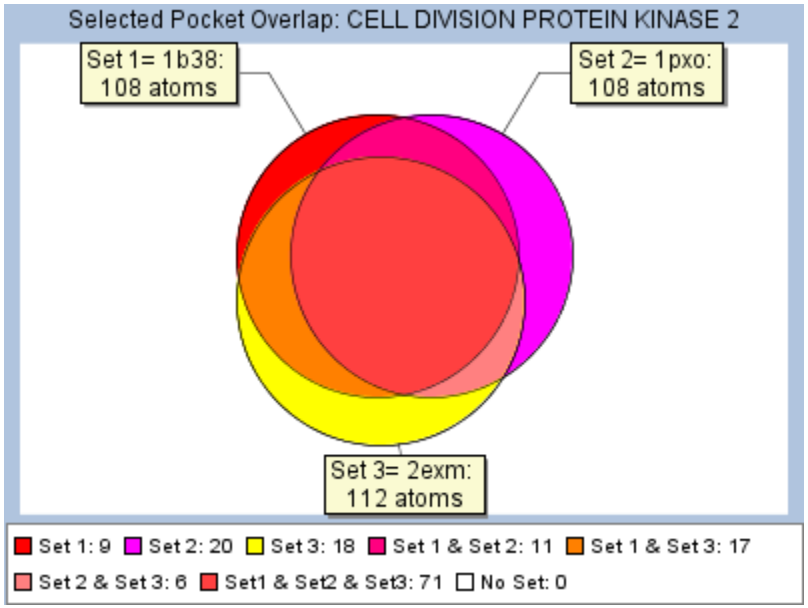
CastP

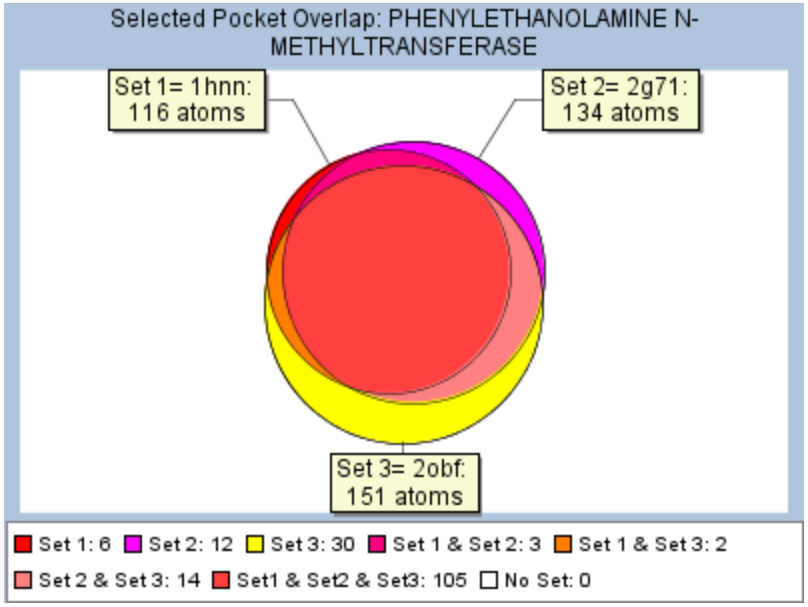

Selected Pocket Overlap: PENICILLIN AMIDOHYDROLASE
Set 1= 1ai5: 353 atoms
Set 2= 1ajp: 353 atoms
Set 3= 1ajq: 344 atoms
Set 1: 16
Set 2: 18
Set 3: 25
Set 1 & Set 2: 30
Set 1 & Set 3: 14
Set 2 & Set 3: 12
Set1 & Set2 & Set3: 293
No Set: 0

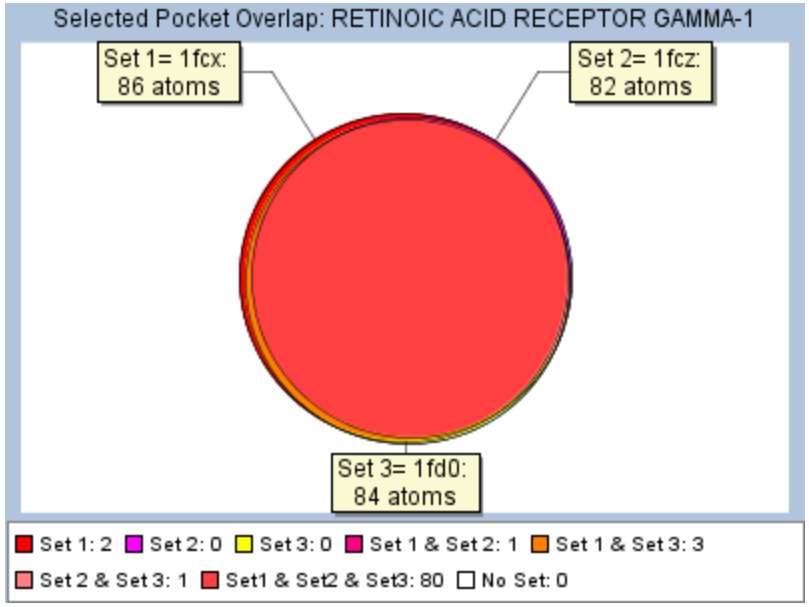

Selected Pocket Overlap: ACETYLCHOLINESTERASE
Set 1= 1e66: 101 atoms
Set 2= 1gpk: 106 atoms
Set 3= 1h23: 111 atoms
Set 1: 5
Set 2: 5
Set 3: 12
Set 1 & Set 2: 8
Set 1 & Set 3: 6
Set 2 & Set 3: 11
Set1 & Set2 & Set3: 82
No Set: 0

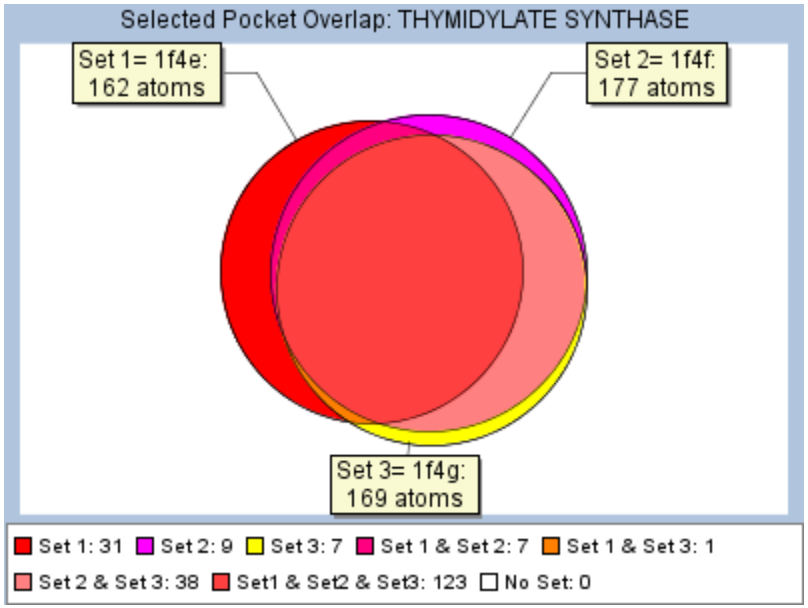

Selected Pocket Overlap: BETA-GLUCOSIDASE A
Set 1= 2cet: 133 atoms
Set 2= 2j77: 68 atoms
Set 3= 2j78: 65 atoms
Set 1: 67
Set 2: 3
Set 3: 1
Set 1 & Set 2: 2
Set 1 & Set 3: 1
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 63
No Set: 0

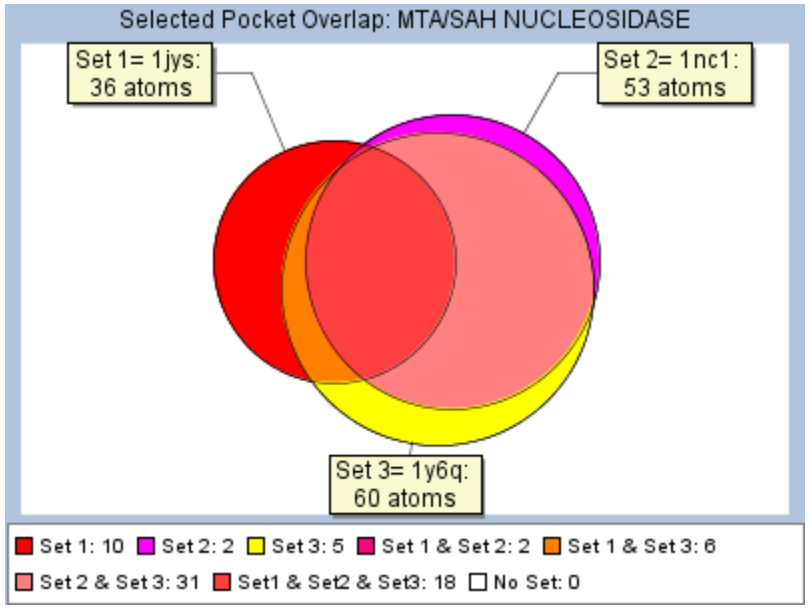

Selected Pocket Overlap: SERINE/THREONINE-PROTEIN KINASE CHK1
Set 1= 1nvq: 59 atoms
Set 2= 2brb: 66 atoms
Set 3= 2c3j: 52 atoms
Set 1: 7
Set 2: 13
Set 3: 2
Set 1 & Set 2: 6
Set 1 & Set 3: 3
Set 2 & Set 3: 4
Set1 & Set2 & Set3: 43
No Set: 0


Selected Pocket Overlap: CELL DIVISION PROTEIN KINASE 2
Set 1= 1b38: 108 atoms
Set 2= 1pxo: 108 atoms
Set 3= 2exm: 112 atoms
Set 1: 9
Set 2: 20
Set 3: 18
Set 1 & Set 2: 11
Set 1 & Set 3: 17
Set 2 & Set 3: 6
Set1 & Set2 & Set3: 71
No Set: 0

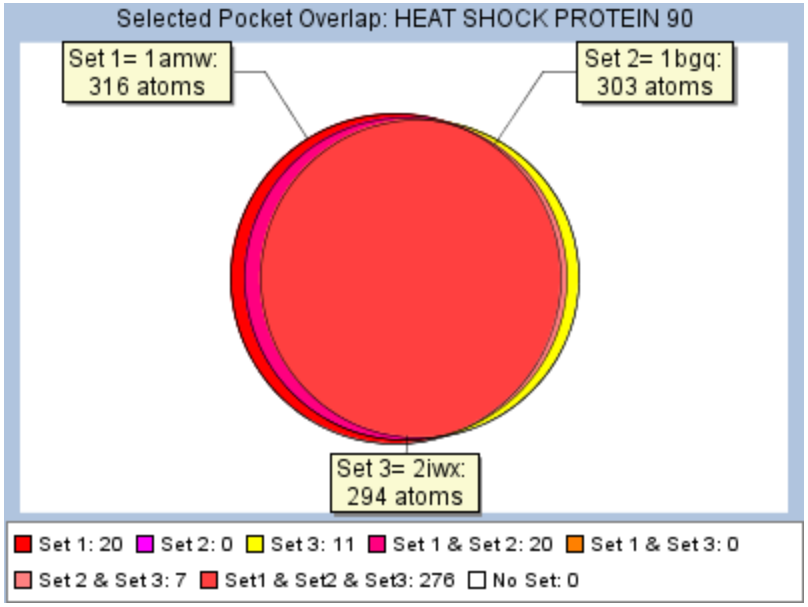

Selected Pocket Overlap: PHENYLETHANOLAMINE N-METHYLTRANSFERASE
Set 1= 1hnn: 116 atoms
Set 2= 2g71: 134 atoms
Set 3= 2obf: 151 atoms
Set 1: 6
Set 2: 12
Set 3: 30
Set 1 & Set 2: 3
Set 1 & Set 3: 2
Set 2 & Set 3: 14
Set1 & Set2 & Set3: 105
No Set: 0

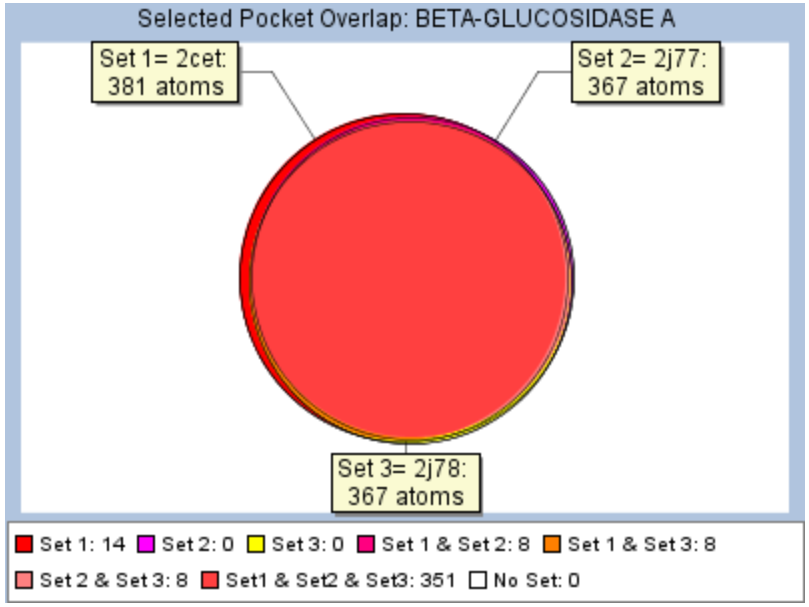

Selected Pocket Overlap: RETINOIC ACID RECEPTOR GAMMA-1
Set 1= 1fcx: 86 atoms
Set 2= 1fcz: 82 atoms
Set 3= 1fd0: 84 atoms
Set 1: 2
Set 2: 0
Set 3: 0
Set 1 & Set 2: 1
Set 1 & Set 3: 3
Set 2 & Set 3: 1
Set1 & Set2 & Set3: 80
No Set: 0

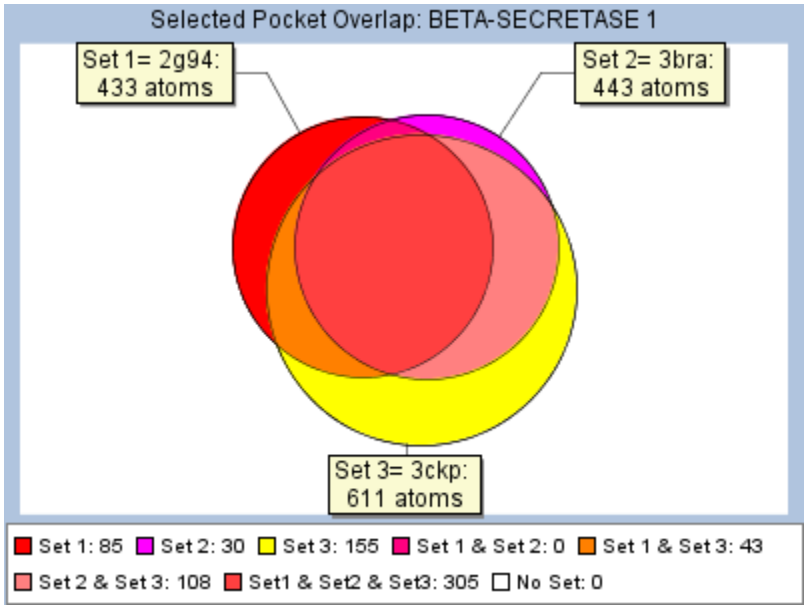

Selected Pocket Overlap: THYMIDYLATE SYNTHASE
Set 1= 1f4e: 162 atoms
Set 2= 1f4f: 177 atoms
Set 3= 1f4g: 169 atoms
Set 1: 31
Set 2: 9
Set 3: 7
Set 1 & Set 2: 7
Set 1 & Set 3: 1
Set 2 & Set 3: 38
Set1 & Set2 & Set3: 123
No Set: 0

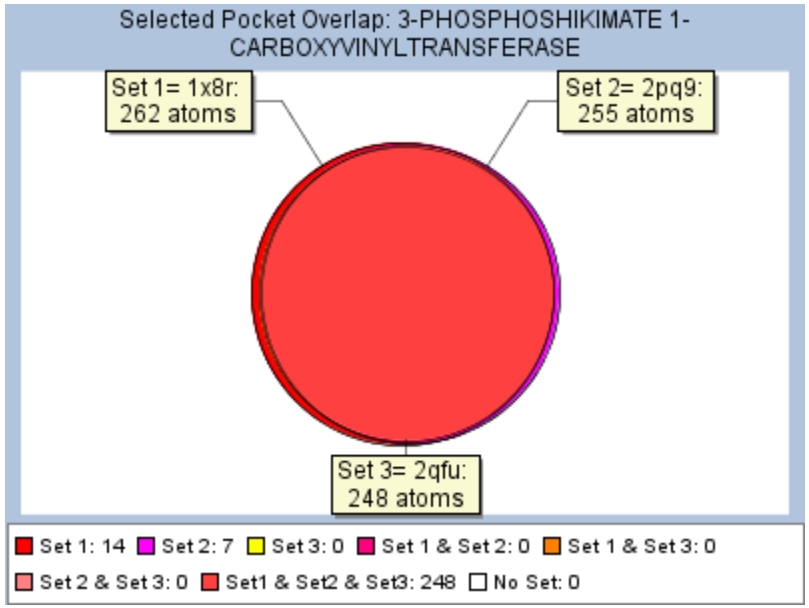

Selected Pocket Overlap: MTA/SAH NUCLEOSIDASE
Set 1= 1jys: 36 atoms
Set 2= 1nc1: 53 atoms
Set 3= 1y6q: 60 atoms
Set 1: 10
Set 2: 2
Set 3: 5
Set 1 & Set 2: 2
Set 1 & Set 3: 6
Set 2 & Set 3: 31
Set1 & Set2 & Set3: 18
No Set: 0

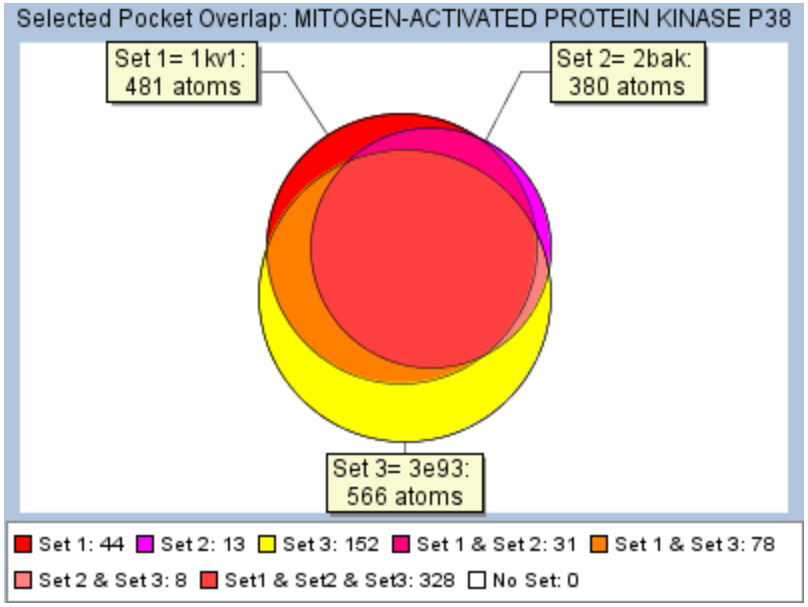
SCREEN

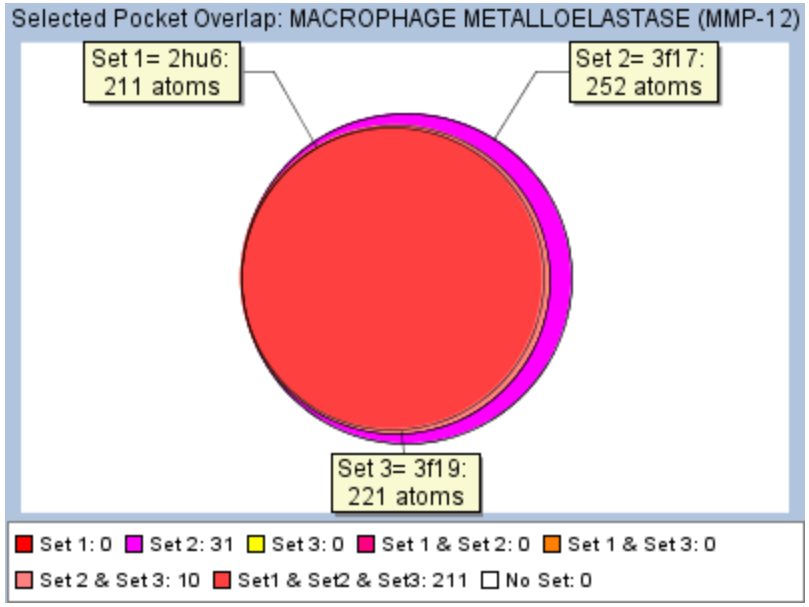

Selected Pocket Overlap: HEAT SHOCK PROTEIN 90
Set 1= 1amw: 316 atoms
Set 2= 1bgq: 303 atoms
Set 3= 2iwx: 294 atoms
Set 1: 20
Set 2: 0
Set 3: 11
Set 1 & Set 2: 20
Set 1 & Set 3: 0
Set 2 & Set 3: 7
Set1 & Set2 & Set3: 276
No Set: 0

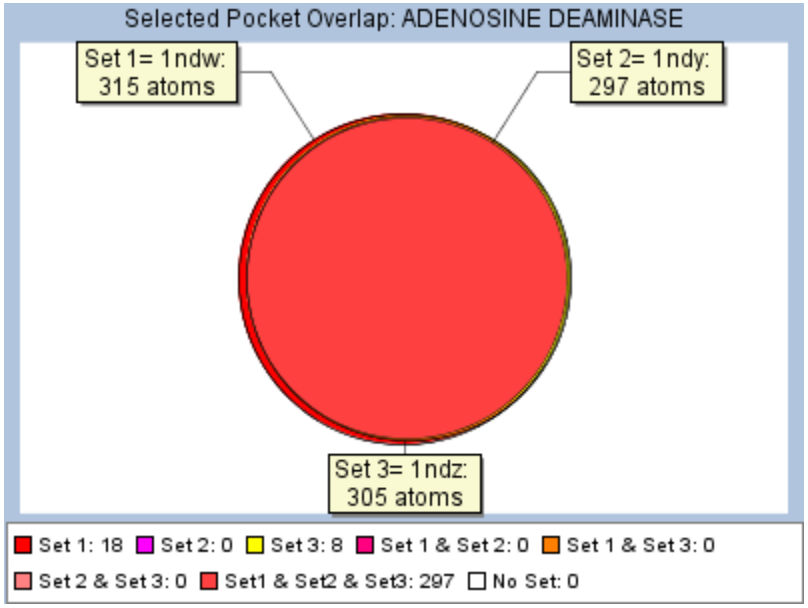

Selected Pocket Overlap: BETA-GLUCOSIDASE A
Set 1= 2cet: 381 atoms
Set 2= 2j77: 367 atoms
Set 3= 2j78: 367 atoms
Set 1: 14
Set 2: 0
Set 3: 0
Set 1 & Set 2: 8
Set 1 & Set 3: 8
Set 2 & Set 3: 8
Set1 & Set2 & Set3: 351
No Set: 0

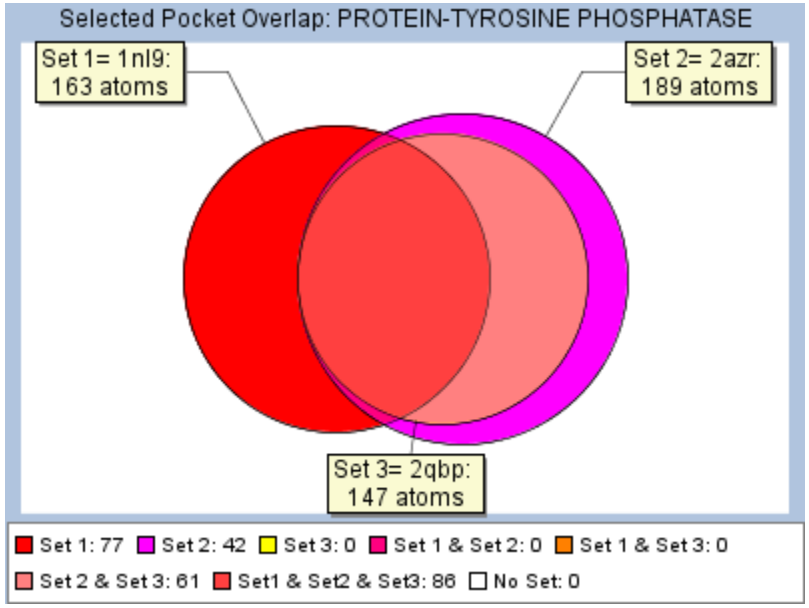

Selected Pocket Overlap: BETA-SECRETASE 1
Set 1= 2g94: 433 atoms
Set 2= 3bra: 443 atoms
Set 3= 3ckp: 611 atoms
Set 1: 85
Set 2: 30
Set 3: 155
Set 1 & Set 2: 0
Set 1 & Set 3: 43
Set 2 & Set 3: 108
Set1 & Set2 & Set3: 305
No Set: 0

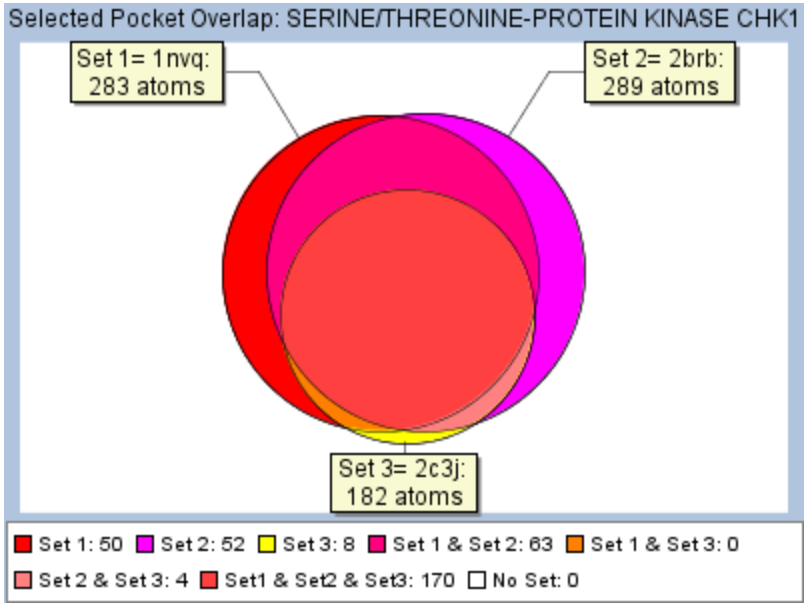

Selected Pocket Overlap: 3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE
Set 1= 1x8r: 262 atoms
Set 2= 2pq9: 255 atoms
Set 3= 2qfu: 248 atoms
Set 1: 14
Set 2: 7
Set 3: 0
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 248
No Set: 0

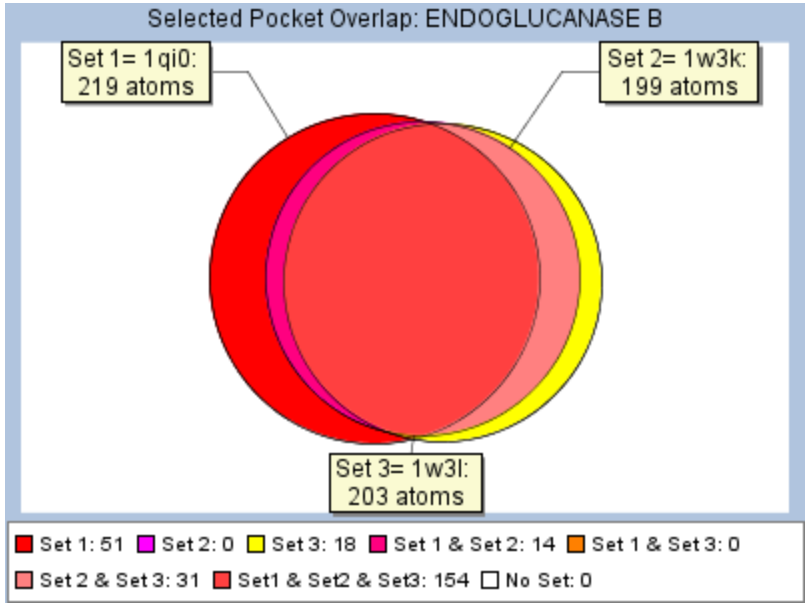

Selected Pocket Overlap: MITOGEN-ACTIVATED PROTEIN KINASE P38
Set 1= 1kv1: 481 atoms
Set 2= 2bak: 380 atoms
Set 3= 3e93: 566 atoms
Set 1: 44
Set 2: 13
Set 3: 152
Set 1 & Set 2: 31
Set 1 & Set 3: 78
Set 2 & Set 3: 8
Set1 & Set2 & Set3: 328
No Set: 0

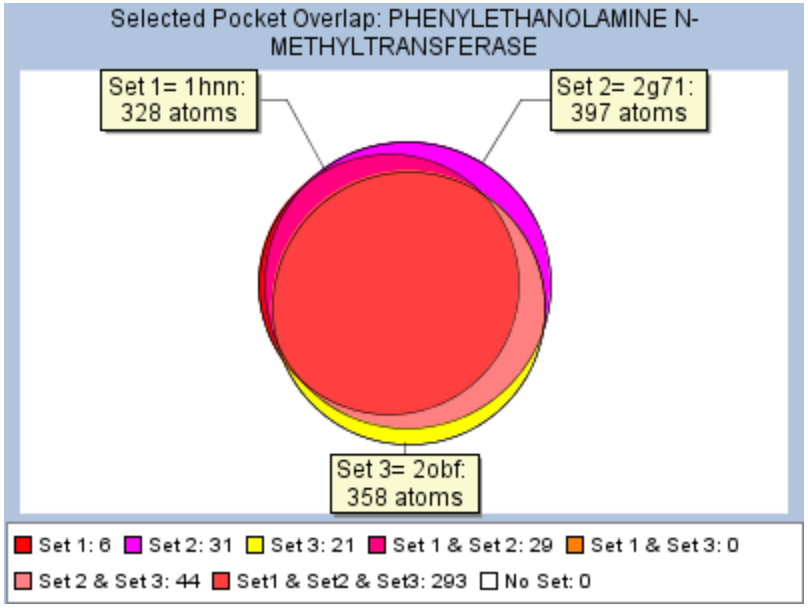

Selected Pocket Overlap: MACROPHAGE METALLOELASTASE (MMP-12)
Set 1= 2hu6: 211 atoms
Set 2= 3f17: 252 atoms
Set 3= 3f19: 221 atoms
Set 1: 0
Set 2: 31
Set 3: 0
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 10
Set1 & Set2 & Set3: 211
No Set: 0

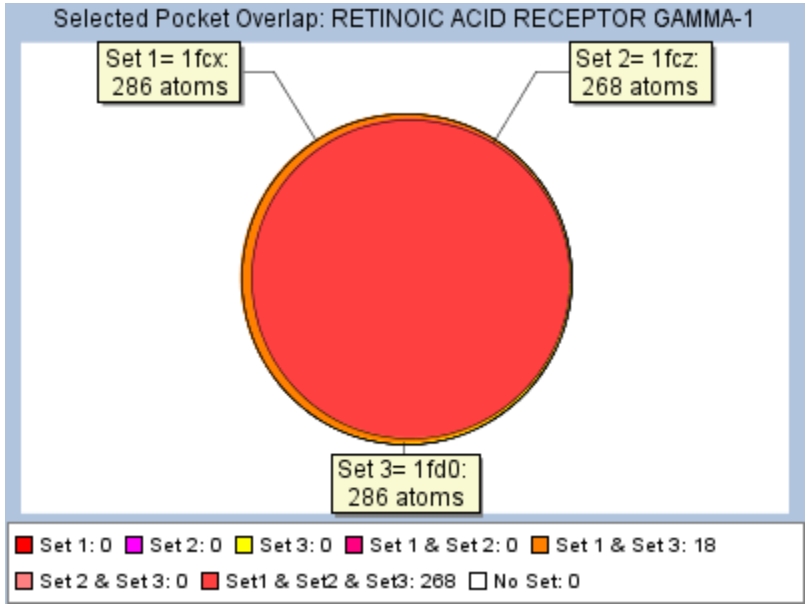

Selected Pocket Overlap: ADENOSINE DEAMINASE
Set 1= 1ndw: 315 atoms
Set 2= 1ndy: 297 atoms
Set 3= 1ndz: 305 atoms
Set 1: 18
Set 2: 0
Set 3: 8
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 297
No Set: 0

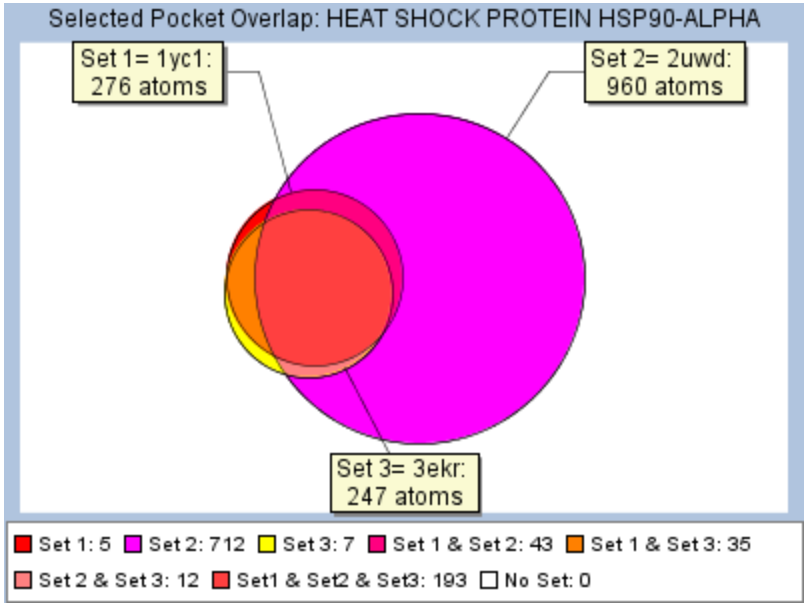

Selected Pocket Overlap: PROTEIN-TYROSINE PHOSPHATASE
Set 1= 1nl9: 163 atoms
Set 2= 2azr: 189 atoms
Set 3= 2qbp: 147 atoms
Set 1: 77
Set 2: 42
Set 3: 0
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 61
Set1 & Set2 & Set3: 86
No Set: 0

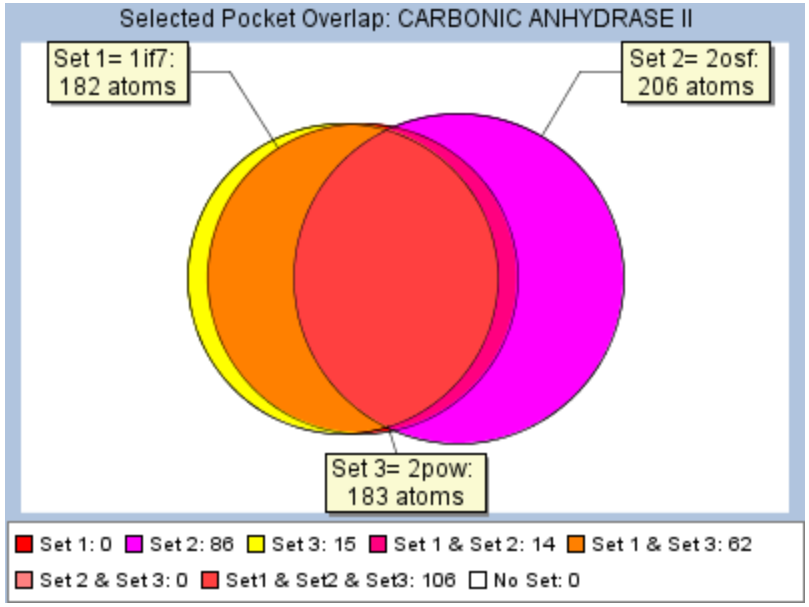

Selected Pocket Overlap: SERINE/THREONINE-PROTEIN KINASE CHK1
Set 1= 1nvq: 283 atoms
Set 2= 2brb: 289 atoms
Set 3= 2c3j: 182 atoms
Set 1: 50 Set 2: 52 Set 3: 8 Set 1 & Set 2: 63 Set 1 & Set 3: 0
Set 2 & Set 3: 4 Set1 & Set2 & Set3: 170 No Set: 0

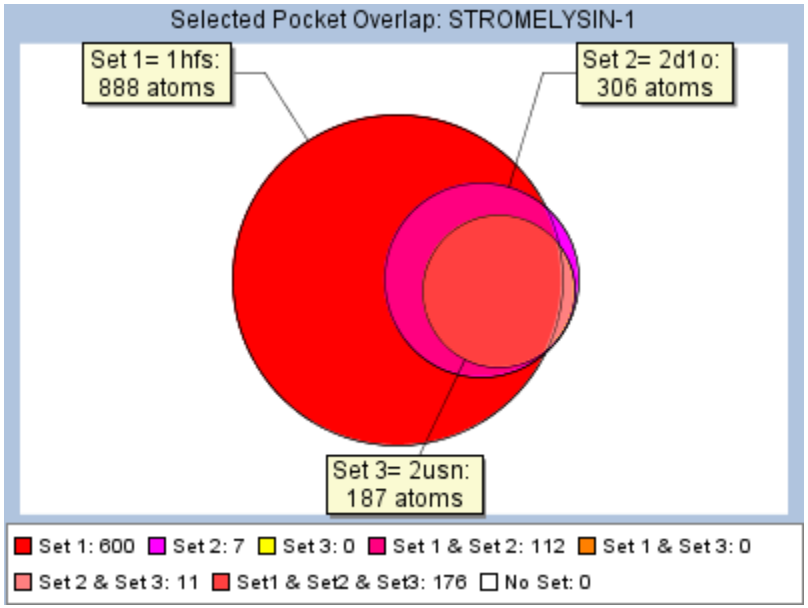

Selected Pocket Overlap: ENDOGLUCANASE B
Set 1= 1qi0: 219 atoms
Set 2= 1w3k: 199 atoms
Set 3= 1w3l: 203 atoms
Set 1: 51 Set 2: 0 Set 3: 18 Set 1 & Set 2: 14 Set 1 & Set 3: 0
Set 2 & Set 3: 31 Set1 & Set2 & Set3: 154 No Set: 0

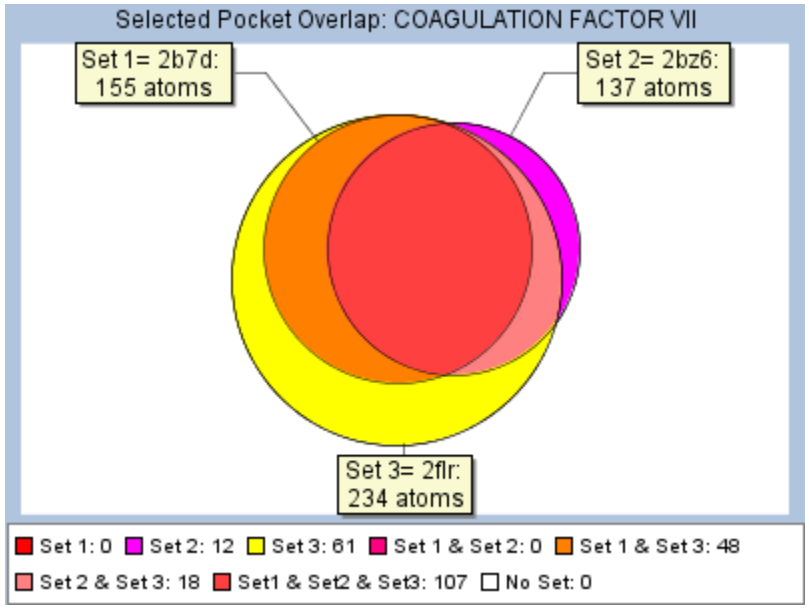

Selected Pocket Overlap: PHENYLETHANOLAMINE N-METHYLTRANSFERASE
Set 1= 1hnn: 328 atoms
Set 2= 2g71: 397 atoms
Set 3= 2obf: 358 atoms
Set 1: 6 Set 2: 31 Set 3: 21 Set 1 & Set 2: 29 Set 1 & Set 3: 0
Set 2 & Set 3: 44 Set1 & Set2 & Set3: 293 No Set: 0

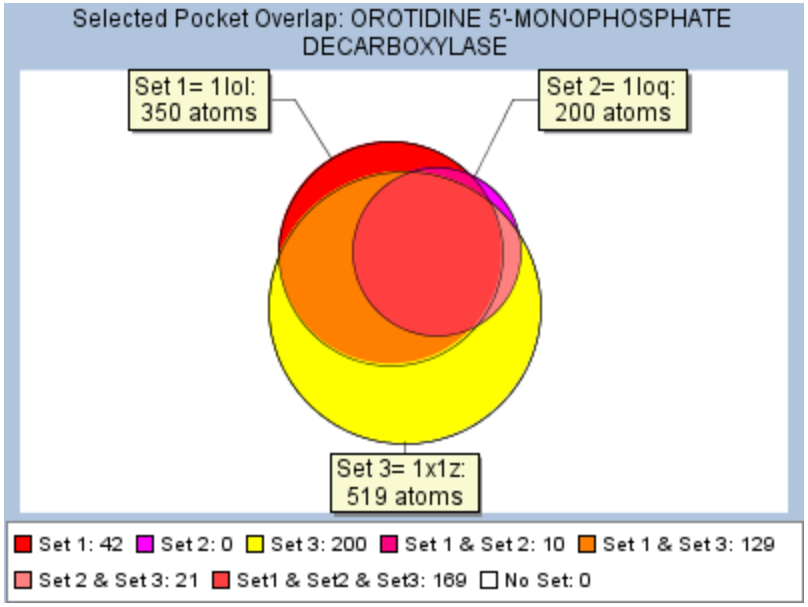

Selected Pocket Overlap: RETINOIC ACID RECEPTOR GAMMA-1
Set 1= 1fcx: 286 atoms
Set 2= 1fcz: 268 atoms
Set 3= 1fd0: 286 atoms
Set 1: 0 Set 2: 0 Set 3: 0 Set 1 & Set 2: 0 Set 1 & Set 3: 18
Set 2 & Set 3: 0 Set1 & Set2 & Set3: 268 No Set: 0

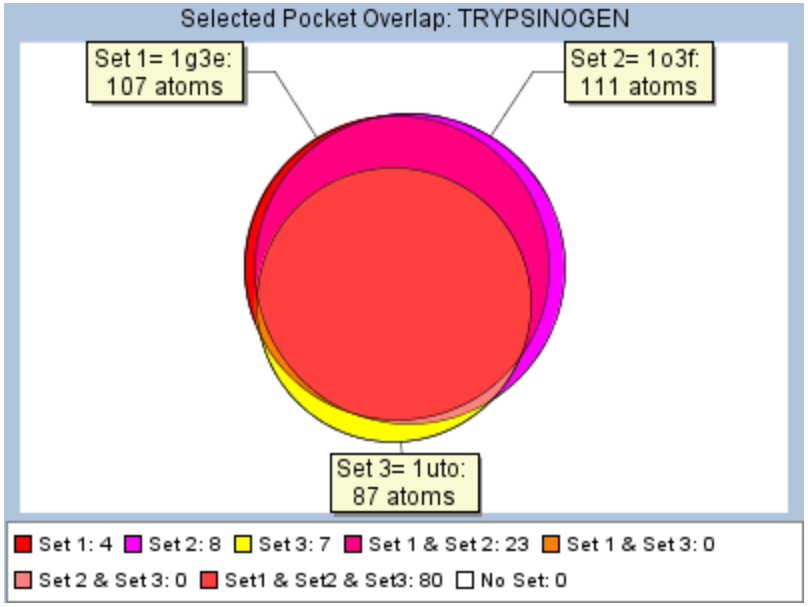

Selected Pocket Overlap: HEAT SHOCK PROTEIN HSP90-ALPHA
Set 1= 1yc1: 276 atoms
Set 2= 2uwd: 960 atoms
Set 3= 3ekr: 247 atoms
Set 1: 5
Set 2: 712
Set 3: 7
Set 1 & Set 2: 43
Set 1 & Set 3: 35
Set 2 & Set 3: 12
Set1 & Set2 & Set3: 193
No Set: 0

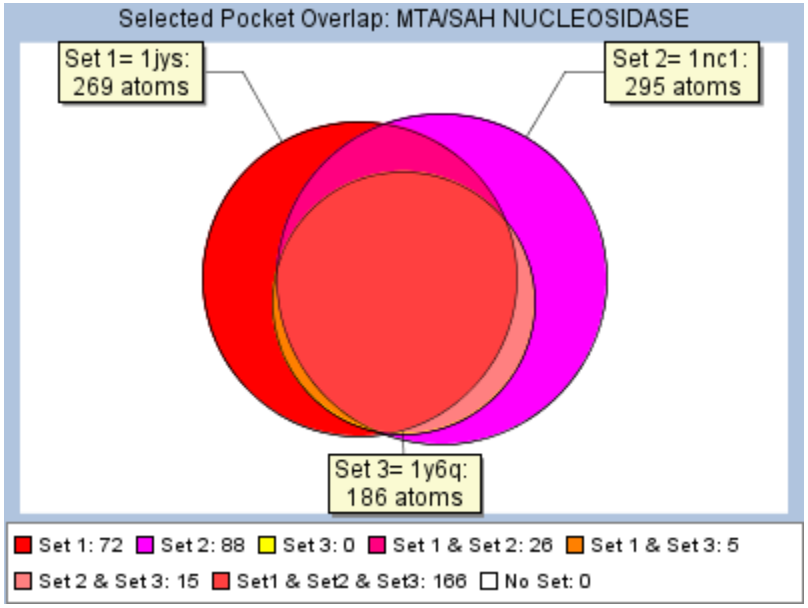

Selected Pocket Overlap: CARBONIC ANHYDRASE II
Set 1= 1if7: 182 atoms
Set 2= 2osf: 206 atoms
Set 3= 2pow: 183 atoms
Set 1: 0
Set 2: 86
Set 3: 15
Set 1 & Set 2: 14
Set 1 & Set 3: 62
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 106
No Set: 0

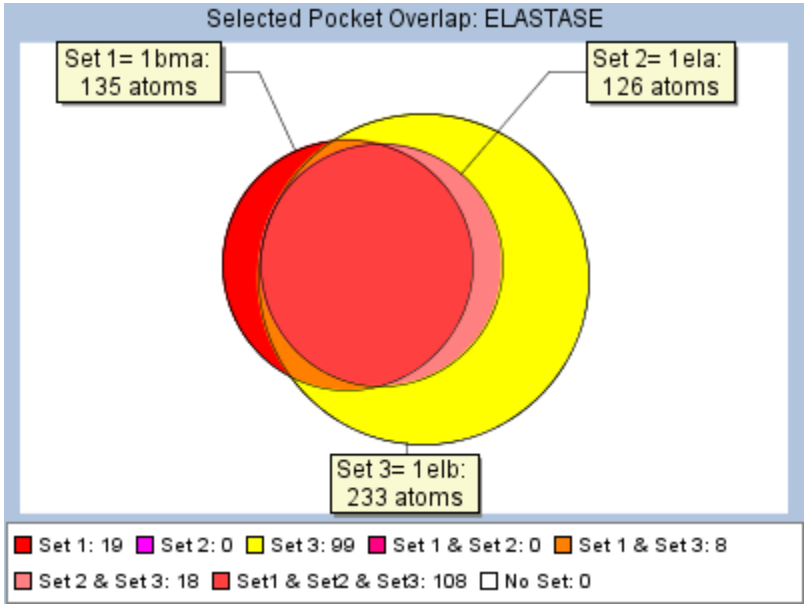

Selected Pocket Overlap: STROMELYSIN-1
Set 1= 1hfs: 888 atoms
Set 2= 2d1o: 306 atoms
Set 3= 2usn: 187 atoms
Set 1: 600
Set 2: 7
Set 3: 0
Set 1 & Set 2: 112
Set 1 & Set 3: 0
Set 2 & Set 3: 11
Set1 & Set2 & Set3: 176
No Set: 0

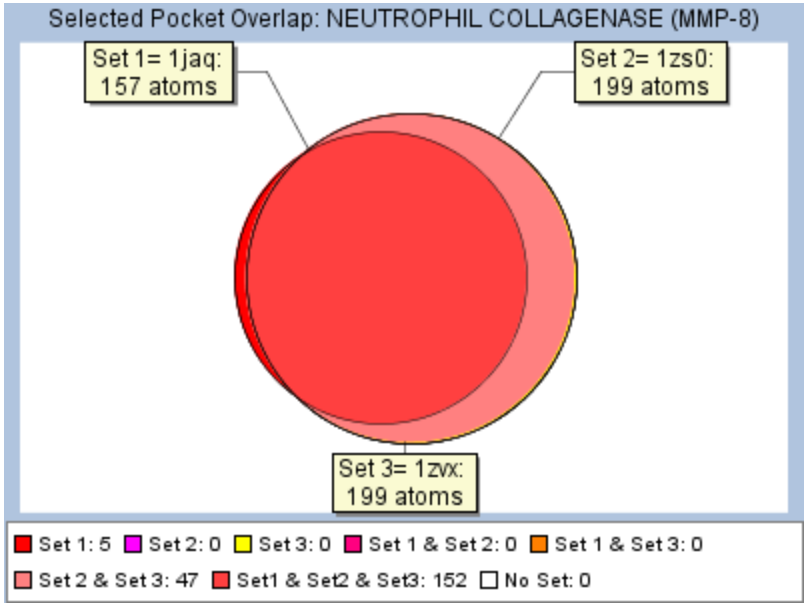

Selected Pocket Overlap: COAGULATION FACTOR VII
Set 1= 2b7d: 155 atoms
Set 2= 2bz6: 137 atoms
Set 3= 2flr: 234 atoms
Set 1: 0
Set 2: 12
Set 3: 61
Set 1 & Set 2: 0
Set 1 & Set 3: 48
Set 2 & Set 3: 18
Set1 & Set2 & Set3: 107
No Set: 0

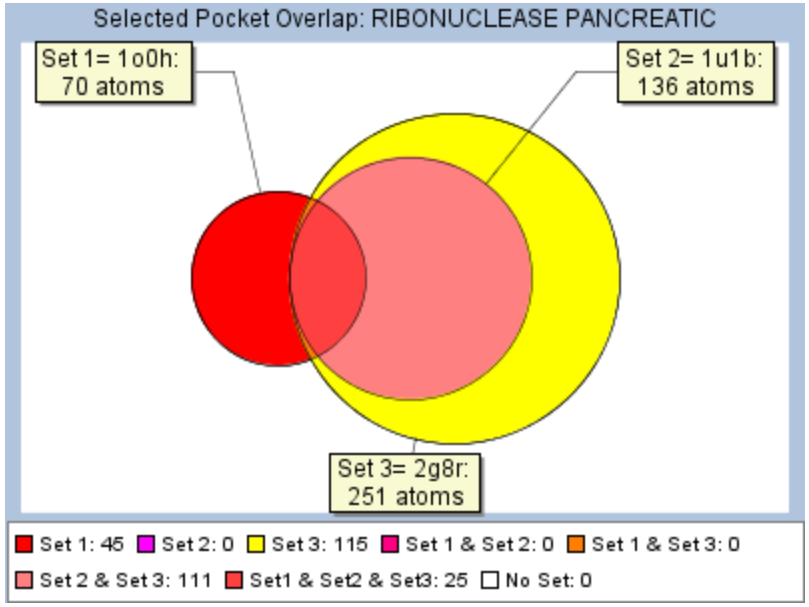

Selected Pocket Overlap: OROTIDINE 5'-MONOPHOSPHATE DECARBOXYLASE
Set 1= 1lol: 350 atoms
Set 2= 1loq: 200 atoms
Set 3= 1x1z: 519 atoms
Set 1: 42
Set 2: 0
Set 3: 200
Set 1 & Set 2: 10
Set 1 & Set 3: 129
Set 2 & Set 3: 21
Set1 & Set2 & Set3: 169
No Set: 0

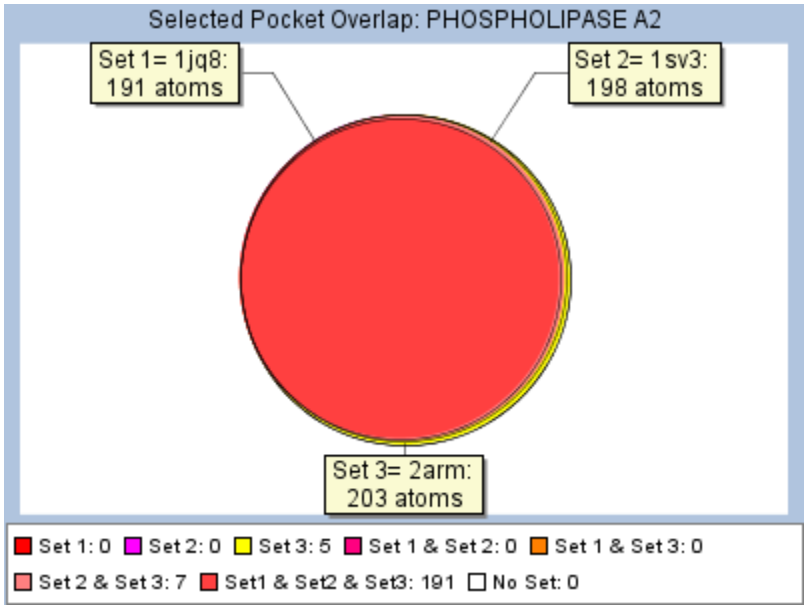

Selected Pocket Overlap: TRYPSINOGEN
Set 1= 1g3e: 107 atoms
Set 2= 1o3f: 111 atoms
Set 3= 1uto: 87 atoms
Set 1: 4
Set 2: 8
Set 3: 7
Set 1 & Set 2: 23
Set 1 & Set 3: 0
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 80
No Set: 0

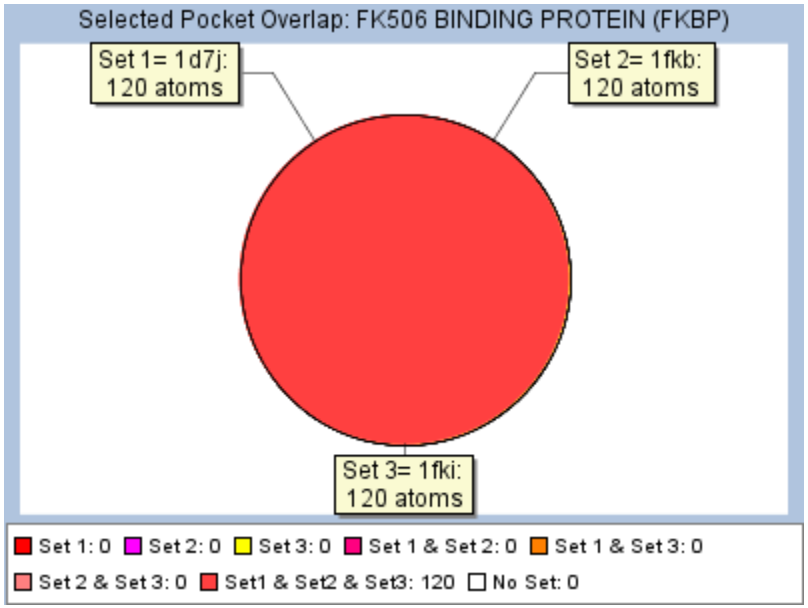

Selected Pocket Overlap: MTA/SAH NUCLEOSIDASE
Set 1= 1jys: 269 atoms
Set 2= 1nc1: 295 atoms
Set 3= 1y6q: 186 atoms
Set 1: 72
Set 2: 88
Set 3: 0
Set 1 & Set 2: 26
Set 1 & Set 3: 5
Set 2 & Set 3: 15
Set1 & Set2 & Set3: 166
No Set: 0


Selected Pocket Overlap: ELASTASE
Set 1= 1bma: 135 atoms
Set 2= 1ela: 126 atoms
Set 3= 1elb: 233 atoms
Set 1: 19
Set 2: 0
Set 3: 99
Set 1 & Set 2: 0
Set 1 & Set 3: 8
Set 2 & Set 3: 18
Set1 & Set2 & Set3: 108
No Set: 0


Selected Pocket Overlap: NEUTROPHIL COLLAGENASE (MMP-8)
Set 1= 1jaq: 157 atoms
Set 2= 1zs0: 199 atoms
Set 3= 1zvx: 199 atoms
Set 1: 5
Set 2: 0
Set 3: 0
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 47
Set1 & Set2 & Set3: 152
No Set: 0


Selected Pocket Overlap: RIBONUCLEASE PANCREATIC
Set 1= 1o0h: 70 atoms
Set 2= 1u1b: 136 atoms
Set 3= 2g8r: 251 atoms
Set 1: 45
Set 2: 0
Set 3: 115
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 111
Set1 & Set2 & Set3: 25
No Set: 0


Selected Pocket Overlap: PHOSPHOLIPASE A2
Set 1= 1jq8: 191 atoms
Set 2= 1sv3: 198 atoms
Set 3= 2arm: 203 atoms
Set 1: 0
Set 2: 0
Set 3: 5
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 7
Set1 & Set2 & Set3: 191
No Set: 0


Selected Pocket Overlap: FK506 BINDING PROTEIN (FKBP)
Set 1= 1d7j: 120 atoms
Set 2= 1fkb: 120 atoms
Set 3= 1fki: 120 atoms
Set 1: 0
Set 2: 0
Set 3: 0
Set 1 & Set 2: 0
Set 1 & Set 3: 0
Set 2 & Set 3: 0
Set1 & Set2 & Set3: 120
No Set: 0

Appendix IV: Virtual Screening Dataset Selection Details

Contained in this appendix is the detailed descriptions of how datasets were extracted, curated, and categorized for the ChEMBL and WOMBAT databases. These descriptions are organized by target with ChEMBL extraction being discussed as the modeling/validation set and WOMBAT as the external set.

ACHE (Acetylcholinesterase)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 93. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 445 active molregnos and 472 inactive molregnos. A total of 8 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 901 (437 active and 464 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 13 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 887 compounds (424 active and 463 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate

where the activity classes were in agreement (removed 22 structures) and no duplicates were found where the activity classes were in disagreement.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname AChE. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 29 of the MIREG could not be converted from SMILES. The remaining compounds were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 405 active MIREG and 344 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 84 active and 6 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 659 compounds (321 active and 338 inactive). No MIREG occurred in both the active and inactive classes. 7 compounds overlap with ChEMBL, leaving a final dataset of 652 compounds (321 active and 331 inactive).

ACK1 (Activated Cdc42-associated Kinase)

Modeling/Validation Set

The Ack1 dataset was curated from patented data from Amgen (US patent 2006- 0040965, US patent US 2007 - 0072851), OSI Pharmaceuticals (US patent 2009 - 0286768) and other sources published in literature. In total, 487 activities were collected. These were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding

124 actives and 68 inactives. After removing salts, standardizing charges, and normalizing stereo information, 16 active and 4 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 172 compounds (108 active and 64 inactive).

External Set

A lack of known ligands for this protein prevented the generation of additional external sets.

AR (Androgen Receptor)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 56. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 305 active molregnos and 149 inactive molregnos. A total of 16 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 422 (289 active and 133 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, no compounds were found to occur more than once in the dataset. Therefore the final dataset consists of 422 compounds (289 active and 133 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each

representative of these “duplicates” was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 19 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 9 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname AR. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound’s smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 449 active MIREG and 68 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 161 active and 7 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 349 compounds (288 active and 61 inactive). A total of 80 MIREG occurred in both the active and inactive classes leaving a final dataset of 269 compounds (248 active and 21 inactive). 11 compounds overlap with ChEMBL, leaving a final dataset of 258 compounds (237 active and 21 inactive).

B2AR (Beta-2 Adrenergic Receptor)

Modeling/Validation Set

All activities with a standard_type of IC50 or Ki were extracted from ChEMBLdb using assay_ids that corresponded to tid 43. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 94 active

molregnos and 157 inactive molregnos. No molregnos occurred in both the active and inactive classes. For each molregno, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 3 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 248 compounds (94 active and 154 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 10 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 3 structures).

External Set

From WOMBAT, all activities with an act_type of IC50, pKi and Ki were extracted using target_fullname 'beta2 adrenergic'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. Compounds were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 60 active MIREG and 88 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 9 active and 2 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 137 compounds (51 active and 86 inactive). No MIREG occurred in both the active and inactive classes.

CA2 (Carbonic Anhydrase II)

Modeling/Validation Set

All activities with a standard_type of Ki were extracted from ChEMBLdb using assay_ids that corresponded to tid 15. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 726 active molregnos and 382 inactive molregnos. A total of 15 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1078 (711 active and 367 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 5 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 1073 compounds (709 active and 364 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 43 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 2 structures). Errors in descriptor calculation identified compounds with carboranes as problematic and 12 compounds were removed.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname 'CA-II'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 16 of the MIREG could not be converted from SMILES. The remaining MIREG were separated into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 953 active MIREG and 251 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 270 active and 61 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 873 compounds (683 active and 190 inactive). A total of 30 MIREG occurred in both the active and inactive classes leaving a final dataset of 843 compounds (668 active and 175 inactive). 65 compounds overlap with ChEMBL, leaving a final dataset of 778 compounds (662 active and 116 inactive).

CDK2 (Cyclin Dependent Kinase 2)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 11678. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 739 active molregnos and 633 inactive molregnos. A total of 6 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1360 (733 active and 627 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information,

no compounds were found to occur more than once in the dataset. Therefore the final dataset consists of 1360 compounds (733 active and 627 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 21 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 2 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname 'CDK2'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 501 active MIREG and 289 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 20 active and 6 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 758 compounds (481 active and 277 inactive). A total of 2 MIREG occurred in both the active and inactive classes leaving a final dataset of 756 compounds (480 active and 276 inactive).

COX2 (Cyclooxygenase-2)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 126. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 699 active molregnos and 759 inactive molregnos. A total of 14 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1430 (685 active and 745 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for each of this compound. The duplicates were found to fall in the same activity class, so one example was retained while the other was deleted. This resulted in a final dataset of 1429 compounds (685 active and 744 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 9 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 2 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname 'COX-2'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into

active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 514 active MIREG and 406 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 84 active and 19 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 817 compounds (430 active and 387 inactive). A total of 6 MIREG occurred in both the active and inactive classes leaving a final dataset of 811 compounds (427 active and 384 inactive).

DHFR (Dihydrofolate Reductase)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 6. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 232 active molregnos and 251 inactive molregnos. A total of 10 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 463 (222 active and 241 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, no compounds were found to occur more than once in the dataset. Therefore the final dataset consists of 463 compounds (222 active and 241 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate

where the activity classes were in agreement (removed 4 structures) and no duplicates were found where the activity classes were in disagreement.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname 'DHFR'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 93 active MIREG and 210 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 29 active and 30 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 244 compounds (64 active and 180 inactive). A total of 2 MIREG occurred in both the active and inactive classes leaving a final dataset of 240 compounds (62 active and 178 inactive).

ESR1 (Estrogen Receptor Alpha)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 19. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 316 active molregnos and 571 inactive molregnos. A total of 3 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 881 (313 active and 568 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After

removing salts, standardizing charges, and normalizing stereo information, 2 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 878 compounds (312 active and 566 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 50 structures) and no duplicates were found where the activity classes were in disagreement. Errors in descriptor calculation identified compounds with carboranes as problematic and 6 compounds were removed.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname 'ERalpha'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 972 active MIREG and 176 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 335 active and 3 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 810 compounds (637 active and 173 inactive). A total of 8 MIREG occurred in both the active and inactive classes leaving a final dataset of 802 compounds

(633 active and 169 inactive). 3 compounds overlap with ChEMBL, leaving a final dataset of 799 compounds (633 active and 166 inactive).

ESR2 (Estrogen Receptor Beta)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 174. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 356 active molregnos and 352 inactive molregnos. A total of 2 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 704 (354 active and 350 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for this compound. The duplicates were found to fall in the same activity class, so one example was retained while the other was deleted. This resulted in a final dataset of 703 compounds (353 active and 350 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 32 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 9 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname 'ERbeta'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 338 active MIREG and 335 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 70 active and 16 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 587 compounds (268 active and 319 inactive). A total of 2 MIREG occurred in both the active and inactive classes leaving a final dataset of 583 compounds (266 active and 317 inactive). 4 compounds overlap with ChEMBL, leaving a final dataset of 579 compounds (266 active and 313 inactive).

F10 (Coagulation Factor X)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 194. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 510 active molregnos and 494 inactive molregnos. A total of 2 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1000 (508 active and 492 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for this compound. The

duplicates were found to fall in the same activity class, so one example was retained while the other was deleted. This resulted in a final dataset of 999 compounds (508 active and 491 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 32 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 2 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%fXa%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 5 of the MIREG could not be converted from SMILES. The remaining MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 1870 active MIREG and 445 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 236 active and 15 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 2064 compounds (1634 active and 430 inactive). A total of 14 MIREG occurred in both the active and inactive classes leaving a final dataset of 2050 compounds (1627 active and 423 inactive).

GR (Glucocorticoid Receptor)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 25. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 210 active molregnos and 206 inactive molregnos. A total of 15 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 386 (195 active and 191 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for this compound. The duplicates were found to fall in the same activity class, so one example was retained while the other was deleted. This resulted in a final dataset of 385 compounds (194 active and 191 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 9 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 3 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%Glucocorticoid receptor%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were

separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 677 active MIREG and 30 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 295 active and 3 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 409 compounds (382 active and 27 inactive). A total of 18 MIREG occurred in both the active and inactive classes leaving a final dataset of 391 compounds (611 active and 71 inactive). 4 compounds overlap with ChemBl, leaving a final dataset of 387 compounds (370 active and 17 inactive).

HIV-Int (HIV Integrase)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 12456. These activities were then filtered into active and inactive classes using thresholds of <= 1000nM and >=50000nM respectively yielding 213 active molregnos and 567 inactive molregnos. A total of 15 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 750 (198 active and 552 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for this compound. The duplicates were found to fall in the same activity class, so one example was retained while the other was deleted. This resulted in a final dataset of 749 compounds (197 active and 552 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 10 structures) and no duplicates were found where the activity classes were in disagreement. Errors in descriptor calculation identified compounds with carboranes as problematic and 1 compound was removed.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%HIV%' and '%IN%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 150 active MIREG and 1631 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 35 active and 766 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 980 compounds (115 active and 865 inactive). A total of 14 MIREG occurred in both the active and inactive classes leaving a final dataset of 966 compounds (108 active and 858 inactive). 12 compounds overlap with ChemBl, leaving a final dataset of 954 compounds (108 active and 846 inactive).

HIV-Pr (HIV Protease)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 191. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 912 active molregnos and 633 inactive molregnos. A total of 9 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1527 (903 active and 624 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for this compound. The duplicates were found to fall in the same activity class, so one example was retained while the other was deleted. This resulted in a final dataset of 1526 compounds (903 active and 623 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 116 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 20 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%HIV%' and '%P%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 4 of the MIREG

could not be converted from SMILES. The remaining MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 3519 active MIREG and 330 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 1113 active and 32 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 2704 compounds (2406 active and 298 inactive). A total of 10 MIREG occurred in both the active and inactive classes leaving a final dataset of 2694 compounds (2401 active and 293 inactive). 3 compounds overlap with ChemBl, leaving a final dataset of 2691 compounds (2400 active and 291 inactive).

HIV-RT (HIV Reverse Transcriptase)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 228. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 513 active molregnos and 664 inactive molregnos. A total of 21 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1135 (492 active and 643 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 1 compound was found to occur more than once in the dataset. Activities were analyzed for this compound. The duplicates were found to fall in opposing activity classes, so both deleted. This resulted in a final dataset of 1133 compounds (491 active and 642 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 32 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 5 structures). Errors in descriptor calculation identified compounds with carboranes as problematic and 1 compound was removed.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%HIV%' and '%RT%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 20 of the MIREG could not be converted from SMILES. The remaining MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 1381 active MIREG and 1053 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 629 active and 273 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 1532 compounds (752 active and 780 inactive). A total of 120 MIREG occurred in both the active and inactive classes leaving a final dataset of 1412 compounds (692 active and 720 inactive). 1 compound overlaps with ChemBl, leaving a final dataset of 1411 compounds (692 active and 719 inactive).

PARP1 (Poly [ADP-ribose] Polymerase-1)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 11663. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 176 active molregnos and 123 inactive molregnos. No molregnos occurred in both the active and inactive classes and after removing salts, standardizing charges, and normalizing stereo information, no compounds were found to occur more than once in the dataset. Therefore the final dataset consists of 299 compounds (176 active and 123 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 6 structures) and no duplicates were found where the activity classes were in disagreement.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%PARP1%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 252 active MIREG and 48 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 4 active and 1 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates

were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 295 compounds (248 active and 47 inactive). A total of 2 MIREG occurred in both the active and inactive classes leaving a final dataset of 293 compounds (247 active and 46 inactive).

PDE5 (Phosphodiesterase 5A)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 3. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 327 active molregnos and 363 inactive molregnos. No molregnos occurred in both the active and inactive classes. For each molregno, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 3 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 687 compounds (324 active and 363 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 5 structures) and no duplicates were found where the activity classes were in disagreement.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%PDE5%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 470 active MIREG and 72 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 42 active and 1 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 499 compounds (428 active and 71 inactive).

PNP (Purine Nucleoside Phosphorylase)

Modeling/Validation Set

All activities with a standard_type of IC50 or Ki were extracted from ChEMBLdb using assay_ids that corresponded to tid 12690. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 89 active molregnos and 86 inactive molregnos. A total of 1 molregno occurred in both the active and inactive classes. This was excluded from the set leaving 173 (88 active and 85 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, no compounds were found to occur more than once in the dataset. Therefore the final dataset consists of 173 compounds (88 active and 85 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 14 structures) and no duplicates were found where the activity classes were in disagreement.

External Set

From WOMBAT all activities with an act_type of IC50 and Ki were extracted using target_fullname '%PNP%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 57 active MIREG and 40 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 15 active and 1 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 81 compounds (42 active and 39 inactive).

PPARG (Peroxisome Proliferator-Activated Receptor Gamma)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 133. These activities were then filtered into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 250 active molregnos and 131 inactive molregnos. A total of 1 molregno occurred in both the active and inactive classes. This was excluded from the set leaving 379 (249 active and 130 inactive) molregnos. For each

of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 3 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 376 compounds (246 active and 130 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 20 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 5 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted from using swissp_id '%PARG%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 224 active MIREG and 155 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 29 active and 8 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 342 compounds (195 active and 147 inactive). A total of 2 MIREG occurred in both the active and inactive classes leaving a final dataset of 340 compounds (194 active and 146 inactive).

REN (Renin)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 11225. These activities were then filtered into active and inactive classes using thresholds of <= 10nM and >=1000nM respectively yielding 801 active molregnos and 468 inactive molregnos. A total of 16 molregnos occurred in both the active and inactive classes. These were excluded from the set leaving 1237 (785 active and 452 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 2 compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 1235 compounds (783 active and 452 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 69 structures) and all replicates were removed for duplicates where the activity classes were in disagreement (removed 11 structures).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%renin%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 17 of the MIREG could not

be converted from SMILES. The remaining MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 676 active MIREG and 52 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 174 active and 3 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 551 compounds (502 active and 49 inactive). A total of 6 MIREG occurred in both the active and inactive classes leaving a final dataset of 545 compounds (499 active and 46 inactive). 9 compounds overlap with ChemBl, leaving a final dataset of 536 compounds (498 active and 38 inactive).

SRC (Tyrosine Protein Kinase SRC)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 10434. These activities were then filtered into active and inactive classes using thresholds of <= 100 and >=10000 respectively yielding 632 active molregnos and 831 inactive molregnos. A total of 8 molregnos occurred in both the active and inactive classes. These were excluded from the modeling set leaving 1447 (624 active and 823 inactive) molregnos. For each of these molregnos, the compounds smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 4 compounds were found to occur more than once in the dataset. For each of these compounds activities were analyzed. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 1443 (623 active and 820 inactive).

When preparing the dataset for QSAR modeling, chirality of compounds was removed. This caused the identification of several more "duplicates". When the activity class of each representative of these "duplicates" was investigated, one replicate was kept for each duplicate where the activity classes were in agreement (removed 16 structures) and no duplicates were found where the activity classes were in disagreement.

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%SRC%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 402 active MIREG and 383 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 56 active and 32 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 697 compounds (346 active and 351 inactive). A total of 4 MIREG occurred in both the active and inactive classes leaving a final dataset of 693 compounds (344 active and 349 inactive). 4 compounds overlap with ChemBl, leaving a final dataset of 689 compounds (344 active and 345 inactive).

F2 (Thrombin)

Modeling/Validation Set

All activities with a standard_type of IC50 were extracted from ChEMBLdb using assay_ids that corresponded to tid 11. These activities were then filtered into active and inactive classes

using thresholds of <= 100nM and >=10000nM respectively yielding 373 active molregnos and 787 inactive molregnos. A total of 1 molregno occurred in both the active and inactive classes. This was excluded from the set leaving 1158 (372 active and 786 inactive) molregnos. For each of these molregnos, the compound's smiles were extracted from ChEMBLdb. After removing salts, standardizing charges, and normalizing stereo information, 8 compounds were found to occur more than once in the dataset. For each of these compounds activities were analyzed. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a final dataset of 1150 (368 active and 782 inactive).

External Set

From WOMBAT, all activities with an act_type of IC50 and Ki were extracted using target_fullname '%factor II%'. The activities were filtered using Pipeline Pilot. For each of the MIREG, the compound's smiles were extracted from WOMBAT. 1 of the MIREG was not able to be converted from SMILES. The remaining MIREG were separated into active and inactive classes using thresholds of <= 100nM and >=10000nM respectively yielding 1194 active MIREG and 973 inactive MIREG. After removing salts, standardizing charges, and normalizing stereo information, 162 active and 63 inactive compounds were found to occur more than once in the dataset. Activities were analyzed for each of these compounds. All duplicates were found to fall in the same activity class, so one example was retained while the others were deleted. This resulted in a dataset of 1942 compounds (1032 active and 910 inactive). A total of 2 MIREG occurred in both the active and inactive classes leaving a final dataset of 1940 compounds (1031 active and 909 inactive). 7 compounds overlap with ChemBl, leaving a final dataset of 1933 compounds (1028 active and 905 inactive).

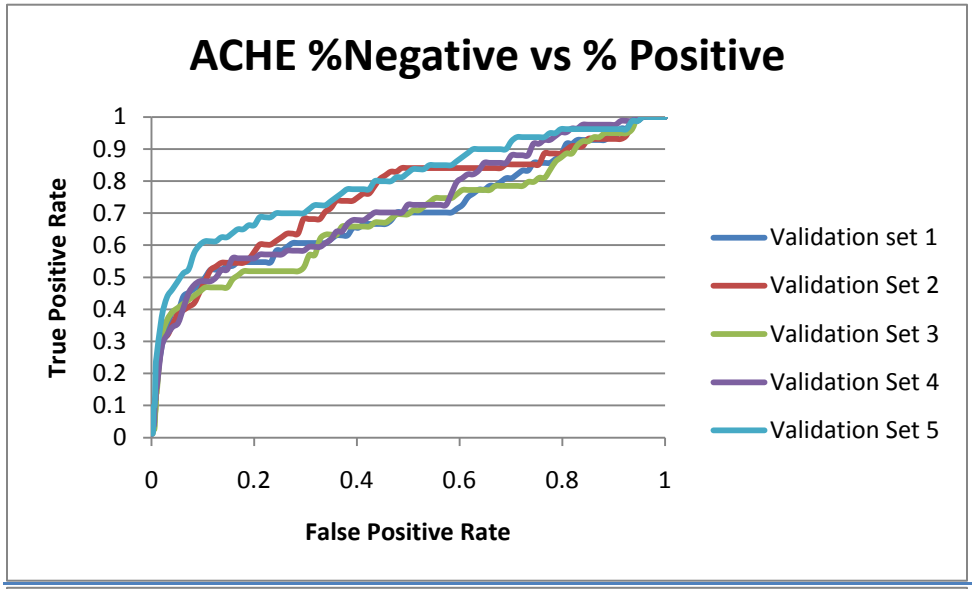
Appendix V: ROC Curves from Benchmark Screening

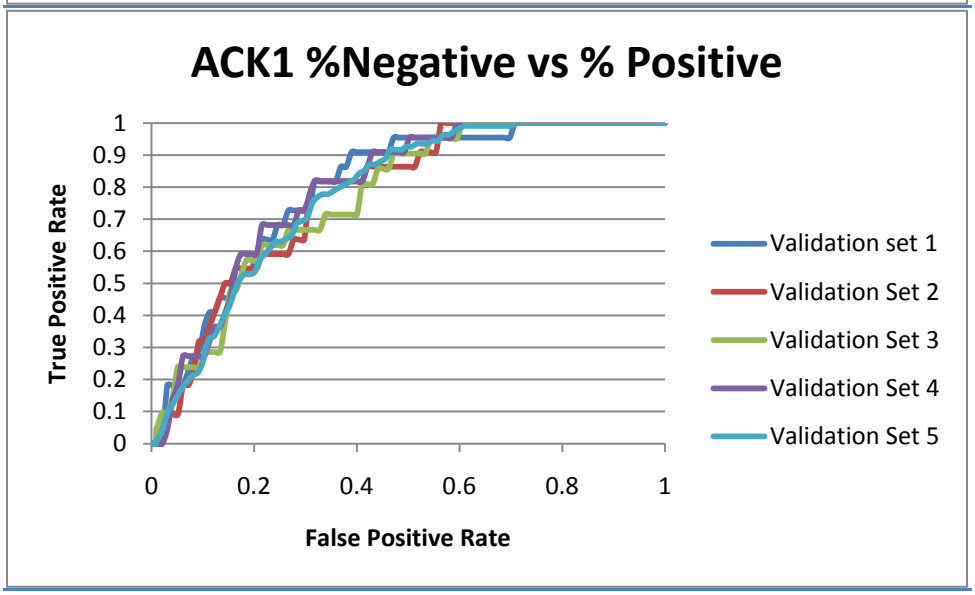
Contained within this appendix are the ROC curves generated when applying different methods to rank the screening sets. The ROC curves are organized by both the applied method and by whether the ROC curve was generated considering the whole screening set or just the compounds with tested activity for the target of interest. More detailed discussion of the results are contained in each subsection of this appendix.

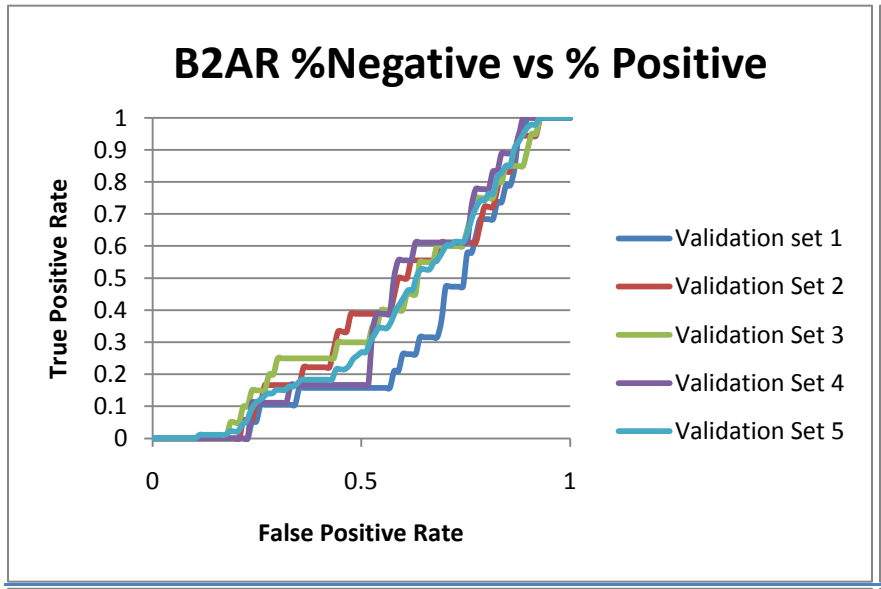
Docking

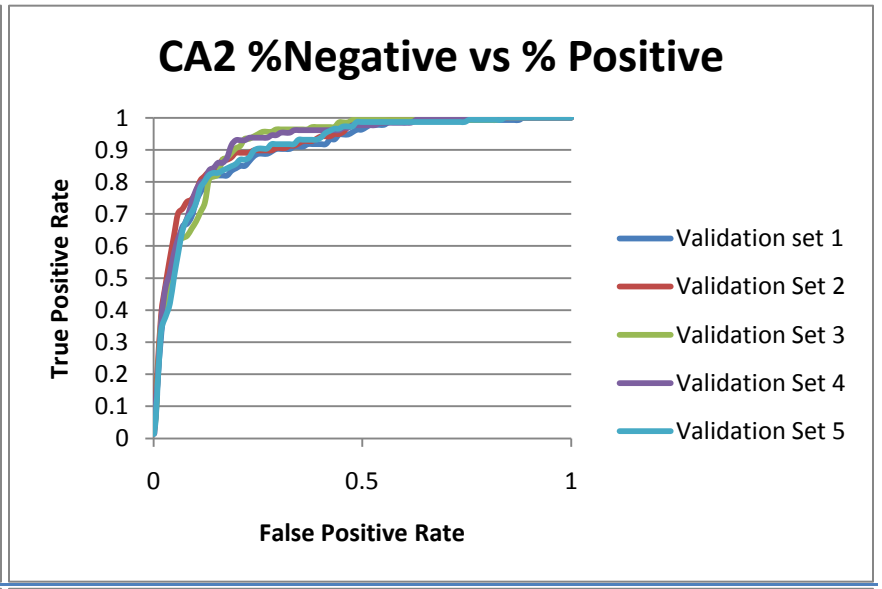
Docking with eHiTS proves to be fairly useful when applied to virtual screening of the full screening library. However, this is only the case when eHiTS was able to identify the protein family and use a family scoring function. In the absence of a family scoring function (see B2AR, HIV-Int, PARP1, PDE5, REN, and PNP), the ranking of compounds is little better than random. The same trend can be seen when looking at the ROC curves for compounds with known activity. Generally, curves are poor when a family scoring function is unavailable. Additionally, docking accuracy is generally lower when looking at only the known compounds indicating that docking performs a better coarse refinement than a fine refinement of a compound library.

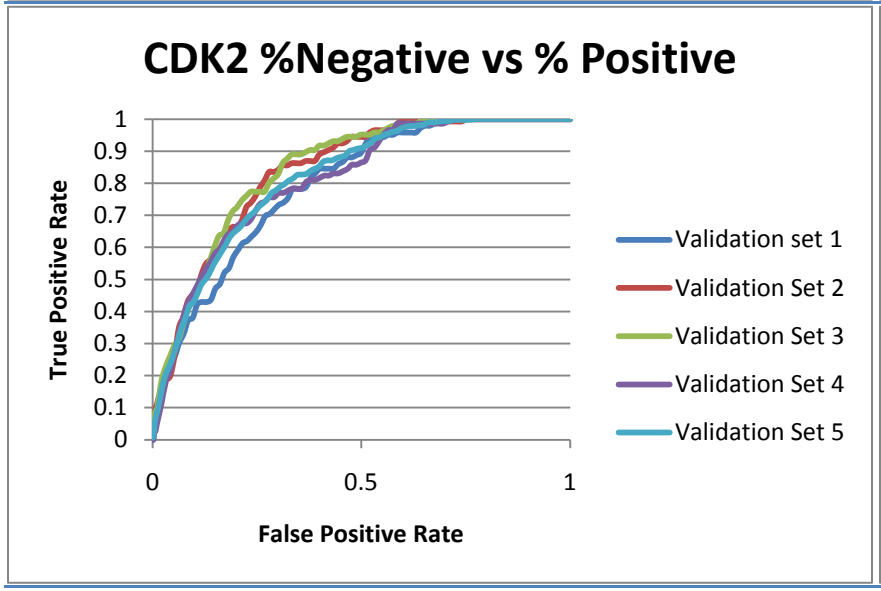
Full Screening Sets

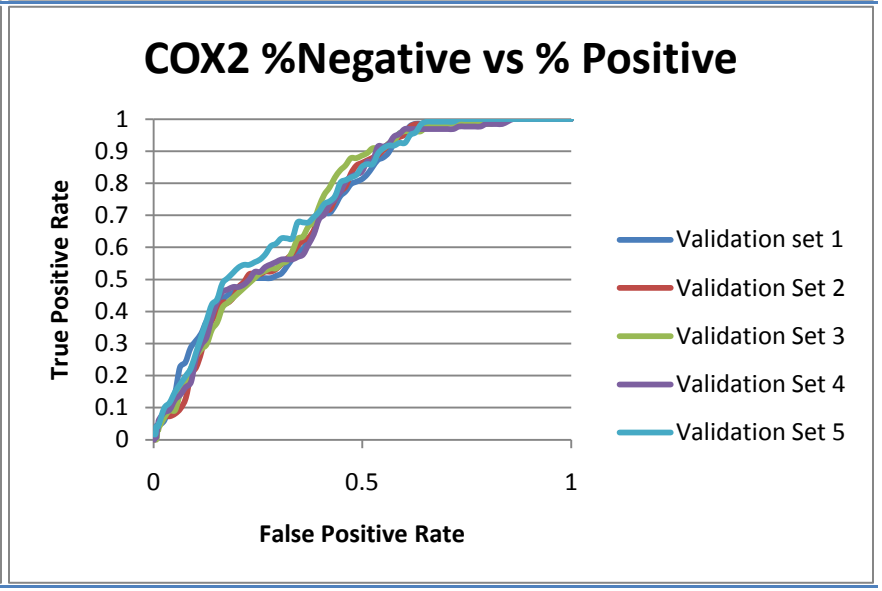


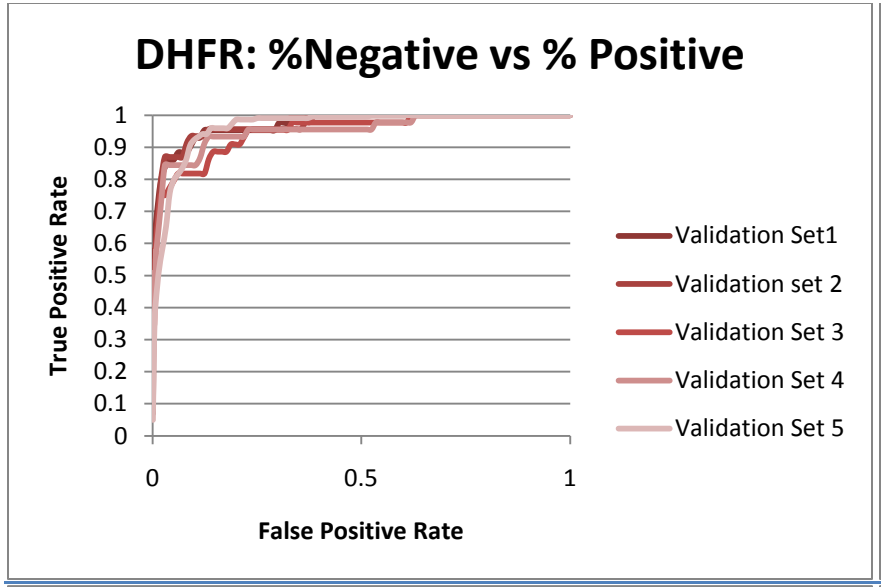


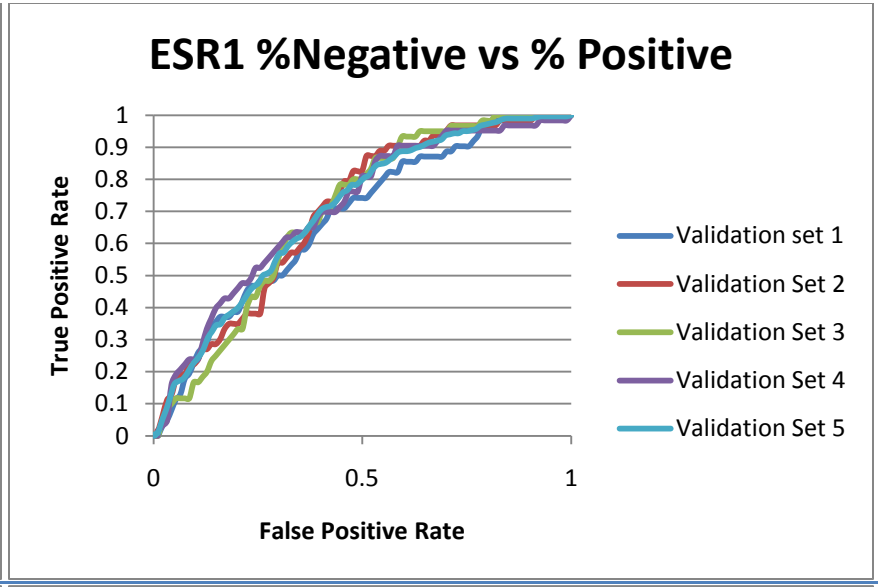

B2AR %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

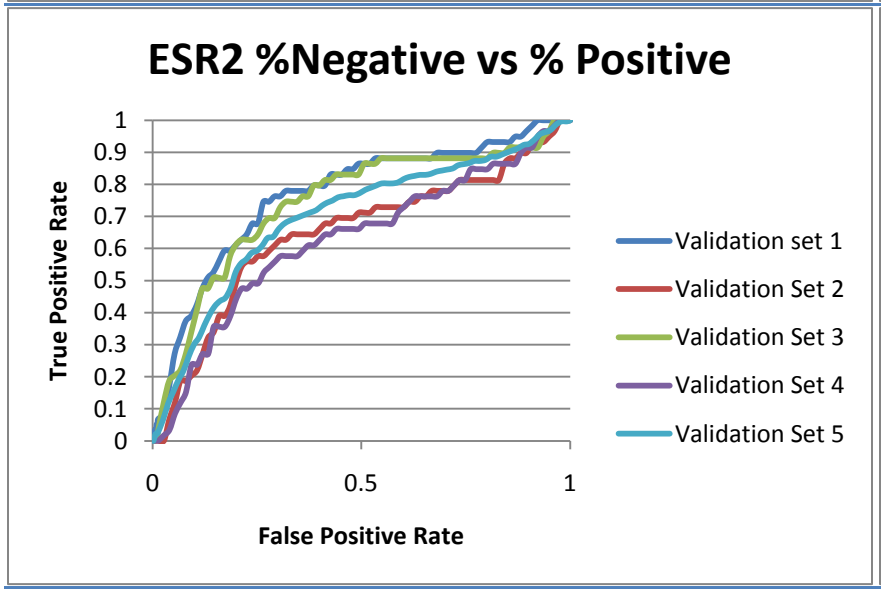

CA2 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

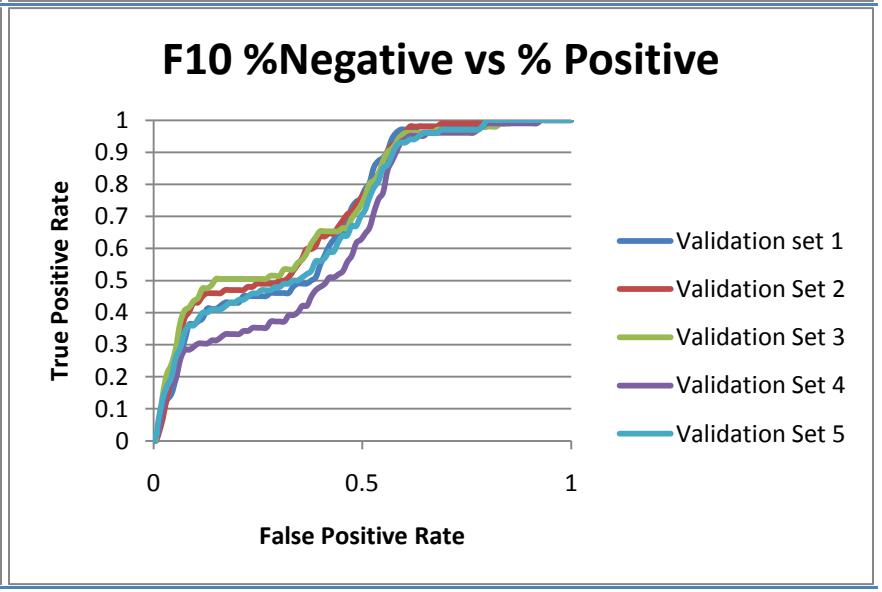

CDK2 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

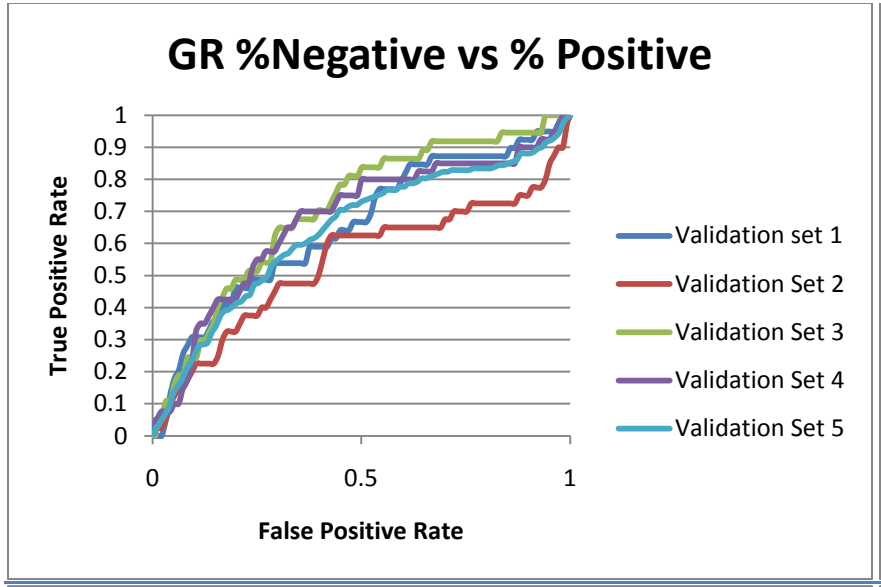

COX2 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

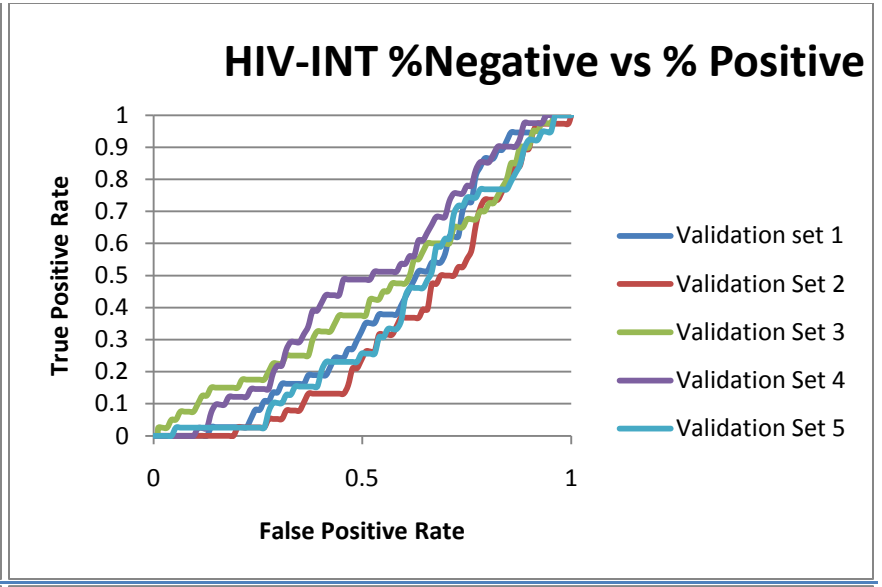

DHFR: %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation Set1
Validation set 2
Validation Set 3
Validation Set 4
Validation Set 5

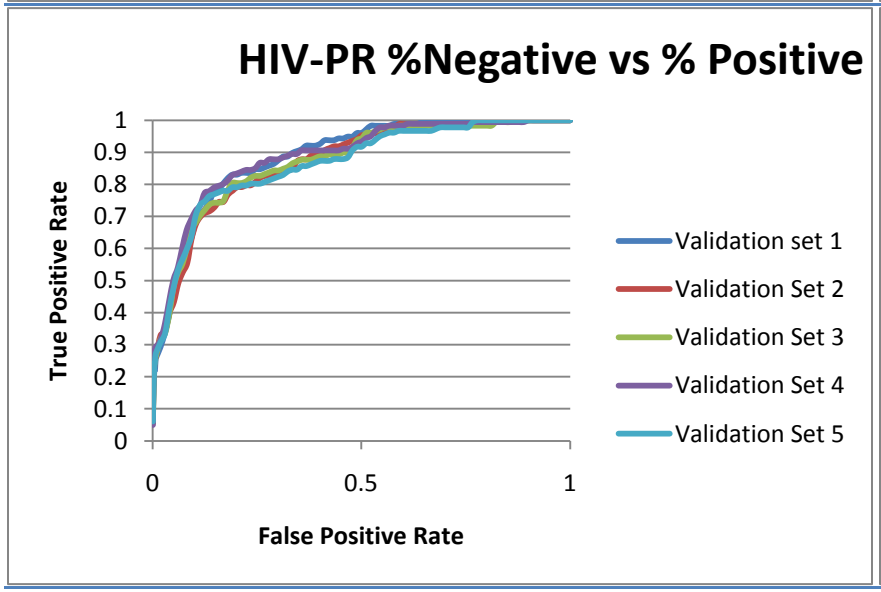

ESR1 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

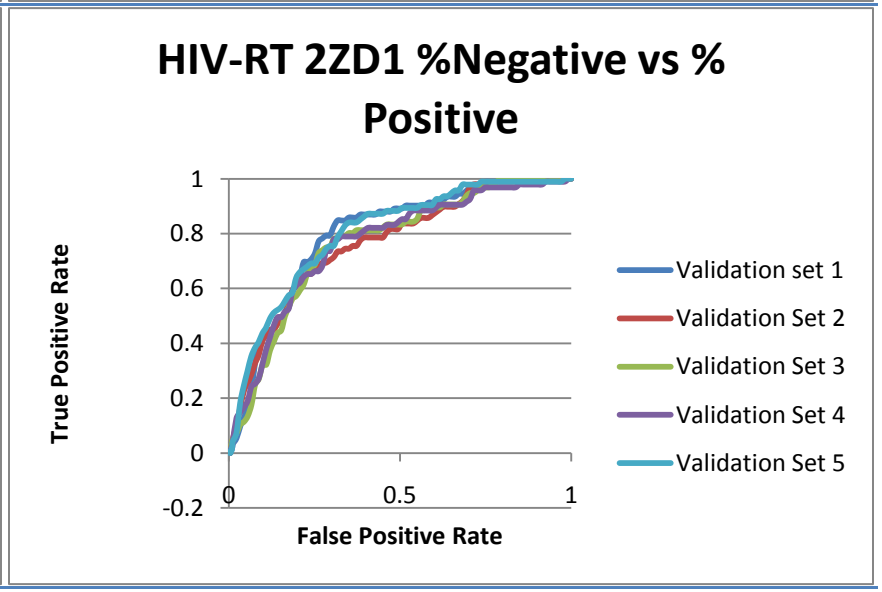

ESR2 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

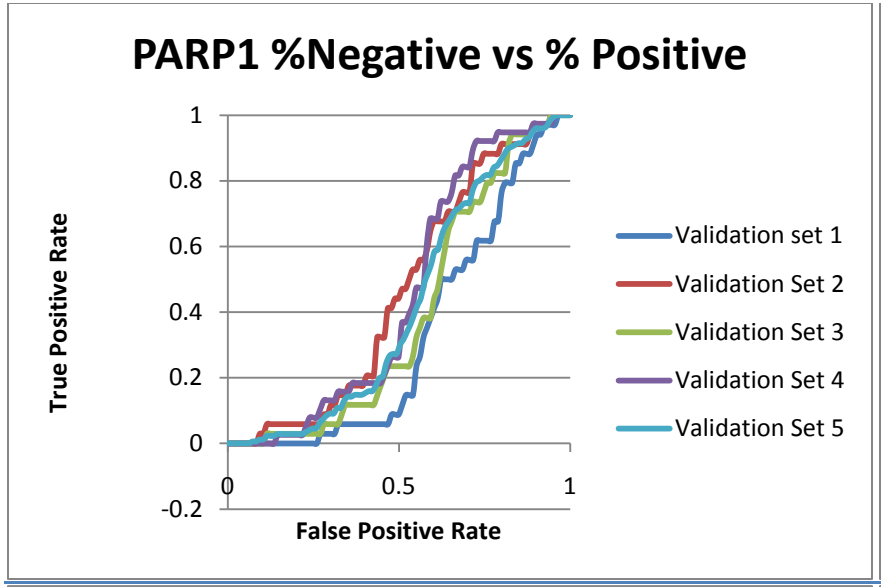

F10 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

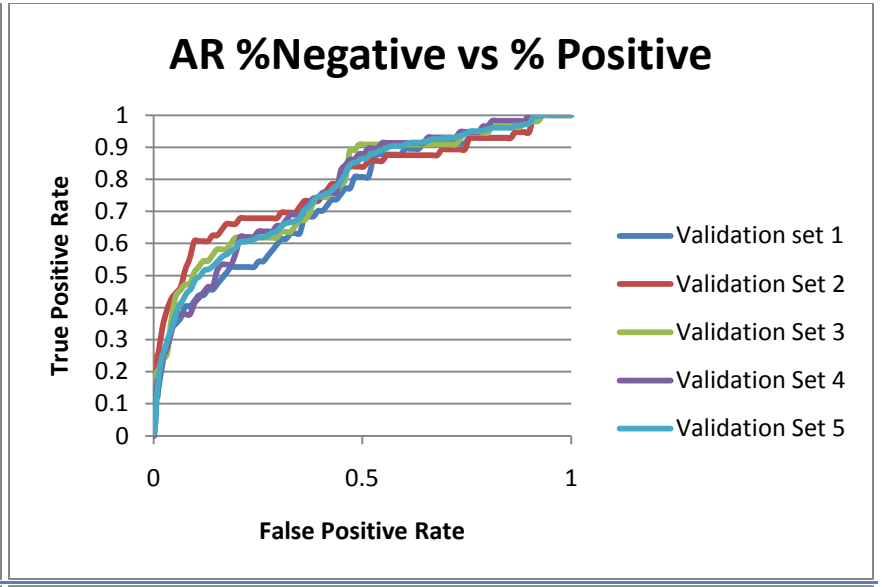

GR %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

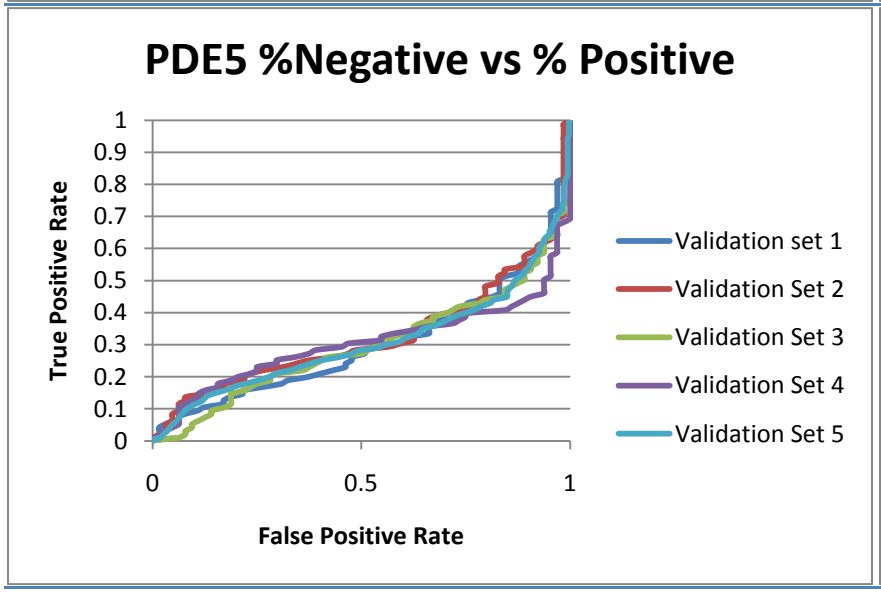

HIV-INT %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

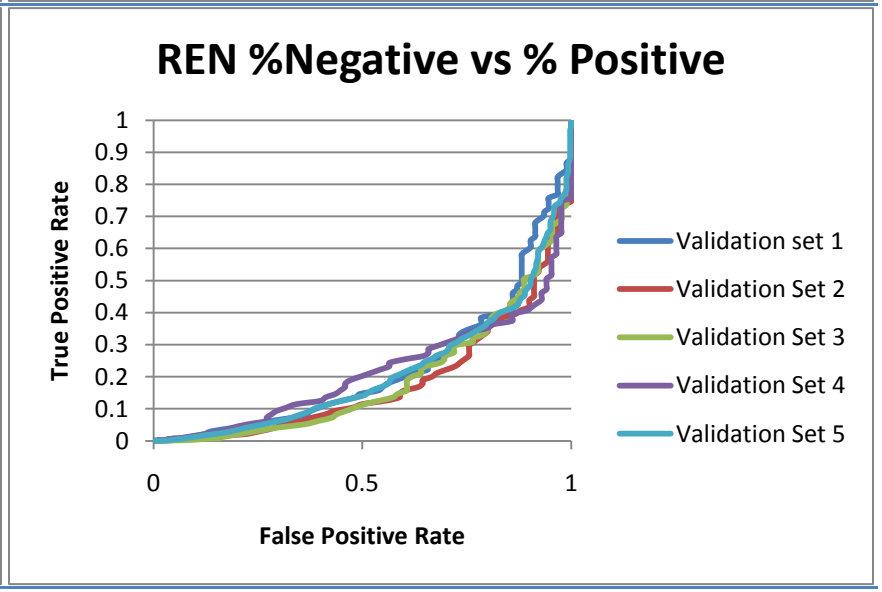

HIV-PR %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

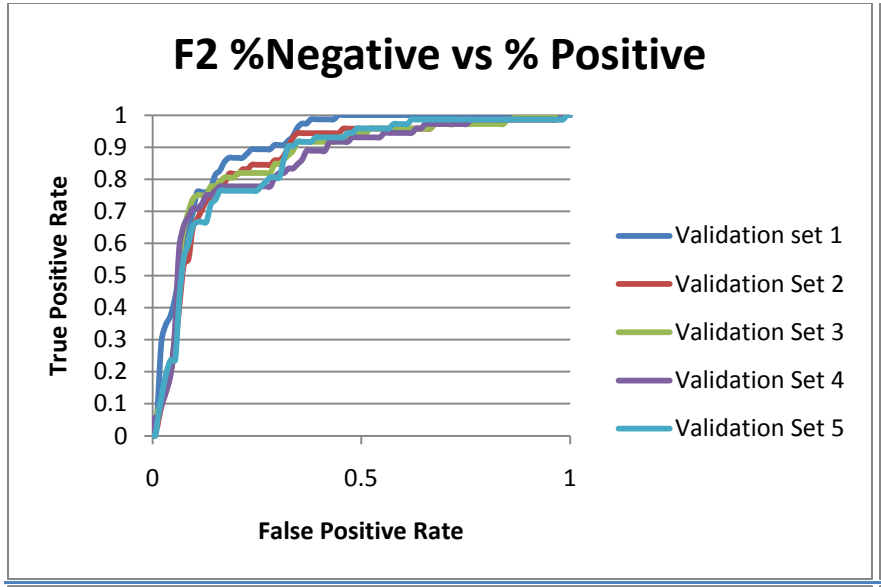

HIV-RT 2ZD1 %Negative vs % Positive
True Positive Rate
False Positive Rate
-0.2
0
0.2
0.4
0.6
0.8
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

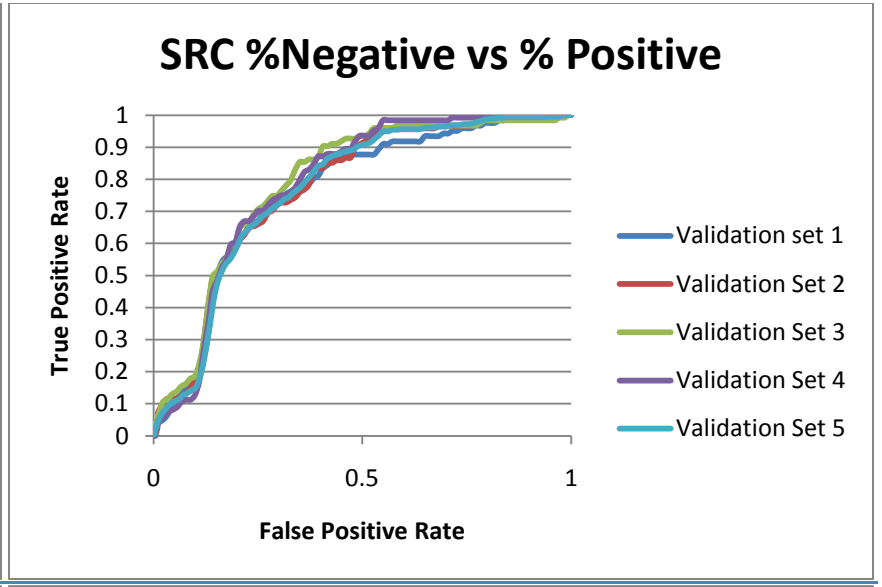

PARP1 %Negative vs % Positive
True Positive Rate
False Positive Rate
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

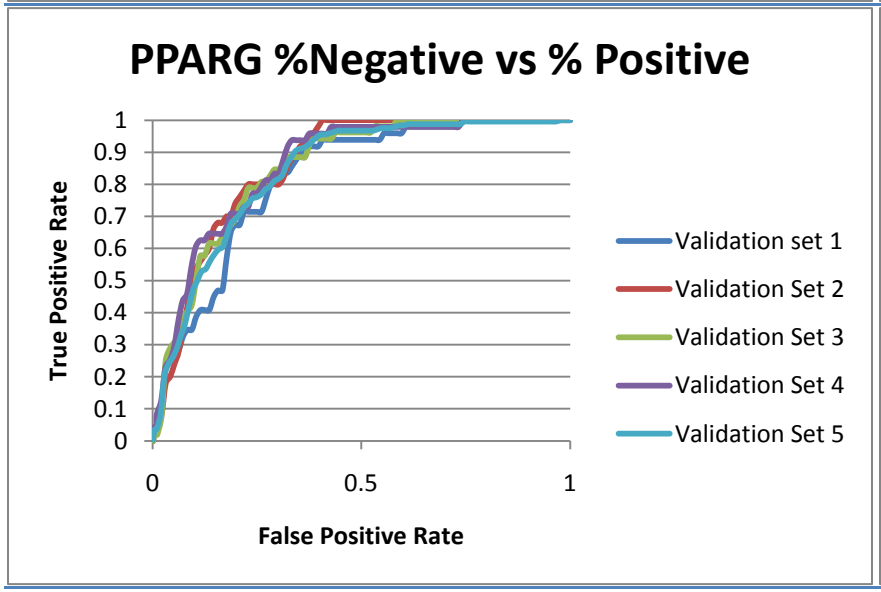

AR %Negative vs % Positive
True Positive Rate
False Positive Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

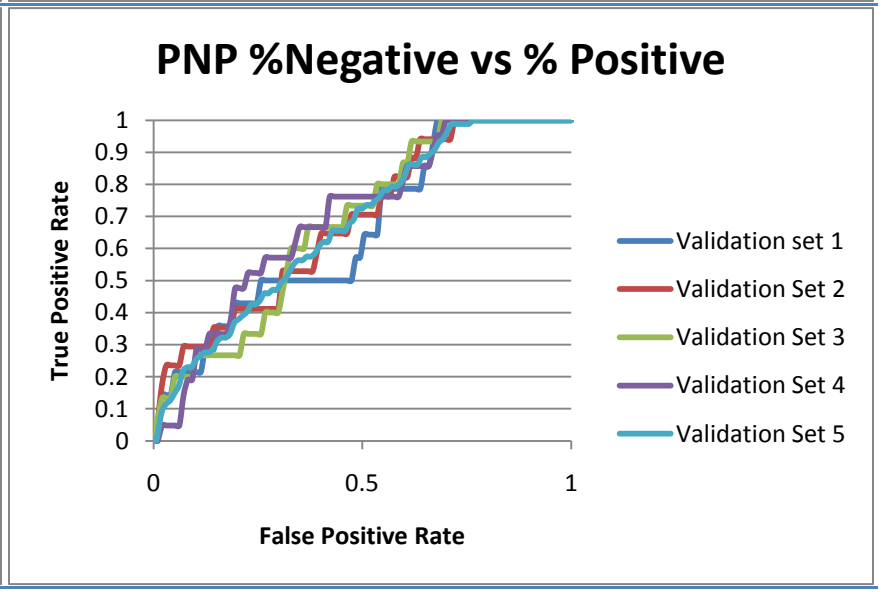

PDE5 %Negative vs % Positive
True Positive Rate
False Positive Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


REN %Negative vs % Positive
True Positive Rate
False Positive Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

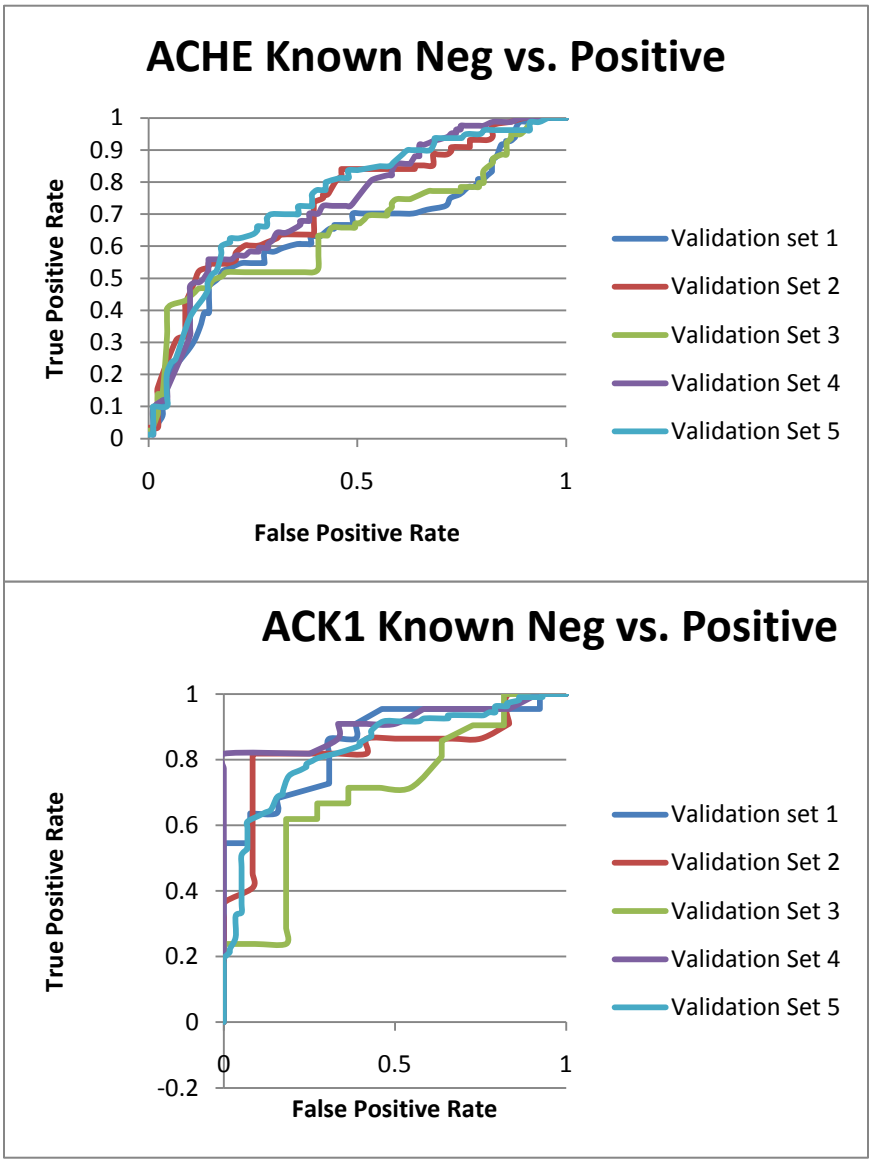

F2 %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

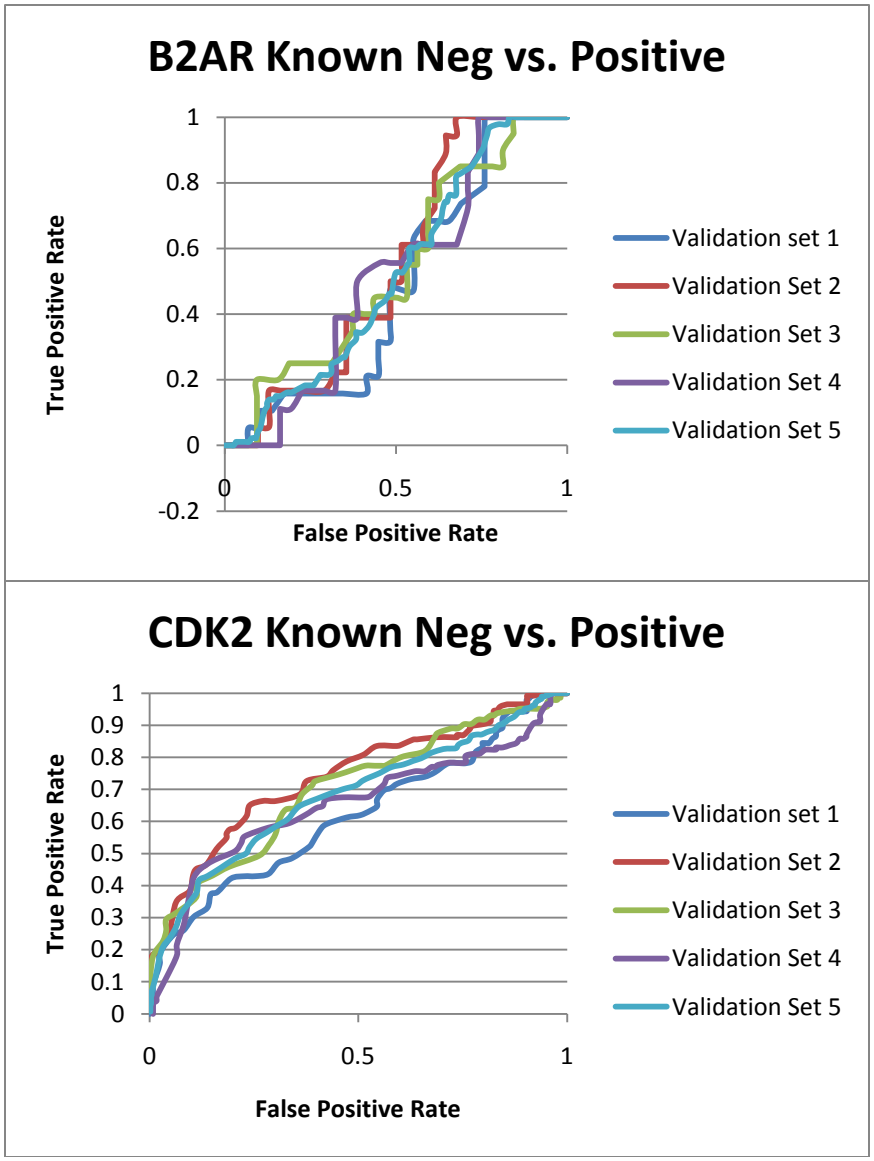

SRC %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

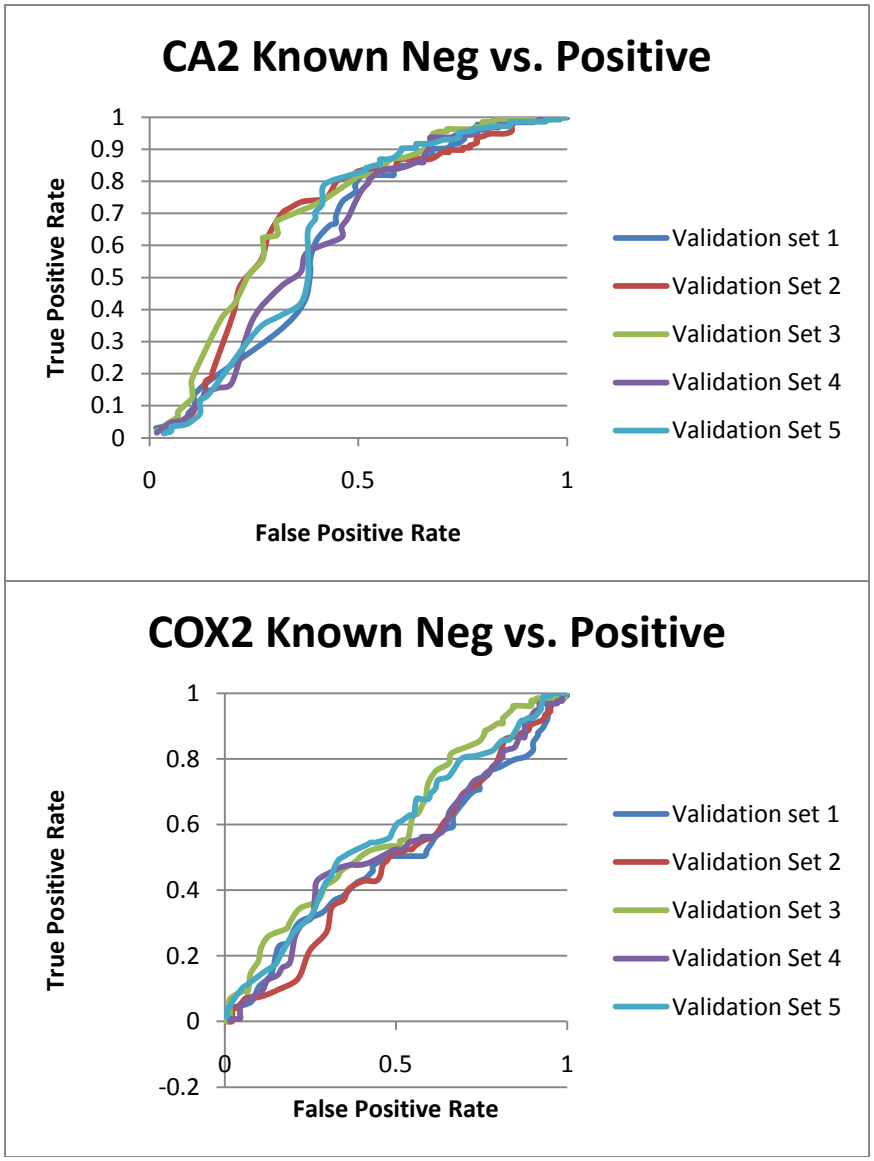

PPARG %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


PNP %Negative vs % Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

Compounds with Known Activity

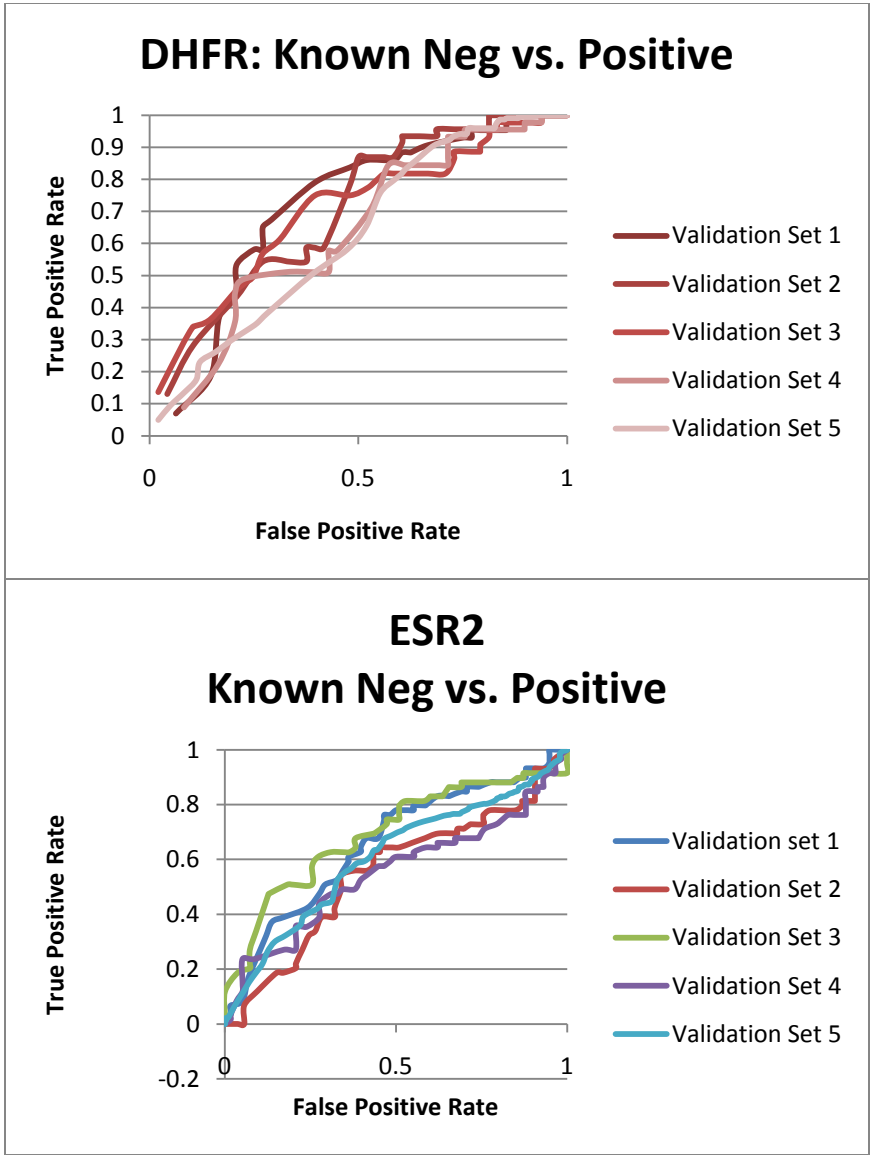


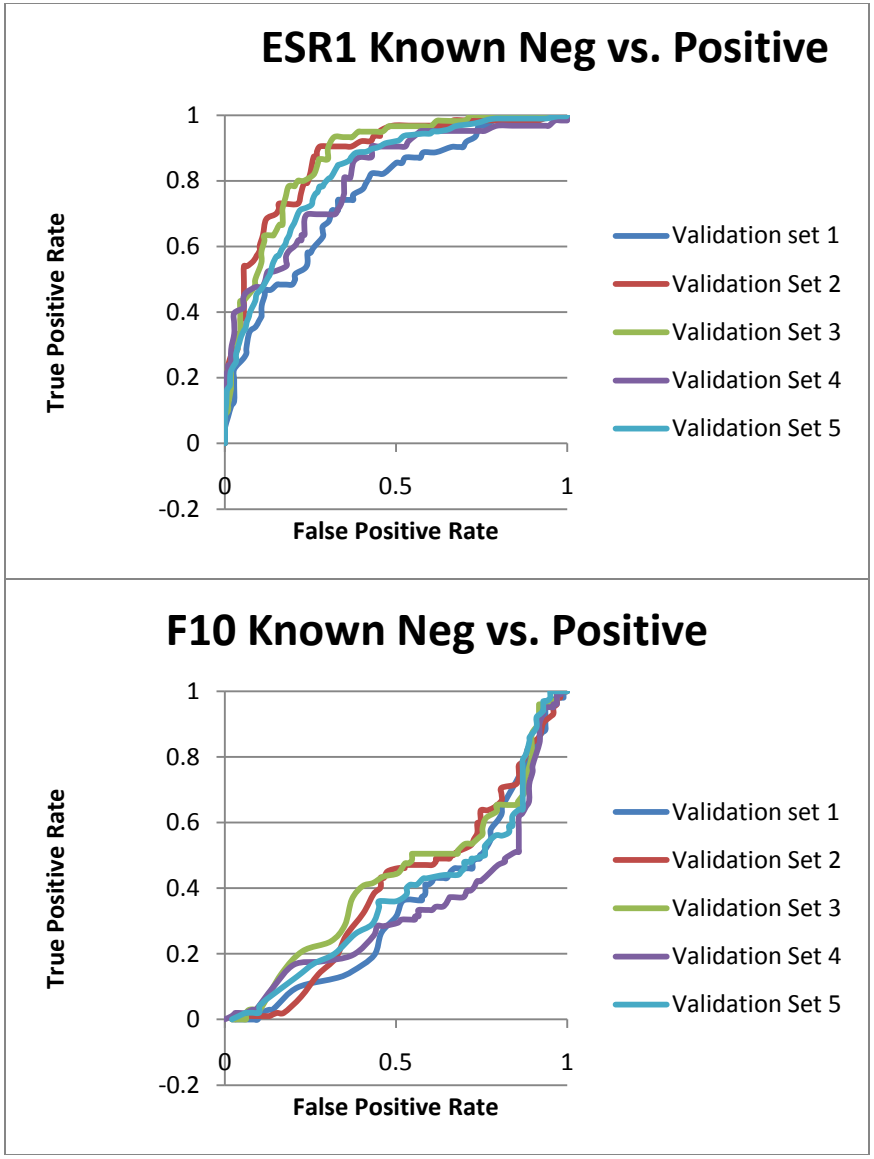

B2AR Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


CA2 Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

CDK2 Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

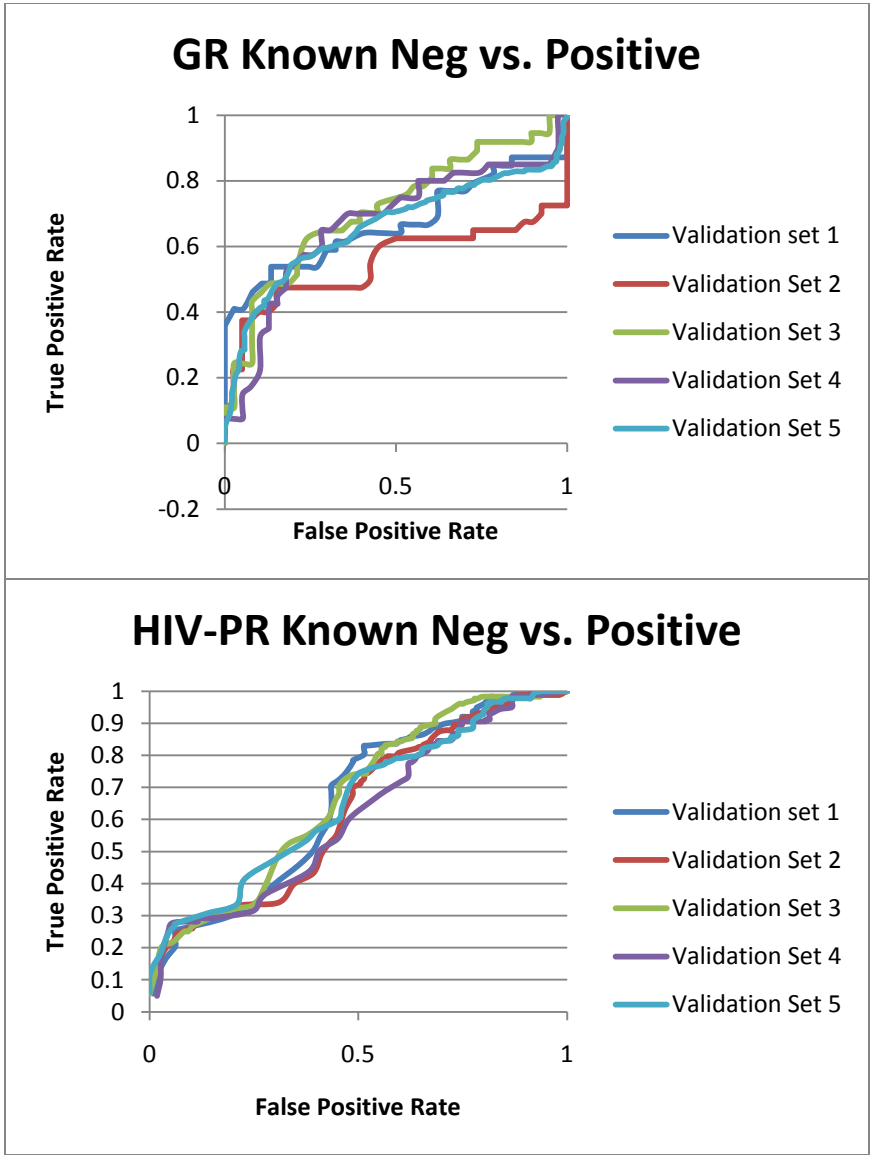
COX2 Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

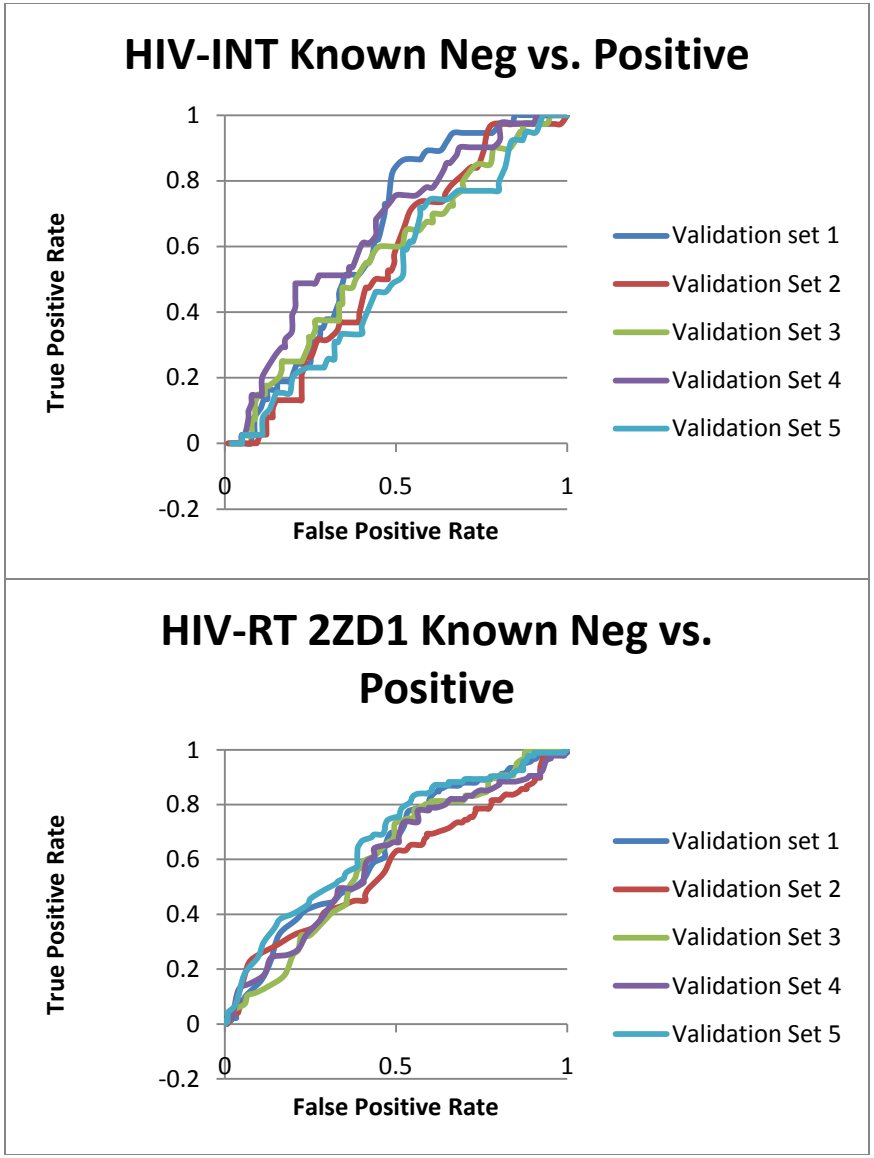

DHFR: Known Neg vs. Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation Set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


ESR1 Known Neg vs. Positive
True Positive Rate
False Positive Rate
-0.2
0
0.2
0.4
0.6
0.8
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

ESR2
Known Neg vs. Positive
True Positive Rate
False Positive Rate
-0.2
0
0.2
0.4
0.6
0.8
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

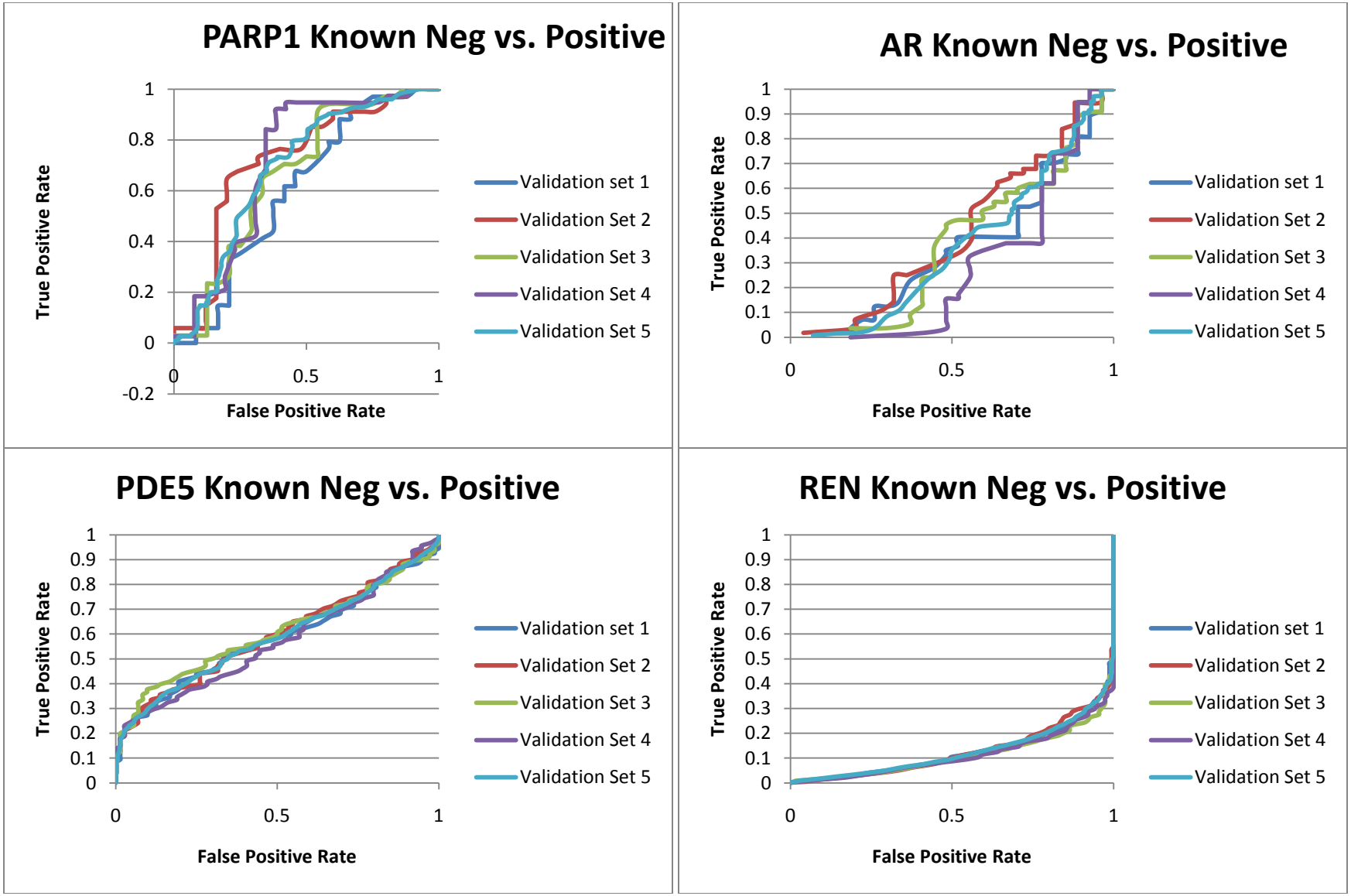
F10 Known Neg vs. Positive
True Positive Rate
False Positive Rate
-0.2
0
0.2
0.4
0.6
0.8
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

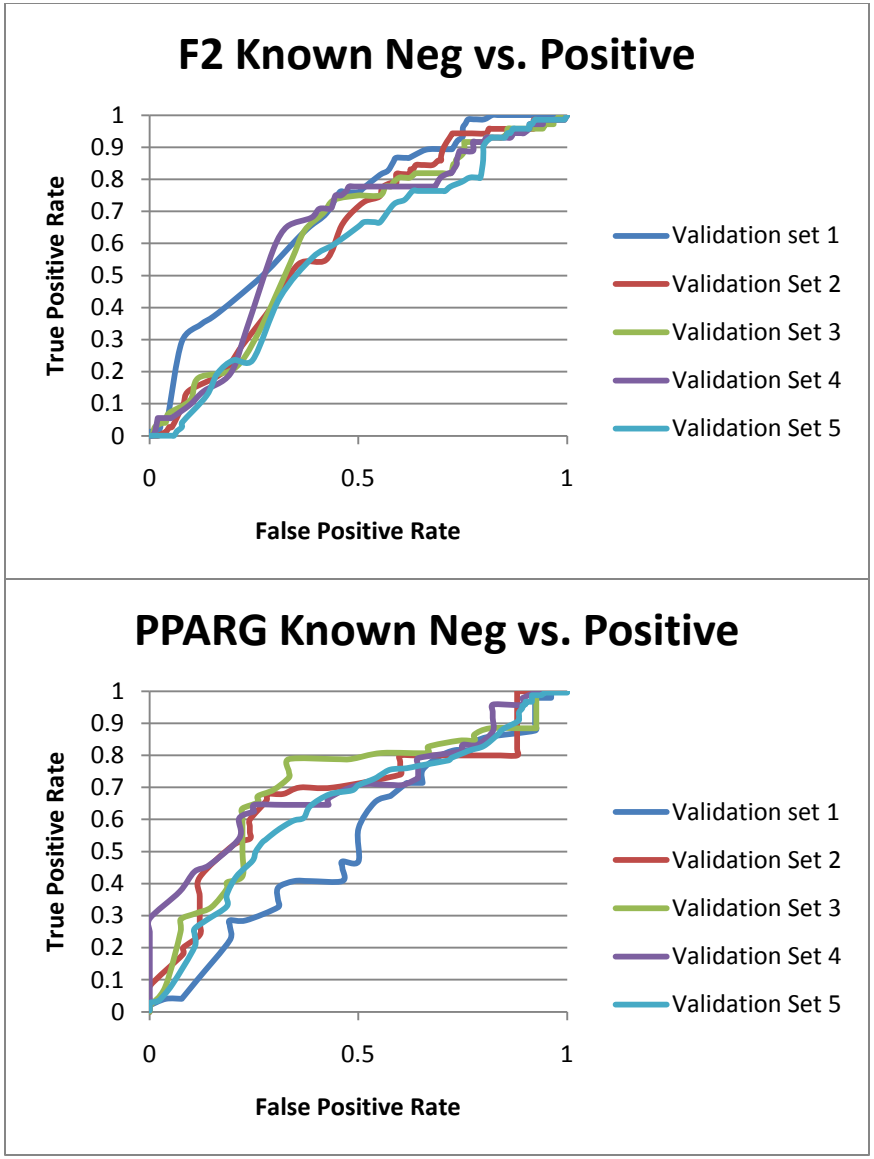

GR Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

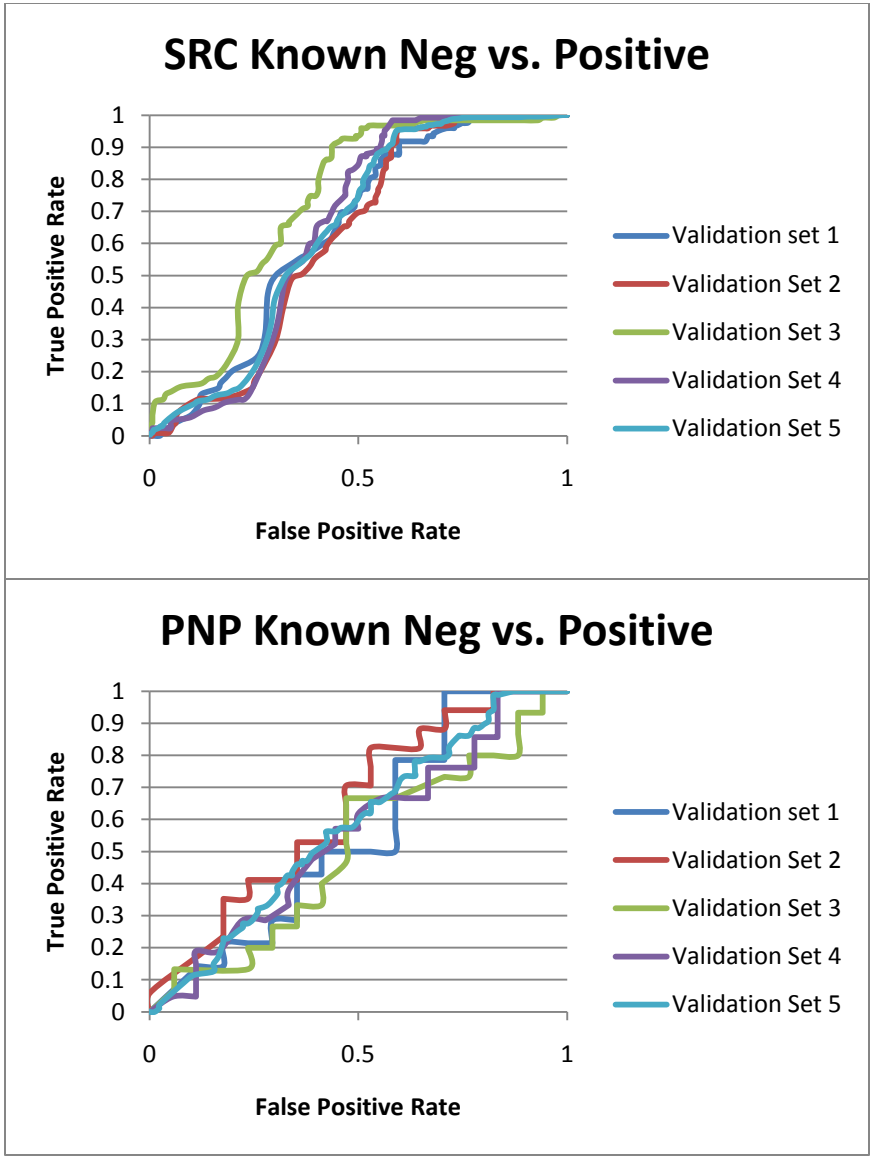

HIV-INT Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

HIV-PR Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

HIV-RT 2ZD1 Known Neg vs. Positive
True Positive Rate
False Positive Rate
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


PARP1 Known Neg vs. Positive
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
AR Known Neg vs. Positive
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PDE5 Known Neg vs. Positive
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
REN Known Neg vs. Positive
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


F2 Known Neg vs. Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


SRC Known Neg vs. Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

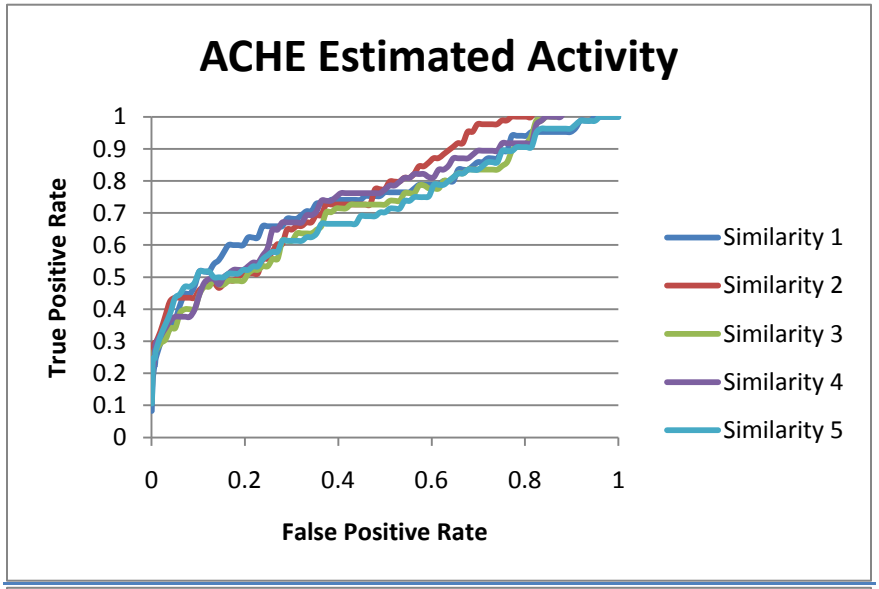
PPARG Known Neg vs. Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

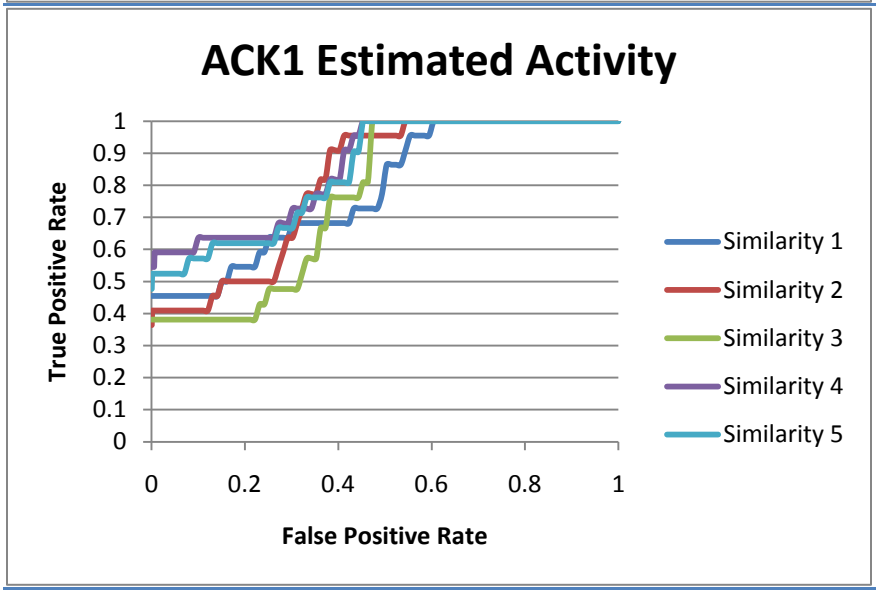
PNP Known Neg vs. Positive
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

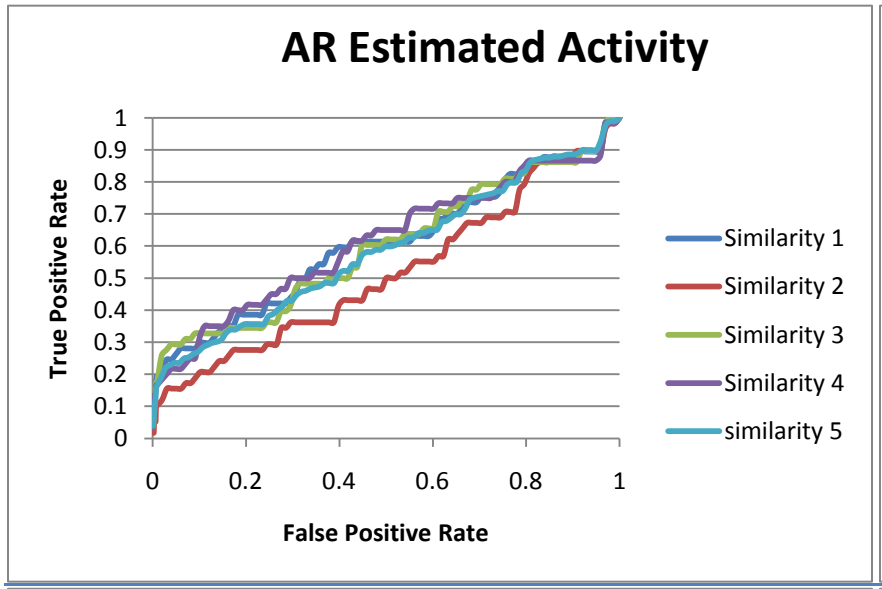
Similarity Searching

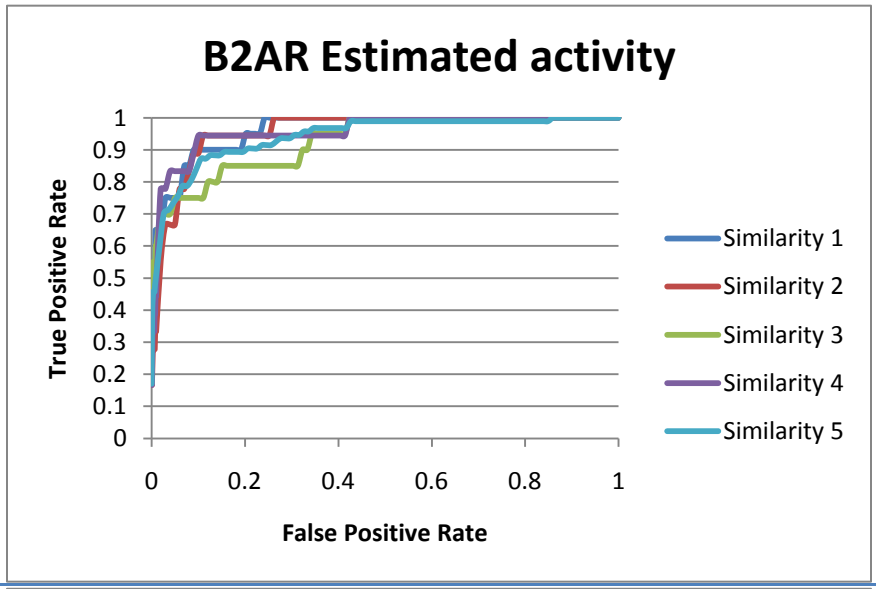
ROC curves were generated for similarity search when done by two different probe sets, the ligand contained within the PDB entry that was used for docking and the full modeling set. When using only the pdb ligand as a probe, the accuracy of ranking is very uncertain. Occasionally the ranking is excellent as in the cases of B2AR, ESR2, and DHFR. However, it is just as frequently terrible as in the cases of GR, HIV-Int, and PNP. This is mirrored in the tanking of compounds with known activities, but accuracy is always lower than that obtained on the entire dataset. Using the entire modeling set as probes, similarity searching provides nearly excellent ranking of the full database for every target. Its ranking of compounds with known activity is not quite as good, but is still very acceptable.

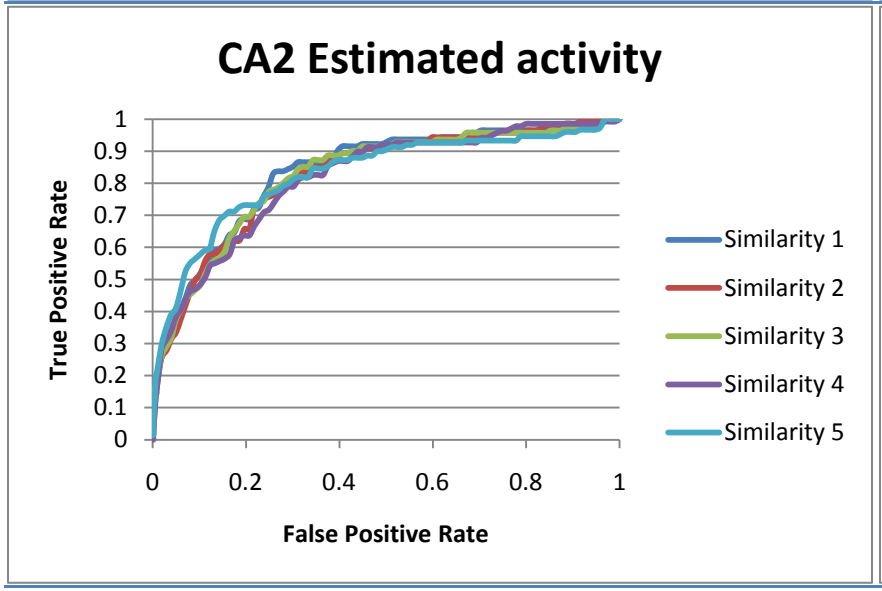
Full Screening Sets-PDB ligand

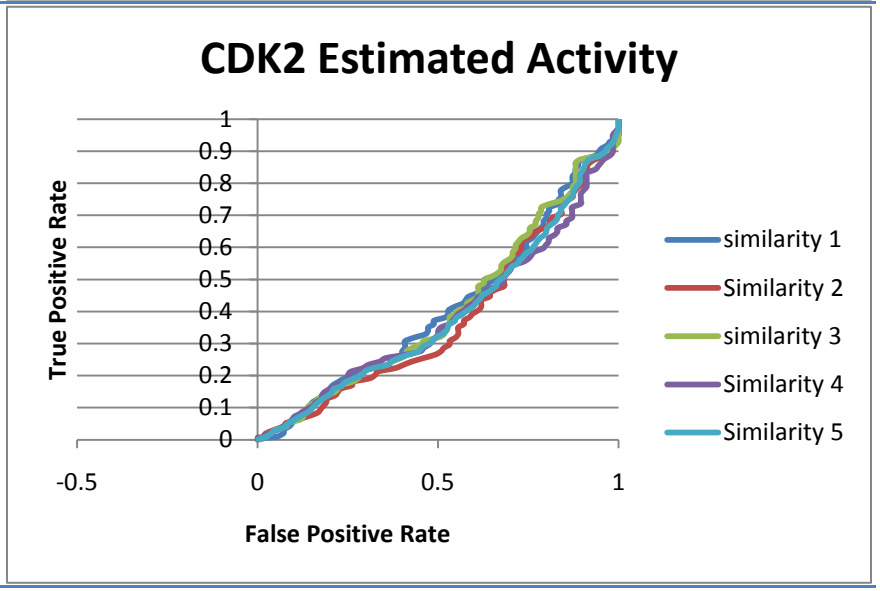


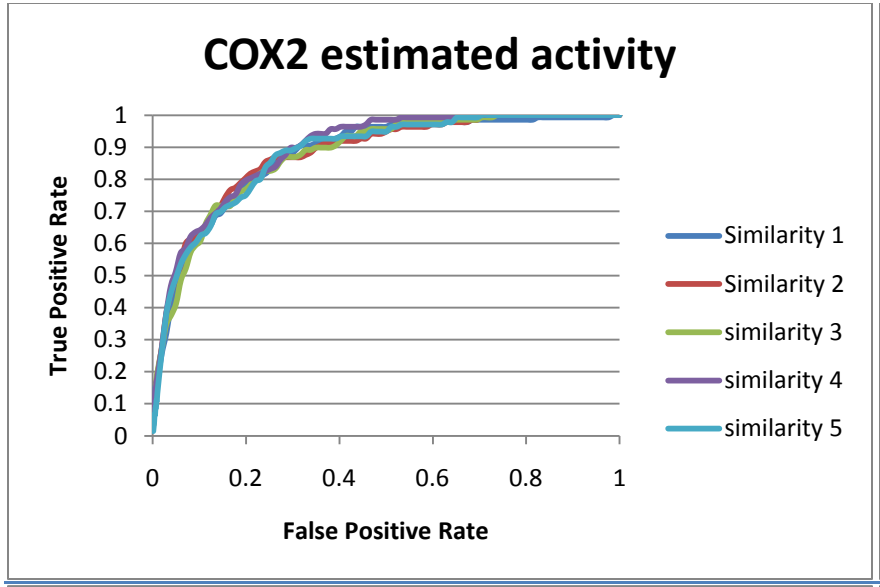


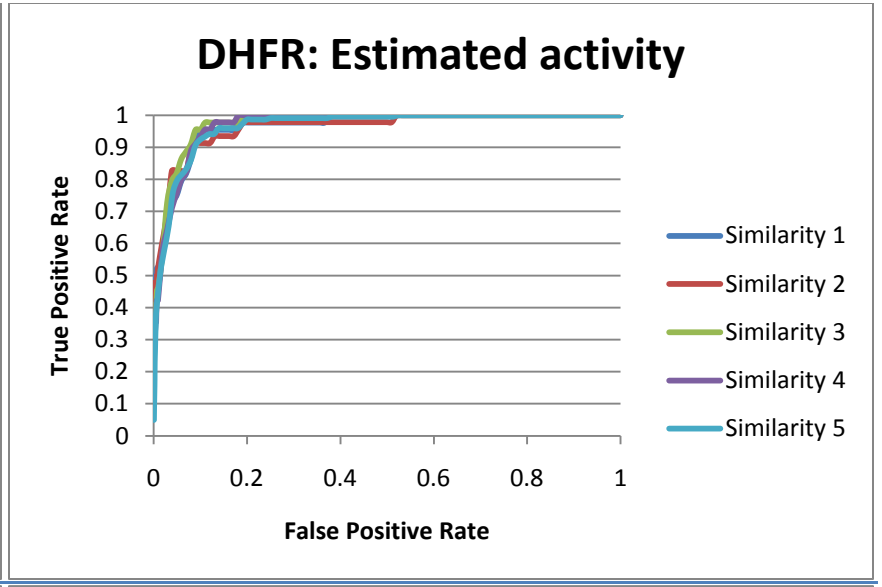

AR Estimated Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
similarity 5

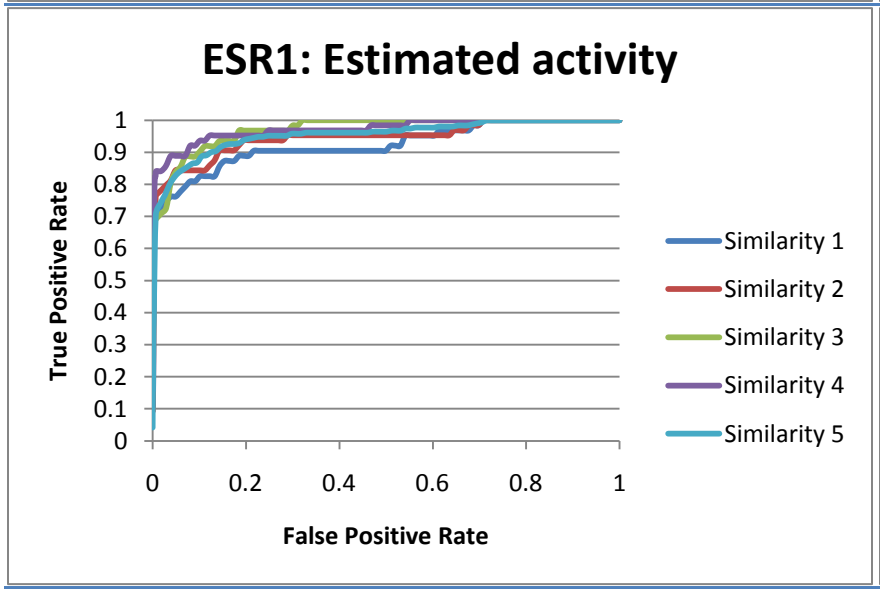

B2AR Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

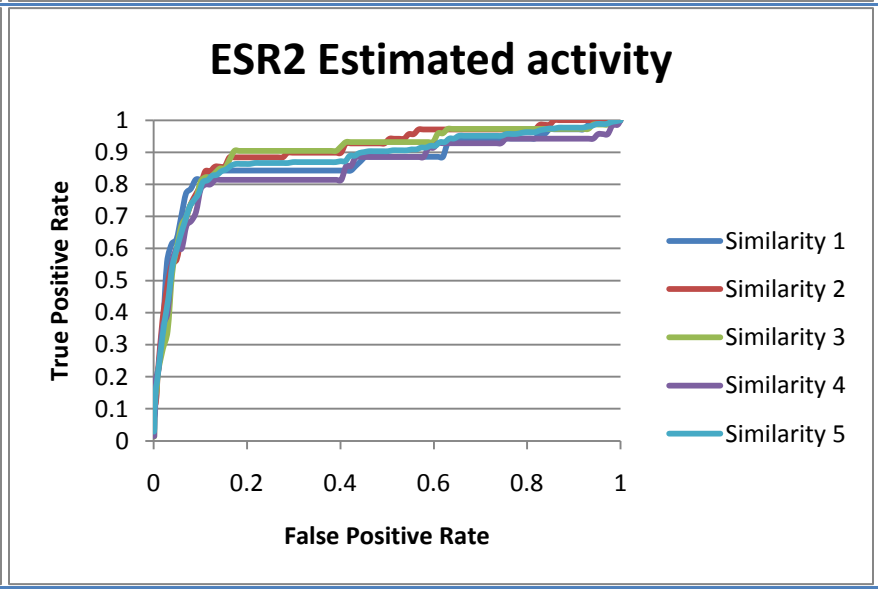

CA2 Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

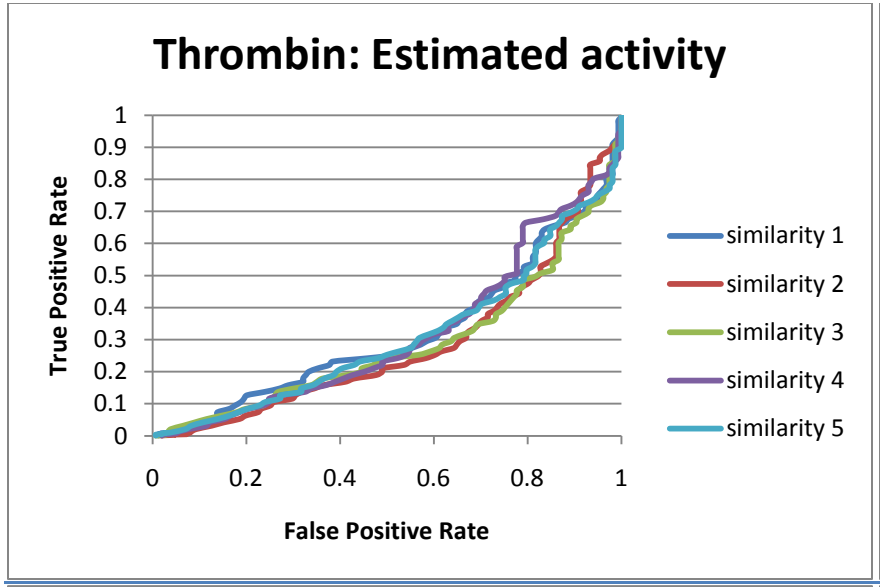

CDK2 Estimated Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-0.5
0
0.5
1
similarity 1
Similarity 2
similarity 3
Similarity 4
Similarity 5

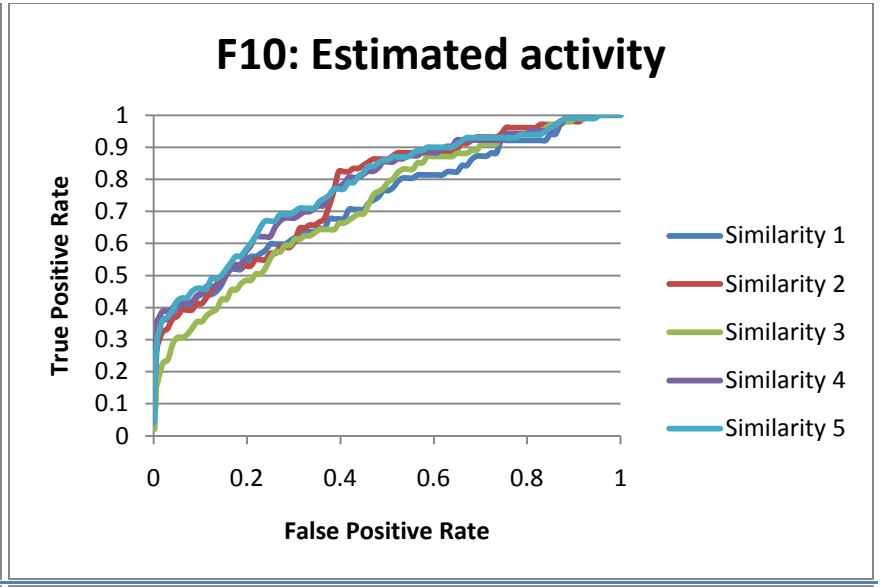

COX2 estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
similarity 3
similarity 4
similarity 5

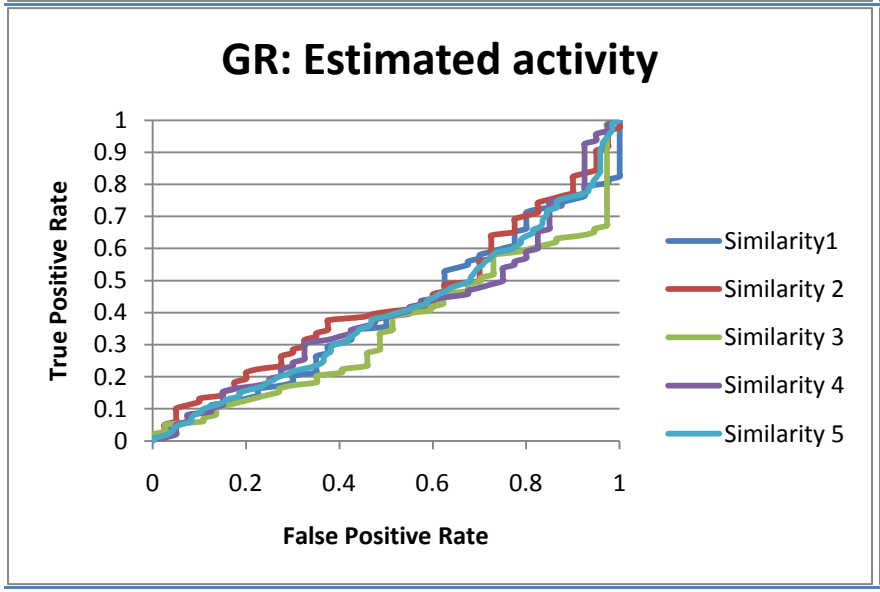

DHFR: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

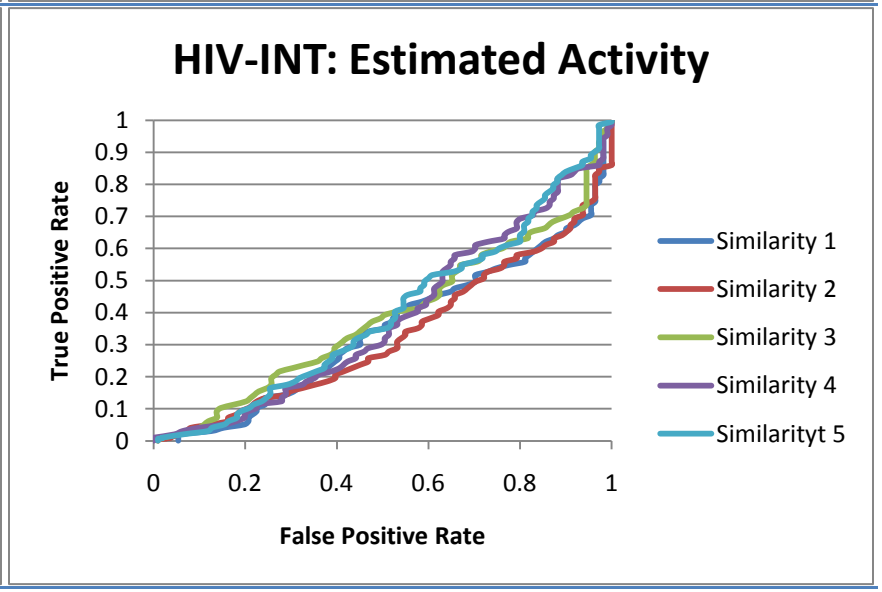

ESR1: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

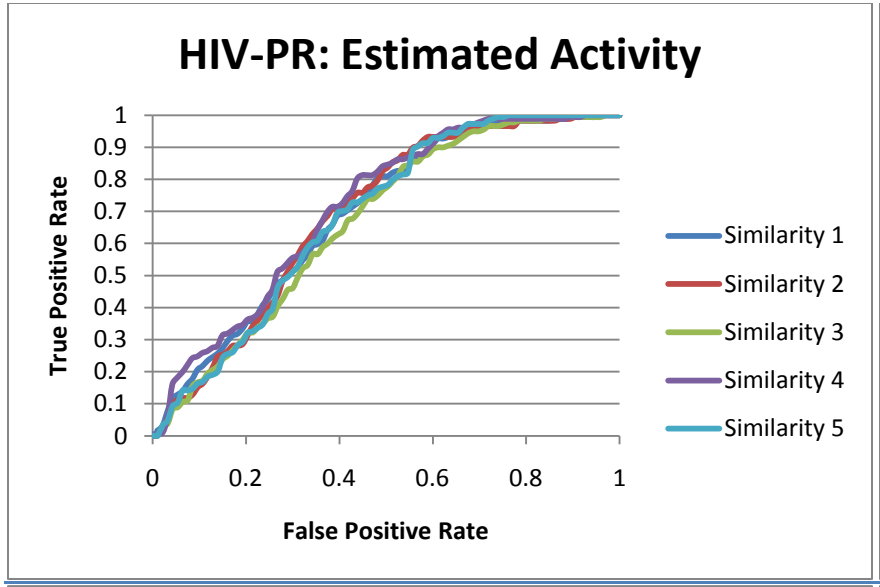

ESR2 Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

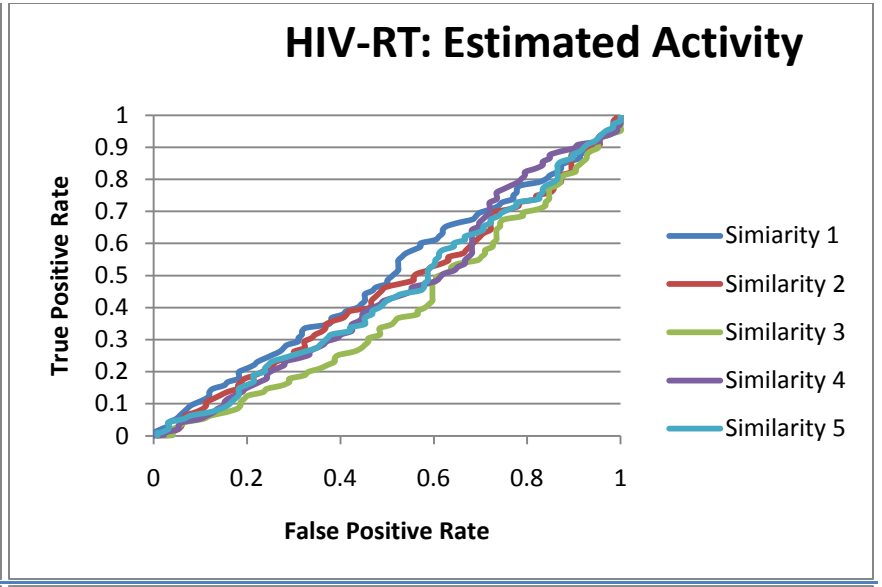

Thrombin: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
similarity 1
similarity 2
similarity 3
similarity 4
similarity 5

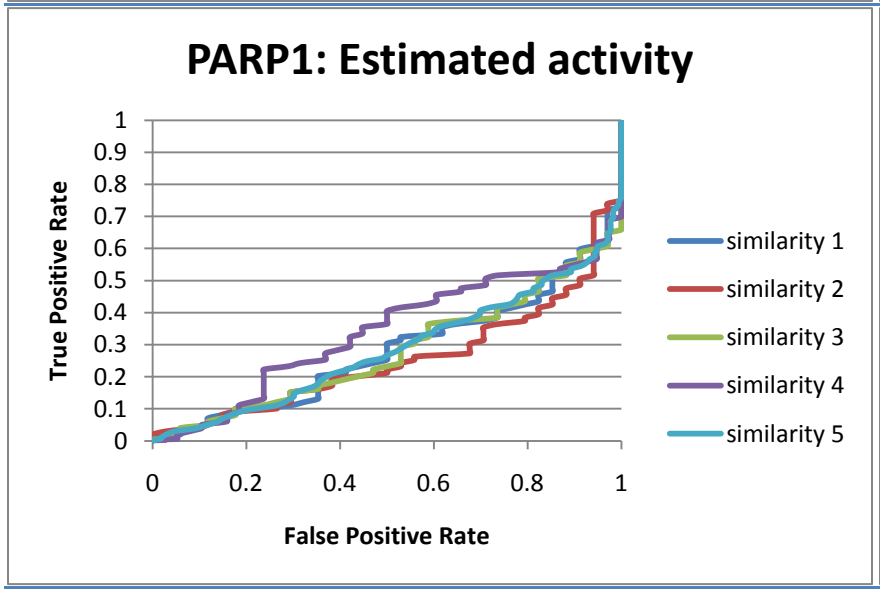

F10: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

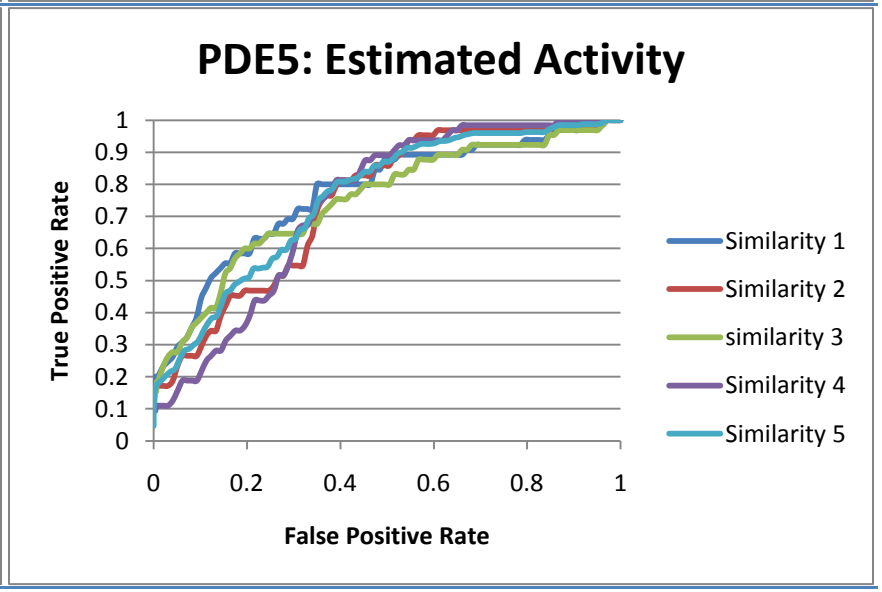

GR: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

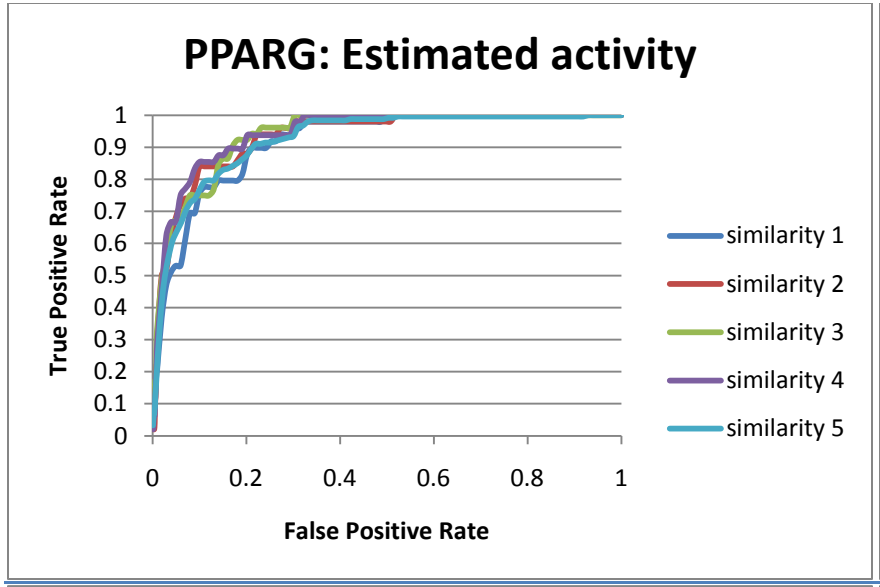

HIV-INT: Estimated Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarityt 5

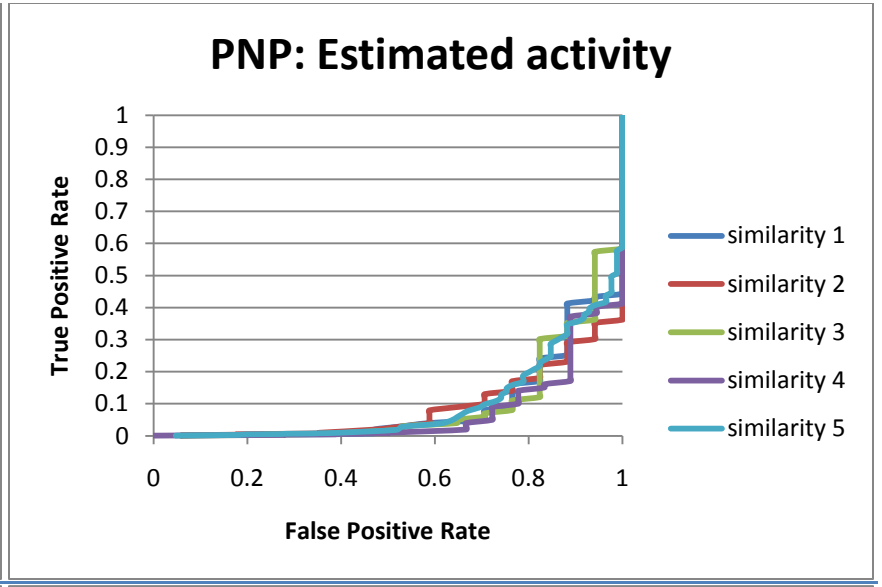

HIV-PR: Estimated Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

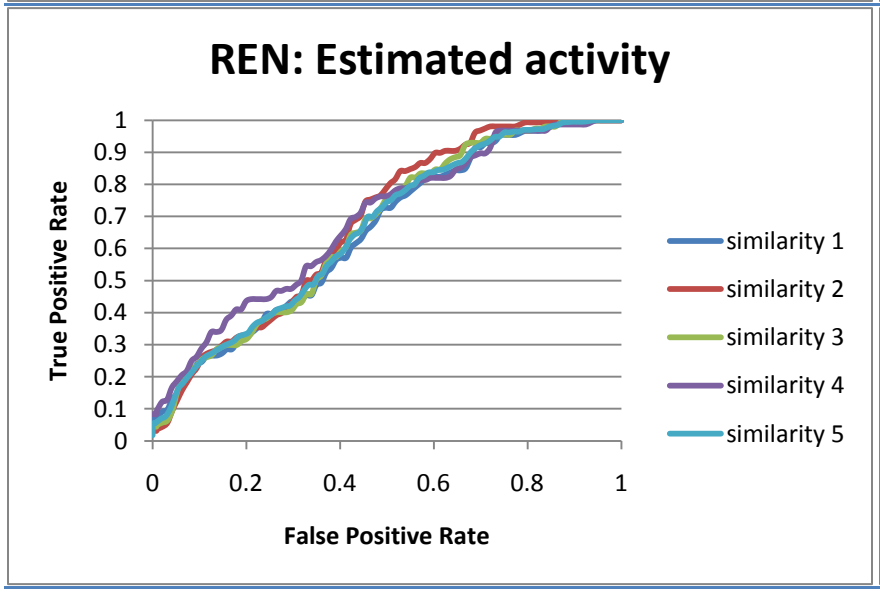

HIV-RT: Estimated Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Simiarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

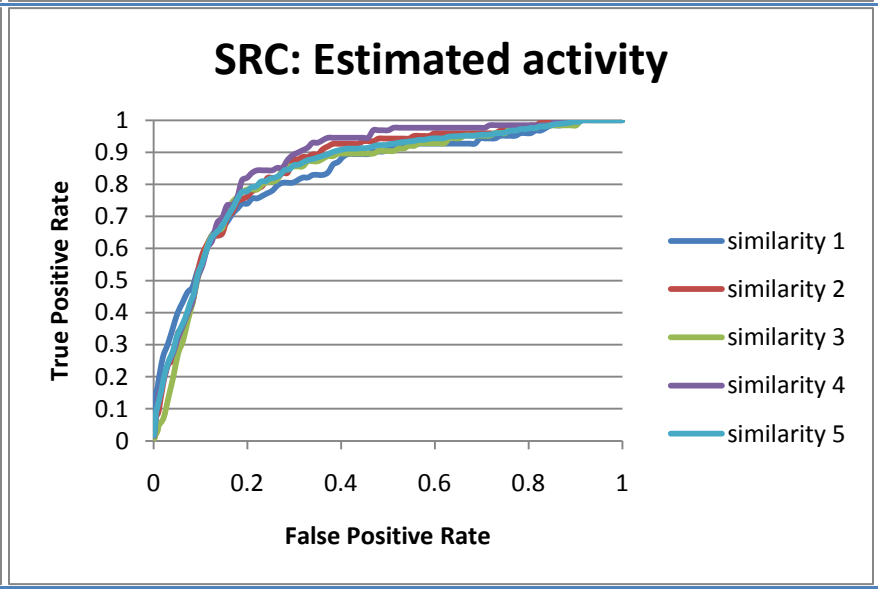

PARP1: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
similarity 1
similarity 2
similarity 3
similarity 4
similarity 5


PDE5: Estimated Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
similarity 3
Similarity 4
Similarity 5

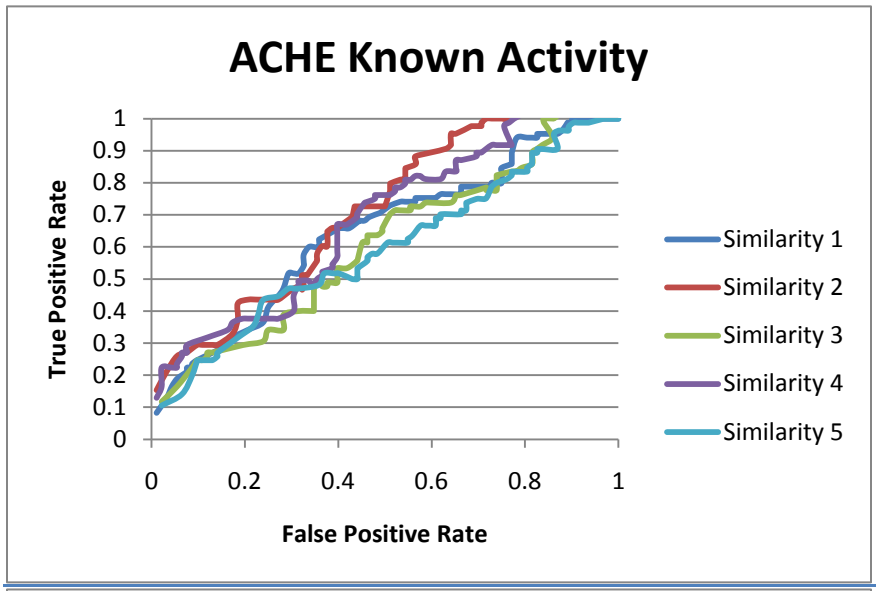

PPARG: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
similarity 1
similarity 2
similarity 3
similarity 4
similarity 5

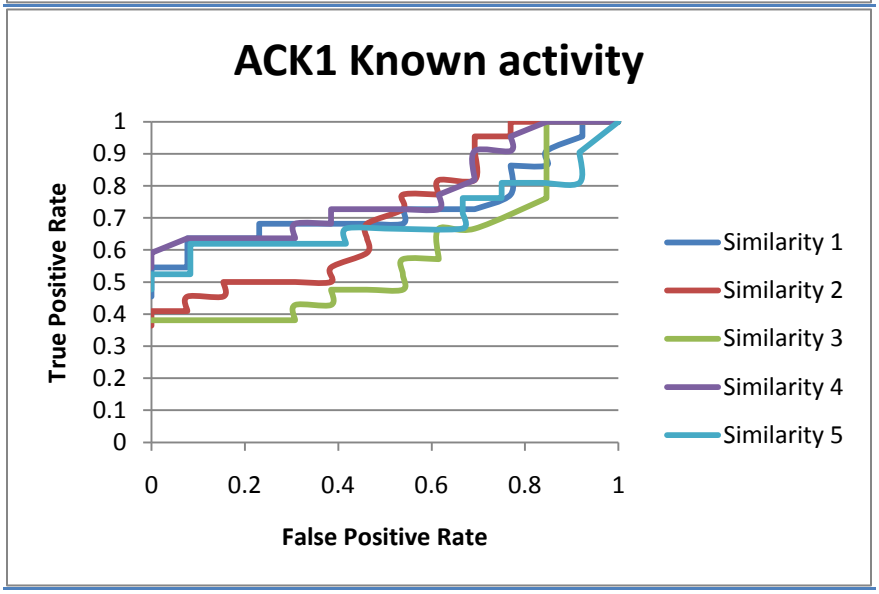

PNP: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
similarity 1
similarity 2
similarity 3
similarity 4
similarity 5

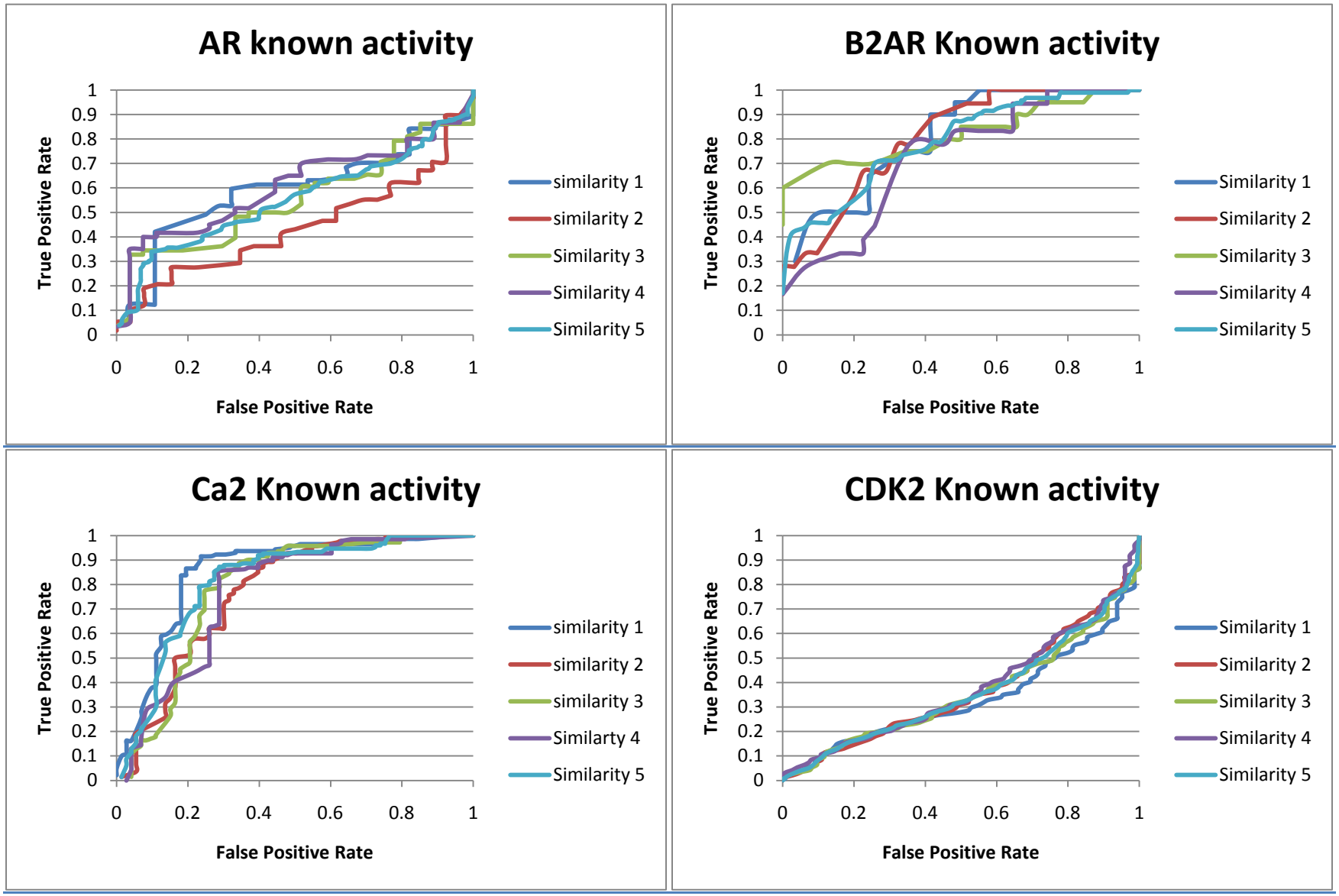

REN: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
similarity 1
similarity 2
similarity 3
similarity 4
similarity 5

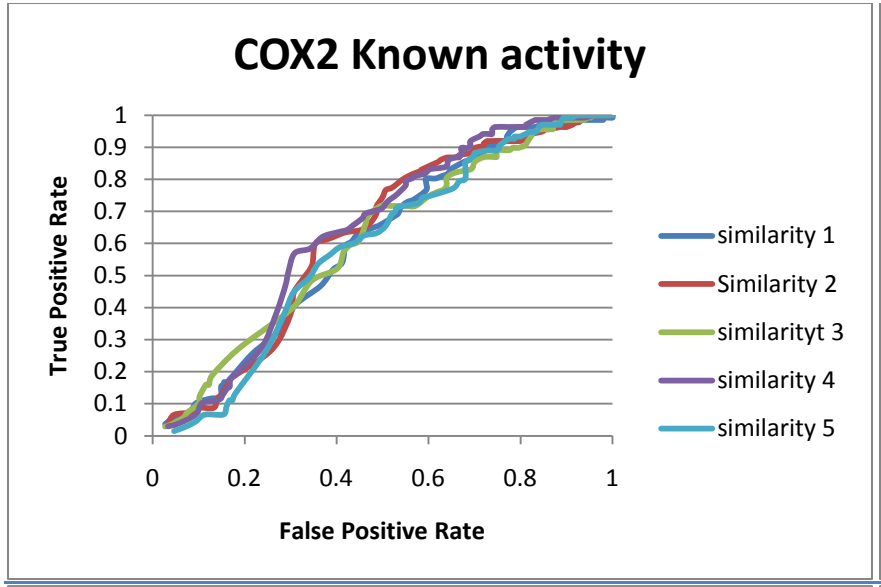

SRC: Estimated activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
similarity 1
similarity 2
similarity 3
similarity 4
similarity 5

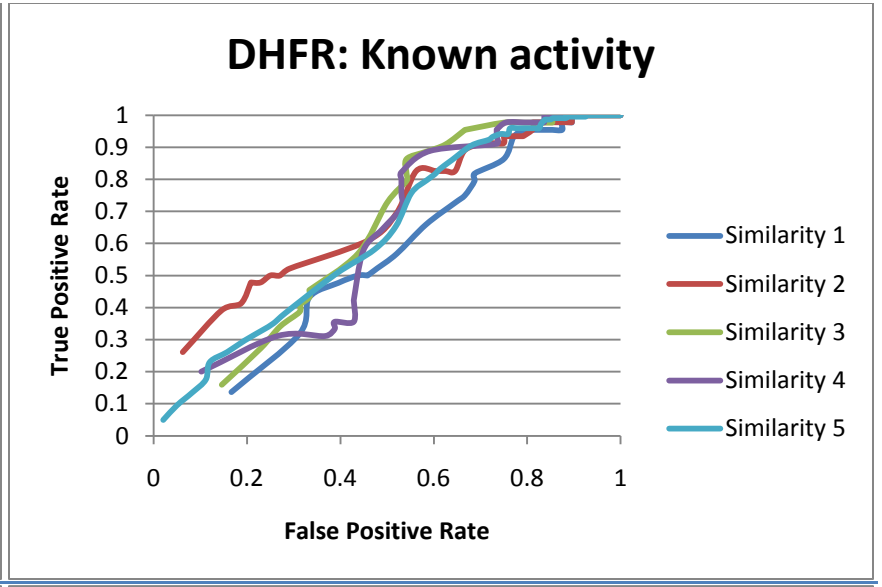
Compounds with Known Activities – PDB Ligand

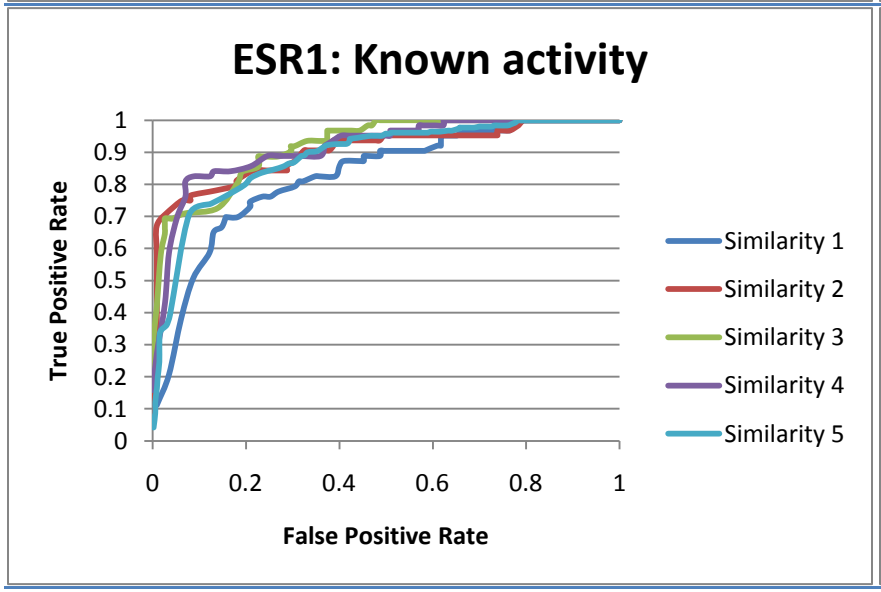


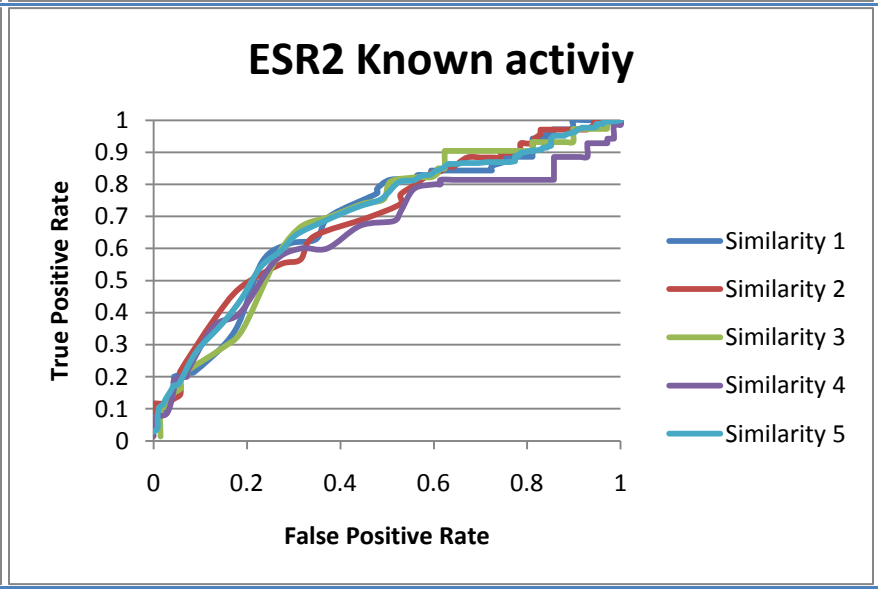


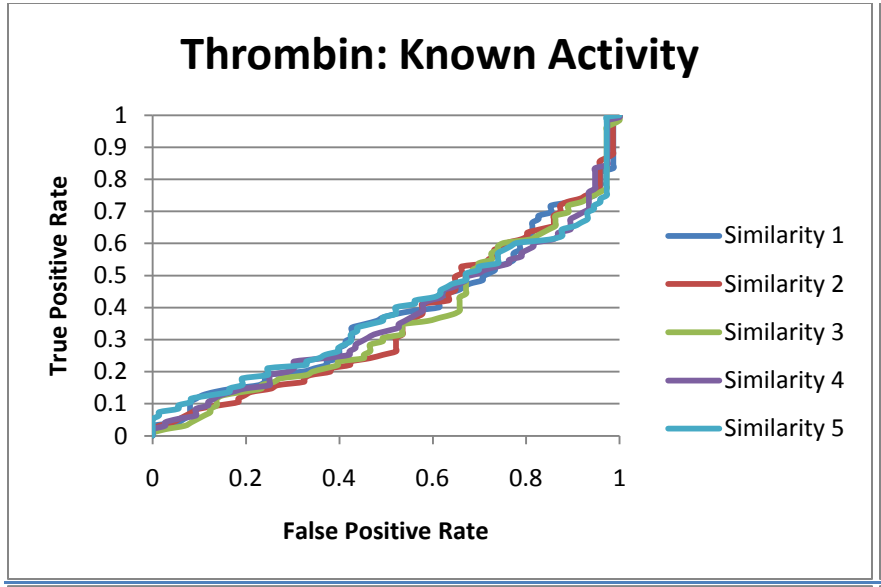

AR known activity
True Positive Rate
False Positive Rate
similarity 1
similarity 2
Similarity 3
Similarity 4
Similarity 5
B2AR Known activity
True Positive Rate
False Positive Rate
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5
Ca2 Known activity
True Positive Rate
False Positive Rate
similarity 1
similarity 2
similarity 3
Similarty 4
Similarity 5
CDK2 Known activity
True Positive Rate
False Positive Rate
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

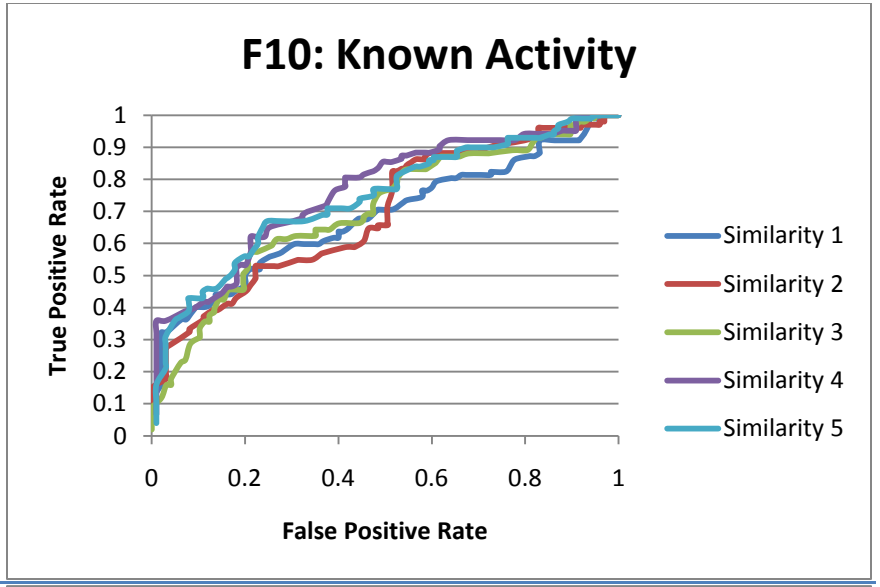

COX2 Known activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
similarity 1
Similarity 2
similarityt 3
similarity 4
similarity 5

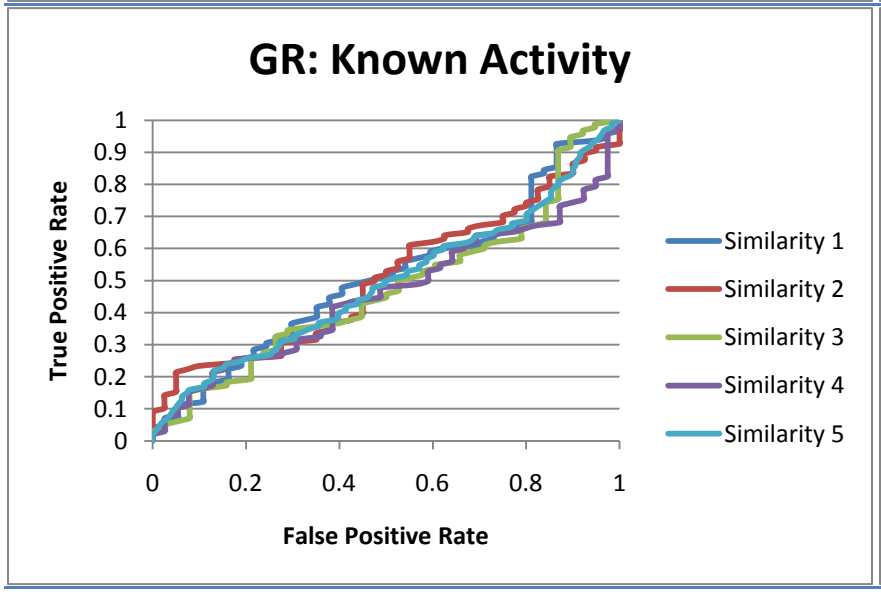

DHFR: Known activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

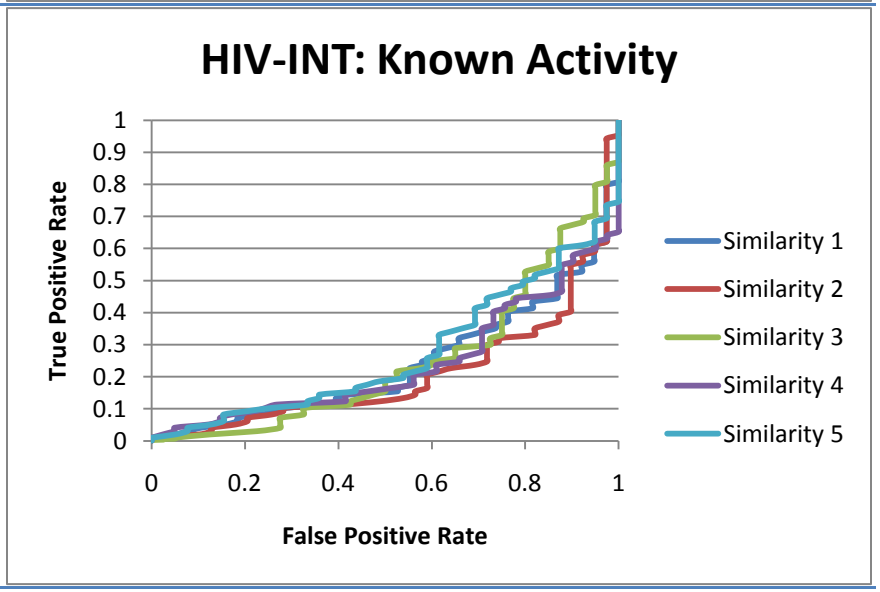

ESR1: Known activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

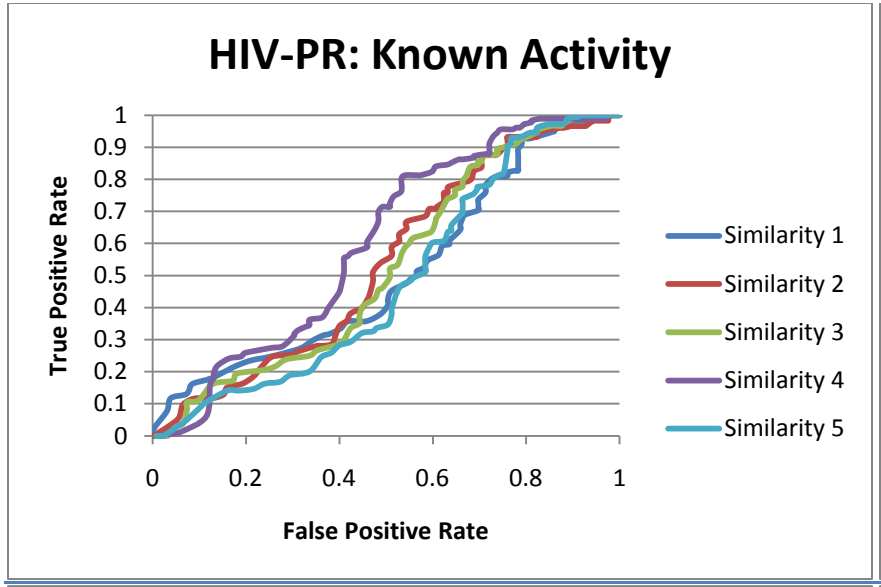

ESR2 Known activiy
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

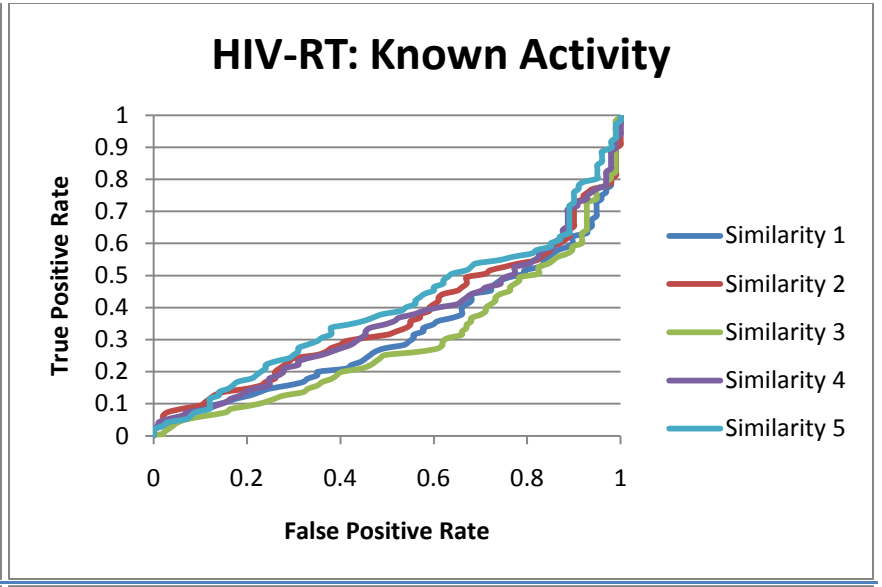

Thrombin: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

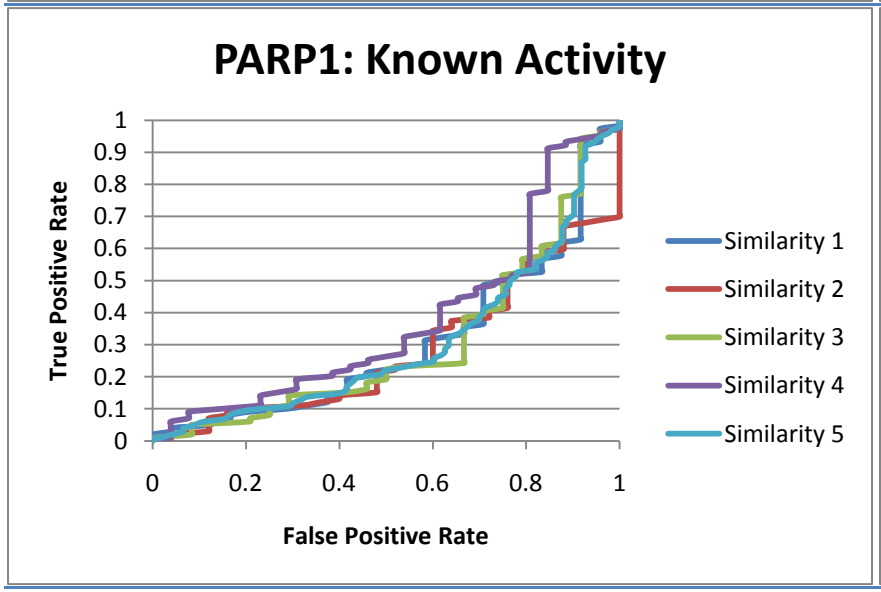

F10: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

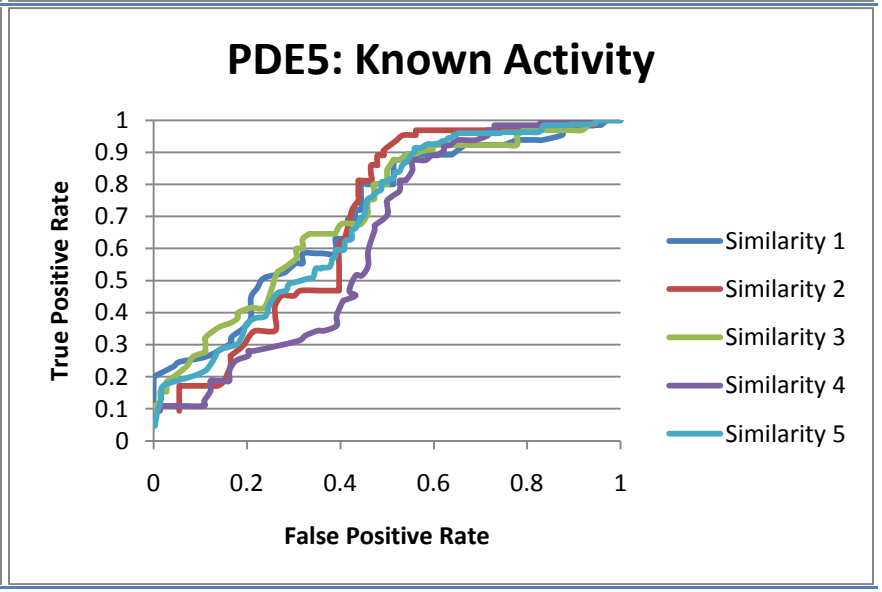

GR: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

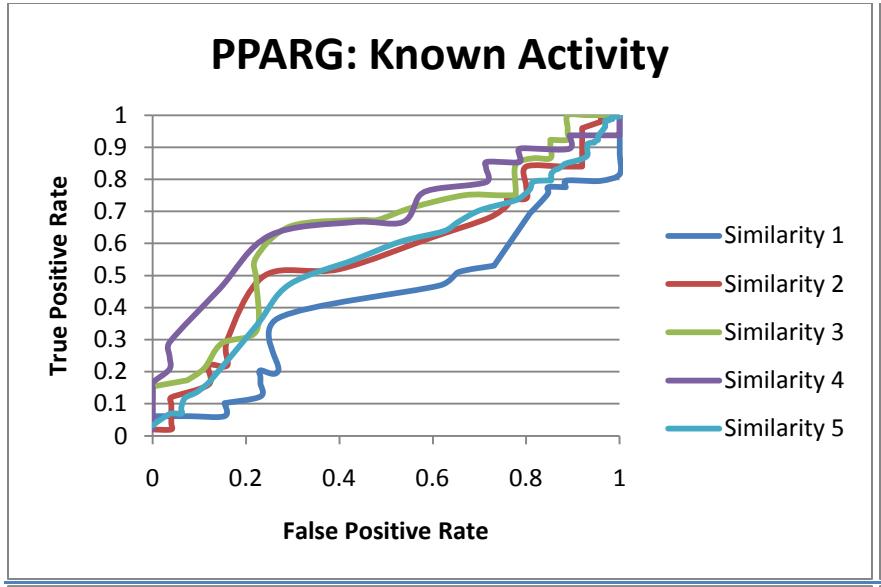

HIV-INT: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

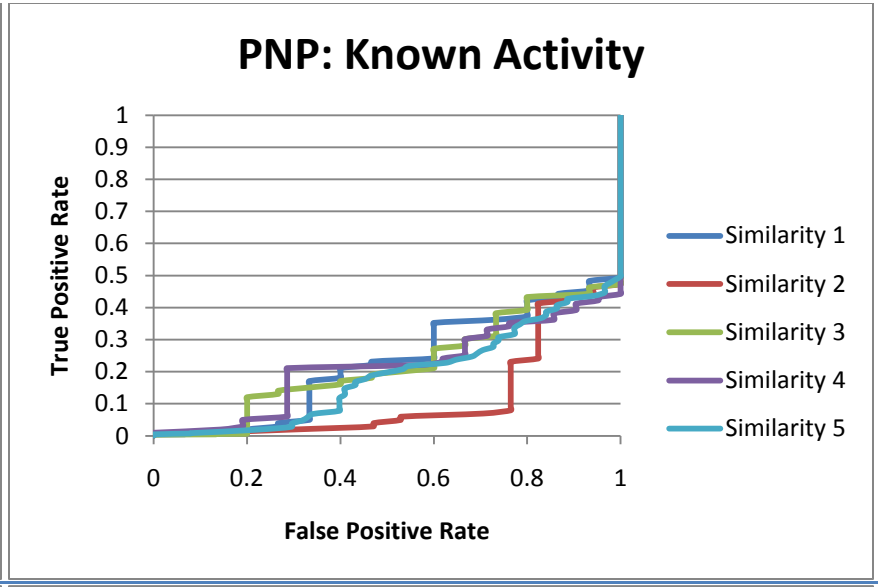

HIV-PR: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

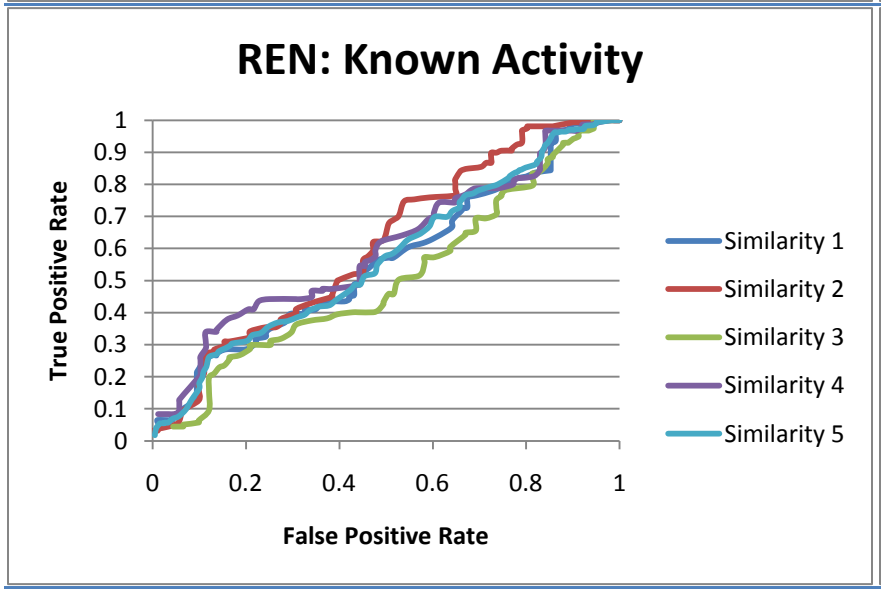

HIV-RT: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

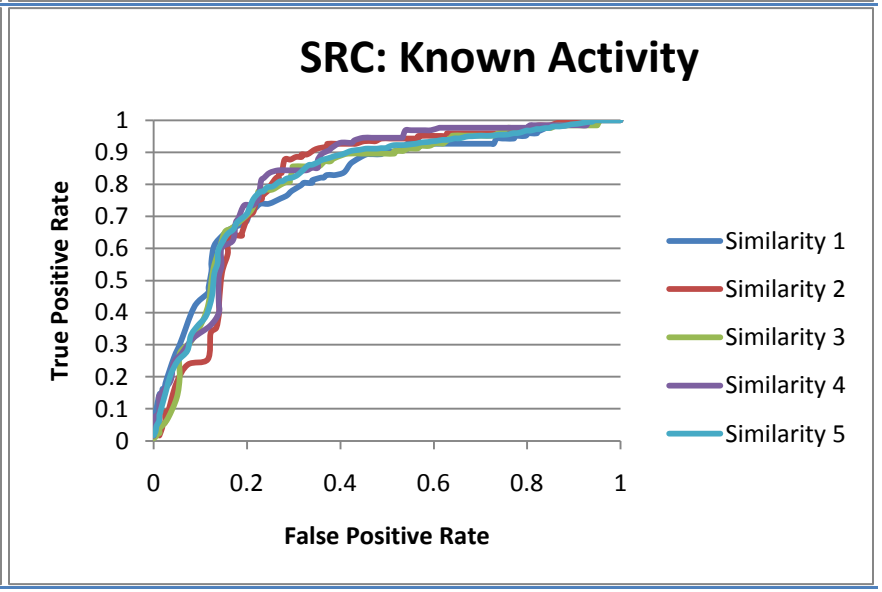

PARP1: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5


PDE5: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

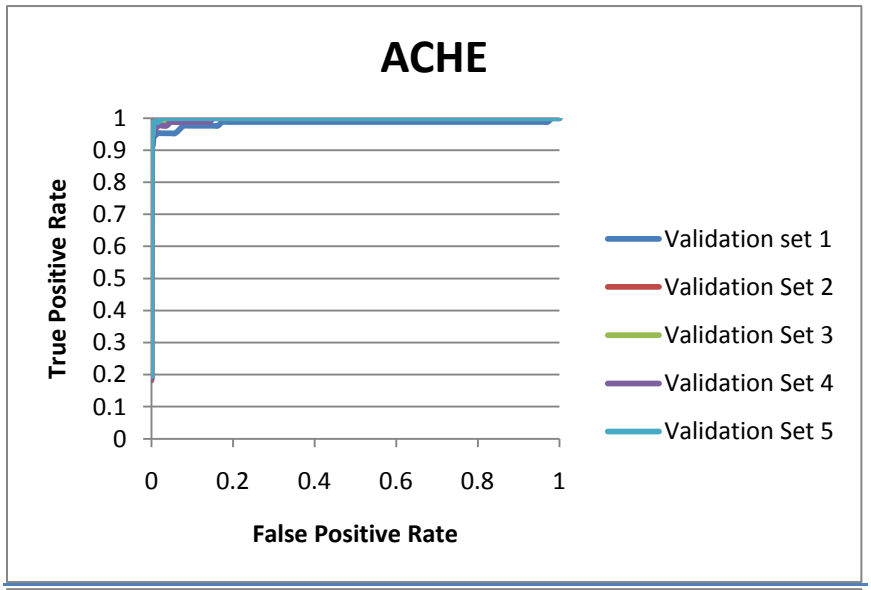

PPARG: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

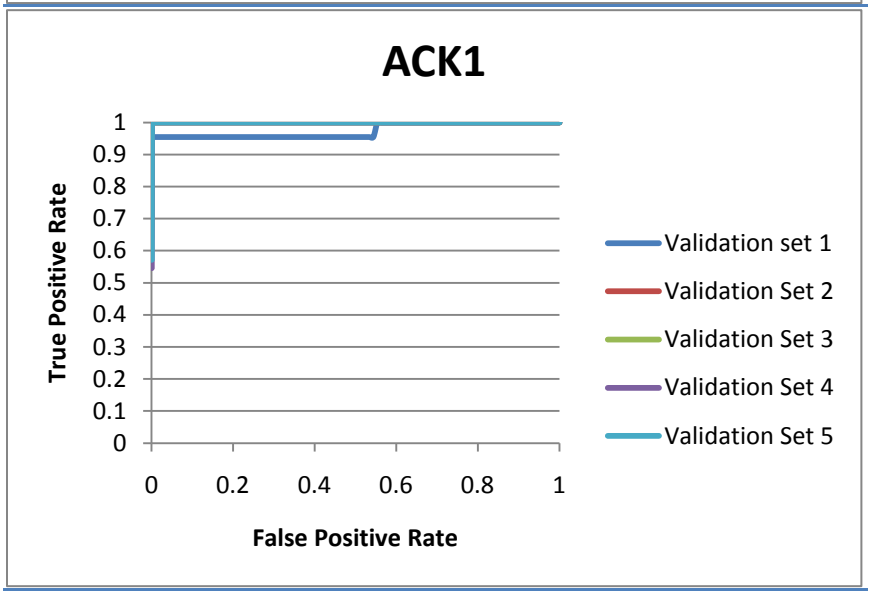

PNP: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

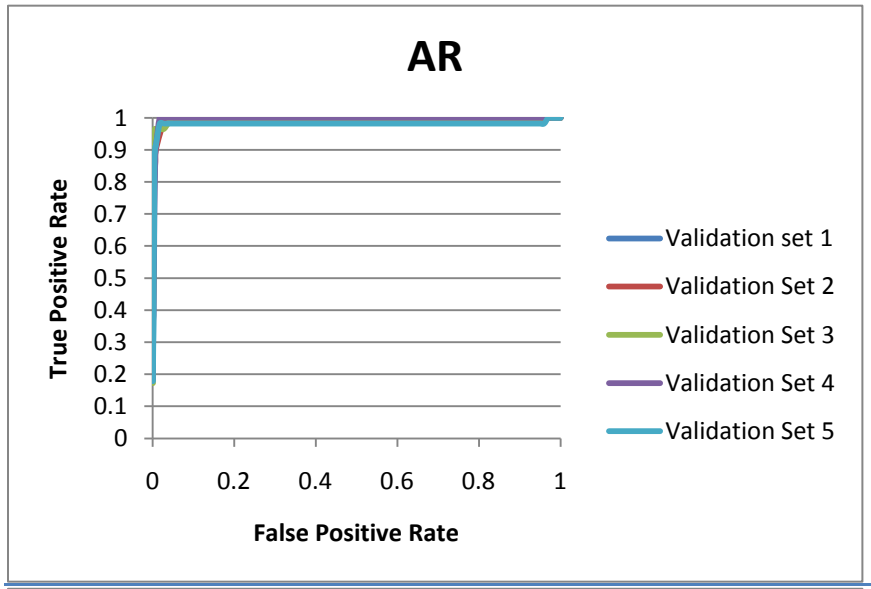

REN: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

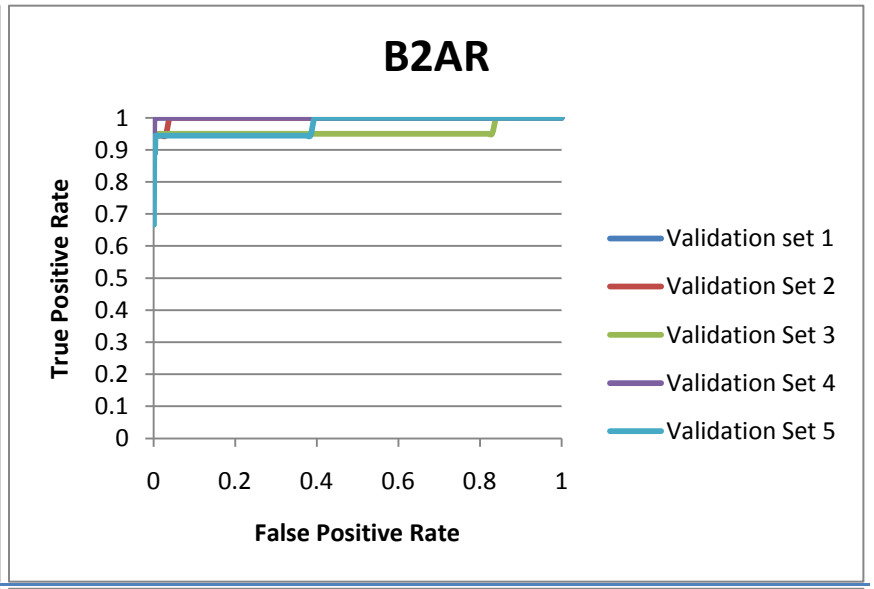

SRC: Known Activity
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Similarity 1
Similarity 2
Similarity 3
Similarity 4
Similarity 5

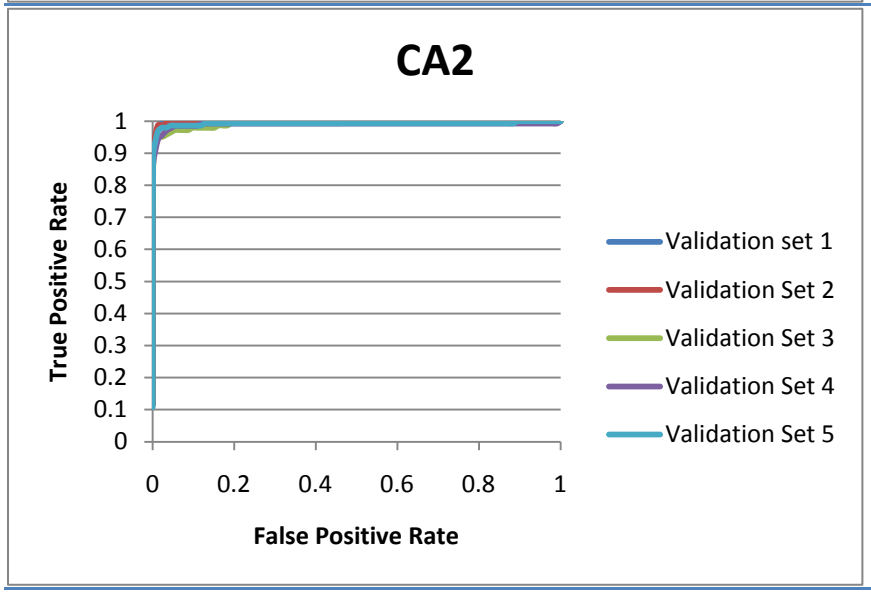
Full Screening Set – Modeling Set Probes

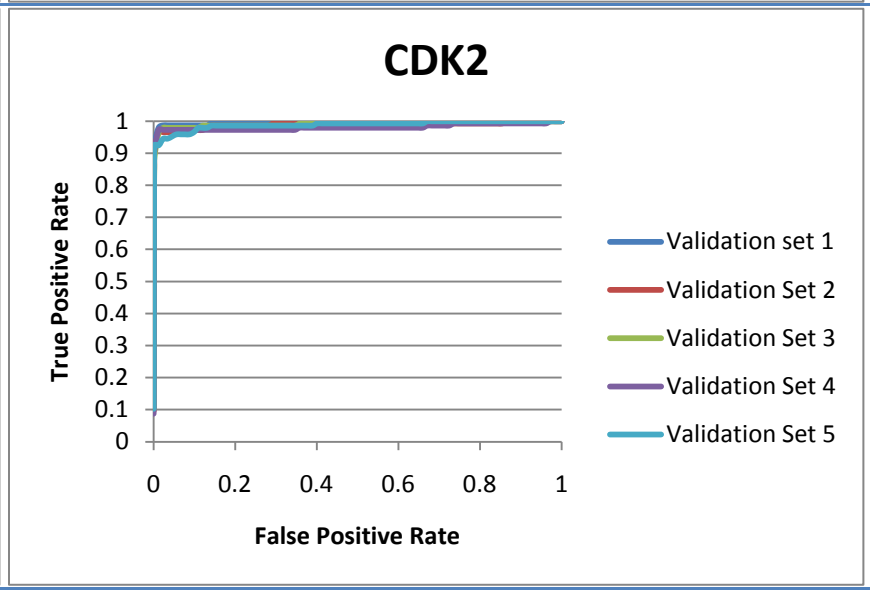


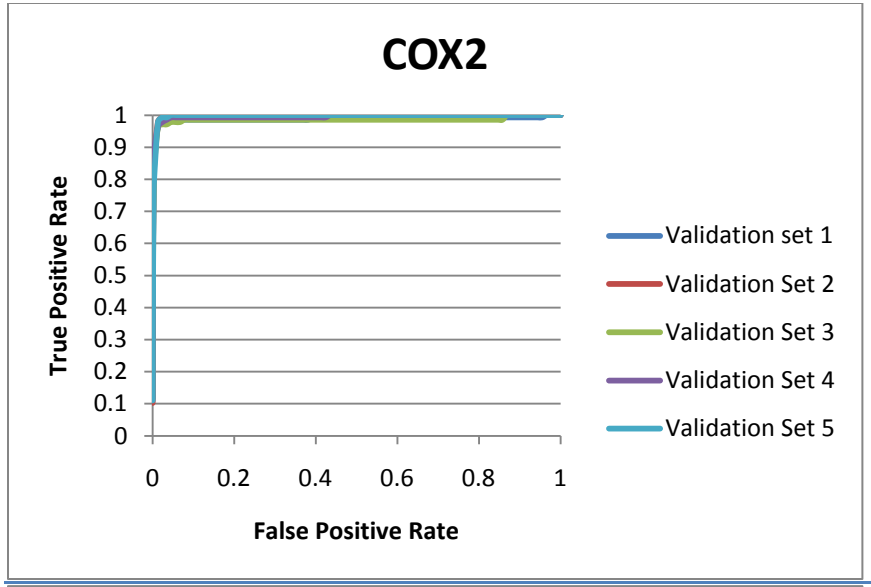


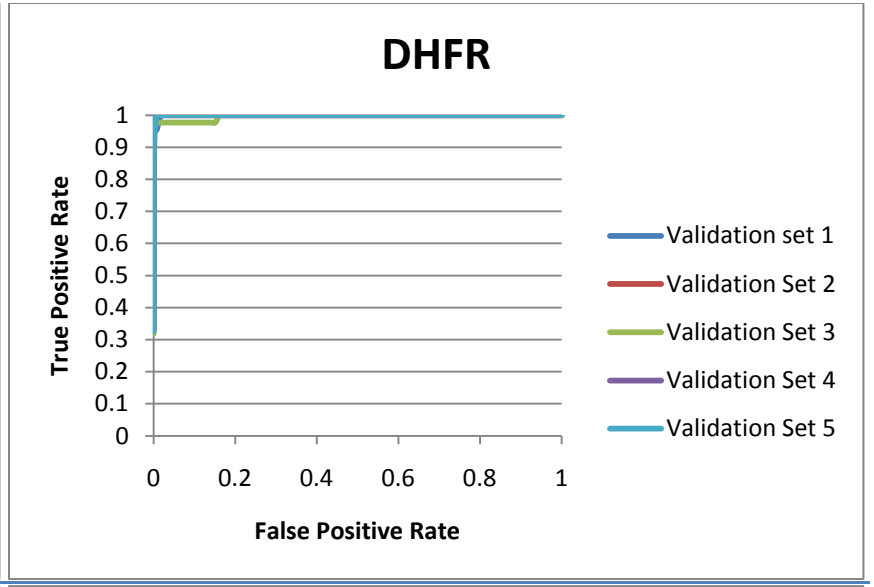

AR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

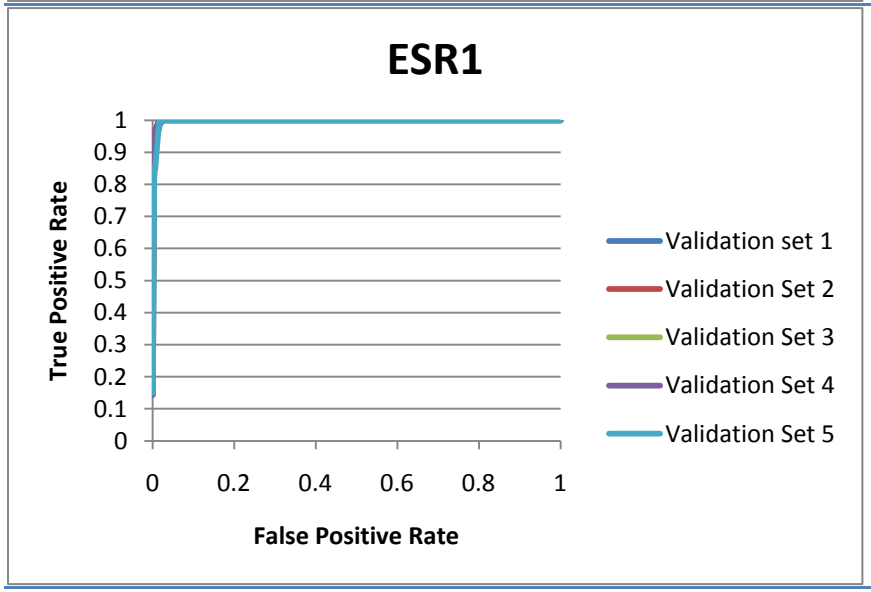

B2AR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

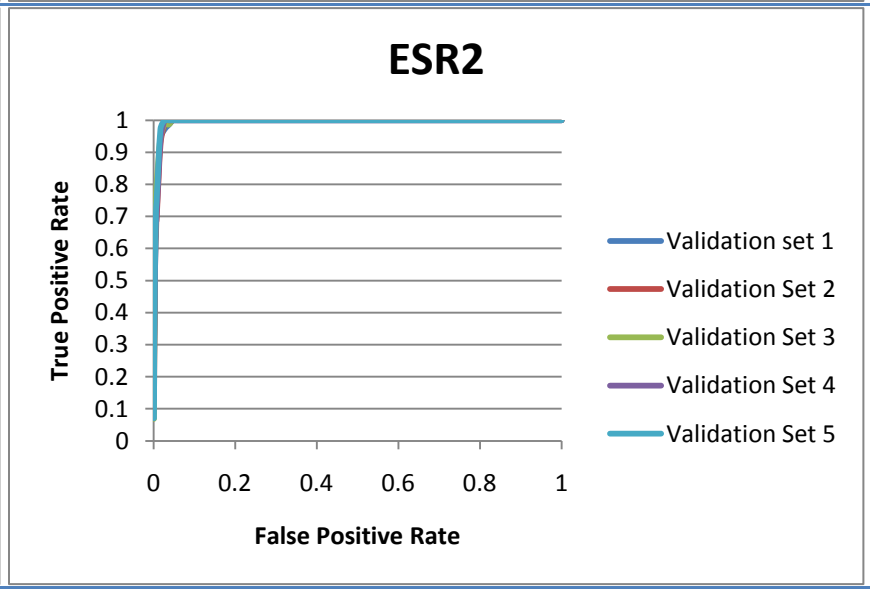

CA2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

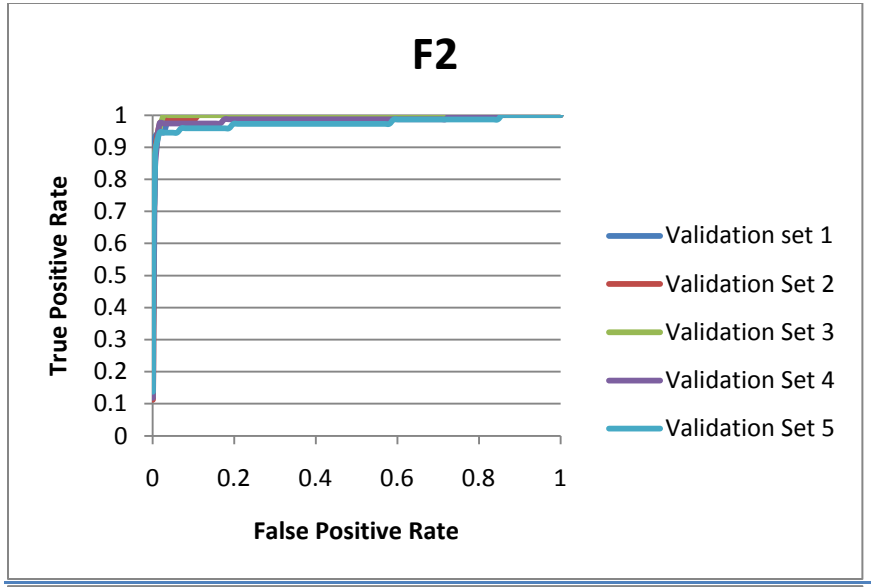

CDK2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

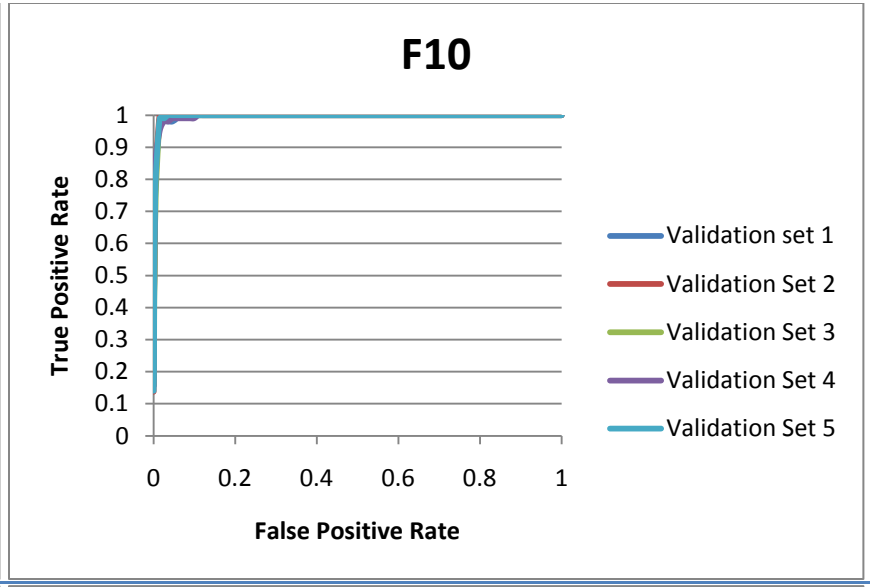

COX2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

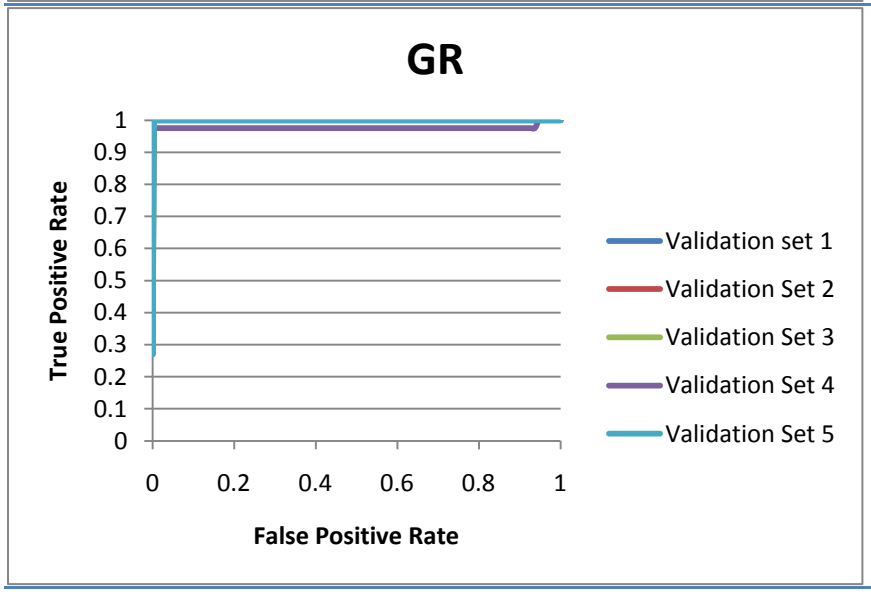

DHFR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

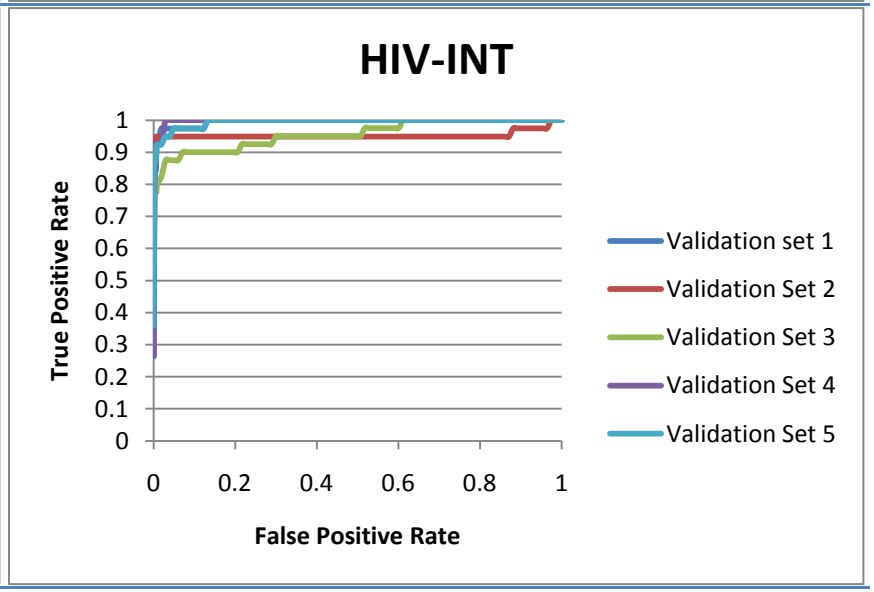

ESR1
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

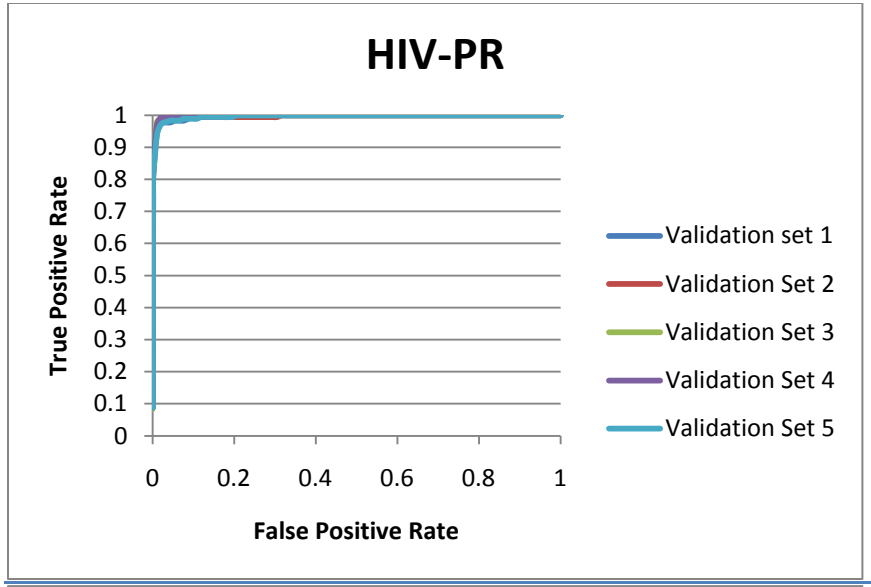

ESR2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

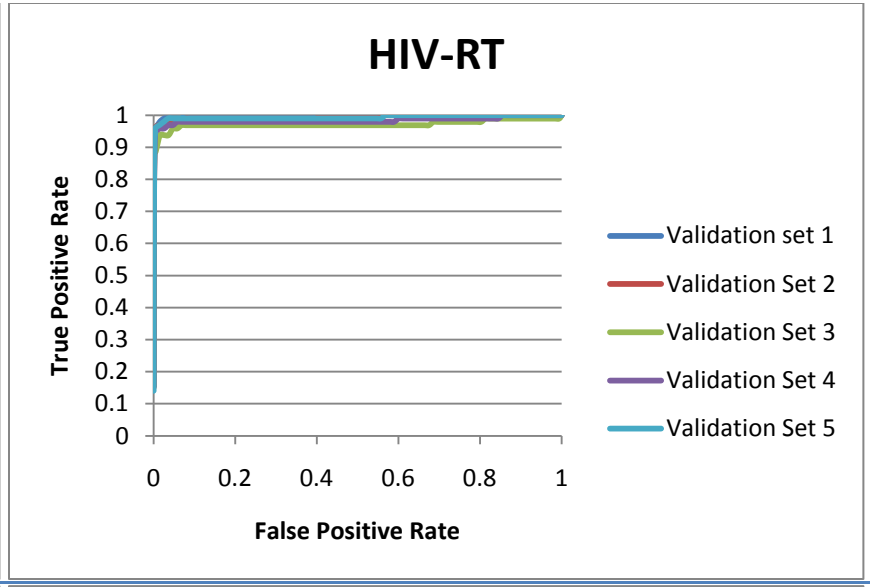

F2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

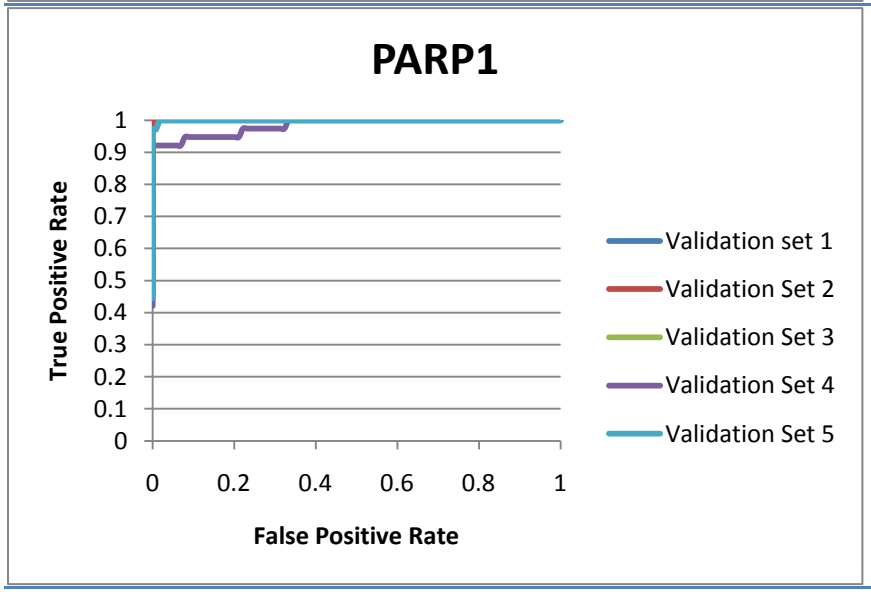

F10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

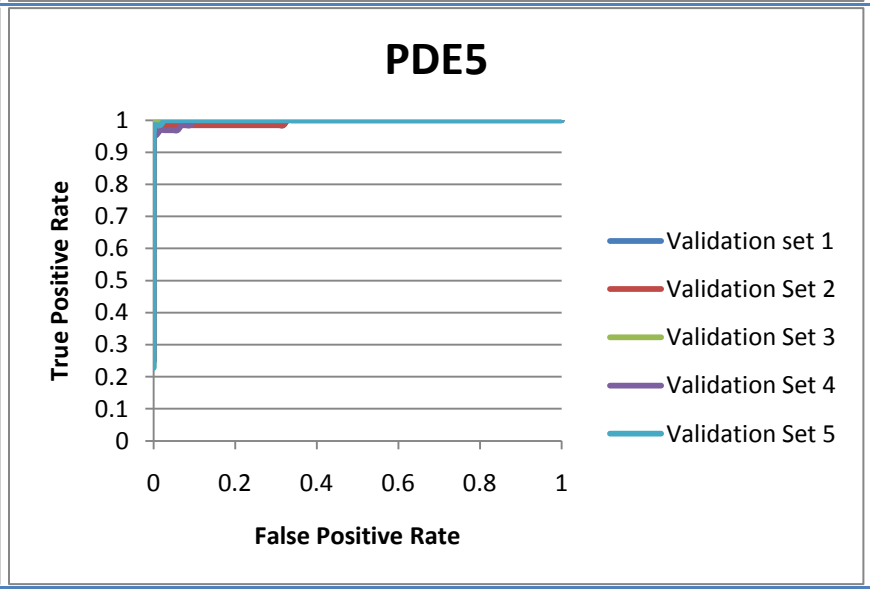

GR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

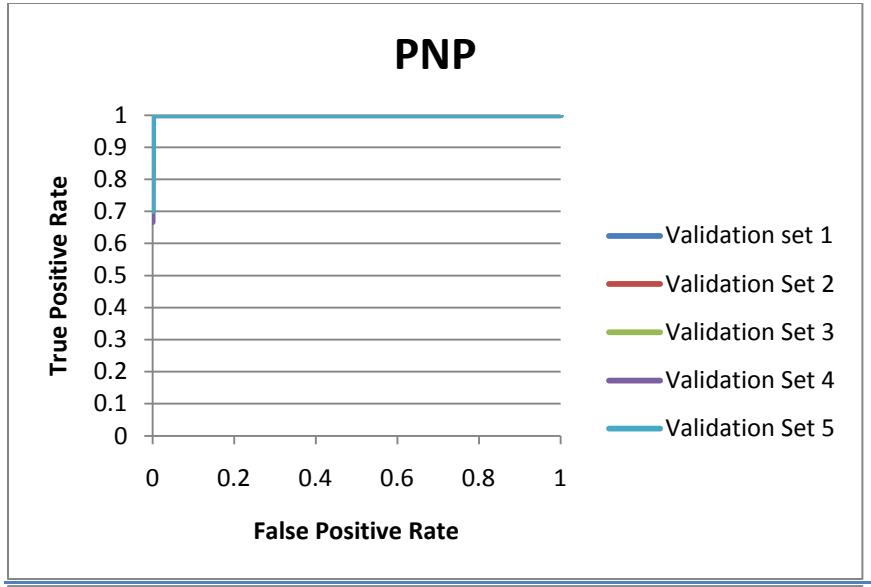

HIV-INT
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

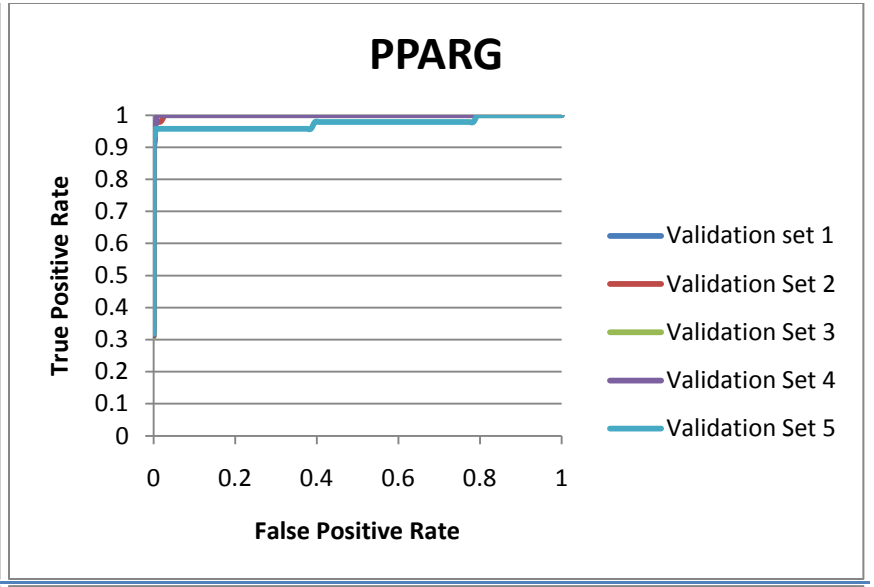

HIV-PR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

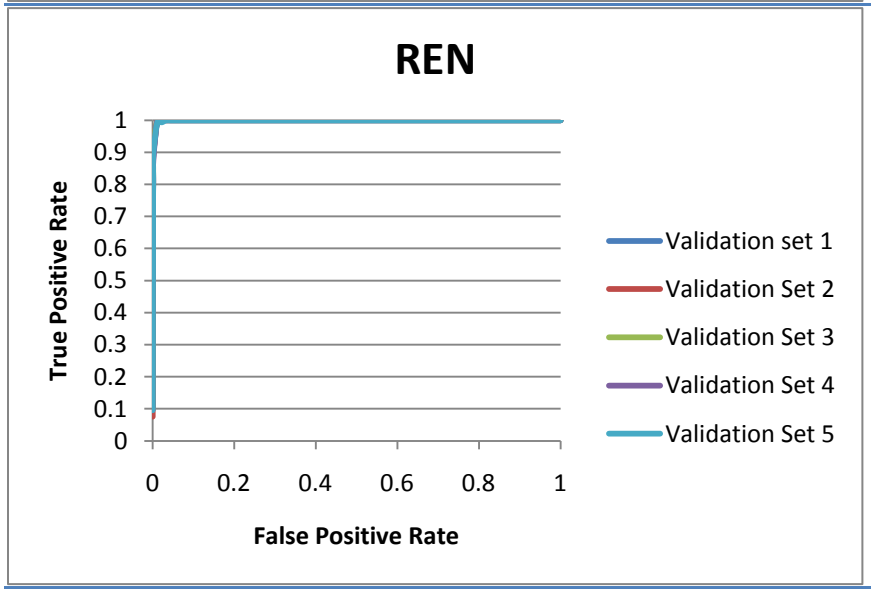

HIV-RT
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

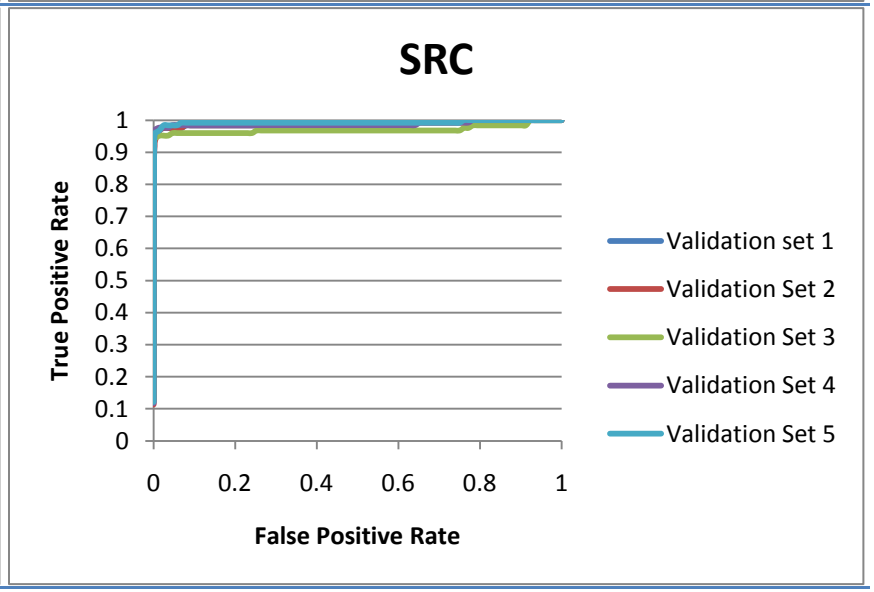

PARP1
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


PDE5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

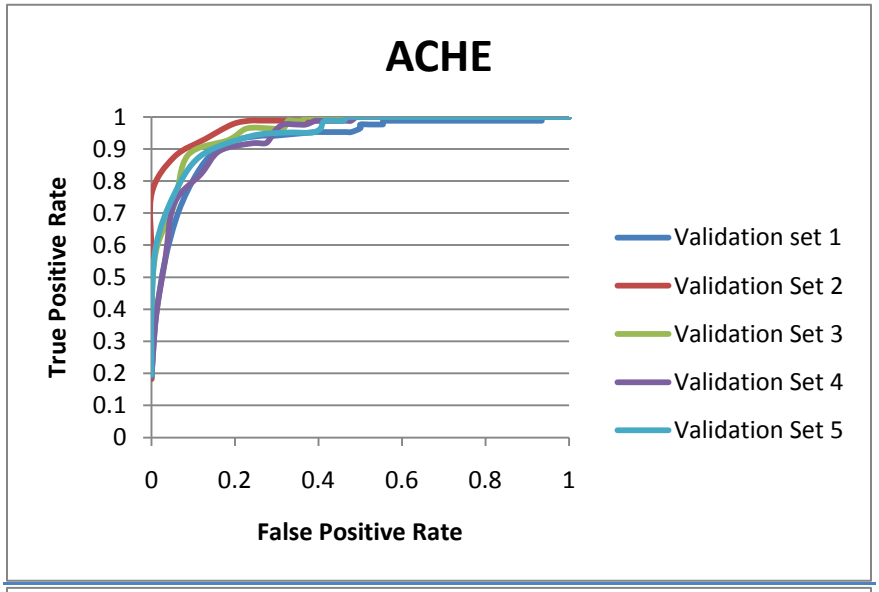

PNP
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

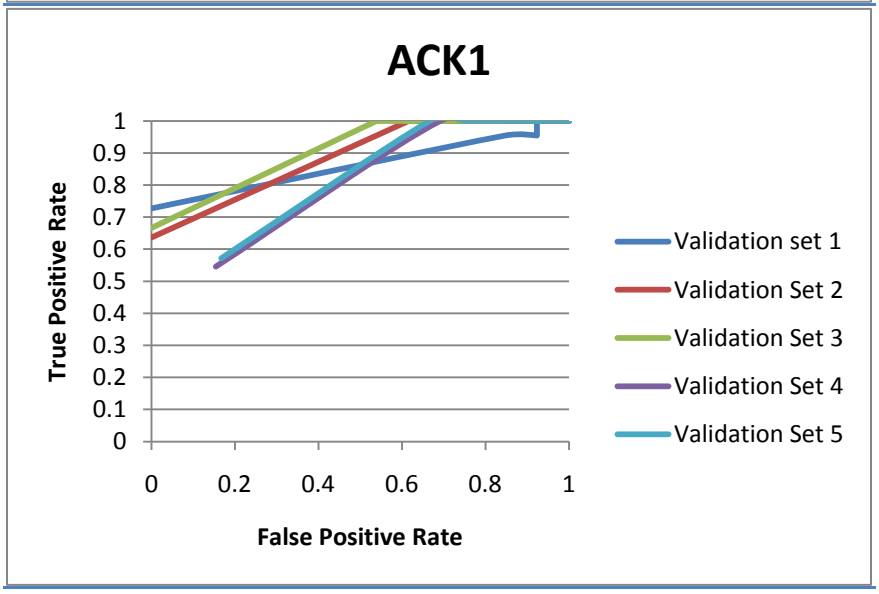

PPARG
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

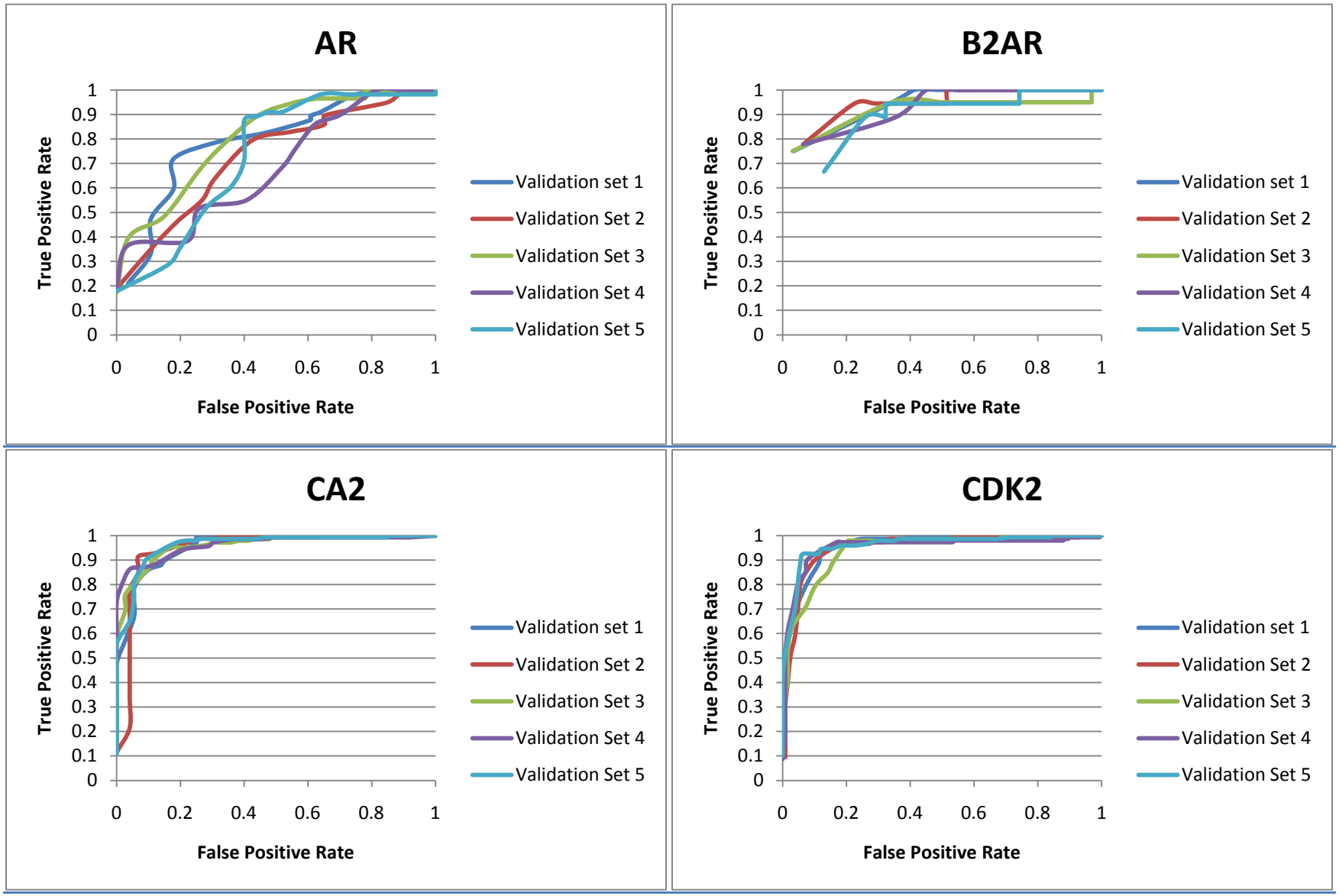

REN
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

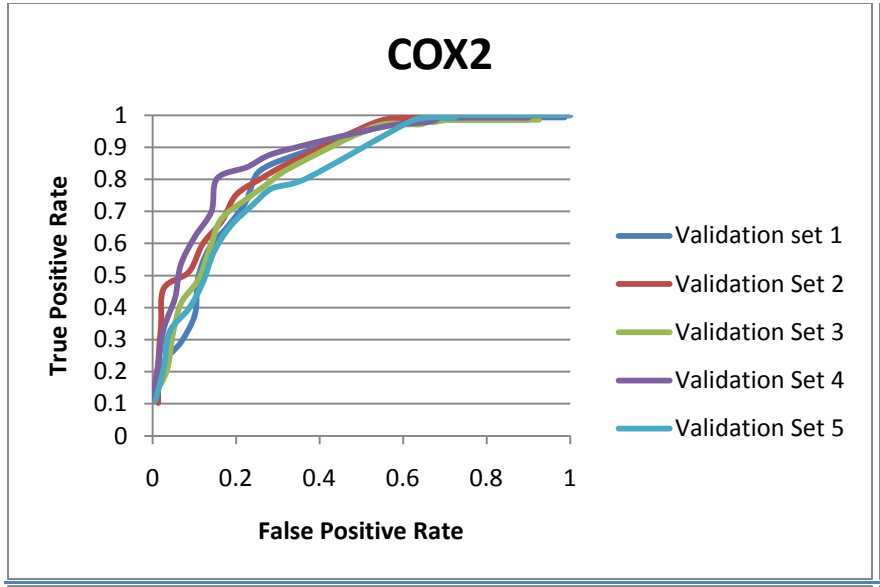

SRC
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

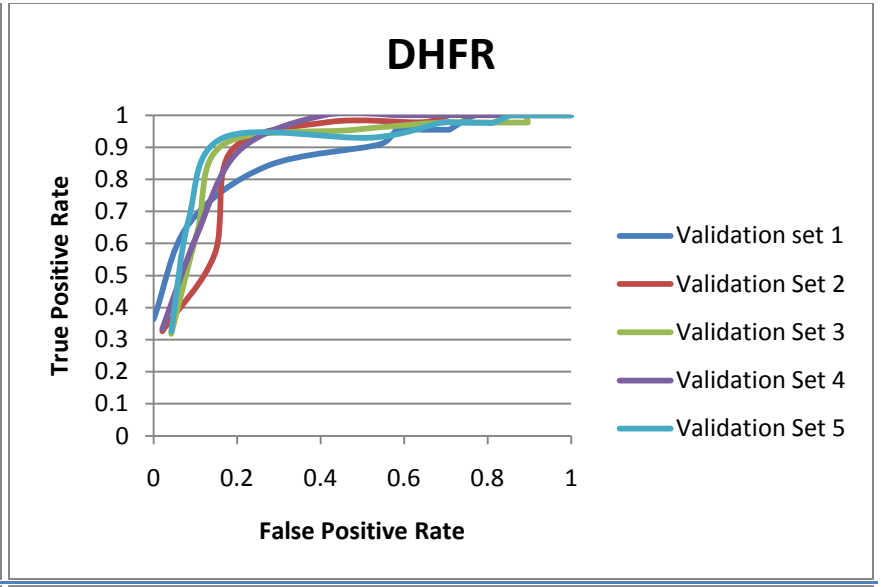
Compounds with Known Activities – Modeling Set Probes

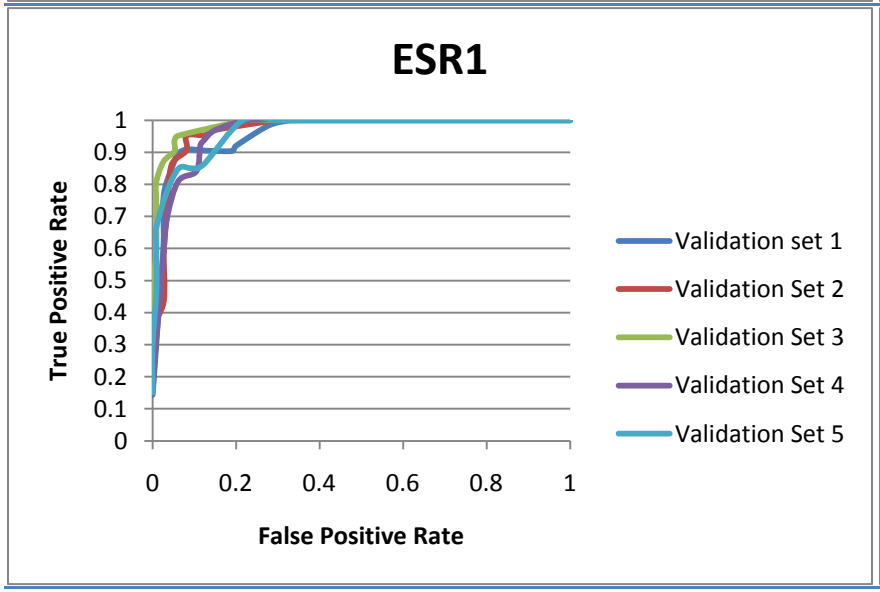


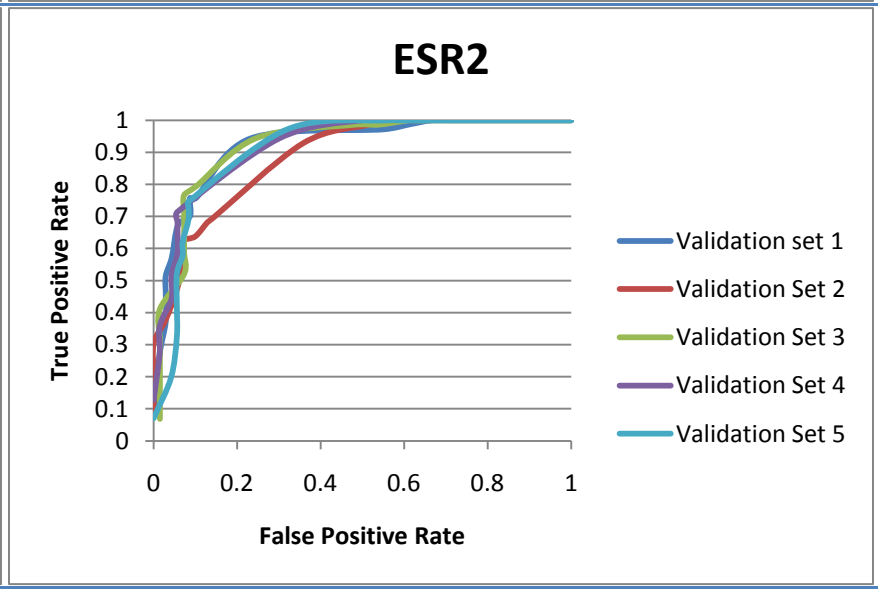


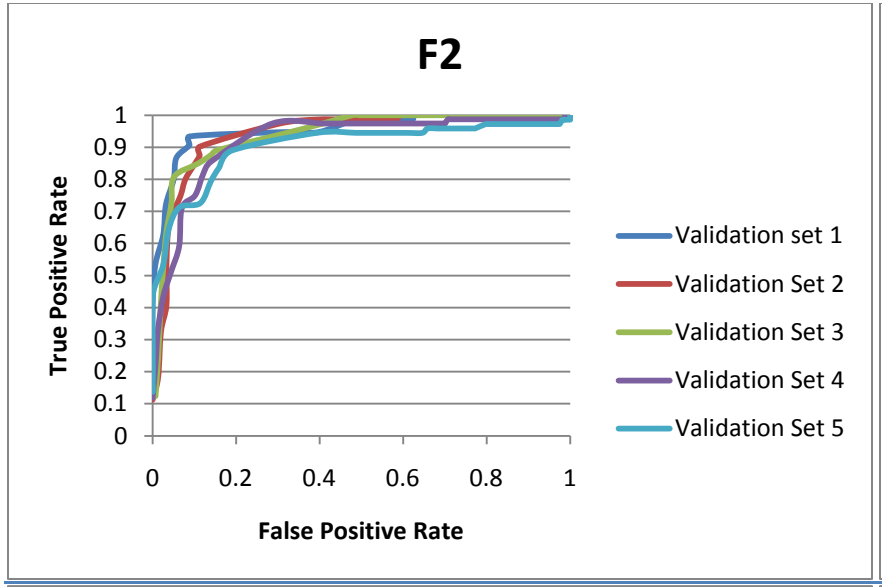

AR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
B2AR
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
CA2
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
CDK2
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

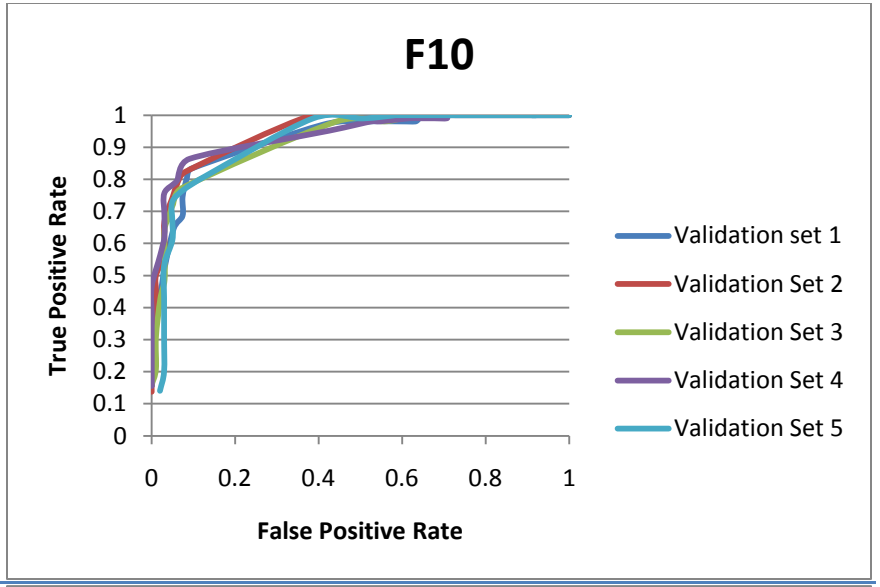

COX2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

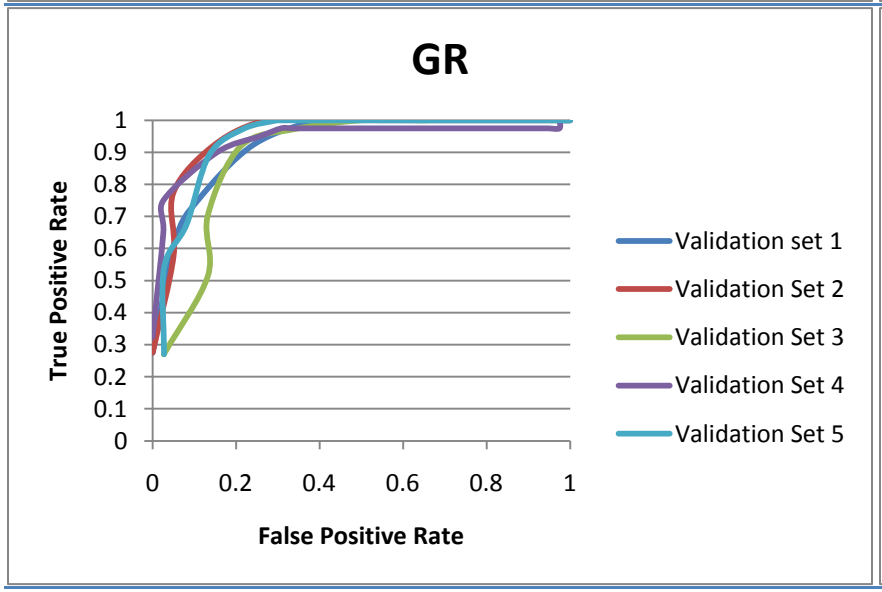

DHFR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

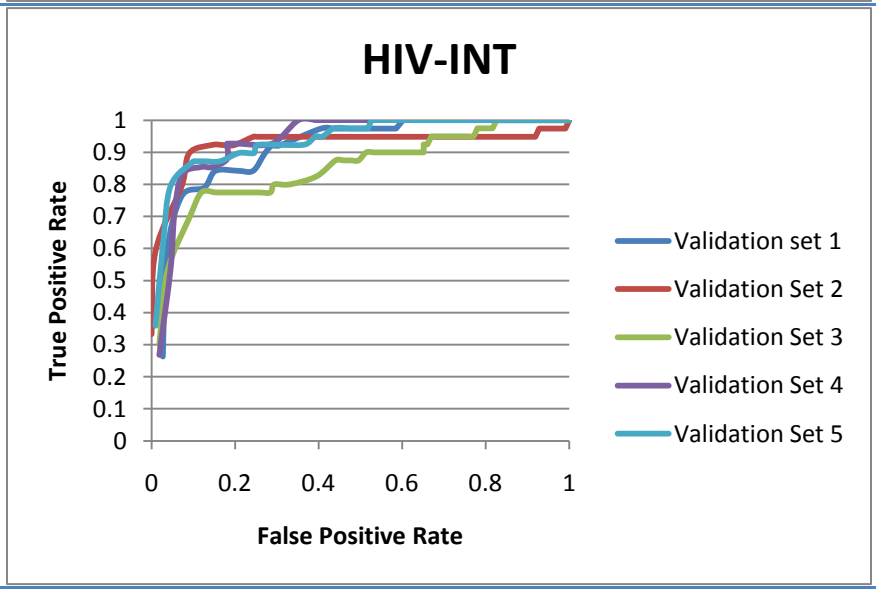

ESR1
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

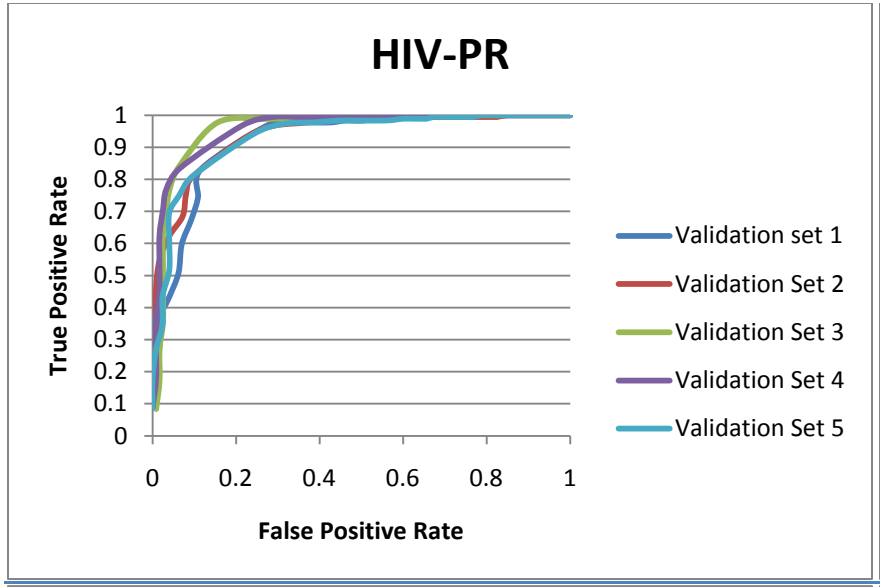

ESR2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

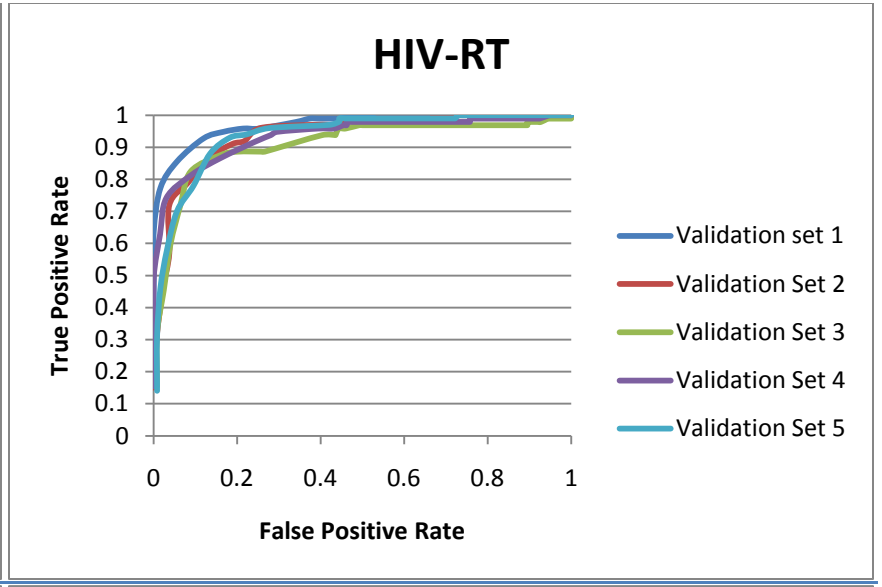

F2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

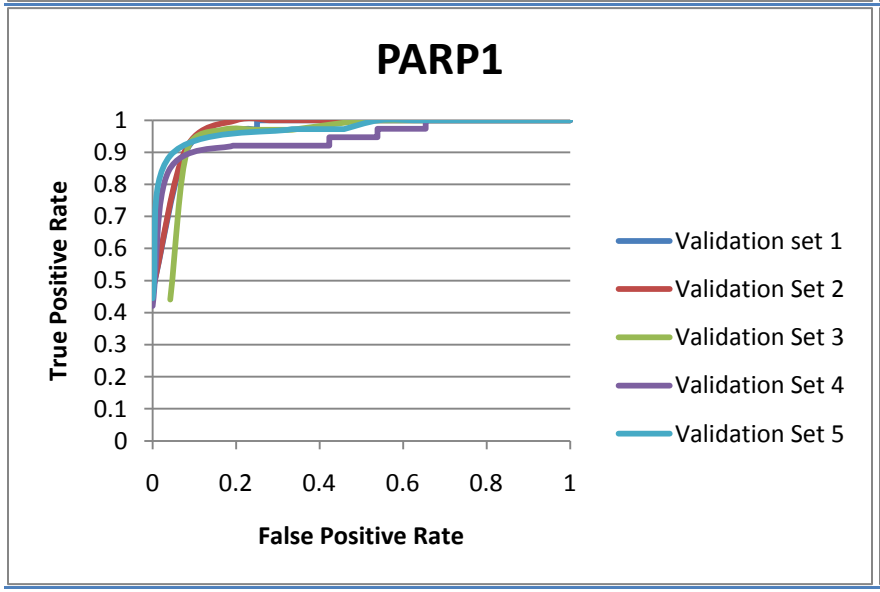

F10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

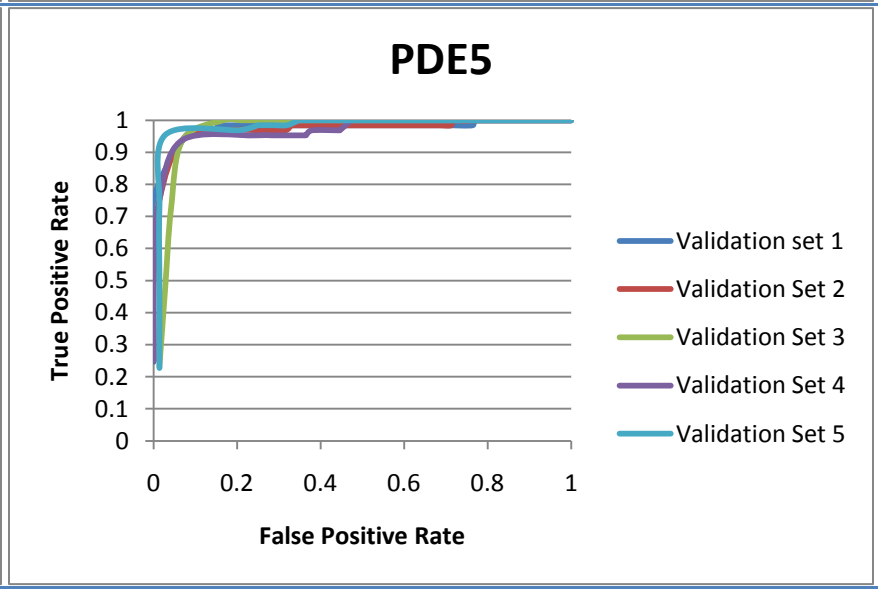

GR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

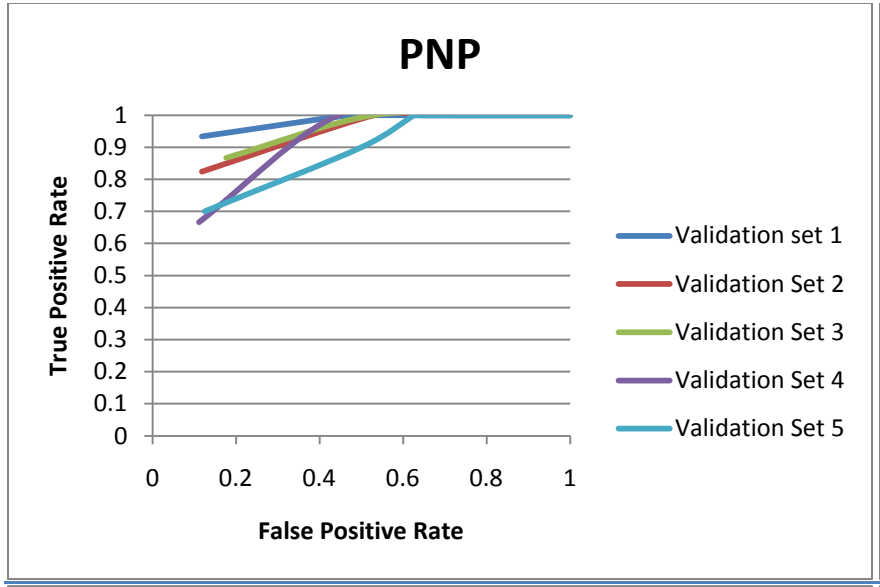

HIV-INT
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

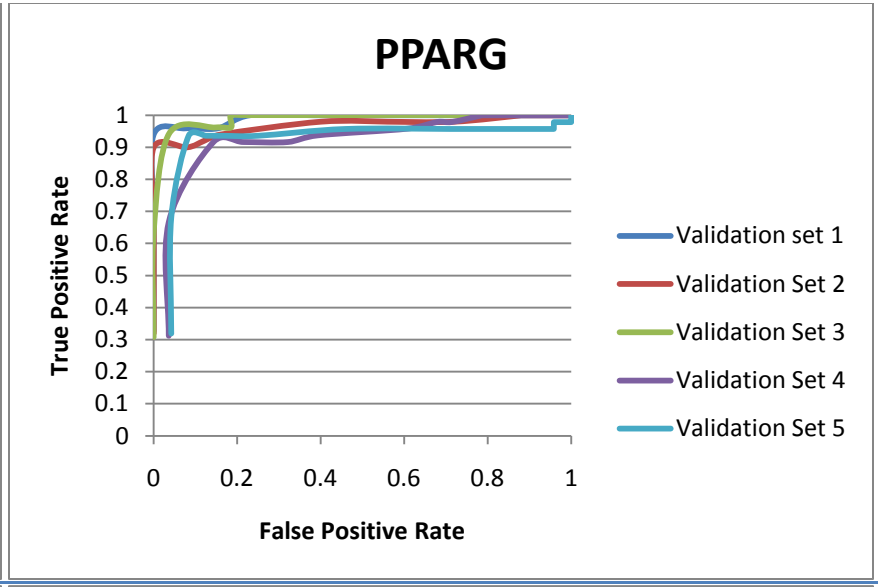

HIV-PR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

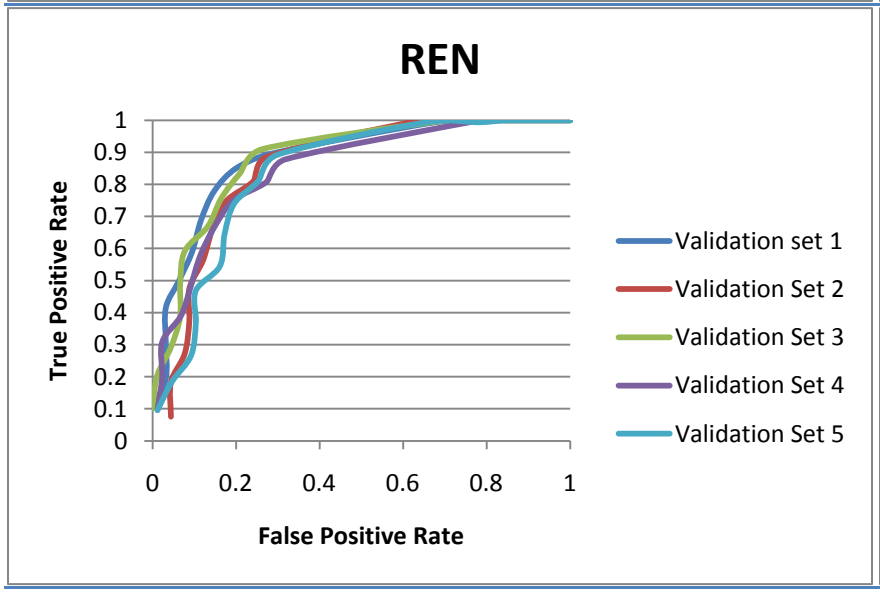

HIV-RT
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

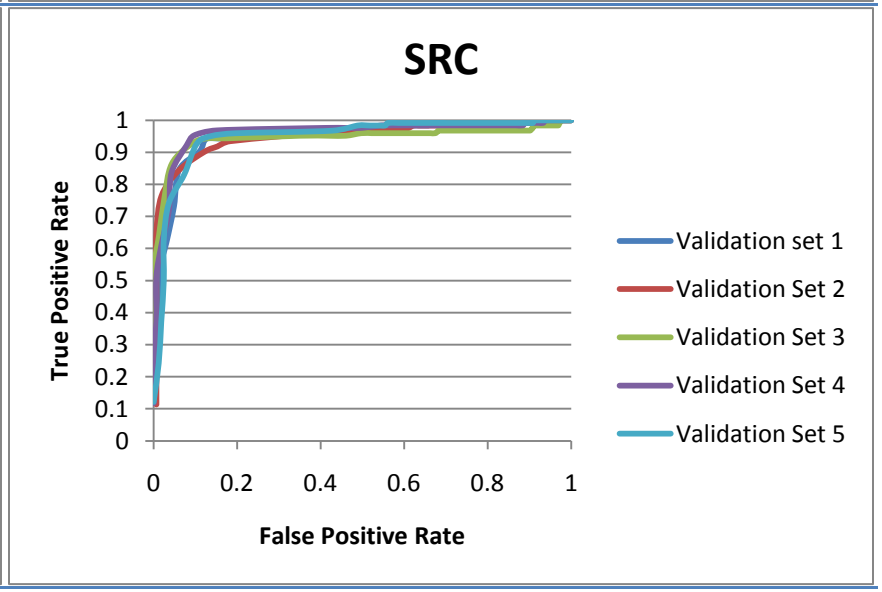

PARP1
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


PDE5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


PNP
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5


PPARG
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

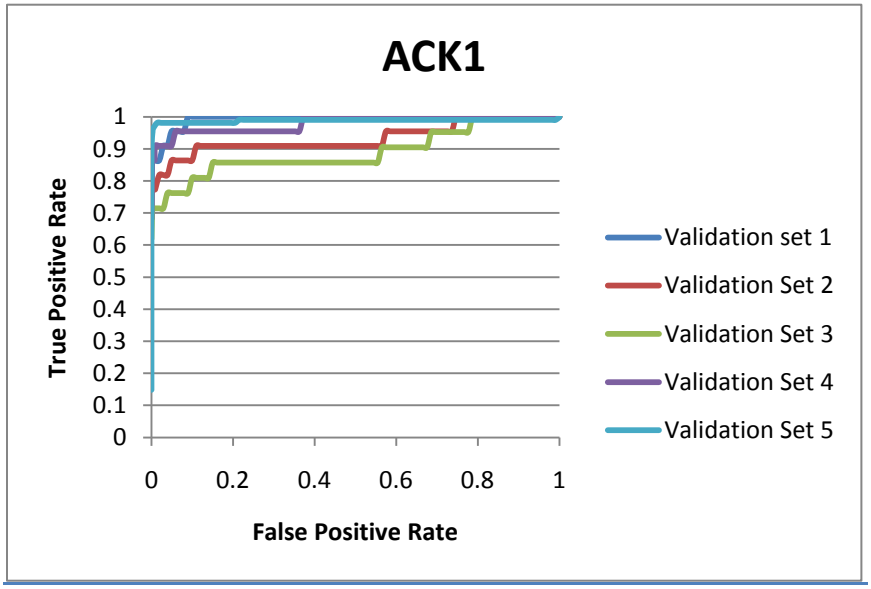

REN
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

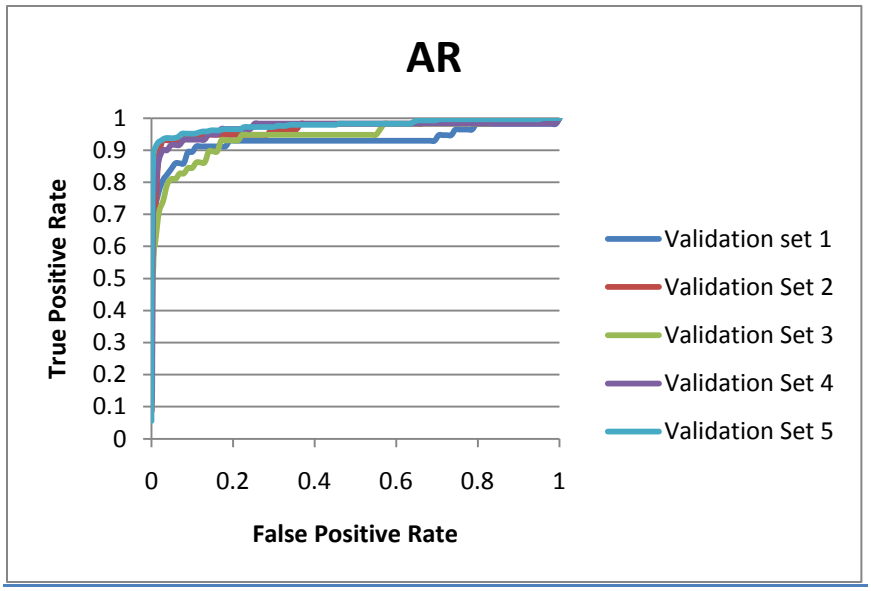

SRC
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

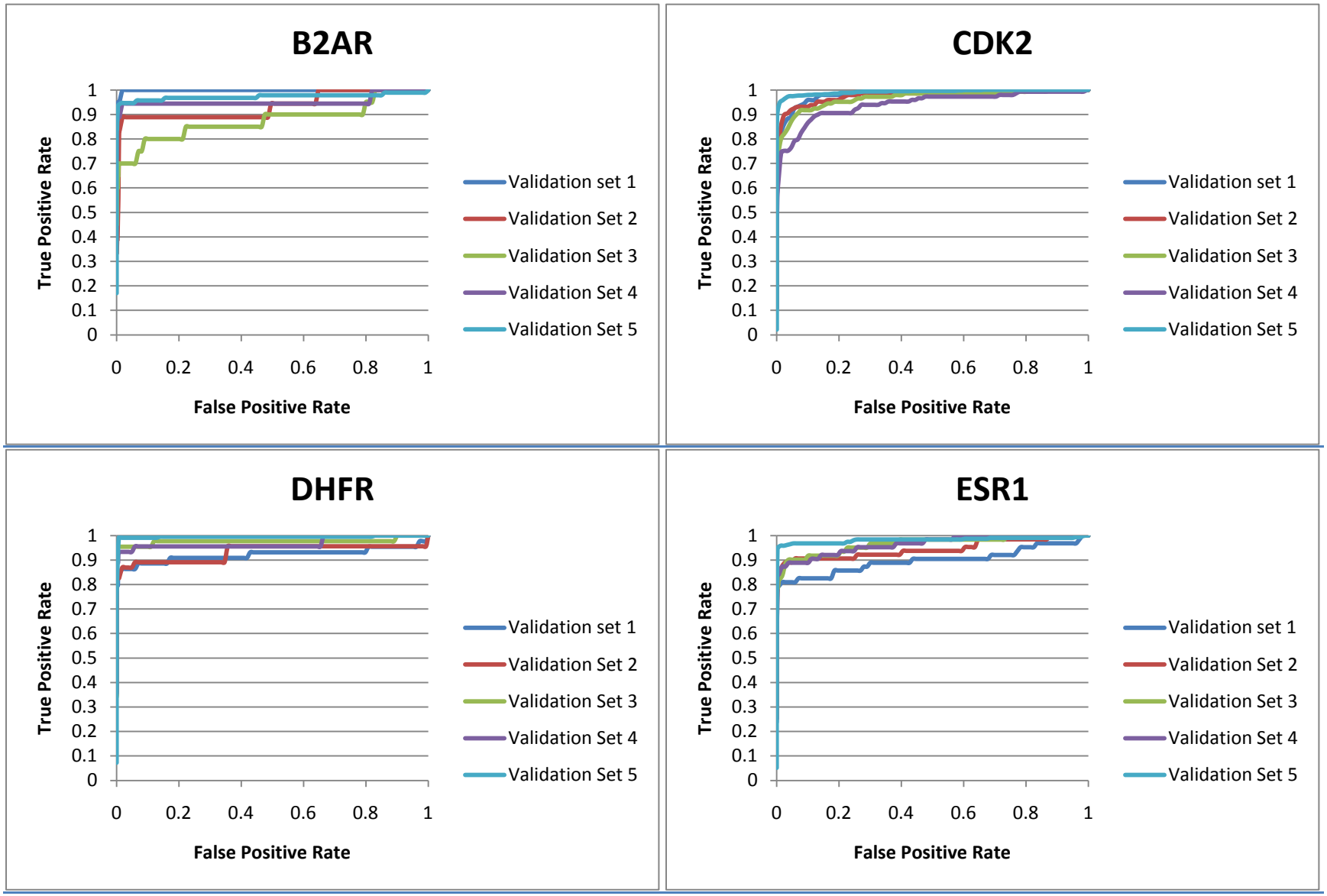
QSAR

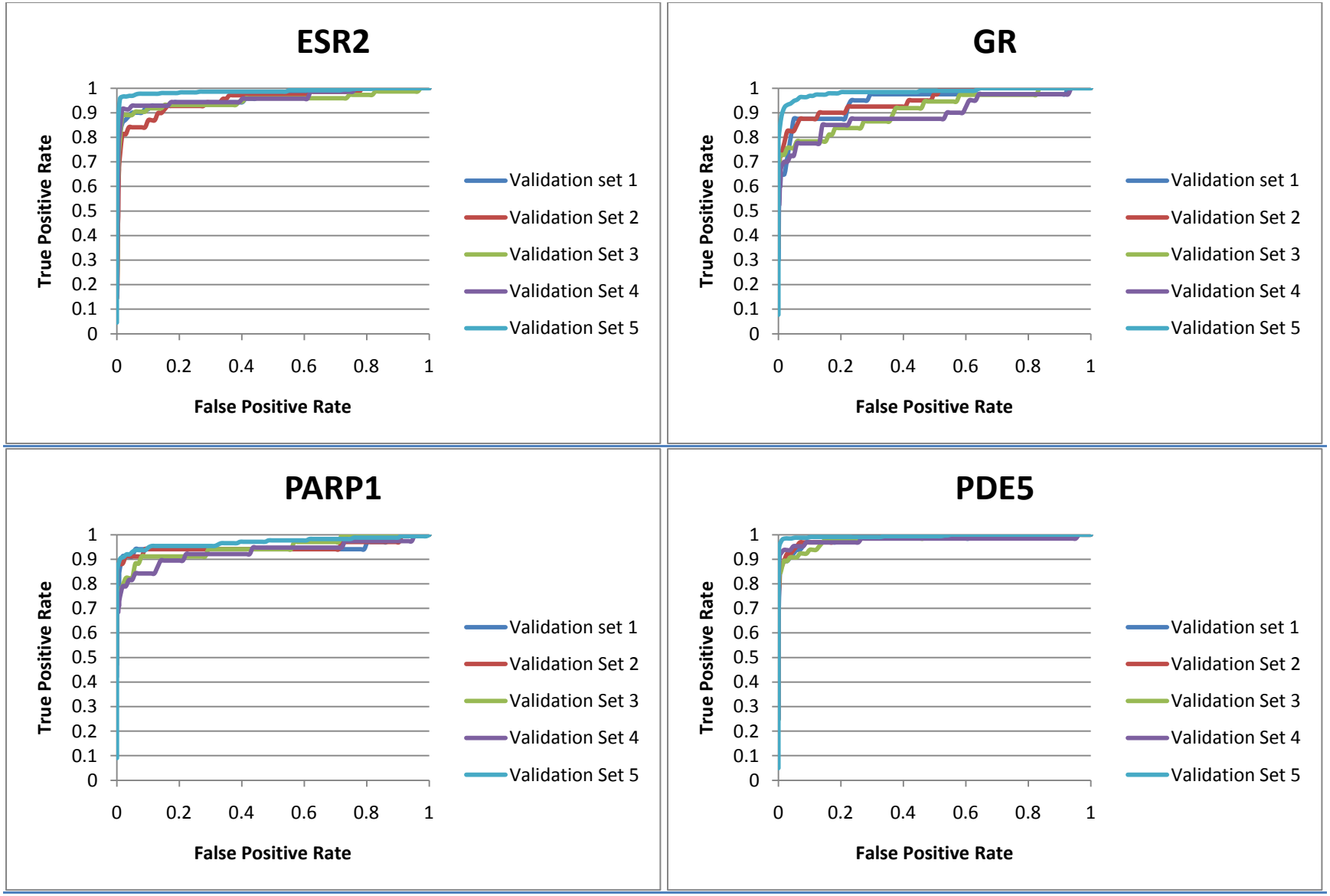
QSAR modeling yielded excellent results when ranking the entire screening library; however, these results were often slightly less exciting than those obtained with similarity searching. On the other hand, the results of ranking compounds with known activities is often as good or better than the ranking of the entire library and usually provides better ranking than similarity searching with the same dataset.

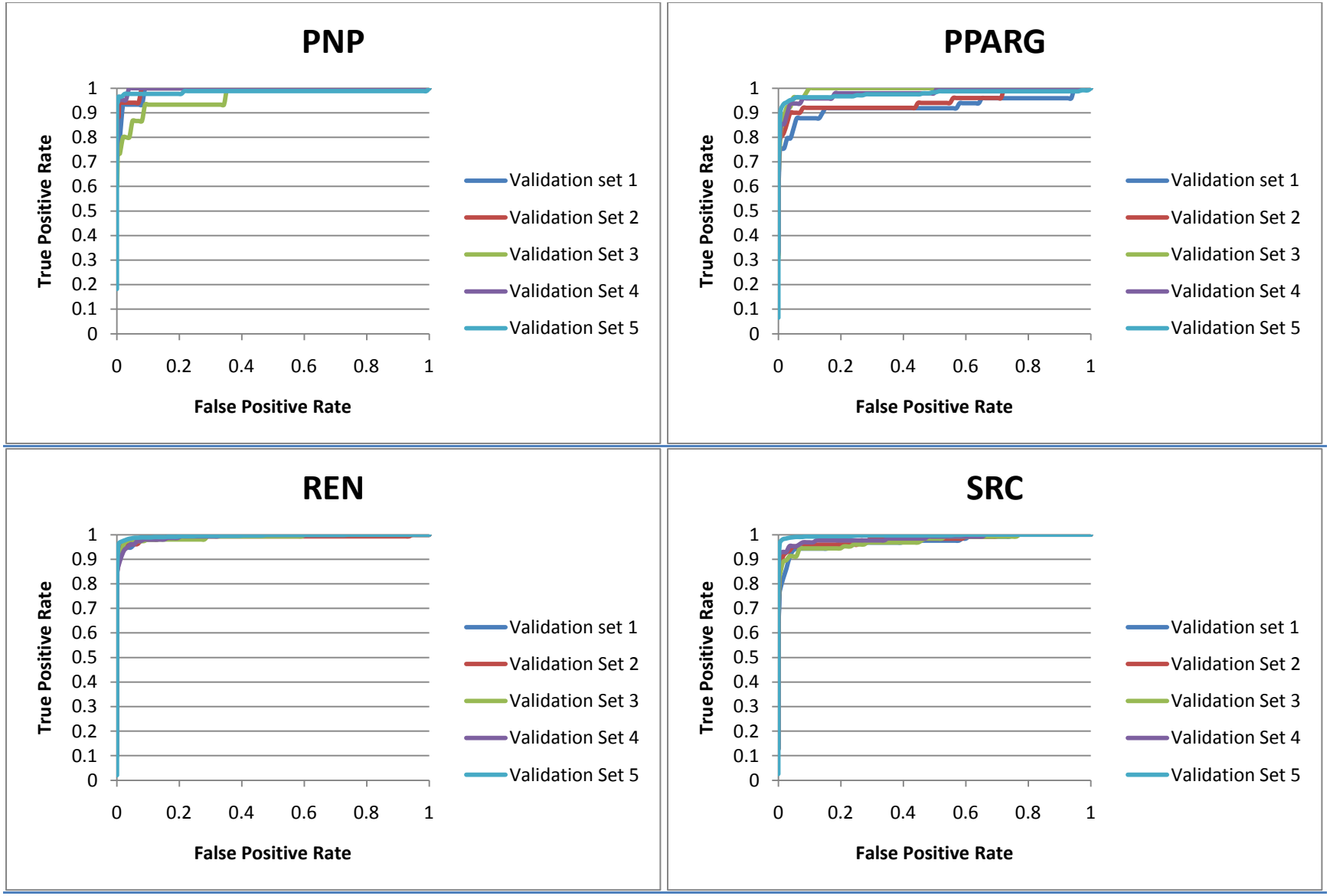
Full Screening Set



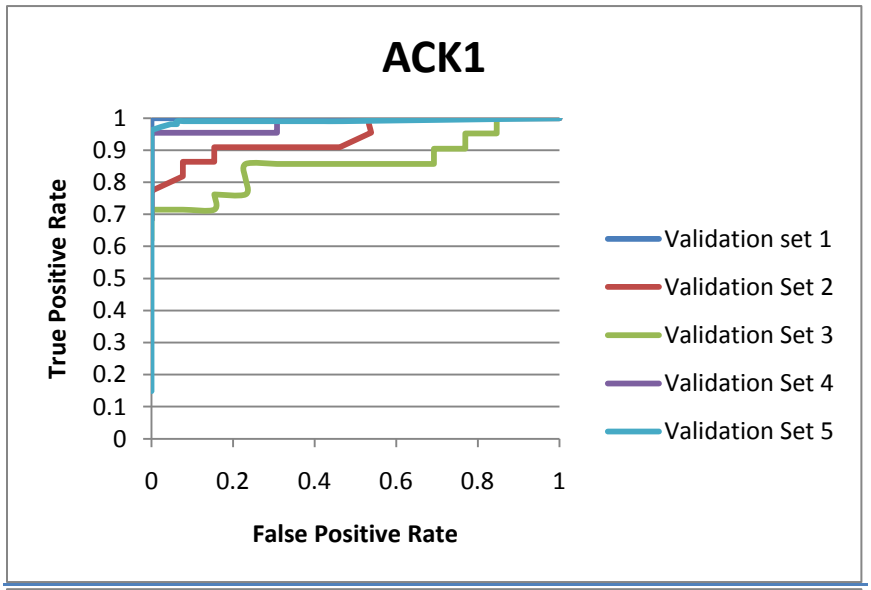


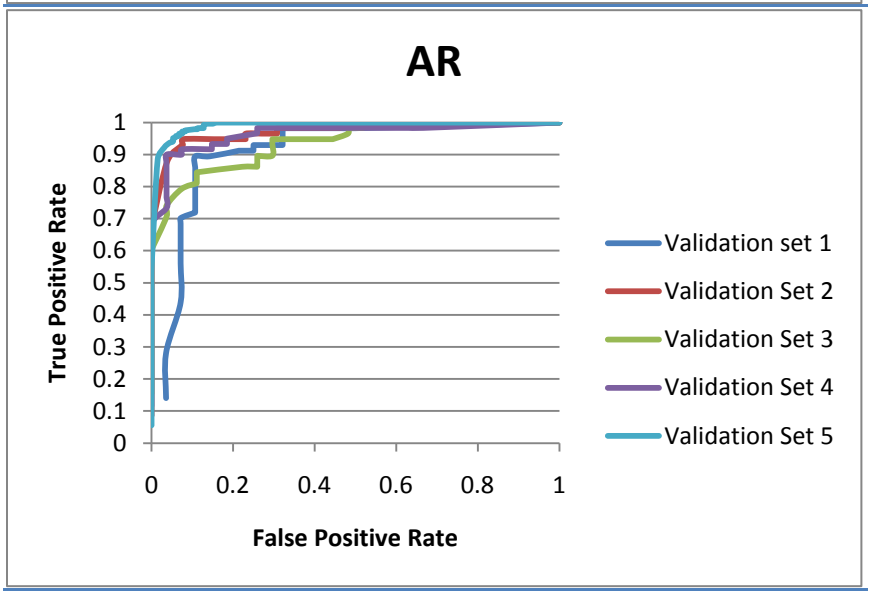

B2AR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
CDK2
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
DHFR
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
ESR1
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

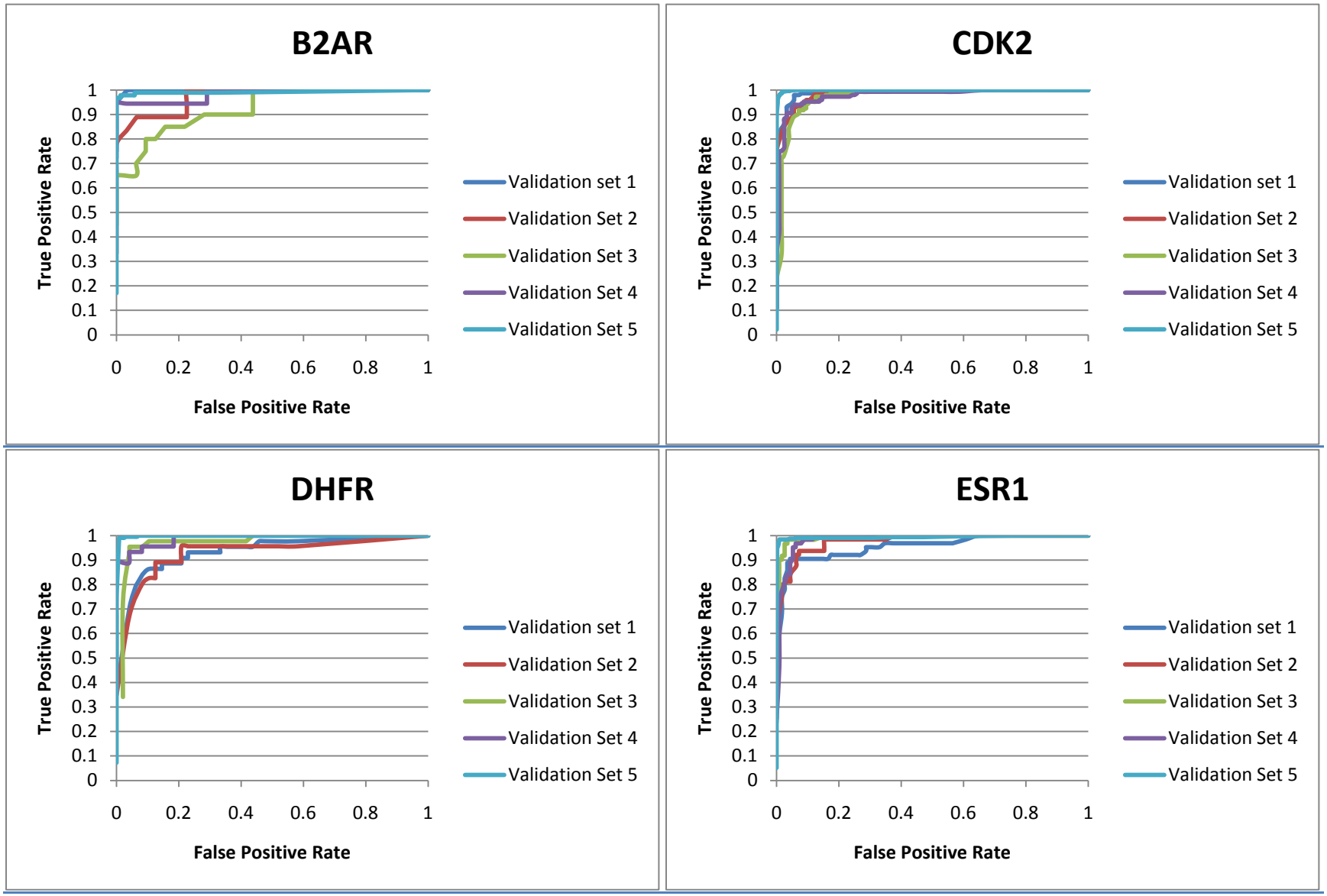

ESR2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
GR
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PARP1
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PDE5
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

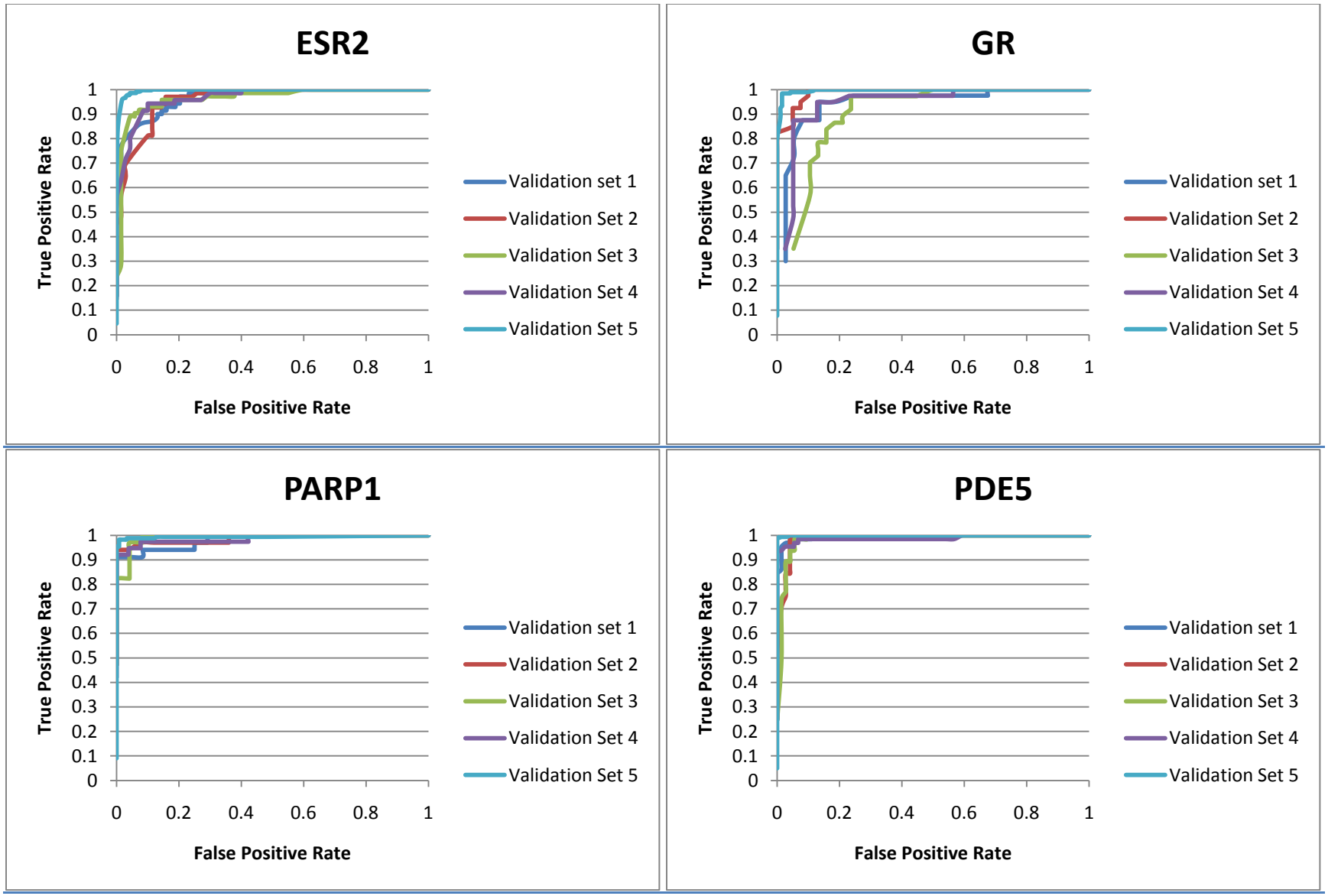

PNP
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PPARG
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
REN
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
SRC
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

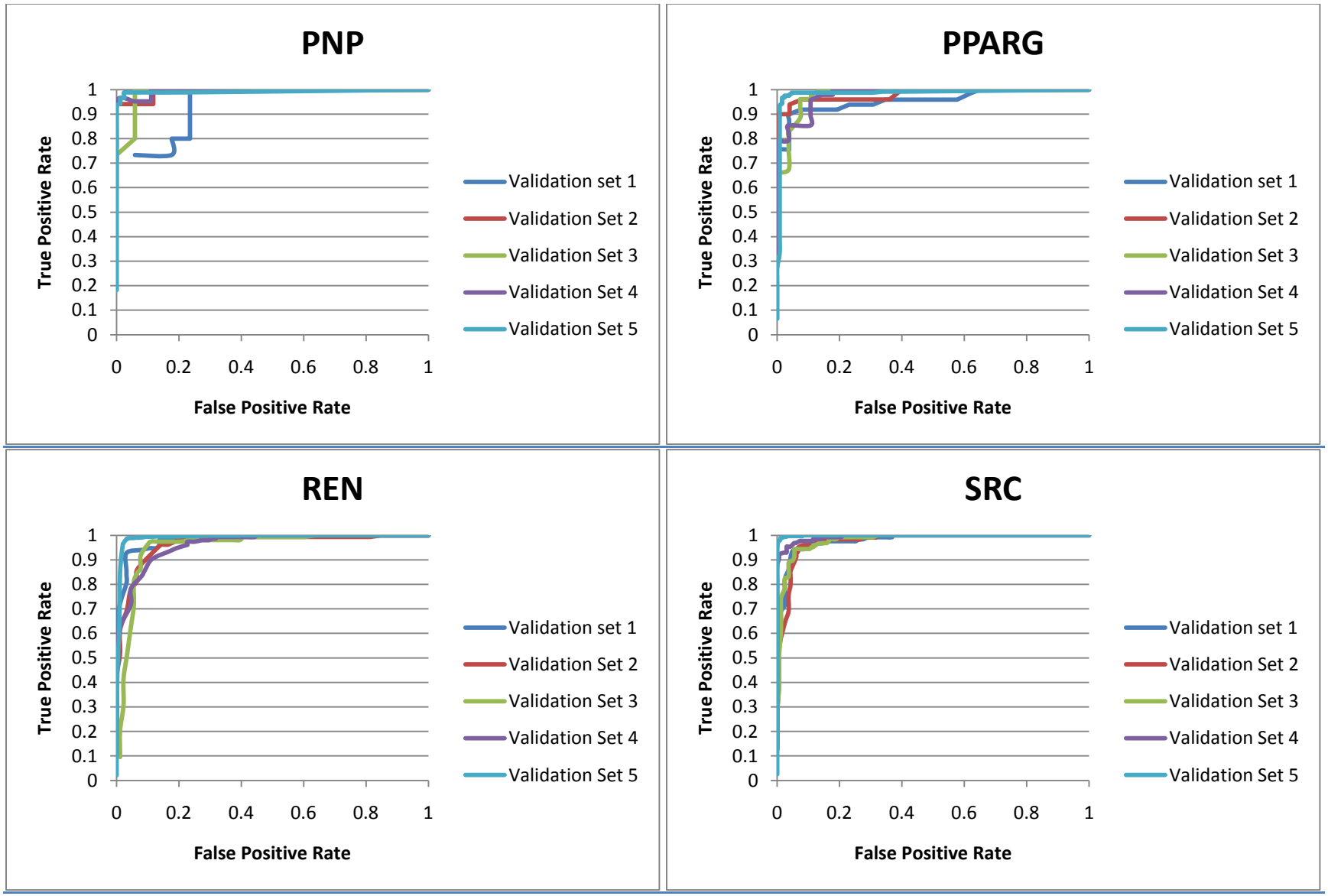
Compounds with Known Activities






B2AR
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
CDK2
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
DHFR
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
ESR1
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

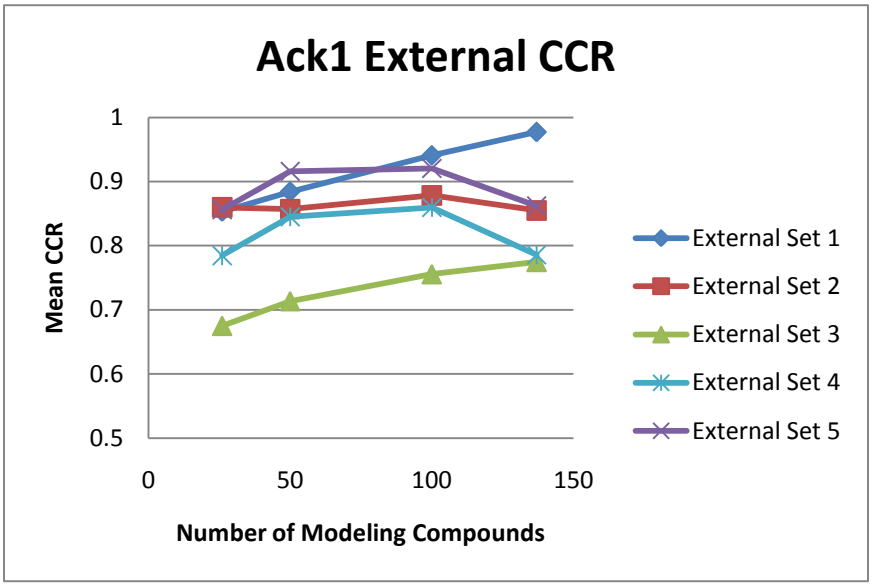

ESR2
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
GR
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PARP1
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PDE5
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

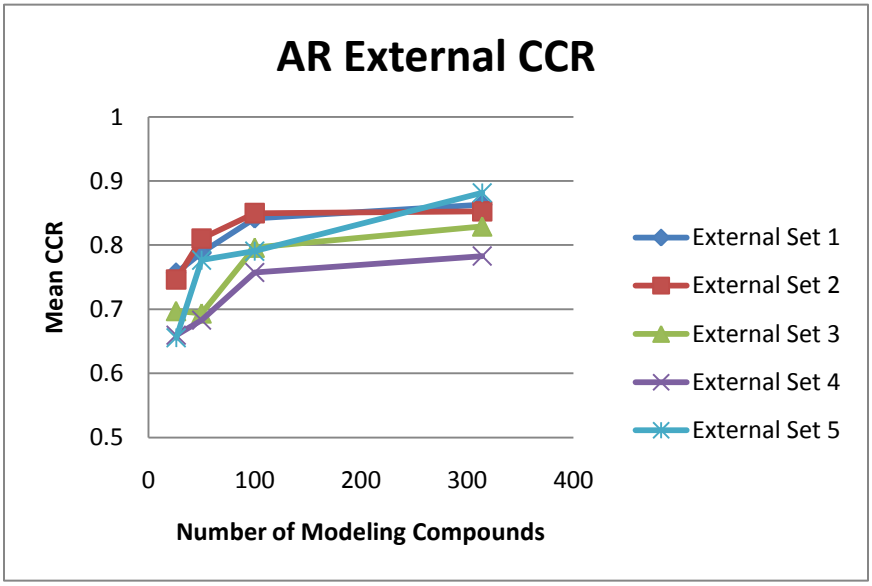

PNP
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
PPARG
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
REN
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5
SRC
True Positive Rate
False Positive Rate
Validation set 1
Validation Set 2
Validation Set 3
Validation Set 4
Validation Set 5

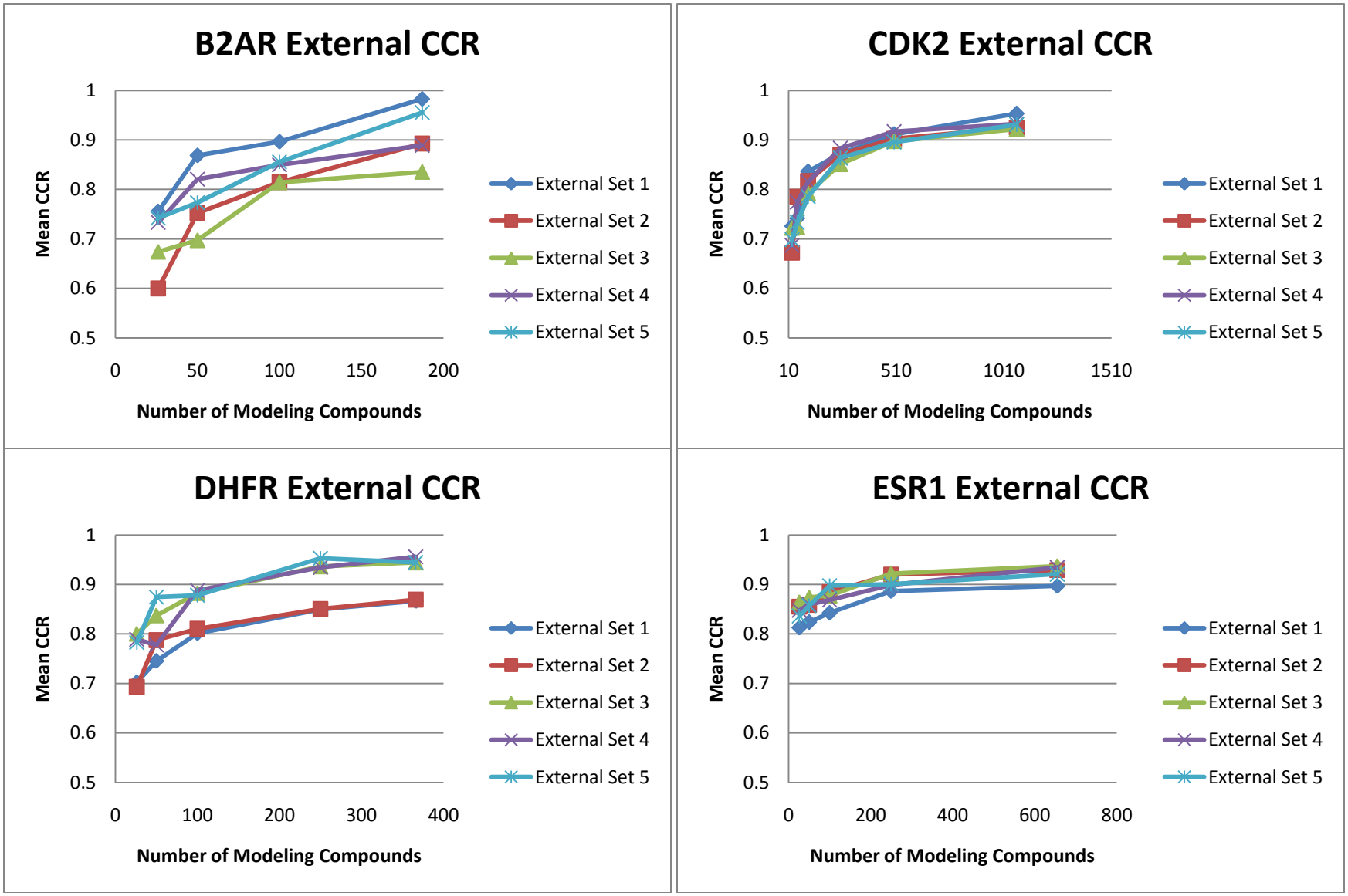
Appendix VI: QSAR Validation Set Statistics

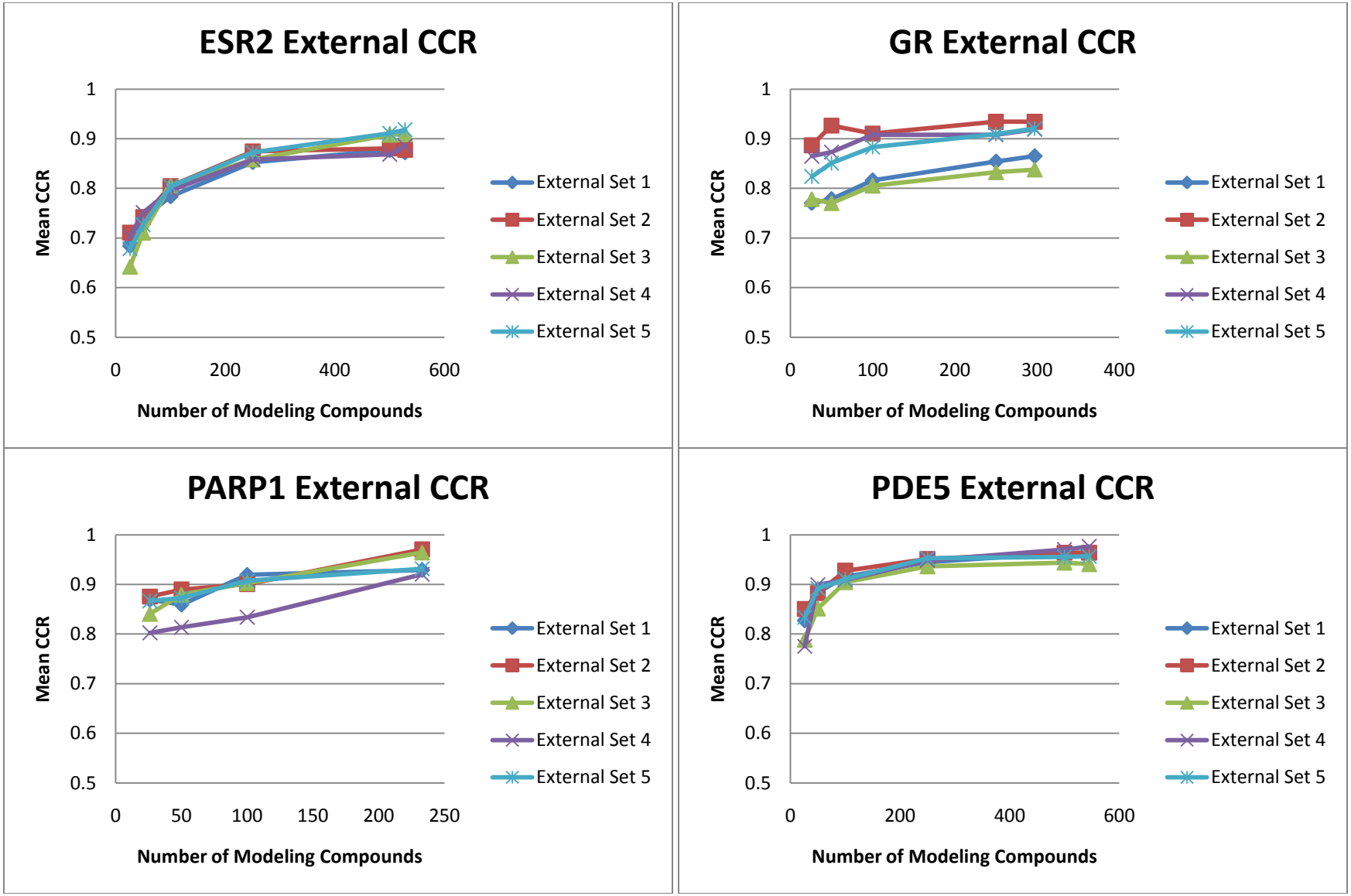
Contained within this appendix are the plots demonstrating how QSAR statistics are effected by the size of the modeling set used to build the QSAR models. The average predictive accuracy of models increases or stays constant as the modeling set size increases. Generally, the stability (i.e. the inverse of the variation in predictive power for multiple samples of the same size) also increases as the modeling set size increases. These results corroborate the expected results that more compounds lead to more predictive models.

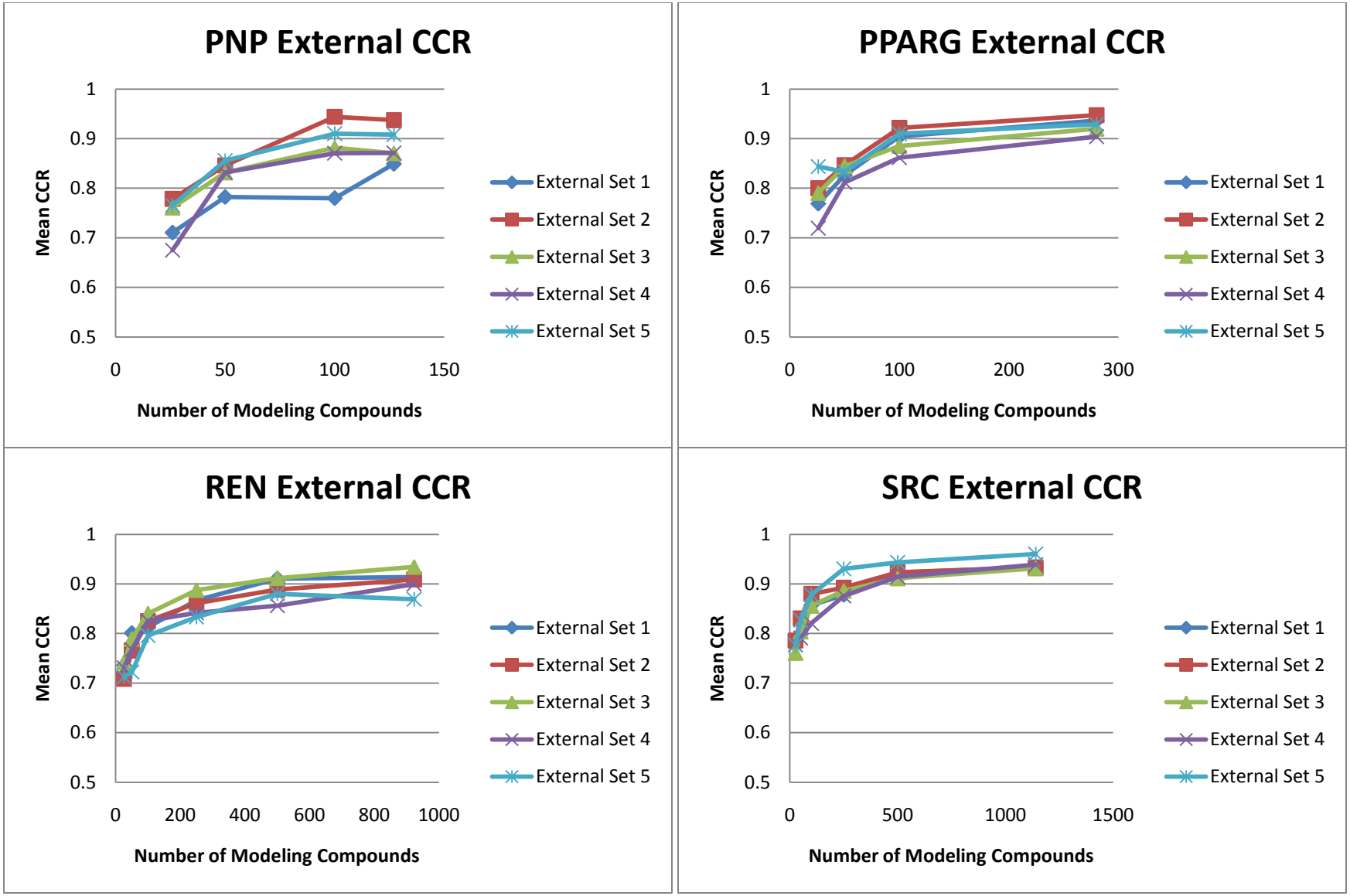
Predictive Power (Mean CCR)



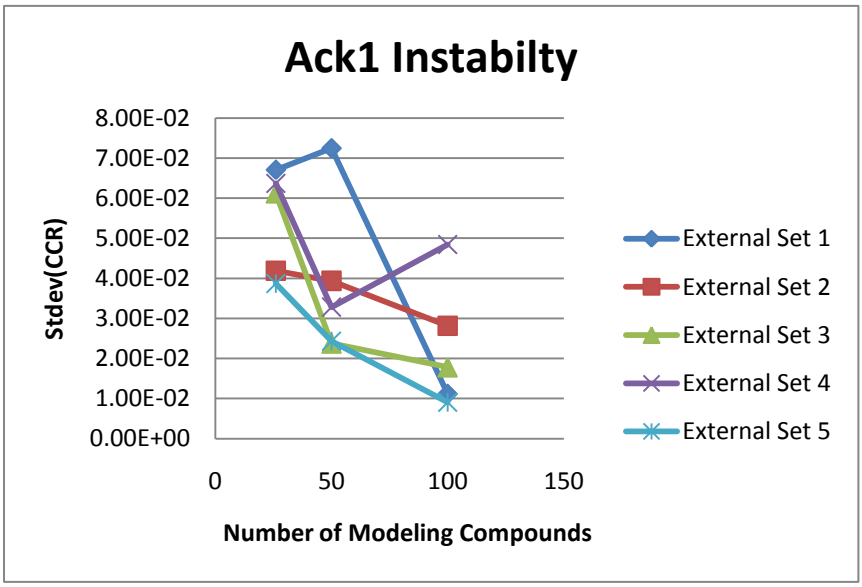


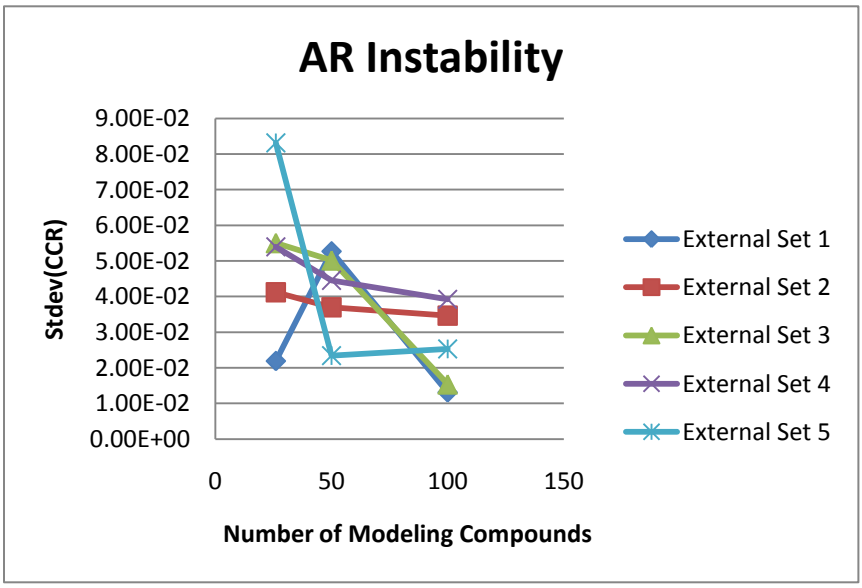

B2AR External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
50
100
150
200
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
CDK2 External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
10
510
1010
1510
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
DHFR External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
100
200
300
400
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
ESR1 External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
200
400
600
800
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

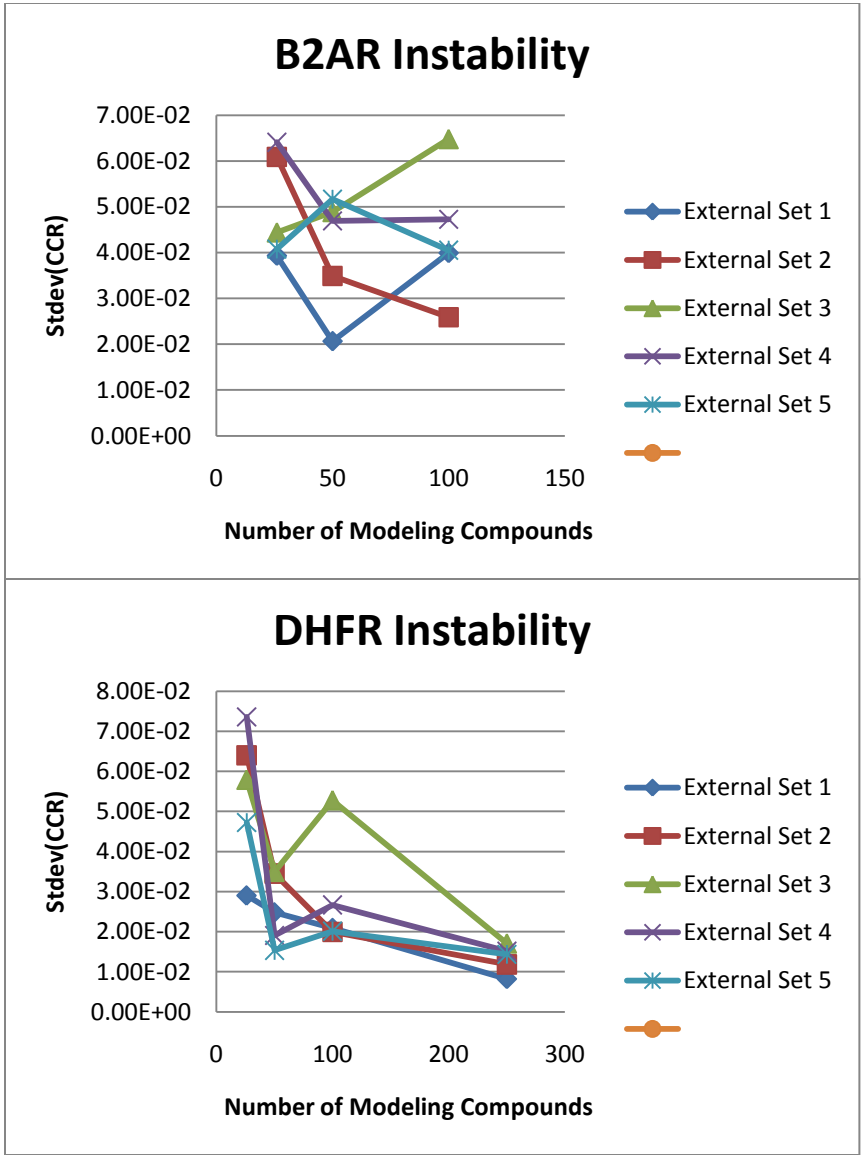

ESR2 External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
200
400
600
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
GR External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
100
200
300
400
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
PARP1 External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
50
100
150
200
250
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
PDE5 External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
200
400
600
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

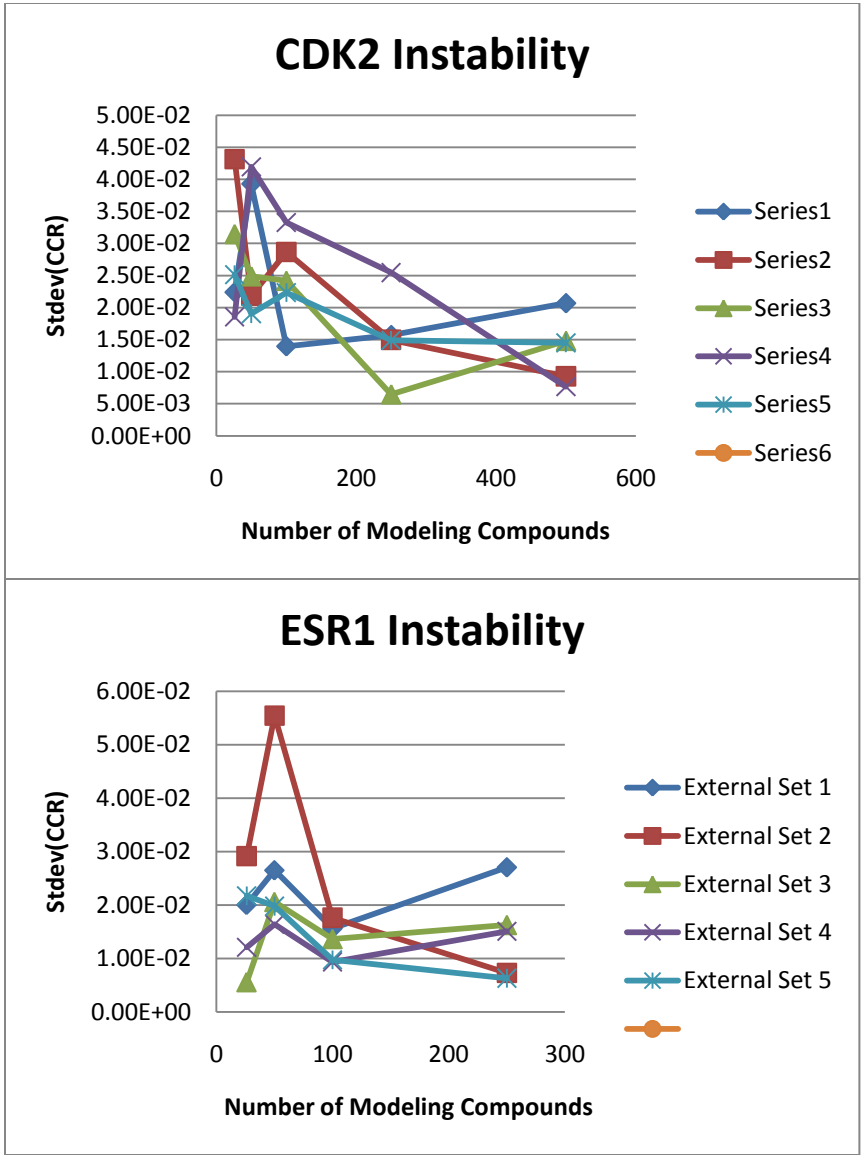

PNP External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
50
100
150
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
PPARG External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
100
200
300
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
REN External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
200
400
600
800
1000
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5
SRC External CCR
Mean CCR
1
0.9
0.8
0.7
0.6
0.5
0
500
1000
1500
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

Model Stability (Stdev CCR)



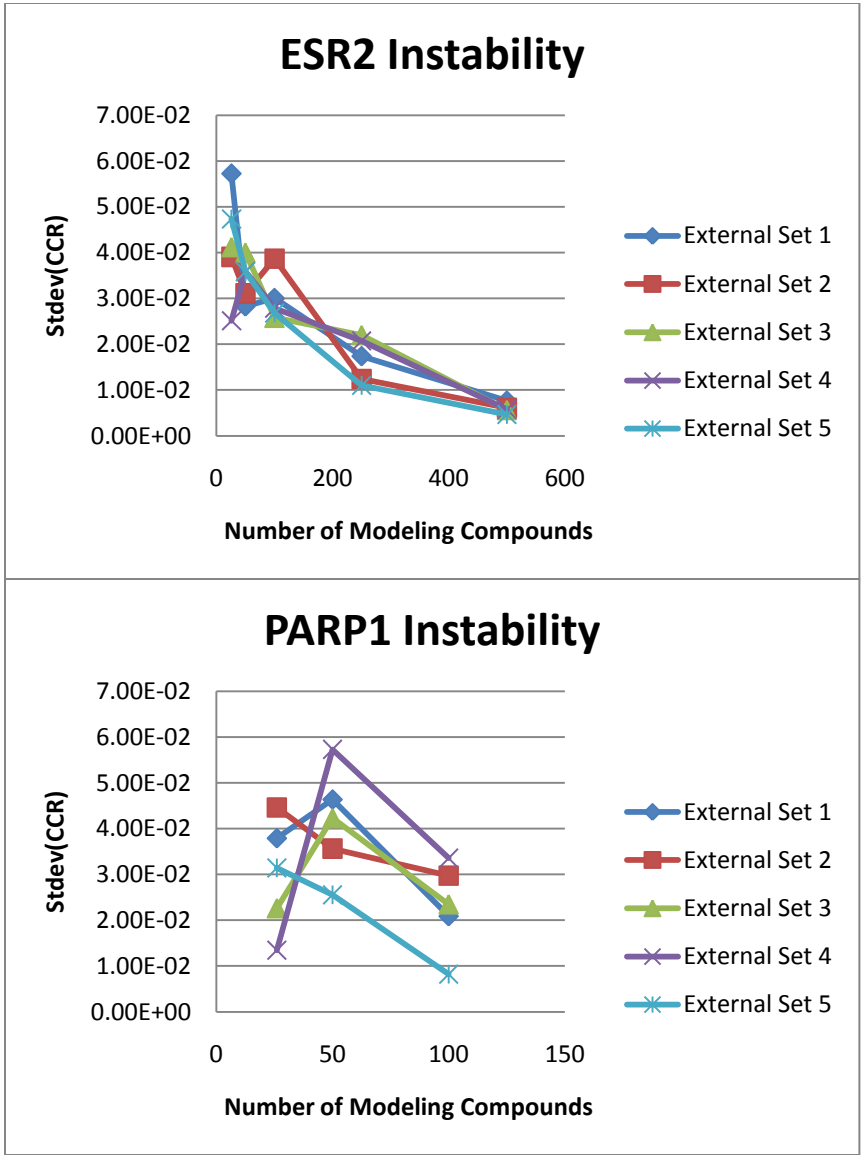


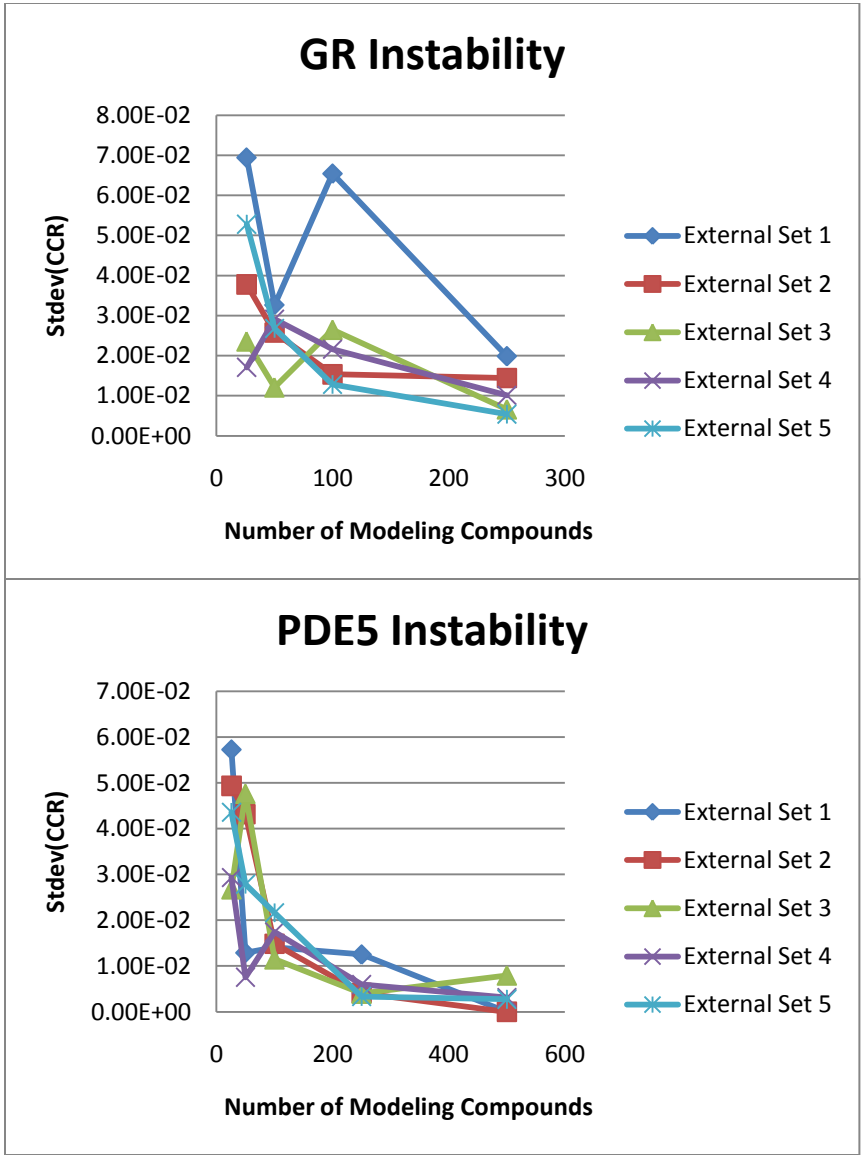

B2AR Instability
Stdev(CCR)
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
50
100
150
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5


CDK2 Instability
Stdev(CCR)
5.00E-02
4.50E-02
4.00E-02
3.50E-02
3.00E-02
2.50E-02
2.00E-02
1.50E-02
1.00E-02
5.00E-03
0.00E+00
0
200
400
600
Number of Modeling Compounds
Series1
Series2
Series3
Series4
Series5
Series6

DHFR Instability
Stdev(CCR)
8.00E-02
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
100
200
300
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

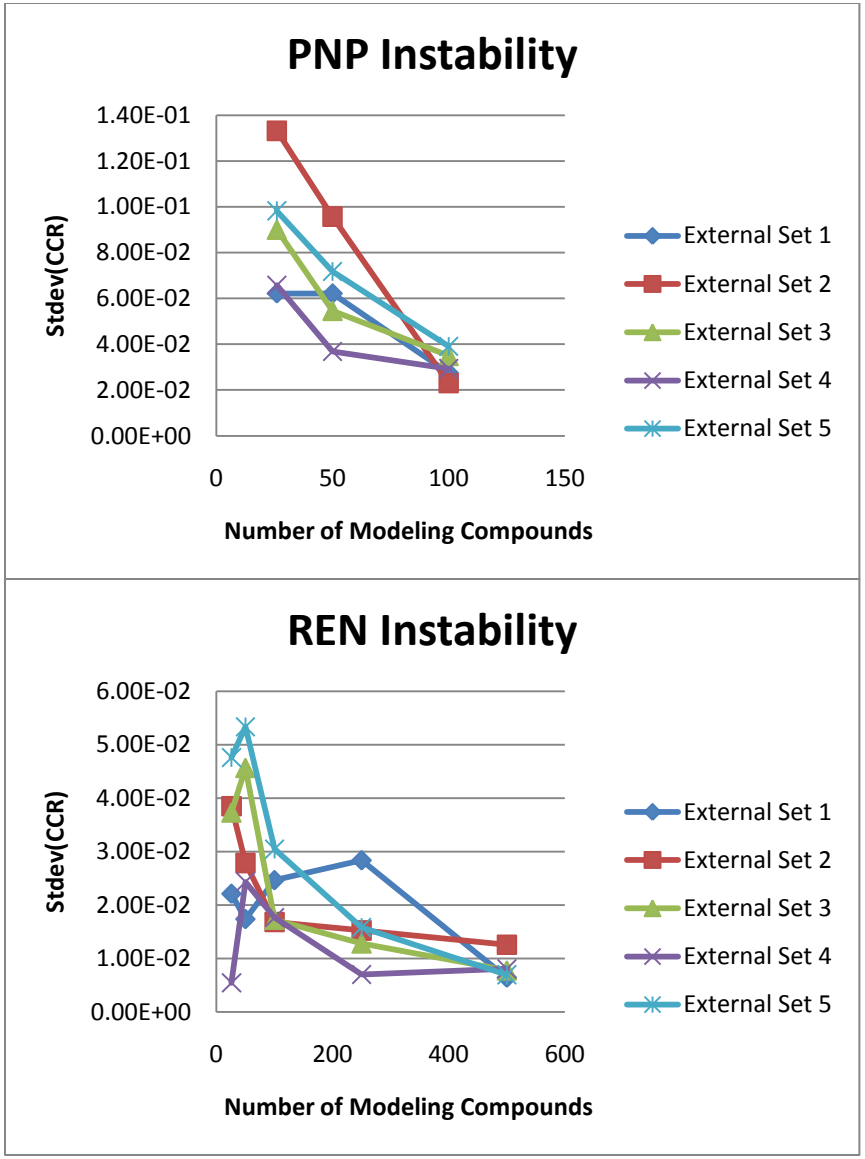
ESR1 Instability
Stdev(CCR)
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
100
200
300
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

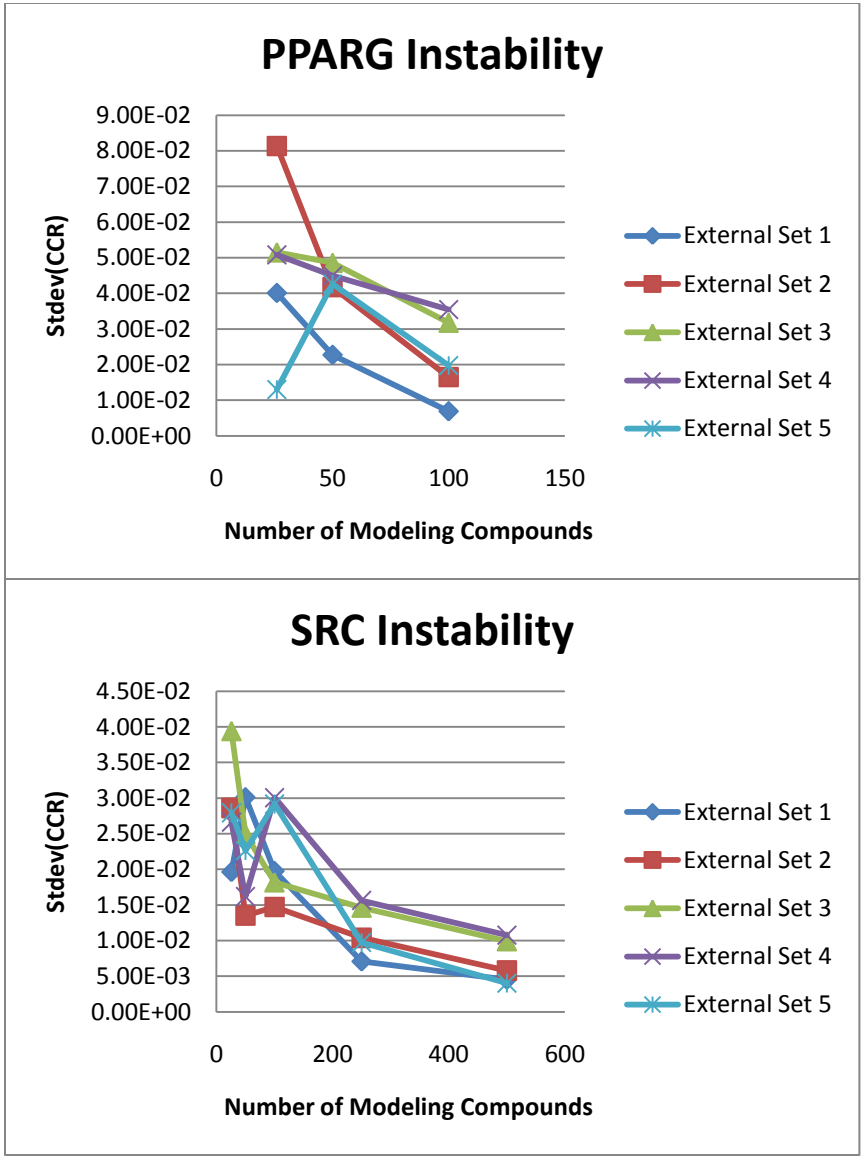

ESR2 Instability
Stdev(CCR)
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
200
400
600
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5


GR Instability
Stdev(CCR)
8.00E-02
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
100
200
300
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

PARP1 Instability
Stdev(CCR)
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
50
100
150
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

PDE5 Instability
Stdev(CCR)
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
200
400
600
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5


PNP Instability
Stdev(CCR)
1.40E-01
1.20E-01
1.00E-01
8.00E-02
6.00E-02
4.00E-02
2.00E-02
0.00E+00
0
50
100
150
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

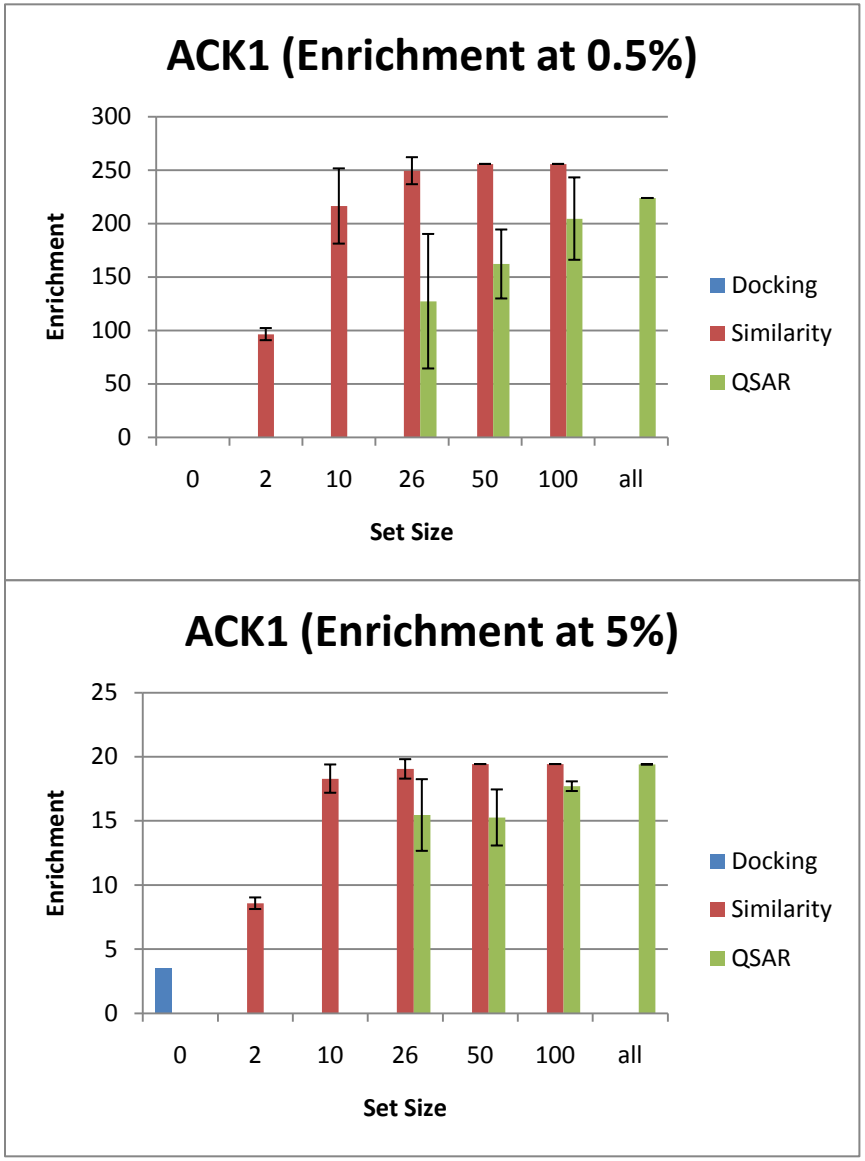

PPARG Instability
Stdev(CCR)
9.00E-02
8.00E-02
7.00E-02
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
50
100
150
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

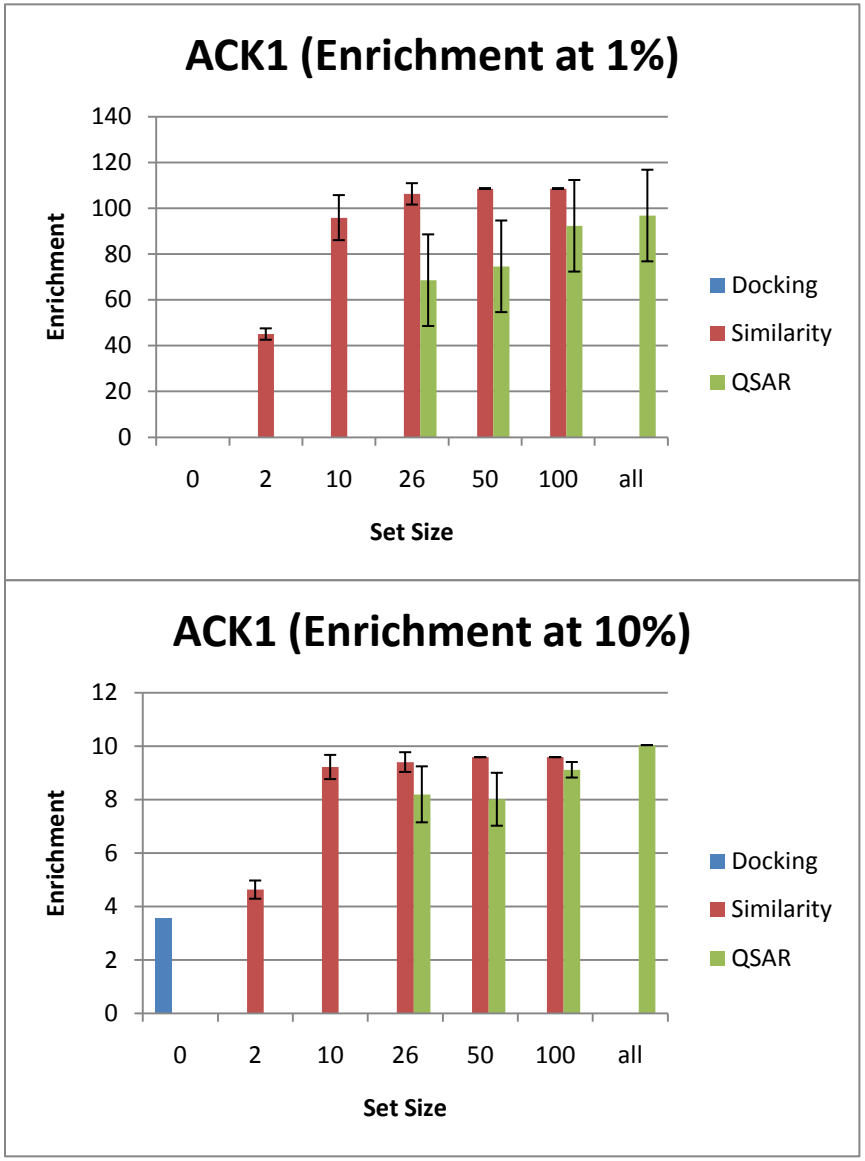
REN Instability
Stdev(CCR)
6.00E-02
5.00E-02
4.00E-02
3.00E-02
2.00E-02
1.00E-02
0.00E+00
0
200
400
600
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

SRC Instability
Stdev(CCR)
4.50E-02
4.00E-02
3.50E-02
3.00E-02
2.50E-02
2.00E-02
1.50E-02
1.00E-02
5.00E-03
0.00E+00
0
200
400
600
Number of Modeling Compounds
External Set 1
External Set 2
External Set 3
External Set 4
External Set 5

Appendix VII: Enrichment Plots

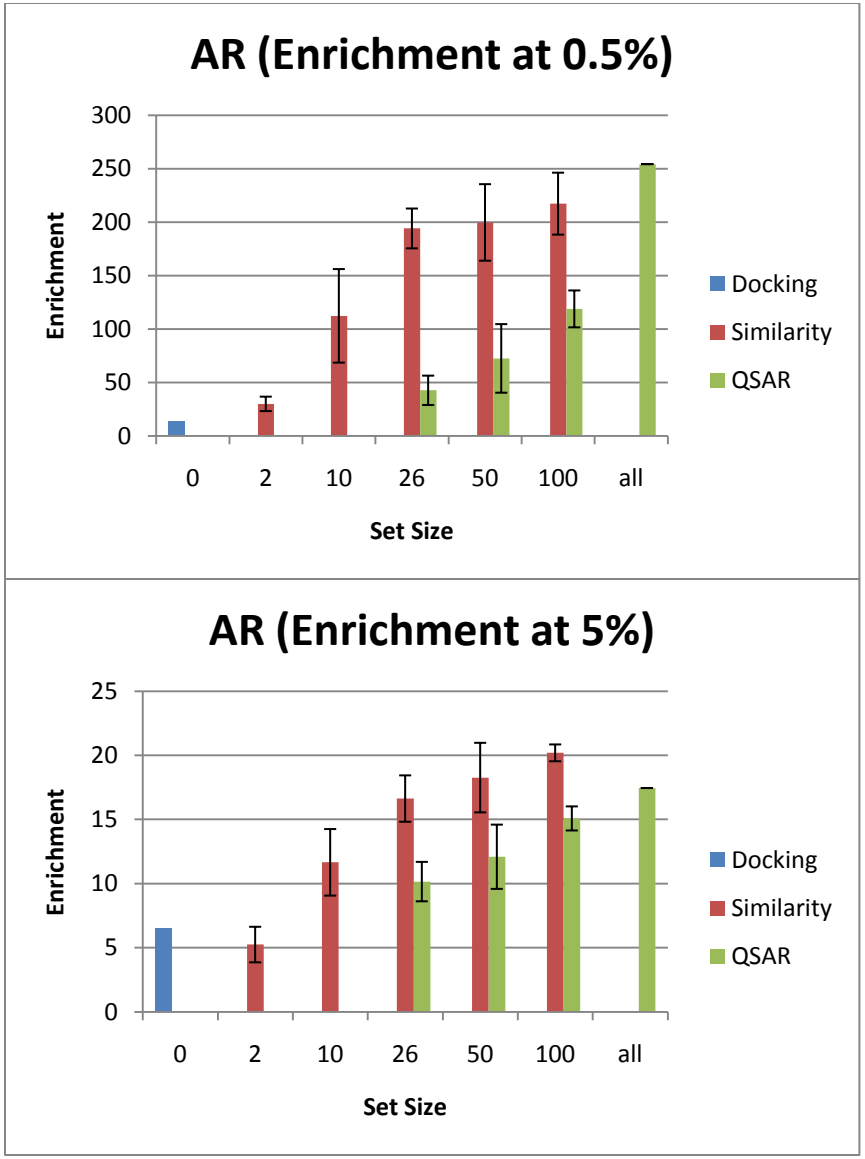
Contained within this appendix are the summary plots of enrichment at different cutoffs in the screening library. Easily compared in these plots are the different methodologies and the effects of the amount of available data on ability to enrich highly ranked compounds. It is clear that using more ligand data provides better enrichment with docking generally yielding enrichments that are significantly less than the best enrichments obtained by ligand-based methods. In every case, similarity searching yielded enrichments much better than QSAR for the same level of input.

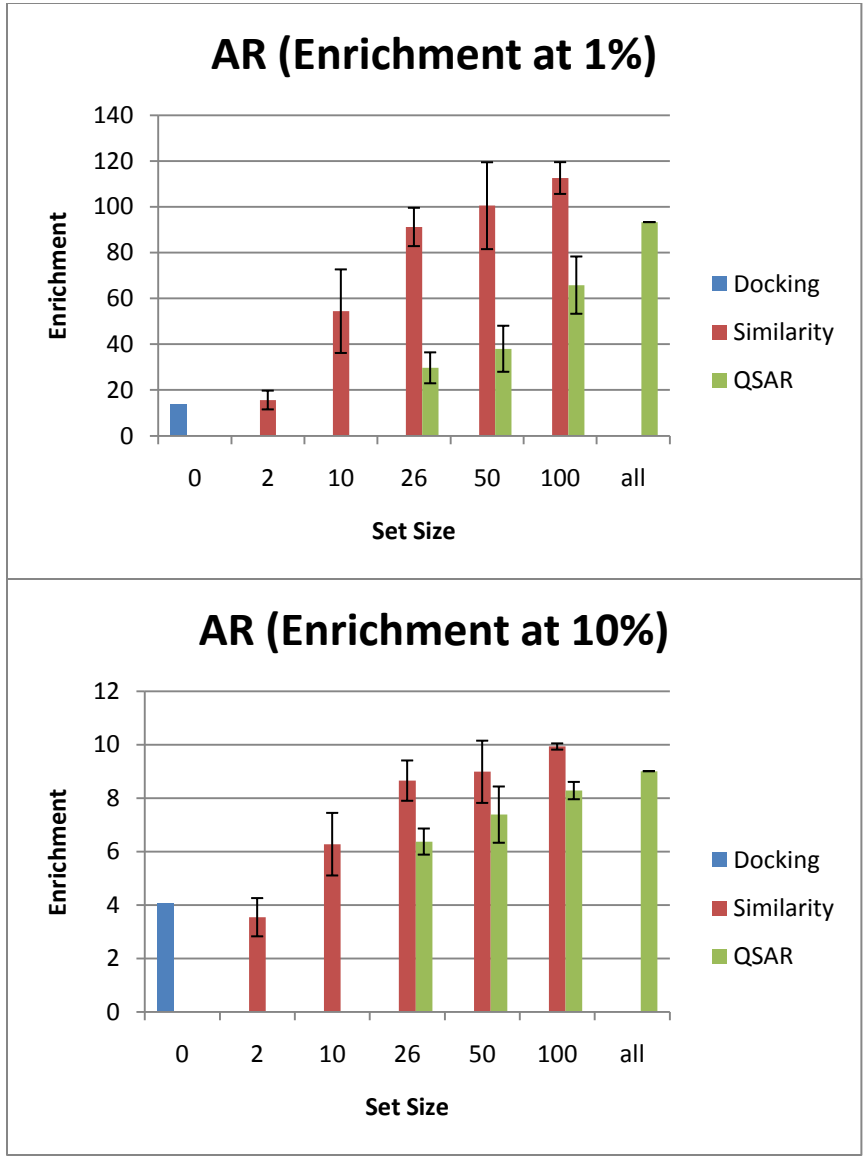

ACK1 (Enrichment at 0.5%)
Enrichment
300
250
200
150
100
50
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR


ACK1 (Enrichment at 1%)
Enrichment
140
120
100
80
60
40
20
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

ACK1 (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

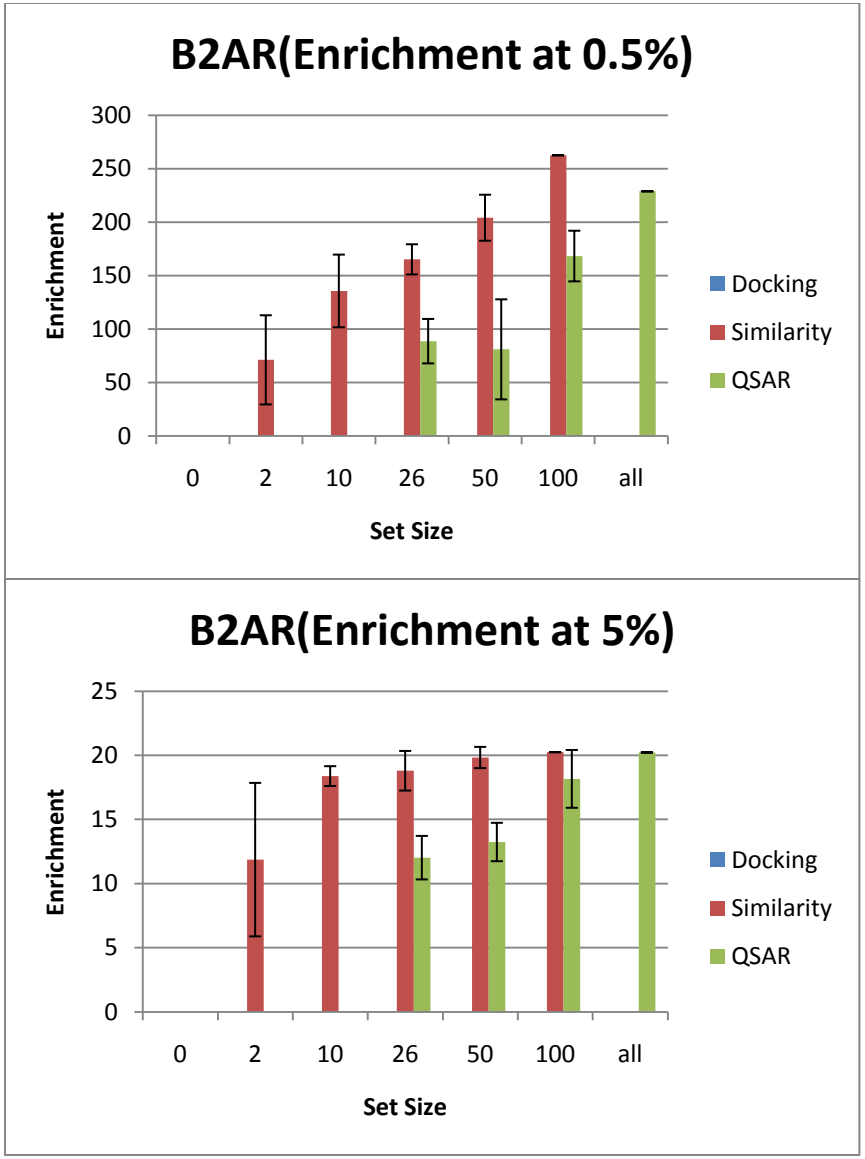
ACK1 (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

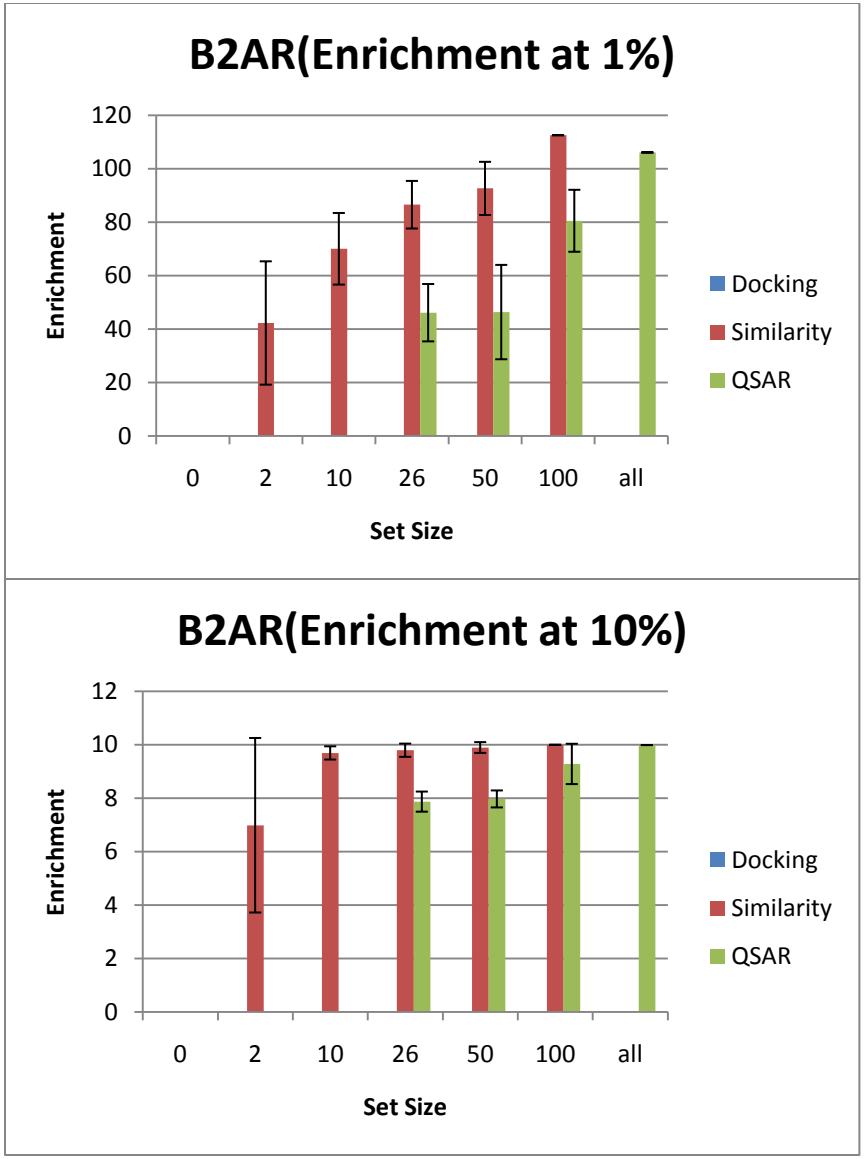

AR (Enrichment at 0.5%)
Enrichment
300
250
200
150
100
50
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR


AR (Enrichment at 1%)
Enrichment
140
120
100
80
60
40
20
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

AR (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

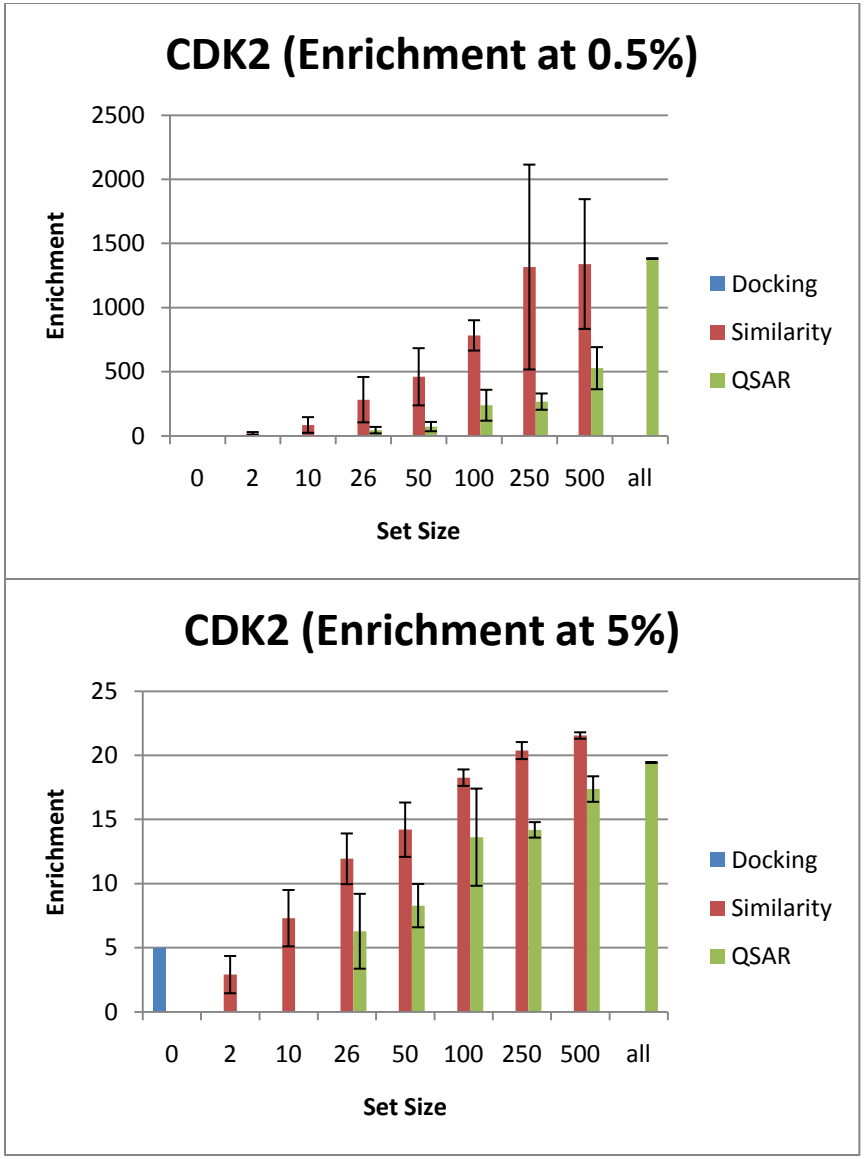
AR (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

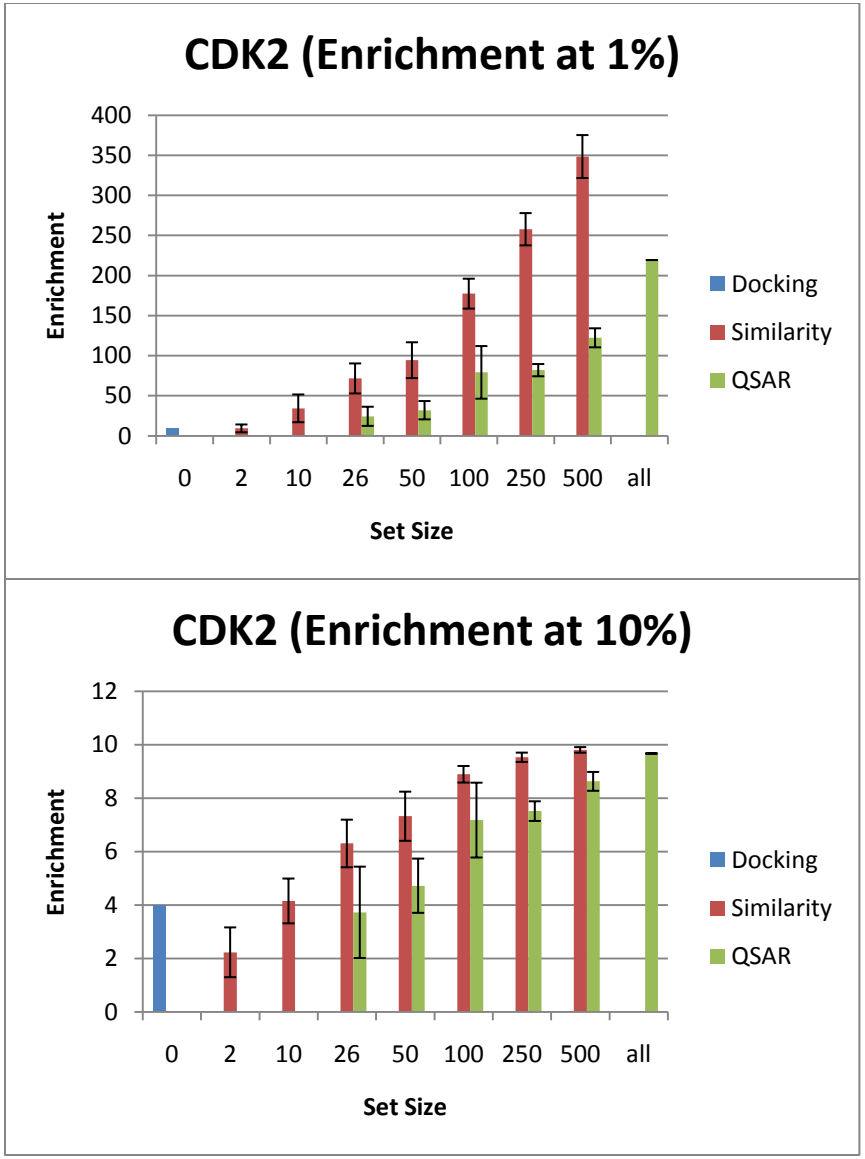

B2AR(Enrichment at 0.5%)
Enrichment
300
250
200
150
100
50
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR


B2AR(Enrichment at 1%)
Enrichment
120
100
80
60
40
20
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

B2AR(Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

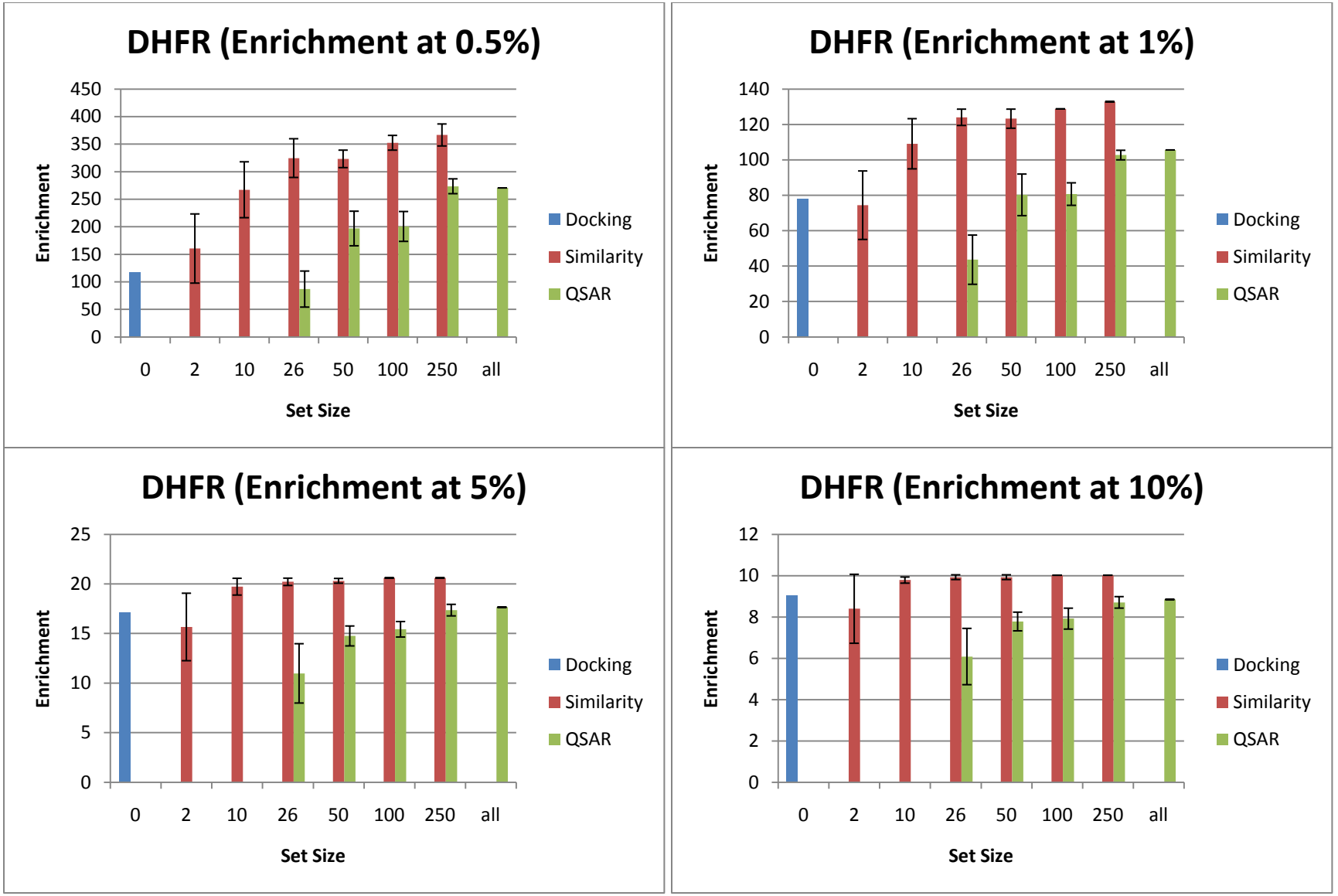
B2AR(Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

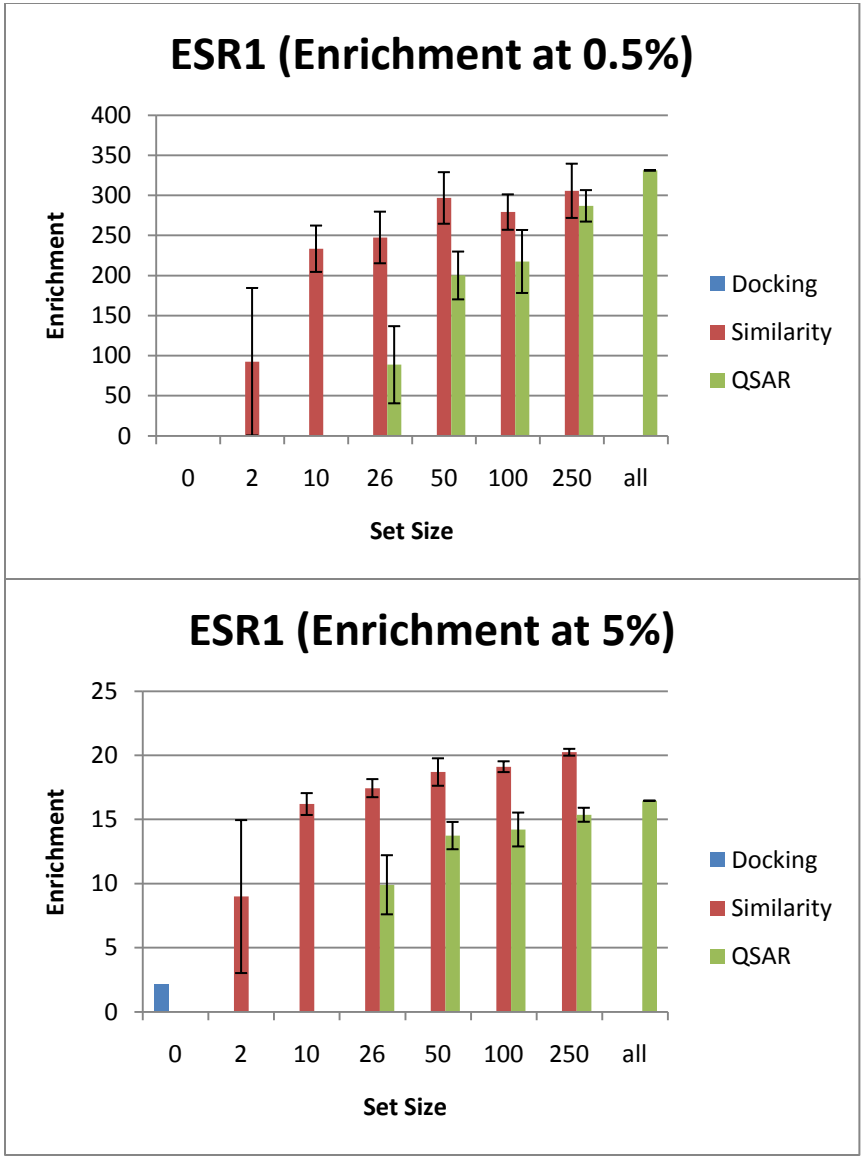

CDK2 (Enrichment at 0.5%)
Enrichment
2500
2000
1500
1000
500
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

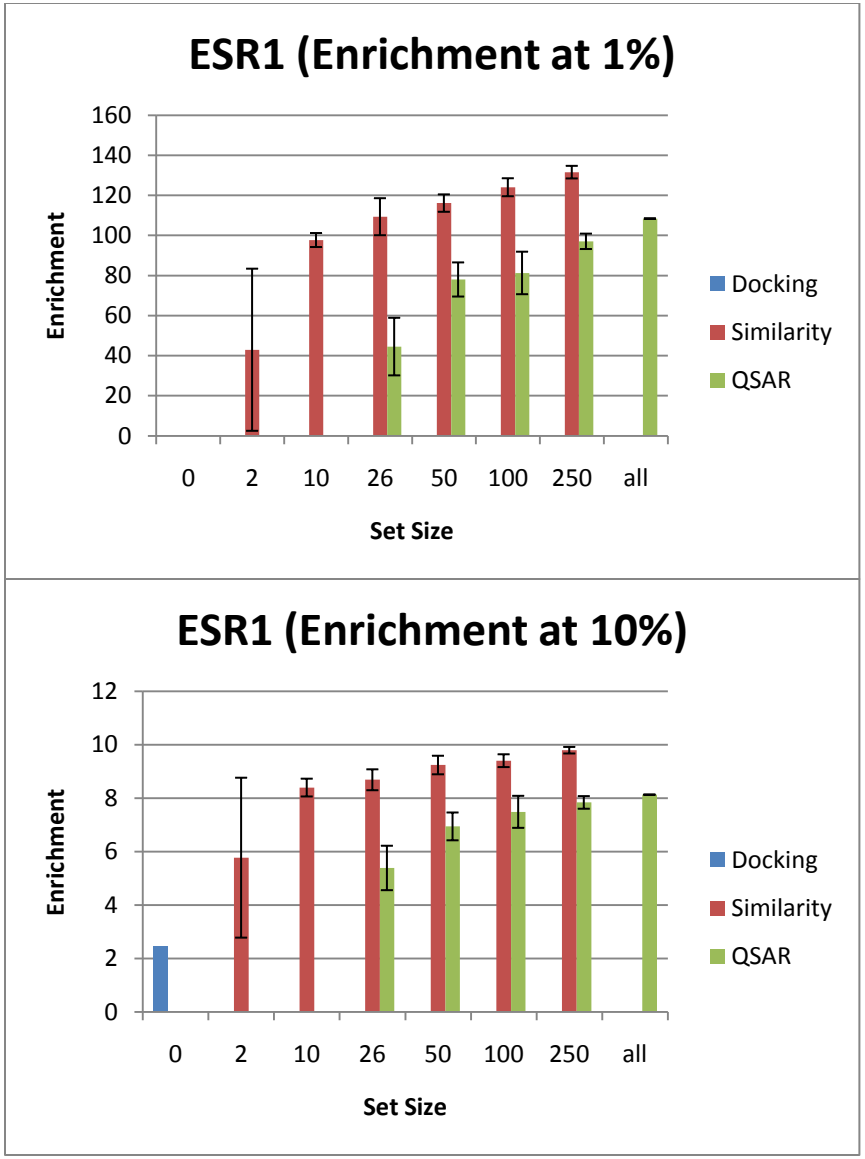

CDK2 (Enrichment at 1%)
Enrichment
400
350
300
250
200
150
100
50
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

CDK2 (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

CDK2 (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

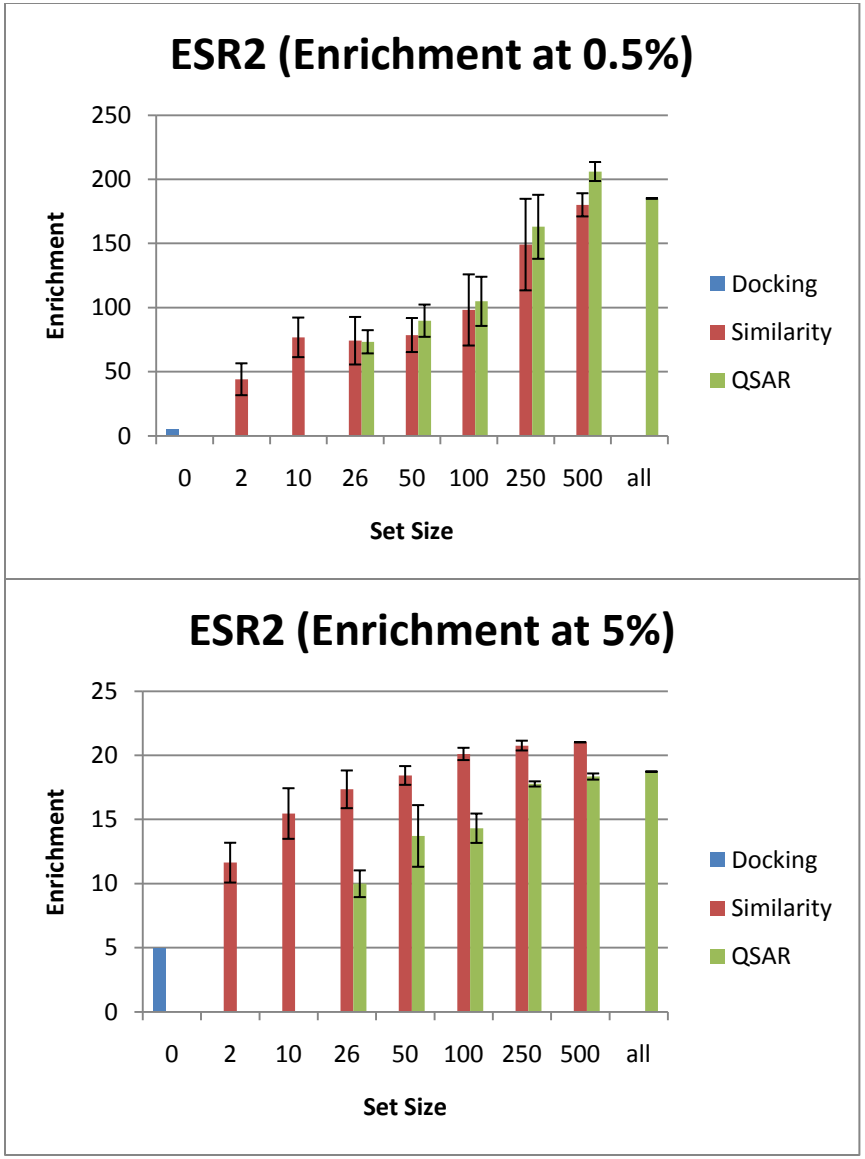

DHFR (Enrichment at 0.5%)
Enrichment
450
400
350
300
250
200
150
100
50
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR
DHFR (Enrichment at 1%)
Enrichment
140
120
100
80
60
40
20
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR
DHFR (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR
DHFR (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR

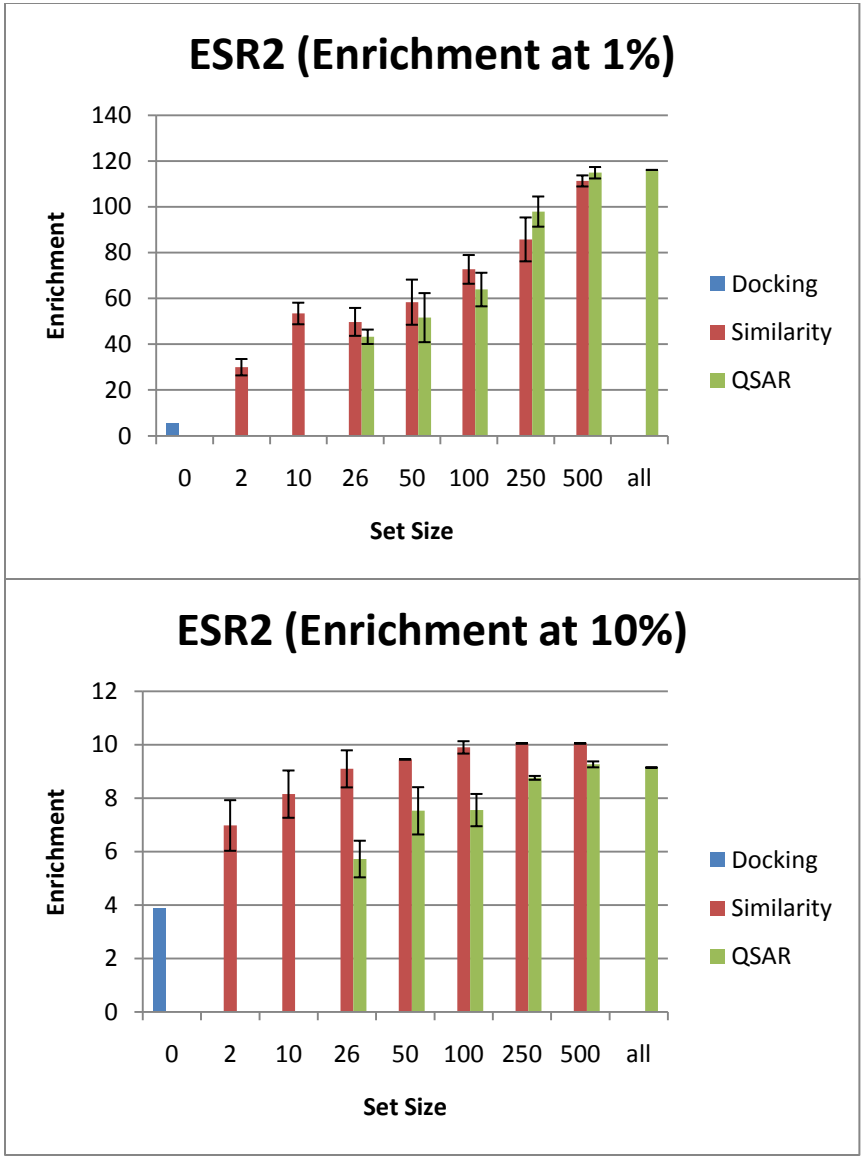

ESR1 (Enrichment at 0.5%)
Enrichment
400
350
300
250
200
150
100
50
0
0 2 10 26 50 100 250 all
Set Size
Docking
Similarity
QSAR


ESR1 (Enrichment at 1%)
Enrichment
160
140
120
100
80
60
40
20
0
0 2 10 26 50 100 250 all
Set Size
Docking
Similarity
QSAR

ESR1 (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0 2 10 26 50 100 250 all
Set Size
Docking
Similarity
QSAR

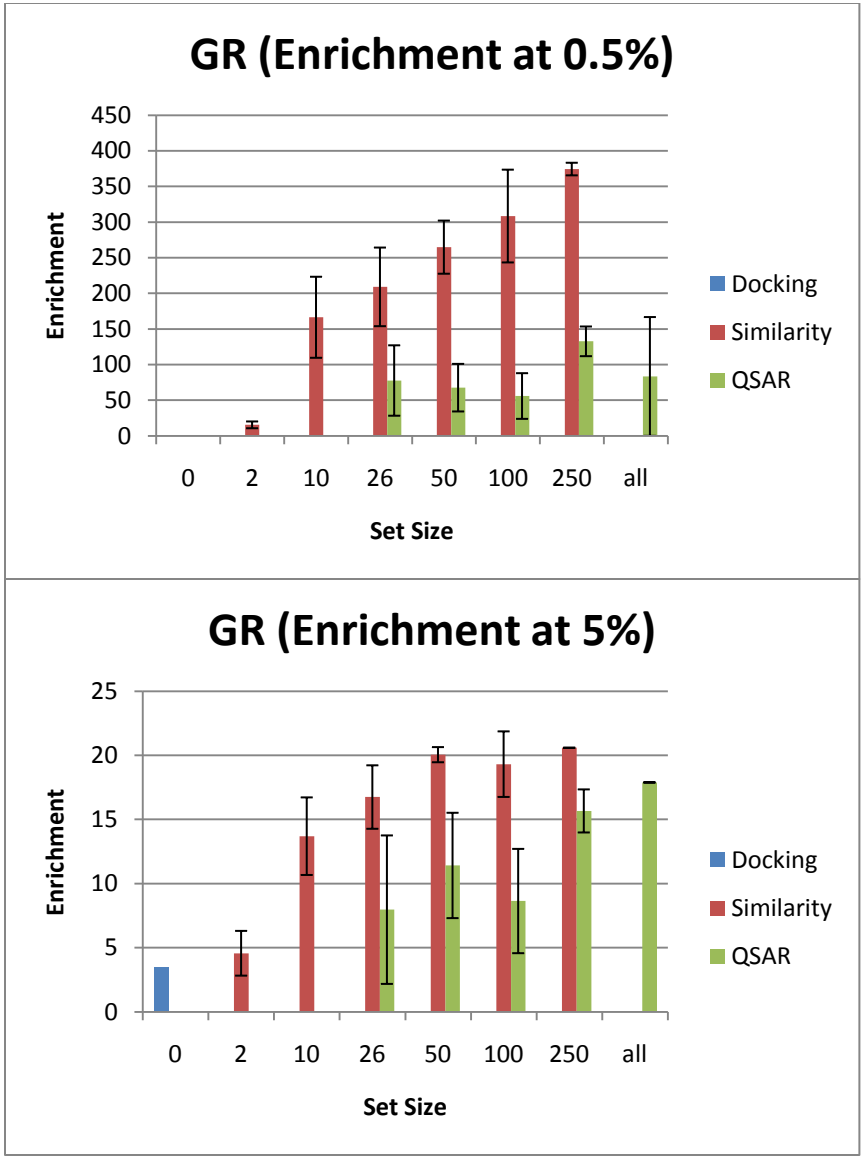
ESR1 (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0 2 10 26 50 100 250 all
Set Size
Docking
Similarity
QSAR

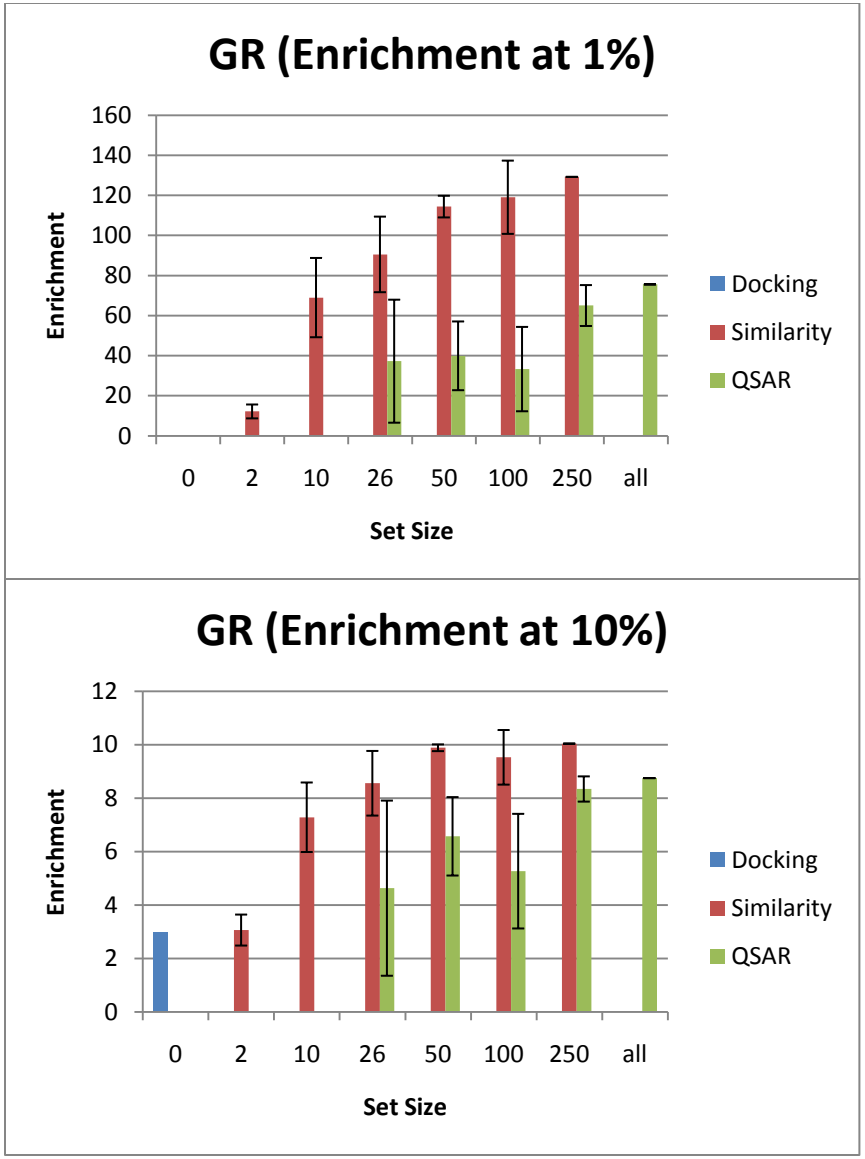

ESR2 (Enrichment at 0.5%)
Enrichment
250
200
150
100
50
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR


ESR2 (Enrichment at 1%)
Enrichment
140
120
100
80
60
40
20
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

ESR2 (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

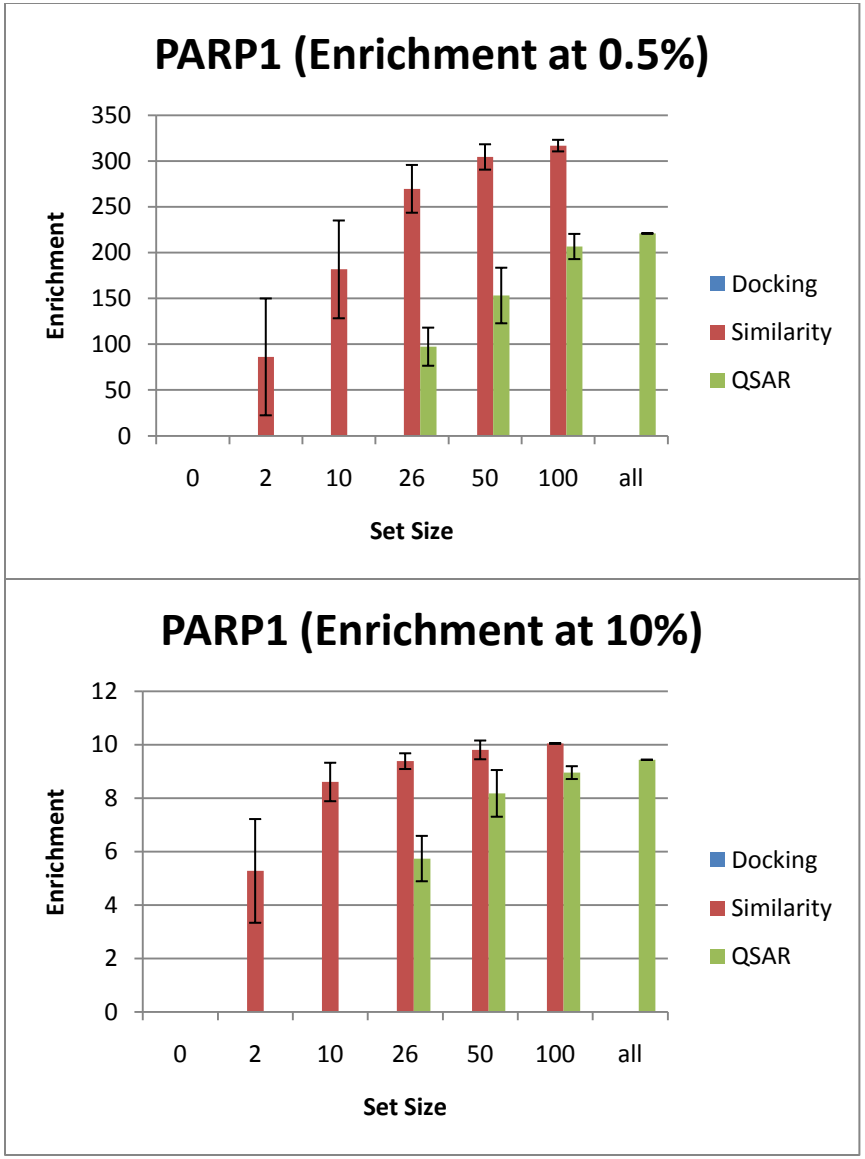
ESR2 (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

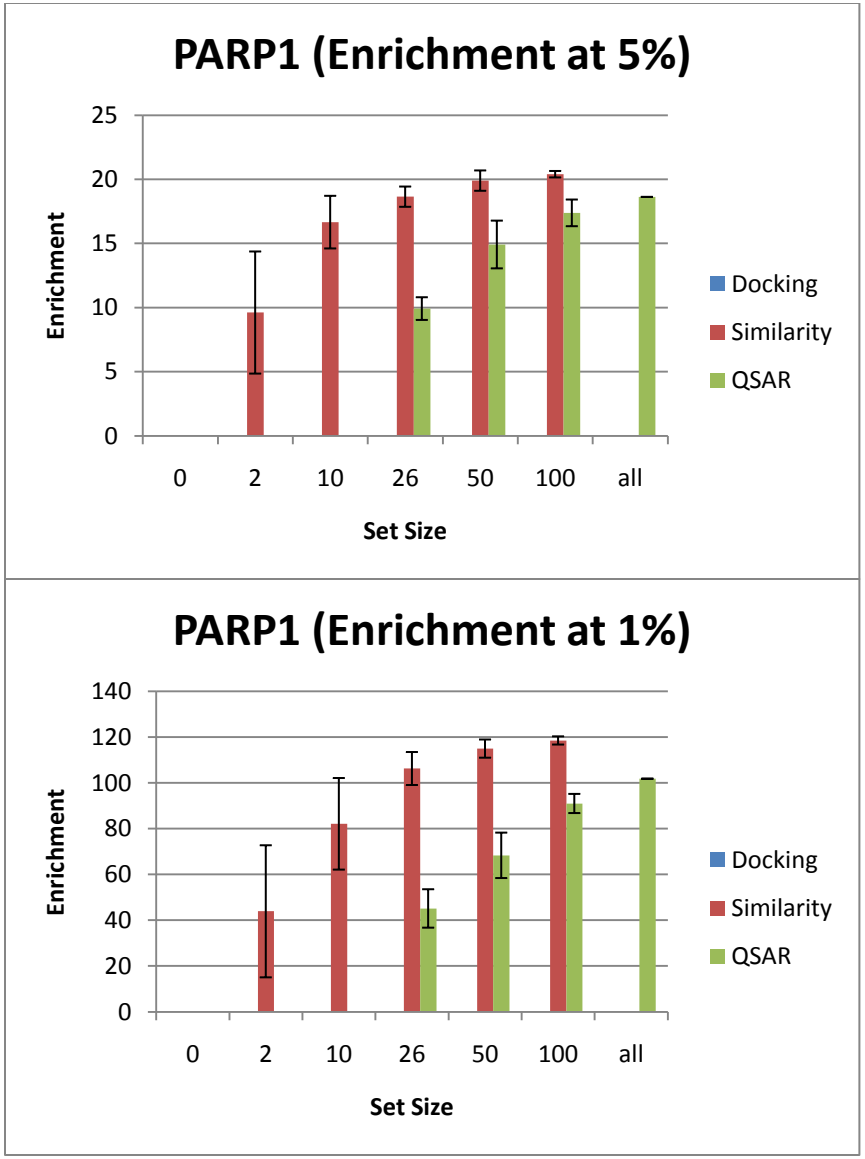

GR (Enrichment at 0.5%)
Enrichment
450
400
350
300
250
200
150
100
50
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR


GR (Enrichment at 1%)
Enrichment
160
140
120
100
80
60
40
20
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR

GR (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR

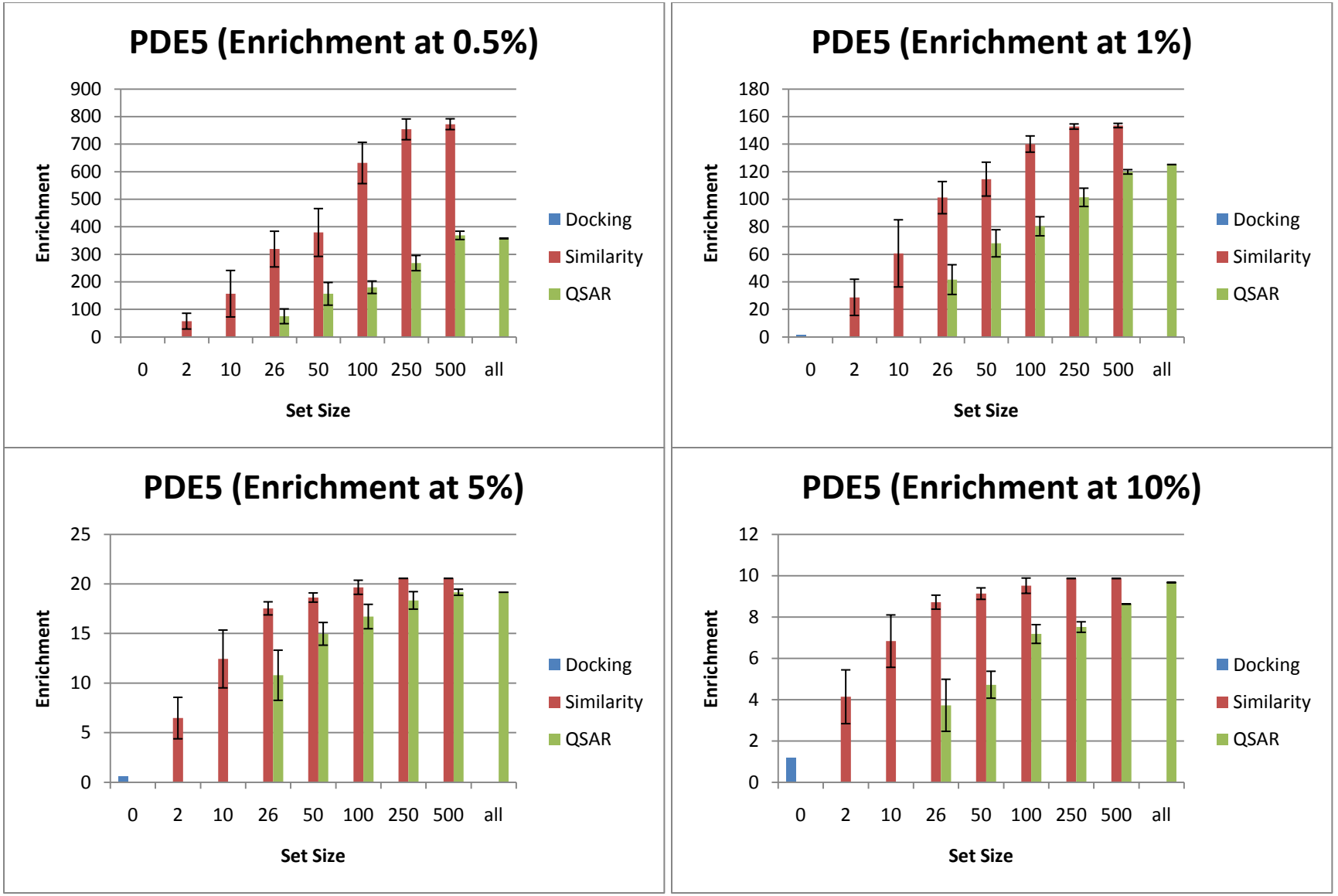
GR (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
250
all
Set Size
Docking
Similarity
QSAR

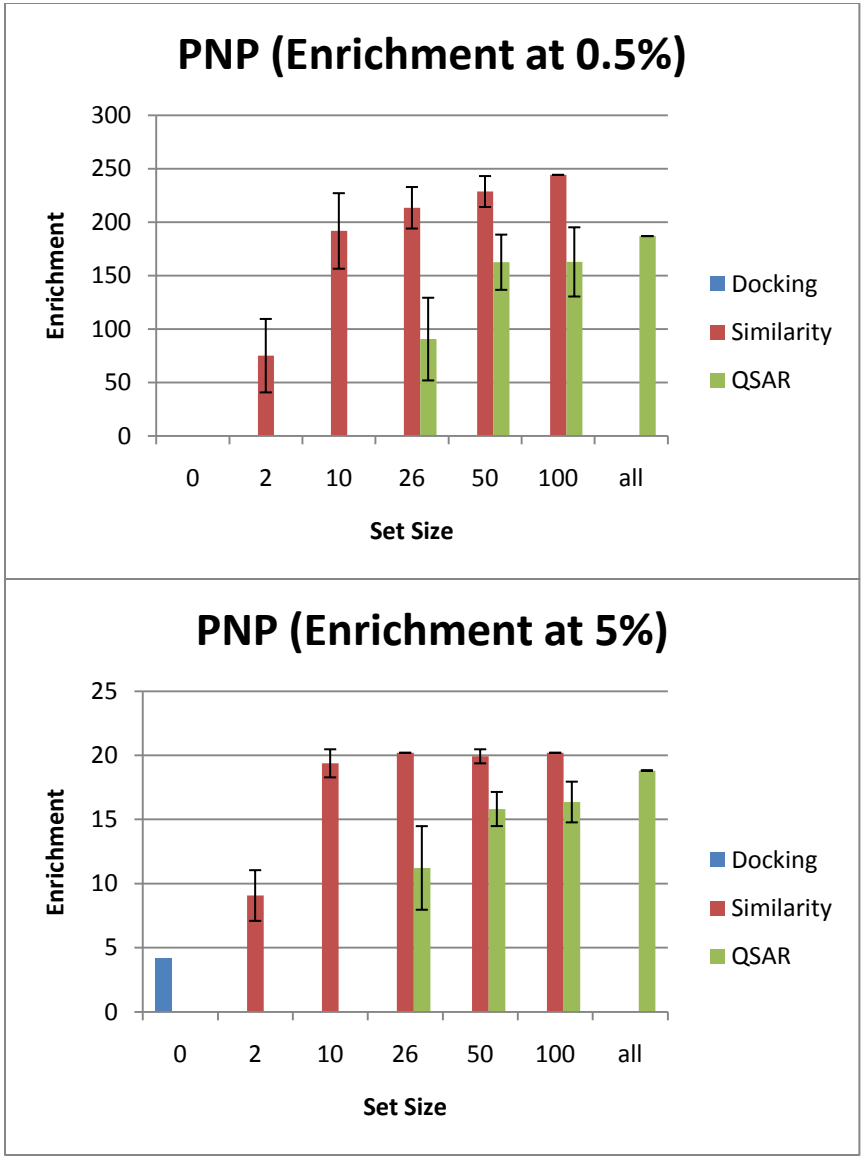

PARP1 (Enrichment at 0.5%)
Enrichment
350
300
250
200
150
100
50
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

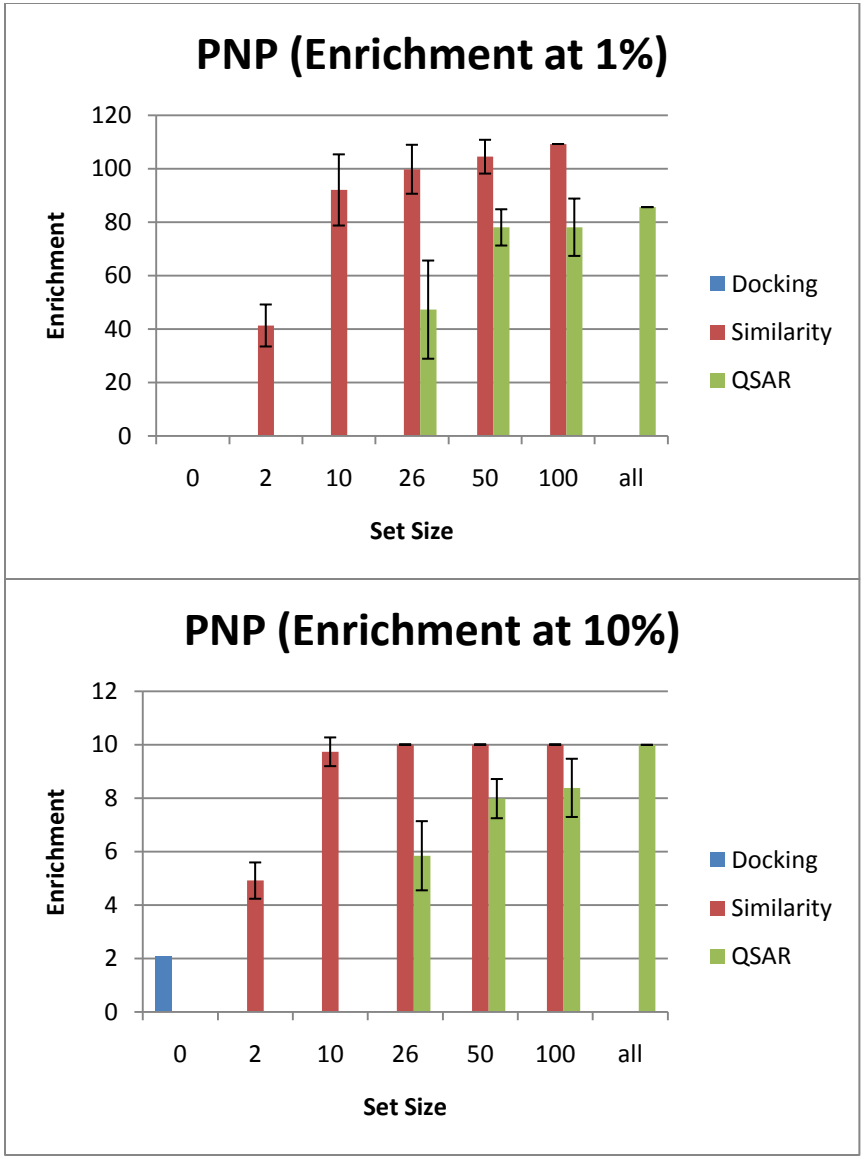

PARP1 (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

PARP1 (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

PARP1 (Enrichment at 1%)
Enrichment
140
120
100
80
60
40
20
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

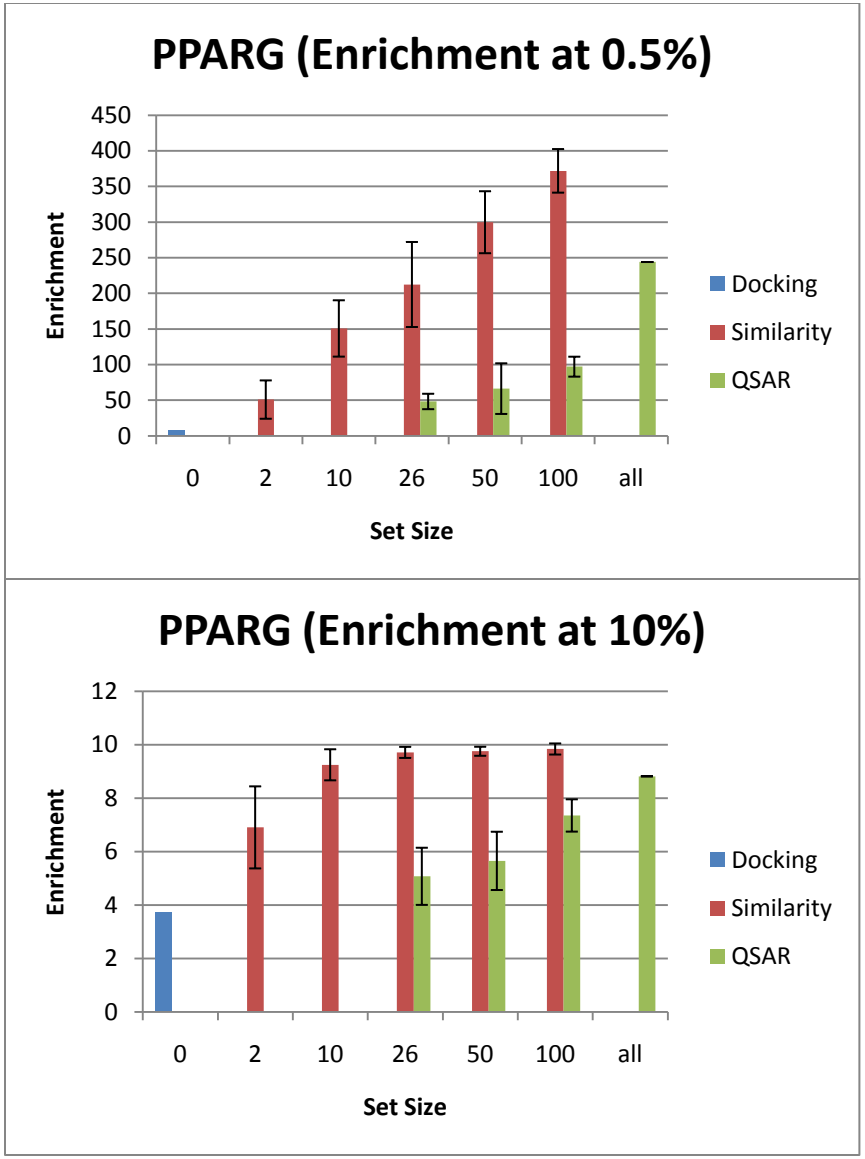

PDE5 (Enrichment at 0.5%)
Enrichment
900
800
700
600
500
400
300
200
100
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR
PDE5 (Enrichment at 1%)
Enrichment
180
160
140
120
100
80
60
40
20
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR
PDE5 (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR
PDE5 (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

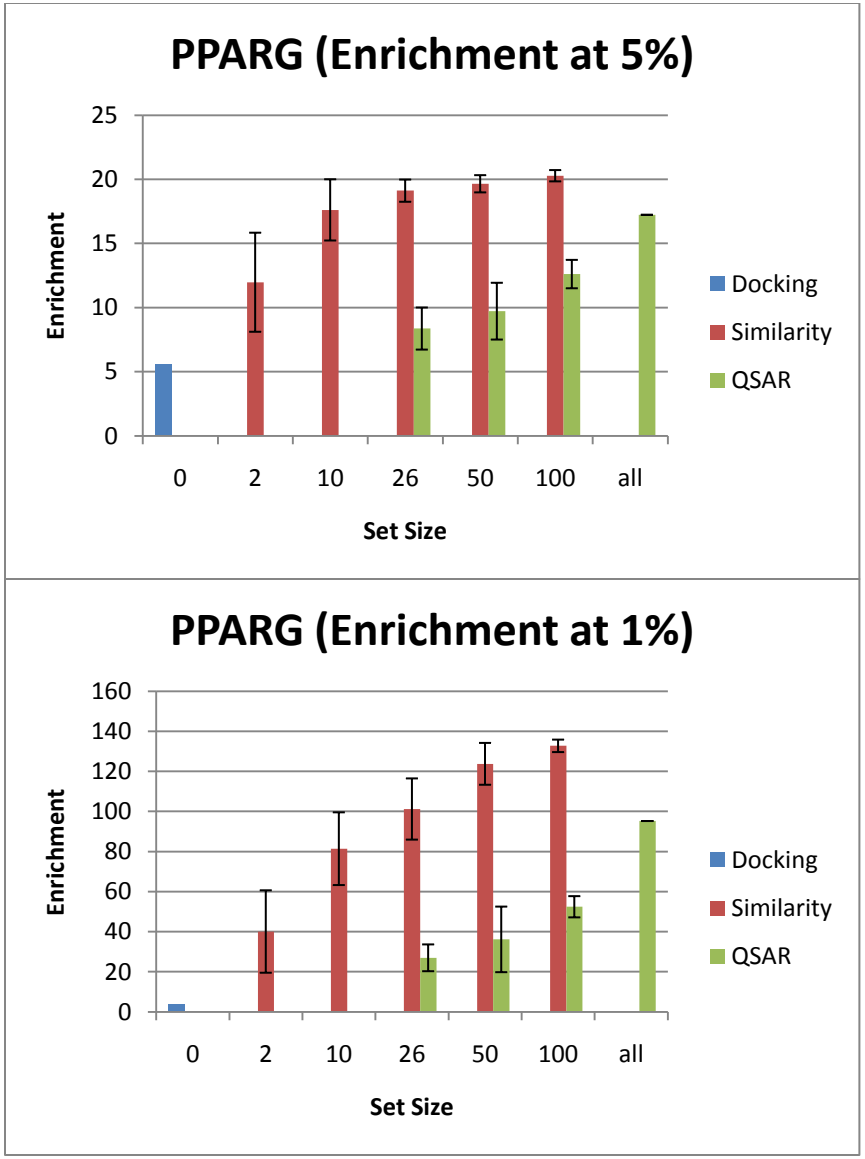

PNP (Enrichment at 0.5%)
Enrichment
300
250
200
150
100
50
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR


PNP (Enrichment at 1%)
Enrichment
120
100
80
60
40
20
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

PNP (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

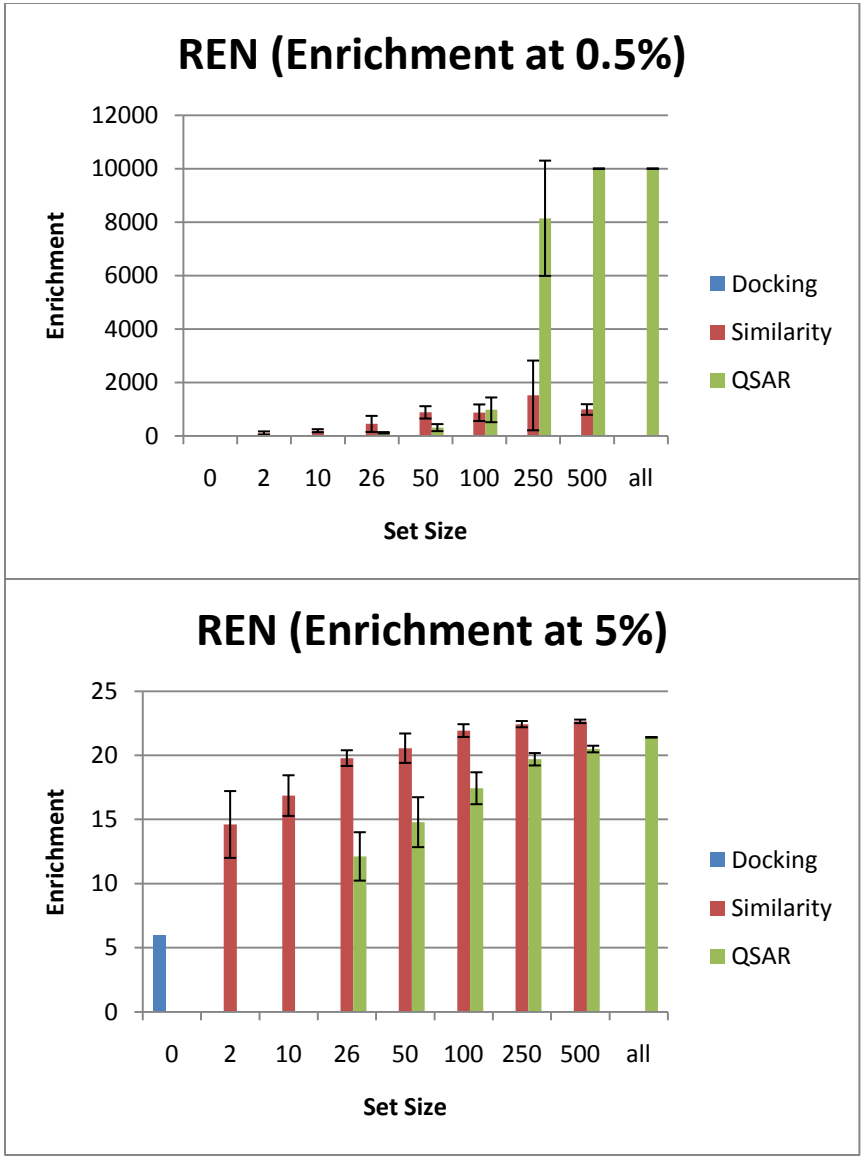
PNP (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

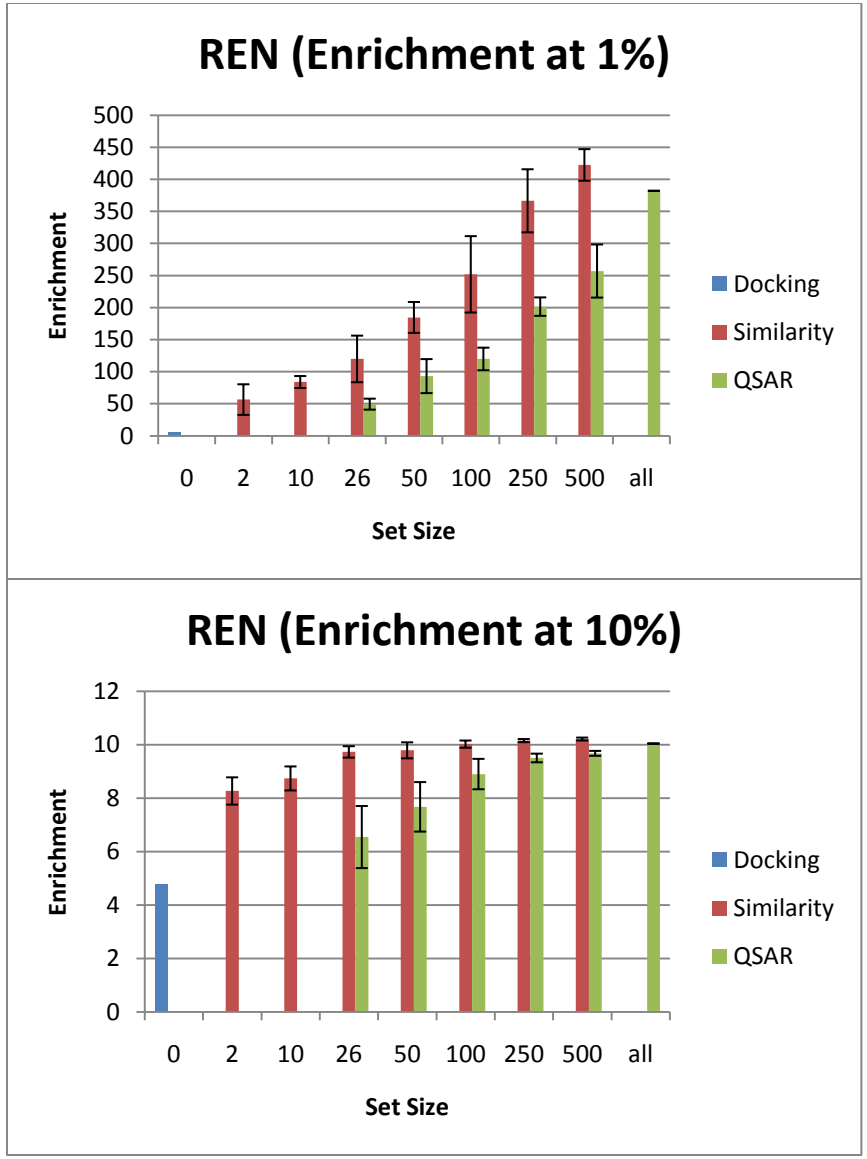

PPARG (Enrichment at 0.5%)
Enrichment
450
400
350
300
250
200
150
100
50
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR


PPARG (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

PPARG (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR

PPARG (Enrichment at 1%)
Enrichment
160
140
120
100
80
60
40
20
0
0
2
10
26
50
100
all
Set Size
Docking
Similarity
QSAR


REN (Enrichment at 0.5%)
Enrichment
12000
10000
8000
6000
4000
2000
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

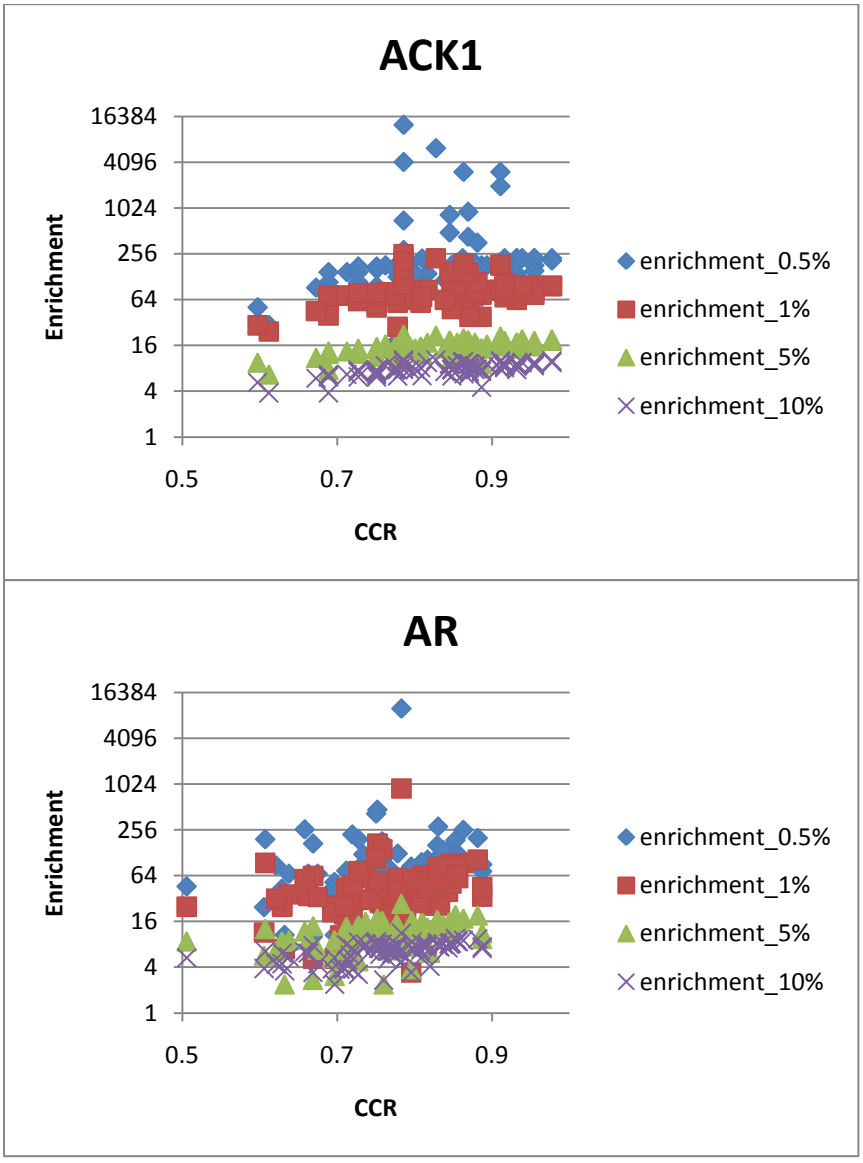

REN (Enrichment at 1%)
Enrichment
500
450
400
350
300
250
200
150
100
50
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

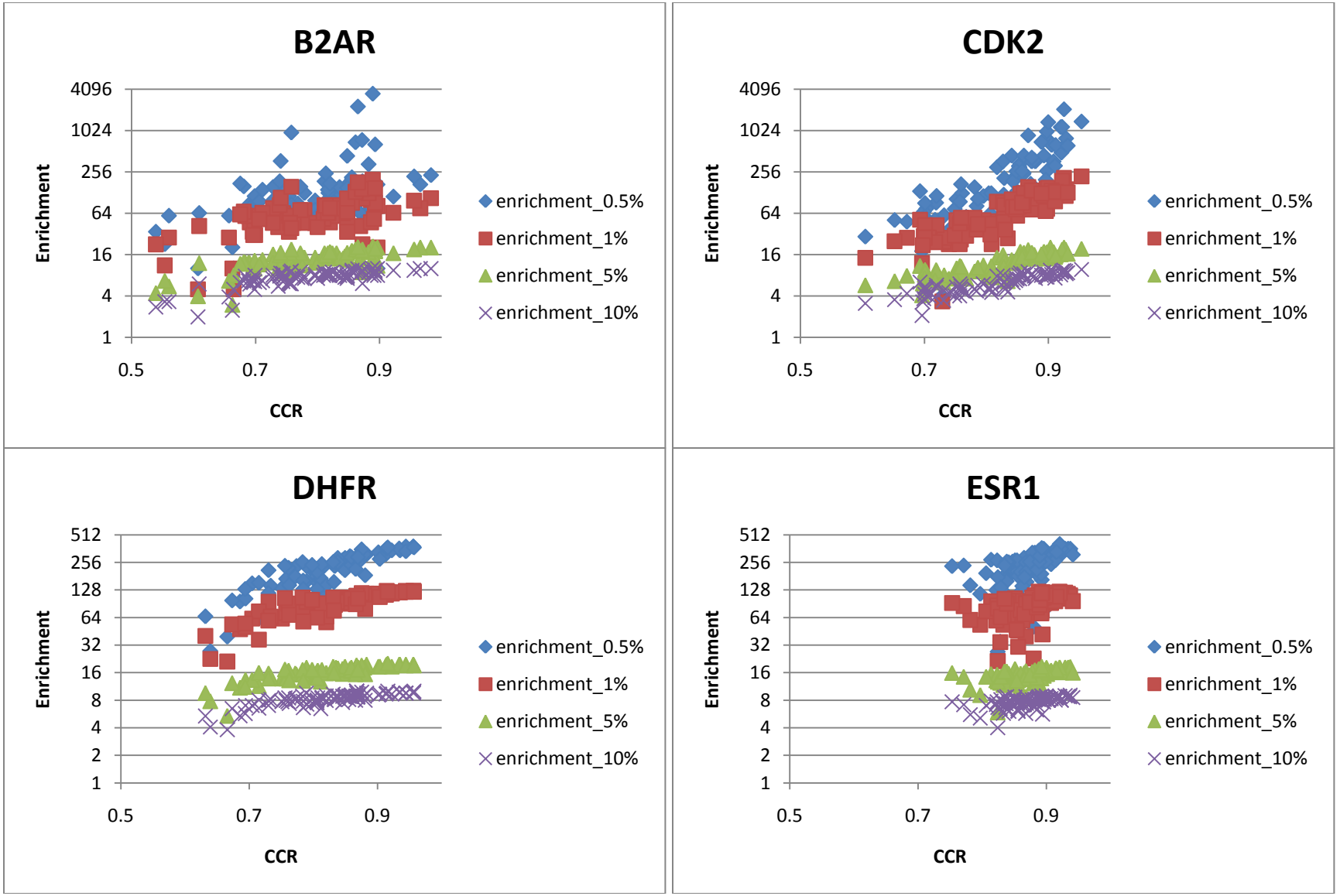
REN (Enrichment at 5%)
Enrichment
25
20
15
10
5
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

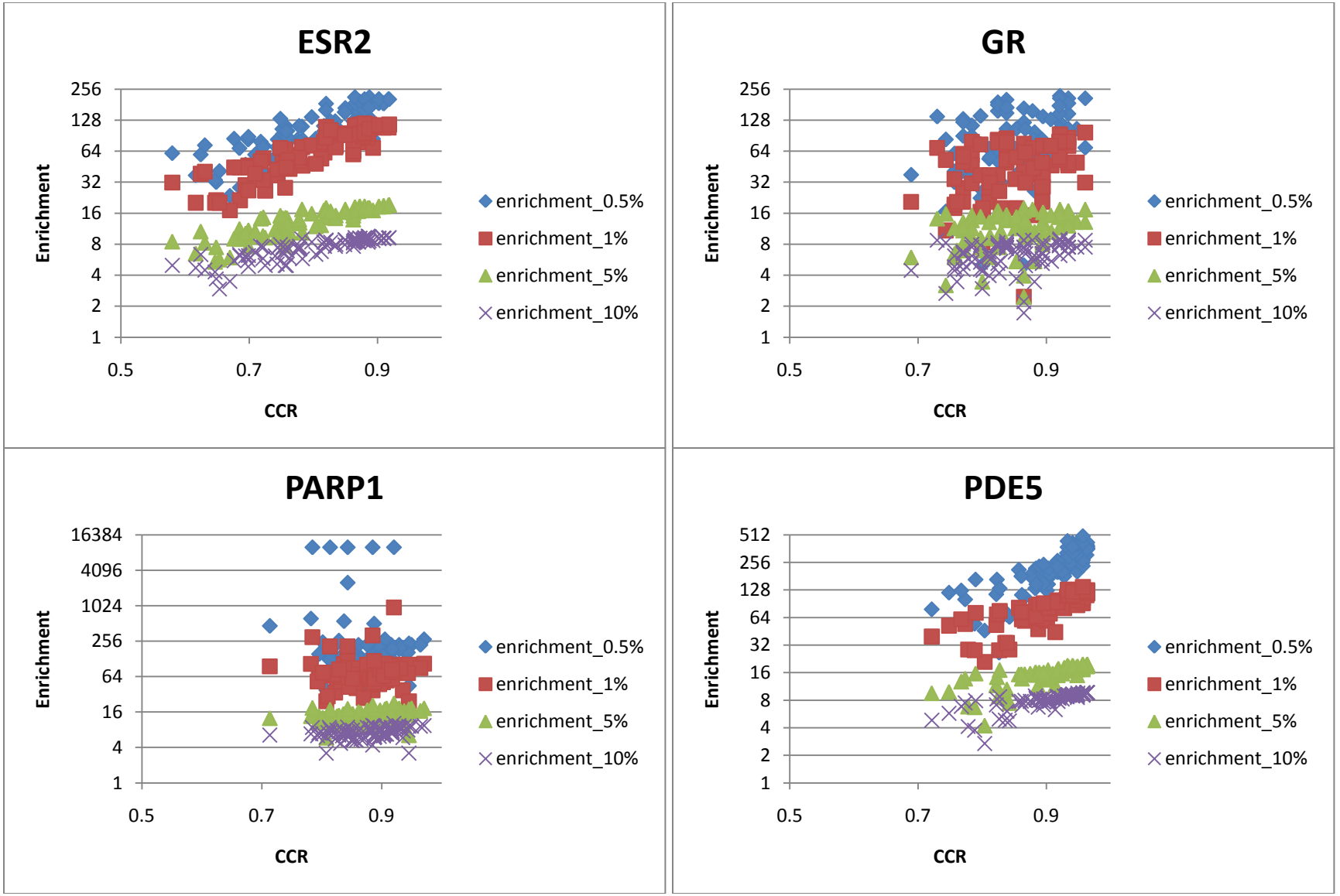
REN (Enrichment at 10%)
Enrichment
12
10
8
6
4
2
0
0 2 10 26 50 100 250 500 all
Set Size
Docking
Similarity
QSAR

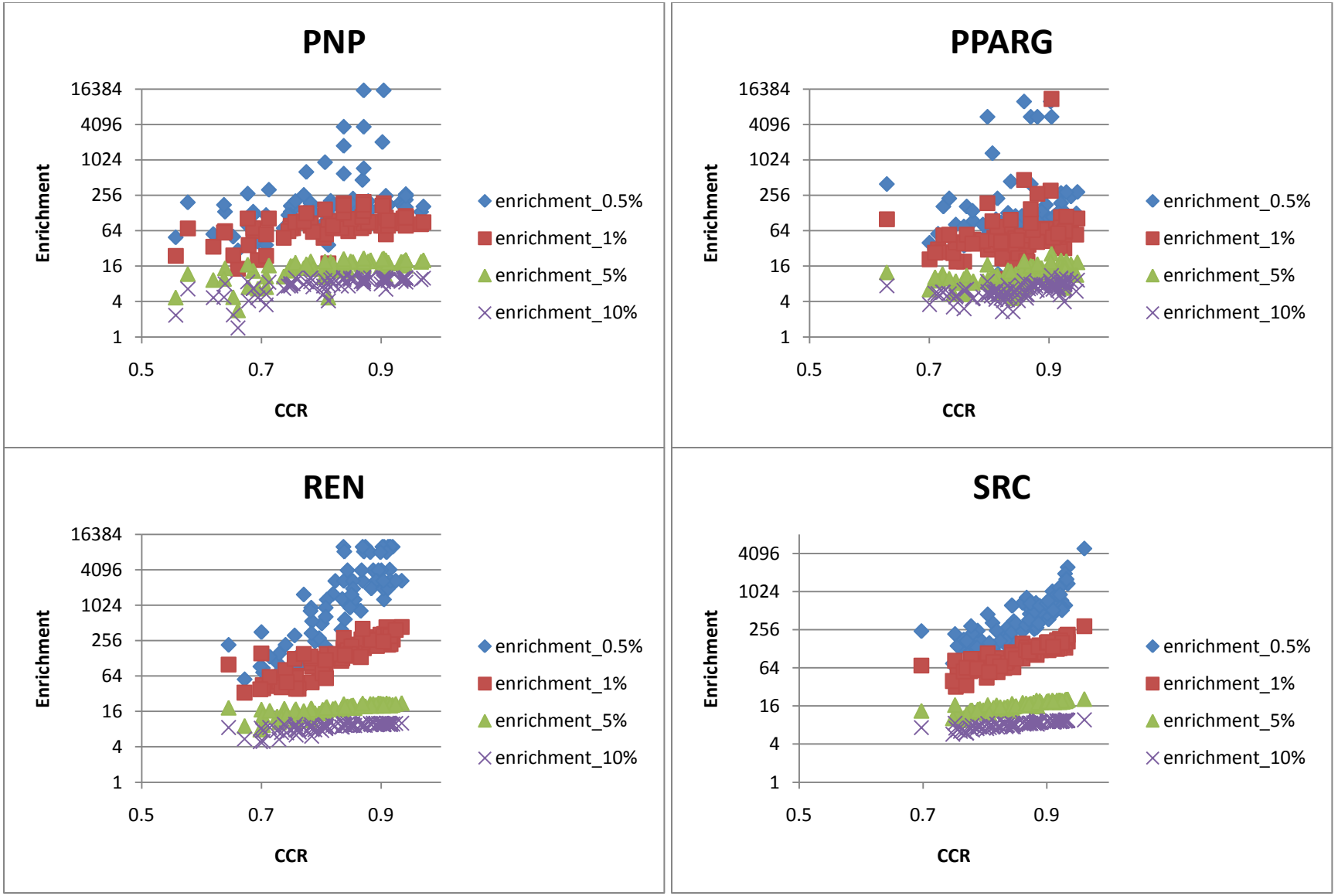
Appendix VIII: CCR vs. Enrichment Plots

Contained within this appendix are plots allow comparison of enrichment and CCR. Typically CCR is statistic used to determine the usefulness of a model; however, these figure seem to indicate that if the goal is to identify models that will provide superior enrichment in virtual screening applications, optimizing CCR may provide little benefit. Generally, enrichment correlates only weakly with CCR.




B2AR
Enrichment
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
CDK2
Enrichment
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
DHFR
Enrichment
512
256
128
64
32
16
8
4
2
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
ESR1
Enrichment
512
256
128
64
32
16
8
4
2
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%


ESR2
Enrichment
256
128
64
32
16
8
4
2
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
GR
Enrichment
256
128
64
32
16
8
4
2
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
PARP1
Enrichment
16384
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
PDE5
Enrichment
512
256
128
64
32
16
8
4
2
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%


PNP
Enrichment
16384
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
PPARG
Enrichment
16384
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
REN
Enrichment
16384
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%
SRC
Enrichment
4096
1024
256
64
16
4
1
0.5
0.7
0.9
CCR
enrichment_0.5%
enrichment_1%
enrichment_5%
enrichment_10%

Appendix IX: Gene Expression Markers for Multidrug Resistance

Contained within this appendix are the hypothetical multidrug biomarkers identified in Section 4.3.1. These biomarkers await experimental validation.

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
222608_s_at	B	anillin, actin binding protein	ANLN	NP_061155
222433_at	B	enabled homolog (Drosophila)	ENAH	NP_001008493 NP_060682
222449_at	B	prostate transmembrane protein, androgen induced 1	PMEPA1	NP_064567 NP_954638 NP_954639 NP_954640
222810_s_at	B	RAS protein activator like 2	RASAL2	NP_004832 NP_733793
222834_s_at	B	guanine nucleotide binding protein (G protein), gamma 12	GNG12	NP_061329
222692_s_at	B	fibronectin type III domain containing 3B	FNDC3B	NP_001128567 NP_073600
223019_at	B	family with sequence similarity 129, member B	FAM129B	NP_001030611 NP_073744
223279_s_at	B	uveal autoantigen with coiled-coil domains and ankyrin repeats	UACA	NP_001008225 NP_060473
223303_at	B	fermitin family homolog 3 (Drosophila)	FERMT3	NP_113659 NP_848537
223315_at	B	netrin 4	NTN4	NP_067052
223322_at	B	Ras association (RalGDS/AF-6) domain family member 5	RASSF5	NP_872604 NP_872605 NP_872606
223639_s_at	B	zinc ribbon domain containing 1	ZNRD1	NP_055411 NP_740753
223640_at	B	hematopoietic cell signal transducer	HCST	NP_001007470 NP_055081
224352_s_at	B	cofilin 2 (muscle)	CFL2	NP_068733 NP_619579

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
224450_s_at	B	RIO kinase 1 (yeast)	RIOK1	NP_113668 NP_694550
224407_s_at	B	serine/threonine protein kinase MST4	RP6-213H19.1	NP_001035917 NP_001035918 NP_057626
224791_at	B	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1	ASAP1	NP_060952
224583_at	B	coactosin-like 1 (Dictyostelium)	COTL1	NP_066972
224663_s_at	B	cofilin 2 (muscle)	CFL2	NP_068733 NP_619579
224895_at	B	Yes-associated protein 1, 65kDa	YAP1	NP_001123617 NP_006097
224911_s_at	B	discoidin, CUB and LCCL domain containing 2	DCBLD2	NP_563615
224917_at	B	microRNA 21	MIR21	---
224955_at	B	TEA domain family member 1 (SV40 transcriptional enhancer factor)	TEAD1	NP_068780
224983_at	B	scavenger receptor class B, member 2	SCARB2	NP_005497
224811_at	B	---	---	---
224840_at	B	FK506 binding protein 5	FKBP5	NP_001139247 NP_001139248 NP_001139249 NP_004108
224856_at	B	FK506 binding protein 5	FKBP5	NP_001139247 NP_001139248 NP_001139249 NP_004108
224996_at	B	aspartate beta-hydroxylase	ASPH	NP_004309 NP_064549 NP_115855 NP_115856 NP_115857
224999_at	B	---	---	---
225080_at	B	myosin IC	MYO1C	NP_001074248 NP_001074419 NP_203693
225091_at	B	zinc finger, CCHC domain containing 3	ZCCHC3	NP_149080

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
225272_at	B	spermidine/spermine N1-acetyltransferase family member 2	SAT2	NP_597998
225502_at	B	dedicator of cytokinesis 8	DOCK8	NP_982272
226425_at	B	CAP-GLY domain containing linker protein family, member 4	CLIP4	NP_078968
227344_at	B	IKAROS family zinc finger 1 (Ikaros)	IKZF1	NP_006051
227346_at	B	IKAROS family zinc finger 1 (Ikaros)	IKZF1	NP_006051
226934_at	B	cleavage and polyadenylation specific factor 6, 68kDa	CPSF6	NP_008938
225701_at	B	AT-hook transcription factor	AKNA	NP_110394
226659_at	B	differentially expressed in FDCP 6 homolog (mouse)	DEF6	NP_071330
226215_s_at	B	lysine (K)-specific demethylase 2B	KDM2B	NP_001005366 NP_115979
226219_at	B	Rho GTPase activating protein 30	ARHGAP30	NP_001020769 NP_859071
226680_at	B	IKAROS family zinc finger 5 (Pegasus)	IKZF5	NP_071911
225802_at	B	topoisomerase (DNA) I, mitochondrial	TOP1MT	NP_443195
225806_at	B	jub, ajuba homolog (Xenopus laevis)	JUB	NP_116265 NP_932352
226245_at	B	potassium channel tetramerisation domain containing 1	KCTD1	NP_001129677 NP_001136202 NP_945342
225842_at	B	pleckstrin homology-like domain, family A, member 1	PHLDA1	NP_031376
226282_at	B	---	---	---
227213_at	B	adenosine deaminase, tRNA-specific 2, TAD2 homolog (S. cerevisiae)	ADAT2	NP_872309
226366_at	B	SNF2 histone linker PHD RING helicase	SHPRH	NP_001036148 NP_775105
227272_at	B	chromosome 15 open reading frame 52	C15orf52	NP_997263
225962_at	B	zinc and ring finger 1	ZNRF1	NP_115644
227811_at	B	FYVE, RhoGEF and PH domain containing 3	FGD3	NP_001077005 NP_149077
228297_at	B	---	---	---

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
228824_s_at	B	prostaglandin reductase 1	PTGR1	NP_001139580 NP_001139581 NP_036344
227473_at	B	---	---	---
227484_at	B	SLIT-ROBO Rho GTPase activating protein 1	SRGAP1	NP_065813
227514_at	B	inositol 1,4,5-triphosphate receptor interacting protein-like 2	ITPRIPL2	NP_001030013
227998_at	B	S100 calcium binding protein A16	S100A16	NP_525127
228009_x_at	B	zinc ribbon domain containing 1	ZNRD1	NP_055411 NP_740753
228496_s_at	B	Cysteine rich transmembrane BMP regulator 1 (chordin-like)	CRIM1	NP_057525
227556_at	B	non-metastatic cells 7, protein expressed in (nucleoside-diphosphate kinase)	NME7	NP_037462 NP_932076
227628_at	B	glutathione peroxidase 8 (putative)	GPX8	NP_001008398
228121_at	B	transforming growth factor, beta 2	TGFB2	NP_001129071 NP_003229
227792_at	B	inositol 1,4,5-triphosphate receptor interacting protein-like 2	ITPRIPL2	NP_001030013
227799_at	B	myosin IG	MYO1G	NP_149043
229670_at	B	---	---	---
229686_at	B	purinergic receptor P2Y, G-protein coupled, 8	P2RY8	NP_835230
230175_s_at	B	---	---	---
230805_at	B	---	---	---
230836_at	B	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	ST8SIA4	NP_005659 NP_778222
229538_s_at	B	IQ motif containing GTPase activating protein 3	IQGAP3	NP_839943
232541_at	B	---	---	---
232843_s_at	B	dedicator of cytokinesis 8	DOCK8	NP_982272
231897_at	B	prostaglandin reductase 1	PTGR1	NP_001139580 NP_001139581 NP_036344
232994_s_at	B	Rho-guanine nucleotide exchange factor	RGNEF	NP_001073948

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
235020_at	B	TAF4b RNA polymerase II, TATA box binding protein (TBP)-associated factor, 105kDa	TAF4B	NP_005631
235072_s_at	B	---	---	---
234339_s_at	B	glioma tumor suppressor candidate region gene 2	GLTSCR2	NP_056525
233496_s_at	B	cofilin 2 (muscle)	CFL2	NP_068733 NP_619579
236565_s_at	B	La ribonucleoprotein domain family, member 6	LARP6	NP_060827 NP_932062
236198_at	B	---	---	---
239294_at	B	---	---	---
242520_s_at	B	chromosome 1 open reading frame 228	C1orf228	NP_001139108
242521_at	B	---	---	---
241879_at	B	---	---	---
244533_at	B	---	---	---
200601_at	A	actinin, alpha 4	ACTN4	NP_004915
200782_at	A	annexin A5	ANXA5	NP_001145
200787_s_at	A	phosphoprotein enriched in astrocytes 15	PEA15	NP_003759
200788_s_at	A	phosphoprotein enriched in astrocytes 15	PEA15	NP_003759
243601_at	B	hypothetical protein LOC285957	LOC285957	---
244654_at	B	myosin IG	MYO1G	NP_149043
200859_x_at	A	filamin A, alpha	FLNA	NP_001104026 NP_001447
200872_at	A	S100 calcium binding protein A10	S100A10	NP_002957
201681_s_at	A	discs, large homolog 5 (Drosophila)	DLG5	NP_004738
202133_at	A	WW domain containing transcription regulator 1	WWTR1	NP_056287
201021_s_at	A	destrin (actin depolymerizing factor)	DSTN	NP_001011546 NP_006861
201022_s_at	A	destrin (actin depolymerizing factor)	DSTN	NP_001011546 NP_006861
201289_at	A	cysteine-rich, angiogenic inducer, 61	CYR61	NP_001545
202431_s_at	A	v-myc myelocytomatosis viral oncogene homolog (avian)	MYC	NP_002458

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
200885_at	A	ras homolog gene family, member C	RHOC	NP_001036143 NP_001036144 NP_786886
202458_at	A	protease, serine, 23	PRSS23	NP_009104
202052_s_at	A	retinoic acid induced 14	RAI14	NP_001138992 NP_001138993 NP_001138994 NP_001138995 NP_001138997 NP_056392
202470_s_at	A	cleavage and polyadenylation specific factor 6, 68kDa	CPSF6	NP_008938
201215_at	A	plastin 3 (T isoform)	PLS3	NP_001129497 NP_005023
201220_x_at	A	C-terminal binding protein 2	CTBP2	NP_001077383 NP_001320 NP_073713
201445_at	A	calponin 3, acidic	CNN3	NP_001830
202071_at	A	syndecan 4	SDC4	NP_002990
201462_at	A	secernin 1	SCRN1	NP_001138985 NP_001138986 NP_001138987 NP_055581
201467_s_at	A	NAD(P)H dehydrogenase, quinone 1	NQO1	NP_000894 NP_001020604 NP_001020605
201468_s_at	A	NAD(P)H dehydrogenase, quinone 1	NQO1	NP_000894 NP_001020604 NP_001020605
201471_s_at	A	sequestosome 1	SQSTM1	NP_001135770 NP_001135771 NP_003891
201059_at	A	cortactin	CTTN	NP_005222 NP_612632
200636_s_at	A	protein tyrosine phosphatase, receptor type, F	PTPRF	NP_002831 NP_569707
200660_at	A	S100 calcium binding protein A11	S100A11	NP_005611
201073_s_at	A	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1	SMARCC1	NP_003065

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
201087_at	A	paxillin	PXN	NP_001074324 NP_002850 NP_079433
201505_at	A	laminin, beta 1	LAMB1	NP_002282
201939_at	A	polo-like kinase 2 (Drosophila)	PLK2	NP_006613
200698_at	A	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 2	KDELR2	NP_001094073 NP_006845
201969_at	A	nuclear autoantigenic sperm protein (histone-binding)	NASP	NP_002473 NP_689511 NP_751896
200663_at	A	CD63 molecule	CD63	NP_001035123 NP_001771
200673_at	A	lysosomal protein transmembrane 4 alpha	LAPTM4A	NP_055528
201125_s_at	A	integrin, beta 5	ITGB5	NP_002204
201585_s_at	A	splicing factor proline/glutamine-rich (polypyrimidine tract binding protein associated)	SFPQ	NP_005057
201976_s_at	A	myosin X	MYO10	NP_036466
201983_s_at	A	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	EGFR	NP_005219 NP_958439 NP_958440 NP_958441
201984_s_at	A	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	EGFR	NP_005219 NP_958439 NP_958440 NP_958441
201995_at	A	exostoses (multiple) 1	EXT1	NP_000118
201590_x_at	A	annexin A2	ANXA2	NP_001002857 NP_001002858 NP_001129487 NP_004030
202011_at	A	tight junction protein 1 (zona occludens 1)	TJP1	NP_003248 NP_783297
201172_x_at	A	ATPase, H+ transporting, lysosomal 9kDa, V0 subunit e1	ATP6V0E1	NP_003936
201360_at	A	cystatin C	CST3	NP_000090
201798_s_at	A	myoferlin	MYOF	NP_038479 NP_579899
200770_s_at	A	laminin, gamma 1 (formerly LAMB2)	LAMC1	NP_002284

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
200771_at	A	laminin, gamma 1 (formerly LAMB2)	LAMC1	NP_002284
200931_s_at	A	vinculin	VCL	NP_003364 NP_054706
201373_at	A	plectin 1, intermediate filament binding protein 500kDa	PLEC1	NP_000436 NP_958780 NP_958781 NP_958782 NP_958783 NP_958784 NP_958785 NP_958786
202237_at	A	nicotinamide N-methyltransferase	NNMT	NP_006160
202238_s_at	A	nicotinamide N-methyltransferase	NNMT	NP_006160
202252_at	A	RAB13, member RAS oncogene family	RAB13	NP_002861
200998_s_at	A	cytoskeleton-associated protein 4	CKAP4	NP_006816
200999_s_at	A	cytoskeleton-associated protein 4	CKAP4	NP_006816
201242_s_at	A	ATPase, Na+/K+ transporting, beta 1 polypeptide	ATP1B1	NP_001001787 NP_001668
201243_s_at	A	ATPase, Na+/K+ transporting, beta 1 polypeptide	ATP1B1	NP_001001787 NP_001668
201251_at	A	pyruvate kinase, muscle	PKM2	NP_002645 NP_872270 NP_872271
202551_s_at	A	cysteine rich transmembrane BMP regulator 1 (chordin-like)	CRIM1	NP_057525
202552_s_at	A	cysteine rich transmembrane BMP regulator 1 (chordin-like)	CRIM1	NP_057525
203705_s_at	A	frizzled homolog 7 (Drosophila)	FZD7	NP_003498
203706_s_at	A	frizzled homolog 7 (Drosophila)	FZD7	NP_003498
204992_s_at	A	profilin 2	PFN2	NP_002619 NP_444252
205417_s_at	A	dystroglycan 1 (dystrophin-associated glycoprotein 1)	DAG1	NP_004384
203323_at	A	caveolin 2	CAV2	NP_001224 NP_937855
203324_s_at	A	caveolin 2	CAV2	NP_001224 NP_937855

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
204116_at	A	interleukin 2 receptor, gamma (severe combined immunodeficiency)	IL2RG	NP_000197
204560_at	A	FK506 binding protein 5	FKBP5	NP_001139247 NP_001139248 NP_001139249 NP_004108
202733_at	A	prolyl 4-hydroxylase, alpha polypeptide II	P4HA2	NP_001017973 NP_001017974 NP_001136070 NP_001136071 NP_004190
202756_s_at	A	glypican 1	GPC1	NP_002072
202822_at	A	LIM domain containing preferred translocation partner in lipoma	LPP	NP_005569
203262_s_at	A	family with sequence similarity 50, member A	FAM50A	NP_004690
204489_s_at	A	CD44 molecule (Indian blood group)	CD44	NP_000601 NP_001001389 NP_001001390 NP_001001391 NP_001001392
202377_at	A	---	---	---
202381_at	A	ADAM metallopeptidase domain 9 (meltrin gamma)	ADAM9	NP_001005845 NP_003807
203411_s_at	A	lamin A/C	LMNA	NP_005563 NP_733821 NP_733822
203416_at	A	CD53 molecule	CD53	NP_000551 NP_001035122
204490_s_at	A	CD44 molecule (Indian blood group)	CD44	NP_000601 NP_001001389 NP_001001390 NP_001001391 NP_001001392
203002_at	A	angiomotin like 2	AMOTL2	NP_057285
202587_s_at	A	adenylate kinase 1	AK1	NP_000467
203038_at	A	protein tyrosine phosphatase, receptor type, K	PTPRK	NP_001129120 NP_002835
204066_s_at	A	ArfGAP with GTPase domain, ankyrin repeat and PH domain 1	AGAP1	NP_001032208 NP_055729

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
203065_s_at	A	caveolin 1, caveolae protein, 22kDa	CAV1	NP_001744
203499_at	A	EPH receptor A2	EPHA2	NP_004422
203510_at	A	met proto-oncogene (hepatocyte growth factor receptor)	MET	NP_000236 NP_001120972
204513_s_at	A	engulfment and cell motility 1	ELMO1	NP_001034548 NP_055615 NP_569709
202598_at	A	S100 calcium binding protein A13	S100A13	NP_001019381 NP_001019382 NP_001019383 NP_001019384 NP_005970
202609_at	A	epidermal growth factor receptor pathway substrate 8	EPS8	NP_004438
204517_at	A	peptidylprolyl isomerase C (cyclophilin C)	PPIC	NP_000934
204951_at	A	ras homolog gene family, member H	RHOH	NP_004301
204960_at	A	protein tyrosine phosphatase, receptor type, C-associated protein	PTPRCAP	NP_005599
202949_s_at	A	four and a half LIM domains 2	FHL2	NP_001034581 NP_001441 NP_963849 NP_963851
202957_at	A	hematopoietic cell-specific Lyn substrate 1	HCLS1	NP_005326
204657_s_at	A	Src homology 2 domain containing adaptor protein B	SHB	NP_003019
204411_at	A	kinesin family member 21B	KIF21B	NP_060066
204425_at	A	Rho GTPase activating protein 4	ARHGAP4	NP_001657
206752_s_at	A	DNA fragmentation factor, 40kDa, beta polypeptide (caspase-activated DNase)	DFFB	NP_004393
204852_s_at	A	protein tyrosine phosphatase, non-receptor type 7	PTPN7	NP_002823 NP_542155
205213_at	A	ArfGAP with coiled-coil, ankyrin repeat and PH domains 1	ACAP1	NP_055531
204220_at	A	glia maturation factor, gamma	GMFG	NP_004868

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
204237_at	A	GULP, engulfment adaptor PTB domain containing 1	GULP1	NP_057399
204341_at	A	tripartite motif-containing 16	TRIM16	NP_006461
204248_at	A	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	GNA11	NP_002058
203965_at	A	ubiquitin specific peptidase 20	USP20	NP_001008563 NP_001103773 NP_006667
205038_at	A	IKAROS family zinc finger 1 (Ikaros)	IKZF1	NP_006051
204688_at	A	sarcoglycan, epsilon	SGCE	NP_001092870 NP_001092871 NP_003910
204798_at	A	v-myb myeloblastosis viral oncogene homolog (avian)	MYB	NP_001123644 NP_001123645 NP_001155128 NP_001155129 NP_001155130 NP_001155131 NP_001155132 NP_005366
204152_s_at	A	MFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase	MFNG	NP_002396
204153_s_at	A	MFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase	MFNG	NP_002396
203760_s_at	A	Src-like-adaptor	SLA	NP_001039021 NP_001039022 NP_006739
203761_at	A	Src-like-adaptor	SLA	NP_001039021 NP_001039022 NP_006739
205266_at	A	leukemia inhibitory factor (cholinergic differentiation factor)	LIF	NP_002300
205269_at	A	lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)	LCP2	NP_005556
205270_s_at	A	lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)	LCP2	NP_005556
203910_at	A	Rho GTPase activating protein 29	ARHGAP29	NP_004806

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
204306_s_at	A	CD151 molecule (Raph blood group)	CD151	NP_001034579 NP_004348 NP_620598 NP_620599
206414_s_at	A	ArfGAP with SH3 domain, ankyrin repeat and PH domain 2	ASAP2	NP_001128663 NP_003878
206660_at	A	immunoglobulin lambda-like polypeptide 1	IGLL1	NP_064455 NP_690594
208862_s_at	A	catenin (cadherin-associated protein), delta 1	CTNND1	NP_001078927 NP_001078928 NP_001078929 NP_001078930 NP_001078931 NP_001078932 NP_001078933 NP_001078934 NP_001078935 NP_001078936 NP_001078937 NP_001078938 NP_001322
205739_x_at	A	zinc finger protein 107	ZNF107	NP_001013768 NP_057304
207238_s_at	A	protein tyrosine phosphatase, receptor type, C	PTPRC	NP_002829 NP_563578 NP_563579 NP_563580
205884_at	A	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)	ITGA4	NP_000876
205885_s_at	A	integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)	ITGA4	NP_000876
205573_s_at	A	sorting nexin 7	SNX7	NP_057060 NP_689424
206116_s_at	A	tropomyosin 1 (alpha)	TPM1	NP_000357 NP_001018004 NP_001018005 NP_001018006 NP_001018007 NP_001018008 NP_001018020
208816_x_at	A	annexin A2 pseudogene 2	ANXA2P2	---

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
208820_at	A	PTK2 protein tyrosine kinase 2	PTK2	NP_005598 NP_722560
206039_at	A	RAB33A, member RAS oncogene family	RAB33A	NP_004785
209263_x_at	A	tetraspanin 4	TSPAN4	NP_001020405 NP_001020406 NP_001020407 NP_001020408 NP_001020409 NP_001020410 NP_003262
209264_s_at	A	tetraspanin 4	TSPAN4	NP_001020405 NP_001020406 NP_001020407 NP_001020408 NP_001020409 NP_001020410 NP_003262
207522_s_at	A	ATPase, Ca++ transporting, ubiquitous	ATP2A3	NP_005164 NP_777613 NP_777614 NP_777615 NP_777616 NP_777617 NP_777618
209734_at	A	NCK-associated protein 1-like	NCKAP1L	NP_005328
208540_x_at	A	S100 calcium binding protein A11	S100A11	NP_005611
209289_at	A	nuclear factor I/B	NFIB	NP_005587
209290_s_at	A	nuclear factor I/B	NFIB	NP_005587
208770_s_at	A	eukaryotic translation initiation factor 4E binding protein 2	EIF4EBP2	NP_004087
207525_s_at	A	GIPC PDZ domain containing family, member 1	GIPC1	NP_005707 NP_974196 NP_974197 NP_974198 NP_974199 NP_974223
208056_s_at	A	core-binding factor, runt domain, alpha subunit 2; translocated to, 3	CBFA2T3	NP_005178 NP_787127
207738_s_at	A	NCK-associated protein 1	NCKAP1	NP_038464 NP_995314

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
208456_s_at	A	related RAS viral (r-ras) oncogene homolog 2	RRAS2	NP_001096139 NP_036382
208683_at	A	calpain 2, (m/II) large subunit	CAPN2	NP_001139540 NP_001739
207957_s_at	A	protein kinase C, beta	PRKCB	NP_002729 NP_997700
208885_at	A	lymphocyte cytosolic protein 1 (L-plastin)	LCP1	NP_002289
208898_at	A	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D	ATP6V1D	NP_057078
207467_x_at	A	calpastatin	CAST	NP_001035905 NP_001035906 NP_001035907 NP_001035908 NP_001035909 NP_001035910 NP_001035911 NP_001741 NP_775083 NP_775084 NP_775086
208908_s_at	A	calpastatin	CAST	NP_001035905 NP_001035906 NP_001035907 NP_001035908 NP_001035909 NP_001035910 NP_001035911 NP_001741 NP_775083 NP_775084 NP_775086
209684_at	A	Ras and Rab interactor 2	RIN2	NP_061866
209685_s_at	A	protein kinase C, beta	PRKCB	NP_002729 NP_997700
208711_s_at	A	cyclin D1	CCND1	NP_444284
208712_at	A	cyclin D1	CCND1	NP_444284
208613_s_at	A	filamin B, beta	FLNB	NP_001157789 NP_001157790 NP_001157791 NP_001448
208206_s_at	A	RAS guanyl releasing protein 2 (calcium and DAG-regulated)	RASGRP2	NP_001092140 NP_001092141 NP_722541

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
208636_at	A	actinin, alpha 1	ACTN1	NP_001093 NP_001123476 NP_001123477
208637_x_at	A	actinin, alpha 1	ACTN1	NP_001093 NP_001123476 NP_001123477
212185_x_at	A	metallothionein 2A	MT2A	NP_005944
209386_at	A	transmembrane 4 L six family member 1	TM4SF1	NP_055035
209834_at	A	carbohydrate (chondroitin 6) sulfotransferase 3	CHST3	NP_004264
209835_x_at	A	CD44 molecule (Indian blood group)	CD44	NP_000601 NP_001001389 NP_001001390 NP_001001391 NP_001001392
209154_at	A	Tax1 (human T-cell leukemia virus type I) binding protein 3	TAX1BP3	NP_055419
208949_s_at	A	lectin, galactoside-binding, soluble, 3	LGALS3	NP_002297
208951_at	A	aldehyde dehydrogenase 7 family, member A1	ALDH7A1	NP_001173
209488_s_at	A	RNA binding protein with multiple splicing	RBPMS	NP_001008710 NP_001008711 NP_001008712 NP_006858
212195_at	A	interleukin 6 signal transducer (gp130, oncostatin M receptor)	IL6ST	NP_002175 NP_786943
209083_at	A	coronin, actin binding protein, 1A	CORO1A	NP_009005
209879_at	A	selectin P ligand	SELPLG	NP_002997
210519_s_at	A	NAD(P)H dehydrogenase, quinone 1	NQO1	NP_000894 NP_001020604 NP_001020605
209108_at	A	tetraspanin 6	TSPAN6	NP_003261
209213_at	A	carbonyl reductase 1	CBR1	NP_001748
209432_s_at	A	cAMP responsive element binding protein 3	CREB3	NP_006359
210427_x_at	A	annexin A2	ANXA2	NP_001002857 NP_001002858 NP_001129487 NP_004030

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
209135_at	A	aspartate beta-hydroxylase	ASPH	NP_004309 NP_064549 NP_115855 NP_115856 NP_115857
212169_at	A	FK506 binding protein 9, 63 kDa	FKBP9	NP_009201
210038_at	A	protein kinase C, theta	PRKCQ	NP_006248
210039_s_at	A	protein kinase C, theta	PRKCQ	NP_006248
212061_at	A	U2-associated SR140 protein	SR140	NP_001073884
211160_x_at	A	actinin, alpha 1	ACTN1	NP_001093 NP_001123476 NP_001123477
210876_at	A	annexin A2 pseudogene 1	ANXA2P1	---
213539_at	A	CD3d molecule, delta (CD3-TCR complex)	CD3D	NP_000723 NP_001035741
211986_at	A	AHNAK nucleoprotein	AHNAK	NP_001611 NP_076965
212086_x_at	A	lamin A/C	LMNA	NP_005563 NP_733821 NP_733822
212089_at	A	lamin A/C	LMNA	NP_005563 NP_733821 NP_733822
212097_at	A	caveolin 1, caveolae protein, 22kDa	CAV1	NP_001744
212104_s_at	A	RNA binding motif protein 9	RBM9	NP_001026865 NP_001076045 NP_001076046 NP_001076047 NP_001076048 NP_055124
210986_s_at	A	tropomyosin 1 (alpha)	TPM1	NP_000357 NP_001018004 NP_001018005 NP_001018006 NP_001018007 NP_001018008 NP_001018020

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
210987_x_at	A	tropomyosin 1 (alpha)	TPM1	NP_000357 NP_001018004 NP_001018005 NP_001018006 NP_001018007 NP_001018008 NP_001018020
210896_s_at	A	aspartate beta-hydroxylase	ASPH	NP_004309 NP_064549 NP_115855 NP_115856 NP_115857
212014_x_at	A	CD44 molecule (Indian blood group)	CD44	NP_000601 NP_001001389 NP_001001390 NP_001001391 NP_001001392
210835_s_at	A	C-terminal binding protein 2	CTBP2	NP_001077383 NP_001320 NP_073713
213036_x_at	A	ATPase, Ca++ transporting, ubiquitous	ATP2A3	NP_005164 NP_777613 NP_777614 NP_777615 NP_777616 NP_777617 NP_777618
211919_s_at	A	chemokine (C-X-C motif) receptor 4	CXCR4	NP_001008540 NP_003458
211938_at	A	eukaryotic translation initiation factor 4B	EIF4B	NP_001408
213944_x_at	A	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	GNA11	NP_002058
212587_s_at	A	protein tyrosine phosphatase, receptor type, C	PTPRC	NP_002829 NP_563578 NP_563579 NP_563580
212588_at	A	protein tyrosine phosphatase, receptor type, C	PTPRC	NP_002829 NP_563578 NP_563579 NP_563580
212589_at	A	related RAS viral (r-ras) oncogene homolog 2	RRAS2	NP_001096139 NP_036382

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
212590_at	A	related RAS viral (r-ras) oncogene homolog 2	RRAS2	NP_001096139 NP_036382
210644_s_at	A	leukocyte-associated immunoglobulin-like receptor 1	LAIR1	NP_002278 NP_068352
211240_x_at	A	catenin (cadherin-associated protein), delta 1	CTNND1	NP_001078927 NP_001078928 NP_001078929 NP_001078930 NP_001078931 NP_001078932 NP_001078933 NP_001078934 NP_001078935 NP_001078936 NP_001078937 NP_001078938 NP_001322
211651_s_at	A	laminin, beta 1	LAMB1	NP_002282
213503_x_at	A	annexin A2	ANXA2	NP_001002857 NP_001002858 NP_001129487 NP_004030
211864_s_at	A	myoferlin	MYOF	NP_038479 NP_579899
211945_s_at	A	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	ITGB1	NP_002202 NP_389647 NP_391987 NP_391988 NP_391989 NP_596867
212294_at	A	guanine nucleotide binding protein (G protein), gamma 12	GNG12	NP_061329
212724_at	A	Rho family GTPase 3	RND3	NP_005159
213746_s_at	A	filamin A, alpha	FLNA	NP_001104026 NP_001447
212285_s_at	A	agrin	AGRN	NP_940978
212413_at	A	septin 6	6-Sep	NP_055944 NP_665798 NP_665799 NP_665801
212738_at	A	Rho GTPase activating protein 19	ARHGAP19	NP_116289

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
212973_at	A	ribose 5-phosphate isomerase A	RPIA	NP_653164
213666_at	A	septin 6	6-Sep	NP_055944 NP_665798 NP_665799 NP_665801
213766_x_at	A	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	GNA11	NP_002058
212415_at	A	septin 6	6-Sep	NP_055944 NP_665798 NP_665799 NP_665801
212420_at	A	E74-like factor 1 (ets domain transcription factor)	ELF1	NP_001138825 NP_758961
213888_s_at	A	TRAF3 interacting protein 3	TRAF3IP3	NP_079504
212658_at	A	lipoma HMGIC fusion partner-like 2	LHFPL2	NP_005770
212662_at	A	poliovirus receptor	PVR	NP_001129240 NP_001129241 NP_001129242 NP_006496
212765_at	A	calmodulin regulated spectrin-associated protein 1-like 1	CAMSAP1L1	NP_982284
212873_at	A	histocompatibility (minor) HA-1	HMHA1	NP_036424
212885_at	A	M-phase phosphoprotein 10 (U3 small nucleolar ribonucleoprotein)	MPHOSPH10	NP_005782
212992_at	A	AHNAK nucleoprotein 2	AHNAK2	NP_612429
213358_at	A	KIAA0802	KIAA0802	NP_056025
213455_at	A	family with sequence similarity 114, member A1	FAM114A1	NP_612398
213901_x_at	A	RNA binding motif protein 9	RBM9	NP_001026865 NP_001076045 NP_001076046 NP_001076047 NP_001076048 NP_055124
213915_at	A	natural killer cell group 7 sequence	NKG7	NP_005592
213160_at	A	dedicator of cytokinesis 2	DOCK2	NP_004937
213029_at	A	nuclear factor I/B	NFIB	NP_005587

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
214752_x_at	A	filamin A, alpha	FLNA	NP_001104026 NP_001447
212698_s_at	A	septin 10	10-Sep	NP_653311 NP_848699
212364_at	A	myosin IB	MYO1B	NP_001123630 NP_001155291 NP_036355
212254_s_at	A	dystonin	DST	NP_001138241 NP_001138242 NP_001138243 NP_001714 NP_056363 NP_065121 NP_899236
212919_at	A	DCP2 decapping enzyme homolog (S. cerevisiae)	DCP2	NP_689837
212825_at	A	PAX interacting (with transcription-activation domain) protein 1	PAXIP1	NP_031375
217028_at	A	chemokine (C-X-C motif) receptor 4	CXCR4	NP_001008540 NP_003458
215464_s_at	A	Tax1 (human T-cell leukemia virus type I) binding protein 3	TAX1BP3	NP_055419
215016_x_at	A	dystonin	DST	NP_001138241 NP_001138242 NP_001138243 NP_001714 NP_056363 NP_065121 NP_899236
214039_s_at	A	lysosomal protein transmembrane 4 beta	LAPTM4B	NP_060877
215091_s_at	A	general transcription factor IIIA	GTF3A	NP_002088
217419_x_at	A	agrin	AGRN	NP_940978
214882_s_at	A	splicing factor, arginine/serine-rich 2	SFRS2	NP_003007
214679_x_at	A	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	GNA11	NP_002058
217892_s_at	A	LIM domain and actin binding 1	LIMA1	NP_001107018 NP_001107019 NP_057441
216264_s_at	A	laminin, beta 2 (laminin S)	LAMB2	NP_002283

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
218733_at	A	male-specific lethal 2 homolog (Drosophila)	MSL2	NP_001138889 NP_060603
218738_s_at	A	ring finger protein 138	RNF138	NP_057355 NP_937761
216215_s_at	A	RNA binding motif protein 9	RBM9	NP_001026865 NP_001076045 NP_001076046 NP_001076047 NP_001076048 NP_055124
219191_s_at	A	bridging integrator 2	BIN2	NP_057377
216226_at	A	TAF4b RNA polymerase II, TATA box binding protein (TBP)-associated factor, 105kDa	TAF4B	NP_005631
217849_s_at	A	CDC42 binding protein kinase beta (DMPK-like)	CDC42BPB	NP_006026
218870_at	A	Rho GTPase activating protein 15	ARHGAP15	NP_060930
218418_s_at	A	KN motif and ankyrin repeat domains 2	KANK2	NP_001129663 NP_056308
218656_s_at	A	lipoma HMGIC fusion partner	LHFP	NP_005771
221059_s_at	A	coactosin-like 1 (Dictyostelium)	COTL1	NP_066972
217996_at	A	pleckstrin homology-like domain, family A, member 1	PHLDA1	NP_031376
218793_s_at	A	sex comb on midleg-like 1 (Drosophila)	SCML1	NP_001032624 NP_001032625 NP_001032629 NP_006737
218028_at	A	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 1	ELOVL1	NP_073732
218368_s_at	A	tumor necrosis factor receptor superfamily, member 12A	TNFRSF12A	NP_057723
218581_at	A	abhydrolase domain containing 4	ABHD4	NP_071343
218168_s_at	A	chaperone, ABC1 activity of bc1 complex homolog (S. pombe)	CABC1	NP_064632
220704_at	A	IKAROS family zinc finger 1 (Ikaros)	IKZF1	NP_006051

Probe Name	Gene Array	Gene Name	Gene Symbol	RefSeq Protein ID
220865_s_at	A	prenyl (decaprenyl) diphosphate synthase, subunit 1	PDSS1	NP_055132
219944_at	A	CAP-GLY domain containing linker protein family, member 4	CLIP4	NP_078968
221007_s_at	A	FIP1 like 1 (S. cerevisiae)	FIP1L1	NP_001128409 NP_001128410 NP_112179
219862_s_at	A	nuclear prelamin A recognition factor	NARF	NP_001033707 NP_001077077 NP_036468 NP_114174
220330_s_at	A	SAM domain, SH3 domain and nuclear localization signals 1	SAMSN1	NP_071419
221676_s_at	A	coronin, actin binding protein, 1C	CORO1C	NP_055140
35974_at	A	lymphoid-restricted membrane protein	LRMP	NP_006143
40562_at	A	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	GNA11	NP_002058
221293_s_at	A	differentially expressed in FDCP 6 homolog (mouse)	DEF6	NP_071330
222258_s_at	A	SH3-domain binding protein 4	SH3BP4	NP_055336
221606_s_at	A	nucleosomal binding protein 1	NSBP1	NP_110390
222154_s_at	A	spermatogenesis associated, serine-rich 2-like	SPATS2L	NP_001093892 NP_001093893 NP_001093894 NP_056350
564_at	A	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	GNA11	NP_002058
57163_at	A	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-like 1	ELOVL1	NP_073732

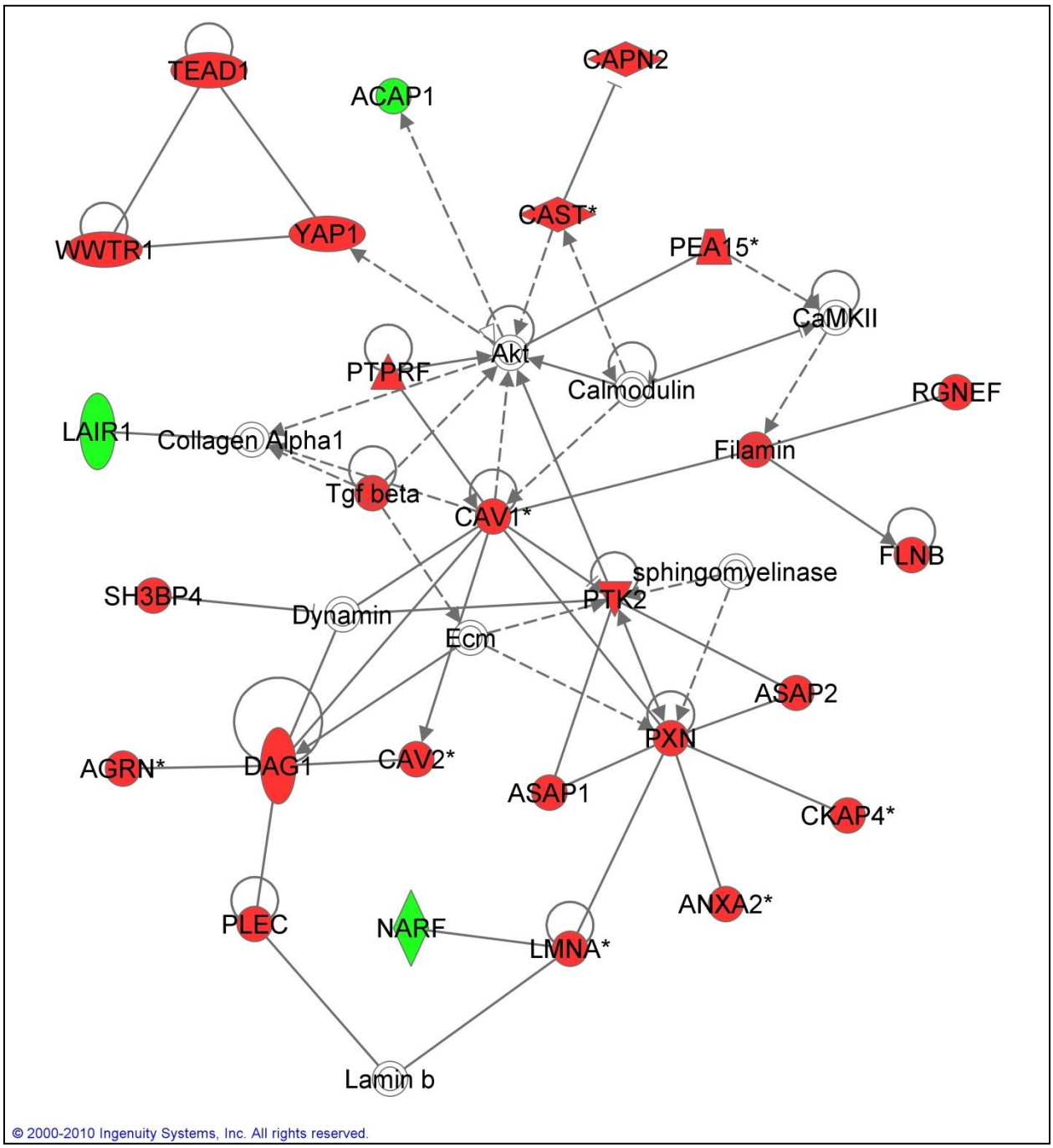
Appendix X: Networks of Gene Expression Markers

Listed below are the networks identified by ingenuity pathway analysis of hypothetical biomarkers of multidrug resistance. For each network, a figure elucidating the connectivity of the network's proteins is provided.

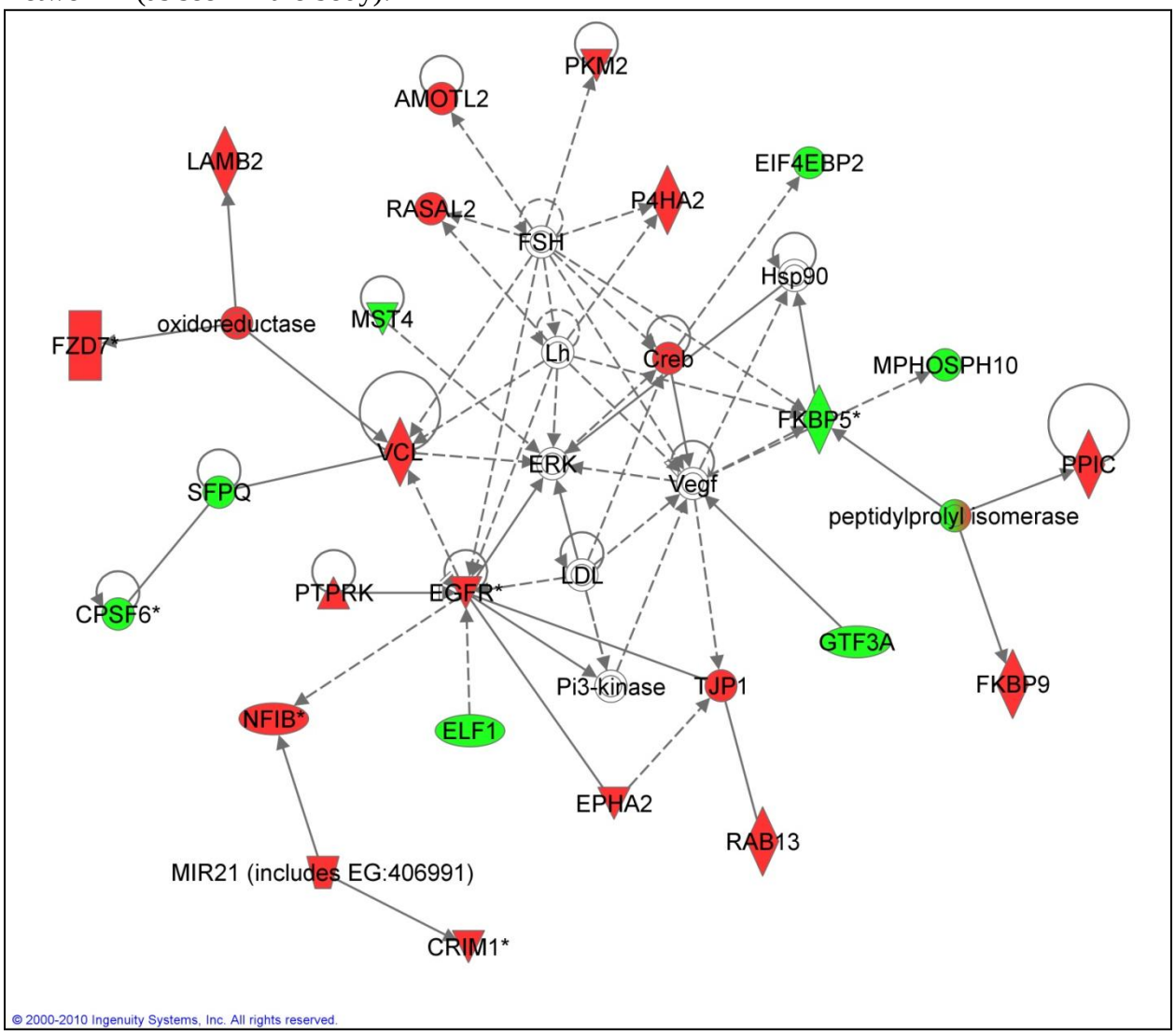
ID	Molecules in Network	Score	Focus Molecules	Top Functions
1	ACAP1, AGRN, Akt, ANXA2, ASAP1, ASAP2, Calmodulin, CaMKII, CAPN2, CAST, CAV1, CAV2, CKAP4, Collagen Alpha1, DAG1, Dynamin, Ecm, Filamin, FLNB, LAIR1, Lamin b, LMNA, NARF, PEA15, PLEC, PTK2, PTPRF, PXN, RGNEF, SH3BP4, sphingomyelinase, TEAD1, Tgf beta, WWTR1, YAP1	43	25	Cellular Assembly and Organization, Cellular Function and Maintenance, Cellular Movement
2	AMOTL2, CPSF6, Creb, CRIM1, EGFR, EIF4EBP2, ELF1, EPHA2, ERK, FKBP5, FKBP9, FSH, FZD7, GTF3A, Hsp90, LAMB2, LDL, Lh, MIR21 (includes EG:406991), MPHOSPH10, MST4, NFIB, oxidoreductase, P4HA2, peptidylprolyl isomerase, Pi3-kinase, PKM2, PPIC, PTPRK, RAB13, RASAL2, SFPQ, TJP1, VCL, Vegf	43	25	Cell-To-Cell Signaling and Interaction, Cellular Assembly and Organization, Nervous System Development and Function
3	Actin, Actin-Actn-Ptk2-Pxn-Vcl, ACTN1, ACTN4, Alpha actin, Alpha Actinin, Alpha catenin, Arp2/3, Bcl9-Cbp/p300-Ctnnb1-Lef/Tcf, CFL2, Cofilin, DFFB, DSTN, ENAH, EPS8, Erm, F Actin, G-Actin, GIPC1, GPC1, IQGAP3, JUB, LIMA1, MYC, MYO1B, MYO1C, NCKAP1, PFN2, PLS3, Profilin, Rock, S100A11, SR140, TPM1, Vla-4	34	21	Cellular Assembly and Organization, Cell Morphology, Cellular Compromise
4	ALDH7A1, CD3, CD3D, CORO1A, CREB3, Gap, HCLS1, IGLL1, IKZF1, IKZF5, IL2RG, LCP2, Mek, Nfat (family), NFkB (complex), PLK2, Ptk, PTPRC, PTPRCAP, PVR, Rap1, Ras, RASGRP2, RASSF5, RHOH, RIN2, RRAS2, SDC4, Sfk, SHB, SLA, Sos, TCR, TNFRSF12A, VAV	34	23	Hematological System Development and Function, Tissue Morphology, Cellular Development
5	AHNAK, Alpha tubulin, ANXA5, ARHGAP29, CBR1, CD44, Collagen type I, Collagen type IV, CTNND1, CTTN, CXCR4, CYR61, DCBLD2, DOCK2, ELMO1, Fgf, hCG, MAP2K1/2, MET, Mmp, P38 MAPK, Pak, PAXIP1, Pdgf, PDGF BB, PHLDA1, PLC gamma, PP2A, PTPN7, Rac, RND3, S100A10, SELPLG, Shc, TM4SF1	32	20	Cellular Movement, Cancer, Cardiovascular System Development and Function
6	AHNAK, ALOX5, AMOTL2, ARHGAP19, ATXN2, CHI3L1, CHST3, CNN3, CORO1C, COTL1, CYFIP1, CYFIP2 (includes EG:26999), DAZAP2 (includes EG:9802), FAM50A, FNDC3B, FXR2, GAS7, KDELR2, KIAA0182, KIAA1217, LCP1, MYOF, NCKAP1, NCKAP1L, NNMT, PDLIM4, QKI, RBM9, RBPMS, RERE, RHOXF2, RPIA, SF1, STK16, TGFB1	28	18	Lipid Metabolism, Small Molecule Biochemistry, Cellular Assembly and Organization

ID	Molecules in Network	Score	Focus Molecules	Top Functions
7	AHNAK, AK1, AKNA, ATP6V0E1, BTG3, CCDC80, CD40LG, CDCA7L, CDKN2A, CHEMOKINE, CSF1, ELOVL1, ERBB2, FAM129B, GULP1, HRAS, IL1A, ITGB1, JAM2, KANK2, LAMB1, LRMP, LXN, MFNG, MGAT5, MIR124, MYO10, MYO1G, NPNT, P4HA2, PMEPA1, PTRF, RIN2, S100A13, SLC29A1	28	18	Cell Cycle, Cellular Growth and Proliferation, Cell-To-Cell Signaling and Interaction
8	ADAM9, Calpain, Caveolin, CD53, CD63, CD151, Collagen(s), ERK1/2, FERMT3, FHL2, Fibrinogen, Focal adhesion kinase, Integrin, Integrin alpha 3 beta 1, Integrin alpha 6 beta 1, Integrin alpha V beta 3, Integrinα, Integrinβ, ITGA4, ITGB1, ITGB5, LAMB1, LAMC1, Laminin, Laminin1, Laminin2, LPP, Metalloprotease, MYO10, NTN4, SHPRH, Talin, TSPAN, TSPAN4, TSPAN6	26	17	Nervous System Development and Function, Tissue Development, Cell-To-Cell Signaling and Interaction
9	amino acids, ANLN, ARHGAP4, ARHGAP8, ARHGAP15, CAMSAP1L1, CDC42, CDC42BPA, CDC42BPB, DCP2, DEF6, DST, E2F1, EZR, FGD1, FGD3, FGD1/3, FIP1L1, GMFG, GNG12, GSK3B, HDAC4, MIR1, MIRLET7A1, MYO18A, RAC1, RHOA, RNF138, RNPS1, RUVBL2, SEPT6, SFRS2, TRAF3IP3, YWHAZ, ZCCHC3	25	17	Cellular Assembly and Organization, Cell Morphology, Cell Signaling
10	Ap1, ATP2A3, Caspase, CCND1, CTBP2, Cyclin A, Cyclin E, E2f, EIF4B, Estrogen Receptor, FLNA, Growth hormone, Gsk3, HCST, Hsp70, Ifn gamma, IL1, IL6ST, Insulin, Interferon alpha, JAK, Jnk, LGALS3, LIF, MT2A, MYB (includes EG:4602), NASP, NQO1, p85 (pik3r), PI3K, PI3K p85, PRKCQ, SMARCC1, STAT, STAT5a/b	22	15	Cellular Development, Cellular Growth and Proliferation, Cell Morphology
11	ANKS1B, beta-estradiol, CXCR7, DLG4, DLGAP4, EIF3D, FLT4, FN1, GLUL, GRB2, HTRA1, KIF21B, KRT17, LHFPL2, LIMA1, LMO7, MATN2, MIR23B (includes EG:407011), MSL2, MYO1B, PKM2, PRSS23, RAI14, RAPSN, RPS13, SCRN1, SHANK3, SLC25A3, SLC25A12, SMAD7, SNX7, TAF4B, UACA, ZNF107, ZNRF1	20	14	Cellular Assembly and Organization, Embryonic Development, Organ Development
12	BRF1, C16ORF53, CABC1, CLTCL1, CTNNB1, CTNNβ-TCF/LEF, DLG5, GART, GAS1, GLTSCR2, HNF4A, IFNA2, IPO13, KCTD1, KDM2B, KIF20A, LAPTM4A, LAPTM4B, NBR1, NFE2L3, NFYB, NME3, NME7, NOSIP, PERP, PRRG2, RFC3, RIOK1, RPL41, SAMSN1, SAT2, SGCE, TAX1BP3, TP53, ZNRD1	18	14	Cell Cycle, Gene Expression, Cancer
13	ABHD4, AGAP1, ASPH, ATP6V1D, ATP6V1E1, CD48, CDH13, CHP, DEFB103A, FABP7, Focal adhesion kinase, FUCA1, GNA11, HIF1A, HTRA1, HTT, IL13, LDHA, LONP1, MTSS1, NFkB (complex), oleic acid, P4HA1, PKM2, PRDX3, RAB33A, RTN3, SCARB2, SEPT9, SLC25A3, SLC25A11, SLC2A4, SRGAP1, ST8SIA4, USP20	13	11	Energy Production, Molecular Transport, Nucleic Acid Metabolism
14	2' 5' oas, Androgen-AR, ARHGAP24, ARHGAP26, ATP1B1, Bcl10-Card10-Malt1, CACNB2, CBFA2T3, Ck2, CST3, EXT1, Histone h3, Histone h4, IgG, IKK (complex), IL12 (complex), Immunoglobulin, MAP1LC3A, Mapk, MHC Class II (complex), MYO9B, NGF, PARP10, Pka, Pkc(s), PRKCB, Ras homolog, RHOC, RHPN1, RNA polymerase II, SPATS2L, SQSTM1, TGFB2, TRIM16, Ubiquitin	12	10	Cardiac Necrosis/Cell Death, Cell Death, Cellular Assembly and Organization

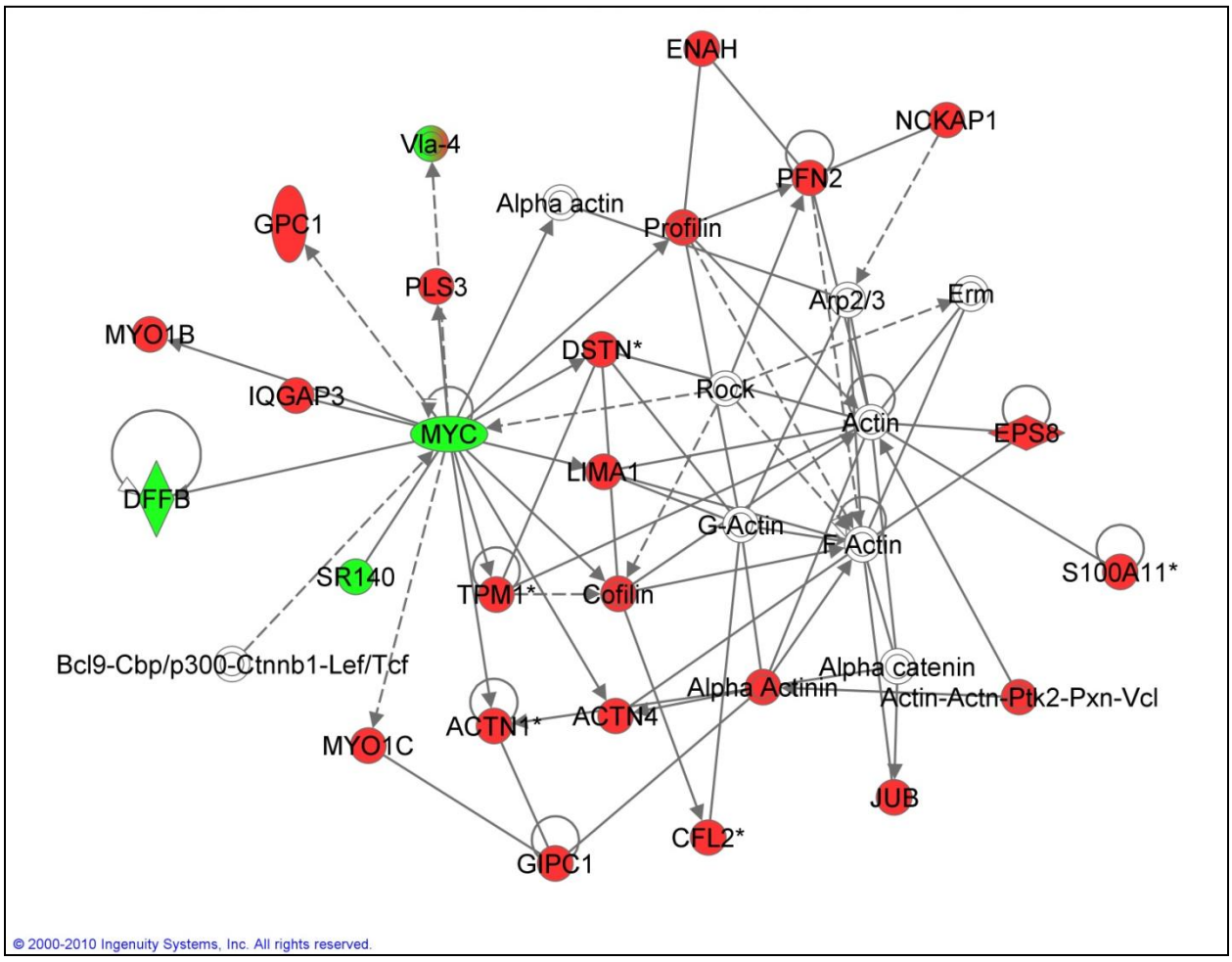
Network 1:


TEAD1
ACAP1
CAPN2
CAST*
WWTR1
YAP1
PEA15*
CaMKII
Akt
PTPRF
Calmodulin
RGNEF
LAIR1
Collagen Alpha1
Filamin
Tgf beta
CAV1*
FLNB
sphingomyelinase
SH3BP4
Dynamin
PTK2
Ecm
ASAP2
PXN
AGRN*
DAG1
CAV2*
ASAP1
CKAP4*
PLEC
NARF
ANXA2*
LMNA*
Lamin b
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

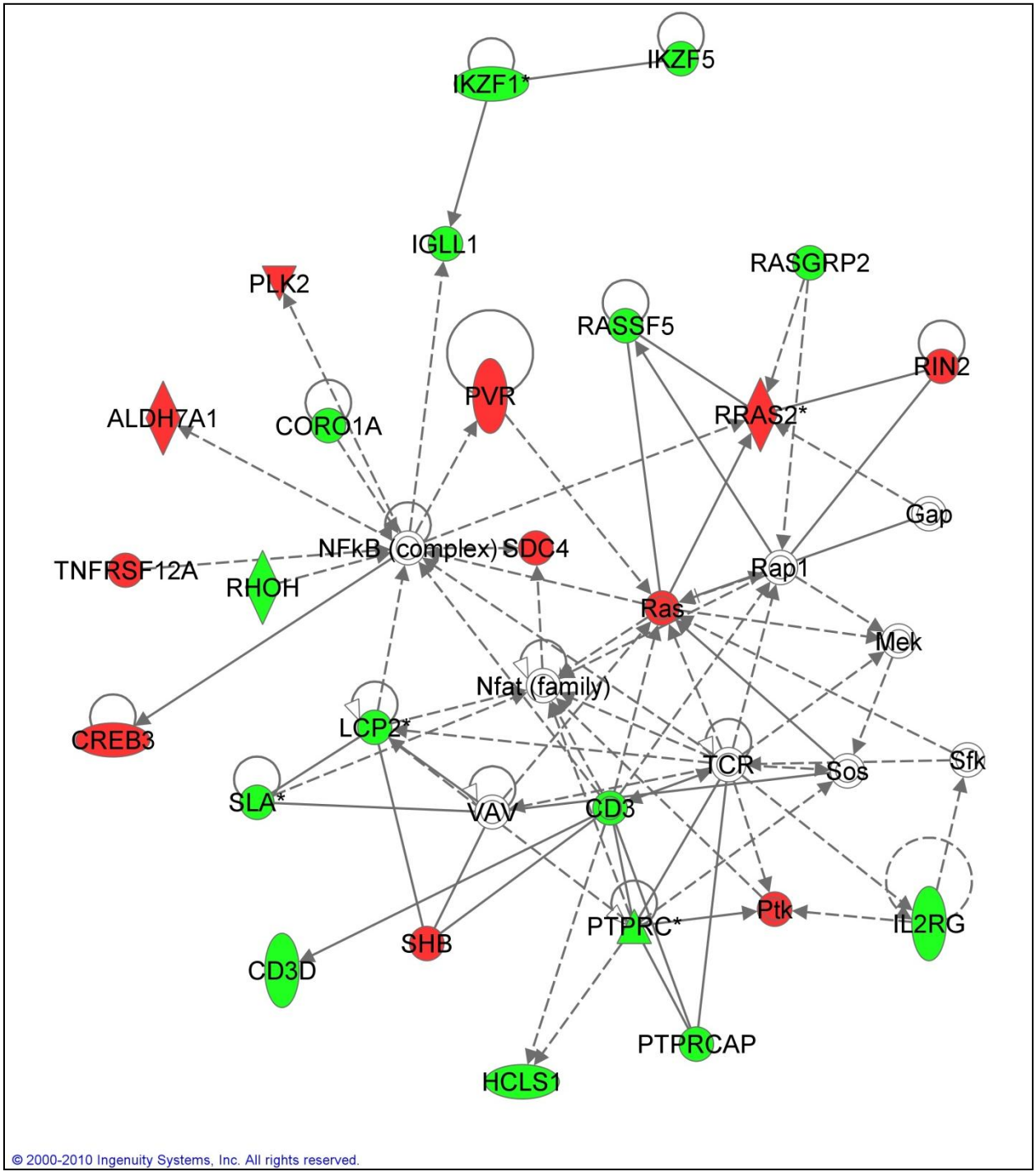
Network 2 (as seen in the body):



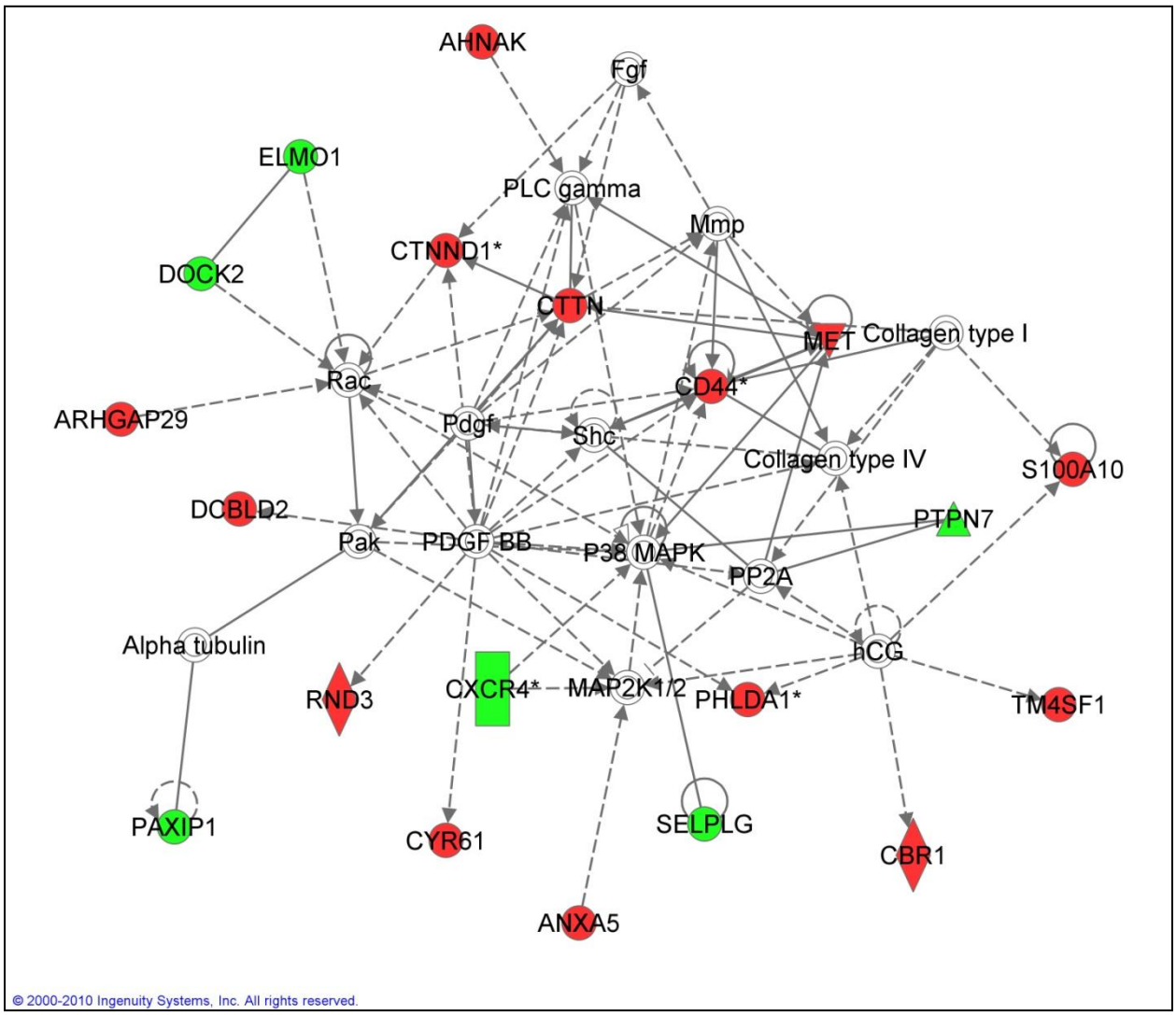
Network 3:


ENAH
NCKAP1
Vla-4
PFN2
Alpha actin
GPC1
Profilin
PLS3
Arp2/3
Erm
MYO1B
DSTN*
IQGAP3
Rock
Actin
EPS8
MYC
LIMA1
DFFB
G-Actin
F-Actin
SR140
TPM1*
Cofilin
S100A11*
Bcl9-Cbp/p300-Ctnnb1-Lef/Tcf
Alpha catenin
Alpha Actinin
Actin-Actn-Ptk2-Pxn-Vcl
ACTN1*
ACTN4
MYO1C
JUB
CFL2*
GIPC1
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

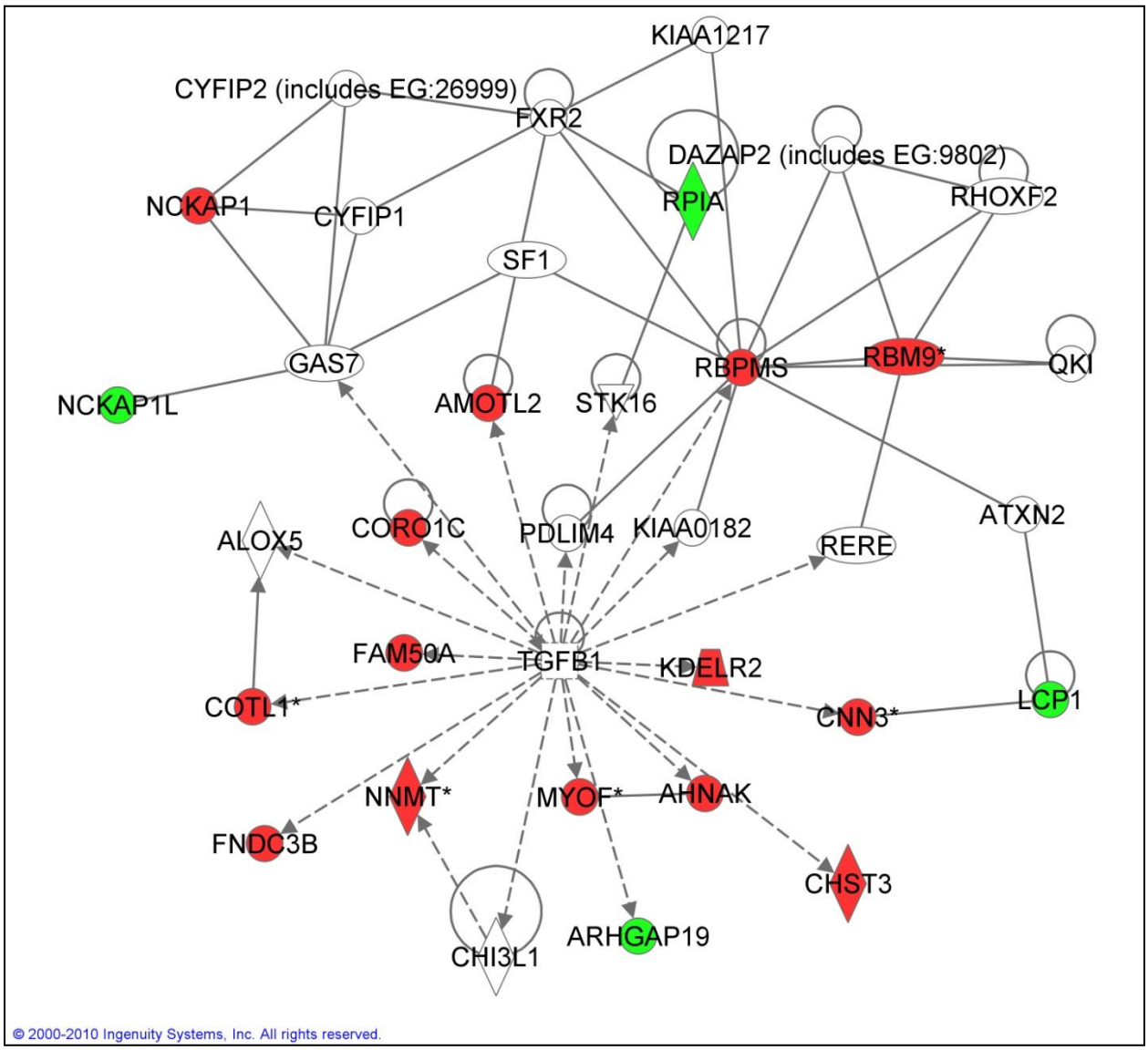
Network 4:


IKZF1*
IKZF5
IGLL1
PLK2
RASGRP2
RASSF5
RIN2
PVR
ALDH7A1
CORO1A
RRAS2*
Gap
NFkB (complex)
SDC4
TNFRSF12A
RHOH
Rap1
Ras
Mek
Nfat (family)
LCP2*
CREB3
TCR
Sos
Sfk
SLA*
VAV
CD3
Ptk
PTPRC*
IL2RG
SHB
CD3D
PTPRCAP
HCLS1
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

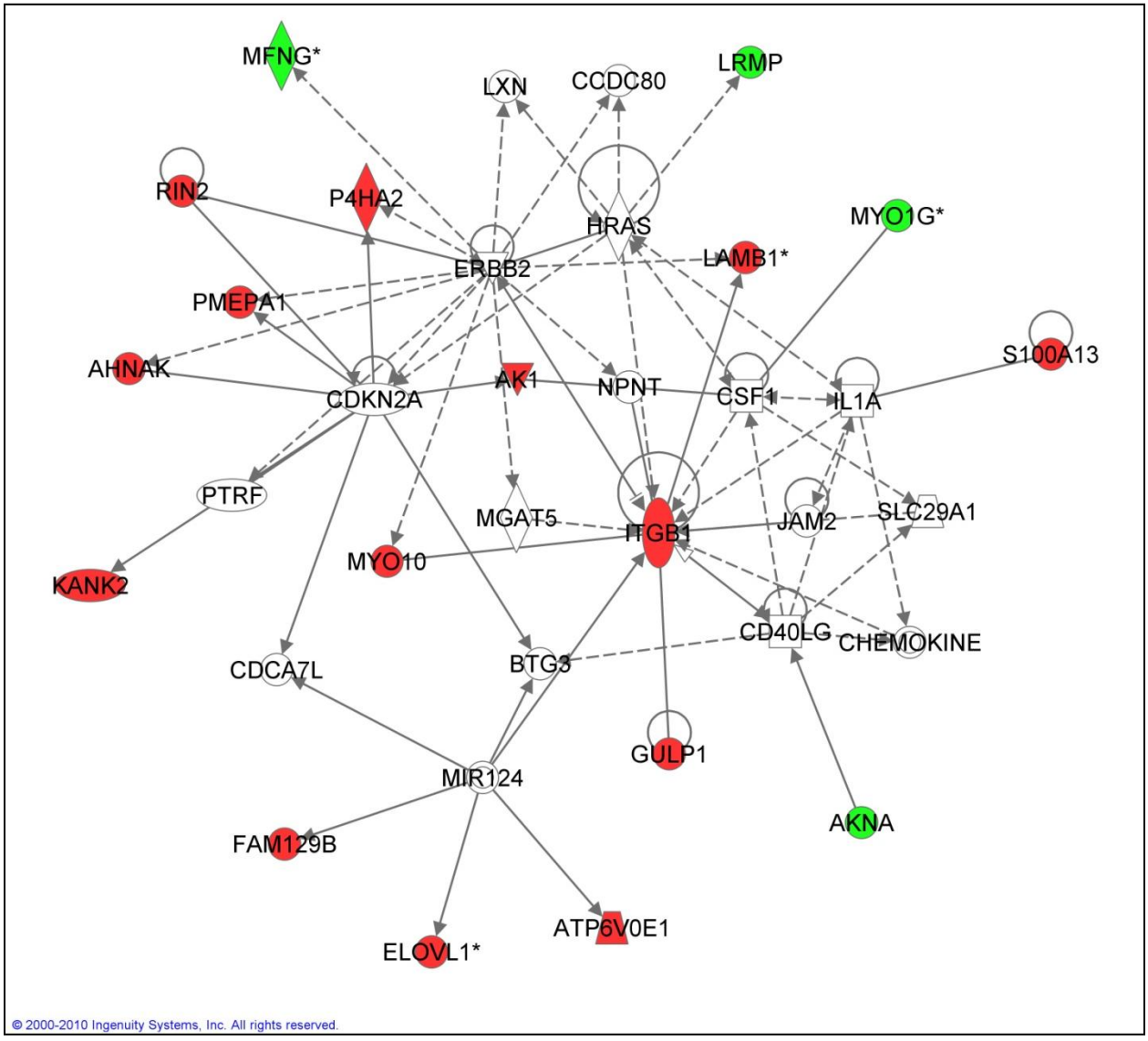
Network 5:


AHNAK
Fgf
ELMO1
PLC gamma
Mmp
CTNND1*
DOCK2
CTTN
MET
Collagen type I
Rac
CD44*
ARHGAP29
Pdgf
Shc
Collagen type IV
S100A10
DCBLD2
PTPN7
Pak
PDGF BB
P38 MAPK
PP2A
Alpha tubulin
hCG
RND3
CXCR4*
MAP2K1/2
PHLDA1*
TM4SF1
PAXIP1
CYR61
SELPLG
CBR1
ANXA5
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

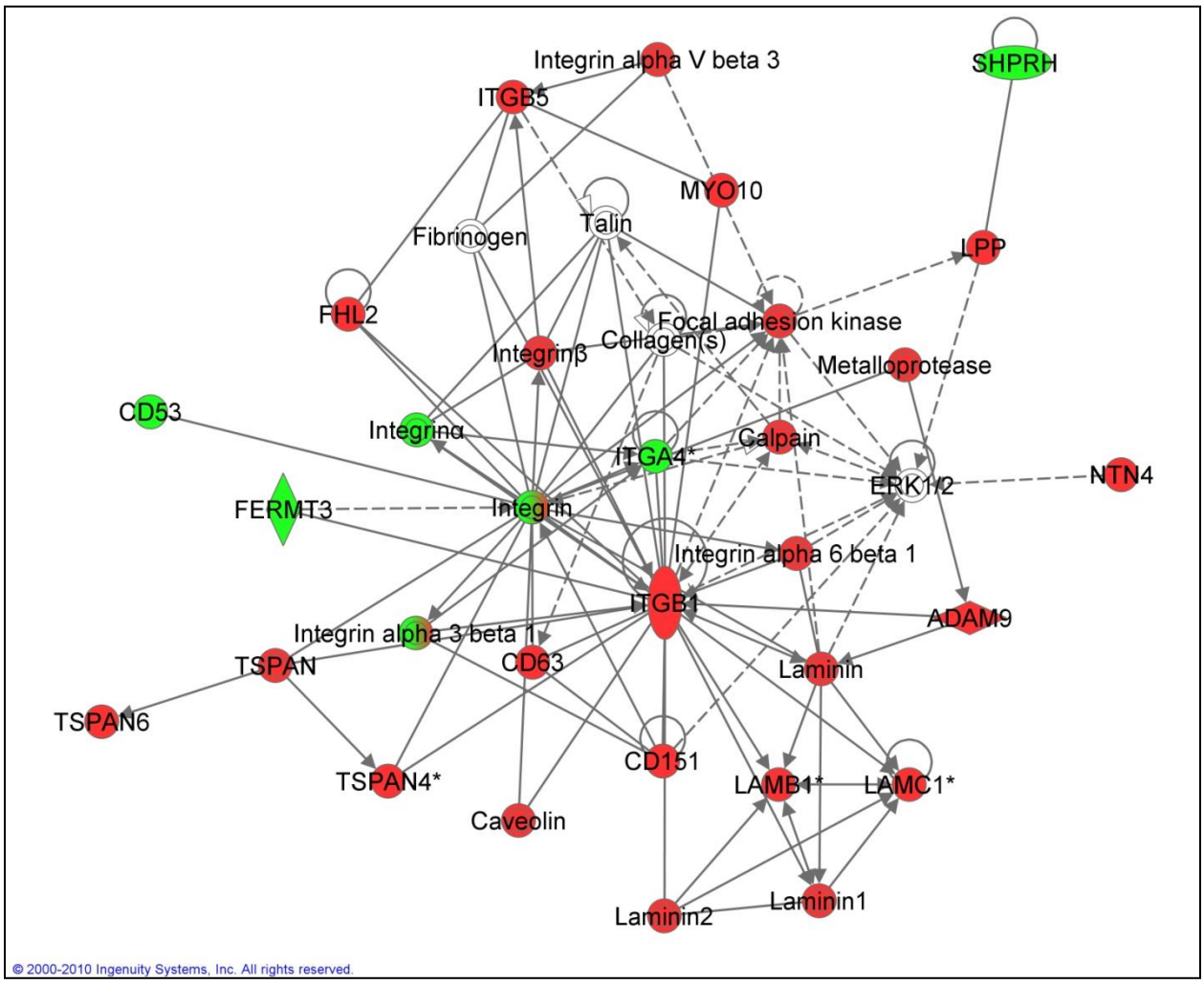
Network 6:


KIAA1217
CYFIP2 (includes EG:26999)
FXR2
DAZAP2 (includes EG:9802)
NCKAP1
CYFIP1
RPIA
RHOXF2
SF1
GAS7
RBPMS
RBM9*
QKI
NCKAP1L
AMOTL2
STK16
ATXN2
ALOX5
CORO1C
PDLIM4
KIAA0182
RERE
FAM50A
TGFB1
KDELR2
COTL1*
CNN3*
LCP1
NNMT*
MYOF*
AHNAK
FNDC3B
CHST3
ARHGAP19
CHI3L1
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

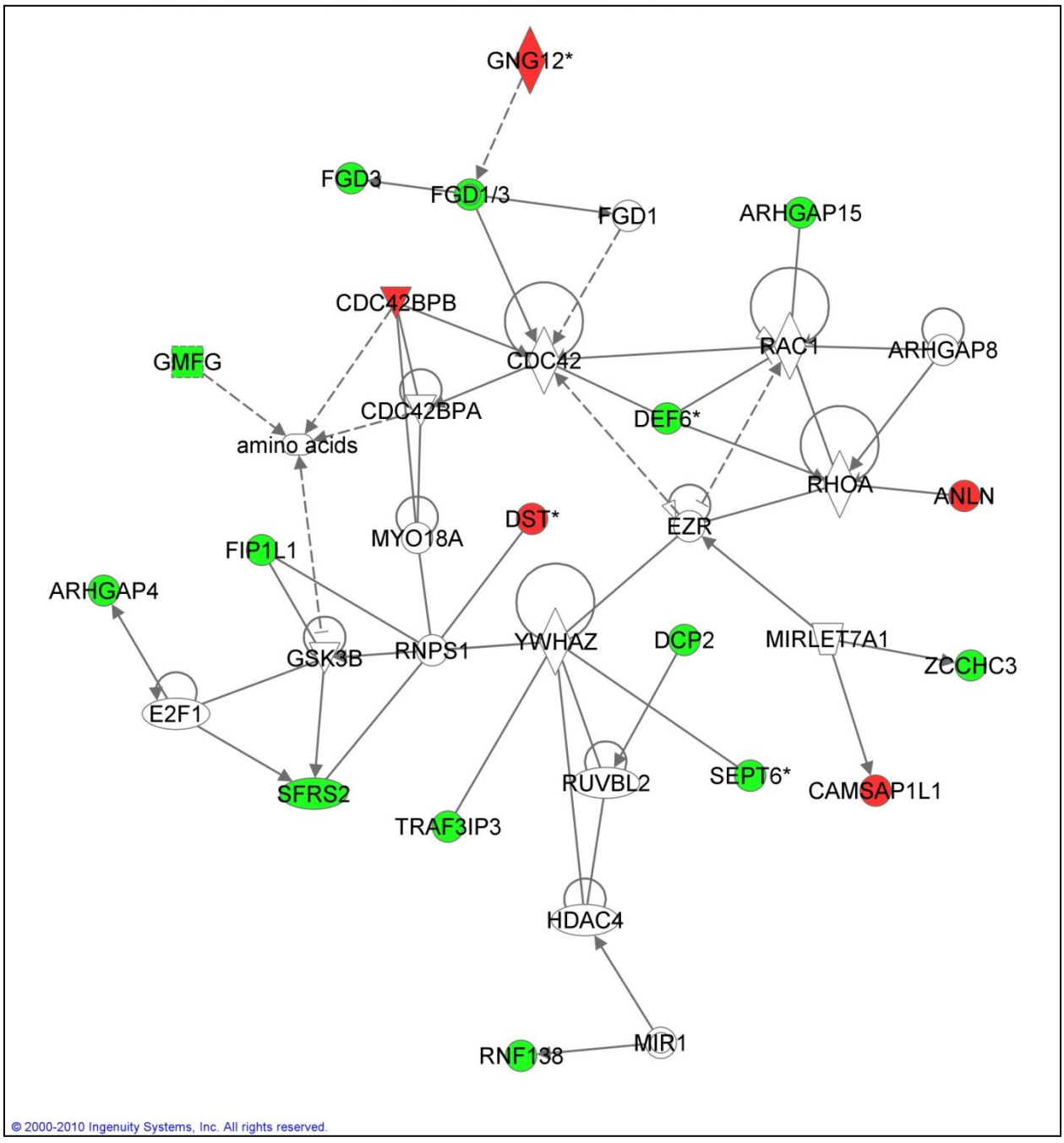
Network 7:


MFNG*
LXN
CCDC80
LRMP
RIN2
P4HA2
HRAS
MYO1G*
ERBB2
LAMB1*
PMEPA1
AHNAK
CDKN2A
AK1
NPNT
CSF1
IL1A
S100A13
PTRF
MGAT5
ITGB1
JAM2
SLC29A1
MYO10
KANK2
CD40LG
CHEMOKINE
CDCA7L
BTG3
GULP1
MIR124
AKNA
FAM129B
ATP6V0E1
ELOVL1*
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

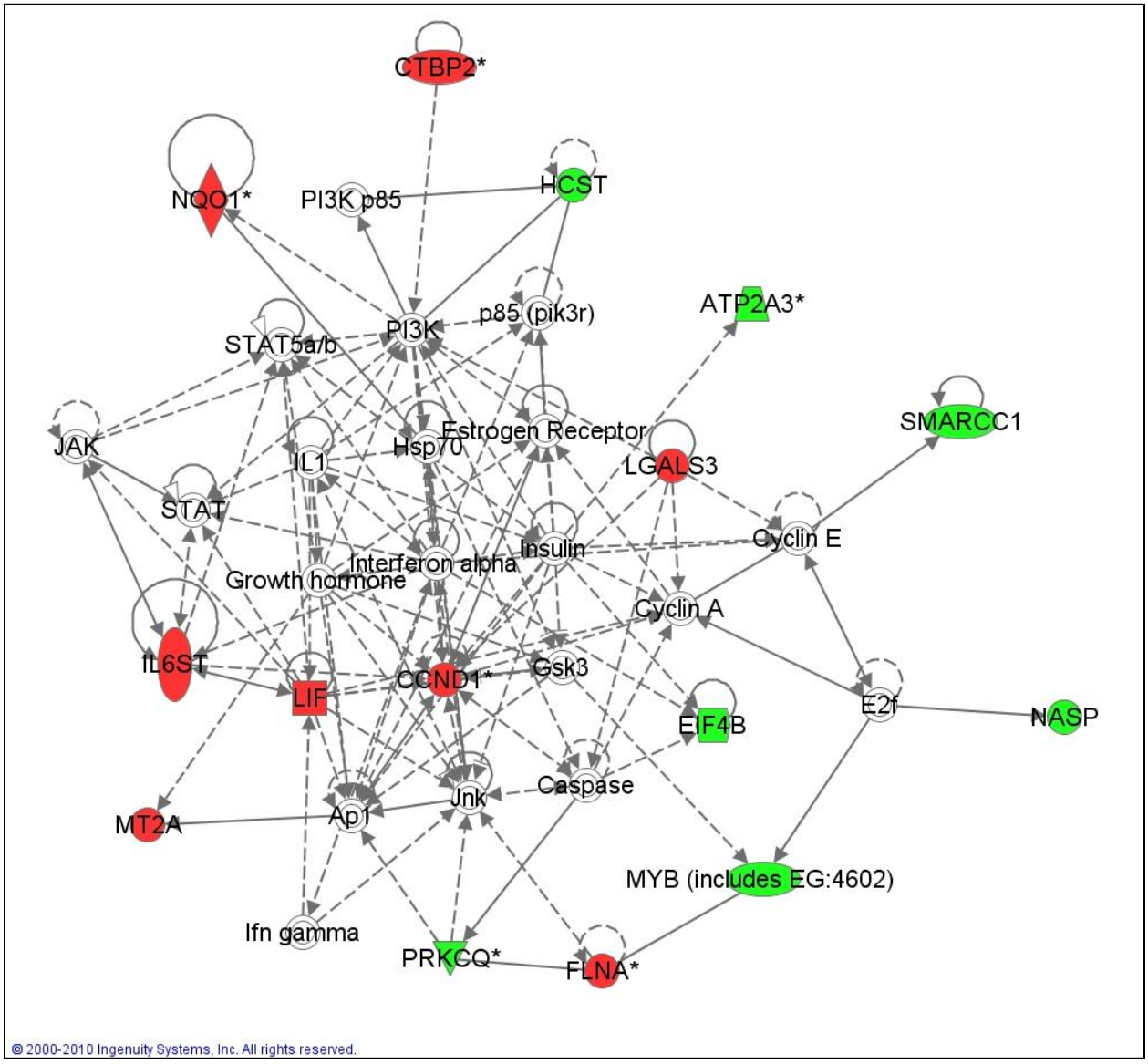
Network 8:


Integrin alpha V beta 3
SHPRH
ITGB5
MYO10
Fibrinogen
Talin
LPP
FHL2
Focal adhesion kinase
Integrinβ
Collagen(s)
Metalloprotease
CD53
Integrinα
Calpain
ITGA4*
NTN4
ERK1/2
FERMT3
Integrin
Integrin alpha 6 beta 1
ITGB1
ADAM9
Integrin alpha 3 beta 1
TSPAN
CD63
Laminin
TSPAN6
CD151
TSPAN4*
LAMB1*
LAMC1*
Caveolin
Laminin1
Laminin2
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

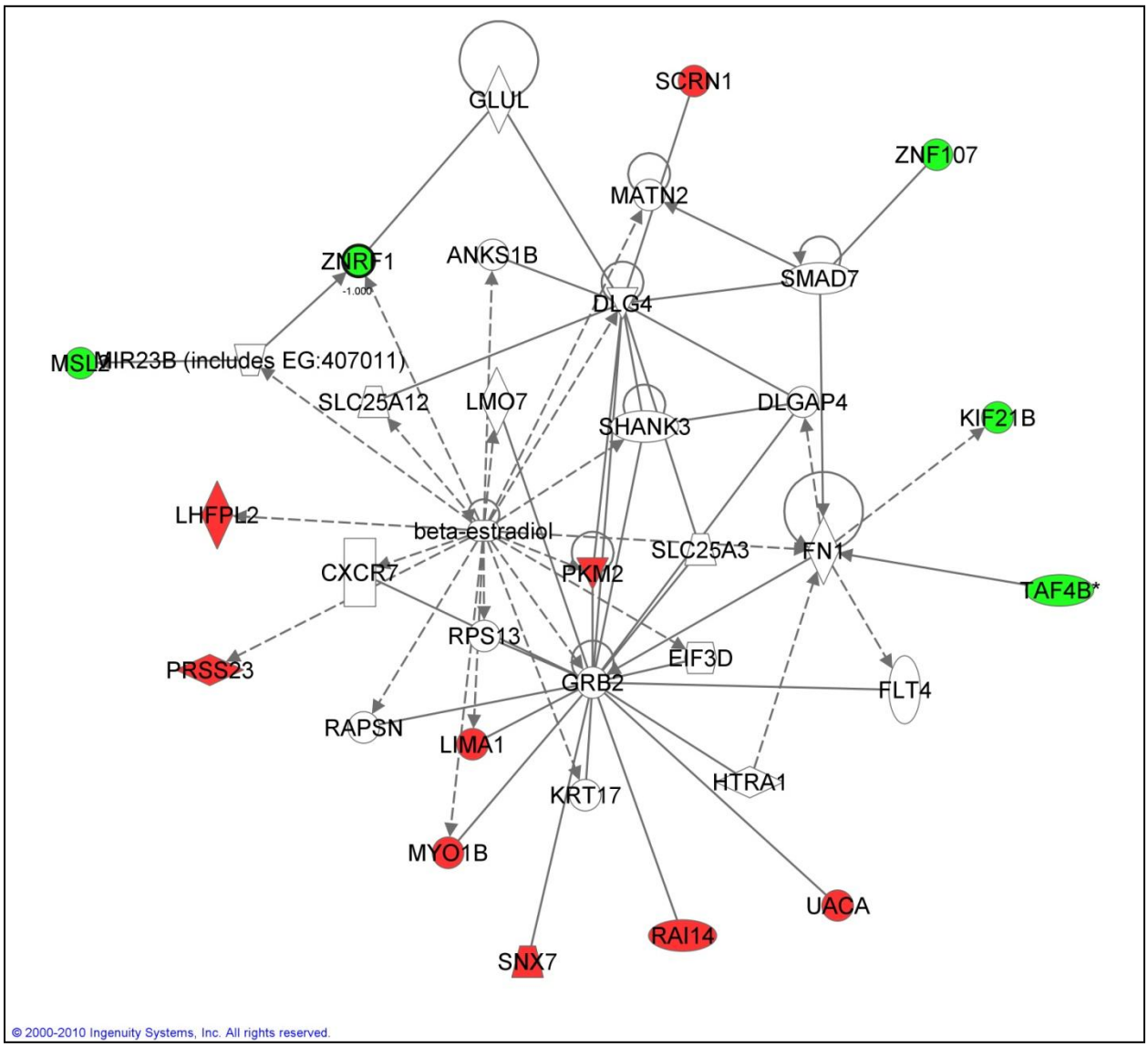
Network 9:


GNG12*
FGD3
FGD1/3
FGD1
ARHGAP15
CDC42BPB
GMFG
CDC42
RAC1
ARHGAP8
CDC42BPA
DEF6*
amino acids
RHOA
ANLN
DST*
MYO18A
EZR
FIP1L1
ARHGAP4
GSK3B
RNPS1
YWHAZ
DCP2
MIRLET7A1
ZCCHC3
E2F1
SFRS2
RUVBL2
SEPT6*
CAMSAP1L1
TRAF3IP3
HDAC4
RNF138
MIR1
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

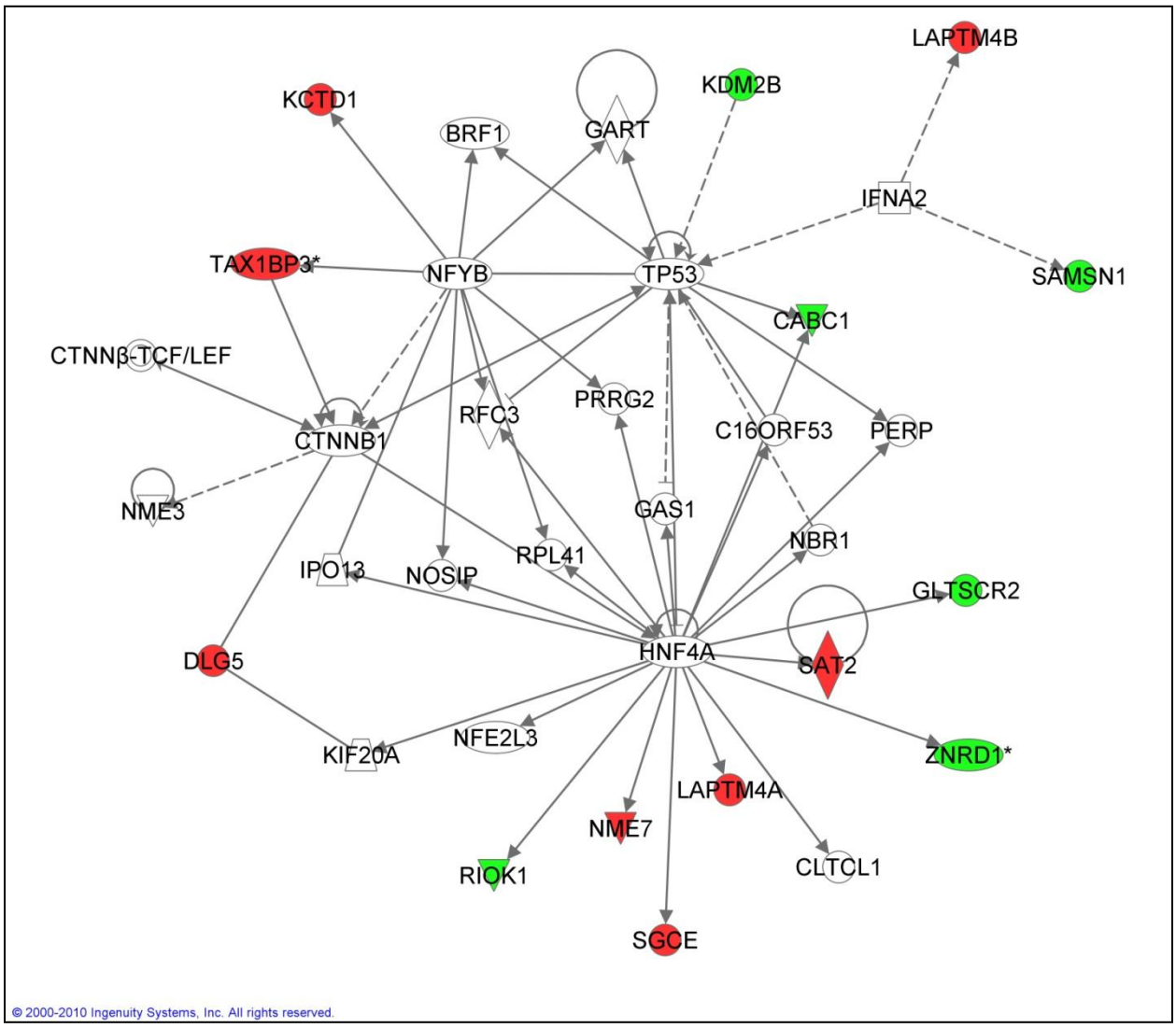
Network 10:


CTBP2*
NQO1*
PI3K p85
HCST
ATP2A3*
p85 (pik3r)
PI3K
STAT5a/b
SMARCC1
Estrogen Receptor
JAK
Hsp70
IL1
LGALS3
STAT
Cyclin E
Insulin
Interferon alpha
Growth hormone
Cyclin A
IL6ST
Gsk3
CCND1*
LIF
E2f
NASP
EIF4B
Caspase
Jnk
Ap1
MT2A
MYB (includes EG:4602)
Ifn gamma
PRKCQ*
FLNA*
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

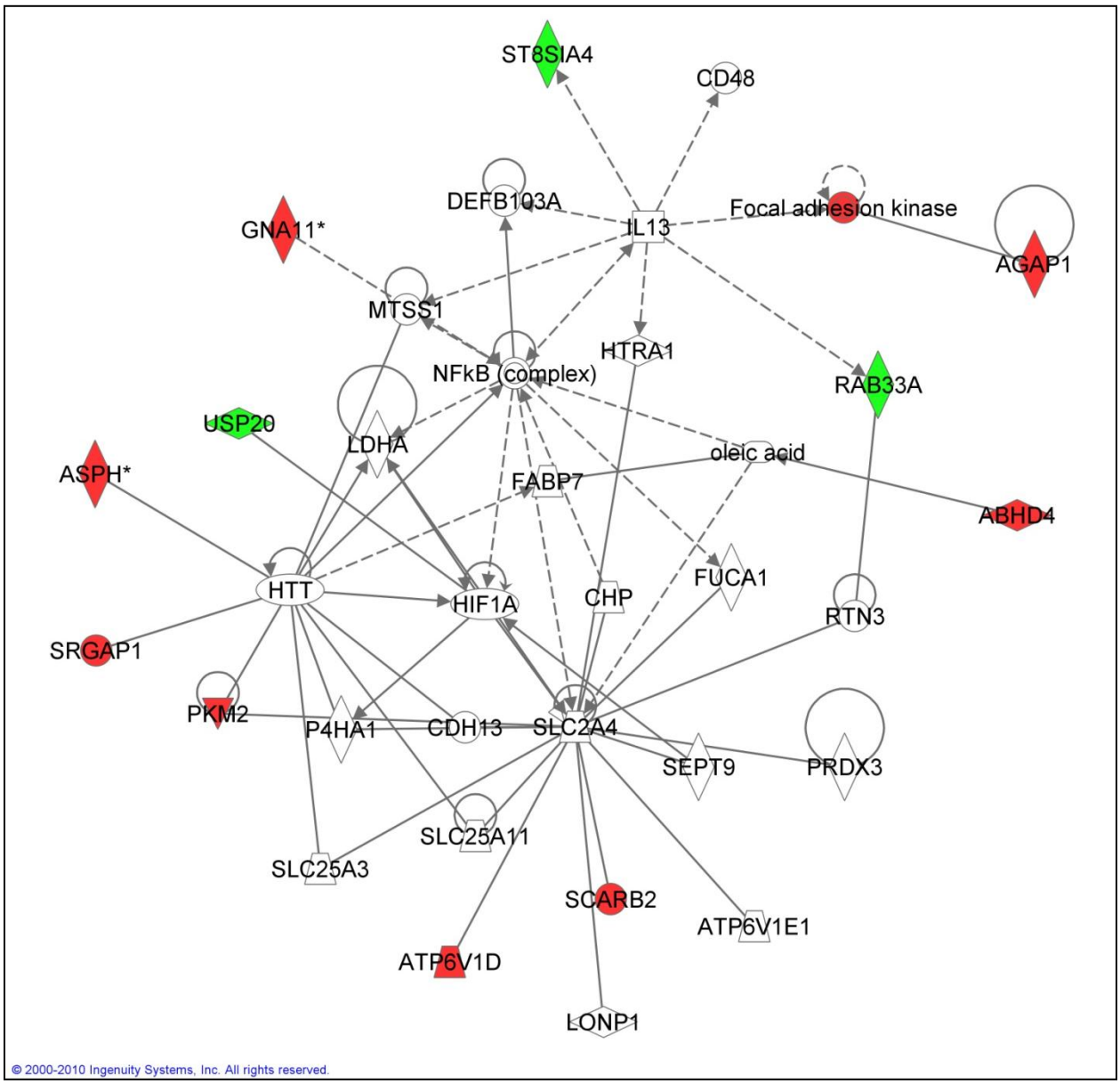
Network 11:


GLUL
SCRN1
ZNF107
MATN2
ZNRF1
-1.000
ANKS1B
SMAD7
DLG4
MSL2
MIR23B (includes EG:407011)
SLC25A12
LMO7
SHANK3
DLGAP4
KIF21B
LHFPL2
beta-estradiol
SLC25A3
FN1
CXCR7
PKM2
TAF4B*
RPS13
PRSS23
EIF3D
GRB2
FLT4
RAPSN
LIMA1
HTRA1
KRT17
MYO1B
UACA
RAI14
SNX7
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

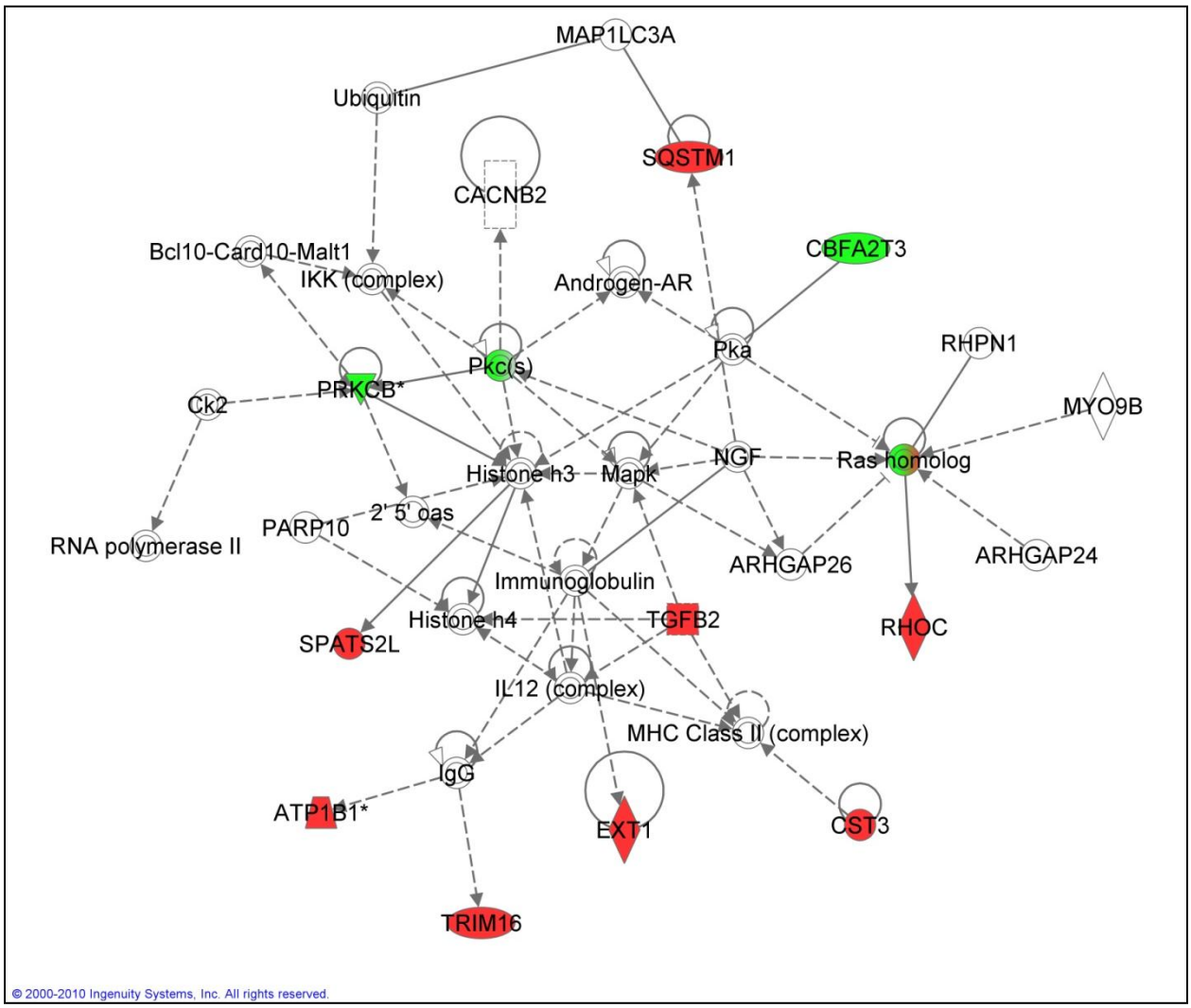
Network 12:


LAPTM4B
KCTD1
KDM2B
BRF1
GART
IFNA2
TAX1BP3*
NFYB
TP53
SAMSN1
CABC1
CTNNβ-TCF/LEF
PRRG2
RFC3
C16ORF53
PERP
CTNNB1
NME3
GAS1
NBR1
RPL41
IPO13
NOSIP
GLTSCR2
HNF4A
DLG5
SAT2
NFE2L3
KIF20A
ZNRD1*
LAPTM4A
NME7
RIOK1
CLTCL1
SGCE
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

Network 13:


ST8SIA4
CD48
DEFB103A
GNA11*
IL13
Focal adhesion kinase
AGAP1
MTSS1
HTRA1
NFkB (complex)
RAB33A
USP20
LDHA
oleic acid
ASPH*
FABP7
ABHD4
HTT
HIF1A
CHP
FUCA1
RTN3
SRGAP1
PKM2
P4HA1
CDH13
SLC2A4
SEPT9
PRDX3
SLC25A11
SLC25A3
SCARB2
ATP6V1E1
ATP6V1D
LONP1
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

Network 14:


MAP1LC3A
Ubiquitin
SQSTM1
CACNB2
Bcl10-Card10-Malt1
IKK (complex)
Androgen-AR
CBFA2T3
RHPN1
Pka
Pkc(s)
PRKCB*
Ck2
MYO9B
NGF
Ras homolog
Histone h3
Mapk
2' 5' oas
PARP10
RNA polymerase II
ARHGAP26
ARHGAP24
Immunoglobulin
Histone h4
TGFB2
RHOC
SPATS2L
IL12 (complex)
MHC Class II (complex)
IgG
ATP1B1*
EXT1
CST3
TRIM16
© 2000-2010 Ingenuity Systems, Inc. All rights reserved.

Reference List

1. Szalay, A.; Gray, J. 2020 Computing: Science in an exponential world. *Nature* **2006,** *440*, 413-414.

2. Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004,** *306*, 1138-1139.

3. PubChem. http://pubchem.ncbi.nlm.nih.gov/ . 2010.
   Ref Type: Electronic Citation

4. ChemSpider. ChemSpider. http://www.chemspider.com . 2009.
   Ref Type: Electronic Citation

5. ChEMBL Database. http://www.ebi.ac.uk/chembldb/ . 2010.
   Ref Type: Electronic Citation

6. PDSP. PDSP. http://pdsp.med.unc.edu . 2009.
   Ref Type: Electronic Citation

7. Oprea, T.; Tropsha, A. Target, Chemical and Bioactivity Databases – Integration is Key. Drug Discov.Today 3, 357-365. 2006.
   Ref Type: Journal (Full)

8. Hughes, B. 2009 FDA drug approvals. *Nat Rev Drug Discov* **2010,** *9*, 89-92.

9. Hughes, B. 2008 FDA drug approvals. *Nat Rev Drug Discov* **2009,** *8*, 93-96.

10. Sams-Dodd, F. Research & market strategy: how choice of drug discovery approach can affect market position. *Drug Discovery Today* **2007,** *12*, 314-318.

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

12. DiMasi, J. A.; Feldman, L.; Seckler, A.; Wilson, A. Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs. *Clin Pharmacol Ther* **2010,** *87*, 272-277.

13. DiMasi, J. A.; Grabowski, H. G. The cost of biopharmaceutical R&D: is biotech different? *Manage. Decis. Econ.* **2007,** *28*, 469-479.

14. Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology* **2007,** *152*, 9-20.

15. Clark, D. E. What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discov.* **2008,** *3*, 841-851.

16. Kubinyi, H. Chance Favors the Prepared Mind - From Serendipity to Rational Drug Design. *Journal of Receptors and Signal Transduction* **1999,** *19*, 15-39.

17. von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Phan, T. V.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993,** *363*, 418-423.

18. Schames, J. R.; Henchman, R. H.; Siegel, J. S.; Sotriffer, C. A.; Ni, H.; McCammon, J. A. Discovery of a Novel Binding Trench in HIV Integrase. *Journal of Medicinal Chemistry* **2004,** *47*, 1879-1881.

19. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004,** *3*, 935-949.

20. Brooijmans, N.; Kuntz, I. D. MOLECULAR RECOGNITION AND DOCKING ALGORITHMS. *Annu. Rev. Biophys. Biomol. Struct.* **2003,** *32*, 335-373.

21. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* **1982,** *161*, 269-288.

22. Wlodawer, A.; Vondrasek, J. INHIBITORS OF HIV-1 PROTEASE: A Major Success of Structure-Assisted Drug Design1. *Annu. Rev. Biophys. Biomol. Struct.* **1998,** *27*, 249-284.

23. Wong, C. F.; McCammon, J. A. Protein flexibility and computer sided drug design. *Annual Review of Pharmacol. Toxicol* **2003,** *43*, 31-45.

24. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des* **2002,** *16*, 151-166.

25. Muegge, I. Selection criteria for drug-like compounds. *Medicinal research reviews* **2003,** *23*, 302-321.

26. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998,** *19*, 1639-1662.

27. Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **2001,** *308*, 377-395.

28. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algroithm for flexible docking. *Journal of Molecular Biology* **1997,** *267*, 727-748.

29. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004,** *47*, 2356-2364.

30. Cherkasov, A.; Shi, Z.; Fallahi, M.; Hammon, G. L. Successful in silico discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *Journal of Medicinal Chemistry* **2005,** *48*, 3203-3213.

31. Ghoneim, O. M.; Legere, J. A.; Golbraikh, A.; Tropsha, A.; Booth, R. G. Novel ligands for the human histamine H1 receptor: synthesis, pharmacology, and comparative molecular field analysis studies of 2-dimethylamino-5-(6)-phenyl-1,2,3,4-tetrahydronaphthalenes. *Bioorg. Med. Chem.* **2006,** *14*, 6640-6658.

32. Zhang, S.; Wei, L.; Bastow, K.; Zheng, W.; Brossi, A.; Lee, K. H.; Tropsha, A. Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J. Comput. Aided Mol. Des* **2007,** *21*, 97-112.

33. Tropsha, A. Predictive QSAR (Quantitative Structure Activity Relationships) Modeling. In *Comprehensive Medicinal Chemistry II*; Martin, Y. C. Ed.; Elsevier: 2006; pp 113-126.

34. Vert, J. P.; Jacob, L. Machine Learning for In Silico Virtual Screening and Chemical Genomics: New Strategies. *Combinatorial Chemistry & High Throughput Screening* **2008,** *11*, 677-685.

35. Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des* **2001,** *7*, 567-597.

36. Muegge, I.; Oloff, S. Advances in virtual screening. *Drug Discovery Today: Technologies* **2006,** *3*, 405-411.

37. Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhancing Drug Discovery Through In Silico Screening: Strategies to Increase True Positives Retrieval Rates. *Current Medicinal Chemistry* **2008,** *15*, 2040-2053.

38. Zhang, S.; Golbraikh, A.; Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J. Med. Chem.* **2006,** *49*, 2713-2724.

39. Wu, G.; Vieth, M. SDOCKER: A Method Utilizing Existing X-ray Structures To Improve Docking Accuracy. *Journal of Medicinal Chemistry* **2004,** *47*, 3142-3148.

40. Brockman, R. W. Mechanisms of Resistance to Anticancer Agents. In *Advances in Cancer Research*, Volume 7 ed.; Alexander Haddow and Sidney Weinhouse Ed.; Academic Press: 1963; pp 129-234.

41. L°nning, P. E. Molecular basis for therapy resistance. *Molecular Oncology* **2010,** *4*, 284-300.

42. Powis, G. Achieving Personalized Cancer Medicine: Trials and Tribulations. *Journal of Investigative Medicine* **2006,** *54*.

43. van 't Veer, L. J.; Bernards, R. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* **2008,** *452*, 564-570.

44. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997,** *25*, 3389-3402.

45. Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007,** *23*, 2947-2948.

46. Falquet, L.; Pagni, M.; Bucher, P.; Hulo, N.; Sigrist, C. J.; Hofmann, K.; Bairoch, A. The PROSITE database, its status in 2002. *Nucleic Acids Res.* **2002,** *30*, 235-238.

47. Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **2003,** *31*, 3692-3697.

48. Cammer, S. SChiSM2: creating interactive web page annotations of molecular structure models using Jmol. *Bioinformatics* **2007,** *23*, 383-384.

49. Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **2006,** *63*, 892-906.

50. Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic acids research* **2003,** *31*, 3352-3355.

51. Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology* **2006,** *6*, 19.

52. Laurie, A. T. R.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of proteinΓÇôligand binding sites. *Bioinformatics 21*, 1908-1916.

53. Lee, J.; Sinkovits, R.; Mock, D.; Rab, E.; Cai, J.; Yang, P.; Saunders, B.; Hsueh, R.; Choi, S.; Subramaniam, S.; Scheuermann, R.; in collaboration with the Alliance for Cellular Signaling Components of the antigen processing and presentation

pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. *BMC bioinformatics* **2006,** *7*, 237.

54. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006,** *49*, 5912-5931.

55. Bar-Haim, S.; Aharon, A.; Ben Moshe, T.; Marantz, Y.; Senderowitz, H. SeleX-CS: A New Consensus Scoring Algorithm for Hit Discovery and Lead Optimization. *J. Chem. Inf. Model.* **2009,** *49*, 623-633.

56. Takahashi, O.; Masuda, Y.; Muroya, A.; Furuya, T. Theory of docking scores and its application to a customizable scoring function. *Sar and Qsar in Environmental Research* **2010,** *21*, 547-558.

57. Tang, H.; Wang, X. S.; Huang, X. P.; Roth, B. L.; Butler, K. V.; Kozikowski, A. P.; Jung, M.; Tropsha, A. Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J. Chem. Inf. Model.* **2009,** *49*, 461-476.

58. Peterson, Y. K.; Wang, X. S.; Casey, P. J.; Tropsha, A. Discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation. *J. Med. Chem.* **2009,** *52*, 4210-4220.

59. Oloff, S.; Zhang, S.; Sukumar, N.; Breneman, C.; Tropsha, A. Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). *J Chem. Inf. Model.* **2006,** *46*, 844-851.

60. Prymula, K.; Jadczyk, T.; Roterman, I. Catalytic residues in hydrolases: analysis of methods designed for ligand-binding site prediction. *Journal of Computer-Aided Molecular Design* **2010,** 1-17.

61. Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004,** *47*, 2977-2980.

62. sybyl7.1. 2006.
Ref Type: Computer Program

63. CCG. Molecular Operation Environment. 2010.
Ref Type: Computer Program

64. Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein science* **1998,** *7*, 1884-1897.

65. Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research* **2006,** *34*, W116-W118.

66. Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. Electron Density Modeling of Large Systems using the Transferable Atom Equivalent Method. *Comput. Chem.* **1995,** *19*, 161-169.

67. Mazza, C. B.; Sukumar, N.; Breneman, C. M.; Cramer, S. M. Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal. Chem.* **2001,** *73*, 5457-5461.

68. Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; Tugcu, N. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 1347-1357.

69. Yang, Z. P. Pharmacophore Fingerprint Analysis of Small Molecules and Proteins for Virtual High Throughput Screening. 2003. University of California, Santa Cruz. Ref Type: Thesis/Dissertation

70. Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrational Spectroscopy* **1999,** *19*, 151-164.

71. Molecular Networks GmbH. (http://www.molecular-networks.com/products) . 2010. Ref Type: Electronic Citation

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

73. Gasteiger, J.; Hutchings, M. G. Quantification of effective polarisability. Applications to studies of X-ray photoelectron spectroscopy and alkylamine protonation. *J. Chem. Soc. , Perkin Trans. 2* **1984,** 559-564.

74. Hammarstr+¦m, L. G.; Liljefors, T.; Gasteiger, J. Electrostatic interactions in molecular mechanics (MM2) calculations via PEOE partial charges I. Haloalkanes. *Journal of Computational Chemistry* **1988,** *9*, 424-440.

75. Hutchings, M. G.; Gasteiger, J. A quantitative description of fundamental polar reaction types. Proton- and hydride-transfer reactions connecting alcohols and carbonyl compounds in the gas phase. *J. Chem. Soc. , Perkin Trans. 2* **1986,** 447-454.

76. Mortier, W. J.; Van Genechten, K.; Gasteiger, J. Electronegativity equalization: application and parametrization. *Journal of the American Chemical Society* **1985,** *107*, 829-835.

77. Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem. Res. Toxicol.* **2010,** *23*, 724-732.

78. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002,** *45*, 2811-2823.

79. Hotelling, H. Relations between two sets of variables. *Biometrica* **1936,** *28*, 312-377.

80. Samarov, D. V. The analysis and advanced extensions of canonical corellation analysis. 2009. University of North Carolina at Chapel Hill.
Ref Type: Thesis/Dissertation

81. Brown, P. J. *Measurement, Regression and Calibration*; Oxford University Press, USA: 1994.

82. Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005,** *26*, 11-22.

83. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* **2003,** *17*, 241-253.

84. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010,** *29*, 476-488.

85. World Drug Index. http://scientific.thomsonreuters.com/products/wdi/. http://scientific.thomsonreuters.com/products/wdi/ . 2008.
Ref Type: Electronic Citation

86. Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic acids research* **2006,** *34*, W32-W37.

87. Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006,** *34*, D668-D672.

88. Zhang, S.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *J Chem. Inf. Model.* **2006,** *46*, 1984-1995.

89. Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness,

Agrochemical-likeness, and Enzyme Inhibition Predictions. *Journal of chemical information and computer sciences* **2003,** *43*, 2048-2056.

90. Netzeva, T. I.; Gallegos, S. A.; Worth, A. P. Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. *Environ. Toxicol. Chem.* **2006,** *25*, 1223-1230.

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

92. Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry* **2006,** *49*, 6789-6801.

93. Mittal, R. R.; McKinnon, R. A.; Sorich, M. J. Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization. *Journal of Chemical Information and Modeling* **2009,** *49*, 1810-1820.

94. Venkatraman, V.; Pe¦ürez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *Journal of Chemical Information and Modeling* **2010,** *50*, 2079-2093.

95. von Korff, M.; Freyss, J.; Sander, T. Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. *Journal of Chemical Information and Modeling* **2009,** *49*, 209-231.

96. Sunset Molecular Discovery, L. Wombat. http://sunsetmolecular.com/products/?id=4 . 2008.
Ref Type: Electronic Citation

97. MDDR.SYMYX technologies.
http://www.mdl.com/products/knowledge/drug_data_report/index.jsp . 2009.
Ref Type: Electronic Citation

98. Pipeline Pilot. 2010. San Diego, CA, USA, Accelerys, Inc.
Ref Type: Computer Program

99. Zsoldos, Z.; Reid, D.; Simon, A.; Sadjad, S. B.; Johnson, A. P. eHiTS: A new fast, exhaustive flexible ligand docking system. *Journal of Molecular Graphics and Modelling* **2007,** *26*, 198-212.

100. DRAGON. http://www.disat.unimib.it/chm/Dragon.htm . 2008.
Ref Type: Electronic Citation

101. Savas, S.; Briollais, L.; Ibrahim-zada, I.; Jarjanazi, H.; Choi, Y. H.; Musquera, M.; Fleshner, N.; Venkateswaran, V.; Ozcelik, H. A Whole-Genome SNP Association

Study of NCI60 Cell Line Panel Indicates a Role of Ca²⁺ Signaling in Selenium Resistance. *PLoS ONE* **2010,** *5*, e12601.

102. Huang, Y.; Blower, P.; Liu, R.; Dai, Z.; Pham, A. N.; Moon, H.; Fang, J.; Sad+¬e, W. Chemogenomic Analysis Identifies Geldanamycins as Substrates and Inhibitors of ABCB1. *Pharmaceutical Research* **2007,** *24*, 1702-1712.

103. Shankavaram, U. T.; Reinhold, W. C.; Nishizuka, S.; Major, S.; Morita, D.; Chary, K. K.; Reimers, M. A.; Scherf, U.; Kahn, A.; Dolginow, D.; Cossman, J.; Kaldjian, E. P.; Scudiero, D. A.; Petricoin, E.; Liotta, L.; Lee, J. K.; Weinstein, J. N. Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics* **2007,** *6*, 820-832.

104. Shankavaram, U.; Varma, S.; Kane, D.; Sunshine, M.; Chary, K.; Reinhold, W.; Pommier, Y.; Weinstein, J. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* **2009,** *10*, 277.

105. Golub, G.; Reinsch, C. Singular value decomposition and least squares solutions. *Numerische Mathematik* **1970,** *14*, 403-420.

106. Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98*, 5116-5121.

107. Bucci, B.; D'Agnano, I.; Amendola, D.; Citti, A.; Raza, G. H.; Miceli, R.; De Paula, U.; Marchese, R.; Albini, S.; Felsani, A.; Brunetti, E.; Vecchione, A. Myc Down-Regulation Sensitizes Melanoma Cells to Radiotherapy by Inhibiting MLH1 and MSH2 Mismatch Repair Proteins. *Clinical Cancer Research* **2005,** *11*, 2756-2767.

108. Liu, X.; Li, P.; Widlak, P.; Zou, H.; Luo, X.; Garrard, W. T.; Wang, X. The 40-kDa subunit of DNA fragmentation factor induces DNA fragmentation and chromatin condensation during apoptosis. *Proceedings of the National Academy of Sciences of the United States of America* **1998,** *95*, 8461-8466.

109. McDonald, J. M.; Dunmire, V.; Taylor, E.; Sawaya, R.; Bruner, J.; Fuller, G.; Aldape, K.; Zhang, W. Attenuated Expression of DFFB is a Hallmark of Oligodendrogliomas with 1p-Allelic Loss. *Molecular Cancer* **2005,** *4*, 35.

110. Nishiyama, M.; Ozturk, M.; Frohlich, M.; Mafune, K. i.; Steele, G.; Wands, J. R. Expression of Human +¦-Actinin in Human Hepatocellular Carcinoma. *Cancer Research* **1990,** *50*, 6291-6294.

111. Bae, Y. H.; Ding, Z.; Zou, L.; Wells, A.; Gertler, F.; Roy, P. Loss of profilin-1 expression enhances breast cancer cell motility by Ena/VASP proteins. *J. Cell. Physiol.* **2009,** *219*, 354-364.

112. Wittenmayer, N.; Jandrig, B.; Rothkegel, M.; Schluter, K.; Arnold, W.; Haensch, W.; Scherneck, S.; Jockusch, B. M. Tumor Suppressor Activity of Profilin Requires a Functional Actin Binding Site. *Mol. Biol. Cell* **2004,** *15*, 1600-1608.

113. Hirota, T.; Kunitoku, N.; Sasayama, T.; Marumoto, T.; Zhang, D.; Nitta, M.; Hatakeyama, K.; Saya, H. Aurora-A and an Interacting Activator, the LIM Protein Ajuba, Are Required for Mitotic Commitment in Human Cells. Cell 114[5], 585-598. 9-5-2003.
Ref Type: Abstract

114. Benzinger, A.; Muster, N.; Koch, H. B.; Yates, J. R.; Hermeking, H. Targeted Proteomic Analysis of 14-3-3-é, a p53 Effector Commonly Silenced in Cancer. *Molecular & Cellular Proteomics* **2005,** *4*, 785-795.

115. Strell, C.; Entschladen, F. Extravasation of leukocytes in comparison to tumor cells. *Cell Communication and Signaling* **2008,** *6*, 10.

116. Wang, J.; Jing, Z.; Zhang, L.; Zhou, G.; Braun, J.; Yao, Y.; Wang, Z. Z. Regulation of acetylcholine receptor clustering by the tumor suppressor APC. *Nat Neurosci* **2003,** *6*, 1017-1018.

117. Desgrosellier, J. S.; Cheresh, D. A. Integrins in cancer: biological implications and therapeutic opportunities. *Nat Rev Cancer* **2010,** *10*, 9-22.

118. Partek, I. Partek® Genomics Suite™. [6.5]. 2010. St Louis, Partek Inc.
Ref Type: Computer Program

119. Brown, F. Editorial opinion: chemoinformatics - a ten year update. *Curr. Opin. Drug Discov. Devel.* **2005,** *8*, 298-302.

120. Olsson, T.; Oprea, T. I. Cheminformatics: a tool for decision-makers in drug discovery. *Curr. Opin. Drug Discov. Devel.* **2001,** *4*, 308-313.

121. Varnek, A.; Tropsha, A. *Cheminformatics Approaches to Virtual Screening*; RSC: London, 2008.

122. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010,** *50*, 1189-1204.

123. ChemAxon. ChemAxon JChem (http://www.chemaxon.com) . 2010.
Ref Type: Electronic Citation

124. Talete s.r.l. Dragon. [5.4.2006]. 2007. Milan, Italy.
Ref Type: Computer Program

125. MolconnZ. http://www.edusoft-lc.com/molconn/ . 2006.
Ref Type: Electronic Citation

126. Kier, L. B.; Hall, L. H. *Molecular connectivity in structure–activity analysis*; Wiley: New York, 1986.

127. Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.

128. Randic, M. Characterization of Molecular Branching. *Journal of the American Chemical Society* **1975,** *97*, 6609-6615.

129. Kier, L. B. Inclusion of Symmetry As A Shape Attribute in Kappa-Index Analysis. *Quantitative Structure-Activity Relationships* **1987,** *6*, 8-12.

130. Kier, L. B. A Shape Index from Molecular Graphs. *Quantitative Structure-Activity Relationships* **1985,** *4*, 109-116.

131. Hall, L. H.; Kier, L. B. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quantitative Structure-Activity Relationships* **1990,** *9*, 115-131.

132. Kier, L. B.; Hall, L. H. *Molecular structure description: The electrotopological state*; Academic Press: New York, 1999.

133. Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. E-state fields: Applications to 3D QSAR. *Journal of Computer-Aided Molecular Design* **1996,** *10*, 513-520.

134. Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State - Structure Information at the Atomic Level for Molecular Graphs. *Journal of Chemical Information and Computer Sciences* **1991,** *31*, 76-82.

135. Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State - An Atom Index for Qsar. *Quantitative Structure-Activity Relationships* **1991,** *10*, 43-51.

136. Kier, L. B.; Hall, L. H. A Differential Molecular Connectivity Index. *Quantitative Structure-Activity Relationships* **1991,** *10*, 134-140.

137. Petitjean, M. Applications of the Radius Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical-Compounds. *Journal of Chemical Information and Computer Sciences* **1992,** *32*, 331-337.

138. Wiener, H. J. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947,** *69*, 17-20.

139. Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* **1947,** *15*, 419-420.

140. Shannon, C.; Weaver, W. *In mathematical theory of communication.*; University of Illinois: 1949.

141. Bonchev, D.; Mekenyan, O.; Trinajstic, N. Isomer Discrimination by Topological Information Approach. *Journal of Computational Chemistry* **1981,** *2*, 127-148.

142. MOE. Chemical Computing Group [2007.09]. 2008.
Ref Type: Electronic Citation

143. Kier, L. B. Inclusion of Symmetry As A Shape Attribute in Kappa-Index Analysis. *Quantitative Structure-Activity Relationships* **1987,** *6*, 8-12.

144. Petitjean, M. Applications of the Radius Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical-Compounds. *Journal of Chemical Information and Computer Sciences* **1992,** *32*, 331-337.

145. Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chemical Physics Letters* **1982,** *89*, 399-404.

146. Balaban, A. T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta.* **1979,** 355-375.

147. MOE. Chemical Computing Group [2007.09]. 2008.
Ref Type: Electronic Citation

148. MOE. Chemical Computing Group [2007.09]. 2008.
Ref Type: Electronic Citation

149. Zheng, W.; Tropsha, A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000,** *40*, 185-194.

150. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953,** *21*, 1087-1092.

151. Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983,** *220*, 671-680.

152. Forrest, S. Genetic algorithms: principles of natural selection applied to computation. *Science* **1993,** *261*, 872-878.

153. Vapnik, V. The Nature of Statistical Learning Theory. 1995. Springer Verlag.
Ref Type: Generic

154. Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. *Journal of chemical information and computer sciences* **2003,** *43*, 2048-2056.

155. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. An accurate QSPR study of O-H bond dissociation energy in substituted phenols

based on support vector machines. *Journal of chemical information and computer sciences* **2004,** *44*, 669-677.

156. Chang, C.-C.; lin, chih-jen. LIBSVM : a library for support vector machines. 2001. Ref Type: Computer Program

157. Breiman, L. Random forests. *Machine Learning* **2001,** *45*, 5-32.