Published in final edited form as:

J Med Chem. 2012 June 28; 55(12): 5704–5719. doi:10.1021/jm2011657.

A Chemocentric Informatics Approach to Drug Discovery: Identification and Experimental Validation of Selective Estrogen Receptor Modulators as ligands of 5-Hydroxytryptamine-6 Receptors and as Potential Cognition Enhancers

Rima Hajjo[‡], Vincent Setola[†], Bryan L. Roth^{‡,†}, and Alexander Tropsha^{‡,*}

[‡]Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

[†]National Institute of Mental Health Psychoactive Drug Screening Program, Department of Pharmacology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Abstract

We have devised a chemocentric informatics methodology for drug discovery integrating independent approaches to mining biomolecular databases. As a proof of concept, we have searched for novel putative cognition enhancers. First, we generated Quantitative Structure-Activity Relationship (QSAR) models of compounds binding to 5-hydroxytryptamine-6 receptor (5HT₆R), a known target for cognition enhancers, and employed these models for virtual screening to identify putative 5-HT₆R actives. Second, we queried chemogenomics data from the Connectivity Map (<http://www.broad.mit.edu/cmap/>) with the gene expression profile signatures of Alzheimer's disease patients to identify compounds putatively linked to the disease. Thirteen common hits were tested in 5-HT₆R radioligand binding assays and ten were confirmed as actives. Four of them were known selective estrogen receptor modulators that were never reported as 5-HT₆R ligands. Furthermore, nine of the confirmed actives were reported elsewhere to have memory-enhancing effects. The approaches discussed herein can be used broadly to identify novel drug-target-disease associations.

Introduction

Target-oriented drug discovery is one of the most popular modern drug discovery approaches^{1–5}. Target-oriented approaches rely on established functional associations between activation or inhibition of a molecular target and a disease. Modern genomics approaches including gene expression profiling, genotyping, genome-wide association, and mutagenesis studies continue to serve as useful sources of novel hypotheses linking genes (proteins) and diseases and providing novel putative targets for drug discovery.

In recent years, functional genomics approaches have been increasingly complemented by chemical genomics^{6–11} i.e., large scale screening of chemical compound libraries in multiple biological assays^{12–16}. The resulting data (either generated within chemical genomics

*To whom correspondence should be addressed: Alexander Tropsha, CB #7360, Beard Hall, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7360, Tel: 919-966-2955, Fax: 919-966-0204, alex_tropsha@unc.edu.

Supporting Information Available: 5-HT₆R datasets, virtual screening hits, cmap negative connections with Alzheimer's disease gene signatures, connectivity scores, and further computational details. LC/MS purity spectra for all tested compounds are also provided. This material is available free of charge via the Internet at <http://pubs.acs.org>.

centers or collected and curated from published literature) have been deposited in many public and private databases such as the NIMH Psychoactive Drug Screening Program K_i Database (K_i -DB)¹⁷, PubChem¹⁸, ChEMBL¹⁹, WOMBAT²⁰ and others (reviewed by²¹).

Various *in silico* techniques have been exploited for analyzing target-specific biological assay data. A recent publication by Kortagere and Ekins²² could serve as a good summary of most common target-oriented computational drug discovery approaches including: (1) structure based virtual screening (docking and scoring) using either experimentally characterized (with X-ray or NMR) or predicted by homology modeling structure of the target protein, (2) chemical similarity searching using known active compounds as queries, (3) pharmacophore based modeling and virtual screening, (4) quantitative structure-activity relationship (QSAR) modeling, and (5) network or pathway analysis.

Data resulting from large-scale gene or protein expression or metabolite profiling (often collectively referred to as 'omics' approaches^{23–26}) can be explored not only for specific target identification but also in the context of systems pharmacology to identify networks of genes (or proteins) that may collectively define a disease phenotype. For example, 'omics' data can be used to query genes or proteins, or post-translationally modified states of proteins that are over- (or under-) expressed in patients suffering from a particular disease. These types of data can be found in a number of public repositories such as the Gene Expression Omnibus (GEO)^{27;28}, GEOmetadb²⁹, the Human Metabolome Database (HMDB)^{30;31}, Kinase SARfari³², the Connectivity Map (cmap)^{33;34}, the Comparative Toxicogenomics Database (CTD)³⁵, STITCH^{36;37}, GenBank^{38;39}, and others. Importantly, many of these databases integrate, in some way, chemical effects on biological systems providing an opportunity to explore diverse computational approaches, individually or in parallel, to modeling and predicting the relationships between drug structure, its bioactivity profile in short term biological assays, and its effects *in vivo*.

Recently, a group of scientists at the Broad Institute established the Connectivity Map (cmap) database to catalog the biological responses of a large number of diverse chemicals in terms of their gene expression profiles³³. Indeed, insights into disease pathology and underlying mechanisms can be revealed by the disease 'gene signature', i.e., those genes whose expression varies consistently between patients and healthy individuals (controls)⁴⁰. Gene-expression profiling has been often applied to elucidate the mechanisms underlying the roles of biological pathway in a disease^{41;42}, reveal arcane subtypes of a disease,^{43;44} and estimate cancer prognosis^{45;46}. At the same time, the treatment of cultured human cells with chemical compounds that target a disease can produce a drug related 'gene signature', i.e., differential expression profile of genes in response to the chemical^{40;47–49}. It has been shown that examining the correlations between gene expression profiles characteristic of a disease and those modulated by drugs may lead to novel hypotheses linking chemicals to either etiology or treatments for a disease^{33;43;50–56}.

The cmap database provides an unusual but intriguing example of what we shall call a *chemocentric* 'omics' database and methodology for generating independent and novel drug discovery hypotheses. Indeed, there exists a wealth of information buried in the biological literature and numerous specialized chemical databases^{17–20;57} linking chemical compounds and biological data (such as targets, genes, experimental biological screening results; cf.⁵⁸). The chemocentric exploration of these sources, either individually or in parallel opens up vast possibilities for formulating novel drug discovery hypotheses concerning the predicted biological or pharmacological activity of investigational chemical compounds or known drugs. The integration and cross-validation of such independent structural hypotheses can increase the quality of the final hit list of predicted actives.

Herein, we describe a novel integrative *chemocentric informatics* approach to drug discovery that integrates computational hits generated from independent analysis of both traditional target-specific assay data and those resulting from large scale genomics and chemical genomics studies. As a proof of concept, we have focused on the Alzheimer's disease as one of the most debilitating neurodegenerative diseases with complex etiology and polypharmacology. We have considered and cross-examined two independent but complementary approaches to the discovery of novel putative anti-Alzheimer's drugs. First, we have employed a traditional target-oriented cheminformatics approach to discovering anti-Alzheimer's agents. We have built QSAR models of ligands binding to 5-hydroxytryptamine-6 receptor (5-HT₆R). It has been shown that 5-HT₆R antagonists can produce cognitive enhancement in animal models⁵⁹, and it has been suggested that this receptor may be a potential target for treating cognitive deficits in Alzheimer's disease⁶⁰. We have then used models developed with the rigorous predictive QSAR modeling workflow established and implemented in our laboratory⁶¹ for virtual screening (VS) of the World Drug Index database (WDI)⁵⁷ and DrugBank⁶² to identify putative cognition enhancing agents as compounds predicted to interact with 5-HT₆R. Second, we have explored (chemo)genomic data available from the cmap project^{33;34} to link chemical compounds and the Alzheimer's disease without making explicit hypotheses about target-specific mechanisms of action, i.e., treating Alzheimer's disease as a complex polypharmacological disease.

We then cross-examined and combined common hits regarded as structural hypotheses resulting from both approaches towards common integrated hits supported by two independent lines of computationally-based evidence. Thirteen common hits (Figure 1) were tested in 5-HT₆R binding assays using the resources of the NIMH Psychoactive Drug Screening Program (PDSP)¹⁷ and ten were confirmed experimentally as actives. Unexpectedly, we found that the confirmed actives included several selective estrogen receptor modulators (SERMs) that were never reported earlier as 5-HT₆R ligands suggesting that they may be potential cognitive enhancers. Indeed, we have identified clinical evidence in biomedical literature in support of this hypothesis. We believe that approaches discussed in this study can be applied to a large variety of systems to identify novel drug-target-disease associations.

Materials and Methods

Integrative Chemocentric Informatics Approach

We have devised an integrative workflow focused on the discovery of new drug candidates and finding new uses for existing drugs by integrating predictions generated from different data types and methods. Currently, the workflow (Figure 2) incorporates three major components: (1) a module for QSAR-based VS of chemical libraries to identify new ligands for target proteins, (2) a network-mining module to identify small molecule therapeutics for specific diseases without necessarily knowing the underlying target-specific mechanism; this module explicitly relies on cmap^{33;34}, an external online database (www.broadinstitute.org/cmap/) that links the effects of different drugs and diseases using gene expression profiles, and (3) ChemoText⁵⁸, an in-house repository of relationships between chemicals, diseases, proteins, and biological processes. The first two modules have been employed extensively for studies reported herein.

We start our study with identifying established disease-target associations (e.g., 5-HT₆R is implicated in treating Alzheimer's disease). Then we mine the biological literature and specialized databases to extract ligands known to interact with the biological target of interest. Activity data could be either binding affinities (K_i values) or functional data (IC₅₀ values for agonists and antagonists). Binding and functional data could be either continuous

(e.g., K_i and IC_{50} values) or categorical (e.g., active vs. non-active or agonist vs. antagonist) in nature. At this stage we use our QSAR-based VS module (see Figure 3; predictive QSAR workflow) to generate robust predictive QSAR models that can be employed for VS of chemical libraries to derive new hypotheses about putative actives (agonists or antagonists).

In parallel, we mine the biological literature for gene signatures associated with the disease and/or for all related protein targets implicated in the disease state. We use these disease related genes and proteins to query specialized databases to extract information about disease-protein (gene)-chemical connections. For example, we use disease gene signatures to query the cmap for putative treatments, and we use related proteins to query ChemoText for related chemicals to establish new disease-protein (gene)-chemical connections. After a thorough analysis of all data, we select hit compounds that are expected to be novel treatments for the disease (cf. Figure 2).

Finally, we integrate hypotheses derived from the QSAR-based VS approach with those derived from text/network mining. The common structural hits identified by both approaches are considered for further experimental validation. We assume that the quality of the final structural hypotheses resulting from independent approaches to knowledge mining in chemocentric databases is intrinsically better than that in any computational hit generated in respective independent studies.

Databases and Datasets

PDSP K_i -DB—PDSP K_i -DB¹⁷ (<http://pdsp.med.unc.edu/pdsp.php>) includes published binding affinities (K_i) of drugs and chemical compounds for receptors, neurotransmitter transporters, ion channels, and enzymes. It currently lists more than 47000 K_i values for more than 700 molecular targets. K_i -DB represents a curated, fully searchable database of both published data and data internally-derived from the NIMH-PDSP. The experimental data for Alzheimer's disease related target 5-HT₆R were extracted from the PDSP K_i -DB available in the public domain. The complete 5-HT₆R dataset included binding affinity data for 250 compounds

World Drug Index (WDI)—WDI⁵⁷ is an authoritative database for marketed and developmental drugs providing information about internationally recognized drug names, synonyms, trade names, trivial names, trial preparation codes, compound structures, and activity data. Herein, we used WDI for QSAR-based VS to identify putative 5HT₆R ligands.

DrugBank—DrugBank⁶² (<http://www.drugbank.ca>) is a unique bioinformatics and cheminformatics resource that combines detailed drug data (i.e., chemical, pharmacological, and pharmaceutical) with comprehensive drug target information (i.e., sequence, structure, and pathway). Currently, the database contains nearly 4800 drug entries. Herein, we used DrugBank for virtual screening using QSAR models to identify putative 5HT₆R ligands among known drugs

PubChem—PubChem¹⁸ is a public repository of chemical structures and their activities obtained from a variety of biological assays. The PubChem compound repository presently contains more than 30 million unique structures with biological property information provided for many of the compounds. Herein, we used PubChem to obtain all chemical structures for our datasets in SDF file format.

The Connectivity Map (cmap)—The cmap^{33;34} (<http://www.broadinstitute.org/cmap/>) is a unique database for using chemical genomics in drug discovery framework. It provides researchers with a systematic solution for the discovery of the functional connections

between drugs, genes, and diseases. The database used in this study (cmap build 02) included 7056 genome-wide expression profiles representing 6100 individual treatment instances with 1309 bioactive small molecules (i.e., drugs and other biologically active compounds). All gene expression profiles included in the cmap were derived from treating cultured human cells (MCF7, PC3, HL60, SKMEL5, HepG2, SHSY5Y) with chemical compounds.

NetAffx—NetAffx^{63;64} (<http://www.affymetrix.com>) gene ontology mining tool is a web-based, interactive tool that permits traversal of the gene ontology graph in the context of microarray data. It accepts a list of Affymetrix probe sets and renders a gene ontology graph as a heat map colored according to significance measurements. It also details and annotates probe sets on Affymetrix GeneChip microarrays. In this study we used NetAffx to populate our disease gene signatures with Affymetrix U133A probe sets.

ChemoText—ChemoText⁵⁸ is an in-house repository of chemical entities, and activity terms (indicating biological effects) extracted from annotations provided in Medline records. This resource has different applications in drug discovery projects. First, we can use ChemoText in a discovery-mode to formulate independent hypotheses about chemical-disease associations according to Swanson's ABC rule.⁶⁵ Secondly, we can use it as an information retrieval tool to gather relevant data about chemical-protein (or gene)-disease connections derived from biomedical literature. In this study, we used ChemoText to retrieve all available biological information about the final computational hits predicted by our integrative approach. This analysis helped us in assessing the novelty of the produced hypotheses and in validating some of them.

Computational Methods

1. QSAR Modeling and QSAR-based Virtual Screening

Preprocessing of the Dataset: We used a workflow for chemical data curation that was developed in our lab and published recently⁶⁶. First, all molecules were “washed” using the “Wash Molecules” application in MOE (v.2007.09)⁶⁷. Using this tool, we processed chemical structures by carrying out several standard operations including 2D depiction layout, hydrogen correction, salt and solvent removal, chirality and bond type normalization (all details can be found in the MOE manual⁶⁷). Second, we used ChemAxon's Standardizer (v. 5.2.6)⁶⁸ to harmonize the representation of aromatic rings. Third, we checked all normalized molecular structures in the dataset for duplicated compounds using the “Sort Unique Entries” application in MOE. Our analysis resulted in the detection and removal of 56 duplicate chemical entries leaving 194 unique normalized molecular structures. These 194 unique, organic compounds, including 102 actives and 94 non-actives (see Table S1 of Supporting Information) were used for binary QSAR studies. We assigned the ‘activity’ class for each compound based on its K_i value(s) obtained from the PDSP and according to PDSP specifications as reported at the PDSP website (<http://pdsp.med.unc.edu/>). Compounds with K_i values ≥ 10 μ M were considered non-actives and assigned to class 0, whereas compounds with K_i values < 10 μ M were considered actives and assigned to class 1.

Dataset Division for Model Building and Validation: Following our predictive QSAR modeling workflow (summarized in a recent review⁶¹) all QSAR models generated to classify 5-HT₆R actives vs. non-actives were validated by predicting both test and external validation sets. The original dataset of 194 compounds (102 actives and 92 non-actives) was randomly split into 5 different subset of nearly equal size to allow for external 5-fold cross validation (CV)^{69;70} where the dataset compounds were ranked from 1 to n (n = total number of compounds in the dataset) then the first compound went to the external set and

the following 4 compounds were included in the modeling set. Then the 6th compound went to the external set and the following 4 compounds went to the modeling set. This process was continued until all compounds were divided between the external and the modeling sets. In this protocol, each subset including 20% of the original dataset was systematically employed as the external validations set while the remaining 80% of the compounds constituted the modeling set.

Another level of internal validation was achieved by comparing model performance for training and test sets. This approach is always employed as part of our predictive QSAR modeling workflow^{61;71} to emphasize the fact that training-set-only modeling is not sufficient to obtain reliable models that are externally predictive⁷². Thus, for each collection of descriptors, the modeling sets (each including 80% of the original dataset) were further partitioned into multiple chemically diverse training and test sets of different sizes using the Sphere Exclusion method implemented in our laboratory⁷³. Only models with external predictive accuracy for test sets above certain threshold (see Selection and Validation of QSAR Models in Supporting Information) were retained for the consensus prediction of the external validation sets. Finally, models that demonstrated the highest predictive power for both evaluation sets were used in consensus fashion for virtual screening of external compound libraries. The model building and validation approach is illustrated schematically in Figure 3.

Combinatorial QSAR Modeling: Two QSAR modeling approaches of different nature were used concurrently to generate classification models for 5-HT₆R actives vs. non-actives (Figure 3). The first approach relied on *k*-nearest neighbor (*k*NN) model optimization method combined with Dragon descriptors, and the second employed classification based on association (CBA) and subgraphs (SG) descriptors. All details about QSAR methods, evaluation of generated models, and consensus prediction are available in the Supporting Information.

Virtual Screening: To identify putative ligands, validated consensus *k*NN-Dragon models generated for 5-HT₆R ligands were used for virtual screening of both the 59000 molecules within the WDI⁵⁷ chemical library and 1300 DrugBank⁶² compounds included in the cmap database. The identified hits (by consensus agreement between all accepted *k*NN-Dragon models) were then evaluated additionally using CBA-SG classifier when it was a need to reduce the size of the VS library generated with *k*NN-Dragon models.

2. Biological Network Mining

Querying the cmap with Alzheimer's Disease Gene Signatures: The cmap^{33;34} was used to discover unexpected connections between chemicals, genes and the Alzheimer's disease by generating a detailed map that links gene patterns associated with Alzheimer's to corresponding patterns produced by drug candidates and a variety of genetic perturbations included in the cmap database. The effects of different drugs and diseases are described using "genomic signatures" — the full complement of genes that are turned on and off by a particular drug or disease. We start by querying the online database (cmap: <http://www.broadinstitute.org/cmap/>) with gene signatures characteristic of the Alzheimer's disease. Then, a computer program, that uses sophisticated pattern-matching methods, matches the barcodes based on the patterns shared among Alzheimer's gene signature and drugs included in the cmap.

Alzheimer's Disease Gene Signatures: In order to query the cmap, a disease gene signature should exist. Two lists of genes are required to perform the query: a list of up-regulated genes and a list of down-regulated genes characteristic of a disease. Query

signatures can be obtained from two major sources: (1) biological literature: gene signatures of diseases can be extracted through the National Library of Medicine's PubMed system (<http://www.ncbi.nlm.nih.gov/pubmed>), (2) GEO^{27,28} database: a gene expression/molecular abundance repository supporting MIAME⁷⁴ (Minimum Information About a Microarray Experiment) compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. For the purposes of this study, two independent reports of gene-expression changes in brain tissues from Alzheimer's patients were used to derive gene signatures (i.e., lists of genes up- and down- regulated in Alzheimer's disease) to query the cmap. Signature 1 (from hippocampus) consisted of 40 genes reported by Hata, R. *et al*⁷⁵, and signature 2 (from cerebral cortex) consisted of 25 genes reported by Ricciarelli, R. *et al*⁷⁶. NetAffx was then used to map gene symbols and Unigene identifiers to populate gene signature lists with Affymetrix U133A probe sets to query the cmap.

3. Hypothesis Integration—We cross-examined and integrated structural hypotheses generated independently from both QSAR-based VS and biological network mining efforts to identify and accept common hits only. This step of hypotheses integration was based on structural identity comparisons. All chemical structures of cmap compounds were retrieved from DrugBank⁶² using their DrugBank identifiers. Identical structures only were then accepted for further analysis. All chemical structures labeled as identical were also subjected to a manual curation step where structures and names of the chemical compounds were compared in different databases to make sure they both refer to the same chemical entity. Common hits were then considered for further experimental validation.

4. Experimental Validation in Radioligand Binding Assays—Final common hit compounds from QSAR-based VS and cmap negative connections with Alzheimer's were purchased and submitted to PDSP for experimental target validation. The experimental details of radioligand binding assays are available at the PDSP website¹⁷.

Results and Discussion

QSAR Modeling of 5-HT₆R Actives vs. Non-actives

kNN with Dragon descriptors was employed to classify modeling set compounds into 5-HT₆R actives vs. non-actives. As part of 5-fold cross validation process, the dataset was divided, into five subsets of nearly equal size. Four parts (selected systematically) formed modeling sets with 155 compounds each and the fifth part, containing 39 compounds, was considered as a validation set. Each of the five modeling sets derived to collect 5-fold CV statistics was additionally subdivided into multiple training and test sets (28-40 divisions) using the Sphere Exclusion algorithm as described in Methods. Multiple QSAR models were generated independently for all training sets and applied to the test sets. Generally, we accept models with CCR values above or equal to 0.70 for both the training and test sets.

However, because we were able to generate thousands of acceptable models, we used more conservative criteria (i.e., CCR_{train} and CCR_{test} above or equal to 0.90) for model selection to predict external compounds. Basically, each train/test split, kNN will generate at least one model. Increasing the "Number of Runs" option will multiply the number of generated models. The "descriptors per model" parameter can again multiply the number of generated models; if kNN tries to make a model with 5 descriptors, a model with 6 descriptors, and a model with 7 descriptors in each case, the total number of models will be tripled. So, to calculate the number of models that will be generated: (Number of train-test splits) * (Number of Runs) * (Number of different values "Descriptors per Model" can take). This is an important thing to consider before starting a modeling run: Too few models will be unlikely to create a good predictor, while too many models can take a very long time to

generate. If the dataset is a high quality dataset (i.e. there are very good distinctive features between actives and inactives) then we may be able to generate thousands of acceptable models. However, these models differ in their statistical parameters. See “Selection and validation of QSAR models” in Supporting Information.

Results of Y-randomization tests confirmed that *k*NN-Dragon classification models with CCR_{train} and CCR_{test} values above or equal to 0.90 were robust. None of the models with randomized class labels of the training set compounds had CCR_{train} and CCR_{test} above 0.65 or CCR_{evs} above 0.55 for any split.

The CBA method was used to classify the dataset using SG descriptors. The modeling sets (described above) were used to build the classifier in CBA⁷⁷ using an initial pool of about 400 SG descriptors. The classifier gave an average CCR_{train} of 0.92 (i.e., the average resulted from five different tests). Then, the external validation sets consisting of 39 compounds were used to assess the robustness of the classifier. The average CCR_{test} was 0.78, which is not as high the CCR value for the training set, but is still statistically acceptable.

Clearly, *k*NN (mean CCR_{evs} = 0.92) performed better than CBA (mean CCR_{evs} = 0.78) on the external validation sets. Therefore, we chose to use *k*NN-Dragon models for VS of external drug libraries. Nevertheless, we maintained CBA-SG models as an additional filter to suggest smaller sets of compounds as 5-HT₆R putative actives selected from the list of virtual hits obtained with *k*NN-Dragon models and therefore predicted by both models as putative actives.

QSAR-based Virtual Screening

Since our models proved reasonably accurate based on external validation sets, we used the best models to mine two external databases of approved and potential drugs for putative 5-HT₆R ligands. The use of applicability domain assures reliable predictions by the models. Therefore, we used two types of applicability domains in the virtual screening of compound databases. First, we used a global applicability domain that acted as a filter and ensured some level of global similarity between the predicted compounds and the compounds in the modeling set. Second, we defined a local applicability domain for each of the individual classification models.

We first screened the WDI database⁵⁷ of about 59000 compounds (approved or investigational drugs) (Figure 5). This original collection had many duplicates (i.e., many salt forms for the same chemical entity), and these duplicates were removed using MOE. We also removed all compounds that were duplicates of those molecules that were included in our modeling and external validation sets. Dragon descriptors were generated for the remaining 46859 unique compounds in the database; of these, 9732 compounds were excluded because Dragon was unable to calculate at least one of the descriptors generated for the modeling set. The remaining 37127 compounds were then subjected to a global applicability domain filter for the modeling set using a strict Z cutoff of 0.5 (which formally places the allowed pairwise distance threshold at the mean of all pairwise distance distribution for the training set plus one-half of the standard deviation). Then, all *k*NN-Dragon models with CCR_{train} and CCR_{test} above or equal to 0.70 were employed in consensus fashion to predict 1500 compounds remaining after several filtering steps, which resulted in the identification of the 600 predicted actives. In an effort to reduce the number of hits, we have generated SG descriptors for these 600 molecules and applied the CBA-SG classifier which filtered out half of these compounds, leaving 300 compounds as putative actives for 5-HT₆R. None of these hits was tested in this study since we explicitly focused on compounds from DrugBank⁶² that were employed in the cmap project. These VS hits

from WDI should be viewed as hypothetically active compounds awaiting the experimental confirmation; the list of the top scoring hits is included with the Supporting Information (Table S2).

Additionally, we screened 1300 DrugBank compounds included in the cmap database. Dragon descriptors were computed for 1273 unique compounds. These compounds were then subjected to a global applicability domain filter for the modeling set using a strict Z cutoff of 0.5. Consequently, we placed the allowed pairwise distance threshold at the mean of all pairwise distance distribution for the training set plus one-half of the standard deviation which resulted in 577 predictions within the applicability domain. Next, validated consensus *k*NN-Dragon models (i.e., all models with CCR_{train} and CCR_{test} above or equal to 0.90) were used to predict these 577 compounds, resulting in the identification of 140 unique compounds predicted to be 5-HT₆R actives. We did not apply the CBA-filter here because, for the subsequent integration with the cmap mining results, we wanted to explore a larger set of all 140 compound hits (i.e., putative 5-HT₆R actives) included in the cmap datasets predicted by *k*NN-Dragon models.

Searching the Connectivity Map for Potential Anti-Alzheimer's Agents

We used two gene signatures for the Alzheimer's disease (designated as S1 and S2) to query the cmap database in an attempt to link genes associated with the disease to potential therapeutic agents. These two signatures were based on two independent rank-ordered gene lists provided by two different Gene Set Enrichment Analysis (GSEA) studies^{75;76}. The two disease signatures were compared with predefined signatures of therapeutic compounds included in the cmap and ranked according to a connectivity score (ranging from +1 to -1), representing relative similarity to the disease gene lists. The connectivity score itself is derived using a nonparametric, rank-based, pattern-matching strategy based on the Kolmogorov-Smirnov statistic⁷⁸. Connectivity scores are calculated using the online tools available at the cmap (<http://www.broadinstitute.org/cmap/>). All instances in the database are then ranked according to their connectivity scores; those at the top (+) are most strongly correlated to the query signature and looked at as disease causes, and those at the bottom (-) are most strongly anticorrelated and considered as possible therapeutics.

The majority of chemicals included in the cmap database are represented by multiple independent replicates. Most compounds are profiled in three different cell lines, some at different concentrations. These are called 'instances' for the same chemical and defined as "a treatment and control pair and the list of probe sets ordered by their extent of differential expression between this treatment and control pair"⁷⁹. The instance is the basic unit of data and metadata in cmap. Instances of the same compound might have similar or dissimilar connectivity scores with the query signature. We have higher confidence in the derived connections when gene signatures are conserved across diverse cell types and experimental settings. However, Lamb and colleagues^{33;34} indicated that the nonconsistent scoring of different instances of the same chemical may represent either (1) a cellular-context dependent difference in activity, (2) a concentration-discriminated effect, or (3) poor reproducibility between replicates. Therefore, 'best' connections are those where multiple, autonomous instances of the same chemical have consistently high (or low) scores. However, inconsistently scoring compounds should not necessarily be dismissed since their significance as potential treatments for a disease can be boosted by additional evidence, such as predictions from QSAR models.

In this study, we were interested in compounds whose chemogenomics profiles were negatively correlated with the Alzheimer's disease gene signatures. Hits with statistically significant, negative connectivity scores could be potential treatments for the Alzheimer's disease; however, the list of negatively correlated molecules might be long and must be

analyzed carefully before suggesting hypotheses of possible mechanisms for controlling or mediating the disease. Examples of top negative connections with both signatures S1 and S2 are shown in Tables 1 and 2, respectively (see Tables S2 and S3 in Supporting Information for full listings of connections and their connectivity scores). Although the two gene signatures (i.e., for the Alzheimer's disease) used to query the cmap shared no common genes, both queries resulted in a common list of negative connections that were given a higher confidence in our studies. All chemical structures for each chemical compound included in the cmap were obtained from the DrugBank⁶² and mapped based on the DrugBank identifiers provided by the cmap database.

Hypothesis Generation: Integrating Independent Hypotheses from QSAR-Based VS and cmap Analysis

We combined hypotheses produced from two different datasets and using two different computational methods (Figure 2): (1) QSAR-based datamining of chemical databases in an effort to identify novel ligands for 5-HT₆R, and (2) Network-mining using two signatures for Alzheimer's disease to query the cmap and identify possible anti-Alzheimer's therapeutics. Our procedure for integrating hypotheses was based on structural identity for chemical compounds derived from both approaches mentioned above. Compounds with negative connectivity scores, representing genes expressed in an opposite fashion to the imported Alzheimer's disease query—which implies their potential benefits to be candidate treatments, were compared with 5-HT₆R hits predicted from QSAR-based VS, and identical compounds were regarded as common hits.

The primary goal for integrating hit lists produced by two independent approaches in this study was initially to overcome some of the inherent hit scoring problems in classification QSAR, and achieve higher success rates in experimental testing of the VS hits. In other words; we often select for further experimental validation those QSAR hits with consensus scores above or equal to 0.90 (referred to as consensus scores in Methods). However, many novel scaffolds that are significantly different (i.e., structurally and possibly, therapeutically) from the training set compounds, might have lower consensus scores ranging from 0.50 to 0.90 despite the fact that they might be actives too. Thus, this process of integrating hypotheses derived independently from different types of data and using multiple prediction methods, allowed us to fish out these low-confidence QSAR hits (that yet could be highly important ligands) for further analysis. As a result, we posit that integrating independent hypotheses is likely to improve the overall experimental success rates of hit compounds identified *in silico*.

Scoring and Integrating Structural Hypotheses to Identify Putative Anti-Alzheimer's Agents

Our method for integrating hit lists was derived from a combination of voting and statistical metrics. In the first step, we used two different scoring functions to rank the computational hits generated independently from both QSAR and cmap. In the QSAR study, we used the *k*NN 'consensus score' which takes into account the total number of models used to predict compound's activity, and the number of models that predicted the compound to belong to a specific class correctly. We considered all computational hits that had an average predicted value (i.e., consensus score) above or equal to 0.50 for further inspection. Our analysis resulted in 140 putative 5-HT₆R actives among cmap compounds and with *k*NN consensus scores ranging from 0.50–1.00 (see Figure 7)

On the other hand, we used the connectivity scores³³ to rank the hits resulting from querying the cmap with Alzheimer's disease gene signatures. Because we were interested in identifying novel treatments for Alzheimer's disease, we ranked hits with larger (–)

connectivity scores at the top and gave them higher confidence. Such compounds were hypothesized to have higher chances to reverse the Alzheimer's gene signatures and therefore might have immense therapeutic value in Alzheimer's disease. We considered for further analysis all compounds that had at least one instance of negative connection with any of the two gene signatures used to query the cmap (S1 and S2) so that not to miss any important connections. Our analysis resulted in identifying 881 negative connectivity instances with S1 and 861 instances with S2 (Figure 6).

Finally, we combined the hypotheses generated from both QSAR and cmap analyses and accepted common hits only. We identified 97 compounds that were both predicted to be active at 5-HT₆R and had at least one instance of negative connectivity with S1 and 106 compounds that had at least one instance of negative connectivity with S2. Accepting only common hits among S1 and S2 resulted in 73 putative hits (see Figure 7). At this stage we applied a manual curation where we inspected all available data for these 73 hits. Each of the 73 common hits had three scores (*k*NN consensus score, cmap connectivity score with S1, and cmap connectivity score with S2) to be considered in the final decision to prioritize hits for further testing. Therefore, we estimated the average connectivity scores for all predicted hits across all treatment instances for each of the S1 and S2 hits. Then we excluded those compounds that had high positive connectivity scores in some treatment instances of the same compound. Finally, we retained 39 compounds that had acceptable negative average connectivity scores at least with one signature (see Figure 7). We hypothesized that these compounds could be tested as putative 5-HT₆R hits and potential cognition enhancing agents in the Alzheimer's disease. One of the final 39 hits, vinpocetine, worth special attention as there is a new evidence that this compound may play a role in the treatment of Parkinson's disease and Alzheimer's disease^{80;81}. Details on all these 39 VS hits are provided in Tables 3 and 4.

Each of the 39 common hits had three scores (*k*NN consensus score, cmap connectivity score with S1, and cmap connectivity score with S2) to be considered in the final decision to prioritize hits for further experimental testing. We plotted the mean connectivity scores *vs.* *k*NN QSAR consensus scores generating separate plots for S1 and S2 (see Figure 8) to analyze these hits in further details.

Additionally, another level of confidence was achieved (besides considering both *k*NN CS and cmap scores) by giving more emphasis to molecules that belonged to the same pharmacological or therapeutic group or had very high structural similarity to hits of higher confidence. This step permitted the retrieval of some compounds that had less significant negative connectivity scores with the disease (e.g., null connectivity or even low positive connectivity scores in few instances). We noticed that the 39 putative actives belonged to several major therapeutic groups (see Table 4): antipsychotics, antidepressants, anti-histamines, selective estrogen receptor modulators (SERMs) and calcium channel blockers. Among these groups, predicting SERMs to have activity at 5-HT₆R was the most surprising.

Hypothesis Testing: Evaluation of Computational Hits at Human Cloned 5-HT₆ Receptors

Common hits from QSAR-VS studies and cmap were taken forward for biological validation, in binding assays, for 5-HT₆ receptor. As discussed above, we identified 39 chemicals, out of 59000 molecules included in the WDI⁵⁷, and 1300 compounds included in the cmap, as consensus hits and putative actives for 5-HT₆R with higher chances of having potential cognition enhancement effects in Alzheimer's disease; none of these hits was included in the training set used to develop QSAR models. Then, we prioritized twelve compounds from a list of 39 molecules plus an additional compound; a SERM that was among the top QSAR-VS hits that was not included in the cmap (**1-13**⁵⁷, see Tables 3 and 5) for further experimental validation in 5-HT₆R radioligand binding assays. It should be noted

that compound 12 was tested because it was a SERM predicted with a high CS of 0.93 and we wanted to test it because three other SERMs were among our common hit list of 39 compounds. Our final selection was based on different criteria: (1) we tested some compounds with high consensus scores and stronger negative connectivity with Alzheimer's disease, (2) some compounds were selected because they belonged to the same therapeutic class as several other predicted hits and were not known before to bind to 5-HT₆R such as selective estrogen receptor modulators (SERMs) (e.g. **1**, **11** and **12**), (3) we tested some compounds with low *k*NN CS (e.g., **10** having a *k*NN CS of 0.58) if other hits that belonged to the same therapeutic class had high consensus scores (e.g., **11** and **12** having *k*NN CS above to or equal 0.93 and **1** having a *k*NN CS of 0.91), (4) we also tried to test predictions that had strong negative connectivity scores with one query signature but had much weaker negative connectivity with the second signature to see if there is one specific signature that was generating better results.

We found that ten of these thirteen predicted actives were confirmed experimentally to inhibit 5-HT₆R radioligand binding thereby achieving a success hit rate of 77 % in this proof-of-concept study (see Table 5). One of these ten confirmed hits was clozapine, which is known to bind 5-HT₆R but was not included in our training set. Binding affinity (*K_i*) values for the nine predicted hits were in the range 17 – 4125 nM, with six compounds having *K_i* values < 1 μM. These six highest affinity compounds were: **3** (*K_i*=17 nM)¹⁷, **4** (*K_i*=105 nM, Figure 9 (A)), **2** (*K_i*=112 nM, Figure 9 (B)), **13** (*K_i*=169 nM, Figure 9 (C)), **8** (*K_i*=214 nM, Figure 9 (D)) and **10** (*K_i*=750 nM, Figure 9 (E)).

Among the tested compounds, we found that compounds having negative connectivity scores and *k*NN CS above 0.90 were all true actives at 5-HT₆R achieving a success rate of 100%. We also found that lowering the threshold to 0.50 resulted in 3 false positives which decreased the success rate down to 77 %. It was strikingly important that we were able to prioritize a VS hit (i.e., **10**) with very low *k*NN CS of 0.56 and insignificant negative connectivity scores with Alzheimer's (see Table 5) and validate that this compound was a true active of 5-HT₆R and a potential cognition enhancer. This is a clear example on the importance of integrating independent hypotheses to increase the confidence of otherwise less significant computational hits.

Mining of the biomedical literature using ChemoText identified possible neuroprotective, in addition to cognitive- and memory-enhancing, effects for most of the computational hits (see Table 6), although there is no evidence that 5-HT₆R –active compounds are neuroprotective. The list of all 39 compounds predicted by our integrative approach as putative 5-HT₆R actives with possible anti- Alzheimer's effects is shown in Table 5.

SERMs Identified as 5-HT₆R ligands

Several selective estrogen receptor modulators (SERMs) were predicted as 5-HT₆R ligands and also had negative connections with the Alzheimer's disease gene signatures. **1**, **10** and **11** had negative connections with Alzheimer's disease gene signatures in the cmap database^{33;34}. **12** was not included in the cmap but was predicted as 5-HT₆R active by QSAR-based VS. Although anti-Alzheimer's effects of these drugs were observed previously and attributed to their modulation of estrogen receptors (ERs), the evidence about ER modulators or hormone replacement therapy in postmenopausal women to prevent or treat the Alzheimer's disease has been inconclusive and sometimes even contradictory^{82–84}. Although postmenopausal estrogen depletion is a known risk factor for Alzheimer's disease, estrogen-containing hormone therapy initiated during late postmenopausal period does not improve episodic memory (an important early symptom of Alzheimer's disease), leads to no improvement or adverse effect on overall cognitive performance and Alzheimer's disease in

postmenopausal women^{84–86}, and it increases the risk of dementia^{83;84}. Be that as it may, there is still substantial evidence from both pre-clinical and human studies that ovarian steroids have significant effects on neuroregulatory pathways^{87–94}. However, critical gaps exist in our knowledge of both the effects on brain function of declining ovarian steroid secretion during reproductive aging, and the role of ovarian steroid hormone therapy in the prevention or treatment of brain diseases⁸².

Raloxifene Identified as a 5-HT₆R Antagonist and Agent with Potential Utility in Alzheimer's Disease

Raloxifene is a selective estrogen receptor modulator used to prevent or treat osteoporosis; recently it was also approved by the FDA as an anti-cancer drug for reducing the risk of invasive breast cancer in postmenopausal women⁹⁵. It was one of the low confidence QSAR-based VS hits because of the low structural similarity with modeling set compounds. Therefore, we would have avoided testing this compound if it had not been predicted from the cmap to have a strong negative connection with Alzheimer's disease. Another level of confidence was obtained from having other compounds that belonged to the same pharmacological group (SERMs) that were predicted as 5-HT₆R actives with high confidence (i.e., consensus scores above 0.90) and had negative connections with Alzheimer's disease. This example highlights the value of the integrated informatics approach in increasing the hit rates of QSAR-based VS. Experimental testing had indeed confirmed that raloxifene binds to 5-HT₆R with a K_i of 750 nM (Table 5, Figure 9 (E)).

Yaffe and coworkers examined the data from the Multiple Outcomes of Raloxifene Evaluation (MORE) trial and indicated that raloxifene given at a dose of 120 mg/day, but not 60 mg/day, led to reduced risk of cognitive impairment in postmenopausal women⁹⁶. Additionally, recent studies pointed out that raloxifene enters the brain in relevant quantities and exerts a measurable effect in humans⁹⁷. It is possible that raloxifene's anticipated anti-Alzheimer's effects could be due to complex polypharmacological profile effecting several protein targets and signaling pathways involved in memory, cognition, inflammation, oxidative control and other important biological processes underlying the Alzheimer's disease etiology, and not limited to its canonical targets (i.e., estrogen receptors). Currently, raloxifene is in phase II clinical trials for Alzheimer's disease in postmenopausal women⁹⁸. This example can thus be considered as a proof of concept for the ability of our approach to increase the confidence in the identified hits that are structurally and biologically dissimilar to training set compounds.

Conclusions

We have developed a novel integrative chemocentric informatics approach that could be used as a tool for generating and cross-validating drug discovery hypotheses. Our approach integrates different *in silico* strategies and different data types and sources to increase the confidence in the final hypotheses. The study design was composed of three major parts: (1) QSAR-based datamining of chemical libraries to identify new ligands for target proteins, (2) Network-mining to identify chemicals that could treat specific diseases; and (3) Integrating hits derived from (1) and (2).

This approach has been applied to study the 5-HT₆R system in relation to cognition enhancement strategies which may be useful for Alzheimer's and, perhaps, similar diseases with impaired cognition. Disease gene signatures for Alzheimer's disease have been used to query the cmap database to formulate testable hypotheses about potential treatments. Common compound hits from QSAR/VS studies against 5-HT₆R and the cmap were tested in at 5-HT₆R. Our approach identified 39 drugs, as potential 5-HT₆R antagonists, out of 59000 molecules included in the WDI⁵⁷. Thirteen hits with the highest confidence level

were tested in binding assays and ten compounds were confirmed as 5-HT₆R ligands achieving a success rate of 77%. We noticed that this study design can be applied to many other protein targets and families of targets involved in the etiology of Alzheimer's disease.

We shall emphasize that our approach could be used to aid in the process of prioritizing computational hits that would not be picked by an individual model contributing to the integrative approach presented in this study. For instance, QSAR/VS hits could emphasize connections from the cmap that one would not focus on otherwise, especially as the size of the database continues to grow. In diseases like the Alzheimer's, with little knowledge about specific etiology, and the lack of drug gene signatures generated from neuronal cell lines, it is hard to decide *a priori* which negative connections are more important to be viewed as potential therapeutics. Reciprocally, one can use strong connections revealed by cmap to focus on weak hits resulting from QSAR/VS studies as was demonstrated here for raloxifene. Thus, integrating hypotheses derived independently from cmap and QSAR should enable us to increase the confidence in common hits. Herein, we hypothesized and proved that integrating results generated from the cmap with predictions generated from QSAR-based VS increased the confidence in the final hit list of predicted actives.

Experimental Section

Radioligand Binding Assays

This screen was performed by the National Institute of Mental Health Psychoactive Drug Screening Program (PDSP)¹⁷. Radioligands were purchased by PDSP from Perkin-Elmer or GE Healthcare. Competition binding assays were performed using transfected or stably expressing cell membrane preparations as previously described (Shapiro *et al.* 2003⁹⁹; Roth *et al.* 2002¹⁰⁰) and are available online (<http://pdsp.med.unc.edu>). All experimental details are available online (<http://pdsp.med.unc.edu/UNC-CH%20Protocol%20Book.pdf>).

Chemistry

Chemical compounds predicted as hits from the virtual screening were obtained from commercial suppliers according to their availability. All compounds were ordered to have above or equal to 95% purity. Additionally, all compounds were subjected to purity assessment using LC/MS by the Center for Integrative Chemical Biology and Drug Discovery at UNC-Chapel Hill (see Supporting Information). LC/MS spectra of all compounds were acquired from an Agilent 6110 Series system with UV detector set to 220 nm. Samples were injected (5 μ L) onto an Agilent Eclipse Plus 4.6 $\times$ 50 mm, 1.8 μ M, C18 column at room temperature. A linear gradient from 10% to 100% B (MeOH + 0.1% Acetic Acid) in 5.0 min was followed by pumping 100% B for another 2 minutes with A being H₂O + 0.1% acetic acid. The flow rate was 1.0 mL/min.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Tripos, Chemical Computing Group, and eduSoft for software grants. We also thank Ms. Xin Chin from the Center for Integrative Chemical Biology and Drug Discovery at UNCChapel Hill for performing the purity control tests. Finally, we acknowledge the access to the computing facilities at the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The studies reported in this paper were supported in part by the NIH research grant GM066940 (awarded to AT); RO1MH61887, U19MH82441, an NIH contract which supports the NIMH Psychoactive Drug Screening Program and the Michael Hooker Chair in Protein Therapeutics and Translational Proteomics (awarded to BR); and by the University of Jordan scholarship (awarded to RH).

Abbreviations

5-HT₆R	5-Hydroxytryptamine-6 Receptors
CARs	Class Association Rules
CBA	Classification Based on Association
CCR	Correct Classification Rate
CCR_{train}	Correct Classification Rate for training set
CCR_{test}	Correct Classification Rate for test set
CCR_{evs}	Correct Classification Rate for external validation set
CCR_{rand}	Correct Classification Rate of the random models using the external validation set
cmap	The Connectivity Map
Cp.	Compound
CS	Consensus Score
CV	Cross Validation
E	Enrichment
En	Normalized Enrichment
FN	False Negative
FP	False Positive
HTS	High Throughput Screen
kNN	<i>k</i> Nearest Neighbor
LOO-CV	Leave-One-Out Cross Validation
MFD	Most Frequent Descriptors
MOE	Molecular Operating Environment
MZ	MolConnZ descriptors
PDSP	NIMH Psychoactive Drug Screening Program
QSAR	Quantitative Structure-Activity Relationships
S1	Alzheimer's disease gene signature 1
S2	Alzheimer's disease gene signature 1
SA	Simulated Annealing
SE	Sensitivity
SG	Subgraph
SP	Specificity
TP	True Positive
TN	True Negative
VS	Virtual Screening
WDI	World Drug Index

References

1. Connor SC, Hansen MK, Corner A, Smith RF, Ryan TE. Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes. *Mol. Biosyst.* 2010; 6:909–921. [PubMed: 20567778]
2. Nicholson JK, Holmes E, Lindon JC, Wilson ID. The challenges of modeling mammalian biocomplexity. *Nat. Biotechnol.* 2004; 22:1268–1274. [PubMed: 15470467]
3. Petak I, Schwab R, Orfi L, Kopper L, Keri G. Integrating molecular diagnostics into anticancer drug discovery. *Nat. Rev. Drug Discov.* 2010
4. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, Van DK, Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 2001; 19:45–50. [PubMed: 11135551]
5. Yang R, Niepel M, Mitchison TK, Sorger PK. Dissecting variability in responses to cancer chemotherapy through systems pharmacology. *Clin. Pharmacol. Ther.* 2010; 88:34–38. [PubMed: 20520606]
6. Brenner C. Chemical genomics in yeast. *Genome Biol.* 2004; 5:240. [PubMed: 15345040]
7. Darvas F, Dorman G, Krajcsi P, Puskas LG, Kovari Z, Lorincz Z, Urge L. Recent advances in chemical genomics. *Curr. Med. Chem.* 2004; 11:3119–3145. [PubMed: 15579004]
8. Nislow C, Giaever G. "Chemogenomics: tools for protein families" and "Chemical genomics: chemical and biological integration". *Pharmacogenomics.* 2003; 4:15–18. [PubMed: 12517281]
9. Salemme FR. Chemical genomics as an emerging paradigm for postgenomic drug discovery. *Pharmacogenomics.* 2003; 4:257–267. [PubMed: 12718717]
10. Zheng XF, Chan TF. Chemical genomics: a systematic approach in biological research and drug discovery. *Curr. Issues Mol. Biol.* 2002; 4:33–43. [PubMed: 11931568]
11. Zheng XF, Chan TF. Chemical genomics in the global study of protein functions. *Drug Discov. Today.* 2002; 7:197–205. [PubMed: 11815236]
12. Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S, Brouwer C, Harland L. Visualizing the drug target landscape. *Drug Discov. Today.* 2010; 15:3–15. [PubMed: 19840866]
13. Hamadeh HK, Todd M, Healy L, Meyer JT, Kwok AM, Higgins M, Afshari CA. Application of genomics for identification of systemic toxicity triggers associated with VEGF-R inhibitors. *Chem. Res. Toxicol.* 2010; 23:1025–1033. [PubMed: 20521778]
14. Kiessling LL, Splain RA. Chemical approaches to glycobiology. *Annu. Rev. Biochem.* 2010; 79:619–653. [PubMed: 20380561]
15. Ogorevc J, Dovc P, Kunej T. Comparative Genomics Approach to Identify Candidate Genetic Loci for Male Fertility. *Reprod. Domest. Anim.* 2010
16. Wagner BK, Clemons PA. Connecting synthetic chemistry decisions to cell and genome biology using small-molecule phenotypic profiling. *Curr. Opin. Chem. Biol.* 2009; 13:539–548. [PubMed: 19825513]
17. PDSP. PDSP Ki Database. 2011. <http://pdsp.med.unc.edu>
18. PubChem. NIH's Molecular Libraries Roadmap Initiative. 2011.
19. ChEMBL. European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI). 2011.
20. Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, TI. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In: Schreiber, SL.; Kapoor, TM.; Wess, G., editors. *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*. New York: Wiley-VCH; 2007. p. 760-786.
21. Oprea T, Tropsha A. Target, Chemical and Bioactivity Databases – Integration is Key. *Drug Discov. Today.* 2006; 3:357–365.
22. Kortagere S, Ekins S. Troubleshooting computational methods in drug discovery. *J. Pharmacol. Toxicol. Methods.* 2010; 61:67–75. [PubMed: 20176118]
23. Burgun A, Bodenreider O. Accessing and integrating data and knowledge for biomedical research. *Yearb. Med. Inform.* 2008:91–101. [PubMed: 18660883]

24. Kandpal R, Saviola B, Felton J. The era of 'omics unlimited. *Biotechniques*. 2009; 46:351–355. [PubMed: 19480630]
25. Polychronakos C. New applications of microarray data analysis: integrating genetics with 'Omics'. Organized by the Cambridge Healthtech Institute, 15–17 August 2007, Washington DC, USA. *Pharmacogenomics*. 2008; 9:15–17. [PubMed: 18154445]
26. Vangala S, Tonelli A. Biomarkers, metabonomics, and drug development: can inborn errors of metabolism help in understanding drug toxicity? *AAPS J*. 2007; 9:E284–E297. [PubMed: 17915830]
27. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30:207–210. [PubMed: 11752295]
28. Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*. 2006; 411:352–369. [PubMed: 16939800]
29. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*. 2008; 24:2798–2800. [PubMed: 18842599]
30. Wishart DS. Human Metabolome Database: completing the 'human parts list'. *Pharmacogenomics*. 2007; 8:683–686. [PubMed: 18240899]
31. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De SA, Zuniga A, Dawe M, Xiong Y, Clive D, Greiner R, Nazzyrova A, Shaykhtudinov R, Li L, Vogel HJ, Forsythe I. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res*. 2009; 37:D603–D610. [PubMed: 18953024]
32. Kinase SARfari. The European Bioinformatics Institute (EMBL-EBI). 2010.
33. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006; 313:1929–1935. [PubMed: 17008526]
34. Lamb J. Innovation - The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*. 2007; 7:54–60.
35. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res*. 2009; 37:D786–D792. [PubMed: 18782832]
36. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von MC, Jensen LJ, Beyer A, Bork P. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res*. 2009
37. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research*. 2008; 36:D684–D688. [PubMed: 18084021]
38. Burks C, Cassidy M, Cinkosky MJ, Cumella KE, Gilna P, Hayden JE, Keen GM, Kelley TA, Kelly M, Kristofferson D. GenBank. *Nucleic Acids Res*. 1991; (19 Suppl):2221–2225. [PubMed: 2041806]
39. Burks C, Cinkosky MJ, Gilna P, Hayden JE, Abe Y, Atencio EJ, Barnhouse S, Benton D, Buenafe CA, Cumella KE. GenBank: current status and future directions. *Methods Enzymol*. 1990; 183:3–22. [PubMed: 2314279]
40. Palfreyman MG, Hook DJ, Klimczak LJ, Brockman JA, Evans DM, Altar CA. Novel directions in antipsychotic target identification using gene arrays. *Curr. Drug Targets CNS Neurol. Disord*. 2002; 1:227–238. [PubMed: 12769629]
41. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*. 1997; 278:680–686. [PubMed: 9381177]
42. Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, Kittrell FS, Zahnow CA, Patterson N, Golub TR, Ewen ME. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*. 2003; 114:323–334. [PubMed: 12914697]
43. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class

- discovery and class prediction by gene expression monitoring. *Science*. 1999; 286:531–537. [PubMed: 10521349]
44. Perou CM, Sorlie T, Eisen MB, van de RM, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. *Nature*. 2000; 406:747–752. [PubMed: 10963602]
  45. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*. 2002; 415:436–442. [PubMed: 11807556]
  46. van't V, Dai H, van d V, He YD, Hart AA, Mao M, Peterse HL, van der KK, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002; 415:530–536. [PubMed: 11823860]
  47. Altar CA, Vawter MP, Ginsberg SD. Target identification for CNS diseases by transcriptional profiling. *Neuropsychopharmacology*. 2009; 34:18–54. [PubMed: 18923405]
  48. Ogden CA, Rich ME, Schork NJ, Paulus MP, Geyer MA, Lohr JB, Kuczenski R, Niculescu AB. Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: an expanded convergent functional genomics approach. *Mol. Psychiatry*. 2004; 9:1007–1029. [PubMed: 15314610]
  49. Le-Niculescu H, McFarland MJ, Mamidipalli S, Ogden CA, Kuczenski R, Kurian SM, Salomon DR, Tsuang MT, Nurnberger JI Jr, Niculescu AB. Convergent Functional Genomics of bipolar disorder: from animal model pharmacogenomics to human genetics and biomarkers. *Neurosci. Biobehav. Rev.* 2007; 31:897–903. [PubMed: 17614132]
  50. Garman KS, Acharya CR, Edelman E, Grade M, Gaedcke J, Sud S, Barry W, Diehl AM, Provenzale D, Ginsburg GS, Ghadimi BM, Ried T, Nevins JR, Mukherjee S, Hsu D, Potti A. A genomic approach to colon cancer risk stratification yields biologic insights into therapeutic opportunities. *Proc. Natl. Acad. Sci. U S A*. 2008; 105:19432–19437. [PubMed: 19050079]
  51. Hassane DC, Guzman ML, Corbett C, Li X, Abboud R, Young F, Liesveld JL, Carroll M, Jordan CT. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood*. 2008; 111:5654–5662. [PubMed: 18305216]
  52. Hieronymus H, Lamb J, Ross KN, Peng XP, Clement C, Rodina A, Nieto M, Du J, Stegmaier K, Raj SM, Maloney KN, Clardy J, Hahn WC, Chiosis G, Golub TR. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell*. 2006; 10:321–330. [PubMed: 17010675]
  53. Riedel RF, Porrello A, Pontzer E, Chenette EJ, Hsu DS, Balakumaran B, Potti A, Nevins J, Febbo PG. A genomic approach to identify molecular pathways associated with chemotherapy resistance. *Mol. Cancer Ther.* 2008; 7:3141–3149. [PubMed: 18852117]
  54. Setlur SR, Mertz KD, Hoshida Y, Demichelis F, Lupien M, Perner S, Sboner A, Pawitan Y, Andren O, Johnson LA, Tang J, Adami HO, Calza S, Chinnaiyan AM, Rhodes D, Tomlins S, Fall K, Mucci LA, Kantoff PW, Stampfer MJ, Andersson SO, Varenhorst E, Johansson JE, Brown M, Golub TR, Rubin MA. Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J Natl. Cancer Inst.* 2008; 100:815–825. [PubMed: 18505969]
  55. Zimmer M, Ebert BL, Neil C, Brenner K, Papaioannou I, Melas A, Tolliday N, Lamb J, Pantopoulos K, Golub T, Iliopoulos O. Small-molecule inhibitors of HIF-2 α translation link its 5'UTR iron-responsive element to oxygen sensing. *Mol. Cell*. 2008; 32:838–848. [PubMed: 19111663]
  56. Zimmer M, Lamb J, Ebert BL, Lynch M, Neil C, Schmidt E, Golub TR, Iliopoulos O. The connectivity map links iron regulatory protein-1-mediated inhibition of hypoxia-inducible factor-2 α translation to the anti-inflammatory 15-deoxy- $\Delta^{12,14}$ -prostaglandin J₂. *Cancer Res.* 2010; 70:3071–3079. [PubMed: 20354189]
  57. Daylight. World Drug Index (WDI). 2004.
  58. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J Biomed. Inform.* 2010; 43:510–519. [PubMed: 20348023]

59. Holenz J, Pauwels PJ, Diaz JL, Merce R, Codony X, Buschmann H. Medicinal chemistry strategies to 5-HT₆ receptor ligands as potential cognitive enhancers and antiobesity agents. *Drug Discov. Today*. 2006; 11:283–299. [PubMed: 16580970]
60. Geldenhuys WJ, Van der Schyf CJ. The serotonin 5-HT₆ receptor: a viable drug target for treating cognitive deficits in Alzheimer's disease. *Expert Rev. Neurother*. 2009; 9:1073–1085. [PubMed: 19589055]
61. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*. 2010; 29:476–488.
62. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*. 2011; 39:D1035–D1041. [PubMed: 21059682]
63. Cheng J, Sun S, Tracy A, Hubbell E, Morris J, Valmeekam V, Kimbrough A, Cline MS, Liu GY, Shigeta R, Kulp D, Siani-Rose MA. NetAffx gene ontology mining tool: A visual approach for microarray data analysis. *Bioinformatics*. 2004; 20:1462–1463. [PubMed: 14962933]
64. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*. 2003; 31:82–86. [PubMed: 12519953]
65. Swanson DR. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc*. 1990; 78:29–37. [PubMed: 2403828]
66. 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
67. Molecular Operating Environment (MOE). Version 2008.10. Chemical Computing Group (CCG). 2011.
68. JChem, Version 5.2.6. ChemAxon. 2010.
69. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*. 2003; 43:579–586. [PubMed: 12653524]
70. Kohavi RA. Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence (IJCAI)*. 1995; 2(12):1137–1143.
71. Tropsha A, Golbraikh A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des*. 2007; 13:3494–3504. [PubMed: 18220786]
72. Golbraikh A, Tropsha A. Beware of q²! *J. Mol. Graph. Model*. 2002; 20:269–276. [PubMed: 11858635]
73. Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput. Aided Mol. Des*. 2002; 16:357–369. [PubMed: 12489684]
74. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FCP, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature Genetics*. 2001; 29:365–371. [PubMed: 11726920]
75. Hata R, Masumura M, Akatsu H, Li F, Fujita H, Nagai Y, Yamamoto T, Okada H, Kosaka K, Sakanaka M, Sawada T. Up-regulation of calcineurin Abeta mRNA in the Alzheimer's disease brain: assessment by cDNA microarray. *Biochem. Biophys. Res. Commun*. 2001; 284:310–316. [PubMed: 11394878]
76. Ricciarelli R, d'Abramo C, Massone S, Marinari U, Pronzato M, Tabaton M. Microarray analysis in Alzheimer's disease and normal aging. *IUBMB Life*. 2004; 56:349–354. [PubMed: 15370883]
77. Liu, B.; Hsu, W.; Ma, Y. Classification Based on Association (CBA). [v2.1]. School of Computing, National University of Singapore; 2001.
78. Hollander, M.; Wolfe, D. Nonparametric Statistical Methods. New York: Wiley; 1999. p. 178-185.
79. The Connectivity Map Help. The Connectivity Map. Broad Institute; 2011.
80. Jeon KI, Xu X, Aizawa T, Lim JH, Jono H, Kwon DS, Abe J, Berk BC, Li JD, Yan C. Vinpocetine inhibits NF-kappaB-dependent inflammation via an IKK-dependent but PDE-independent mechanism. *Proc. Natl. Acad. Sci. U S A*. 2010; 107:9795–9800. [PubMed: 20448200]

81. Medina AE. Vinpocetine as a potent antiinflammatory agent. *Proc. Natl. Acad. Sci. U S A.* 2010; 107:9921–9922. [PubMed: 20495091]
82. Asthana S, Brinton RD, Henderson VW, McEwen BS, Morrison JH, Schmidt PJ. Frontiers proposal. National Institute on Aging "bench to bedside: estrogen as a case study". *Age (Dordr.).* 2009; 31:199–210. [PubMed: 19277902]
83. Henderson VW. Estrogens, episodic memory, and Alzheimer's disease: a critical update. *Semin. Reprod. Med.* 2009; 27:283–293. [PubMed: 19401959]
84. Shumaker SA, Legault C, Rapp SR, Thal L, Wallace RB, Ockene JK, Hendrix SL, Jones BN III, Assaf AR, Jackson RD, Kotchen JM, Wassertheil-Smoller S, Wactawski-Wende J. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA.* 2003; 289:2651–2662. [PubMed: 12771112]
85. Pinkerton JV, Henderson VW. Estrogen and cognition, with a focus on Alzheimer's disease. *Semin. Reprod. Med.* 2005; 23:172–179. [PubMed: 15852203]
86. Rapp SR, Espeland MA, Shumaker SA, Henderson VW, Brunner RL, Manson JE, Gass ML, Stefanick ML, Lane DS, Hays J, Johnson KC, Coker LH, Dailey M, Bowen D. Effect of estrogen plus progestin on global cognitive function in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA.* 2003; 289:2663–2672. [PubMed: 12771113]
87. Schmidt PJ, Rubinow DR. Sex Hormones and Mood in the Perimenopause. *Glucocorticoids and Mood Clinical Manifestations, Risk Factors, and Molecular Mechanisms.* 2009; 1179:70–85.
88. Benmansour S, Piotrowski JP, Altamirano AV, Frazer A. Impact of Ovarian Hormones on the Modulation of the Serotonin Transporter by Fluvoxamine. *Neuropsychopharmacology.* 2009; 34:555–564. [PubMed: 18322468]
89. Frye CA. Progestogens influence cognitive processes in aging. *Future Medicinal Chemistry.* 2009; 1:1215–1231. [PubMed: 21426099]
90. Ledoux VA, Smejkalova T, May RM, Cooke BM, Woolley CS. Estradiol Facilitates the Release of Neuropeptide Y to Suppress Hippocampus-Dependent Seizures. *Journal of Neuroscience.* 2009; 29:1457–1468. [PubMed: 19193892]
91. Woolley CS. Estrogen and synapses in the hippocampus. *Faseb Journal.* 2007; 21:A89.
92. Hart SA, Snyder MA, Smejkalova T, Woolley CS. Estrogen mobilizes a subset of estrogen receptor-alpha-immunoreactive vesicles in inhibitory presynaptic boutons in hippocampal CA1. *Journal of Neuroscience.* 2007; 27:2102–2111. [PubMed: 17314305]
93. Woolley CS. Acute effects of estrogen on neuronal physiology. *Annual Review of Pharmacology and Toxicology.* 2007; 47:657–680.
94. Woolley CS, Schwartzkroin PA. Hormonal effects on the brain. *Epilepsia.* 1998; 39:S2–S8. [PubMed: 9915614]
95. Evista Approved for Reducing Breast Cancer Risk. U.S. Food and Drug Administration; <http://www.fda.gov/>. 9-17-2007.
96. Yaffe K, Krueger K, Cummings SR, Blackwell T, Henderson VW, Sarkar S, Ensrud K, Grady D. Effect of raloxifene on prevention of dementia and cognitive impairment in older women: The multiple outcomes of raloxifene evaluation (MORE) randomized trial. *American Journal of Psychiatry.* 2005; 162:683–690. [PubMed: 15800139]
97. Littleton-Kearney MT, Ostrowski NL, Cox DA, Rossberg MI, Hurn PD. Selective estrogen receptor modulators: tissue actions and potential for CNS protection. *CNS Drug Rev.* 2002; 8:309–330. [PubMed: 12353060]
98. Can Alzheimer's disease be prevented?. The National Institute on Aging (NIA). 5-10-2010
99. Shapiro DA, Renock S, Arrington E, Chiodo LA, Liu LX, Sibley DR, Roth BL, Mailman R. Aripiprazole, a novel atypical antipsychotic drug with a unique and robust pharmacology. *Neuropsychopharmacology.* 2003; 28:1400–1411. [PubMed: 12784105]
100. Roth BL, Baner K, Westkaemper R, Siebert D, Rice KC, Steinberg S, Ernsberger P, Rothman RB. Salvinorin A: A potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proceedings of the National Academy of Sciences of the United States of America.* 2002; 99:11934–11939. [PubMed: 12192085]

101. Hwang J, Zheng LT, Ock J, Lee MG, Kim SH, Lee HW, Lee WH, Park HC, Suk K. Inhibition of glial inflammatory activation and neurotoxicity by tricyclic antidepressants. *Neuropharmacology*. 2008; 55:826–834. [PubMed: 18639562]
102. Pfeifer, A.; Schrattenholz, A.; Muhs, A. Therapeutic Compound. PCT/EP2008/058527[WO 2009/004038 A2], 1-80. 1-8-2009. USA. 7-2-2008.
103. Ong J, Kerr DI. Clinical potential of GABAB receptor modulators. *CNS Drug Rev*. 2005; 11:317–334. [PubMed: 16389296]
104. Ong J, Parker DA, Marino V, Kerr DI, Puspawati NM, Prager RH. 3-Chloro,4-methoxyfendiline is a potent GABA(B) receptor potentiator in rat neocortical slices. *Eur. J Pharmacol*. 2005; 507:35–42. [PubMed: 15659292]
105. Zhao W, Wang J, Ho L, Ono K, Teplow DB, Pasinetti GM. Identification of antihypertensive drugs which inhibit amyloid-beta protein oligomerization. *J Alzheimers Dis*. 2009; 16:49–57. [PubMed: 19158421]
106. Zhang L, Yu J, Pan H, Hu P, Hao Y, Cai W, Zhu H, Yu AD, Xie X, Ma D, Yuan J. Small molecule regulators of autophagy identified by an image-based highthroughput screen. *Proc. Natl. Acad. Sci. U S A*. 2007; 104:19023–19028. [PubMed: 18024584]
107. Horwood JM, Dufour F, Laroche S, Davis S. Signalling mechanisms mediated by the phosphoinositide 3-kinase/Akt cascade in synaptic plasticity and memory in the rat. *Eur. J Neurosci*. 2006; 23:3375–3384. [PubMed: 16820027]
108. Pezet S, Marchand F, D'Mello R, Grist J, Clark AK, Malcangio M, Dickenson AH, Williams RJ, McMahon SB. Phosphatidylinositol 3-kinase is a key mediator of central sensitization in painful inflammatory conditions. *J Neurosci*. 2008; 28:4261–4270. [PubMed: 18417706]
109. Doraiswamy PM, Krishnan KR, Oxman T, Jenkyn LR, Coffey DJ, Burt T, Clary CM. Does antidepressant therapy improve cognition in elderly depressed patients? *J Gerontol. A Biol. Sci. Med. Sci*. 2003; 58:M1137–M1144. [PubMed: 14684712]
110. O'Neill K, Chen S, Diaz BR. Impact of the selective estrogen receptor modulator, tamoxifen, on neuronal outgrowth and survival following toxic insults associated with aging and Alzheimer's disease. *Exp. Neurol*. 2004; 188:268–278. [PubMed: 15246826]
111. Khalifa AE. Zuclopenthixol facilitates memory retrieval in rats: possible involvement of noradrenergic and serotonergic mechanisms. *Pharmacol. Biochem. Behav*. 2003; 75:755–762. [PubMed: 12957216]

Figure 1.
Computational hits.

Figure 2.

Study design for the integrated informatics approach for drug discovery integrating network mining, text mining of biological literature, the analysis of disease gene signatures and efficient cheminformatics techniques, to discover novel drugs with desired polypharmacology.

Figure 3.
The workflow for QSAR model building, validation and virtual screening as applied to 5-HT₆R dataset.

Figure 4. Comparison of the QSAR approaches to classify 5-HT₆R actives vs. non-actives based on CCR_{evs} .

Figure 5. QSAR-based virtual screening of two chemical databases: the WDI and DrugBank compounds included in the cmap.

Figure 6.
Querying the connectivity map with Alzheimer's disease gene signatures (S1 and S2).

Figure 7.
Integrating hypotheses from QSAR modeling and cmap negative connections.

Figure 8.

Plots for *k*NN scores vs. cmap connectivity scores for 39 final common hits from QSAR-based VS and cmap for: (A) Alzheimer's disease signature S1, and (B) Alzheimer's disease signature S2. Squares: compounds predicted and validated as 5-HT₆R actives having negative connectivity scores with Alzheimer's disease gene signatures; diamonds: compounds predicted and experimentally validated as 5-HT₆R actives but having positive connectivity scores with one of the Alzheimer's disease gene signatures; triangles: compounds predicted as 5-HT₆R actives having negative connectivity scores with Alzheimer's disease gene signatures but found non-actives in radioligand binding assays against 5-HT₆R; circles: compounds predicted as 5-HT₆R actives which have negative

connectivity scores with Alzheimer's disease gene signatures but were not experimentally tested.

A

B

C

Figure 9.

Competition binding isotherms at 5-HT₆R for several predicted actives: (A) **2** (red triangle) and chlorpromazine (square), and **4** (blue triangle) and chlorpromazine (square); (B) **8** (red triangle) and chlorpromazine (square), and **10** (blue triangle) and chlorpromazine (square); (C) **13** (triangle) and chlorpromazine (square), versus [³H]LSD.

Table 1

Top twenty negative connections from the cmap with S1.

Compound	Rank ^a	Cell	Score	Instance_ID
Naproxen	6100	PC3	−1	7146
Sulfacetamide	6099	MCF7	−0.990	1695
Amprolium	6098	HL60	−0.930	1979
Aminoglutethimide	6097	MCF7	−0.913	7463
Ioxaglic acid	6096	HL60	−0.897	2966
Dexpanthenol	6095	MCF7	−0.871	7455
Suxibuzone	6094	MCF7	−0.870	7163
Chlorphenesin	6093	HL60	−0.862	1432
Metixene	6092	HL60	−0.853	2451
Fulvestrant	6091	MCF7	−0.843	5565
Seneciophylline	6090	MCF7	−0.841	2797
Troglitazone	6089	MCF7	−0.839	6991
Dicloxacillin	6088	HL60	−0.834	2445
Phentolamine	6087	HL60	−0.831	2362
Monocrotaline	6086	MCF7	−0.828	6771
Lymecycline	6085	HL60	−0.823	2953
Bezafibrate	6084	PC3	−0.815	6653
6-Benzylaminopurine	6083	HL60	−0.812	2351
Terbutaline	6082	MCF7	−0.811	3202
Clorgiline	6081	MCF7	−0.805	3219

^aThe rank order is generated from estimating the connectivity scores of 6100 individual treatment instances with S1. A rank order of 6100 corresponds to the compound with the strongest negative connectivity S1, while a rank order of 1 corresponds to the compound with the strongest positive connectivity with S1.

Table 2

Top twenty negative connections from the cmap with S2.

Compound	Rank ^a	Cell	Score	Instance_ID
Trifluoperazine	6100	HL60	-1	2389
1	6099	MCF7	-0.982	4994
Ethotoin	6098	HL60	-0.977	2196
Sulfafurazole	6097	HL60	-0.973	1603
Quercetin	6096	MCF7	-0.964	4846
Triflusal	6095	HL60	-0.925	1717
Alfuzosin	6094	PC3	-0.903	4644
Metitepine	6093	HL60	-0.890	1616
Trioxysalen	6092	MCF7	-0.885	6216
7	6091	MCF7	-0.883	258
Tanespimycin	6090	HL60	-0.873	6184
Spironolactone	6089	MCF7	-0.871	6255
Nifurtimox	6088	MCF7	-0.859	4953
Iobenguane	6087	HL60	-0.847	1729
Undecanoic acid (U0125)	6086	PC3	-0.845	663
Monorden	6085	MCF7	-0.841	5947
Primidone	6084	PC3	-0.833	6723
Calcium pantothenate	6083	MCF7	-0.828	4775
Phthalylsulfathiazole	6082	HL60	-0.826	3033
Ceforanide	6081	PC3	-0.824	6751

^aThe rank order is generated from estimating the connectivity scores of 6100 individual treatment instances with S2. A rank order of 6100 corresponds to the compound with the strongest negative connectivity S2, while a rank order of 1 corresponds to the compound with the strongest positive connectivity with S2.

Table 3

Final thirty nine computational hits from QSAR-based VS and cmap.

cmap Name	cmap Score 1	cmap Score 2	Num. kNN Models	kNN CS	kNN Pred	CBA Pred
Acpromazine	-0.528	-0.496	441	0.93	B	B
Alimemazine	-0.121	-0.117	438	1.00	B	B
Astemizole	-0.349	-0.237	328	0.91	B	B
Bepridil	-0.134	-0.409	393	0.89	B	B
Bromperidol	-0.239	-0.213	428	0.83	B	B
Cetirizine	-0.495	-0.327	421	0.92	B	B
Chlorprothixene	-0.277	-0.298	442	0.90	B	B
Cinchocaine	-0.004	-0.335	423	0.58	B	B
Cinnarizine	-0.349	-0.149	414	0.98	B	B
Citalopram	-0.003	-0.260	429	0.71	B	NB
1	-0.378	-0.265	409	0.91	B	B
2	-0.192	-0.310	437	0.96	B	B
Cloperastine	-0.273	-0.353	443	0.88	B	B
3	0.093	-0.058	422	0.97	B	B
Diltiazem	-0.128	-0.336	433	0.72	B	NB
4	0.027	-0.259	444	0.95	B	B
5	-0.303	-0.228	393	0.84	B	NB
Flavoxate	-0.127	-0.112	403	0.71	B	NB
6	0.055	-0.138	351	0.98	B	B
Imipramine	-0.400	-0.214	427	0.96	B	B
Laudanosine	-0.226	-0.174	411	0.78	B	NB
7	-0.028	-0.078	428	0.71	B	B
Meclozine	-0.365	-0.171	439	0.95	B	B
Mepacrine	-0.236	-0.301	418	0.53	B	B
Methylethergometrine	-0.400	-0.509	441	0.98	B	B
Naftifine	-0.198	-0.148	359	0.85	B	B
8	0.011	-0.354	433	0.93	B	B
Phenoxybenzamine	-0.461	-0.309	444	0.79	B	NB

cmap Name	cmap Score 1	cmap Score 2	Num. kNN Models	kNN CS	kNN Pred	CBA Pred
Piperidolate	-0.168	-0.119	431	0.69	B	NB
9	-0.247	-0.206	435	0.96	B	B
Nitratine dihydrochloride	-0.086	-0.213	381	0.72	B	B
Promazine	-0.210	-0.307	424	0.97	B	B
10	0.047	-0.058	356	0.56	B	NB
11	0.300	-0.220	435	0.93	B	B
Telenzepine	-0.419	-0.114	387	0.68	B	B
Terfenadine	-0.183	-0.512	416	0.51	B	NB
Vanoxerine	-0.450	-0.233	374	1.00	B	NB
Vinpocetine	-0.177	-0.132	376	0.76	B	B
13	-0.152	0.144	434	0.98	B	B

Table 4

Therapeutic classes of the thirty nine final computational hits from QSAR-based VS and cmap.

cmap Name	Therapeutic Class/Use
Acepromazine	Antipsychotic
Alimemazine	Antipruritic, sedative, hypnotic and anti-emetic
Astemizole	Anti-Histamine
Bepidil	Calcium channel blocker once used to treat angina
Bromperidol	Neuroleptic, used as an antipsychotic in the treatment of schizophrenia
Cetirizine	Second-generation antihistamine
Chlorprothixene	Typical antipsychotic drug of the thioxanthene class
Cinchocaine	Local anesthetic
Cinnarizine	Antihistamine which is mainly used for the control of nausea and vomiting due to motion sickness
Citalopram	Antidepressant drug of the selective serotonin reuptake inhibitor (SSRI) class
1	SERM
2	Tricyclic antidepressant
Cloperastine	Cough suppressant.
3	Atypical antipsychotics
Diltiazem	Calcium channel blocker
4	Psychotropic agent with tricyclic antidepressant and anxiolytic properties
5	Calcium channel blocker
Flavoxate	Anticholinergic with antimuscarinic effects
6	Antipsychotic
Imipramine	Tricyclic antidepressant
Laudanosine	Benzyltetrahydroisoquinoline alkaloid. Interacts with GABA, opioid, and nicotinic acetylcholine receptors
7	Morpholino derivative of quercetin. It is a potent inhibitor of phosphoinositide 3-kinase s (PI3Ks)
Meclozine	Antihistamine considered to be an antiemetic
Mepacrine	Antiprotozoal, antirheumatic and an intrapleural sclerosing agent. It is known to act as a histamine N-methyltransferase inhibitor
Methylergometrine	Psychedelic alkaloid
Naftifine	Allylamine antifungal drug
8	Second-generation tricyclic antidepressant
Phenoxybenzamine	Non-specific, irreversible alpha antagonist
Piperidolate	Antimuscarinic.
9	Drug used in scientific research which acts as a moderately selective dopamine D ₃ receptor partial agonist.
Nitratine dihydrochloride	Hypotensive, spasmolytic, coronary dilator and sedative
Promazine	Antipsychotic
10	SERM
11	SERM
Telenzepine	Anticholinergic or sympatholytic
Terfenadine	Antihistamine formerly used for the treatment of allergic conditions
Vanoxerine	Piperazine derivative which is a potent and selective dopamine reuptake inhibitor (DRI)
Vinpocetine	Vinpocetine has been identified as a potent anti-inflammatory agent that might have a potential role in the treatment of Parkinson's disease and Alzheimer's disease ^{80,81}

cmap Name	Theraputic Class/Use
13	Typical antipsychotic drug

Table 5

Experimental validation results for the thirteen computational hits predicted as 5-HT₆R ligands and had negative connections with Alzheimer's disease gene signatures.

Compound	PDSP ID	Score ^b /Cell		kNN	CBA	Ki
		CID ^a	Score ^c /Cell	CS ^d	Pred. ^e	(nM)
1	13499	13499	-0.602/PC3	0.91	B ^f	1,956.0
		1548953	-0.982/MCF7			
2	13494	13494	-0.768/PC3	0.96	B	112.0
		2801	-0.814/MCF7			
3	24842		-0.590/PC3	0.97	B	17.0 ^g
4	13495	2818	-0.652/MCF7			
		667477	-0.463/MCF7	0.95	B	105.0
5	14821	14821	-0.777/HL60			
		3336	-0.520/MCF7	0.84	NB ^h	NB
6	14815	3336	-0.683/HL60			
		3396	-0.493/MCF7	0.98	B	1,188.0
7	13502	3396	-0.551/HL60			
		3973	-0.790/MCF7	0.69	B	NB
8	13503	13503	-0.883/MCF7			
		4543	-0.555/PC3	0.96	B	214.0
9	13498	4543	-0.586/MCF7			
		3038495	-0.741/MCF7	0.79	B	NB
10	13505	3038495	-0.619/HL60			
		5035	-0.626/HL60	0.56	NB	750.0
11	13506	5035	-0.619/HL60			
		2733526	0/MCF7	0.93	B	1,041.0
12ⁱ	16514	2733526	-0.531/MCF7			
		3005573	N/A	0.93	B	4,125.0
13	13510	3005573	N/A			
			-0.609/PC3	0.98	B	169.0

Compound	PDSP ID	Score ^b /Cell	kNN	CBA	Ki
	CID ^a	Score ^{2c} /Cell	CS ^d	Pred. ^e	(nM)
	5311507	-0.746/HL60			
Success Rate		77 % for predictions with kNN CS	0.5	100% for predictions with kNN CS	0.9

^a CID, PubChem compound ID;

^b cmap score1, the highest negative connectivity score for this compound with S1 (or the smallest positive in case all other scores are positive);

^c cmap score2, the highest negative connectivity score for this compound with S2 (or the smallest positive in case all other scores are positive);

^d CS, consensus score;

^e CBA pred., predicted binding to 5-HT₆ receptors by CBA;

^f B, active;

^g PDSP certified data;

^h NB, non-active;

ⁱ Toremifene was not included in the cmap database but was prioritized because 3 other related SERMs were hits from both cmap and QSAR-based VS.

Table 6

The significance of the tested hits in relation to cognition, neuroprotection and anti-Alzheimer's effects.

Compound	Predicted	K_i	Significance to Alzheimer's disease prevention/treatment
1	Active	1956.0	Unknown
2	Active	112.0	Neuroprotective ¹⁰¹
3	Active	17.0	Used in combination therapy for Alzheimer's ¹⁰²
4	Active	105.0	Unknown
5	Active	NB	GABA receptor modulator ^{103;104} and may inhibit amyloid-beta protein oligomerization as other related antihypertensives ¹⁰⁵
6	Active	1188.0	Possible anti-Alzheimer's effects ¹⁰⁶
7	Active	NB	Can inhibit central sensitization and neuroinflammation ^{107;108}
8	Active	214.0	Possible anti-Alzheimer's effects ¹⁰⁹
9	Active	NB	Unknown
10	Active	750.0	Possible anti-Alzheimer's effects ⁹⁶
11	Active	1041.0	Neuroprotective ¹¹⁰
12	Active	4125.0	Unknown
13	Active	169.0	Facilitates memory in rats ¹¹¹