National Technical University of Athens

Department of Chemical Engineering

Technical University of Denmark

Department of Systems Biology

Diploma Thesis

A Chemoinformatics Approach for the In Silico Identification of Interactions Between Psychiatric Drugs and Plant-Based Food

Anna Maria Tsakiroglou

Thesis Supervisors:
Assoc. Prof.: Kouskoumvekaki Irene
Assoc. Prof.: Topakas Evangelos

February 2016

Dedication

To all the people who made the realization of this work possible.
My supervisor for the most useful input and advice along the way. My parents for their unending support. To Katerina, and last, but not least, to Panagiotis, for always being there for me, even from so far away.
You are awesome.

Abstract

The overall objective of this project was the development of a systematic approach for identifying potential interactions between plant-based food and marketed psycholeptics and psychoanaleptic drugs. The problem was addressed, initially, by pairing psychiatric agents and phytochemicals to their common protein targets, using information available in online databases, such as NutriChem 1.0 [21] and Drugbank [16] and constructing the food-drug interaction networks. Thereupon, a search for additional phytochemicals that would be expected to interact with targets of psychiatric drugs was carried out. More specifically, three protein targets, P07550, P28222 and P14416 were selected, based on their frequency of interaction with both psycholeptics and psychoanaleptics and the availability of 3D structural information in PDB [19]. For these protein targets, ligand-based pharmacophoric hypothesis were generated, using experimental activity data from the literature, and the HypoGen feature of Accelrys Discovery Studio (DS) [7]. The pharmacophore models were validated using external test sets and Fisher's method. Subsequently, the models were used as queries to screen the NutriChem 1.0 database for more potentially active phytochemicals. Finally, based on a protocol introduced by Vilar et al. [37], a similarity-based prediction of psychiatric drugs' interactions with nutrients, present in plant-based food, was carried out. For that purpose, a reference database was constructed, incorporating information about 1763 FDA approved drugs and 69,356 drug-drug interactions (DDIs). Drug-drug interaction profile similarity, adverse effects similarity information, as well as 2D structural and target similarity information was gathered for the reference data set and used to train a SVM classification model. This model was later used to predict interactions between 64 phytochemicals and 85 psychiatric drugs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Determining drug and plant-based food interactions

Diet constitutes one of the most dynamic expressions of the ”exposome”, a term used to describe the sum of all environmental exposures (e.g. diet, air pollutants, lifestyle factors) over the life course of an individual (Jensen et al. [21]). Moreover, it is one of the most challenging to assess its effects in health homeostasis and disease development, considering its myriad components and their temporal variation. Through anecdotal experience, vegetables and fruits are considered beneficial to our health. Indeed, it is estimated that almost 80% of chronic diseases could be avoided by consumption of healthier food, whereas meta-analysis of observational studies has shown a dose-response effect of fruits and vegetables on cardiovascular diseases and stroke risk (Jensen et al. [21]). In several diseases the treatment effect is augmented by the combination of certain dietary components with pharmacotherapy. However, interference of plant-based foods with drug performance and pharmacological activity may also potentially contribute to an increased risk of side effects or treatment failure. Enzyme-inducing antiepileptic drugs such as carbamazepine and phenytoin may decrease serum vitamin D concentration by increasing the metabolism of vitamin D. An association among their use, reduced bone mineral density, and increased fracture risk has been suggested by several observational studies.(Chan [10]). In some cases drug intake may cause a negative impact on nutrient homeostasis, although the opposite can also happen; changes in nutrient intake can significantly alter a patient's response to a drug. A well-known example of an unfavorable drug-food interaction, the inhibitory effect of grapefruit juice on cytochrome P450, results in increased bioavailability of drugs such as felodipine, cyclosporine and saquinavir, which could lead to drug toxicity and poisoning (Jensen et al. [22]).

The causes of drug-food interactions are multifactorial and can also depend on sex, ethnicity, environmental factors, and genetic polymorphisms (Chan [10]). The majority of high risk treatment failure due to drug-food interactions is associated with reduced bioavailability of the drug in the fed state. Possible explanations include chelation with components in food, gastric acid secretion during food intake, or other direct drug-food interactions. On other occasions, food intake may result in an increase in drug bioavailability either because of a food-induced increase in drug solubility or because of the secretion of gastric acid or bile in response to food intake. In those cases, the effect of increased bioavailability of the drug tend to be ambiquous (Schmidt LE [36]).

Since drug-food associations are well recognized as an important element in pharmaceutical treatment, a need arises to systematically identify, predict and manage potential interactions between food and and marketed or novel drugs. Such an approach would pave the way for further applications in personalized medicine.

1.1.1 Definition of a drug-nutrient interaction

A drug-nutrient interaction, often referred to as a drug-food interaction, is defined as a physical, chemical, physiologic, or pathophysiologic relationship between a drug and a nutrient (Chan [10]). Assessing the dietary intervention in drug therapy can be better founded in a nutrient-based approach, as it allows for a more effective understanding of its mechanism. Additionally, clinical significance is noted when a drug-nutrient interaction becomes associated with an altered physiologic response, which may lead to malnutrition, treatment failure, adverse events, or, in the most serious case, a life-threatening event, including death (Chan [10]). Drug-nutrient interactions are often associated with a quantifiable alteration of the pharmacokinetic and/or ppharmacodynamic profile of a drug or a nutrient. The term kinetics (pharmacokinetics, nutrikinetics) refers to the quantitative description of the disposition of a drug or nutrient in the human body, through the processes of absorption, distribution, metabolism, and excretion of the compound (ADME)(Chan [10]). Based on the physiological sequence of events after a drug or nutrient has entered the body and the subsequent mechanism of interaction Chan [10] identified four broad categories of interactions; *I. Ex Vivo Bioinactivations*: This type of interactions involves chemical or physical reactions that take place before the drug and nutrient involved have entered the body (eg. in enteral nutrition or parenteral nutrition). *II. Absorption Phase–Associated Interactions*: Here, the nutrient may modify the function of an enzyme (type A interaction) or a transport protein (type B interaction) that is responsible for the biotransformation or transport of the drug prior to reaching the systemic circulation. According to the norm, many type II interactions can be avoided by allowing an adequate amount of time between drug intake and the consumption of the nutrient. *III. Physiologic Action–Associated Interactions*: There interactions occur after the absorption phase is complete for at least 1 of the interaction pairs. The mechanisms involve changing the cellular or tissue distribution, systemic metabolism or transport, or penetration to specific organs or tissues of the nutrient/drug (pharmacodynamic alterations). On this occasion, separation of administration times is not expected to resolve the problem. *IV. Elimination Phase–Associated Interactions*: These interactions may involve the modulation, antagonism, or impairment of renal or enterohepatic elimination.

1.1.2 Obstacles in identifying and predicting herbal-drug interactions

The task of adverse events reporting for interactions, positive or negative, between drugs and plant-based food is challenging, even for those phytochemicals that are already marketed as nutraceuticals or dietary supplements. A report published by the Department of Health and Human Services (DHHS) estimated that less than 1% of all drug – dietary supplement interactions are reported to the FDA through the MedWatch system for reporting adverse events(Chavez et al. [11]). Moreover, in a published systematic review, aiming to assess the interactions between herbal and conventional drugs (Fugh-Berman and Ernst [18]), it was found that only 13% of the suspected interactions that were extracted from online databases and evaluated, could be characterized as well-documented. The existing data that guide the clinical management of most drug-nutrient interactions are mostly anecdotal experience, uncontrolled observations, and opinions, whereas the science in understanding the mechanism of drug- nutrient interactions remains limited. This scarcity of published clinical evidence constitutes a major setback in the process of identifying and predicting herbal-drug interactions.

Another barrier in identifying potential interactions resides in the nature of plant-based food, itself. First of all, natural products usually consist of complex mixtures of bioactive compounds, whose exact chemical composition is often unknown. Further complicating the matter, the chemical constituents of plant-based food may vary according to the seasonality, growing conditions, or the specific part of the plant that is examined. For herbal dietary supplements, the variation in manufacturing processes, which are not standardized, nor regulated by the FDA, augments the overall complexity, as well (Chavez et al. [11]).

1.1.3 Overview of available methods

Known or suspected pharmacological activity, data derived from in vitro or animal studies, or isolated case reports are the common sources for gaining knowledge on the interference of dietary components with the pharmacokinetics and/or pharmacodynamic processes of medical substances (Chavez et al. [11]). However, the scarcity of clinical based evidence, combined with the frequent lack of comprehensive documentation, underlines the importance of using data mining computational techniques for effective extraction of large scale, relevant information and, furthermore, stresses the potential of developing in silico models for the prediction of unwanted side effects (SEs), caused by the intervention of phytochemicals with drug protein targets. In this respect, several SE prediction, systems biology computational methods that are widely employed in the pharmaceutical industry to predict drug-drug interactions (DDIs), such as pathway based methods or chemical similarity based methods, might also prove to be pertinent in predicting drug-food interactions. A recently published study by Jensen et al. [22] presented a systems chemical biology approach that integrated data from the scientific literature and online databases, to gain a global view of the associations between diet and dietary molecules with drug targets, metabolic enzymes, drug transporters and carriers currently deposited in DrugBank. Additionaly ,disease areas and drug targets that are most prone to the negative effects of drug-food interactions were identified, showcasing a platform for making recommendations in relation to foods that should be avoided under certain medications. Lastly, by investigating the correlation of gene expression signatures of foods and drugs novel drug-diet interactome map was generated.

1.2 Psycholeptic and Psychoanaleptic drugs

Since the discovery of the first psychiatric drugs, more than 50 years ago, pharmaceutical companies have been regarding psychopharmaceuticals as a major part of the overall prescription drug market. Although many efforts have been invested in discovering antidepressant, anxiolytic or antipsychotic agents with new, alternative, mechanisms of action, the original mechanisms on monoamine or GABAergic systems remain the basis of all currently available drugs. New approaches to discover new, cognitive enhancing, drugs, among many different ideas, have attempted to enhance cholinergic function, using selective muscarinic or nicotinic agonists. Other approaches include the development of selective dopamine D1 receptor agonists, or drugs that alter glutamatergic function by suppressing NMDA or AMPA receptor function (Iversen [20]). In the present study, two main subcategories of psychopharmaceuticals are examined for their interactions with plant-based food; the psycholeptics and the psychoanaleptics.

In pharmacology, a psycholeptic is a tranquilizer, or a medication which produces a calming effect upon a person. The psycholeptics are classified under N05 in the Anatomical Therapeutic Chemical Classification System (ATC), a system of alphanumeric codes developed by the WHO for the classification of drugs and other medical products. In ATC classification system, the active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. A psychoanaleptic, on the other hand, is a medication that produces an arousing effect. The psychoanaleptics are classified under N06 in the Anatomical Therapeutic Chemical Classification System. Subcategories of N06 include Psychostimulants (ex. amphetamines), agents used for ADHD, nootropics, anti-dementia drugs and antidepressants (for Drug Statistics Methodology [17], DS. et al. [16]).

Table 1.1: ATC N05 Psycholeptics (DS. et al. [16])

PSYCHOLEPTICS			
ANTIPSYCHOTICS		**HYPNOTICS AND SEDATIVES**	**ANXIOLYTICS**
Benzamides	***Phenothiazines with piperidine structure***	***Aldehydes and derivatives***	***Azaspirodecanedione derivatives***
Sulpiride	Propericiazine	Dichloralphenazone	Buspirone
Remoxipride	Thioridazine	Chloral hydrate	***Benzodiazepine derivatives***
Amisulpride	Pipotiazine	Paraldehyde	Oxazepam
Butyrophenone derivatives	Mesoridazine	***Barbiturates, plain***	Lorazepam
Melperone	***Thioxanthene derivatives***	Barbital	Adinazolam
Pipamperone	Thiothixene	Secobarbital	Tofisopam
Droperidol	Zuclopenthixol	Talbutal	Alprazolam
Haloperidol	Chlorprothixene	Pentobarbital	Chlordiazepoxide
Diazepines, oxazepines, thiazepines and oxepines	Flupentixol	Thiopental	Clobazam
Quetiapine		Amobarbital	Halazepam
Loxapine		Aprobarbital	Camazepam
Asenapine		Butethal	Ethyl loflazepate
Olanzapine		Heptabarbital	Cloxazolam
Clozapine		Hexobarbital	Bromazepam
Diphenylbutylpiperidine derivatives		Methohexital	Clotiazepam
Fluspirilene		***Benzodiazepine derivatives***	Fludiazepam
Pimozide		Estazolam	Ketazolam
Indole derivatives		Midazolam	Prazepam
Molindone		Flurazepam	Etizolam
Lurasidone		Triazolam	Diazepam
Sertindole		Temazepam	***Carbamates***
Ziprasidone		Brotizolam	Mebutamate
Lithium		Flunitrazepam	Meprobamate
Lithium		Quazepam	***Dibenzo-bicyclo-octadiene derivatives***
Other antipsychotics		Cinolazepam	Benzoctamine
Zotepine		Nitrazepam	***Diphenylmethane derivatives***
Aripiprazole		***Benzodiazepine related drugs***	Hydroxyzine
Paliperidone		Eszopiclone	Captodiame
Iloperidone		Zolpidem	***Other anxiolytics***
Risperidone		Zaleplon	Etifoxine
Cariprazine		Zopiclone	
Phenothiazines with aliphatic side-chain		***Melatonin receptor agonists***	
Acepromazine		Melatonin	
Promazine		Ramelteon	
Cyamemazine		***Other hypnotics and sedatives***	
Methotrimeprazine		Ethchlorvynol	
Chlorpromazine		Methaqualone	
Triflupromazine		Triclofos	
Phenothiazines with piperazine structure		Scopolamine	
Thioproperazine		Propiomazine	
Perphenazine		Dexmedetomidine	
Acetophenazine		clomethiazole	
Prochlorperazine		***Piperidinedione derivatives***	
Fluphenazine		Methyprylon	
Trifluoperazine		Glutethimide	

Table 1.2: ATC N06 Psychoanaleptics (DS. et al. [16])

PSYCHOANALEPTICS		
PSYCHOSTIMULANTS, AGENTS USED FOR ADHD AND NOOTROPICS	**ANTIDEPRESSANTS**	**ANTI-DEMENTIA DRUGS**
Centrally acting sympathomimetics	Desipramine	***Anticholinesterases***
Amphetamine	Protriptyline	Galantamine
Pemoline	Dosulepin	Donepezil
Dextroamphetamine	***Other antidepressants***	Tacrine
Methamphetamine	Oxitriptan	Rivastigmine
Atomoxetine	Tianeptine	***Other anti-dementia drugs***
Lisdexamfetamine	Venlafaxine	Memantine
Methylphenidate	Milnacipran	Ginkgo biloba
Modafinil	Nomifensine	
Dexmethylphenidate	Mianserin	
Fencamfamine	Vortioxetine	
Fenethylline	Reboxetine	
Other psychostimulants and nootropics	Agomelatine	
Aniracetam	Vilazodone	
Acetylcarnitine	Nefazodone	
Adrafinil	Duloxetine	
Idebenone	Bupropion	
Piracetam	Desvenlafaxine	
Xanthine derivatives	L-Tryptophan	
Caffeine	Minaprine	
propentofylline	Trazodone	
Monoamine oxidase A inhibitors	Mirtazapine	
Toloxatone	Viloxazine	
Moclobemide	***Selective serotonin reuptake inhibitors***	
Monoamine oxidase inhibitors, non-selective	Paroxetine	
Iproclozide	Zimelidine	
Isocarboxazid	Citalopram	
Nialamide	Sertraline	
Tranylcypromine	Fluoxetine	
Phenelzine	Escitalopram	
Non-selective monoamine reuptake inhibitors	Fluvoxamine	
Nortriptyline	Etoperidone	
Amoxapine		
Clomipramine		
Trimipramine		
Amineptine		
Amitriptyline		
Maprotiline		
Dimetacrine		
Imipramine		
Butriptyline		
Doxepin		

Chapter 2

Food Interaction Networks for Psycholeptic and Psychoanaleptic Drugs

2.1 Introduction

2.1.1 Assessing the pharmacodynamics and pharmacokinetics effects of drug-food interactions through identification of shared protein targets

While attempting to assess the effects (positive or negative) of plant-based food on the pharmacological action of drugs, a significant dichotomy is observed. Herbal–drug interactions can be characterized as either pharmacodynamic (PD) or pharmacokinetic (PK) in nature.

PK interactions imply an alteration of the absorption, distribution, metabolism, or elimination properties (ADME) of a conventional drug by an herbal product. Understanding ADME requires an insight in the bioavailability of the drug candidate from the route of administration to the ultimate site of activity, for the required duration of time, in order to elicit the intended pharmacology. On the first stages, factors that influence the absorption of a drug (usually for oral administration), are the dissolution and solubility within the gastrointestinal lumen, luminal behavior, enterocyte permeability, and intestinal and liver metabolism. Later on, the distribution phase initiates, once the drug reaches the systemic circulation. Thereupon, the intended ADME are dictated by plasma pharmacokinetics. On the final stage, the duration of drug activity is influenced by drug metabolism and elimination (Dowty et al. [15]). In order to study the pharmacokinetics of drugs, the interactions of phytochemicals with proteins involved with ADME (metabolic enzymes, transporters, carriers) should be reviewed.

Pharmacodynamic (PD) interactions may occur when constituents of herbal products have either synergistic or antagonist activity in relation to a conventional drug. As a result, the bioavailability of a therapeutic molecule is altered at the site of action, at the drug-receptor level. To get an insight of the pharmacodynamic processes that are affected by a phytochemical, interactions with drug targets should be studied (Chan [10]).

2.2 Methodology

2.2.1 Data availability

There are not many available collections of resources, regarding the molecular composition of food and the related biological activities of its phytochemical content. Examples of ongoing efforts include the Danish Food Composition Database, the Phenol-Explorer [32], the KNApSAcK Family Database [1] and the MAPS database. However, a significant shortage of these databases, from a systems biology scope, is the lack of chemical structures and high-throughput retrieval of data [21].

In order to better understand the association of phytochemicals with disease related protein targets, the NutriChem database was utilized, developed by Jensen et al. [21], and aiming to provide an exhaustive resource, listinig the molecular content of plant-based food and the effects of dietary interventions. To build NutriChem, 21 million MEDLINE abstracts (1908-2012) were processed by text mining, to detect information concerning links of plant - based food with its constituents and links of food with human disease phenotypes. A Naive - Bayes classifier was trained to recognize these links and, subsequently, validated using an external dataset of 250 abstracts, yielding an accuracy of

88.4% in discovering food - phytochemical pairs and 84.5% for food - disease pairs. An association of phytochemicals with disease on a molecular level ensued, by identifying phytochemicals with experimental bioactivity against disease associated protein targets, using ChEMBL database and the Fisher's exact test to systematically discover associations from the mined data pairs of food - phytochemicals and food - diseases. This effort resulted in a content of 18487 pairs of 1772 plant - based food and 7898 phytochemicals, along with 6442 pairs of 1066 plant - based foods and 751 diseases phenotypes. Moreover, predicted associations were generated for 548 phytochemicals and 252 diseases. The current version of NutriChem 1.0 database allows querying by plant- based food name of id, by disease, or by phytochemical coumpound. In the present study, NutriChem was used to construct the networks of drug - food space interactions, by extracting data concerning which phytochemicals interact with relevant drug targets and in what plant - based food these phytochemicals can be found.

In order to construct the drug - protein target interaction network for psycholeptic and psychoaanaleptics, data was extracted from the Drugbank database (DS. et al. [16]), including information about primary and secondary human protein targets, enzymes, transporters and carriers.

2.2.2 Data processing

To promote a broader understanding of the interactions between the plant-food space and the drug space of psycholeptic and psychoanaleptic agents, the protein target networks were constructed.
The plant-based food compounds and their human protein targets were retrieved from NutriChem 1.0 database . Thereupon, the matching of phytochemicals to their specific plant-based food of origin was verified by manually curating the data retrieved from NutriChem. This process was carried out by closely examining the references cited for each food-compound pairing. Because of the innate characteristics of the algorithm used to mine data from PubMed abstracts, a small number of false positive food-compound pairings were observed. Those errors occurred usually for one of the following reasons:

- The reference text included an experimental assay of more than one plant-based foods. In that case the algorithm assigned the phytochemicals present in the text to each and every plant-based food that was also mentioned.

- The phytochemicals were associated with parts of the plant-based food that are not edible. For example Serotonin was mentioned to be present in leaves and seeds of the Solanum tuberosum, but not in the tubers, which are the edible part of the potato.
- The mined text referred to an experimental assay testing a plant-based food for specific compounds of interest, and concluded that they were not found in the plant.
- The chemical compound was artificially added and used as part of a pretreatment method for an experimental assay.

It is noted that all proteins in the interactions networks are represented by their Uniprot ID [13], all drugs are mentioned by their Drugbank codes [16], and finally, all plant-based food is cited by its id in NCBI Food Taxonomy (`http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=advisors`) . The full data tables for the networks are attached as supplementary material S7 and S8.

2.3 Drugs to protein target interactions network

In order to construct the drug - protein target interaction network for psycholeptic and psychoanaleptics, data was extracted from the Drugbank database (DS. et al. [16]), including information about primary and secondary human protein targets, enzymes, transporters and carriers. The pharmacological action of each drug-protein pair was either known, or unknown.

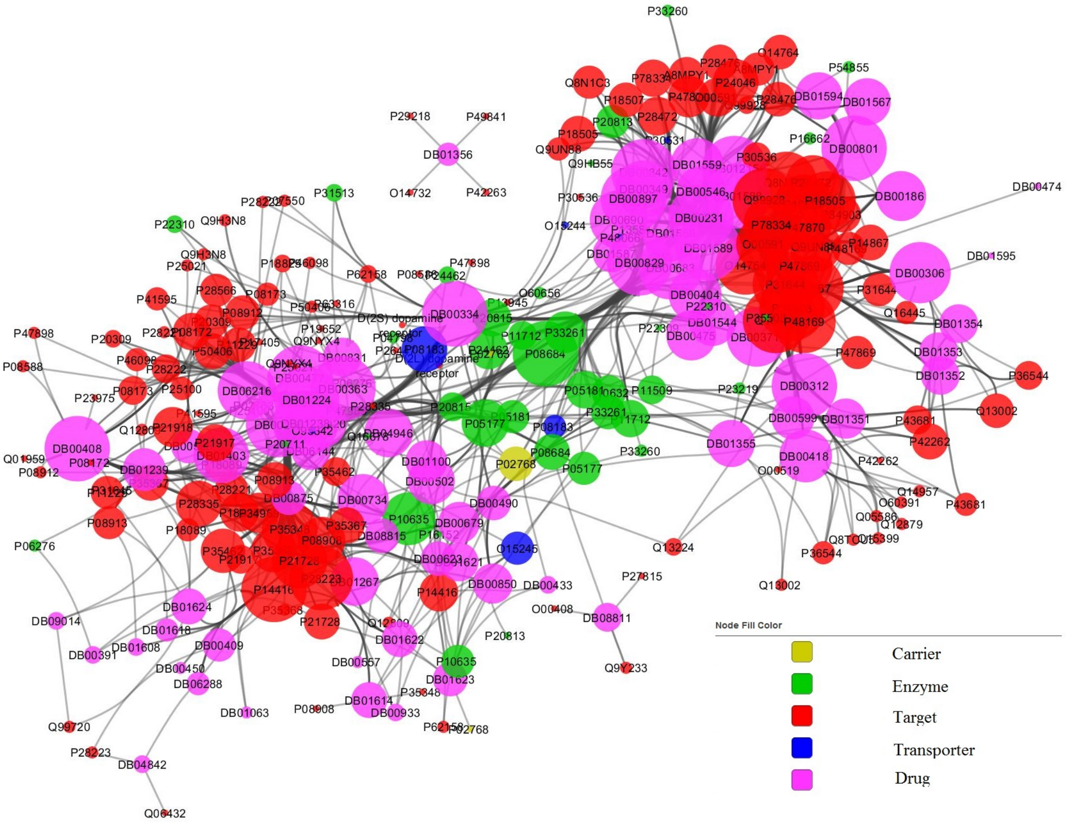


Figure 2.1: Network mapping of psycholeptic drugs to their human protein targets, enzymes, transporters and carriers. Node size is proportionate to the degree of connectivity for each node.

A total of 77 psycholeptic drugs were found to be interacting with a summation of 191 unique protein targets. Among the proteins, 7 transporters, 141 targets, 3 carriers and 40 enzymes could be identified. In Figure 2.1 the interaction network is presented. In this graph it becomes apparent that two major subcategories of drugs can be distinguished, based on their targeting of different human proteins targets. The first one occupies the left, lower, subset area of the network, and comprises mostly of antipsychotic agents, while the second one occupies the right subset area, and is made up of anxiolytics, hypnotics and sedatives. However, most of the enzymes,carriers and transporters are located roughly in the middle of the interaction network, indicating that they are targeted often by antipsychotics, anxiolytics, hypnotics and sedatives, alike.

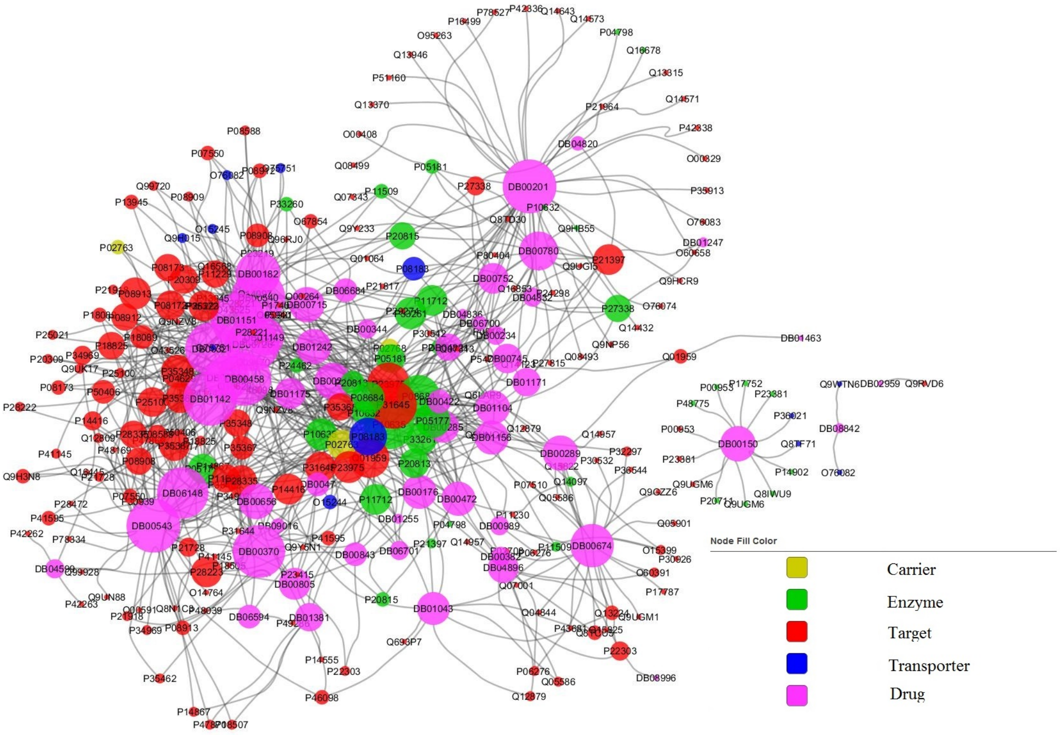


Figure 2.2: Network mapping of psychoanaleptic drugs to their human protein targets, enzymes, transporters and carriers. Node size is proportionate to the degree of connectivity for each node.

Next, the ATC N06 subcategory of 57 psychoanaleptic drugs was matched to 246 human proteins, of which, 13 were transporters, 3 carriers, 41 enzymes and 189 protein targets. The interaction network for psychoanaleptics is presented in Figure 2.2. While observing the networks, it is noted that several proteins can qualify as targets of both psycholeptic and psychoanaleptic drugs.

2.4 Food to protein target interactions network

Several constituents of plant-based food were found to be active against some of the protein targets of psycholeptic and psychoanaleptic drugs that were mentioned above. The network mapping of plant-based food to their human protein targets was carried out, including only the protein targets, enzyme, carriers and transporters that were also targeted by psycholeptic or psychoanaleptic drugs.

For psycholeptic drugs, 12 protein targets and 1 metabolic enzyme (P06276) were also targeted by a variety of plant-based food nutrients. As each food node is connected with a protein node by more than one edges, forming a multigraph, it is apparent that some foods may interfere with psycholeptic drugs by means of multiple protein-nutrient interactions. The most commonly targeted protein by nutrients is the enzyme P06276 or human Cholinesterase BCHE, an esterase with broad substrate specificity that is known to contribute to the inactivation of the neurotransmitter acetylcholine. It can, also, degrade neurotoxic organophosphate esters [13]. Sorting the protein nodes by degree of connectivity, proteins P49841 and P21728 emerged as the next most common proteins targeted by food. Figure 2.3 illustrates the food-protein network for psycholepic drugs. It is worth noting that the proteins often targeted by phytochemicals do not coincide with popular protein targets in the network of psycholeptics, and furthermore, most of them target proteins related to antipsychotic drugs, rather than sedatives/ anxiolytics .

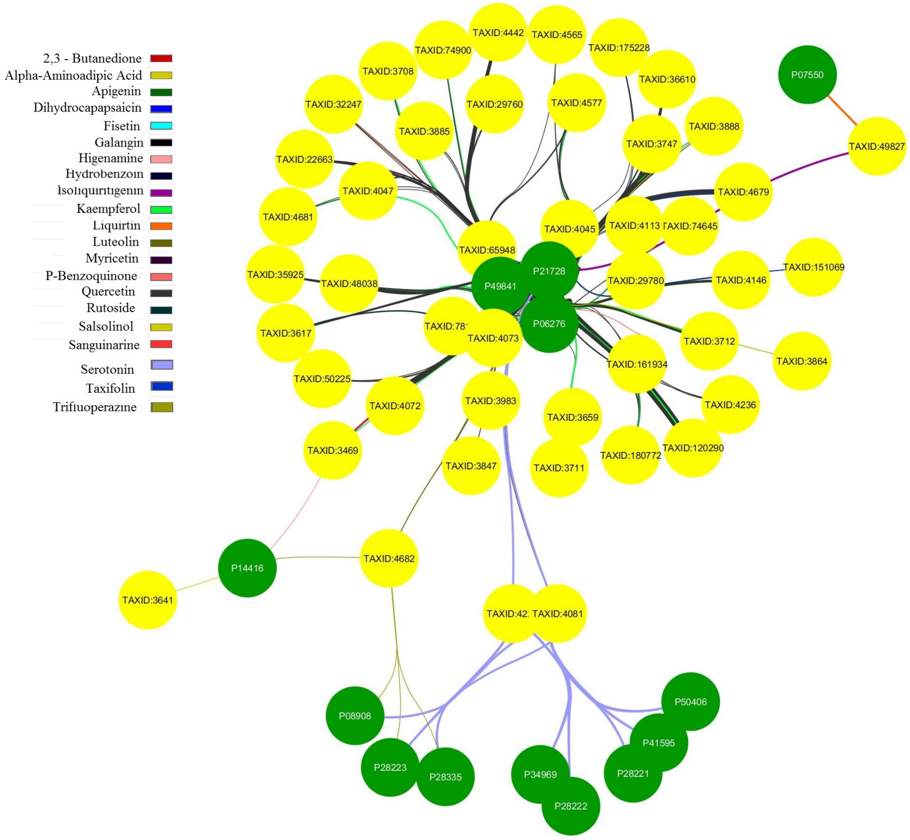


Figure 2.3: Network mapping of plant-based food (yellow marked nodes) to their human protein targets (green marked nodes). Included are only the protein targets, enzyme, carriers and transporters that are also targeted by psycholeptic drugs. Different colored edges represent different active compounds contained in plant-based food.

The equivalent network for psychoanaleptic drugs features 21 protein targets and 2 metabolic enzymes (P27338, P06276). Furthermore, the proteins most commonly targeted by phytochemicals are P22303, P21728, P06276, P41143 and P41145. Figure 2.4 illustrates the food-protein network for psychoanalepic drugs.

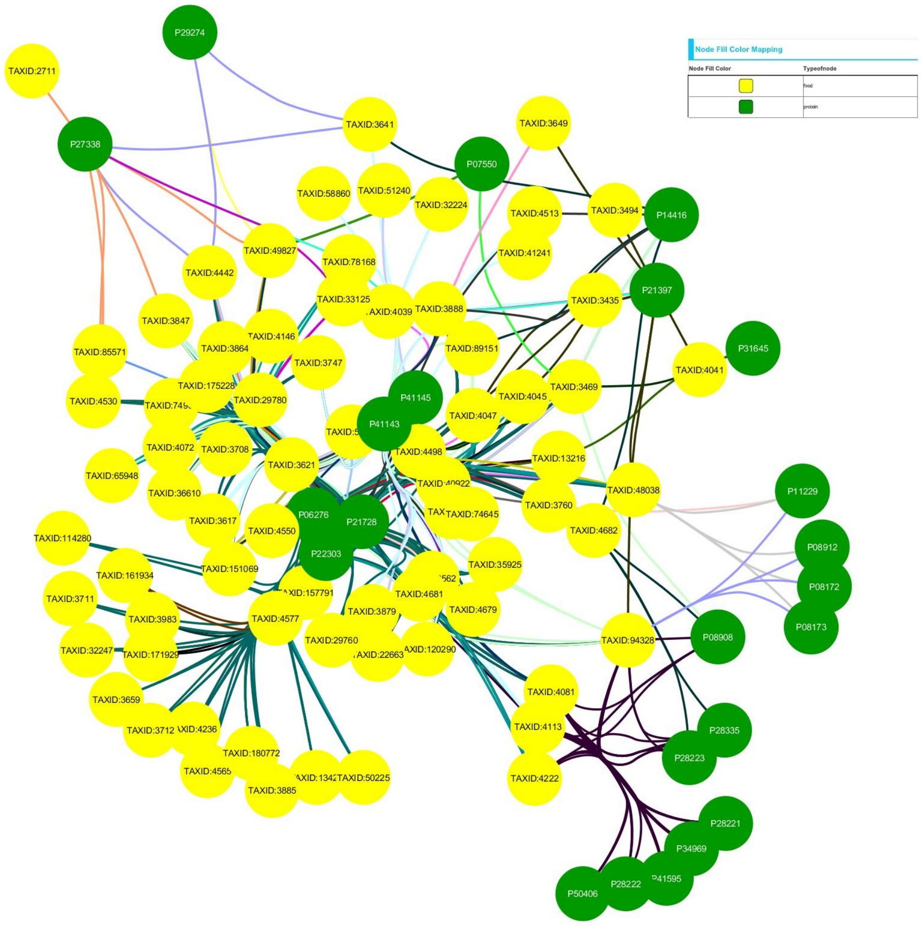


Figure 2.4: Network mapping of plant-based food (yellow marked nodes) to their human protein targets (green marked nodes). Included are only the protein targets, enzyme, carriers and transporters that are also targeted by psychoanaleptic drugs. Different colored edges represent different active compounds contained in plant-based food.

2.5 The drug - food space interaction network

All the available data can be combined in the construction of the food-drug space interaction networks in Figures 2.5 and 2.6. Most of these interactions indicate possible alterations of the pharmacodynamic properties, since they mostly revolve around protein targets, and less around transporters, carriers or metabolic enzymes. Additionally, not all proteins that are co-targeted by phytochemicals have a known pharmacological action for the drugs in question.

In Figure 2.5 a large subgroup of psycholeptic drugs is predicted to interact with several herbal dietary components, whereas a smaller subgroup is found to be targeted only by 3 plant-based foods; safflower (TAXID:4222),tomato (TAXID:4081) and garlic (TAXID:4682). Those are also the foods with the higher degree of connectivity in the entire graph, interacting with the majority of psycholeptic drugs, through multiple proteins.

In Figure 2.6 two subcategories of food nodes are distinguished. The first one, featuring strawberry (TAXID:3747), onion (TAXID:4679), quinces (TAXID:36610), sweet-pepper (TAXID:4072) and swede (TAXID:3708), among others, seems to interact mainly with a group of 8 psycoanaleptics (DB00382, DB00989, DB00370, DB08996, DB00321, DB00543, DB00674, DB00805, DB00726). The second one, where ginger (TAXID:94328), poppy-seed (TAXID:3469)and safflower (TAXID:4222) are the most prominent, shares protein targets with all drugs in the network.

In Figures 2.7 and 2.8 the complete interaction networks are demonstrated. The protein targets in these graphs are all targeted by food nutrients and drugs, alike. However, different topologies are present in the graphs, where targets are mostly interacting with food, as opposed to drugs, and vice versa.

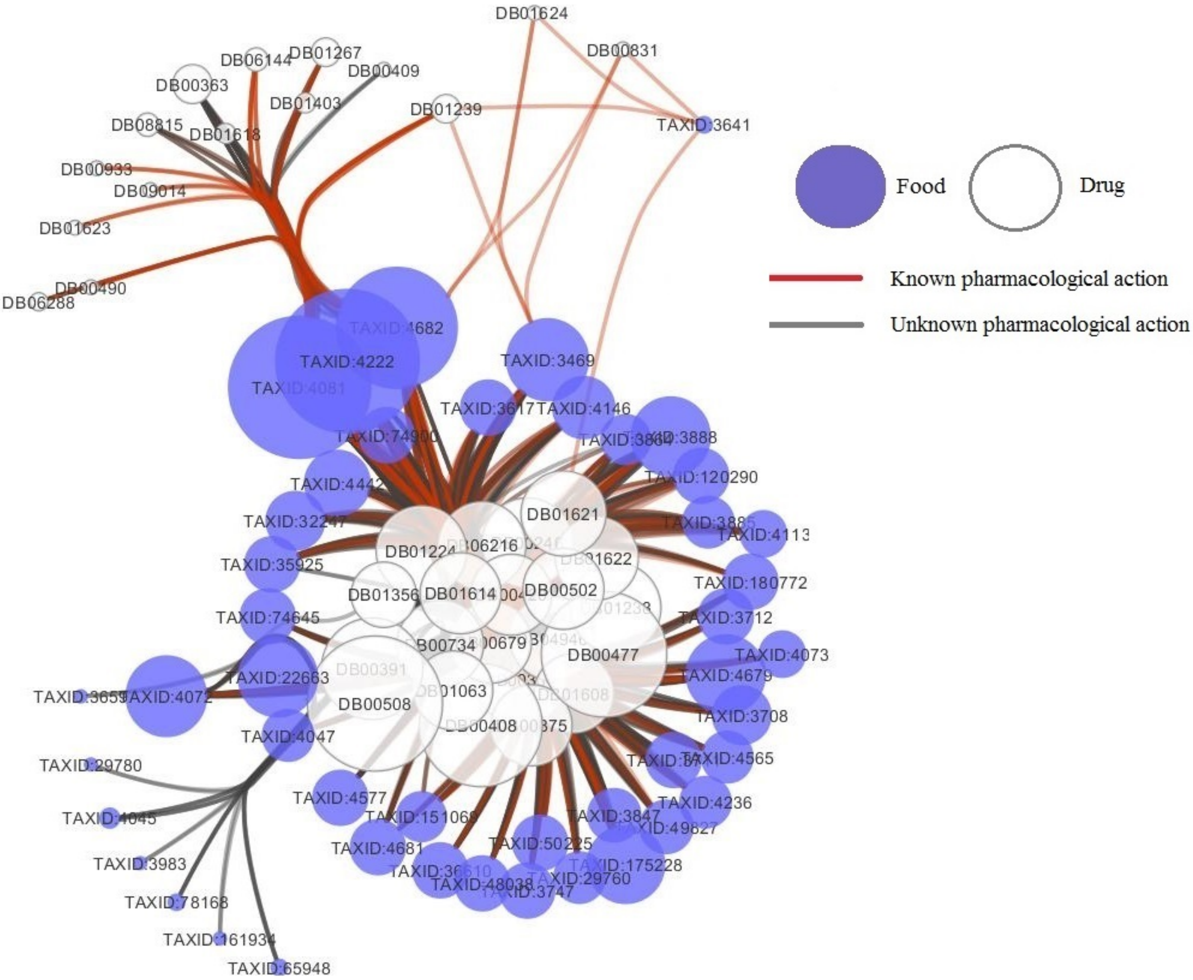


Figure 2.5: Drug to plant-based food interaction network, where each edge represents a common protein target between a psycholeptic drug and a phytochemical, belonging to an edible plant. Node size is proportionate to the connectivity degree of the node. All edges have been bundled in order to obtain a clear, representative image of the dense network. Marked red are the edges that signify a known pharmacological action of the drug on the protein target.

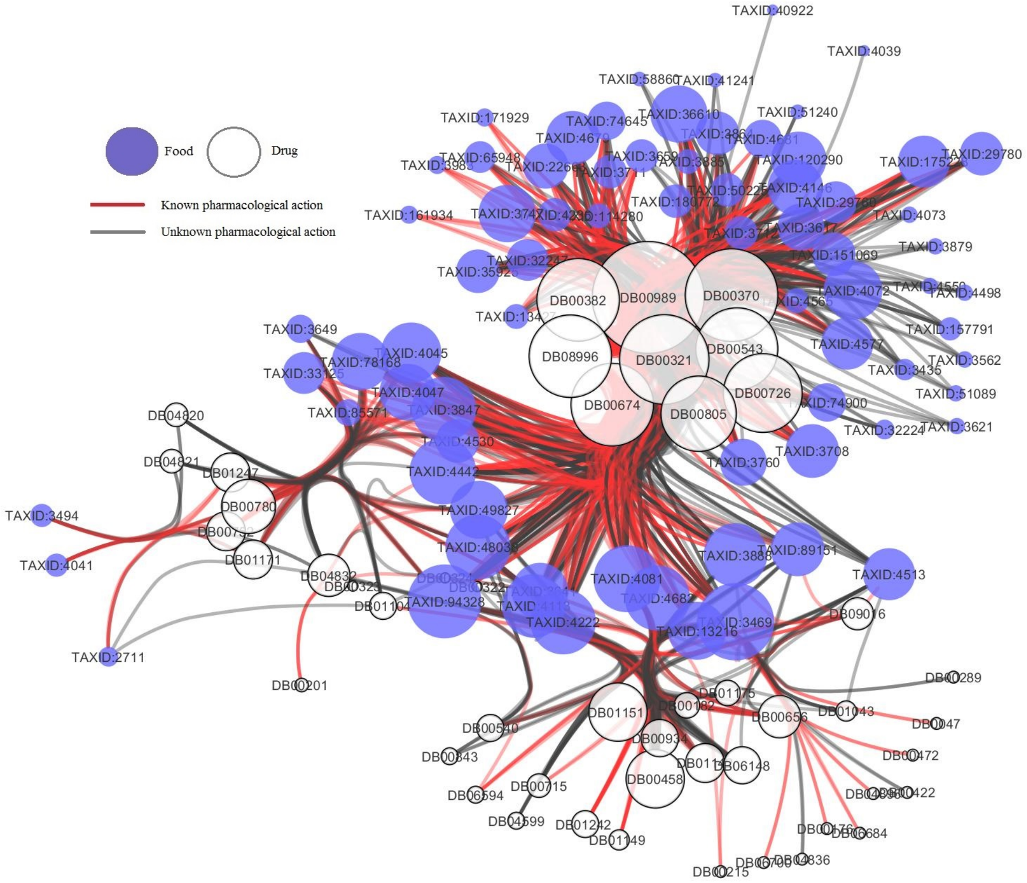


Figure 2.6: Drug to plant-based food interaction network, where each edge represents a common protein target between a psychoanaleptic drug and a phytochemical, belonging to an edible plant. Node size is proportionate to the connectivity degree of the node. All edges have been bundled in order to obtain a clear, representative image of the dense network. Marked red are the edges that signify a known pharmacological action of the drug on the protein target.

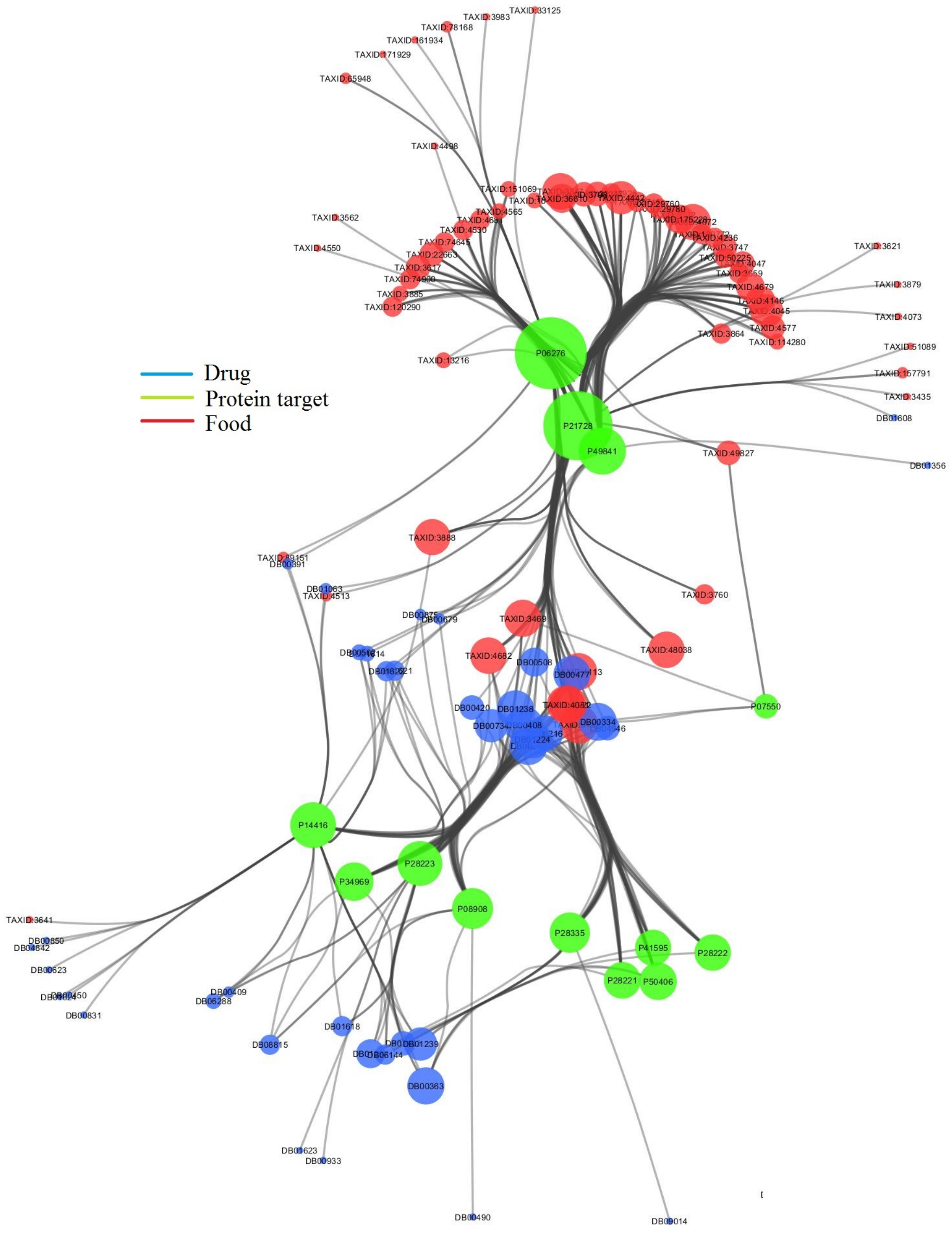


Figure 2.7: Network mapping of psycholeptic drugs and plant-based food to the human proteins that they target. Node size indicates the degree of connectivity of each node.

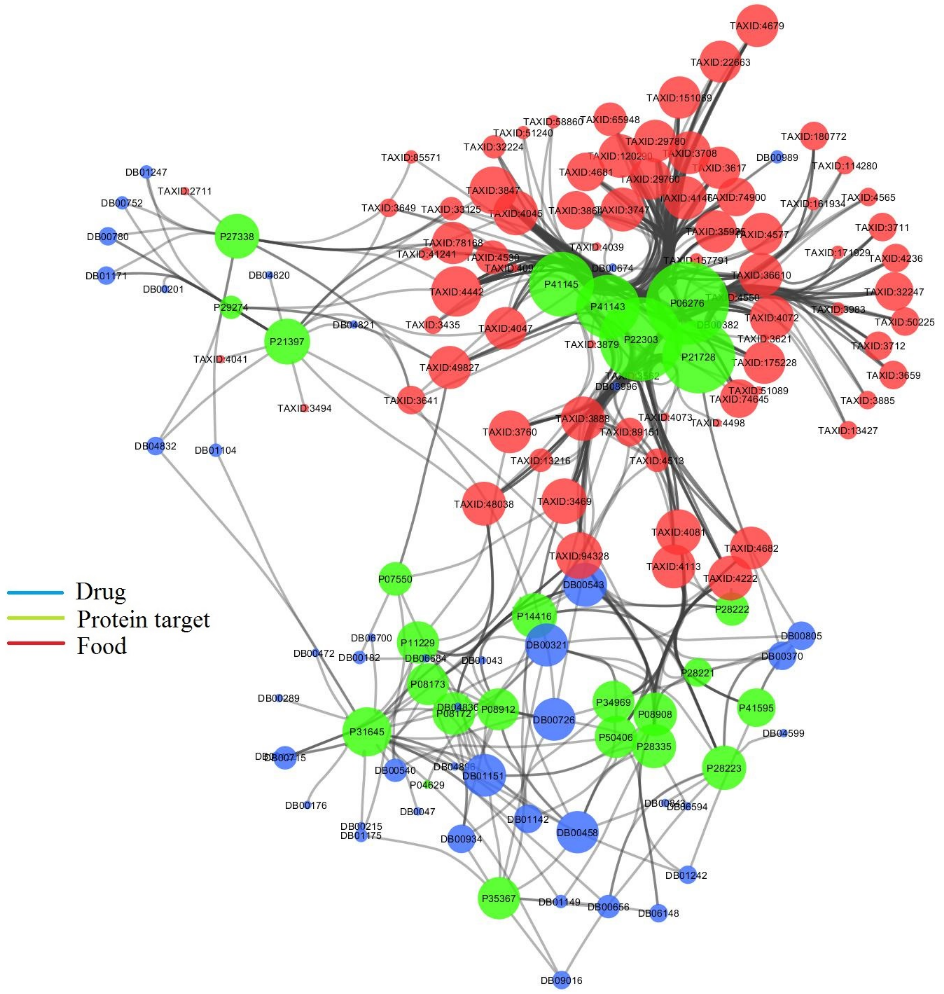


Figure 2.8: Network mapping of psychoanaleptic drugs and plant-based food to the human proteins that they target. Node size indicates the degree of connectivity of each node.

Chapter 3

Pharmacophores for phytochemicals interacting with psychiatric drug targets

3.1 Introduction

In the next part of this study, a search for additional phytochemicals that would be expected to interact with targets of psychiatric drugs was carried out. More specifically, three protein targets, P07550, P28222 and P14416 were selected, based on their frequency of interaction with both psycholeptics and psychoanaleptics and the availability of 3D structural information in PDB. For these protein targets, ligand-based pharmacophoric hypothesis were generated, using experimental activity data from the literature (ChEMBL database), and the HypoGen feature of Accelrys Discovery Studio (DS). The pharmacophore models were validated using external test sets and Fisher's method to evaluate the ability of pharmacophores to differentiate between experimentally known actives. Finally, the valid pharmacophores were used as queries to screen for more potentially active phytochemicals from the NutriChem 1.0 database.

3.1.1 Storing 3D structures

A fundamental problem in Chemoinformatics is the lack of a priori information available, that would reveal what the 3D structure of a compound would be. Moreover, all compounds are flexible to some degree, so the 3D structure might change over time. Experimental 3D structure data, derived either from X-ray crystallography or NMR spectroscopy, can only get us so far, and they are often unable to predict the form that the compound will take for example when binding to a protein target. Instead, computational 3D flexibility analysis can be employed to tackle this problem [43]. Most compounds have rotatable bonds, which means that the whole molecule can flex into many different poses in 3D, which are referred to as different ”conformers”. An empirical definition for a rotatable bond is: any single bond which is not part of a ring, is not terminal (e.g. Methyl) and is not in a conjugated system (e.g. an Amide). It is understood that a single molecule can have an infinite or less than infinite number of conformers (if discrete rotation units are considered). Conformers of a compound can be unequal in regard to their energy levels, and usually the most low energy (and thus stable) conformers are encountered. Conformation flexibility can be dealt with either at the representation level, by storing multiple conformers, or at the algorithm level, by providing just one representation of the 3D structure and letting algorithms flex the molecule as needed [43]. While only information about the atoms and how they are connected by bonds is required in order to store 2D structural data, the coordinates of atoms relative to some origin and conformation analysis is additionally required to create 3D databases. Usually a connection-table type file format, often either an MDL MOL or SD File or a Sybyl MOL2 file is used to store 3D structural information. Other formats can be used too, such as CML, PDB and for the coordinates simply an XYZ file [43].

3.1.2 QSAR and Virtual Screening

QSAR stands for Quantitative Structure-Activity Relationships. It refers to the mathematical correlation of structural features with experimental activity results, for multiple compounds, usually in the same series. Biological activity is then translated into a function of molecular descriptors (structural or physicochemical properties). If the model generated is successful, we can use this function as a predictive tool for compounds, where the activity is unknown, but the descriptors can be calculated. QSAR may assume linear, or non linear functions to describe activity. When non linear QSAR is carried out, methods such as Neural Networks, Decision Trees, Support Vector Machines or Bayesian classifiers can be employed [43].

One very common application of QSAR models, particularly nonlinear models is in virtual screening. This means using informatics to predict which of a large number of compounds will bind to a protein target or inhibit a cellular assay. It is, in other words, a computational equivalent to high throughput biological assaying. Other methods, alternative to QSAR for virtual screening, include simple chemical similarity to known active compounds, and molecular docking when protein target structure information is available. When QSAR methods are used for virtual screening, it is common to use classification models, that discern between active and inactive compounds, as well as quantitative prediction models (which attempt to predict a strength of binding) [43].

3.1.3 Pharmacophores

A pharmacophore is a generic set of molecular features that is required for binding to a biological macro-molecule. It is used to refer to structural features (or derivatives such as hydrogen bonding potential), and is usually used in reference to 3D structures. It may also be defined as set of features and distance bounds of these features from each other in 3D [43]. It does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their targets structure. Thus, a pharmacophore represents chemical functions, valid not only for one currently bound ligand, but also for unknown molecules [42].

A pharmacophore can be represented in a variety of ways: e.g. a distance matrix of pharmacophore points (with a dictionary for point types which may contain coordinates of 3D substructures or SMARTS of 2D features). The generic structure described by the pharmacophore needs to able to represent distance ranges (rather than exact ranges) and incorporate ambiguity in pharmacophore points [43]. Pharmacophoric features can be defined in the 2D or 3D space. When 2D pharmacophores are used, the information concerning the relative position of features can be stored as a pharmacophore fingerprint (1D vector storing relative distances between the features). Common pharmacophore features include the H-bond acceptor, H-bond donor, anionic, cationic, hydrophobic and aromatic features. An atom is considered an acceptor if it can attract an hydrogen (nitrogen, oxygen or sulfur and not an amide nitrogen, aniline nitrogen and sulfonyl sulfur and nitro group nitrogen), and a donor if it can give an hydrogen [34].

A pharmacophore can be used as a potent tool in virtual screening, to query 3D databases of compounds, in order to find molecules that bind to a particular protein. A pharmacophore search resembles a substructure search, as it comprises of a sub-graph query on a fully-connected distance matrix graph [43]. It can be used to filter a ligand database, prior to docking simulations, or as a post filtering tool of docking results to remove compounds that don't bind according to the pharmacophore query. Moreover, the pharmacophore alignment can be used to guide the placement during a docking session. Other applications of pharmacophores include target identification, prediction of drug side-effects, ADME-tox profiling and 3D QSAR analysis [34].

An instance of pharmacophore based virtual screening was applied in the early stages of drug discovery for alternatives to Rimonabant. Rimonabant is an anorectic anti-obesity drug produced and marketed by Sanofi-Aventis. It is an inverse agonist for the cannabinoid receptor CB1. Its main avenue of effect is reduction in appetite. Rimonabant is the first selective CB1 receptor blocker to be approved for use anywhere in the world. It is approved in 38 countries including the E.U., Mexico, and Brazil. However, it was rejected for approval for use in the United States. This decision was made after a U.S. advisory panel recommended the medicine not be approved because it may increase suicidal thinking and depression [16]. During the lead discovery of alternatives for Rimonabant, compounds were examined as potential cannabinoid receptor CB1 antagonists. The cannabinoid receptor's crystal structure was yet unknown and only some homology models were available in the literature. In order to build the pharmacophore for CB1, 8 CB1 selective antagonists/inverse agonists were selected from the literature and a maximum of 250 unique conformations were generated for each molecule (with Macromodel using the MMFF94s force field). The pharmacophore was generated using the Catalyst software and the resulting pharmacophore was used to screen a library of about 500k compounds (max. of 150 conformations per molecule, generated with Catalyst). The pharmacophore search resulted in 22794 hits (approx. 5% of the database), which were subsequently filtered based on the desired physicochemical properties. The remaining 2100 compounds were then clustered into 420 groups, using a maximum dissimilarity clustering algorithm and one compounds from each group was isolated, based on a Bayesian ranking model. Finally, the 420 compounds were screened at a single concentration. Out of these, only five compounds showed more than 50% inhibition. All five compounds were confirmed in the full curve assay. The screening process yielded 1 compound with a K_i activity of less than 100 nM [40].

Challenges in pharmacophore modeling are encountered due to the complexity of protein structures, when for example, multiple binding sites are present. Additionally, one active site may permit several binding modes and therefore, several pharmacophores may be fit to describe the protein-ligand binding [34].

3.1.4 Pharmacophore generators

A pharmacophoric hypothesis can be generated using either ligand - based approaches or structure - based methods. On the first occasion, where biological activities of multiple compounds are known, a sophisticated class of computational techniques can be used to deduce features required for biological activity and identify pharmacophores. A drawback of these ligand - based technigues is their inability to provide detailed structural information, required to design new molecules in drug discovery. On the second occasion, structure - based pharmacophore methodology is more reliable, as it imposes constraints required for interaction selectivity [2]. However, structure - based methods require comprehensive knowledge of the protein's structure and the ligand - protein interactions in the binding pocket.

There are several commercial packages that enable pharmacophore modeling, such as the Accelrys Discovery Studio (former Catalyst), Tripo's GASP and GALAHAD, Ligandscout by Inte:ligand, MOE by the Chemical Computing Group and Schrödinger's Phase software [34].

Three pharmacophore generation protocols are provided by the Accelrys Discovery Studio (the commercial successor of the Catalyst software); the *Receptor - Ligand Phar-*

macophore Generation protocol, the *Common Feature Pharmacophore Generation* protocol and the *3D QSAR Pharmacophore Generation* feature [7]. The commercial program Discovery Studio has a well documented process of pharmacophore generation and statistical analyses to give indications of the validity of the hypotheses generated. While much of the statistical analysis is automatically generated, the interpretation of that output as well as bias from user input can greatly affect the outcome, and therefore interpretation of the results forms an important component of such studies. Below is a brief summary of the pharmacophore generation features of Discovery Studio.

The *Receptor - Ligand Pharmacophore Generation* feature utilizes interactions between receptor-ligand complexes to generate a hypothesis. The X-ray crystal structures of such complexes become increasingly available in online databases (e.g. PDB database overseen by the Worldwide Protein Data Bank organization). This structured-based method generates selective pharmacophore models based on receptor-ligand interactions. To build the pharmacophore, a set of features is identified from the binding ligand and models derived from different combinations of these features are ranked, based on measures of sensitivity and specificity. Selectivity for the models is estimated, using a Genetic Function Approximation (GFA) model. Descriptors for the GFA are the the number of total features in pharmacophore models and the feature-feature distance bin values [2].

The *Common Feature Pharmacophore Generation* protocol utilizes the HipHop algorithm. It is able to generate pharmacophore models only by identification of the common chemical features shared by the active molecules received as input and their relative alignment to the common feature set, without having to consider biological data [28]. The algorithm can also optionally use information from inactive ligands to place excluded volume features. HipHop identifies configurations or three-dimensional spatial arrangements of chemical features that are common to molecules in a training set. The configurations are identified by a pruned exhaustive search, starting with small sets of features and extending them until no larger common configuration is found. Training set members are evaluated on the basis of the types of chemical features they contain, along with the ability to adopt a conformation that allows those features to be superimposed on a particular configurations. The user can define how many molecules must map completely or partially to the pharmacophore. This option allows broader and more diverse pharmacophores to be generated. The resultant pharmacophores are ranked as they are built. The ranking is a measure of how well the molecules map onto the proposed pharmacophores, as well as the rarity of the pharmacophore model. If a pharmacophore model is less likely to map to an inactive compound, it will be given a higher rank; the reverse is also true. [7]

The *3D QSAR Pharmacophore Generation* feature utilizes the HypoGen algorithm to derive Structure Activity Relationship (SAR) 3D pharmacophore models from a set of molecules for which activity values are known (predictive pharmacophores). It consist of a stepwise process, that receives as input data concerning ligand 3D structure and associated biological activity. The algorithm can be modified to place excluded volumes in key locations in an attempt to model unfavorable steric interactions.[7]

3.1.5 The HypoGen Module of Discovery Studio

When using the 3D QSAR Pharmacophore Generation feature in Discovery Studio, the quality of the training set substantially affects the significance of the hypothesis generated. It should consist of at least 16 compounds, spanning a minimum of 4 orders of magnitude of activity. Redundant data (i.e. compounds whose structural information – and therefore biological activity – essentially explain the same structure/activity outcome) should be removed as its inclusion can bias the output. The training set should not contain compounds known to be inactive due to steric interactions with the receptor, that is, exclusion volume problems, as DS is not equipped to handle such cases as it does not have the capability to understand features that have a negative impact on activity. Inclusion of these compounds would therefore lead to a bias in the pharmacophore.[29]

The *3D QSAR Pharmacophore Generation* protocol carries out conformation analysis for the input ligands, using an algorithm developed specifically to ensure good coverage of conformational space within a minimal number of conformers. The program generates a maximum of 256 different poses for each molecule, all within a specified energy range, and selected so that differences in inter-function distances are maximized. Chemical features from all the conformations are considered by HypoGen. A maximum of five features obtained, and can include hydrogen bond donors, hydrogen bond acceptors, hydrophobic features (aliphatic and aromatic) and ionisable groups, among others. These chemical features are defined within Discovery Studio in a dictionary using the CHM language and are based on atomic characteristics.[29] The in-built HypoGen module is then able to use all this information to generate the top ten scoring hypothesis models. This is performed in three phases [29]:

1. The constructive phase generates hypotheses that are common among the most active compounds. This is done in several steps; first the most active compounds are identified, then all hypotheses among the two most active compounds are determined and stored, those that fit the remaining active compounds are kept.

2. The subtractive phase then removes the hypotheses that fit the inactive compounds as well. This is performed by determining the inactive compounds, defined as having an activity 3.5 orders of magnitude greater than the most active compound. Any hypothesis that matched more than half the compounds identified as being inactive is removed.

3. The final phase is the optimization phase. This involves each hypothesis undergoing small perturbations in an attempt to improve the cost of the model. Examples of some of the alterations include, rotating vectors attached to features, translating pharmacophore features, adding a new feature or removing a feature. The ten highest scoring unique hypotheses are then exported.

These ten returned hypothesis models are analyzed to determine the best model. This process involves a mapping analysis, as well as a thorough cost analysis (statistical analysis) to determine which hypotheses are the most likely to be an accurate representation of the data.

The mapping analysis of HypoGen makes the assumption that an active molecule should map more features than an inactive molecule. Therefore a molecule should be inactive because it either does not contain an important feature, or misses the feature as it cannot be orientated correctly in space. Based on this assumption the most active compounds should map all features of the hypothesis model. [29]

The most accurate hypothesis model can be distinguished by plotting for each hypothesis a graph of the estimated activities against the actual activities. By calculating the line of best fit, a correlation value is obtained for each different hypothesis, allowing for a direct evaluation of the models performance. [29]

A major assumption used within DS in the generation of hypothesis models is based upon Ocram's razor, which states that between otherwise equivalent alternatives, the simplest model is the best. Aiming to quantify the simplicity of the models, costs are assigned to hypotheses in terms of the number of bits required to generate them. The total hypothesis cost is calculated using the three cost factors [29]:

1. The weight cost - increases as the feature weights in a model deviate from an ideal value
2. The error cost - increases as the RMS difference between the estimated and actual activities for the training set molecules increases.
3. Configuration cost - a fixed cost that depends on the complexity of the hypothesis space being optimized. Therefore, the lower the cost of a hypothesis the less bits required to generate it and the simpler the model.

An analysis of the costs of generating the pharmacophore can also serve as a means to validate the significance of the model. The greater the difference between the null cost and the total cost the more statistically valid the hypothesis, and thus, the greater the probability of this model being a true representation of the data. The null cost is the cost of generating a hypothesis where the error cost is high. The total cost is the actual cost of hypothesis generation, and the fixed cost is where the error cost is minimal (perfect pharmacophore). If the difference between the total cost and the null hypothesis cost is more than 60 bits, there is greater than 90% probability that the model is a true representation of the data. If the difference is 40-60 bits, there is a 75-90% chance that it represents a true correlation of the data. When the difference becomes less than 40 bits, the probability of the hypothesis being a true representation rapidly falls below 50% and if the total-null cost difference is less than 20 bits there is little chance of it being accurate and the training set should be reconsidered [29].

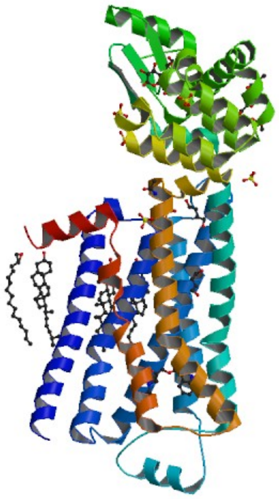


Figure 3.1: Serial X-ray crystallography structure of the Beta2-adrenergic receptor P057550.[19] Complex with ligands dodecathylene glycol, acetamide, 1,4 butanediol, (S)-carazolol, cholesterol, beta-maltose, palmitic acid, sulfate ion.

3.2 HypoGen 3D QSAR pharmacophore generation for P07550

The P07550 protein target is a human Beta-2 adrenergic receptor, that mediates the catecholamine-induced activation of adenylate cyclase through the action of G proteins. The beta-2-adrenergic receptor binds epinephrine with an approximately 30-fold greater affinity than it does norepinephrine [13]. The 3D structure of P07550 is presented in figure 3.1.

The psycholeptic and psychoanaleptic drugs interacting with P07550, as well as the plant-based food that has been found to target the same protein are listed in the following tables 3.1 and 3.2. This data has been derived from NutriChem 1.0 and Drugbank, in a similar process, as in chapter 2., where the comprehensive interaction networks were constructed. It can be observed that P07550 is frequently targeted by psycholeptic and psychoanaleptic drugs, though the pharmacological action of those interactions is yet unclear. At the same time, nutrients contained in opium poppy and licorice also interact with the receptor.

Table 3.1: Psycholeptic and psychoanaleptic drug interactions for P07550

Drug interactions for P07550			
Drugbank code	Interaction	Pharmacological action	ATC Category
DB06216	Antagonist	unknown	Psycholeptics
DB00334	Other/Unkown	unknown	Psycholeptics
DB00540	Antagonist	unknown	Psychoanaleptics
DB01151	Antagonist	unknown	Psychoanaleptics
DB00182	Agonist	unknown	Psychoanaleptics

Table 3.2: Plant-based food interactions for P07550

Food interactions for P07550		
Nutrient	Food Taxonomy ID	Common name
Glutamic Acid	TAXID:3469	Opium poppy
Liquiritin	TAXID:49827	Licorice
Wogonin	TAXID:49827	Licorice

3.2.1 Methodology & Model parameters

In order to built the pharmacophore hypothesis for P07550, initially, biological activity data were extracted from the ChEMBL database [5] (accesed: October 2015). The HypoGen algorithm cannot convert different types of activities and therefore cannot compare the activities of molecules, if they are expressed in different reference systems (eg. IC_50, K_i, $logIC_50$, etc.). For that purpose, only compounds with known K_i activities were isolated from the bioassays of ChEMBL. The K_i activity is an intrinsic, thermodynamic quantity that is independent of the substrate (ligand) but depends on the enzyme (target) and characterizes the thermodynamic equilibrium [9]. Lower values of K_i indicate higher biological activity. The ligands with know K_i activity against

P07550 were converted in 3D structures using the MolConvert command line program in Marvin 15.11.30, 2015, ChemAxon (http://www.chemaxon.com), using their canonical SMILES as input. This way, a 3D library of 484 ligands with known activities was constructed.

The training set for HypoGen should consist of at least 16 compounds, spanning a minimum of 4 orders of magnitude of activity. Redundant data should be removed as its inclusion can bias the output. From the pool of 484 ligands, a subset of 50 ligands with satisfactorily diverse properties was selected, using the Cluster Ligands protocol of Discovery Studio. The selection was based on a maximum dissimilarity method, where the distance function between the molecules was a Euclidean distance, because numeric structures were used to describe the 3D structures. When a Euclidean distance is used in the Cluster Ligands protocol of DS, each numeric property is first shifted and scaled so that each property has a mean of 0 and standard deviation of 1. The final distance is then scaled by the square root of the number of dimensions [7]. The 50 diverse molecules were then minimized, using CHARMm Forcefield. Any missing hydrogens were also added at this stage. As training set, 34 ligands from the minimized data set were selected, with activities ranging from 0.16 to 794,328. The remaining 16 minimized ligands were used as an external validation set. The data sets are presented in tables 3.3 and 3.4.

The HypoGen algorithm was adjusted to select maximum 5 of each one of the following features for 3D quantitative pharmacophore modelling; hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), hydrophobic (HY), aromatic ring (RA) and negative ionizable feature(NegIonizable). The FAST method has been applied to collect a representative set of conformers for the training set ligands. The conformers were chosen within a range of energetically reasonable conformations for each compound. In particular, all conformers within a range of 20 kcal/mol with respect to the global minimum, have been employed to build the pharmacophore hypothesis. Maximum 255 conformers were generated per structure. The uncertainty value for the activity data was set to two (maximum uncertainty). Minimum inter-feature distance of $2.97\dot{A}$ was set, as default.

Fischer validation index was set to 95% which was implemented together with the establishment of HypoGen pharmacophore model. The test set of 16 molecules was mapped to candidate hypothesis for validation by using ligand pharmacophore mapping protocol. The rigid fitting method was adopted for ligand pharmacophore mapping of the validation set, allowing for each ligand to miss all but one features of the pharmacophore (e.g. if the pharmacophore consisted of 4 features, a ligand could miss up to 3 features).

Table 3.3: The 16 ligands used as external validation set for P07550, after the minimization process.

Index	ChEMBL ID	K_i Activity	Uncertainty	Forcefield	Partial Charge Method	Intitial Potential Energy	Potential Energy
1	CHEMBL513389	0.631	2	CHARMm	MomanyRone	25.9879	6.794
2	CHEMBL3298330	5.01	2	CHARMm	MomanyRone	44.0581	-44.3315
3	CHEMBL1258599	14	2	CHARMm	MomanyRone	18.8895	-27.0613
4	CHEMBL199824	60.6	2	CHARMm	MomanyRone	33.7525	12.4483
5	CHEMBL1084773	110	2	CHARMm	MomanyRone	61.4172	28.8918
6	CHEMBL2204360	560	2	CHARMm	MomanyRone	22.9094	12.2861
7	CHEMBL1209157	1,000	2	CHARMm	MomanyRone	76.8788	50.0608
8	CHEMBL25856	1,200	2	CHARMm	MomanyRone	27.529	1.01987
9	CHEMBL458002	2,398.83	2	CHARMm	MomanyRone	81.4903	38.2892
10	CHEMBL2068577	6,606.93	2	CHARMm	MomanyRone	62.7838	-22.874
11	CHEMBL569270	10,000	2	CHARMm	MomanyRone	11.886	-15.9227
12	CHEMBL2203551	10,000	2	CHARMm	MomanyRone	27.8513	9.25367
13	CHEMBL1630578	15,578	2	CHARMm	MomanyRone	100.449	44.7566
14	CHEMBL40317	15,800	2	CHARMm	MomanyRone	18.8512	-31.4942
15	CHEMBL1622248	39,810.70	2	CHARMm	MomanyRone	49.3216	20.3858
16	CHEMBL54716	100,000	2	CHARMm	MomanyRone	73.3512	20.1177

3.2.2 Pharmacophore model for P07550

The HypoGen module generated successfully 10 pharmacophore hypothesis. The 4 hypothesis that were ranked as best are presented below in figure 3.2. It can be observed that 4 hypothesis contain one hydrophobic feature and one aromatic ring, however they vary on the number and relative position of the hydrogen bond donors.

The quality of HypoGen pharmacophore hypotheses is described by fixed cost, null cost, total cost and some other statistical parameters. In a significant pharmacophore model, cost difference between the null and total cost should be remarkable (>60). Moreover, higher correlation and lower RMSD are always good statistical indicators for a model's efficiency. [39]. These results for the 10 hypothesis can be found in table 3.5. The first hypothesis is selected as the best available model, according to the above crite-

Table 3.4: The 34 ligands used as training set for P07550, after the minimization process.

Index	ChEMBL ID	K_i Activity	Uncertainty	Forcefield	Partial Charge Method	Intitial Potential Energy	Potential Energy
1	CHEMBL723	0.166	2	CHARMm	MomanyRone	51.7482	23.5237
2	CHEMBL499	0.201	2	CHARMm	MomanyRone	42.1459	0.01061
3	CHEMBL387852	2.82	2	CHARMm	MomanyRone	11.7734	-28.6112
4	CHEMBL29141	5	2	CHARMm	MomanyRone	54.4131	7.67004
5	CHEMBL1800934	12.71	2	CHARMm	MomanyRone	46.9807	-1.23621
6	CHEMBL357995	14	2	CHARMm	MomanyRone	18.0999	-17.2776
7	CHEMBL249359	38	2	CHARMm	MomanyRone	29.2774	-64.4839
8	CHEMBL250352	49	2	CHARMm	MomanyRone	36.5459	10.8418
9	CHEMBL1683936	61	2	CHARMm	MomanyRone	37.9087	-15.2891
10	CHEMBL1221801	151	2	CHARMm	MomanyRone	25.605	-7.0682
11	CHEMBL188622	223	2	CHARMm	MomanyRone	137.603	-7.40492
12	CHEMBL462313	320	2	CHARMm	MomanyRone	34.366	9.55272
13	CHEMBL497963	501.19	2	CHARMm	MomanyRone	68.538	-0.23249
14	CHEMBL281350	1,000	2	CHARMm	MomanyRone	118.903	71.7545
15	CHEMBL1098230	1,000	2	CHARMm	MomanyRone	6.99728	-4.25591
16	CHEMBL1242950	1,000	2	CHARMm	MomanyRone	37.6626	21.8012
17	CHEMBL1242923	1,000	2	CHARMm	MomanyRone	46.6601	24.9096
18	CHEMBL707	1,889	2	CHARMm	MomanyRone	45.4829	19.3003
19	CHEMBL442	1,898	2	CHARMm	MomanyRone	75.696	8.32702
20	CHEMBL787	2,398	2	CHARMm	MomanyRone	58.8235	-5.4701
21	CHEMBL40650	3,870	2	CHARMm	MomanyRone	28.4613	0.82565
22	CHEMBL565547	7,600	2	CHARMm	MomanyRone	16.364	-0.83703
23	CHEMBL229429	10,000	2	CHARMm	MomanyRone	35.8521	3.77239
24	CHEMBL3104093	10,000	2	CHARMm	MomanyRone	52.2738	42.2754
25	CHEMBL6310	10,000	2	CHARMm	MomanyRone	23.9959	10.6719
26	CHEMBL30713	10,000	2	CHARMm	MomanyRone	12.5588	-0.19888
27	CHEMBL555146	10,000	2	CHARMm	MomanyRone	50.7249	32.049
28	CHEMBL495075	10,000	2	CHARMm	MomanyRone	49.2167	21.5371
29	CHEMBL1824265	10,000	2	CHARMm	MomanyRone	27.0155	-17.8479
30	CHEMBL72168	13,900	2	CHARMm	MomanyRone	18.2649	-25.8226
31	CHEMBL2070835	25,118.90	2	CHARMm	MomanyRone	29.6519	7.93831
32	CHEMBL57163	100,000	2	CHARMm	MomanyRone	27.3846	4.78856
33	CHEMBL226292	100,000	2	CHARMm	MomanyRone	28.8216	-47.4604
34	CHEMBL1626178	794,328	2	CHARMm	MomanyRone	40.8516	20.3985

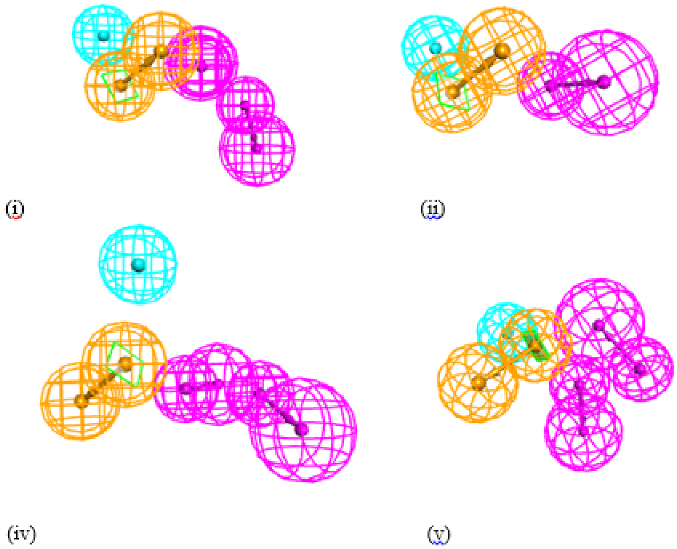


Figure 3.2: The four best pharmacophore hypothesis generated for P07550. The features depicted include the aromatic ring (orange), hydrophobic (light blue) and hydrogen bond donor (purple). It can be observed that 4 hypothesis contain one hydrophobic feature and one aromatic ring, however they vary on the number and relative position of the hydrogen bond donors.

ria. This hypothesis is depicted in figure 3.3.

The mappings of training set compounds and the correlation between their actual and predicted biological activities are elucidated in the plot of figure 3.4, where a line of best fit has been drawn for the two variables.

3.2.3 Model validation & Interpretation of results

The model of hypothesis 1 for protein P07550 was validated in three ways. First, a Fisher's test with a significance of 95% was implemented during the generation of the model. At the same time, the model was tested against the external validation test of table 3.3. Finally, an additional validation method was applied with a database of 484 compounds, with varied activities against the target, to further assess the ability of the model to differentiate between active and inactive ligands.

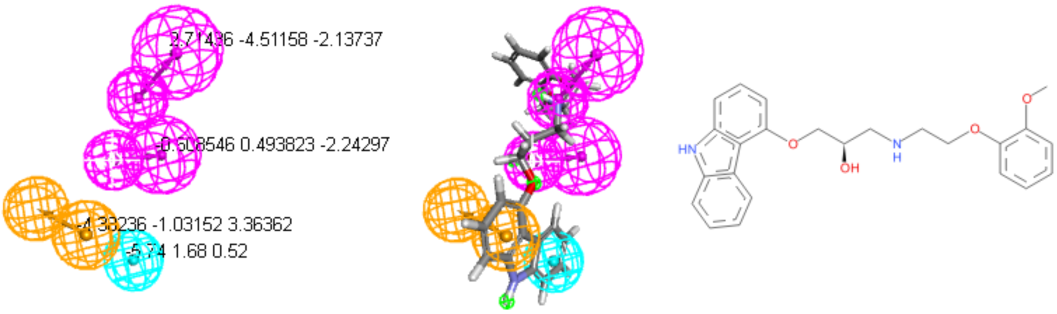


Figure 3.3: The first phamacophore hypothesis, which was selected as best available model for P07550. (a) Coordinates of the features. (b) Mapping of the most active compound CHEMBL723 with $K_i = 0.166$(c) 2D structure of CHEMBL723.

Table 3.5: Results of ten top scored pharmacophore hypotheses generated by HypoGen. (P07550)

Hypothesis	Total cost	Cost difference	RMSD	Correlation coefficient	Features
1	234.504	312.062	2.59087	0.863333	HBD HBD HY RA
2	256.817	289.749	2.65633	0.857371	HBD HY RA
3	266.236	280.33	2.91343	0.823481	HBD HBD HY RA
4	273.068	273.498	3.00371	0.810945	HBD HBD HY RA
5	275.592	270.974	2.78957	0.842347	HBD HY RA
6	282.446	264.12	3.1036	0.796494	HBD HBD HY RA
7	285.612	260.954	3.12473	0.793396	HBD HBD HY RA
8	286.518	260.048	2.95411	0.819972	HBD HY RA
9	287.087	259.479	3.0772	0.800937	HBD HY RA
10	287.367	259.199	3.14223	0.790763	HBD HBD HY RA

Note: Cost difference is the difference between null and total cost; null cost is 546.566; fixed cost is 287.087.

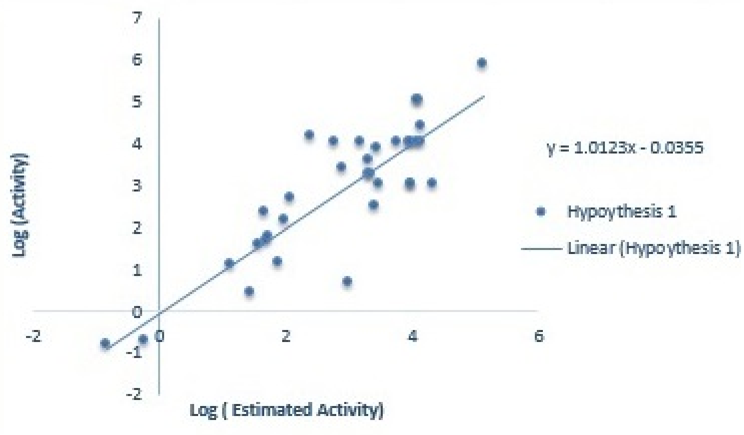


Figure 3.4: Correlation and line of best fit for estimated activity vs actual activity of training set for hypothesis 1 (P07550).

In the first validation method, a Fisher's randomization test was carried out by Discovery Studio. The null hypothesis was that the actual A_{real} and predicted $A_{predict}$ activities for the ligands against the target were unrelated variables. A randomization test was carried out, instead of Fisher's exact test, since the sample size causes the number of possible combinations to increase dramatically, to the point where a computer may have a hard time doing all the calculations in a reasonable period of time. The randomization test works by generating random combinations of numbers in the $A_{real} \times A_{predict}$ table, with the probability of generating a particular combination equal to its probability under the null hypothesis. For each combination, the Pearson's chi-square statistic is calculated. The proportion of these random combinations that have a chi-square statistic equal to or greater than the observed data is the P-value. For a significance of 95%, if the P-value is smaller or equal than 0.05, there is strong evidence against the null hypothesis. Therefore, the predicted activity is strongly correlated to the actual activity. If the P-value is bigger than 0.05, there isn't enough evidence to reject the null hypothesis [30]. In the present statistical problem, it was found that the hypothesis 1., indeed, represented a valid correlation, with a 95% significance.

Against the initial validation set of 16 compounds, the model of the first hypothesis yielded a correlation coefficient of 0.529, an RMSD of 2.463. The behavior of the model against the external validation set is anticipated to be poorer, compared to that of the training set, because the external test set possibly contains structural formations that are very different from the ones used to generate the model. However, this could also potentially indicate an over-fitting of the data.

In the next validation step, the database of 484 compounds was minimized, using CHARMm forcefield. Then, the Build Database feature of Discovery Studio was applied, in order to generate a database, automatically indexed with sub-structure, pharmacophore feature, and shape information to allow fast database searching. For each compound, maximum 255 conformations were generated. The full database is presented as supplementary material (S1). The range of activities for these compounds was between 0.1-1,160,000. The model of hypothesis 1. was selected to screen this database, using the FAST search method, to identify active candidates. By choosing a cut-off at $K_i = 1,000$ (as indicated by Discovery Studio), the ligands were characterized as active or inactive against the target. This permitted the construction of a confusion matrix, in order to evaluate the accuracy, precision and recall of the model. This matrix can be found in table 3.6.

Table 3.6: Confusion (or contingency) table for pharmacophore hypothesis 1 (P07550). Database of 484 compounds of activities ranging between 0.1-1,160,000. Cut-off method at $K_i = 1,000$.

	Predicted Active	Predicted Inactive	Total
Active	137	192	329
Inactive	46	109	155
Total	183	301	484

The confusion matrix records the number of compounds that were predicted active and were actually active (true positives, TP), the number predicted inactive but actually active (false negatives, FN), the number predicted active but actually negative (false negatives, FN) and predicted inactive and actually inactive (true negatives, TN). While evaluating the confusion matrix the following statistical indices are taken into account

[43]:

1. Accuracy, or the proportion of the total number of predictions that were correct:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{3.1}$$

2. Precision, or the fraction of the compounds returned as active which are active:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

3. Recall, or the fraction of actives which are actually identified. Coincides with TP%.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

4. f-score, the harmonic mean of precision and recall:

$$f = \frac{2TP}{2TP + FP + FN} \tag{3.4}$$

The accuracy of the hypothesis was 50.8% and its precision was found to be 74.9%. Moreover, the true positive recall was specified as 41.6% and the f-score was 53.5%. These results show that, even though the model is not as accurate in distinguishing active from inactive ligands, as desired, and a lot of the actives were missed, it has very good precision. Therefore, positive predictions are almost always actual active compounds.

3.3 HypoGen 3D QSAR pharmacophore generation for P28222

P28882 or 5-HT1B-BRIL is a G-protein coupled receptor for 5-hydroxytryptamine (serotonin). Also functions as a receptor for ergot alkaloid derivatives, various anxiolytic and antidepressant drugs and other psychoactive substances, such as lysergic acid diethylamide (LSD). Ligand binding causes a conformation change that triggers signaling via guanine nucleotide-binding proteins (G proteins) and modulates the activity of downstream effectors, such as adenylate cyclase. Signaling inhibits adenylate cyclase activity.

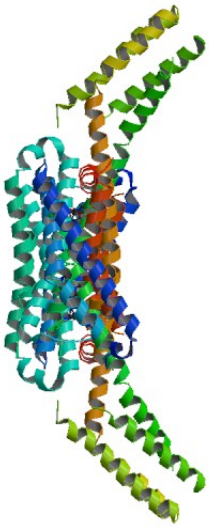


Figure 3.5: Crystal structure of the chimeric protein of 5-HT1B-BRIL in complex with dihydroergotamine (PSI Community Target).[38]

Arrestin family members inhibit signaling via G proteins and mediate activation of alternative signaling pathways. Regulates the release of 5-hydroxytryptamine, dopamine and acetylcholine in the brain, and thereby affects neural activity, nociceptive processing, pain perception, mood and behavior. Besides, plays a role in vasoconstriction of cerebral arteries [13]. The 3D structure of P28222 is presented in figure 3.5. Ligands are usually bound in a hydrophobic pocket formed by the transmembrane helices [13].

The psycholeptic and psychoanaleptic drugs interacting with P28222, as well as the plant-based food that has been found to target the same protein are listed in the following tables 3.7 and 3.8. It can be observed that several psycholeptics and psychoanaleptics bind or act as antagonists to the receptor. Furthermore, interactions with plant-based food are anticipated for ginger, potatoes, safflower and tomato, that contain serotonin.

3.3.1 Methodology & Model parameters

In order to built the pharmacophore hypothesis for P28222, biological activity data were extracted from the ChEMBL database [5] (accesed: October 2015) and processed in the same way, as for P07550. The ligands with know K_i activity against P28222

Table 3.7: Psycholeptic and psychoanaleptic drug interactions for P28222

Drug interactions for P28222			
Drugbank code	Interaction	Pharmacological action	ATC Category
DB01224	OtherUnkown	unknown	Psycholeptics
DB00408	Binder	unknown	Psycholeptics
DB00363	Antagonist	unknown	Psycholeptics
DB00246	Antagonist	unknown	Psycholeptics
DB01238	Antagonist	unknown	Psycholeptics
DB00543	Antagonist	unknown	Psychoanaleptics
DB00321	Binder	unknown	Psychoanaleptics

Table 3.8: Plant-based food interactions for P28222

Food interactions for P28222		
Nutrient	Food Taxonomy ID	Common name
Serotonin	TAXID:94328	Ginger
Serotonin	TAXID:4222	Safflower
Serotonin	TAXID:4081	Tomato

were converted in 3D structures using the MolConvert command line program in Marvin 15.11.30, 2015, ChemAxon (http://www.chemaxon.com), using their canonical SMILES as input. This way, a 3D library of 901 ligands with known activities was constructed. From this library, a subset of 49 ligands with satisfactorily diverse properties was selected, using the Cluster Ligands protocol of Discovery Studio, as before. The 49 diverse molecules were then minimized, using CHARMm Forcefield. Any missing hydrogens were also added at this stage. As training set, 34 ligands from the minimized data set were selected, with activities ranging from 0.96 to 10,000. The remaining 15 minimized ligands were used as an external validation set. The two data sets are presented in tables 3.10 and 3.9.

The parameters specified for the HypoGen algorithm were the same as for protein P07550. The algorith was adjusted to select maximum 5 of each one of the following features for 3D quantitative pharmacophore modelling; hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), hydrophobic (HY), aromatic ring (RA) and negative ionizable feature(NegIonizable). The FAST method has been applied to collect a representative set of conformers for the training set ligands. The conformers were chosen within a range of energetically reasonable conformations for each compound. In particular, all conformers within a range of 20 kcal/mol with respect to the global minimum, have been employed to build the pharmacophore hypothesis. Maximum 255 conformers were generated per structure. The uncertainty value for the activity data was set to two (maximum uncertainty). Minimum inter-feature distance of $2.97\dot{A}$ was set, as default.

Fisher validation index was, again, set to 95% which was implemented together with the establishment of HypoGen pharmacophore model. The test set of 15 molecules was mapped to candidate hypothesis for validation, and the rigid fitting method was adopted for ligand pharmacophore mapping of the validation set, allowing for each ligand to miss all but one features of the pharmacophore.

3.3.2 Pharmacophore model for P28222

The HypoGen module generated successfully 10 pharmacophore hypothesis. The 4 hypothesis that were ranked as best are presented below in figure 3.6. It can be observed that the two first hypothesis contain two hydrophobic features, one aromatic ring, and

Table 3.9: The 34 ligands used as training set for P28222, after the minimization process.

Index	ChEMBL ID	Ki Activity	Uncertainty	Forcefield	Partial Charge Method	Intitial Potential Energy	Potential Energy
1	CHEMBL601013	0.96	2	CHARMm	MomanyRone	56.6094	24.4612
2	CHEMBL442	2.1	2	CHARMm	MomanyRone	75.696	8.32702
3	CHEMBL198488	3.98	2	CHARMm	MomanyRone	118.423	70.0094
4	CHEMBL85	10	2	CHARMm	MomanyRone	59.507	27.2564
5	CHEMBL355517	32	2	CHARMm	MomanyRone	26.105	5.94079
6	CHEMBL319352	48	2	CHARMm	MomanyRone	52.83	14.0227
7	CHEMBL104374	110	2	CHARMm	MomanyRone	25.8071	7.89435
8	CHEMBL107772	150	2	CHARMm	MomanyRone	9.46981	-60.0958
9	CHEMBL425190	199.53	2	CHARMm	MomanyRone	61.1205	49.039
10	CHEMBL393169	300	2	CHARMm	MomanyRone	52.0014	25.9022
11	CHEMBL2413153	411	2	CHARMm	MomanyRone	37.6139	8.06594
12	CHEMBL244946	501.19	2	CHARMm	MomanyRone	73.3816	38.928
13	CHEMBL2424668	700	2	CHARMm	MomanyRone	46.484	17.5512
14	CHEMBL203013	720	2	CHARMm	MomanyRone	50.2642	30.9616
15	CHEMBL458002	776.25	2	CHARMm	MomanyRone	81.4903	38.2892
16	CHEMBL38465	1,000	2	CHARMm	MomanyRone	16.1349	0.16076
17	CHEMBL715	1,000	2	CHARMm	MomanyRone	63.0408	38.3711
18	CHEMBL1242950	1,000	2	CHARMm	MomanyRone	37.6626	21.8012
19	CHEMBL339980	1,060	2	CHARMm	MomanyRone	30.3097	4.65631
20	CHEMBL139230	1,120	2	CHARMm	MomanyRone	23.2803	10.1596
21	CHEMBL360803	1,980	2	CHARMm	MomanyRone	23.3446	7.12036
22	CHEMBL71676	2,024	2	CHARMm	MomanyRone	16.2114	3.57228
23	CHEMBL229429	10,000	2	CHARMm	MomanyRone	35.8521	3.77239
24	CHEMBL1794855	10,000	2	CHARMm	MomanyRone	-2.01882	-101.981
25	CHEMBL1258843	10,000	2	CHARMm	MomanyRone	-8.30133	-35.1111
26	CHEMBL70988	10,000	2	CHARMm	MomanyRone	24.3354	12.8505
27	CHEMBL338115	10,000	2	CHARMm	MomanyRone	16.9741	-2.65065
28	CHEMBL71751	10,000	2	CHARMm	MomanyRone	50.2933	20.719
29	CHEMBL139976	10,000	2	CHARMm	MomanyRone	14.4727	7.74292
30	CHEMBL1189234	10,000	2	CHARMm	MomanyRone	78.1977	66.6188
31	CHEMBL371300	10,000	2	CHARMm	MomanyRone	22.6007	-0.26914
32	CHEMBL2058429	10,000	2	CHARMm	MomanyRone	69.2834	36.5829
33	CHEMBL2181187	10,000	2	CHARMm	MomanyRone	42.1677	26.7867
34	CHEMBL2031737	10,000	2	CHARMm	MomanyRone	83.4966	48.9074

Table 3.10: The 15 ligands used as external validation set for P28222, after the minimization process.

Index	ChEMBL ID	Ki Activity	Uncertainty	Forcefield	Partial Charge Method	Intitial Potential Energy	Potential Energy
1	CHEMBL370110	2.5	2	CHARMm	MomanyRone	91.8835	68.501
2	CHEMBL15933	31.62	2	CHARMm	MomanyRone	73.9576	38.6059
3	CHEMBL2204360	33	2	CHARMm	MomanyRone	22.9094	12.2861
4	CHEMBL467094	100	2	CHARMm	MomanyRone	78.7629	52.0283
5	CHEMBL126340	190	2	CHARMm	MomanyRone	94.6274	19.3007
6	CHEMBL3217984	316.23	2	CHARMm	MomanyRone	65.0853	51.6073
7	CHEMBL73281	600	2	CHARMm	MomanyRone	23.1011	-4.66025
8	CHEMBL2260930	751.8	2	CHARMm	MomanyRone	55.1366	-2.35408
9	CHEMBL311694	1000	2	CHARMm	MomanyRone	65.3363	24.5109
10	CHEMBL233212	1000	2	CHARMm	MomanyRone	86.9175	48.1739
11	CHEMBL376344	1000	2	CHARMm	MomanyRone	30.4082	-8.39994
12	CHEMBL3104091	1348.96	2	CHARMm	MomanyRone	42.5095	31.5714
13	CHEMBL1642866	5000	2	CHARMm	MomanyRone	139.27	63.6545
14	CHEMBL269538	10000	2	CHARMm	MomanyRone	22.3699	6.8177
15	CHEMBL569270	10000	2	CHARMm	MomanyRone	11.886	-15.9227

one hydrogen bond donor. However, the positions and orientations of the features vary. The third and fourth pharmacophores are slightly differentiated, as they feature three hydrophobic groups, instead of two. The statistical results (null cost, total cost, RMSD and correlation coefficient) for the 10 hypothesis can be found in table 3.11. The first hypothesis is selected as the best available model, according to the above criteria. This hypothesis is depicted in figure 3.7.

The mappings of training set compounds and the correlation between their actual and predicted biological activities are depicted in figure 3.8.

3.3.3 Model validation & Interpretation of results

The model of hypothesis 1 for protein P28222 was validated in three ways, similarly to P07550. A Fisher's test with a significance of 95% was implemented during the generation of the model, and an external data test, presented in table 3.10, was used as external validation. Finally, the model was used to screen a database of 200 compounds, with varied activities against the target, to further assess the ability of the model to differentiate between active and inactive ligands.

During the validation with Fisher's randomization test, the null hypothesis stated that

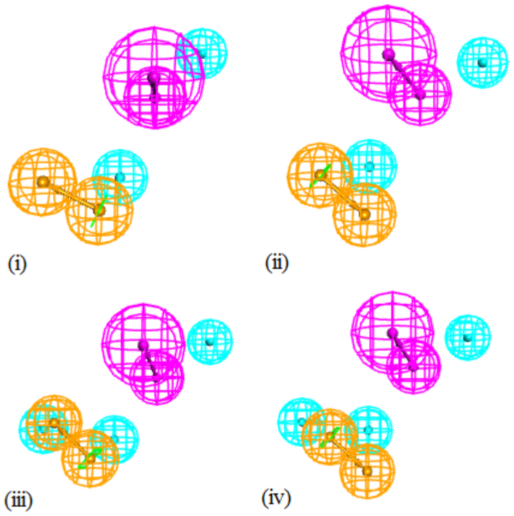


Figure 3.6: The four best pharmacophore hypothesis generated for P28222. The features depicted include the aromatic ring (orange), hydrophobic (light blue) and hydrogen bond donor (purple).

Table 3.11: Results of ten top scored pharmacophore hypotheses generated by HypoGen.(P28222)

Hypothesis	Total cost	Cost difference	RMSD	Correlation coefficient	Features
1	198.883	144.26	2.1958	0.815277	HBD HY HY RA
2	210.691	132.45	2.34836	0.785151	HBD HY HY RA
3	223.601	119.54	2.49353	0.75344	HBD HY HY HY RA
4	225.307	117.83	2.51288	0.748959	HBD HY HY HY RA
5	233.839	109.3	2.61343	0.724623	HBA HBD HY RA
6	235.464	107.68	2.64042	0.717728	HBA HBD HY RA
7	237.056	106.09	2.65639	0.713633	HBA HY RA
8	238.146	105	2.65935	0.712923	HBA HBD HY RA
9	241.958	101.18	2.70405	0.701117	HBA HBD HY RA
10	243.51	99.631	2.7158	0.697981	HBD HY HY HY RA

Note: Cost difference is the difference between null and total cost; null cost is 343.141; fixed cost is 116.911.

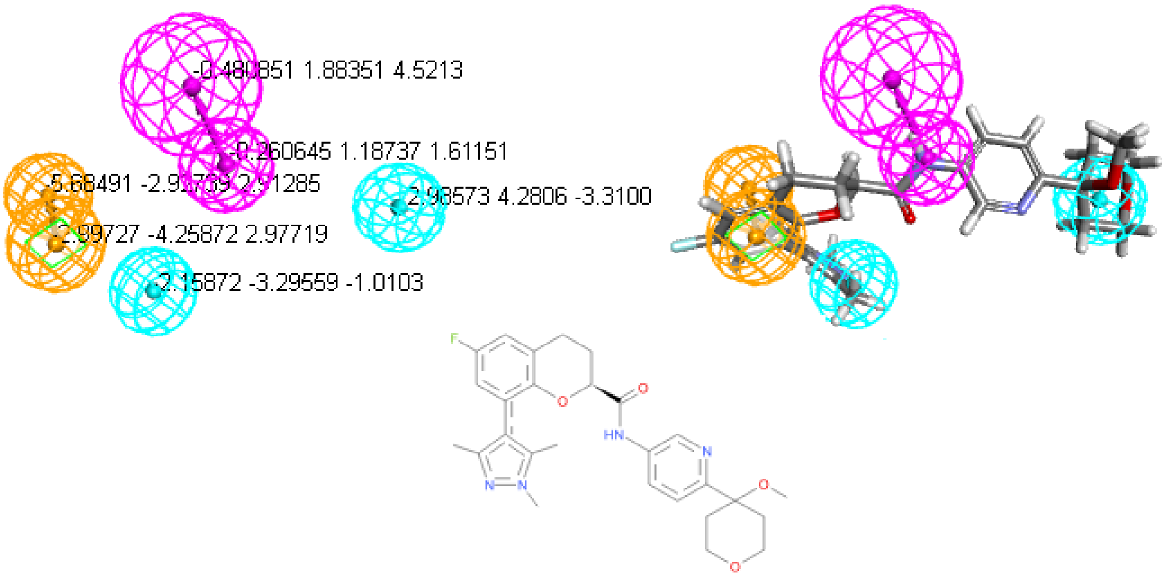


Figure 3.7: The first phamacophore hypothesis, which was selected as best available model for P28222. (a) Coordinates of the features. (b) Mapping of the most active compound CHEMBL601013 with $K_i = 0.96$. (c) 2D structure of CHEMBL601013.

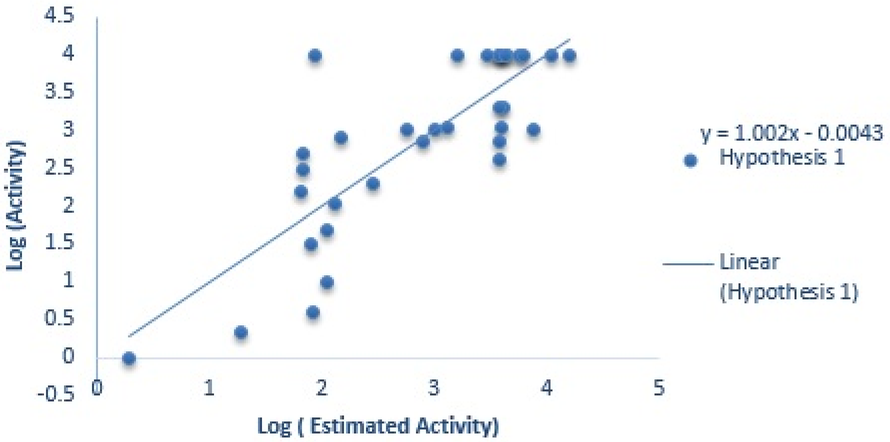


Figure 3.8: Correlation and line of best fit for estimated activity vs actual activity of training set for hypothesis 1 (P28222).

no correlation exists between the actual and predicted activities of the ligands against the P28222. In the present statistical problem, it was found that the P-value for hypothesis 1. was smaller or equal than 0.05, and therefore there was strong evidence against the null hypothesis. A valid correlation between actual and predicted activities was established with 95% significance.

Against the initial validation set of 15 compounds, the model of the first hypothesis yielded a correlation coefficient of 0.478, an RMSD of 1.852. The behavior of the model against the external validation set is very poor, because the external test set possibly contains structural formations that are very different from the ones used to generate the model. However, this could also potentially indicate an over-fitting of the data.

In the next validation step, the Build Database feature of Discovery Studio was applied to the data set of 200 compounds, in order to generate a database, automatically indexed with sub-structure, pharmacophore feature, and shape information to allow fast database searching. For each compound, maximum 255 conformations were generated. The full database is presented as supplementary material (S2). The range of activities for these compounds was between 0.14-30,000. The model of hypothesis 1. was selected to screen this database, using the FAST search method, to identify active candidates. By choosing a cut-off at $K_i = 6,000$ (as indicated by Discovery Studio), the ligands were characterized as active or inactive against the target. This permitted the construction of a confusion matrix, in order to evaluate the accuracy, precision and recall of the model. This matrix can be found in table 3.12.

Table 3.12: Confusion (or contingency) table for pharmacophore hypothesis 1 (P28222). Database of 200 compounds of activities ranging between 0.14-30,000. Cut-off method at $K_i = 6000$.

	Predicted Active	Predicted Inactive	Total
Active	18	147	165
Inactive	1	34	35
Total	19	181	200

From the confusion table, it becomes apparent that the accuracy of the prediction is

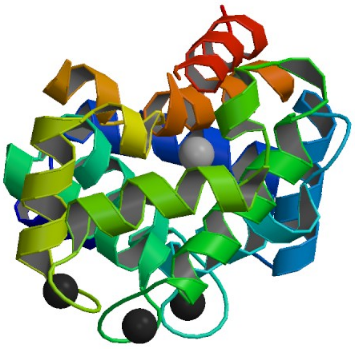


Figure 3.9: Neuronal calcium sensor-1 (NCS-1)from Rattus norvegicus complex with D2 dopamine receptor peptide from Homo sapiens [35]. Calcium and potassium ions are bound as co-factors in the complex.

26.0%, the precision 94.7%, the recall only 10.9% and the f-score 19.6%. This hypothesis is very precise in identifying true actives, however it appears to have a very high false negative rate (low recall). Therefore, a lot of actives in the data set were missed, because the restrictions imposed by the model were too specific. This could indicate an over-fitting of the model on the training set, or could be attributed to different binding modes of ligands present on the data set, that were not incorporated in the initial hypothesis.

3.4 HypoGen 3D QSAR pharmacophore generation for P14416

The D(2) dopamine receptor or P14416 is a dopamine receptor whose activity is mediated by G proteins which inhibit adenylyl cyclase [13]. The 3D structure for P14416 is presented in figure 3.9.

The psycholeptic and psychoanaleptic drugs interacting with P14416, as well as the plant-based food that has been found to target the same protein are listed in the following tables 3.13 and 3.14. This data has been derived from NutriChem 1.0 and Drugbank, in a similar process, as in chapter 2., where the comprehensive interaction networks were constructed. It is noted that P14416 is registered as a very common target for both psycholeptics and psychoanaleptics, whose pharmacological action as antagonists of the dopamine receptor is well established. However, several nutrients seem to interact with the same target; banana, barley, garlic, garden pea, cacao and opium poppy are all cited as plant-based food that contains phytochemicals, active against P14416.

3.4.1 Methodology & Model parameters

In order to built the pharmacophore hypothesis for P14416, biological activity data were extracted from the ChEMBL database [5] and processed in the same way, as for the two previous targest. The ligands with know K_i activity against P14416 were converted in 3D structures, using their canonical SMILES as input. This way, a 3D library of 5650 ligands with known activities was constructed. From this library, a subset of 32 ligands with satisfactorily diverse properties was selected, using the Cluster Ligands protocol of Discovery Studio, as before. The 32 diverse molecules were then minimized, using CHARMm Forcefield. Any missing hydrogens were also added at this stage. As training set, 20 ligands from the minimized data set were selected, with activities ranging from 0.05 to 14,000. The remaining 12 minimized ligands were used as an external validation set. The two data sets are presented in tables 3.16 and 3.15.

The parameters specified for the HypoGen algorithm were the same as for protein P07550 and P28222. The algorith was adjusted to select maximum 5 of each one of the following features for 3D quantitative pharmacophore modelling; hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), hydrophobic (HY), aromatic ring (RA) and negative ionizable feature(NegIonizable). The FAST method has been applied to collect a representative set of conformers for the training set ligands. The conformers were chosen within a range of energetically reasonable conformations for each compound. In particular, all conformers within a range of 20 kcal/mol with respect to the global minimum, have been employed to build the pharmacophore hypothesis. Maximum 255 conformers were generated per structure. The uncertainty value for the activity data was set to two (maximum uncertainty). Minimum inter-feature distance of $2.97\dot{A}$ was set, as default.

Once more, Fisher validation index was set to 95% which was implemented together

Table 3.13: Psycholeptic and psychoanaleptic drug interactions for P14416

Drug interactions for P14416			
Drugbank code	Interaction	Pharmacological action	ATC Category
DB00391	Antagonist	yes	Psycholeptics
DB00409	Antagonist	yes	Psycholeptics
DB06288	Antagonist	yes	Psycholeptics
DB00450	Antagonist	yes	Psycholeptics
DB00502	Antagonist	yes	Psycholeptics
DB01224	Antagonist	yes	Psycholeptics
DB00408	Antagonist	yes	Psycholeptics
DB06216	Antagonist	yes	Psycholeptics
DB00363	Antagonist	yes	Psycholeptics
DB04842	Antagonist	yes	Psycholeptics
DB01618	Antagonist	yes	Psycholeptics
DB08815	Antagonist	yes	Psycholeptics
DB06144	Antagonist	yes	Psycholeptics
DB00246	Antagonist	yes	Psycholeptics
DB01238	Antagonist, Partial agonist	yes	Psycholeptics
DB01267	Antagonist	yes	Psycholeptics
DB00734	Antagonist	yes	Psycholeptics
DB01614	Antagonist	yes	Psycholeptics
DB00420	Antagonist	yes	Psycholeptics
DB01403	Antagonist	yes	Psycholeptics
DB00508	Antagonist	yes	Psycholeptics
DB01622	Antagonist	yes	Psycholeptics
DB00850	Antagonist	yes	Psycholeptics
DB01063	Antagonist	yes	Psycholeptics
DB00623	Antagonist	yes	Psycholeptics
DB00831	Antagonist	yes	Psycholeptics
DB01621	Antagonist	yes	Psycholeptics
DB01624	Antagonist	yes	Psycholeptics
DB01239	Antagonist	yes	Psycholeptics
DB01043	Agonist	unknown	Psychoanaleptics
DB00543	Antagonist	unknown	Psychoanaleptics
DB00726	OtherUnknown	unknown	Psychoanaleptics
DB00458	Binder	unknown	Psychoanaleptics
DB01142	Antagonist	unknown	Psychoanaleptics
DB01151	Binder	unknown	Psychoanaleptics
DB00805	Agonist	unknown	Psychoanaleptics
DB00370	Binder	unknown	Psychoanaleptics
DB00182	Binder	unknown	Psychoanaleptics

Table 3.14: Plant-based food interactions for P14416

Food interactions for P14416		
Nutrient	Food Taxonomy ID	Common name
Salsolinol	TAXID:89151	Banana
Tyramine	TAXID:4513	Barley
Tyramine	TAXID:3888	Garden pea
Trifluoperazine	TAXID:4682	Garlic
Higenamine	TAXID:3469	Opium poppy
Aporphine	TAXID:3469	Opium poppy
Salsolinol	TAXID:3641	Theobroma cacao

with the establishment of HypoGen pharmacophore model. The test set of 12 molecules was mapped to candidate hypothesis for validation, and the rigid fitting method was adopted for ligand pharmacophore mapping of the validation set, allowing for each ligand to miss all but one features of the pharmacophore.

The mappings of training set compounds and the correlation between their actual and predicted biological activities are depicted in figure 3.10.

3.4.2 Pharmacophore model for P14416

The HypoGen module generated successfully 10 pharmacophore hypothesis. The 4 hypothesis that were ranked as best are presented below in figure 3.11. Hypothesis one, two and three are comprised of hydrogen bond acceptors and donors, all with different spatial arrangements. However, hypothesis three, includes one hydrophobic feature, three hydrogen bond acceptors and does not support the existence of hydrogen bond donnors.

The statistical results (null cost, total cost, RMSD and correlation coefficient) for the 10 hypothesis can be found in table 3.17. The first hypothesis is selected as the best available model, according to the above criteria. This hypothesis is presented in figure 3.12.

Table 3.15: The 20 ligands used as training set for P14416, after the minimization process.

Index	ChEMBL ID	Ki Activity	Uncertainty	Forcefield	Partial Charge Method	Intitial Potential Energy	Potential Energy
17	CHEMBL156651	0.05	2	CHARMm	MomanyRone	34.186	9.04991
20	CHEMBL1259129	1.9	2	CHARMm	MomanyRone	78.1434	54.8565
19	CHEMBL147159	2.2	2	CHARMm	MomanyRone	60.4403	27.9359
18	CHEMBL3265321	5.98	2	CHARMm	MomanyRone	55.5861	-252.07
16	CHEMBL1087817	6.3	2	CHARMm	MomanyRone	72.1295	42.4538
15	CHEMBL301265	21	2	CHARMm	MomanyRone	18.7794	-4.28568
14	CHEMBL3220214	32.36	2	CHARMm	MomanyRone	31.662	15.6717
13	CHEMBL3265323	64.16	2	CHARMm	MomanyRone	-125.768	-992.896
12	CHEMBL440512	119	2	CHARMm	MomanyRone	69.9521	39.778
11	CHEMBL1223598	316.23	2	CHARMm	MomanyRone	96.3959	33.1814
10	CHEMBL158973	330	2	CHARMm	MomanyRone	11.6849	6.99085
9	CHEMBL99208	402	2	CHARMm	MomanyRone	35.9677	13.0951
8	CHEMBL1258035	501.19	2	CHARMm	MomanyRone	78.1303	45.8024
7	CHEMBL2325934	830	2	CHARMm	MomanyRone	47.713	17.1369
6	CHEMBL269396	1,400	2	CHARMm	MomanyRone	21.115	6.66341
5	CHEMBL40284	1,600	2	CHARMm	MomanyRone	41.4219	24.7036
4	CHEMBL2205831	2,830	2	CHARMm	MomanyRone	85.0433	73.5694
3	CHEMBL484203	4,100	2	CHARMm	MomanyRone	65.7878	26.3962
2	CHEMBL19331	9,400	2	CHARMm	MomanyRone	19.365	8.20034
1	CHEMBL241101	14,000	2	CHARMm	MomanyRone	47.7179	11.492

Table 3.16: The 12 ligands used as external validation set for P14416, after the minimization process.

Index	ChEMBL ID	Ki Activity	Uncertainty	Forcefield	Partial Charge Method	Intitial Potential Energy	Potential Energy
1	CHEMBL2062848	12,000	2	CHARMm	MomanyRone	39.8627	22.8099
2	CHEMBL141343	5,587	2	CHARMm	MomanyRone	10.8383	-30.3153
3	CHEMBL165381	2,400	2	CHARMm	MomanyRone	-23.7392	-40.7116
4	CHEMBL211026	1,500	2	CHARMm	MomanyRone	58.8052	25.8356
5	CHEMBL92939	984	2	CHARMm	MomanyRone	51.4646	33.1222
6	CHEMBL21731	665	2	CHARMm	MomanyRone	34.0352	25.6469
7	CHEMBL348285	355	2	CHARMm	MomanyRone	24.6282	15.361
8	CHEMBL1909065	234	2	CHARMm	MomanyRone	172.403	59.1701
9	CHEMBL3216146	79	2	CHARMm	MomanyRone	65.8214	19.9372
10	CHEMBL3084517	40.1	2	CHARMm	MomanyRone	173.425	65.5998
11	CHEMBL3234537	23	2	CHARMm	MomanyRone	82.7016	20.506
12	CHEMBL1087744	2.5	2	CHARMm	MomanyRone	53.5237	23.202

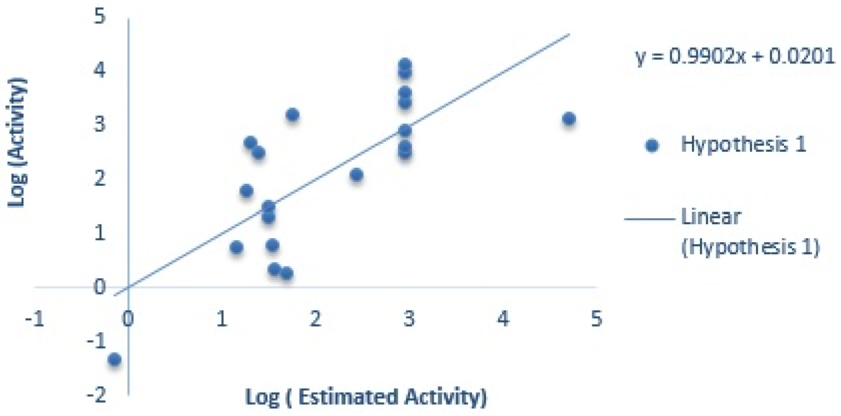


Figure 3.10: Correlation and line of best fit for estimated activity vs actual activity of training set for hypothesis 1 (P14416).

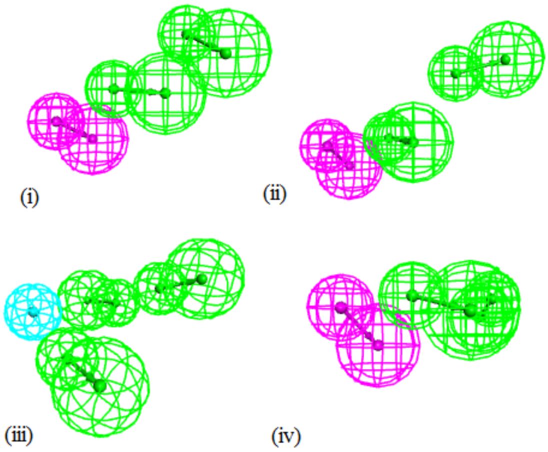


Figure 3.11: The four best pharmacophore hypothesis generated for P14416. The features depicted include the hydrogen bond acceptor (green), the hydrophobic (light blue) and hydrogen bond donor (purple).

Table 3.17: Results of ten top scored pharmacophore hypotheses generated by HypoGen. (P14416)

Hypothesis	Total cost	Cost difference	RMSD	Correlation coefficient	Features
1	165.107	103.68	3.07381	0.742764	HBA HBA HBD
2	165.668	103.12	3.08537	0.740472	HBA HBA HBD
3	167.308	101.48	3.10949	0.735683	HBA HBA HBA HY
4	168.12	100.67	3.11668	0.734272	HBA HBA HBD
5	169.647	99.145	3.14744	0.727983	HBA HBA HBA HY
6	172.734	96.058	3.18791	0.719626	HBA HBA HBD
7	175.173	93.619	3.2266	0.711394	HBA HBA HBD
8	175.292	93.5	3.23369	0.709834	HBA HBA HBA HY
9	176.173	92.619	3.23918	0.708692	HBA HBA HBD
10	177.909	90.883	3.27765	0.700175	HBA HBA HBA HY

Note: Cost difference is the difference between null and total cost; null cost is 268.792; fixed cost is 70.2819.

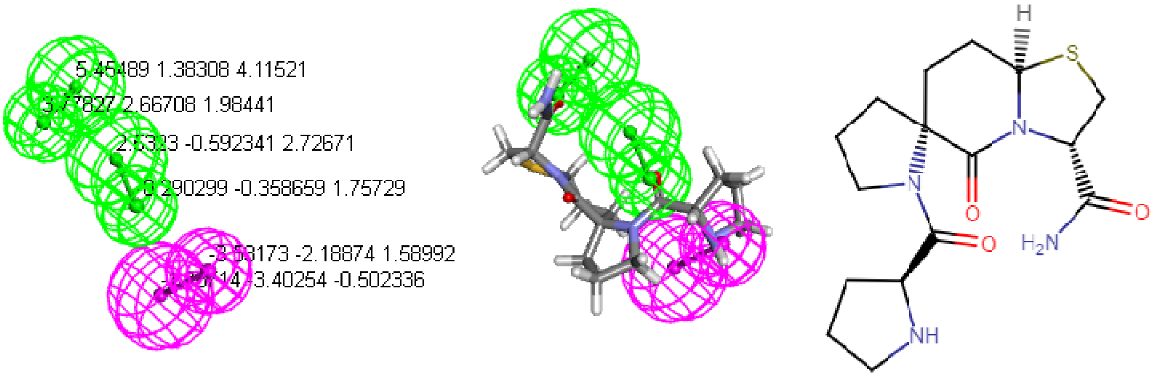


Figure 3.12: The first phamacophore hypothesis, which was selected as best available model for P14416. (a) Coordinates of the features. (b) Mapping of the most active compound CHEMBL156651 with $K_i = 0.05$. (c) 2D structure of CHEMBL156651.

3.4.3 Model validation & Interpretation of results

During the validation with Fisher's randomization test, the null hypothesis stated that no correlation exists between the actual and predicted activities of the ligands against the P14416. It was found that the P-value for hypothesis 1. was smaller or equal than 0.35, and therefore the null hypothesis could be rejected with a significance of 65%. Therefore, we can assume that there is a weak correlation between actual and predicted activities.

Against the initial validation set of 12 compounds, the model of the first hypothesis yielded a correlation coefficient of 0.278, an RMSD of 2.201. The behavior of the model against the external validation set is very poor, because the external test set possibly contains structural formations that are very different from the ones used to generate the model. However, this could also potentially indicate an over-fitting of the data.

In the next validation step, the Build Database feature of Discovery Studio was applied to the data set of 199 compounds, in order to generate a database, automatically indexed with sub-structure, pharmacophore feature, and shape information to allow fast database searching. For each compound, maximum 255 conformations were generated. The full database is presented as supplementary material (S3). The range of activities for these compounds was between 0.031-88,000. The model of hypothesis 1. was selected to screen this database, using the FAST search method, to identify active candidates. By choosing a cut-off at $K_i = 300$ (as indicated by Discovery Studio), the ligands were characterized as active or inactive against the target. This permitted the construction of a confusion matrix, in order to evaluate the accuracy, precision and recall of the model. This matrix can be found in table 3.18.

Table 3.18: Confusion (or contingency) table for pharmacophore hypothesis 1 (P14416). Database of 199 compounds of activities ranging between 0.031-88,000. Cut-off method at $K_i = 300$.

	Predicted Active	Predicted Inactive	Total
Active	15	74	89
Inactive	14	96	110
Total	29	170	199

The accuracy, as calculated from the contingency table, is 55.8%, the recall is 16.9%, while the precision and f-score are, respectively, 51.7% and 25.4%. The performance of the model in distinguishing active from inactive compounds is overall mediocre. The low recall might also be an indicator of a bias towards predicting inactive compounds, which is not always a desirable characteristic, especially during the early stages of lead discovery for new active compounds.

3.5 Screening NutriChem 1.0 Database to identify potentially active phytochemicals

In the next part of this study, the pharmacophore models constructed for the three proteins, were used to screen the database of phytochemicals that were stored in NutriChem 1.0. Through this process, additional nutrients, that are constituents of plant-based food, could be identified as potentially active for the three targets of psychiatric drugs.

3.5.1 Building a 3D conformation database for NutriChem 1.0

The database of phyrotchemicals in NutriChem was available in the form of canonical SMILES, that had to be converted into 3D structures. This process was carried out using the MolConvert command line program in Marvin 15.11.30, 2015, ChemAxon (http://www.chemaxon.com). The molecules were then minimized, using CHARMm Forcefield, and the minimization feature of Discovery Studio. Any missing hydrogens were also added at this stage. That way, a database of 4,076 ligands was built. Following that, the Build Database feature of Discovery Studio was also applied, in order to generate a library, automatically indexed with sub-structure, pharmacophore feature, and shape information to allow fast database searching. For each of the 4,076 compounds, a maximum 255 conformations were generated. This was the final library, utilized for the pharmacophore screening.

3.5.2 Active phytochemicals against P07550

Screening with hypothesis 1 for protein P07550 returned 550 hits, of which 218 had an estimated activity of less than $K_i = 1,000$. The list of compounds can be found in supplementary materials S4. The first 20 phytochemicals that were estimated to have the highest biological activity are cited in table 3.19.

Table 3.19: The 20 most active phytochemicals against P07550, as predicted by the respective pharmacophore model. The chemical IDs of the compounds (CDBNO, CHEMLIST, CID etc.) refer to their IDs in different online databases. The common names of the nutrients can be found, by using these IDs as queries in the online version of Nutrichem 1.0. The fit value is a measure of how well a compound maps to the pharmacophore features. A fit value of 10 represents an ideal mapping.

Index	Phytochemical ID	Estimated Activity	Fit value
1	CDBNO:49744	0.518744	9.49205
2	CDBNO:15590;CHEMLIST:4081603;CID:486612	0.979762	9.21588
3	CDBNO:49991	1.13665	9.15137
4	CDBNO:5346	1.15189	9.14559
5	CDBNO:35992;CHEMLIST:4076356;CID:100528	1.23077	9.11682
6	CHEM003089;CID:503732	1.28145	9.0993
7	CDBNO:54376	1.32841	9.08367
8	CID:636544	2.10969	8.88278
9	CID:16109834	2.16519	8.8715
10	CDBNO:18098;CHEMLIST:4195944;CHEMLIST:4217368;CID:145948;CID:163910	2.21789	8.86106
11	CID:637309	2.45399	8.81713
12	CDBNO:3581;CHEMLIST:4012292;CID:101712	2.5283	8.80417
13	CID:9917512	2.53903	8.80233
14	CDBNO:49745	2.83817	8.75396
15	CDBNO:42991;CHEMLIST:4253861;CID:193876	2.84968	8.7522
16	CDBNO:23164	3.25694	8.69419
17	CID:160355	3.31273	8.68681
18	CDBNO:41634	3.80728	8.62638
19	CDBNO:1223;CHEM000762;CHEMLIST:4013160;CID:164648	3.85899	8.62053
20	CDBNO:22388;CDBNO:22389;CDBNO:22390;CHEBI:53663;CHEMLIST:4021760;CID:14579	3.98236	8.60686

3.5.3 Active phytochemicals against P28222

Screening with hypothesis 1 for protein P2822 returned 322 hits, of which 135 had an estimated activity of less than $K_i = 6,000$. The list of compounds can be found in supplementary materials S5. The first 20 phytochemicals that were estimated to have the highest biological activity are cited in table 3.20.

Table 3.20: The 20 most active phytochemicals against P28222, as predicted by the respective pharmacophore model. The chemical IDs of the compounds (CDBNO, CHEMLIST, CID etc.) refer to their IDs in different online databases. The common names of the nutrients can be found, by using these IDs as queries in the online version of Nutrichem 1.0. The fit value is a measure of how well a compound maps to the pharmacophore features. A fit value of 10 represents an ideal mapping.

Index	Phytochemical ID	Estimated Activity	Fit value
1	CDBNO:49744	3.08888	7.1399
2	CDBNO:14395	6.40887	6.82292
3	CDBNO:39663	8.56582	6.69693
4	CDBNO:49745	9.47557	6.65309
5	CDBNO:48871	9.62313	6.64638
6	CHEMLIST:4223196;CID:4698	11.5751	6.56618
7	CHEMLIST:4231504;CID:6451137	11.5791	6.56602
8	CHEMLIST:4249758;CID:21701	14.2701	6.47527
9	CDBNO:49747	15.5695	6.43743
10	CHEMLIST:4259864;CID:52999	16.324	6.41687
11	CDBNO:20788;CID:46173976	16.3249	6.41685
12	CDBNO:43854;CDBNO:43856;CHEMLIST:4254560;CID:321311;CID:394846	17.4212	6.38862
13	CDBNO:48870	21.0255	6.30695
14	CDBNO:432	26.1479	6.21226
15	CDBNO:22584	29.8699	6.15447
16	CDBNO:13832	34.6159	6.09042
17	CDBNO:14745	34.9813	6.08586
18	CDBNO:20357	41.5108	6.01154
19	CID:11980866	42.0726	6.0057
20	CDBNO:20789	51.9176	5.91439

3.5.4 Active phytochemicals against P14416

Screening with hypothesis 1 for protein P14416 returned 2061 hits, of which 1722 had an estimated activity of less than $K_i = 300$. The list of compounds can be found in supplementary materials S6. The first 20 phytochemicals that were estimated to have the highest biological activity are cited in table 3.21.

Table 3.21: The 20 most active phytochemicals against P14416, as predicted by the respective pharmacophore model. The chemical IDs of the compounds (CDBNO, CHEMLIST, CID etc.) refer to their IDs in different online databases. The common names of the nutrients can be found, by using these IDs as queries in the online version of Nutrichem 1.0. The fit value is a measure of how well a compound maps to the pharmacophore features. A fit value of 10 represents an ideal mapping.

Index	Phytochemical ID	Estimated Activity	Fit value
1	CID:401298	0.56832	4.95381
2	CHEMLIST:4054705;CID:439221	0.573688	4.94972
3	CDBNO:34944	0.583902	4.94206
4	CDBNO:30695;CHEBI;31856;CHEBI;52513;CHEMLIST:4276090;CID:163659	0.590597	4.93711
5	CDBNO:24573;CDBNO:47429	0.600097	4.93018
6	CDBNO:11825;CID:5315931	0.635241	4.90546
7	CID:10770608	0.647479	4.89717
8	CDBNO:6005	0.655568	4.89178
9	CDBNO:19098;CHEM000755	0.656343	4.89127
10	CDBNO:50869	0.668715	4.88316
11	CDBNO:39888;CHEMLIST:4250130;CID:5380394	0.685889	4.87215
12	CDBNO:40864;CHEMLIST:4075312;CID:108011	0.717281	4.85271
13	CDBNO:45129	0.723306	4.84908
14	CDBNO:17844	0.733558	4.84297
15	CDBNO:32296;CHEMLIST:4220693;CID:471121	0.7351	4.84205
16	CDBNO:44710;CDBNO:45095;CDBNO:45097	0.737451	4.84067
17	CDBNO:56252	0.765303	4.82457
18	CDBNO:3714	0.7861	4.81292
19	CDBNO:34946;CHEMLIST:4093417;CID:442544	0.811273	4.79923
20	CID:25203645	0.812658	4.79849

Chapter 4

Similarity-based modeling in large scale prediction of drug-nutrient interactions

4.1 Introduction

4.1.1 Background: In silico prediction of drug-drug interactions

Drug-drug interactions (DDIs) may occur whenever a concurrent consumption of different drugs is taking place, and their effects may be either synergistic, antagonistic or coalistic (when a new effect appears, that wouldn't have been caused by the consumption of either drug on its own). Drug-drug interactions, much like herbal-drug interactions, can be caused by either pharmacokinetic, or pharmacodynamic associations [37]. The significance of DDIs becomes apparent, when we consider that 30% of all Adverse Effects (ADEs) are caused by DDIs (ADEs are one of the primary reason that drugs can fail clinical trials). Furthermore, DDIs are regarded responsible for 0.57-4.87% of all hospital admissions [37]. This is why, DDI evaluation is considered rather an important aspect of Pharmacovigilance (PV), the science and activities related to detection, assessment, understanding and prevention of ADEs [4]. DDIs can be evaluated in different stages of drug development, in silico, in vitro and in vivo.

In vitro studies usually focus on modeling the interactions of drugs with cytochrome enzymes P450. A majority of drugs is cleared via P450 mediated metabolism, thus the inhibition of P450 enzymes can be responsible for serious drug interactions. When concomitant drugs are metabolized by the same P450 enzymes, their metabolism could be affected. However, recently, the importance of other mechanisms, such as transporters, has also been recognized and addressed by in vitro assessments [8]. Experimental detection of ADEs using extensive in vitro safety pharmacology profiling remains challenging in terms of cost and efficiency [25].

In vivo studies include clinical trials and further studies, after the drugs enter the marketplace. Unfortunately, DDIs may go undetected in clinical trials, because of the limited number of participants, in comparison with the high number of drugs and possible combinations. Moreover, DDIs may still go undetected in post-market assessments, due to the many different comorbidities, drug combinations and confounding factors that exist [37].

The problem of in silico, preclinical prediction of DDIs remains an open research challenge. Informatics has an important role in the discovery of new DDIs. Cheminformatic methods, such as 2D/3D quantitative structure-activity relationships (QSAR) and molecular docking, can be useful to predict DDIs [27]. Computational modeling has also been used to predict CYP metabolism–based DDIs [31]. Furthermore, data mining of scientific literature, electronic medical records or adverse event databases is an emergent approach for DDI detection [3],[24].

As part of a growing trend, many studies have also incorporated similarity based modeling in identifying known, and predicting unknown DDIs. Vilar et al. [37] described a protocol applicable on a large scale to predict novel DDIs based on similarity of drug interaction candidates to drugs involved in established DDIs. The model they proposed integrated structural 2D and 3D similarity measures, with known interaction profile, target and side-effect similarities. It was trained against a well established reference standard database of known DDIs and used to predict new potential DDIs and their implications.

4.1.2 Work-flow: Similarity-based prediction of psychiatric drug-nutrient interactions

Based on the protocol presented by Vilar et al. [37], a similar procedure was carried out to attempt a similarity-based prediction of psychiatric drugs' interactions with nutrients present in plant-based food. For that purpose, a reference standard database was constructed, incorporating information about 1763 FDA approved drugs and 69,356 DDIs. Drug-drug interaction profile similarity, ADE similarity information, as well as 2D structural and target similarity information was gathered for the reference data set and used to train a SVM classification model. This model was later used to predict interactions between the 64 phytochemicals and 85 drugs, classified as psycholeptics and psychoanaleptics, that were used to create the networks in figures 2.5 - 2.8.

4.1.3 Presentation of the SVM algorithm

SVM is an algorithm which belongs to the supervised learning models and is able to analyze data sets and recognize patterns. It can be used for both classification and regression. This algorithm can build a model that assigns new examples into one category or the other given a set of examples that belong to two categories. An SVM model is basically a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. SVM algorithms can also perform linear and non-linear classification. There is also a big drawback that the SVM is only directly applicable for two-class tasks, therefore algorithms that reduce the multi-class task to several binary problems have to be applied.

The SVM algorithms have many applications in Biology and Chemistry, especially in Chemoinformatics. Some of these applications are the classification of objects as diverse as proteins and DNA sequences, microarray expression profiles and mass spectra results. Also some other applications nowadays of SVMs are potency and active state and similarity searching of several compounds in combination with QSAR models [14] [33] [41].

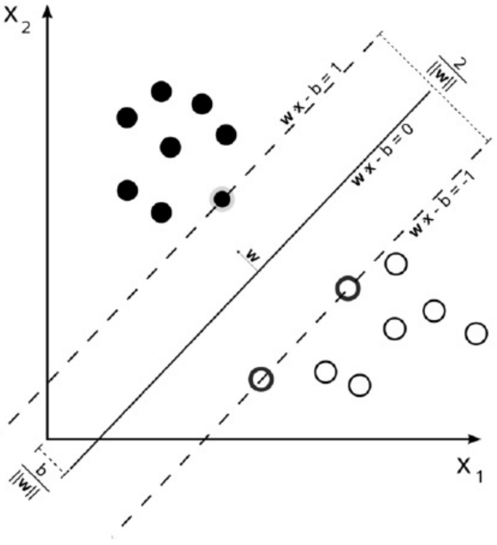


Figure 4.1: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors [12].

4.2 Materials and Methods

The protocol that was implemented involves the generation of the reference standard DDI database (matrix M_1) and the drug similarity databases (matrix M_2). These data are integrated through a straightforward process consisting of the extraction of the maximum value in each array of the matrix multiplication to generate the set of potential new DDIs (matrix M_3). The generation and handling of all matrices was carried out using Microsoft Excel's VBA. Finally, the information provided by M_3 is used to train an SVM model to predict new potential interactions. In the last stage of the procedure, the performance of the final model is assessed. This model is used to predict interactions in the database of psychiatric drugs and phytochemicals.

4.2.1 Generation of the reference standard DDI database (matrix M_1)

By using the Interax Interaction Search feature of Drugbank [16], the interactions for 1763 FDA approved drugs were downloaded and used to construct a symmetric

1763x1763 matrix (M_1), where each cell represented a drug-drug combination. If a interaction was documented in Drugbank for a specific combination, the respective cell was annotated as 1. An absence of documented interaction was annotated by 0. As the method is based on matrix multiplication and maximization, the values in the diagonal of the matrix needed to be set to 0 to avoid the growth of data noise caused by the ”similarity” of drugs with themselves. As part of the analysis, the pharmacological or clinical effects associated with the DDIs were also stored.

4.2.2 Generation of the drug similarity databases (matrix M_2)

The matrices M_2 were also symmetrical 1763x1763, where this time, each cell contained similarity information about the respective combination of drugs. Four different similarity types were evaluated, and thus, four different M_2 matrices were constructed.

Similarity based on 2D structural fingerprints

The basic 2D molecular structure fingerprint technique consists of representing a molecule as a bit vector that codifies the presence or absence of different substructural or pharmacophoric features in each bit position. Essentially a fingerprint is a codified vector version of the 2D molecular structure. There are different types of systems for codifying fingerprints, however in the present method MACCS (for Molecular Access System) structural keys were used [23]. As a way to represent a sparse binary vector, only the positions codifying the fragments present in the molecule are stored in the final fingerprint [37].

For each drug, a MACCS fingerprint was calculated, using Python 2.7 and the open source software OpenBabel. Following the instructions in the protocol of Vilar et al. [37], a Python script was written, that given a SMILES.txt list of tab-delimited chemical IDs and SMILES codes, e.g., such as;

```
CHEMBL973  C\C(=C(/C#N)\C(=O)Nc1ccc(cc1)C(F)(F)F)\O
CHEMBL1382 CC(C)N(CC[C@H](c1ccccc1)c2cc(C)ccc2O)C(C)C
```

it calculated all MACCs fingerprints for the drugs. The fingerprint type can be changed by specifying a different string for FINGERPRINT. Available fingerprints are FP2, FP3, FP4 and MACCS [37].

In order to calculate the similarity between two fingerprints, belonging to two different drugs, the script went on to calculate the Tanimoto coefficient for each pair. The TC is also known by the term Jaccard index. The TC can adopt values in the range between 0 (maximum dissimilarity) and 1 (maximum similarity). The TC between fingerprints A and B is defined as [43]:

$$TC = \frac{N_{AB}}{N_A + N_B + -N_{AB}} \tag{4.1}$$

where , N_A and N_B is the number of features present in fingerprints A and B, respectively, and N_{AB} is the number of features present in common to both fingerprints A and B.

This script generates an output TC_results.csv file containing all pairwise TCs above a specified cutoff (T_CUTOFF, default 0.00).

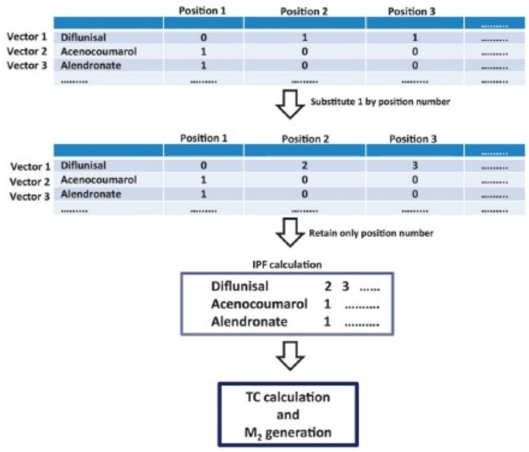
Script 4.1: Using Python to calculate molecular fingerprints and TC between all drug pairs with Open Babel

```
# -*- coding: cp1252 -*-
import csv
import subprocess
import re
import os
T_CUTOFF = 0.00
FINGERPRINT = "MACCS"
FILENAME = "SMILES.txt"
input_read = open(FILENAME,"r")
input_temp = open("temp_smi_file.txt","w")
input_dict = dict()
for line in input_read:
    newline =line.split()
    id = newline[0]
    smiles = newline[1]
    input_dict[id] = smiles
    input_temp.write("%s\t%s\n" %(smiles, id) )
```

```
input_read.close()
input_temp.close()
f = open("TC_results.csv", "w")
writer = csv.writer(f)
writer.writerow(["chemical1", "chemical2", "TC"])

for chemical1 in input_dict:
    babel_command= "obabel -ismi -: %s  temp_smi_file.txt -ofpt -xf%
       s" %(input_dict[chemical1], FINGERPRINT)
    output = subprocess.Popen(babel_command, shell=True, stdout=
       subprocess.PIPE, stderr=subprocess.PIPE)
    TC_list = []
    n=0
    while True and n<1000000:
        n=n+1
        line = output.stdout.readline()
        #line example:  >CHEMBL1382 Tanimoto from CHEMBL973 = 0.2
        if line != "":
            newline = re.split(">|=", line)
            #newline: [ , "CHEMBL1382 Tanimoto from CHEMBL973 ", "
               0.2\n"]
            #indices: [0] [1] [2]
            if len(newline) > 2:
                id_catcher = newline[1].split()
                chemical2 = id_catcher[0]
                TC = float(newline[2].strip())
                TC_list.append((chemical2, TC))
        else:
            break
    for chemical2,TC in TC_list:
        if TC > T_CUTOFF and chemical1 != chemical2:
            writer.writerow([chemical1, chemical2, TC])
        else:
            writer.writerow([chemical1, chemical2, 0])
f.close()
os.remove("temp_smi_file.txt")
```

The calculated Tanimoto coefficients for all pairs of drugs were then transformed into



		Position 1	Position 2	Position 3	....
Vector 1	Diflunisal	0	1	1	....
Vector 2	Acenocoumarol	1	0	0	....
Vector 3	Alendronate	1	0	0	....
	....	....	....	....	....

		Position 1	Position 2	Position 3	....
Vector 1	Diflunisal	0	2	3	....
Vector 2	Acenocoumarol	1	0	0	....
Vector 3	Alendronate	1	0	0	....
	....	....	....	....	....

Diflunisal	2	3	......
Acenocoumarol	1	....	
Alendronate	1	....	

Figure 4.2: Workflow of the steps carried out during the calculation of Tanimoto coefficients for target profile similarity. Graph adapted from Vilar et al. [37].

the symmetrical M_2 matrix, where each element (i, j) contained the TC between the drug in row i and the drug in column j. All diagonal values were again set to 0.

Similarity based on target profile fingerprints

A second M_2 matrix was generated, by evaluating the similarity of drugs, in respect to whether they target the same proteins. For that purpose all the targets, enzymes, carriers, transporters for each FDA approved drug were downloaded from Drugbank [16]. Next, a table was constructed, where each row represented a drug and each column represented one of the 1990 protein targets. Each cell (i, j) was marked as 1 if the drug i was found to target protein j. This information was then coded into ”target profile fingerprints” in the following manner; Each different protein was assigned a number, from 1-1990. Then, values 1 in each cell in the matrix were substituted with their column’s target number. As a result, each drug in the matrix was matched to a serial number (”fingerprint”), comprising of a series of numbers, corresponding to the proteins it targets. This procedure is depicted in figure 4.2.

These target profile fingerprints were used to calculate Tanimoto coefficients for each pair of drugs, using Python (script 4.2) and following the instructions of Vilar et al. [37]. The input of this script is a M2input.txt file with the following format:

```
id1, 1 3 45 234 237
id2, 4 5 355 546
id3, 1 2 3 87 1080
```

The pairwise TCs were then exported in the M2_TC_results.txt and, eventually, transformed into yet another symmetrical M_2 matrix, where each element (i, j) contained the target similarity TC, between the drug in row i and the drug in column j. All diagonal values were again set to 0.

Script 4.2: Using Python to calculate TCs between fingerprints

```
# -*- coding: utf-8 -*-
import os
import sys
from collections import defaultdict

def tanimoto(a, b):
    if len(a)==0 and len(b)==0:
        return 0
    else:
        return len(a&b)/float(len(a|b))



fingerprints = defaultdict(set)

FILENAME=open("M2input.txt","r")
sys.stdin=FILENAME

OUTPUTFILE=open("M2_TC_results.txt","w")
sys.stdout=OUTPUTFILE
```

```
for line in sys.stdin:
    identifier, fpt = line.split(",")
    fingerprints[identifier] = set(fpt.split())



for id1 in sorted(fingerprints.keys()):
    for id2 in sorted(fingerprints.keys()):
        if id1 > id2:
            continue
        similarity = tanimoto(fingerprints[id1], fingerprints[id2])
        print >> sys.stdout, "%s\t%s\t%f" % (id1, id2,similarity)

OUTPUTFILE.close()
```

Similarity based on interaction profile fingerprints (IPFs)

Another way, proposed by Vilar et al. [37], to evaluate the similarity between two drugs was by comparing their interaction profiles. More specifically, two drugs were considered similar if they interacted with the same drugs in the initial M_1 reference interaction matrix. For that purpose, a new matrix was constructed, similar to M_1, where every row and column represented one of the 1763 FDA approved drugs, and known interactions between drugs were marked with 1 in the corresponding cells. The procedure following that was identical to the one carried out in the evaluation of target similarity profiles. Each column was annotated with a position number, which was later used to substitute the values 1 in each particular column. In that manner, a serial number for each row was generated (”interaction profile fingerprint or IPF”). The IPFs were, then, used as input for script 4.2, in order to calculate the new Tanimoto coefficients that measured the interaction profile similarity between the pairs of drugs. Finally, a third M_2 matrix was built, where each element (i, j) contained the interaction profile similarity TC, between the drug in row i and the drug in column j. Diagonal values were again set to 0, as before.

Similarity based on ADE profile fingerprints

Moreover, the similarity between drugs was evaluated based on their adverse effect profiles [37]. Information about drug ADEs was downloaded from the SIDER database [26]. A matrix was then built, where each row represented one of the 1763 drugs and each column one of 5059 side-effects. ADE profile fingerprints and pairwise TCs were then calculated, the same way, as for interaction and target profiles. This way the fourth M_2 matrix was generated, where each element (i, j) contained the ADE similarity TC, between the drug in row i and the drug in column j. Once again, diagonal values were set to 0.

4.2.3 Generation of the new set of potential DDIs (matrix M_3)

In this stage of the protocol, the two databases M1 and M2 are integrated, in the way that Vilar et al. [37] proposed. The objective here is to obtain the matrix M3 that contains all the possible scored DDIs through the multiplication of M_1 by M_2, retaining only the highest value in the array multiplication in each cell. This procedure is illustrated in figure 4.3. The resulting M_3 was not symmetric (since the action performed was not conventional multiplication), and had to be converted into a symmetric matrix, by keeping the highest value on either side of the diagonal. Vilar et al. [37] claim it is also possible to associate clinical effects with the new DDIs, in the way illustrated in figure 4.4. As four different ways of similarity were examined, the obtained result were four different M_3 matrices. In effect, for each drug-drug combination, four TC variables were assigned.

4.2.4 Training and assessment of the SVM model

The information extracted from the M_3 matrices amounted to 1,553,203 unique drug-drug combinations, each described by four TC variables. Moreover, 69,356 of these combinations were noted as established interactions, which were recorded in the initial reference database. A Support Vector Machine classification algorithm was implemented in Molegro Data Modeller 2013.3.0 [6], by using a subset of 50,322 of drug-drug combinations (10,918 known interactions). The selection was done, by using a N dimensional

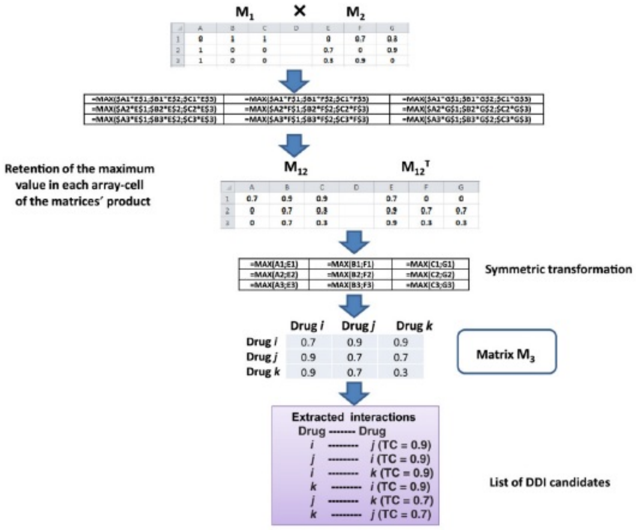


Figure 4.3: Generation of the new set of potential DDIs (matrix M_3). Vilar et al. [37].

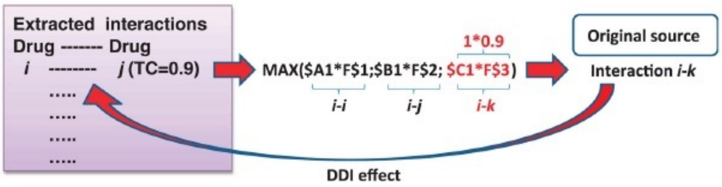


Figure 4.4: DDI effect linkage: list of DDIs extracted from M_3 are associated with the initial source in M_1 and with the clinical or pharmacological effects caused by the interaction.). Vilar et al. [37].

grid method, in order to remove redundant, overlapping data. The SVM was trained by using as a target variable a boolean (1 or 0) variable that indicated whether a data point represented a well established, initial interaction. The Tanimoto coefficients of IPFs, targets, ADEs and structural fingerprints were used as independent variables. A larger subset of training data could have been used, however, due to computational time restrictions, and the good performance of the model generated, this was not attempted in the present analysis.

The model generated had an accuracy of 80.83%. Its precision, recall and f-measure was 82.15%, 96.48% and 88.74% , respectively, in predicting an absence of interaction (0). Using 5-fold validation, the corresponding accuracy was 80,76%, while precision, recall and f-measure were calculated as 82.12%, 96,42% and 88.70%. There results display an adequate capability of the SVM to classify the data. Furthermore, over-fitting the data, is most likely avoided, since the 5-fold validation did not affect significantly the accuracy of the model.

4.2.5 Building the test set of phytochemicals and psychiatric drugs

A test database of phytochemicals and psychiatric drugs was constructed in an identical way as the training database. A M_1 149x149 matrix was generated for the 64 phytochemicals and 85 psychiatric drugs. Known interactions were sought for in Drugbank for the psychiatric drugs, as well as the phytochemicals that were already documented as nutraceuticals. For example, Epigallocatechin and Catechin may be found in Drugbank under the code DB01266 (Sinecatechins), an approved botanical drug product for topical use. Structural TCs were calculated, using MACCS fingerprints, similarly to the training set. In order to calculate target profile fingerprints, the information about the common targets that was obtained from Drugbank and NutriChem, in the first part of this paper, was utilized (Chapter 2.). Finally, ADE information was obtained from SIDER, whenever it was available. Unfortunately, this information was scarce in the case of phytochemicals. The M_3 matrices were calculated the same way, as before, resulting in a data set of 11,026 unique combinations, each described by 4 similarity measures (TCs). Table 4.1 lists the phytochemicals and psychiatric drugs that were studied.

Table 4.1: List of the 64 phytochemicals and 85 psychiatric drugs

Phytochemicals		Psychiatric drugs		
(-)-Epigallocatechin	Iso-Ompa	DB00176	DB00734	DB01623
(+)-Catechin	Kaempferol	DB00182	DB00752	DB01624
(S)-Laudanosine	Liquiritin	DB00215	DB00780	DB04599
2,3-Butanedione	Luteolin	DB00246	DB00805	DB04820
9-Amino-1,2,3,4-Tetrahydroacridine	Myricetin	DB00289	DB00831	DB04821
Alpha-Aminoadipic Acid	N-Acetylcysteine	DB00321	DB00843	DB04832
Aminooxyacetate	Naringenin	DB00334	DB00850	DB04836
Apigenin	Nicotine	DB00363	DB00875	DB04842
Aporphine	Nicotinic Acid	DB00370	DB00933	DB04896
Arecoline	Nordihydroguaiaretic Acid	DB00382	DB00934	DB04946
Atenolol	Papaveraldine	DB00391	DB00989	DB06144
Caffeine	Papaverine	DB00408	DB01043	DB06148
Carbachol	Paroxetine	DB00409	DB01063	DB06216
Cortisone	P-Benzoquinone	DB00420	DB01104	DB06288
Daidzein	Phenylbutazone	DB00422	DB01142	DB06594
Dihydrocapsaicin	Propidium Iodide	DB00450	DB01149	DB06684
Donepezil	Quercetin	DB00458	DB01151	DB06700
Epicatechin Gallate	Rivastigmine	DB00472	DB01171	DB08815
E-Piceatannol	Rutoside	DB00476	DB01175	DB08996
Eriodictyol	Salsolinol	DB00477	DB01224	DB09014
Falcarindiol	Sanguinarine	DB00490	DB01238	DB09016
Fisetin	Serotonin	DB00502	DB01239	
Flavanone	Sodium Nitroprusside	DB00508	DB01242	
Galangin	Spermine	DB00540	DB01247	
Glutamic Acid	Taxifolin	DB00543	DB01267	
Harmine	Thapsigargin	DB00557	DB01356	
Hesperetin	Tramadol	DB00623	DB01403	
Higenamine	Trifluoperazine	DB00656	DB01608	
Hydrobenzoin	Tryptamine	DB00674	DB01614	
Hyperin	Tyramine	DB00679	DB01618	
Isatin	Wogonin	DB00715	DB01621	
Isoliquiritigenin	Xanthotoxin	DB00726	DB01622	

4.3 Results

When implementing the SVM trained model in generating a prediction for the test set of phytochemicals and psychiatric drugs, the following results were obtained. The accuracy, precision, recall and f-score for this model were 94.55%, 98.56%, 94.94% and 96.72% respectively. The confusion table can be found in table 4.2. It was, therefore, observed that the discerning ability of the SVM model for the data set of phytochemicals and psychiatric drugs was very high. Figure 4.5, illustrates the interaction network. This network consists only of the initial interaction data (values ”1”, extracted from Drugbank), and the predicted interactions (values ”1”, from the SVM model), between drugs and phytochemicals.

Table 4.2: Confusion matrix for the interaction predictions of the SVM model on the test set of phytochemicals and psychiatric drugs.

	Predicted non interacting	Predicted interacting	SUM
Non interacting	8849	472	9321
Interacting	129	1576	1705
SUM	8978	2048	11026

According to the information supplied by NutriChem, Paroxetine, Tramadol and Trifluoperazine are compounds that can be found in white-pepper, poppy-seed and garlic, respectively. Donepezil is a constituent of peach fruit, whereas Rivastigmine may be found in cashew nuts. Caffeine, on the other hand is the well known constituent of tea and coffee, that may also be found in cocoa, Typha angustifolia, Caulophyllum robustum, the California fawn lily, Ginkgo biloba and Bracken fern. For Xanthotoxin, the network of plant-based food that is cited by NutriChem is presented in figure 4.6. Carbachol may be found in ginger, Nepeta cataria, Panax ginseng and Viburnum prunifolium. Moreover, Papaverine is found in poppy-seed, and the network for Nicotine is presented in figure 4.6. Finally, Atenolol may be found in nutmeg and Cortisone is a constituent of sprouted lentil [21].

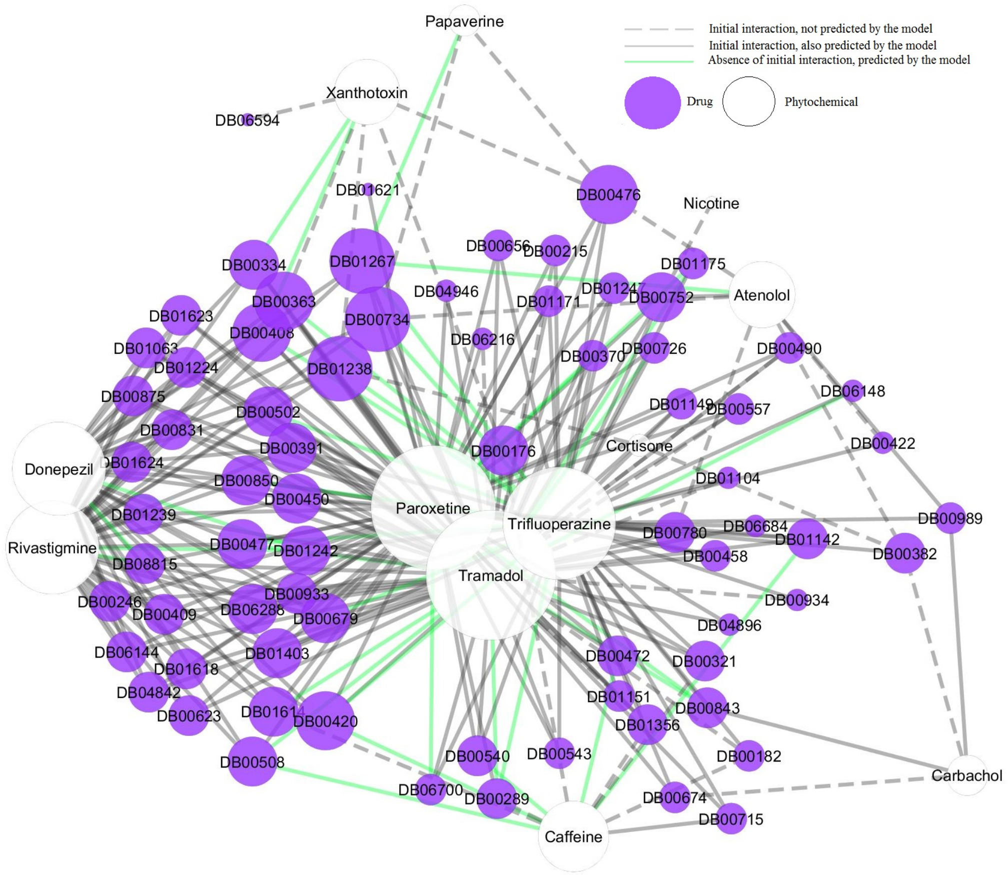


Figure 4.5: Interactions between psycholeptic and psychoanaleptic drugs (Drugbank code names used) and phytochemicals. Data extracted from Drugbank and predicted by the SVM model.

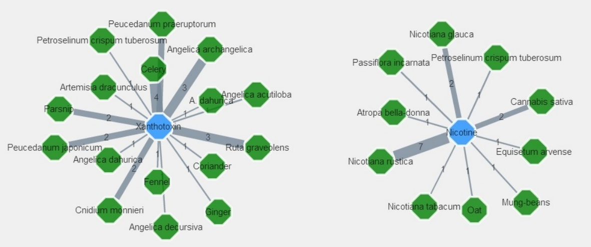


Figure 4.6: Plant-based food that contains Xanthotoxin and Nicotine[21]. Edge width indicates the number of references cited in NutriChem for each association.

Chapter 5

Discussion

Characteristics, limitations and points for improvement of the pharmacophoric screening approach

Because pharmacophoric screening is only a preliminary method in the discovery of relevant compounds, it is meant to be used as a coarse sieve; not too precise, however a high recall of true positives is usually desired. In that sense, the results of pharmacophoric screening are usually large aggregations of molecules that are later filtered based on additional criteria. The models that were built in the present analysis, and used to screen NutriChem for ligands that would bind with proteins P07550, P28222 and P14416, were characterized by mediocre results in recalling true positive hits and higher precision. These qualities would mean that many true hits could be missed, even in the initial stage of the discovery process. Moreover, the models were not proven to be very robust when tested against different validation sets, which could indicate an over-fitting of the models on the training data sets.

As a means to generate improved pharmacophore models, efforts should be made in improving the quality of the training sets. This could be done by docking simulations or literature searches to reveal in which specific binding pockets of the protein each compound binds, and filter out only these compounds that bind in the same pocket as the psychiatric drugs. Special care should be given, also, to filtering out only the specific isomers of a compound that bind with the protein. In the effort to build a robust pharmacophore model, the structure-based approach of the HypoGen algorithm could be integrated with receptor-based approaches. For example, the HypoGen algorithm could be enhanced with excluded volume restrictions, to avoid compounds that cannot bind due to steric obstructions. Alternatively, the pharmacophore models that were generated could be compared and integrated with receptor-based pharmacophore models, since complexes of ligands bounds with the proteins are available in PDB.

Characteristics, limitations and points for improvement of the similarity-based prediction model

The similarity-based approach for the prediciton of drug-nutrient interactions had several significant limitations and caveats. Although the model displayed good robustness in hold-out validation and against the test set of phytochemicals/psychiatric drugs, it's performance in predicting new, potential interactions, that were not documented in the initial reference set, is limited. The approach that was adopted by Vilar et al. [37] depends vastly on the quality of the initial reference standard database, and expands upon the well-established DDIs, known in the literature. Therefore, in cases where limited information is available about established interactions, as is the case with newly marketed drugs, or phytochemicals, the performance of the model is hindered. A second caveat is that the protocol introduced by Vilar et al. [37] was designed for use on large-scale analysis. Therefore, it is not suitable for detection of small variations in the similarity measure that can strongly affect the biological effects of drugs. Moreover, it is expected that different similarity measures might be more suitable to detect interactions in different categories of drugs.

A more comprehensive similarity-based approach to predict nutrient-drug interactions is one that should be better tailored to the available information from the literature, concerning phytochemicals. In the case of compounds found in plant-based food, only partial (and thus biased) information is available to be used in the calculation of IPF similarity or ADE similarity. Therefore, a more solid approach to this particular problem could use only target and structural similarity measures in order to create a good classification model. Moreover, the training reference data set could be better designed to involve phytochemicals, as well as drugs. Due to the nature of the protocol implemented by Vilar et al. [37], a larger-scale approach, that would incorporate inter-class interactions of phytochemicals with many different ATC categories of drugs, would be better suited for analysis with a similarity-based model. Different similarity measures could also be employed, such as 3D structural similarity, provided that available information could be found in the literature.

Bibliography

[1] F.M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, M. Altaf-Ul-Amin, and L.K. et al. Darusman. Knapsack family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.*, 53, 2012.

[2] Mahreen Arooj, Sugunadevi Sakkiah, Songmi Kim, Venkatesh Arulalapperumal, and Keun Woo Lee. A combination of receptor-based pharmacophore modeling & QM techniques for identification of human chymase inhibitors. *PloS one*, 8(4):e63030, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0063030. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3637262{&}tool=pmcentrez{&}rendertype=abstract`.

[3] Percha B, Garten Y, and Altman RB. Discovery and explanation of drug-drug interactions via text mining. *Pac Symp Biocomput.*, page 410–421, 2012.

[4] Asma Ben Abacha, Md Faisal Mahbub Chowdhury, Aikaterini Karanasiou, Yassine Mrabet, Alberto Lavelli, and Pierre Zweigenbaum. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification, 2015. ISSN 15320464.

[5] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J.P. Overington. The chembl bioactivity database: an update. *Nucleic Acids Res.*, 42:1083–1090, 2014. doi: 10.1093/nar/gkt1031.

[6] CLC BIO. Molegro data modeller: User manual, 2013. URL http://www.clcbio.com/files/usermanuals/MDM_manual.pdf.

[7] Dassault Systèmes BIOVIA. Discovery studio modeling environment. San Diego: Dassault Systèmes, 2015. Release 4.5.

[8] T. Bjornsson, J. Gallaghan, H. Einolf, and et. al. The conduct of in vitro and in vivo drug-drug interaction studies: A pharmaceutical research and manufacturer's of america (phrma) perspective. *Drug Metabolism and Disposition*, 31(7):812–832, 2003.

[9] R. Z. Cer, U. Mudunuri, R. Stephens, and F. J. Lebeda. IC50-to-Ki: a web-based tool for converting IC50 to Ki values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Research*, 37(Web Server):W441–W445, 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp253. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkp253.

[10] Lingtak-Neander Chan. Drug-nutrient interactions. *Journal of Parenteral and Enteral Nutrition*, 37(4):450–459, 2013. ISSN 19412444, 01486071. doi: 10.1177/0148607113488799.

[11] Mary L. Chavez, Melanie A. Jordan, and Pedro I. Chavez. Evidence-based drug–herbal interactions. *Life Sciences*, 78(18):2146–2157, 2006. ISSN 00243205. doi: 10.1016/j.lfs.2005.12.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0024320505012440.

[12] Wikimedia Commons. Svm max sep hyperplane with margin. Wikimedia Commons, the free media repository, 2015. URL https://commons.wikimedia.org/wiki/File:Svm_max_sep_hyperplane_with_mar-gin.png(image2).

[13] The UniProt Consortium. Uniprot: a hub for protein information. *Nucleic Acids Res.*, 43:D204–D212, 2015.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[15] Martin E. Dowty, Dean M. Messing, Yurong Lai, and Leonid (Leo) Kirkovsky. *ADME*, pages 145–200. John Wiley & Sons, Inc., 2011. ISBN 9780470915110. doi: 10.1002/9780470915110.ch4. URL `http://dx.doi.org/10.1002/9780470915110.ch4`.

[16] Wishart DS., Knox C., Guo AC., Shrivastava S., Hassanali M., Stothard P., Chang Z., and Woolsey J. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668–72, 2006.

[17] WHO Collaborating Centre for Drug Statistics Methodology. Atc/ddd index 2016. `http://www.whocc.no/`, 2016. Accessed: 2016-01-23.

[18] Adriane Fugh-Berman and E. Ernst. Herb–drug interactions: Review and assessment of report reliability. *British Journal of Clinical Pharmacology*, 52(5):587–595, 2001. ISSN 1365-2125. doi: 10.1046/j.0306-5251.2001.01469.x. URL `http://dx.doi.org/10.1046/j.0306-5251.2001.01469.x`.

[19] C.-Y Huang, V Olieric, P Ma, N Howe, L Vogeley, X Liu, R Warshamanage, T Weinert, E Panepucci, B Kobilka, K Diederichs, M Wang, and M. Caffrey. In meso in situ serial X-ray crystallography of soluble and membrane proteins at cryogenic temperatures. *Acta Crystallogr.,Sect.D*, 72:93–112, 2016. URL `http://www.rcsb.org/pdb/explore/explore.do?structureId=5D5A`.

[20] Leslie L. Iversen. Psychiatric disorders. *Drug Discovery Today: Therapeutic Strategies, Drug Discovery Today. Therapeutic Strategies, Drug Discov. Today Ther. Strateg, Drug Discov Today Ther Strateg, Drug Discovery Today*, 5(3):143–144, 2008. ISSN 17406773. doi: 10.1016/j.ddstr.2009.02.001.

[21] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki. NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Research*, 43(D1):D940–D945, 2015. ISSN 0305-1048. doi: 10.1093/nar/gku724. URL `http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku724`.

[22] Kasper Jensen, Yueqiong Ni, Gianni Panagiotou, and Irene Kouskoumvekaki. Developing a Molecular Roadmap of Drug-Food Interactions. *PLOS Computational Biology*, 11(2):e1004048, 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004048. URL `http://dx.plos.org/10.1371/journal.pcbi.1004048$\backslash$nhttp://www.ncbi.nlm.nih.gov/pubmed/25668218`.

[23] Durant JL, Leland BA, Henry DR, and Nourse JG. Reoptimization of mdl keys for use in drug discovery. *J Chem Inf Comput Sci.*, 42:1273–1280, 2002.

[24] Tari L, Anwar S, Liang S, Cai J, and Baral C. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *J Med Chem.*, 26:i547–i553, 2010.

[25] Mei Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X. Chen, H Xu, and et. al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association : JAMIA*, 19(e1):e28–e35, 2012. URL `http://doi.org/10.1136/amiajnl-2011-000699`.

[26] Kuhn M, Campillos M, Letunic I, and et al. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 6:343, 2010.

[27] Rahnasto M, Raunio H, Poso A, Wittekindt C, and Juvonen RO. Quantitative structure-activity relationship analysis of inhibitors of the nicotine metabolizing cyp2a6 enzyme. *J Med Chem.*, 48:440–449, 2005.

[28] F Manetti, F Corelli, M Biava, R Fioravanti, G C Porretta, and M Botta. Building a pharmacophore model for a novel class of antitubercular compounds. *Farmaco (Società chimica italiana : 1989)*, 55(6-7):484–91, 2000. ISSN 0014-827X. URL `http://www.ncbi.nlm.nih.gov/pubmed/11204750`.

[29] Adam McCluskey, Paul A. Keller, Jody Morganb, and James Garnera. Synthesis, molecular modeling and biological activity of methyl and thiomethyl substituted pyrimidines as corticotropin releasing hormone type 1 antagonists. *The Royal Society of Chemistry*, Supplementary data, 2003.

[30] J.H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, 2 edition, 2009.

[31] Hudelson MG and et al. High confidence predictions of drug-drug interactions: predicting affinities for cytochrome p450 2c9 with multiple computational methods. *J Med Chem.*, 51:648–654, 2008.

[32] V. Neveu, J. Perez-Jiménez, F. Vos, V. Crespy, L. du Chaffaut, L. Mennen, C. Knox, R. Eisner, J. Cruz, D. Wishart, and A. Scalbert. Phenol-explorer: an online comprehensive database on polyphenol contents in foods. *Database*, 2010, 2010. doi: 10.1093/database/bap024. URL `http://database.oxfordjournals.org/content/2010/bap024.abstract`.

[33] W. S. Noble. What is a support vector machine. *Nature biotechnology*, 24(12): 1565–1567, 2006.

[34] Grace Shema Nzabonimpa. Pharmacophore in drug design. Center for Biological Sequence Analysis CBS DTU: Computational Chemical Biology Group, 2015.

[35] S Pandalaneni, V Karuppiah, M Saleem, L.P Haynes, R.D Burgoyne, O Mayans, J.P Derrick, and L. Lian. Neuronal Calcium Sensor-1 Binds the D2 Dopamine Receptor and G-Protein-Coupled Receptor Kinase 1 (Grk1) Peptides Using Different Modes of Interactions. *J.Biol.Chem.*, 290:18744, 2015. doi: 25979333. URL `http://www.rcsb.org/pdb/explore/explore.do?structureId=5AER`.

[36] Dalhoff K. Schmidt LE. Food-drug interactions. *Drugs*, 62(10):1481–1502, 2002.

[37] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, and Nicholas P Tatonetti. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nature Protocols*, 9(9):2147–2163, 2014. ISSN 1754-2189. doi: 10.1038/nprot.2014.151. URL `http://www.nature.com/doifinder/10.1038/nprot.2014.151`.

[38] C. Wang, Y. Jiang, J. Ma, H. Wu, D. Wacker, V. Katritch, G.W. Han, W. Liu, X.P. Huang, E. Vardy, J.D. McCorvy, X. Gao, X.E. Zhou, K. Melcher, C. Zhang, F. Bai,

H. Yang, L. Yang, H. Jiang, B.L. Roth, V. Cherezov, R.C. Stevens, and H.E. Xu. Structural basis for molecular recognition at serotonin receptors. *Science*, 340: 610–614, 2013. doi: 23519210. URL `http://www.rcsb.org/pdb/explore/explore.do?structureId=4iar`.

[39] Fengxiao Wang and Yadong Chen. Pharmacophore models generation by catalyst and phase consensus-based virtual screening protocol against PI3Kα inhibitors. *Molecular Simulation*, 39(7):529–544, 2013. ISSN 0892-7022. doi: 10.1080/08927022.2012.751592. URL `http://www.tandfonline.com/doi/abs/10.1080/08927022.2012.751592`.

[40] Hongwu Wang, Ruth A. Duffy, George C. Boykow, Samuel Chackalamannil, and Vincent S. Madison. Identification of novel cannabinoid cb1 receptor antagonists by using virtual screening with a pharmacophore model. *Journal of Medicinal Chemistry*, 51(8):2439–2446, 2008. doi: 10.1021/jm701519h. URL `http://dx.doi.org/10.1021/jm701519h`. PMID: 18363352.

[41] A. M. Wassermann, K. Heikamp, and J. Bajorath. Potency□directed similarity searching using support vector machines. *Chemical biology & drug design*, 77(1): 30–38, 2006.

[42] C.G Wermuth, C.R. Ganellin, P. Lindberg, and L.A. Mitscher. Glossary of terms used in medicinal chemistry (iupac recommendations 1998). *Pure Appl. Chem.*, 70 (5):1129–1143, 1998. URL `http://dx.doi.org/10.1351/pac199870051129`.

[43] D Wild. *Introducing Cheminformatics*. Lulu, 2012. URL `http://www.citeulike.org/group/14458/article/10885453`.