PROBABILISTIC LATENT FACTOR MODELS FOR TRANSFORMATIVE DRUG DISCOVERY

by

Murat Can Cobanoglu

BS, Sabanci University, 2008

MS, Sabanci University, 2010

Submitted to the Graduate Faculty of

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Murat Can Cobanoglu

It was defended on

June 9th, 2015

and approved by

Ziv Bar-Joseph, PhD, Associate Professor, CMU Machine Learning Department

Zoltán N. Oltvai, MD, Associate Professor, Department of Pathology

Gary Silverman, MD, PhD, Professor, Pediatrics, Cell Biology and Physiology

Andrew Stern, PhD, Associate Professor, University of Pittsburgh Drug Discovery Institute

Dissertation Co-Advisor: Ivet Bahar, PhD, Distinguished Professor, Department of Computational and Systems Biology

Dissertation Co-Advisor: D. Lansing Taylor, PhD, Professor, Director, University of Pittsburgh Drug Discovery Institute

PROBABILISTIC LATENT FACTOR MODELS FOR TRANSFORMATIVE DRUG DISCOVERY

Murat Can Cobanoglu, M.S.

University of Pittsburgh, 2015

ABSTRACT: The cost of discovering a new drug has doubled every 9 years since the 1950s. This can change by using machine learning to guide experimentation. The idea I have developed over the course of my PhD is that using latent factor modeling (LFM) of the drug-target interaction network, we can guide drug repurposable efforts to achieve transformative improvements. By better characterizing the drug-target interaction network, it is possible to use currently approved drugs to achieve therapies for diseases that currently are not optimally treated. These drugs might be directly used through repurposing, or they can serve as a starting point for new drug discovery efforts where they are optimized through medicinal chemistry methods. To achieve this goal, I have developed LFM-based techniques applicable to existing databases of drug-target interaction networks. Specifically, I have started out by establishing that probabilistic matrix factorization (PMF; one type of LFM algorithm) can be used as descriptors by showing they capture therapeutic function similarities that state-of-the-art 3D chemical similarity methods could not capture. Then I have shown that PMF can effectively predict unknown drug-target interactions. Furthermore, I have used newly developed computational techniques for discovering repurposable drugs for two diseases, α1 antitrypsin (α1-AT) deficiency (ATD) and Huntington's disease (HD) leading to successful discoveries in both. For ATD, two sets of data generated by the David Perlmutter and Gary Silverman laboratories have been used as input to deduce potential targets and repurposable drugs: (i) a high throughput screening data from a genome-wide RNAi knockdown in a *C. elegans* model for studying ATZ (Z-allele of α1-AT), and (ii) data from Prestwick library screen for the same model. We have predicted that the antidiabetic drug glibenclamide would be beneficial against ATZ aggregation, and data collected to date in *Mus musculus* models are promising. We have worked on HD with the Robert Friedlander lab, by examining the potential drugs and implicated pathways for 15 neuroprotective (repurposable) drugs that they have identified in a two-stage screening study. Based on LFM-based analysis of the targets of these drugs, we have developed a number of hypotheses to be tested. Among them, the antihypertensive drug sodium nitroprusside appears to be effective against HD based on neuronal cell death inhibition experiments that were conducted at the University of Pittsburgh Drug Discovery Institute as well as the Friedlander lab. Finally, we have built a web server, named BalestraWeb, for facilitating the use of PMF in repurposable drug identification by the broader community. BalestraWeb enables users to extract information on known and potential targets (or drugs) for any approved drug (or target), simply by entering the name of the query drug (or target). I have also laid out the framework for developing an integrated resource for quantitative systems pharmacology, Balestra toolkit (BalestraTK), which would take advantage of existing databases such as STITCH, UniProt, and PubChem. Collectively, our results provide firm evidence for the potential utility of machine learning techniques for assisting in drug discovery.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to acknowledge my advisors, Drs. Ivet Bahar and Lans Taylor, my thesis committee, Drs. Gary Silverman, Ziv Bar-Joseph, Zoltan Oltvai, Andrew Stern, and our collaborators, Drs. David Perlmutter and Robert Friedlander for their valuable contributions at the various stages of my PhD studies. I would like to thank Stephen Pak, Linda O'Reilly and Olivia Long for the amazing collaboration that we have had together on the ATD project. I would like to thank Diane Carlisle and Hossein Mousavi in the Friedlander Lab, and Laura Vollmer in the Drug Discovery Institute (UPDDI) for their valuable experimental work in the HD project. In the UPDDI, I would like to particularly thank Celeste Reese for her valuable effort in teaching me how to design and run drug discovery experiments using the cutting edge tools and techniques available there, Mark Schurdak for his assistance in overseeing my experimental education and work, Seia Comsa for offering help and advice whenever needed and last but not least the entire UPDDI personnel for providing such an enjoyable work environment.

1.0 INTRODUCTION

The pharmaceutical industry is in a crisis: for every one dollar's worth of patent-protected therapeutic revenue expiring, the industry can generate only 26 cents of revenue through new patented therapeutics [1]; bringing a new molecular entity to the market is estimated to cost around $1.8 billion [2]; the rate of new drug discovery per billion dollars of research and development spending has steadily halved every 9 years for the last 60 years giving credence to the so-called "Eroom's Law" which is "Moore's Law" read backwards [3]; and the success rate in the present drug discovery and development process (from beginning to end) is only 4% [2]. Moreover, we are witnessing a major shift in the pharmaceutical industry financial landscape, with major companies being driven toward mergers in order to create a larger "pipeline" of potential drugs to supplement the low R&D productivity with this trend having started more than a decade ago [4]. However even these merged giants could deliver only eleven FDA approved drugs out of the thirty approved in 2011 [5]. The culmination of these difficulties is that traditional pharma companies are reported to have lost 2% sales and 2.5% earnings in the first half of 2014 compared to their reported sales and earnings from one year ago [6].

The main causes of this crisis have been debated for a long time. The arguments put forward range from the perceived lack of managerial success [7] to the improper structuring of R&D divisions [3]. Yet, there is another set of arguments that suggest that the problem at the heart of the matter is more scientific than managerial: Hopkins and Overington have been pioneering of the idea of multiple-target modulation and the replacement of the high affinity binding of a single peptide paradigm by polypharmacology paradigm for years [8-10]. Many other scientists have also emphasized the need for a paradigm shift, as there has been a mounting evidence on the promiscuity of drugs as well as targets [11-15].

The efficiency of drug discovery and development might be improved by adopting a systemic approach that takes into consideration the interaction of existing drugs and candidate compounds with the entire network of proteins and other biomolecules in a cell [16]. This approach can either be used for repurposing, or it can be used to inform *de novo* drug discovery. To give a specific example about the latter use, we can make inferences on the drugs that share a common therapeutic activity with other drugs and use such inferences for hit diversification. Alternatively, we can identify the targets that are similar in their interaction patterns with known drugs, despite being dissimilar in structure or sequence and uncover novel biochemical properties, or potentially even biological pathways. Therefore there is a multitude of ways in which a systematic analysis of the interactome can reveal novel, useful insights.

The often-cited scientific justification for the paradigm shift is the observation that most single-target manipulations do not perturb biological systems: a pioneering work by Hillenmeyer et al. reports that only 34% of gene deletions result in disease or lethality, however 97% of the gene knockouts results in a phenotypic catastrophe when the gene deletion is combined with one or more small molecules under a specific environmental condition [17]. Moreover, the early work of Barabasi and Oltvai has established that most biological networks are essentially scale-free [18], which further corroborates the observation that single-target modulation of biological networks can have limited effect due to the fact that in a scale-free network most single-target perturbations will have minimal effect whereas those that fall on ‘hubs’ will have too strong an effect which in turn would make it hard to use as a therapeutic intervention. The relatively high ratio of drug failures due to safety concerns, reported as accounting for more than half of all failed projects in one recent review [19] and accounting for 20% of all phase II failures in another review [20] can be arguably be attributed, among other reasons, to the toxicity impact of modulating the so-called hub nodes of the scale-free biological network that comprises the cell. Furthermore the observation that the topological organization of the biological networks strongly reflects the underlying functional relationships [21] also helps develop an appreciation of pharmacological therapy as the modulation of a biological network instead of a simple 'lock-and-key' problem. Indeed, the "one gene, one drug, one disease" paradigm is now widely recognized to fail in describing experimental observations [8]. Many drugs act on multiple targets, and many targets are themselves involved in multiple pathways. For example, β-lactam antibiotics and most antipsychotic drugs exert their effect through interactions with multiple proteins [10;22]. Biological networks are highly robust to single-gene knockouts, as recently shown for yeast where 80% of the gene knockouts did not affect cell survival [17]. Similarly, a recent study showed that 81% of the 1,500 genes knocked out in mice would not cause embryonic lethality, further corroborating the robustness of biological networks against single target perturbagens [23].

These results suggest that quantitative systems pharmacology (QSP) strategies that take account of target (and drug) promiscuities can present attractive alternative routes to drug discovery. QSP approaches take into account complex biomolecular interactions in their cellular context. They combine computational and experimental studies in order to develop new compounds [22]. This requires a systems-level understanding of the biological process of interest, with detailed higher resolution modeling of the specific biochemical pathways of interest; along with supporting experimental data to help inform the entire effort.

The dissertation contains work that requires a broad base of understanding in both computational and biological disciplines. A major contribution to the field is the adoption of latent factor modeling (LFM) methods for analyzing the bipartite network of drug-target interactions and making predictions on potential new drug-target association. Therefore, I present below the background for different methods of computational drug-target interaction prediction. Furthermore, I also briefly present the background on two specific diseases, α-1 antitrypsin deficiency (ATD) and Huntington's disease (HD), which have been examined as two application areas of biological significance within the scope of this dissertation.

1.1 BACKGROUND ON COMPUTATIONAL METHODS

Recent years have seen many network-based models adopted to reduce the complexity of, and efficiently explore, drug-target interaction systems [10;22;24;25]. In particular, the development of computational methods that can efficiently assess potential new interactions became an important goal. Computational approaches used to predict unknown drug-target interactions can be divided into roughly four categories:

I. Chemical-similarity-based methods [26-28],

II. Integrative (both target- and chemical-similarity-based) methods [29-35],

III. Holistic approaches [36-41],

IV. Target-similarity-based methods [42-44].

The first two posit that if two entities are chemically or structurally similar they will share interactions – which is an assumption that may hold in multiple cases. However it is not guaranteed to hold universally as dissimilar chemicals can bind to different sites on the same protein and/or have allosteric effects. Furthermore the utility of different methods may depend on the size of the dataset being analyzed, e.g. computing chemical-chemical and target-target similarity matrices can be problematic for large databases like STITCH [45] (STITCH v3.1 has 210 million interactions between 2.6 million proteins and 300,000 chemicals). To overcome these limitations, holistic methods have been introduced, which utilize a number of different data sources such as gene expression perturbation [37;38] or high-throughput screening [40].

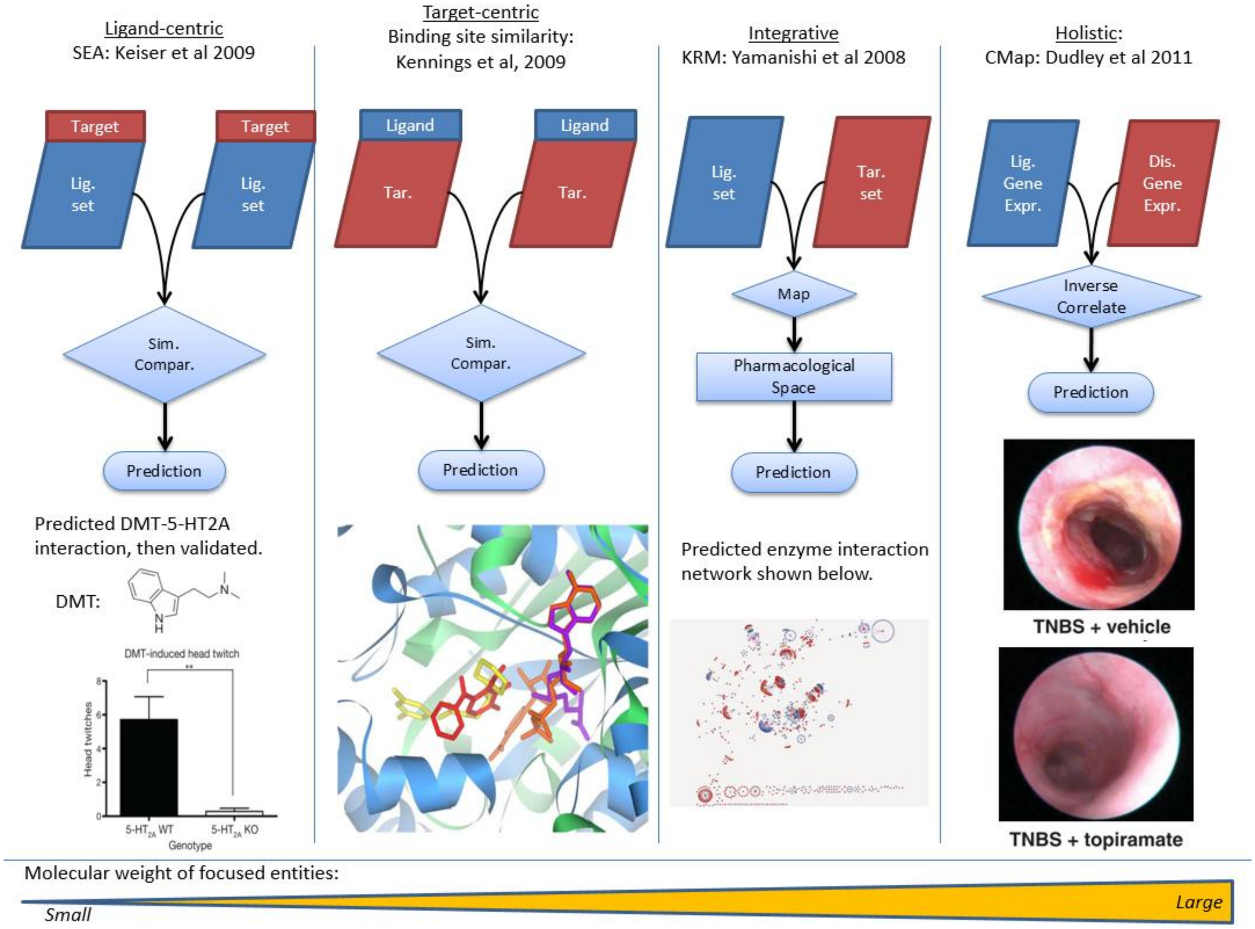


Figure 1: Summary categorization of current computational methods for polypharmacology predictions

One representative study from each main category is shown, along with figure(s) from the cited work to illustrate the results. The bar chart for ligand-centric methods shows that the interaction between DMT and 5-HT2A predicted by the method has been experimentally validated [28]. The figure in the holistic methods section shows that the validation of the prediction that topiramate would be useful in inflammatory bowel disease [37]. On the target-centric methods, the inset figure shows the ligand-binding site similarity between two target proteins, COMT and InhA, which serves as the basis for their subsequently validated prediction that comtan, an inhibitor of the COMT, would also inhibit InhA [42]. (see text for details).

1.1.1 Ligand-Centric Approaches

Among the ligand centric methods, the most significant is the Similarity Ensemble Approach (SEA) [26-28]. The SEA method was introduced by the Shoichet laboratory, in an article where they first used the method to relate protein pairs through similarities between their known ligands [26]. Later, this idea was adapted to drug repurposing predictions by comparing a single query chemical to all the known binders of each known protein [28]. More recently, predictions made by this method on a side effect target set were tested in a high throughput scheme by Novartis, in order to provide an unbiased assessment of the capabilities of the method: about 22% of the experimentally tested predictions turned out to be true predictions [27].

The SEA method is based on the calculation of the chemical similarity between the two sets of ligands known to interact with two different targets. Shoichet and coworkers have used the MDDR database [46] to retrieve data on the chemicals and their targets. They have used the 2D fingerprint similarity method (also known as Tanimoto similarity) to calculate the pairwise similarity between chemicals. This method entails the conversion of a chemical structure into a binary vector where each element of the vector indicates the presence/absence of a specific chemical feature. The similarity between two chemicals is then calculated by dividing the number of shared features by the total number of unique features present in both molecule vectors [47]. They have adapted the BLAST algorithm [48] to calculate the expectation that the chemical similarity between a set of chemicals and a specific query chemical of interest can be observed by chance. Using SEA, the authors have calculated the chemical similarity between a query chemical and all chemicals known to interact with a particular target. If a particular chemical is statistically significantly similar to all the drugs that are known to share a particular target, then it is predicted that this chemical would also work against that target.

The advantages of this method are that it can be easily applied to novel chemicals with no previously known interactions, and that it is rapid since it relies on chemical fingerprint based vector operations (which are efficient). The disadvantages are that it requires the target to have a large set of known and validated interaction partners (i.e. drugs).

1.1.2 Integrative Approaches

The methods that integrate chemical and biological information to generate polypharmacology predictions are termed here "integrative methods". Encapsulating as much information as possible to boost performance is an attractive idea. Consequently there is a significant body of research that focuses on the use of integrative approaches for polypharmacology prediction. There are multiple such methods: the kernel regression method [29], bipartite local models [31], integrated bipartite graph inference [32], SITAR [33], the unified probabilistic framework [35] and the Bayesian Matrix Factorization method [34].

The bipartite graph learning method of Yamanishi *et al.* is a good example of an integrative approach since it fundamentally describes a way of mapping drugs and proteins into the integrated 'pharmacological space' to then use proximity in this space to be indicative of interaction [29]. Moreover, this approach has been shown to work better than related studies [41] and has been the foundation for further techniques. Therefore I will discuss this method in some detail.

The authors employ three methods of generating polypharmacology predictions; all relying on calculating the similarity among chemicals, and likewise among proteins. The similarity between drugs is computed using the Tanimoto score for chemical fingerprints; the similarity between the targets is computed as the normalized Smith-Waterman score between the two sequences. The first method they describe is the nearest profile method, where they assign each new compound the interaction profile of the compound which has the highest similarity to the query. The second method is the weighted profile method where they weigh the interaction of each compound to compute the final interaction vector assigned to a query as the weighted sum of all the interaction vectors for all compounds with the weights being the similarity between the compounds. All the above described operations can also be applied for proteins to predict the drugs that would interact, since drugs and proteins are interchangeable with this methodology. Finally, they describe a novel method called the bipartite graph learning method, which employs a kernel regression model.

The bipartite graph learning method first entails the construction of a distance matrix **K** of size N + M between all compounds and proteins, where N is the number of compounds, M is the number of proteins. The element $K_{i,j}$ is the similarity between elements i,j if they are of the same type (i.e. both drugs or both proteins) or the shortest distance in the bipartite connectivity graph if they are of different types (i.e. one drug and the other protein, or *vice versa*). The matrix **K** is then decomposed into its eigenvalues and eigenvectors:

$$\mathbf{K} = \mathbf{\Gamma}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{\Gamma} = \mathbf{U}\mathbf{U}^{\mathbf{T}} \tag{1}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues and the columns of matrix $\mathbf{\Gamma}$ are the eigenvectors and $\mathbf{U} = \mathbf{\Gamma}\mathbf{\Lambda}^{1/2}$. The row vectors of **U** are then used to represent each drug and protein in the training set in an integrated ‘unified feature space’. Then a weight is learned for each compound and protein in the training set using kernel regression model, which entails finding the set of weights that minimize the loss function:

$$\mathbf{L} = \left\| \mathbf{U}\mathbf{U}^{\mathbf{T}} - \mathbf{S}\mathbf{W}\mathbf{W}^{\mathbf{T}}\mathbf{S}^{\mathbf{T}} \right\|_F^2 \tag{2}$$

where **S** is the similarity matrix and **W** is the weight matrix and $\|.\|_F$ is the Frobenius norm. To map a new compound to the integrated space, the vector corresponding to the new compound in the pharmacological space must be computed. This vector is computed as follows:

$$\mathbf{u}_{c_{new}} = f_c(c_{new}, c_i) = \sum_{i=1}^{n_c} s_c(c_{new}, c_i)\mathbf{w}_{c_i} \tag{3}$$

where $\mathbf{w}_{c_i}$ is a weight vector and $s_c(. \, , \, .)$ is a chemical structure similarity score. Likewise for proteins,

$$\mathbf{u}_{g_{new}} = f_g(g_{new}, g_j) = \sum_{j=1}^{n_g} s_g(g_{new}, g_j)\mathbf{w}_{g_j} \tag{4}$$

where $\mathbf{w}_{g_j}$ is a weight vector and $s_g(. \, , \, .)$ is a sequence similarity score. Finally, when a drug-target pair is queried for interaction, the drug and target are both mapped to the integrated space, and the dot product between their coordinate vectors in the integrated space is used as a measure of closeness between the query drug and target. If the drug and target are closer than a set threshold, they are declared to be interacting. The most important strength of this method is that it requires only the sequence of the proteins and just the chemical structure of the small-molecules (both of which are always available) therefore it is broadly applicable. The method has been later improved upon addition of pharmacological information [32].

The later work of Bleakley and Yamanishi treats polypharmacology predictions as a supervised learning problem [31]. Given a drug-target pair (d_i, t_j), the method entails labeling all proteins known to interact with d_i as one class (labeled +1) and all other proteins as another class (labeled -1), with a classifier trained to distinguish the interactors from the rest based on protein sequence. Then this classifier is used to predict the label of t_j. The same procedure is repeated to train a classifier that distinguishes drugs interacting with t_j (labeled +1) from the rest (labeled -1) based on the chemical structure of the drugs. Perlman *et al.* proposed to integrate many different similarity measures for comparing drugs and targets to define numerous features, which are then used to train classifiers for making polypharmacology predictions [33]. Five sources of information are used to compare a pair of drugs: chemical structure, side effect, perturbation of gene expression, ATC[1] code and finally ligand similarity. The ligand similarity is the overlap between the sets of SEA-predicted targets for each chemical [28]. The proteins are compared using sequence similarity, proximity in a protein-protein interaction network and overlap of Gene Ontology annotations [49]. The features are defined as one (out of five) chemical-chemical comparison method and one (out of three) protein-protein comparison method, for a total of 15 features. Then given a query drug-target pair, the similarity score of a feature is computed as:

$$Score(d,t) = \max_{d',t' \neq d,t} S_1(d,d')^r \cdot S_2(t,t')^{1-r}, 0 \leq r \leq 1 \qquad (5)$$

$Score(d,t) = max_{d',t' \neq d,t}\, S_1(d,d')^r \cdot S_2(t,t')^{1-r},\ 0 \leq r \leq 1$ where S_1 is the feature's drug-drug comparison method and S_2 is the feature's protein-protein comparison method, with *r* optimized through cross-validation. The authors then use forward-propagation feature selection (initially no features, most useful feature is added at each step) and backward-elimination (initially all features, most useless feature is dropped) to select 10 features that both techniques identified as useful. They then trained a logistic regression classifier on this feature set to separate the interacting pairs from the non-interacting ones. The authors report a classification performance of AUC[2] = 0.935 for their method, SITAR; whereas the kernel regression method of [29] is reported to have AUC=0.884 and the bipartite local models method [31] yields AUC=0.814.

[1] ATC: Anatomical, Therapeutic and Chemical classification system

[2] AUC: Area under the Receiver-Operator Characteristic curve; a classifier performance metric where the best possible classifier scores 1 and the worst possible classifier (random classification) scores 0.5.

SITAR is an excellent example of the utility of integrating multiple techniques for comparative analysis of drugs/proteins.

In the same spirit, Swann and coworkers suggested a method for integrating many diverse structure- and ligand-based comparison results to predict protein-chemical interactions in a robust manner [35]. Their method requires knowledge of actives and decoys[3] for each target. Given the actives / decoys for a target and a particular comparison technique they bin the range of scores computed by the method, then divide the number of actives in each bin by the total number of compounds (actives and decoys) in that bin to assign the probability of activity for the bin. They call this probability the 'belief' that the result from this technique is true. They assign such defined activity probabilities for the docking score computed by the FRED [50] and GLIDE [51] docking programs with four different force fields; the ECFP6 chemical fingerprint [52] overlap (Tanimoto score, computed as described when discussing the ligand-centric approaches) and ROCS three-dimensional spatial and physicochemical property overlap. They integrate the entire set of activity probabilities (which they term 'beliefs') to get the cumulative belief score as follows:

$$\textit{cumulative belief} = 1 - \prod_{i=1}^{N} (1 - P_i) \tag{6}$$

where P_i is the belief from the i^{th} technique and N is the total number of techniques used. The authors argue that the strength of the method is in its capability to incorporate new scoring functions. The orthogonality of the data sources that they integrate is a strong advantage. However their assignment of 'beliefs' is dependent on the presence on actives and decoys for a given target, which restricts their method only to targets that are already well-characterized.

[3] Decoys are compounds with no known activity against the target of interest.

Finally, the first fully probabilistic formulation for drug-target interaction network inference is the Bayesian Matrix Factorization method proposed by Gonen [34]. This method entails projecting the drugs and proteins into the same (integrated) subspace, through the use of chemical and genomic kernels respectively. The study uses the same dataset as that of [29] described above, therefore the genomic and chemical kernels are exactly as described above: chemical fingerprint similarity for the chemical similarity kernel and Smith-Waterman based sequence similarity for the genomic kernel. The low-dimensional projections in the integrated space are then used to compute interaction scores between drugs and targets using a factorization of the interaction matrix. Any given new drug or target can be mapped to the integrated space through the use of the relevant kernel and once it is projected onto the integrated space, its interaction scores can be computed as well. This way, the interaction between a new compound and known targets, a new target and known drugs or a new drug and a new target can be estimated. For automatic complexity control, the probabilistic representation has been applied Bayesian treatment by the introduction of priors and therefore exact and optimal inference of the posterior is very hard. There are two techniques that can be applied: variational approximation or sampling procedures. Gonen opted to adopt variational approximation which entails using a factorized version of the posterior distribution of the probabilistic representation as a lower bound on the marginal likelihood and then optimizing that bound. The author reports better AUC values than those acquired earlier [32]. However an earlier study [33] reported higher AUC values for all four types of targets. Secondly, others have reported that sampling-based inference procedures have advantages over variational approximations for Bayesian matrix factorization [53]. Nevertheless, this work is valuable as a first fully probabilistic formulation of the polypharmacology prediction problem.

The main advantage of integrative approaches is that they can utilize drug and target similarity calculation methods in making predictions. This can be beneficial when making predictions on drugs and targets with no other previously known interactions (either newly synthesized chemicals, or newly characterized genes). Furthermore after learning is completed, making a new prediction can be quite efficient. However the reliance on the similarity calculation methods (chemical or genomic) is also a major disadvantage: chemical or genomic similarity does not necessarily imply interaction similarity. Drugs with different chemical moeities can bind different sites on the same protein, thus sharing the same target. Alternatively two targets with similar sequences can have major differences in the ligand recognition site, thereby having different interaction characteristics despite being highly similar in sequence.

1.1.3 Holistic Approaches

Holistic approaches are distinguished by their being independent from information on individual targets. Their advantage is that they allow for a broader assessment of the activity of a compound and they can be used when there is not enough data for using one of the other approaches. Most of these methods take advantage of the high-throughput screening (HTS) methodologies developed in the last two decades. The significant methods that can be categorized under this umbrella are the following: connectivity map (CMap) [36-38], guilt-by-association (GBA) [39], the Bioactivity Profile Similarity Search (BASS) [40], and PREDICT [41]. A comprehensive review of these methods can be found in [16]. Cancer has been highlighted as being a disease where these holistic methods can play a particularly important role in the development of novel therapies [54]. Finally, multi-scale holistic models that integrate data spanning across multiple levels of biological organizations have been described [55].

The CMap approach is a pioneering work that established the idea of anti-correlating the effect of a perturbagen with the impact of a disease for predicting activity [36]. To measure the phenotypic response to different perturbagens and diseases, the authors used a microarray mRNA expression assay and computed the up/down-regulation patterns. The perturbagens that correlate positively mimic the effect of the disease while those that correlate negatively have the potential to restore the normal phenotype. The authors have studied 164 small molecule perturbagens in 4 cell lines (with most of their results in the breast cancer epithelial cell line MCF7). They showed that their method can capture the anti-estrogenic activity of fulvestrant because the response to this perturbagen and that to treatment with estrogen anti-correlate; among other success stories. This work has established the idea behind holistic approaches to polypharmacological predictions.

CMap has been successfully applied to repurpose the anticonvulsant topiramate for inflammatory bowel disease [37] and the antiulcer drug cimetidine as a therapeutic for lung adenocarcinoma (LA) [38]. In these studies, the authors have downloaded gene expression signatures characterizing 100 diseases from the Gene Expression Omnibus [56] and then anti-correlated these signatures with the 164 drug signatures in CMap, as described above. The two images in the lower part of the middle section in Figure 1 are reproduced from [37] and they show the clinical endoscopy of mice that were treated with TNBS to induce inflammatory bowel disease with and without treatment with topiramate. The therapeutic impact of topiramate can be clearly seen. Similarly, the authors of [38] showed that tumors treated with cimetidine shrunk in size. These results serve to illustrate that computational strategies are viable methods for assessing polypharmacology and drug repurposable possibilities. At the very least, these strategies give good starting points at a favorable cost/benefit ratio. The major advantage of

CMap is that it can make clinically relevant predictions without requiring a detailed understanding of the mechanism. However, the major disadvantage is that it requires transcriptomic profiling of the entire chemical library.

The guilt-by-association method was first introduced by Chiang and Butte [39]. Fundamentally this method is based on the idea that when two diseases share a therapy, then the therapies that are known to work for only one of them might also work for the other. With this starting point, the authors investigated 726 diseases and 2,022 drugs for pairs of diseases that share at least one therapeutic using the data in the DRUGDEX system (Thomson Healthcare, Greenwood Village, CO) and the Drug-Disease Knowledge Base (DrDKB). They then predicted that the drugs known to work for one disease but not the other, would work for both diseases. They found that their drug use suggestions were 12 times more likely to be undergoing clinical trials than a random drug-disease pair not within their suggestion set. The main disadvantage of this approach is the high false positive rate.

Predicting drug-disease associations directly has been a direction that the developers of SITAR have also taken with their development of PREDICT [41]. PREDICT compares the drugs using their targets in addition to chemical structure similarity and side effect similarity. The diseases are compared with the text-mining based semantic similarity of disease phenotype information and overlap between human phenotype ontology entries. They used a total of 593 drugs and 313 diseases by merging data from DrugBank [57], KEGG Drug [58], Matador [59], OMIM [60] and UMLS [61] to create the list of drug-disease associations. Each feature consists of one drug-drug comparison method and one disease-disease comparison method. For a given drug-disease pair, the value of each feature is computed using the scoring scheme in SITAR. Then a logistic regression classifier is trained on these features using the known drug-disease associations (from the databases listed above) as training data to classify a given drug-disease association as true or false (this is also highly similar to SITAR). The authors report an AUC performance of 0.9 in 10-fold cross-validation (i.e. 10% of the drugs are hidden, and their associations are predicted using a model trained on the remaining 90%; repeated 10 times each time hiding a different set of drugs). PREDICT compares favorably with the guilt-by-association [39] and CMap [36]. This method stitches together drug-target interactions and target-disease associations to directly make predictions on drug-disease associations. While useful for elucidating more practical predictions, the lack of validation, the lack of mechanistic insight and the use of a small dataset makes it hard to assess the utility of the method.

Cheng *et al.* have developed a new direction, where they use similarity between the bioactivity profiles of compounds to predict unknown targets of known drugs, using a method called bioactivity profile similarity search [40]. Their study is based on the bioactivity data of 4,296 compounds tested in the US National Cancer Institute 60 human tumor cell line anticancer drug screen (NCI-60). For each compound, a bioactivity vector of length 60 is generated, where the i^{th} entry corresponds to the $\log(GI_{50})$[4] value of the compound against the i^{th} cell line. Each drug d_i is compared against every other drug d_j in the dataset by computing the Pearson coefficient between their bioactivity profile vectors. Whenever the similarity between d_i and d_j is 75% or higher, the targets of d_j are assumed to be targeted by d_i as well and *vice versa*. The authors claim that 44.8% of their predictions were verified against publicly available databases. The one criticism of the method is that compounds with more than 75% similarity in their bioactivities are likely to be highly similar in chemistry and the authors do not establish that their

[4] GI_{50}: The concentration required for 50% growth inhibition of tumor cells.

similarity assessments are not easily discoverable through simple chemoinformatics methods (such as Tanimoto scores) that do not require expensive HTS data.

The abundance of data, the increasingly cheaper computational resources and the success of the previously discussed methods have led to increasingly ambitious projects. Bai and Abernethy describe the use of computational data and resources to attempt new therapeutic discovery ranging from the small chemicals and individual biochemical reactions all the way to organism-level responses [55]. They describe a quintipartite (5-compartment) approach for determination of toxicity of drug candidates. They describe the data as being composed of chemicals, proteins, pathways, organs, and phenotypes where the interactions between these parts are in that order: chemical-protein interactions, protein-pathway associations, pathway-organ interactions and finally organ-phenotype mapping. They unify multiple methods that have been used as predictors based on subcategories of this high-level approach and present it as a possible unified approach to predicting toxicity arguing that the integration across scales is going to achieve what individual models cannot.

Finally, Du and Elemento argue that the advent of holistic systems biology approaches present unique opportunities for the advancement of cancer therapeutics [54]. They argue for the use of an integrated approach for cancer that has been recently enabled by the advent of modern technologies, where cancer is probed at the genomic level, protein/post-translational level and tissue level in an iterative and integrated manner is necessary for realizing more effective treatment. They argue that the interplay between the highly person-specific nature of cancer as a disease, the interplay between the tumor microenvironment and the disease, as well as the Darwinian evolution that the cancer cells undergo create unique challenges that can only be overcome by holistic approaches that combine all of these factors together. They argue for the need to develop an approach that involves experimentally characterizing the genome, transcriptome, proteome and the microenvironment whose output are evaluated in a holistic computational model to select optimal treatment strategy as the necessary road for the future.

1.1.4 Target-Centric Approaches

Possibly the most straightforward way of building a target-centric, systems-wide polypharmacology prediction scheme is to dock all drugs to all proteins. Li and colleagues have attempted to do that, by collecting 252 human drug targets, 4,854 small molecule compounds from DrugBank and docking all-to-all [43]. They first identified 13,156 binding pockets in 678 protein drug targets. Then they docked the known drugs for these targets into these binding pockets and evaluated how good the fit was. If their docking software ICM (Molsoft LLC, San Diego, CA) was able to recover the already-known interaction between a drug and its target, then the target was deemed to be ‘reliable-for-docking’ [62]. They identified 252 targets and 2,923 binding pockets to be fit for docking. Then they docked all 4,854 drugs to each pocket and examined the results. They reported that they were able to correctly predict 10 of 14 known interactors of the protein kinase MAPK14, as well as all 4 targets of chemical BIM-8 that were not in the original dataset (DrugBank v1). They also gave a list of 31 interaction predictions that were not in DrugBank v1 but supported by literature. The major drawback of this approach is that it requires protein structure, which is not available for all proteins. Another is the need for extensive computational resources and time if rigorous simulations that take account of the conformational flexibility of the targets are to be carried out.

Another target-centric polypharmacology prediction paradigm is to consider binding pocket similarity. The idea is that when two proteins share similar features in their binding pockets, they will interact with the same ligand. One example of such an approach is the sequence order-independent profile-profile alignment (SOIPPA) method [44]. The idea behind SOIPPA is that the structure and fold might be similar between two proteins, with the same domains in roughly the same three-dimensional arrangement, while their order in the sequence might be different. Since the tertiary structure of the protein is more relevant to the ligand-protein interaction than the primary structure, SOIPPA aims to capture these domain similarities irrespective of sequence properties. This method has later been used to capture binding site similarities and enable proteome-wide polypharmacology screens [42]. The authors first extracted the binding site of a drug from a known structure, then used SOIPPA to screen for other proteins with similar binding sites, and finally performed docking to evaluate the fit. They were able to demonstrate similarity between the binding sites of human catechol-O-methyltransferase (COMT) and the *M. tuberculosis* enoyl-acyl carrier protein reductase (InhA). COMT is targeted by entacapone and tolcapone while InhA is reportedly targeted by isoniazid and ethionamide. The authors postulated then that entacapone would interact with InhA too – which would mean that entacapone could treat multi-drug resistant (MDR) tuberculosis. Their preliminary experiments have shown that Comtan tablets (which contain entacapone as the active ingredient) have slowed the growth of *M. tuberculosis* in culture. The advantage of this method is the mechanistic and rational basis for the predictions. However the requirement of structural data limits applicability to only structurally resolved proteins. It also does not take account of the conformational flexibility of proteins.

Finally, recent developments in sequencing technology have given rise to a new approach called phenome-wide association studies (PheWAS) where the diseases that are of interest for a particular genetic variant are searched in addition to the more-established genome-wide association studies (GWAS) where the genes of importance for a particular disease are searched [63]. The central idea with these methods is to find links between genes and the diseases (or more broadly, phenotypes) of interest and then use the information on known drugs targeting these genes to make new drug repurposing predictions.

1.2 BIOMEDICAL BACKGROUND

1.2.1 α-1 Antitrypsin Deficiency

α1-Antitrypsin (α1-AT) is a member of the serine protease inhibitor superfamily, also called serpins, which regulates the activity of trypsins (in the digestive system) and neutrophil elastase (in the lungs). AT deficiency (ATD, also known as A1AD) is an inherited autosomal co-dominant disorder that causes lung and liver diseases. It affects 1 in 2,000 to 5,000 individuals of Northern European descent [64;65]. It is one of the most common genetic cause of liver disease in children, and causes cirrhosis and hepatic fibrosis and carcinoma in adults [66;67]. Furthermore the aggregation phenotype in ATD has been recognized as a model for conformational diseases, including many common neurodegenerative diseases such as Alzheimer's disease [68]

The primary cause for ATD is the E342K mutation in the *SERPINA1* gene that encodes AT, which causes the production of the aggregation-prone Z variant of AT, called ATZ, that accumulates in the endoplasmic reticulum (ER) of the liver cells. AT/SERPINA1 is the prototypical member of the serpin superfamily and a major anti-protease in the circulation and extracellular fluids [69]. The function of AT is to protect tissues from collateral damage by neutralizing leukocyte-derived peptidases [70;71]. A structural depiction of the work of the serpins, as reported in the PDB [72], is shown in Figure 2. On the *left*, the serpin is shown immediately after its interaction with the trypsin molecule, with the serpin shown in blue and the proteinase shown in green (PDB:1K9O) [73]. Upon cleavage by trypsin, the serpin's recruiting arm quickly undergoes a structural reorganization, embedding the recruiting arm in a sheet of β-strands; which is being shown in the structure on the *right* [74]. The structural reorganization prevents the trypsin from completing its reaction and releasing itself, thereby trapping the protein in a mouse-trap fashion.

Hepatocytes are the major biosynthetic source of AT, where the protein normally enters the constitutive secretory pathway [75]. However, the Z-mutation delays native folding and impairs secretion, which leads to polymerization and aggregation of ATZ by a domain swapping mechanism [76]. Consequently, ATZ is retained within the endoplasmic reticulum (ER) as large inclusions that cause fibrosis/cirrhosis and hepatocellular carcinoma [77-79]. In ATD patients, therefore, a loss of serpin inhibitory activity underlies the lung disease, whereas a gain-of-toxic-function triggers liver disease.

ATZ aggregation induces a reduction in circulating AT, and predisposes adults to developing emphysema and chronic obstructive pulmonary disease [80-82] because of the lack of the proteinase inhibitory function in the lungs. In addition, ATD patients homozygous for the most common mutation, Z (E342K), are at increased risk of developing liver disease throughout their lifetime due to the ATZ aggregation in the hepatocytes [66;67]. Simply stated, ATZ leads to two major disease phenotypes (i) the gain-of-toxic-function due to ATZ aggregation causes liver damage; (ii) the loss-of-function due to reduced secretion of AT from the liver leads to lung diseases. The marked accumulation of mutant ATZ has been demonstrated in the PiZ transgenic mouse to lead to liver damage, closely resembling that in human disease [83;84]. As known from earlier studies, only a subpopulation of ATD patients develop liver disease [85], suggesting that genetic and/or environmental modifiers determine the susceptibility of an ATD individual to liver disease [83].

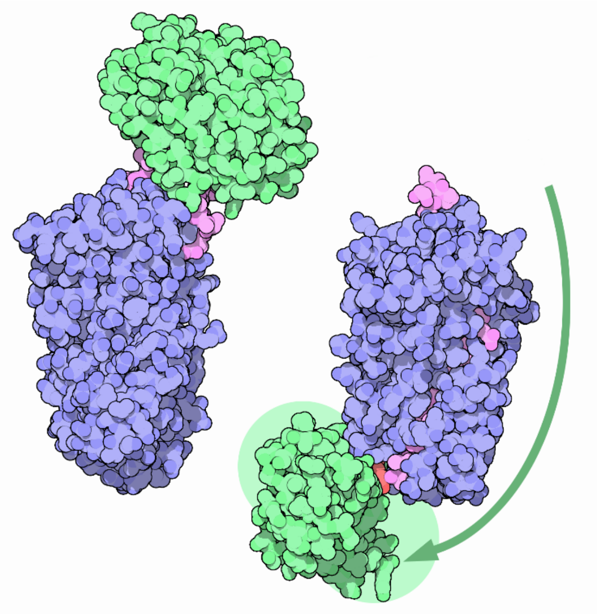


Figure 2: The 'mouse-trap' mechanism of serpins

Serpins operate as proteinase inhibitors by recruiting and trapping the proteinases as reported and shown in the figure above adopted from the PDB [72]. The structure on the left (PDB:1K9O) [73] shows the serpin-trypsin complex (serpin shown in purple, proteinase shown in green) immediately after binding; whereas the structure on the right (PDB:1EZX) [74] shows the trypsin after the serpin has stabilized, inactivating the serpin. Upon cleavage, the serpin undergoes a structural reorganization, embedding the recruiting arm in a sheet of β-strands, with this change preventing the proteinase from dissociation thus trapping the proteinase.

1.2.2 Huntington's Disease

Huntington's disease (HD) is an autosomal dominant genetic neurodegenerative disease, caused by an expanded CAG repeat in the huntingtin gene, that affects 4-10 out of 100,000 people in the western world with many others at the risk of disease [86]. Higher than 40 CAG repeats cause nearly full penetrance at about 65 years of age, while the average onset of disease is at the age of 40 [87]. Disease onset usually occurs during the fourth or fifth decade in life and mean survival of onset being 15 to 20 years after onset; furthermore the disease is universally fatal, and despite best efforts, there is currently no known cure for HD [88]. The clinical phenotypes the disease presents involve characteristic movement disorder (Huntington's chorea), cognitive disorders, and psychiatric symptoms. The etiology of the disease is described as selective regional neuron loss and gliosis in striatum, cerebral cortex, thalamus, subthalamus and hippocampus [89]. Owing to the discovery of the causal mutation of the disease, transgenic mouse models of the disease have been made possible [90]. In these mice models of disease, selective regional neuronal loss accompanying motor symptoms has been demonstrated as observed in the human disease [89].

The Friedlander lab has screened the library of the Neurodegeneration Drug Screening Consortium [91] in isolated mitochondria for cytochrome *c* release inhibition, and tested the hits resulting from this first screening for their neuroprotective activities in ST14A cell lines [92]. These were immortalized striatal cells stably expressing a mutant huntingtin fragment to serve as a model of HD [92]. In total they have identified 21 drugs that inhibit cytochrome *c* release, 15 of which subsequently demonstrated neuronal cell death inhibition activity in ST14A HD model cell lines serum deprivation and heat insult assays. Among them methazolamide also showed a dose-dependent delay in HD progression *in vivo*, in a mice model of HD (specifically R6/2) [90].

1.3 SCOPE OF CONTRIBUTION

Most computational methods for predicting drug-target interactions rely on similarity. However, there are multiple shortcomings with basing interaction inferences mostly on chemical and/or genomic similarity; primarily that global similarity is not always a good predictor of specific binding behaviour. There can be proteins with highly similar sequence (and even structure) but with a very small, varible ligand-binding region (such as membrane-bound receptors) that give rise to critically different interaction patterns. Since small-molecule compounds used as drugs are usually much smaller, the converse is harder yet there are cases where minor modifications can lead to widely different physiological phenotypic differences. A good example can be found in steroidogenesis in humans: Testesterone and Estradiol have 74% chemical similarity based on the MACCS fingerprints, calculated using Pybel [93], despite having radically different phenotypic effects. The contribution presented in this paper is completely independent of any chemical/protein similarity methods and relies on the interaction network therefore bringing a novel and complementary approach that avoids the pitfalls of other methods relying on similarity.

I have demonstrated that a latent factor based drug-target interaction prediction method has successful descriptive and predictive power. I have validated the predictive characteristics with many different cross-validation setups. I have also tested the descriptive characteristics by comparing the drug-drug similarities calculated by the latent variables to those calculated by 3D chemical similarities. Finally, I have shown that such a method can perform remarkably well in directing experimentation in an active learning setting.

I performed both computational and experimental studies towards elucidating the mechanism of action of these drugs and designing new, more potent inhibitors or HD. Specifically, I helped develop a neuronal cell death inhibition assay using the Q7/Q111 striatal neurons derived from murine cell lines which respectively express 7- and the 111-CAG-repeat human huntingtin protein. I characterized the apoptotic response under heat and serum deprivation induced stress conditions, and helped develop a high content screening (HCS) based workflow for assessing the level of neuroprotection through neuronal cell death inhibition in response to chemical intervention. Computationally, I have used the descriptive function of latent variables within the context of HD in order to discover other drugs that can work effectively. I developed a method for analyzing the drugs that were observed by the Friedlander lab to be preventing cytochrome *c* release from the mitochondria and/or to be neuroprotective, and identified other drugs that could potentially be helpful in this disease. This work has given rise to the discovery of a novel repurposable candidate sodium nitroprusside (SNP). SNP is traditionally used as an antihypertensive owing to the fact that it breaks down in circulation and releases nitric oxide (NO), which results in vascular smooth muscle relaxation and vessel dilation. SNP has been experimentally shown to be an effective inhibitor of neuronal cell death in the Q111 HD model cell line, initially in the experiments done in the Friedlander lab by Hossein Mousavi. This phenotype was later reproduced in the University of Pittsburgh Drug Discovery Institute using the assays developed under the guidance of Lans Taylor, Andrew Stern, Mark Schurdak and with the work of Celeste Reese, Laura Vollmer and myself.

I have also analyzed the whole-genome RNAi knockdown data in a *C. elegans* model of ATD to identify the genes that significantly impact disease progression, matched those nematode genes to the druggable human genome, and identified the best candidate drug for modulating the disease, glibenclamide (traditionally used as an antidiabetic), as a potential repurposable drug against ATD. Building on this central idea, we identified a set of 104 known proteostasis network (PN) modifier genes, and mapped them onto their human orthologs using two different databases/compendia available for *C. elegans* genes: Wormbase and Ortholist [94;95]. We mapped the human orthologs to interacting drugs, and filtered for targets of drugs that occur in the Library of Pharmaceutically Active Compounds (LOPAC™) for feasibility. There were four such targets: Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit alpha isoform (PIK3CA), Transthyretin (TTR), ATP-binding cassette (ABC), and Nociceptin receptor (OPRL1) and we tested two drugs for each of these targets identifying four that were shown to reduce ATZ aggregation.

Furthermore there is a need for user-friendly tools that an experimental scientist could use to rapidly search for known and predicted protein/targets using as query a given drug or a target of interest. These tools need to be easy-to-use, accessible, efficient, yet highly robust and low in false positives in order to help build reasonable hypotheses for further experimentation. This is important because the experimental scientists are an important audience, if not the key audience, to which these methods are addressed to. Yet they cannot be expected to possess the technical expertise required to develop and implement algorithms, neither download or implement existing tools, and then to run the code simply to get one prediction of interest. I contributed a new web server, BalestraWeb, to facilitate the broad dissemination and usage of the PMF-based computational prediction tools developed within the scope of this doctoral studies where the execution of the complicated machine learning is abstracted from the user who simply enters the query of interest (drug and/or target) and clicks one ‘Predict’ button. Finally, my work on laying the foundations of BalestraTK can help other scientists conduct research easier by allowing them to easily integrate multiple datasources.

1.4 SPECIFIC AIMS

Below is a summary of the specific aims proposed to be accomplished during the course of my doctoral research studies.

Specific Aim 1: Predicting drug-target interactions using probabilistic latent factor models and validating their use as descriptors of therapeutic effects. The drug-target interaction network can be used to learn probabilistic latent factor models (LFM) about drugs and targets. These latent factor models can be used as (i) descriptors of drugs/targets for therapeutic function similarity comparison, clustering, distance calculation purposes; and (ii) predictors of drug-target interaction likelihood.

Sub Aim 1: Latent variables as descriptors. We will demonstrate the use of LFM as descriptors of drug-target interactions by showing that the LVs can capture therapeutic functional similarity between compounds in cases missed by state-of-the-art similarity based comparison.

Sub Aim 2: Latent variables as predictors. We will validate the use of LFM as predictors by comparing them against state-of-the-art methods on benchmark datasets, in addition to an active learning setting where LFM directs interaction experimentation *in silico*.

Specific Aim 2: Identification of repurposable candidates for α-1 antitrypsin deficiency and Huntington's disease. New drugs that can be repurposed against ATD will be identified using the experimental high content screening data collected on *C. elegans* model of the disease. Furthermore I will diversify previously identified hits against Huntington's disease to identify more effective neuroprotectives using latent factor modeling based methods.

Sub Aim 1: Predict repurposable candidates for A1AD. I will identify potential targets in humans using the genome-wide RNAi screen, and a chemical library screen performed on a *C.elegans* model of ATZ aggregation. The genes that significantly alter ATZ accumulation will be mapped to their human orthologs. The drugs interacting with the human targets will be reported, for experimental verification. The data from an additional chemical screen, Prestwick library [96], for their ATZ elimination activities will be analyzed to identify potential targets, as well as the common chemical patterns that led to anti-aggregation activity, toward identifying new repurposable candidates.

Sub Aim 2: Describe mechanism of action of neuroprotective drugs. Drugs that share one common target with neuroprotective drugs but otherwise have as diverse a target profile as possible will be identified, where diversity is defined as distance within the latent variable space, These drugs will then be tested in a neuroprotection assay, whose development I will assist. The hypothesis therein is that other, more effective drugs can be identified by exploiting the information we have about our currently known active drugs.

Specific Aim 3: Development of new tools to integrate existing data sources and enable fast, efficient prediction of drug-target interactions to expedite drug discovery process. The algorithms, software and tools developed during the course of the doctoral studies will be made available to the larger community of biomedical researchers with the help of user-friendly interfaces.

Sub Aim 1: Website for drug-target interaction prediction. We will build BalestraWeb, (http://balestra.csb.pitt.edu/) a website for latent factor based interaction predictions. The user will be able to acquire predictions for (i) a specific drug-target pair; (ii) the most likely interaction partners of a drug, (iii) the drugs most likely to interact with a given target.

Sub Aim 2: Development of a toolkit, BalestraTK, for chemical, protein and interaction data integration and analysis. We will develop a Python toolkit (https://github.com/mcc-/balestraTK) for interaction information access and prediction. The users of the toolkit will be able to integrate and easily access data stored in the public databases DrugBank, STITCH, UniProt, and PubChem.

1.5 SUMMARY OF FINDINGS

Within the scope of aim 1, subaim 1, I have shown that latent factor based models can accurately describe the interaction profiles of drugs by assessing the similarity of the drugs with known similar therapeutic functions in latent variable space. I have also compared the similarity of these molecules in latent variable space to the results acquired by using state-of-the-art chemoinformatic 3D chemical similarity methods and shown that latent variable methods actually discover therapeutic function similarity better. For aim 1, subaim 2, I have validated the use of LFM based methods for predictive drug-target interaction assessment in two ways: (i) I assessed the recapitulation rate of known interactions after randomly removing 70% of the interactions, (ii) *de novo* prediction performance by assessing the presence of direct and indirect evidence for predictions made after including all the available data [97].

Within the scope of aim 2, subaim 1, I have analyzed the whole-genome RNAi knockdown data from the Perlmutter lab's *C. elegans* model of ATD. Based on these data, I devised a logistic regression based classifier to distinguish the genes that reduce aggregation, mapped these genes to human drug-target orthologs using WormBase and DrugBank and thus identified human drugs. Selecting for maximal match performance at every step, I predicted that the antidiabetic glibenclamide would be effective in A1AD, and this prediction has subsequently been validated by the Perlmutter lab *in vivo* in a murine model of the disease. The same central idea has been applied using more data sources: OrthoList in addition to WormBase; STITCH and MetaCore instead of DrugBank [45;94;95]. This approach has yielded eight other predictions that were tested, four of which have turned out to be active [65]. Finally, I have analyzed the data from the chemical screening and identified three new predictions. One of these has been validated by our collaborators so far, with the other two to be tested. Within the scope of aim 2, subaim 2, I have analyzed the results of a two-stage screen previously conducted by the Friedlander lab [92]. After implementing a novel computational method for computationally selecting new compounds to test, I have identified a set of experimentally feasible compounds to test. Then I have participated in the experiments to build a neuroprotection assay based on the Q7/Q111 HD model cell lines. Our experiments suggest that the antihypertensive SNP is a promising potential repurposable candidate. SNP has shown statistically significantly more neuroprotective activity than the glaucoma medicine methazolamide, which had been shown to be highly neuroprotective in HD *in vitro* and *in vivo* thus being used by the Friedlander lab as a positive control.

Within the scope of aim 3, subaim 1, I have built BalestraWeb [98] which is a publicly accessible website that allows any user to be able to access predictions made by our methods and thus allows us to facilitate the dissemination of the results of our work to a wider audience. Making a prediction is simplified to the point that the user only needs to input the name of the drug and/or target of interest, and click the ‘Predict’ button. The website automatically retrieves the proper prediction made by our latent factor based method and then visualizes the prediction(s) within the context of known interactions; as well as providing the user with links to more information about any drug/target that is either predicted to interact or known to interact. Within the scope of aim 3, subaim 2, I have built BalestraTK, which is open-source and publicly available (https://github.com/mcc-/balestraTK). This toolkit allows the users to easily parse, analyze and integrate publicly available datasets for use in computational systems pharmacology projects.

2.0 LATENT FACTOR MODELING BASED ANALYSIS OF DRUG TARGET INTERACTIONS

In this section, I will describe the various methods used or developed during the course of my doctoral studies. First, I describe the various latent factor based probabilistic modeling techniques that I have adapted to QSP studies. Secondly, I describe the computational techniques we used and implemented for data analysis and new predictions within the scope of the ATD project. Third, I present the algorithms developed for new compound identification within the scope of the HD project. Finally, I describe the methodology behind BalestraWeb and BalestraTK.

2.1 METHODOLOGY

2.1.1 Problem Definition

The drug-target interaction network is a bipartite graph with two types of nodes; drugs, and targets (Figure 4**)**. Each edge represents an interaction between a drug and a target. The drug-target interaction identification problem is to determine the missing edges that are likely to exist given all nodes and some of the edges in the network.

2.1.2 Dataset

We used DrugBank (version of September 20, 2011) as the database [99]. All drugs annotated therein as approved, along with their annotated targets, are included in our dataset (i.e., we excluded compounds annotated as withdrawn or nutraceuticals), resulting in N = 1,413 drugs and M = 1,050 targets with 4,731 interactions among them. The interaction network displays small-world characteristics: many nodes have low degree and a few, very high degree, as illustrated in the panels b and c of Figure 4, in line with previous studies on drug-target networks [100]. On average, there are 3.35 interactions per drug, and 4.50 interactions per target.

2.1.3 Probabilistic Matrix Factorization (PMF)

PMF is a member of the LFM subtype of collaborative filtering family of machine learning algorithms that decomposes the connectivity matrix, $\mathbf{R}_{N \text{ x } M}$, of a bipartite graph of N drugs and M targets as a product of two matrices of latent variables (LVs) [53;101]. $\mathbf{R}_{N \text{ x } M}$ is defined as:

$$R_{ij} = \begin{cases} 1 \text{ if drug } i \text{ interacts with target } j \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

The matrix $\mathbf{R}_{N \text{ x } M}$ is modeled as the product of two matrices $\mathbf{U}^{\mathrm{T}}{}_{N \text{ x } D}$ and $\mathbf{V}_{D \text{ x } M}$, that express each drug/target in terms of D LVs. Our objective is to find the best approximation for LVs, while avoiding over-fitting. The predicted connectivity matrix $\hat{\mathbf{R}}_{N \text{ x } M}$ is then expressed as:

$$\hat{\mathbf{R}}_{N\times M} = \mathbf{U}^{T}_{N\times D}\mathbf{V}_{D\times M} \tag{8}$$

where $\mathbf{U}^{\boldsymbol{T}}$ and $\mathbf{V}$ are composed of N rows $\mathbf{u}_i^{\mathrm{T}}$ and M columns $\mathbf{v}_j$, respectively, each being D-dimensional. The PMF adopts a probabilistic linear model with Gaussian noise to model the interaction. Therefore, the conditional probability over observed interactions is represented as

$$p(\mathbf{R} \mid \mathbf{U}, \mathbf{V}, \sigma^{2}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left[f(R_{ij} \mid \mathbf{u}_i^T \mathbf{v}_j, \sigma^{2}) \right]^{I_{ij}} \qquad (9)$$

where $f((x \mid \mu, \sigma)$ is the Gaussianly distributed probability density function for x, with mean μ and variance σ, and I_{ij} is the indicator function equal to 1 if the entry R_{ij} is known, and 0 otherwise. Therefore, $p(\mathbf{R} \mid \mathbf{U}, \mathbf{V}, \sigma^2)$ gives us a probabilistic representation of the connectivity matrix, $\mathbf{R}$ [101]. Using zero-mean, spherical Gaussian priors on LVs, we can write

$$p(\mathbf{U} \mid \sigma_U^2) = \prod_{i=1}^{N} f(\mathbf{u}_i \mid 0, \sigma_U^2 \mathbf{I}) \qquad (10)$$

and

$$p(\mathbf{V} \mid \sigma_V^2) = \prod_{j=1}^{M} f(\mathbf{v}_j \mid 0, \sigma_V^2 \mathbf{I}) \qquad (11)$$

which leads to the log-likelihood of $\mathbf{U}$ and $\mathbf{V}$ given by

$$\ln(p(\mathbf{U}, \mathbf{V} \mid \mathbf{R}, \sigma^2, \sigma_U^2, \sigma_V^2)) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij} (R_{ij} - \mathbf{u}_i^T \mathbf{v}_j)^2 - \frac{1}{2\sigma_U^2} \sum_{i=1}^{N} \mathbf{u}_i^T \mathbf{u}_i - \frac{1}{2\sigma_V^2} \sum_{j=1}^{M} \mathbf{v}_j^T \mathbf{v}_j + C \qquad (12)$$

where C is a term that does not depend on LVs; the first term on the right-hand side is the squared error function to be minimized; and the two summations over the square magnitudes of $\mathbf{u}_i$ and $\mathbf{v}_j$ are regularization terms that favor simpler solutions and penalize overfitting. The above log-likelihood directly follows from the Bayes' rule where $\mathbf{R}$ stands for data, and $\mathbf{U}$ and $\mathbf{V}$ represent the model. If we assume that the variance of the prior for the drugs and targets are equal, i.e. $\sigma_U^2 = \sigma_U^2 = \sigma^2$, and if we define $\lambda = 1/\sigma^2$ then we can write this optimization function with a single hyper-parameter, λ. Furthermore, maximizing this log-likelihood can be shown to be equivalent to minimizing a squared-loss error function, regularized with the Frobenius norm of the latent variable vectors, as shown in the following equation:

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij} - \mathbf{u}_i^T \mathbf{v}_j)^2 + \frac{\lambda}{2}\left(\sum_{i=1}^{N}\mathbf{u}_i^T\mathbf{u}_i + \sum_{j=1}^{M}\mathbf{v}_j^T\mathbf{v}_j\right) \qquad (13)$$

To learn an optimal model means to find the **U** and **V** matrices, or the *D*-dimensional LV vectors, $\mathbf{u}_i$ $(1 \leq i \leq N)$ and $\mathbf{v}_j$ $(1 \leq j \leq M)$, that maximize the log-likelihood function, together with the identification of the hyper-parameters (λ and D) that yield the optimal learning performance. To identify the best hyper-parameters, we use cross-validation. Specifically, we hid 70% of the data and looked at performance in recapturing the missing interactions. Figure 13 demonstrates the results with D = 30 and D = 50 when predicting 100 and 1000 interactions, using the active learner as well as the passive learner (discussed in more detail below, in the next subsection). We chose D=50 and λ=0.01 based on our cross-validation runs. The PMF code distributed by the authors of [53] can be freely downloaded at http://www.utstat.toronto.edu/~rsalakhu/BPMF.html whereas our code that is partially built on top of that can be freely accessed at http://www.csb.pitt.edu/Faculty/bahar/QSP_PMF_code/ along with all the data we used to run these analyses. A detailed description of the contents of the two files provided in the link to our files is provided in Appendix A.

The PMF method yields the optimal $\mathbf{u}_i$ and $\mathbf{v}_j$ vectors corresponding to each drug, d_i, and each target t_j, respectively. The basic idea is that the model is forced towards making a 'no-interaction' prediction by the regularization – i.e. there is a penalty associated with any non-zero value in the LV matrices. However, there is also a penalty for failing to capture known interactions– i.e., if the dot product of the LV vectors corresponding to an interacting drug-target pair is close to zero. Therefore learning a model means to optimally balance out two objectives:

developing a sufficiently complex model to describe the known interactions, but not overly complex to end up in over-fitting. In this study, we use gradient descent for minimizing the objective (error) function. The adoption of higher D values usually yields more accurate results, although beyond a certain limit the increase in complexity and decrease in efficiency may not warrant the marginal improvement, if any, in prediction accuracy. D = 50 is adopted here as an optimal dimensionality for prediction runs. The method is highly efficient: a 50-dimensional model is trained on the entire DrugBank in approximately 2 seconds using a 2.00 GHz AMD Opteron processor. Moreover, the computing time to learn a PMF model scales linearly with the number of interactions, and as such, the method can be advantageously used for much larger datasets.

2.1.4 Methodology for Active Learning On Drug-Target Interactions Using PMF

The active learning (AL) strategy adopted in the present study is, in part, motivated by the success reported by Warmuth *et al.*[102] who demonstrated that hit maximization is a viable AL strategy applicable to predicting drug-target interactions. The AL strategy adopted here also prioritizes the discovery of unknown interactions. Our method differs in that we aim at capturing the interactions between all drugs and targets, as opposed to predicting activity against a single target.

The procedure is the following: We begin with the set of N drugs and M targets, and known associations, schematically shown in Figure 3 by *black* connectors. The purpose is to identify new associations, indicated by *red* connectors. For each candidate interaction, say the possible interaction between d_i and t_j, we compute the model's estimate, by calculating the dot product $\mathbf{u}_i{}^T\mathbf{v}_j$ which serves as a weight ω_{ij} for the edge/connector between d_i and t_j. Clearly, ω_{ij}, or the likelihood of association between d_i and t_j, is high when $\mathbf{u}_i$ and $\mathbf{v}_j$ have both large values of the same sign at the same dimension(s). For example, a relatively large weight may originate from the 2nd component of both $\mathbf{u}_i$ and $\mathbf{v}_j$, which means that the predicted association is mainly due to latent variable 2. We evaluated the statistical weights $\omega_{ij}(d_i, t_j)$ for the N x M pairs of drug-targets for two purposes: (i) benchmarking the methodology via an iterative AL scheme, and (ii) making *de novo* predictions. In the former case, the method is benchmarked by hiding 70% of known interactions and examining whether the top-ranking prediction is a 'hit', i.e., whether it corresponds to a known (but hidden) interaction. The outcome from this test is fed back to the model, to repeat the calculation for the next prediction. Therefore, the AL model is updated at each iteration using the newly acquired 'hit' or 'miss' data until a predetermined number (m) of predictions are made. The passive learner (PL) makes the m predictions simultaneously without updating its model.

In the case of *de novo* predictions, all DrugBank data were used as input. *De novo* predictions also lend themselves to an AL scheme provided that the top-ranking prediction is experimentally tested and then the new hit or miss data are incorporated in the model to perform a new prediction, and so on, until the experimentation budget is exhausted.

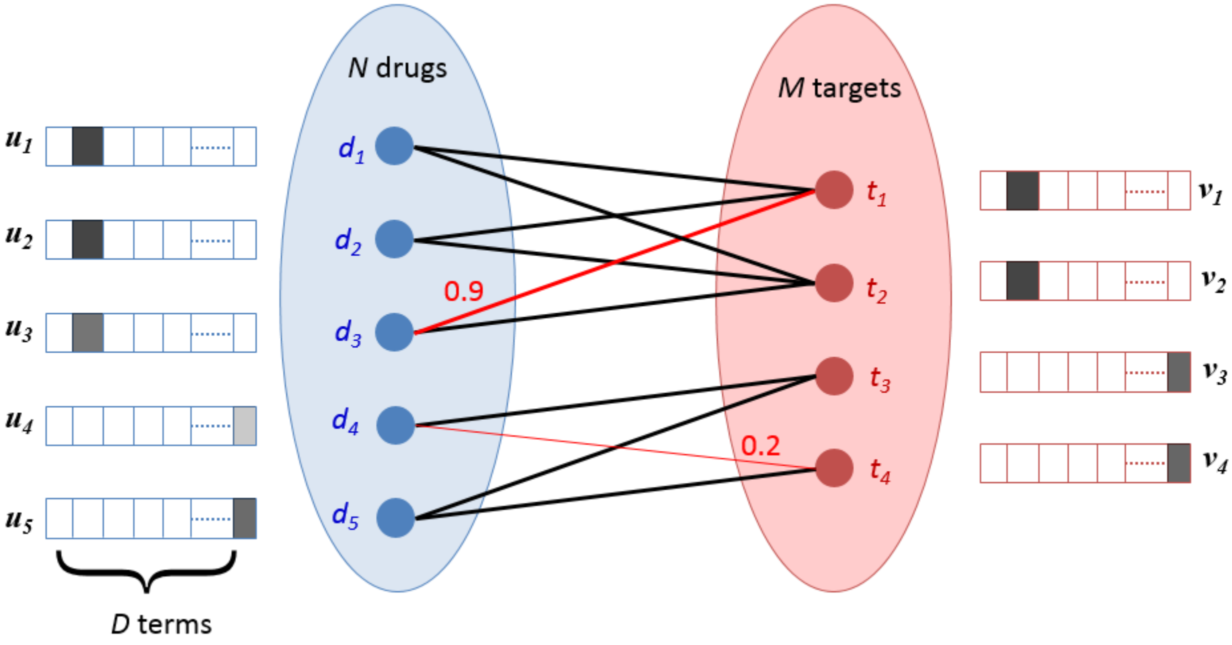


Figure 3: Qualitative illustration of the method for identifying drug-target interactions.

The known interactions between drugs and targets (indicated by the *black* lines) are used to learn the LV vectors (shown adjacent to each node) that describe each drug and target. The dot product $u_i{}^T v_j$ of the LVs for each pair of drug d_i and target t_j define the predicted statistical weight ω_{ij} of corresponding connection. Example predictions are shown in *red*. (From [97])

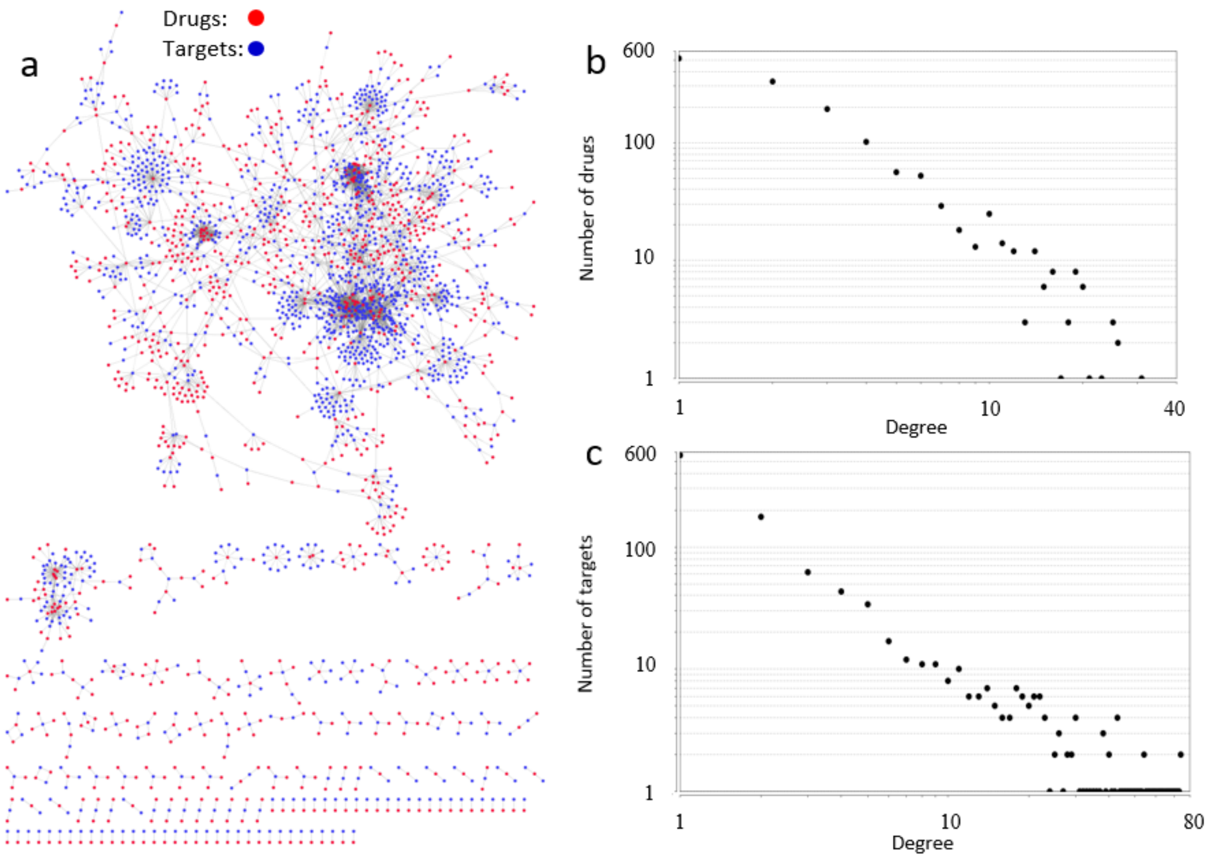


Figure 4: Drug-target interaction network

(a) Network representation of the drug target interaction dataset used in this study. The drugs are shown in *blue*, protein targets, in *red*. Data retrieved from DrugBank [99], Cytoscape used for visualization [103] (b) Distribution of drugs with respect to the number of targets they interact with (i.e. as a function of the number of edges around each drug node). (c) The distribution of targets with respect to the number of distinct drugs they interact with. (From [97])

2.2 RESULTS

Below I am going to discuss the results that demonstrate the power of LFM in descriptive function, predictive function, and in active learning.

2.2.1 Descriptive Power of LFM

To assess whether the LVs provide us with a pharmacologically meaningful metric, we examined the clustering of drugs in the *D*-dimensional space of the latent vectors. The number of clusters was chosen to be 30 as that was the value that gave the lowest Akaike information criterion [104], and using as basis the drug-drug distance $L_1(d_i, d_j) = \boldsymbol{\Sigma}_k |u_{ik} - u_{jk}|$ where u_{ik} designates the k^{th} component of $\mathbf{u}_i$, and the summation is performed over *D* components.

Inasmuch as our method evaluates drugs based on their interaction profiles with targets, which in turn refer to specific therapeutic or phenotypic actions, the similarity of a pair of drugs should be high when their therapeutic effects are comparable and *vice versa*. Thus, the method will tend to cluster drugs that exhibit similar patterns of interactions (with target proteins), which we term as *functionally similar drugs*. The heatmap in Figure 5a displays the resulting organization of drugs in 30 clusters (indicated by different colors and indices along the axes). Table 1 lists the dominant therapeutic action associated with each cluster. The dark regions on the map indicate high functional similarity. The dark blocks along the diagonal show that most clusters include highly similar members, except for two (clusters *29* and *30*), which apparently combine the outliers.

Given that (promiscuous) proteins present more than one site for ligand-binding, different functionalities may be modulated by chemical-structurally different drugs, depending on the binding site on the target (e.g. catalysis, substrate recognition, or allosteric signaling). Furthermore, a shared phenotype may arise from the targeting of different proteins along a given pathway. In order to make a better assessment of the properties of drugs grouped in those clusters, we examined their 3D structural similarities. High similarities would suggest that they bind similar epitopes, if not similar (or identical) structural domains or proteins. If, on the contrary, they are structurally dissimilar, this might indicate a different site on the same protein, or a different target on the same pathway, or other indirect effect due to drug-target network connectivity.

The extent of 3D structure similarity between pairs of drugs was computed using a multitude of the OpenEye™ scientific software products as described below (http://www.eyesopen.com/). We chose 3D similarity because it was reported to be a better predictor than 2D methods for off-target interactions, and to perform equally well in on-target interactions [105]. However 3D methods may suffer from more noise due to the conformational flexibility of the small molecule therefore we generated all possible stereoisomers using OpenEye FLIPPER [106], and up to 200 conformers per stereoisomer using OpenEye OMEGA [106] for every drug. All combinations of conformers accessible to the examined pair of drugs were examined using OpenEye Shape [35] toolkit; and the best matching pair was adopted to assign a 3D similarity score. This computationally expensive task led to the heat map presented in panel b of Figure 5. The drugs (along the axes) are ordered as in panel a to enable visual comparison.

The comparison of Figure 5 shows that some clusters of functionally similar drugs (panel a) also exhibit some 3D similarities (panel b), whereas others display little structural similarity. We examined more closely the individual clusters to see if shared therapeutic functions were captured even when 3D similarities were absent. Figure 6 illustrates the results for *cluster 14*. This cluster essentially consists of anti-anxiety drugs, the majority of which are both functionally (panel b) and structurally (panel c) similar. However, the cluster also includes a structurally dissimilar drug, ethchlorvynol (panel a), which shares the same type of phenotypic action (as a sedative) as the majority of the cluster membership (mostly targeting GABA receptors). The present approach thus detects chemically or structurally distinctive drugs that share common activities, which would have been missed by methods based on ligand fingerprint similarities.

Another interesting observation concerns the cross-correlations between different clusters (i.e. the off-diagonal regions of the heat maps). We note for example that cluster *11* also contains a set of sedatives. LVs are able to capture the commonality between the clusters *11* and *14* as may be seen by the strong signal (dark region) at the off-diagonal region enlarged in Figure 6b. The 3D similarity, on the other hand, cannot recognize the functional similarity and potential interference/side effects between these drugs in these two clusters (Figure 6c). Figure 8 illustrates the same behavior for another cluster, whose members are mostly antineoplastic agents, albeit with various 3D structures. The LVs thus provide information on drug groups that potentially share pathways or exhibit similar activity patterns despite their distinct physicochemical properties.

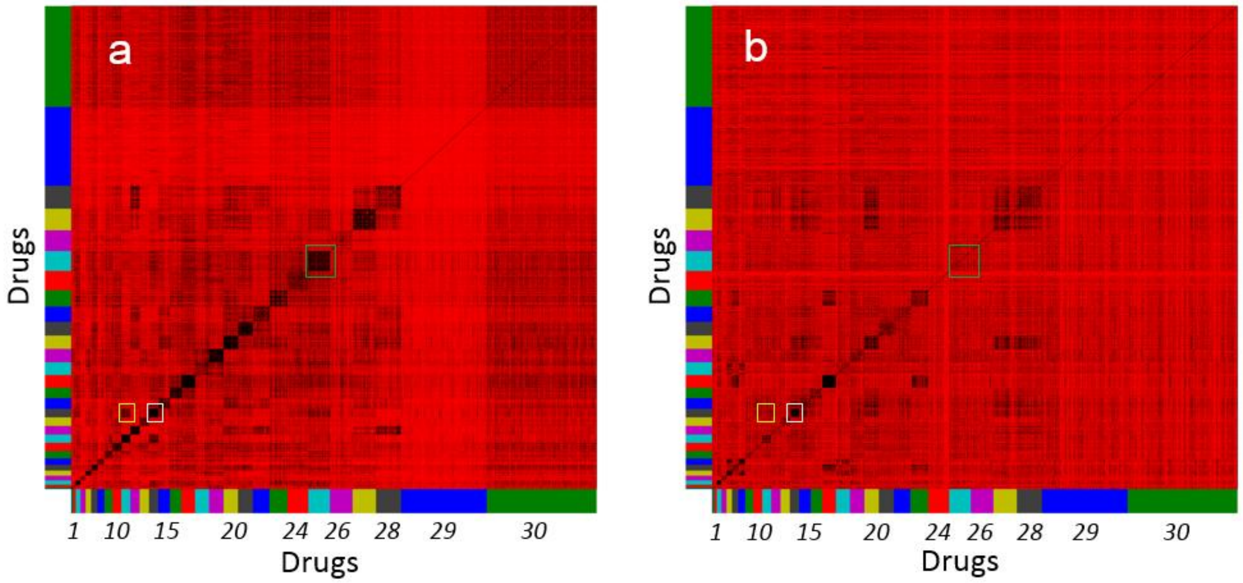


Figure 5: Comparison of pairwise similarities of drugs, based on their therapeutic targets compiled in DrugBank and 3D structure

Panel a displays the 30 clusters of drugs (color-coded along the axes; see Table 1 for their dominant therapeutic indication) deduced from the PMF of 1,413 approved drugs and corresponding 1,050 targets compiled in DrugBank. By definition, drugs belonging to a given cluster share similar interaction patterns with respect to targets. Panel b displays their 3D similarities, with the drugs being ordered as in panel a. *Dark* regions indicate high similarity based on LVs (panel a) or 3D similarities (panel b). Comparison of the panels shows that close proximity in LV space (which indicates functional similarity) does not necessarily imply 3D-structure similarity. LV distances were distributed in the range [0, 1] whereas the 3D distances were distributed in the range [0, 2] ([0-1] from spatial overlap, [0-1] from physicochemical property overlap); with the distribution of values also skewed in different ways. To render the two sets comparable, we performed rank normalization on both the LV similarities and 3D similarities. Selected boxes are enlarged in Figure 6 (*white),* Figure 7 (y*ellow*), and Figure 8 (*green*). (From [97])

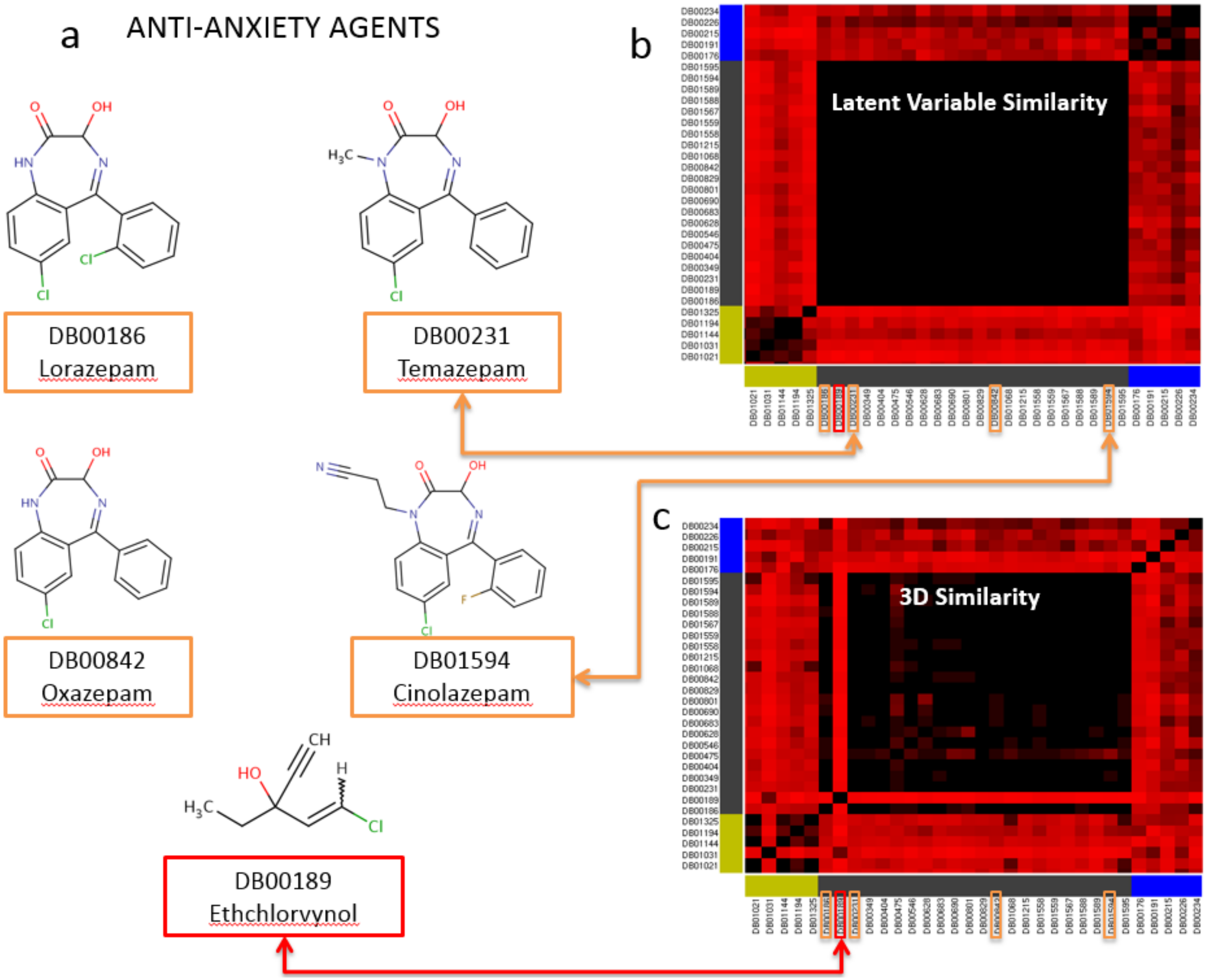


Figure 6: Latent variables can capture therapeutic action similarities when 3D similarity metrics cannot

Closer examination of the similarities between the members of the cluster *14* in Table 1 (enclosed in white boxes in Figure 5, enlarged in panels b and c here) shows that the cluster contains a series of anti-anxiety drugs. A few members of this cluster (indicated by orange boxes along the abscissa of panels b and c) are displayed in panel a, to illustrate their shared structural features, also indicated by the panel c that reflects their 3D similarities. The same cluster however contains ethchlorvynol, also used as a sedative, which would have been missed if we had used exclusively used 3D similarity to identify functionally similar drugs. (From [97])

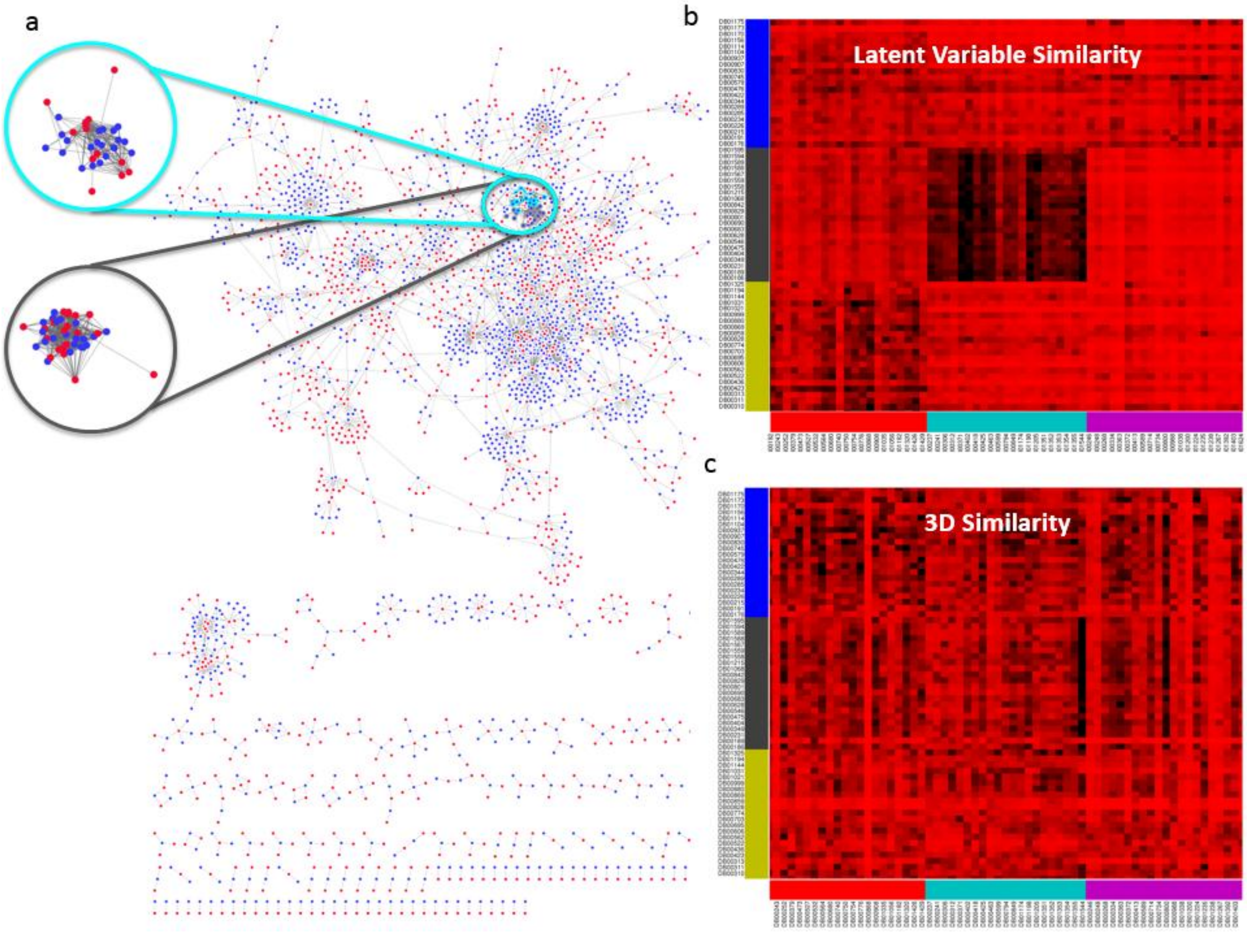


Figure 7: Strong cross-correlations between different clusters of drugs are consistent with their similar therapeutic functions

Cluster 11, color-coded *cyan*, is essentially composed of hypnotics and sedatives. *Cluster 14* (*dark gray*) contains anti-anxiety drugs. The drugs in these two clusters are located very closely on the drug-target interaction network, as shown in panel a, consistent with their similar actions. The LV-derived heat maps capture the functional similarity between these two clusters (as indicated by strong signals, or the *dark* region, in panel b); the maps based on 3D similarity (panel c) do not. In panel a drugs are shown in *blue*, protein targets in *red*. Most drugs and targets are part of a single connected component. Data are retrieved from DrugBank [99]. Cytoscape is used for visualization [103]. (From [97])

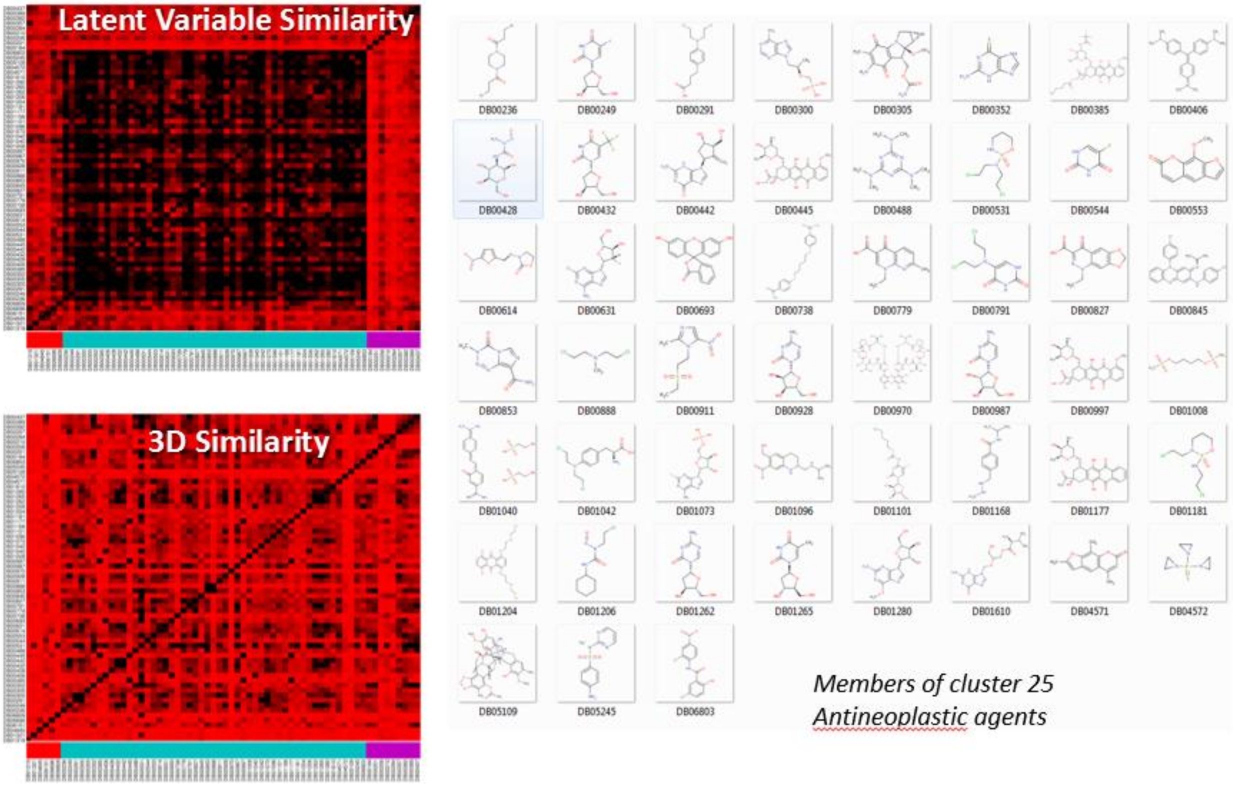


Figure 8: Latent variables capture functional similarity

The figure illustrates how the drugs clustered based on their PMF-derived latent vectors share functional similarity, while their 3-dimensional (3D) structures may vary. The cluster shown in this case is dominated by antineoplastic agents, and they show significant latent variable similarity. The color code in the maps varies from red (no similarity) to black (high similarity). The corresponding 51 drug structures and DrugBank identifiers are presented on the right**.** (From [97])

Table 1: The most dominant therapeutic function in each cluster

Cluster Number	Most Dominant Therapeutic Function	Number of Drugs
1	Antineoplastic Agents	1
2	Gastrointestinal Agents	4
3	Nucleic Acid Synthesis Inhibitors	6
4	Quinolones	11
5	Calcium Channel Blockers	12
6	Anti-Bacterial Agents	13
7	Androgens	15
8	Anti-Bacterial Agents	16
9	Anti-HIV Agents	19
10	Anti-Arrhythmia Agents	21
11	Hypnotics and Sedatives	21
12	Antipsychotic Agents	22
13	Diuretics	22
14	Anti-anxiety Agents	22
15	Adrenergic Uptake Inhibitors	27
16	Analgesics, Opioid	27
17	Anti-inflammatory Agents	32
18	Anti-Bacterial Agents	33
19	Cyclooxygenase Inhibitors	34
20	Antihistamines	34
21	Adrenergic beta-Antagonists	35
22	Sympathomimetics	39
23	Contraceptives, Oral, Synthetic	42
24	Anti-Bacterial Agents	49
25	Antineoplastic Agents	51
26	Cholinesterase Inhibitors	53
27	Muscarinic Antagonists	55
28	Antipsychotic Agents	59
29	Antineoplastic Agents	201
30	Antihypertensive Agents	258

2.2.2 Predictive Power of LFM

To evaluate the performance of the method in comparison to previous work, we considered three important studies in this area, one recently published by Gonen [34] and two by Yamanishi et al. [29;32]. Gonen used a kernel based matrix factorization (KBMF) with chemical and genomic similarities to predict multiple targets. Yamanishi et al., on the other hand, integrated chemical, genomic and pharmacological data to map all drugs and targets to the same unified feature space where each protein-compound pair closer than a predefined threshold was predicted to interact. Our approach differs from both studies, in that PMF assumes an independent LV for each row and column with Gaussian priors; whereas KBMF employs LVs spanning all rows and columns with Gaussian process priors, and Yamanishi et al project drugs and targets into a pharmacological space based on the eigenvalue decomposition of the graph-based similarity matrix.

The benchmarking procedure that we adopted is a five-fold cross-validation of drugs on four target classes: Enzymes, Ion channels, G-protein coupled receptors (GPCRs) and Nuclear Receptors. In order to achieve comparable results, we used the same protocol as that adopted earlier, i.e., we divided our dataset into five subsets, and each was used as a test set, and the others, as training sets. Due to the randomness involved in the selection of subsets, we repeated the cross-validation experiments 100 times with randomly selected subsets and evaluated the average AUC (area under the receiver operating curve) for each subset. The first four rows in Table 2 compare the results (columns *6-10*) for the four classes, and the 5th row lists the average performances weighted by the size of the interaction space. Our method performs best when applied to large datasets (e.g. enzymes and ion channels); whereas Gonen’s performs best in the case of GPCRs, and Yamanishi et al. (2010) exhibits the highest performance for nuclear receptors, where the present method yields a relatively low (0.642) AUC value. Examination of the statistical significance of our results (Figure 9 panel a) indicates that the mean AUC values obtained for all four sets are highly robust. Their variances vary from 2% (Enzymes and Ion Channels) to 11% (Nuclear Receptors). Finally, the application of the same benchmarking protocol to DrugBank yielded an accuracy rate of 79.4 ± 0.01% (Table 2, last row), supporting the utility of the method when applied to large datasets.

In principle, it might be intrinsically harder to make accurate predictions for larger datasets as the size of the potential interaction space *N* x *M* grows quadratically when the number of drugs and targets grow linearly, particularly if the number of known interactions is small. The occupancy of the *N* x *M* interaction matrix is only 1.5% in the Enzyme class, which could make it difficult to learn an informative model. The present PMF technique, however, successfully learned an informative model and handled the complexity of interactions in this space of interactions, apparently due to the availability of a sufficiently large (absolute) number of known interactions (Figure 10 panel b).

The drug class that targets ion channels has the second largest number of known interactions among the four. Although the size of interaction space is one order of magnitude smaller than Enzyme class, there are 776 known interactions leading to a percent occupancy of 5.37% of all possible ion channel-drug associations. The success of our method in this case may be attributed to both the relatively large number of known interactions and the rich annotation of that class of interactions. The two other classes, GPCRs and Nuclear Receptors, are significantly smaller in terms of their interaction space and/or occupancy of that space. Nuclear Receptors comprise only 27 drugs and 22 targets, and 44 interactions. A method that relies solely on connectivity, like ours, cannot presumably formulate an informative model when the set of ‘edges’ to construct the network connectivity matrix is incomplete. In those cases, the data that come from other sources, e.g. chemical similarity and genomic patterns, amend this lack of information. Consequently, methods that incorporate such features [32;34] outperform ours.

To further examine the effect of scarcity of known interactions on the performance of the present method, we performed additional tests by varying the fraction of hidden interactions. The results are presented in Figure 10. Panels a-d show the performance on Ion Channels, Enzymes, GPCRs and Nuclear Receptors, respectively. These results show that the performance depends on the fraction of known interactions. To put the results into perspective, we indicated by a vertical dashed line in each panel the fraction of data (80%) used in previous studies [32;34] for training purposes. Consistent with the above findings, Ion Channels yield the best result: previous AUC values [34] (of 0.799; Table 2) are matched with about only 35% of the data. On the Enzyme group, we match the performance of Yamanishi et al. [32] (AUC of 0.845) with roughly 70% of the data used for training. GPCRs and Nuclear Receptors yield AUC values lower than those previously attained, [32;34] irrespective of the fraction of hidden interactions.

Table 2: Properties of the examined space of proteins-drugs, and performance of the present method in comparison to others

Target type	# of known inter-actions	# of drugs (*N*)	# of targets (*M*)	Size of interaction space (*N*×*M*)	Percent occupancy of the space	Yamanishi (pred pharmacol effects)		Gonen, *2012*[a]	Present method (D = 50)
						2008[a]	*2010*[a]		
Enzymes	1,515	212	478	101,336	1.50%	0.821	0.845	0.832	**0.861 ± 0.02**
Ion Channels	776	99	146	14,454	5.37%	0.692	0.731	0.799	**0.904 ± 0.02**
GPCRs	314	105	84	8,820	3.56%	0.811	0.812	**0.857**	0.771 ± 0.04
Nuclear Receptors	44	27	22	594	7.41%	0.814	**0.830**	0.824	0.650 ± 0.11
All[b]	2,649	443	730	-	-	0.782	0.807	0.825	**0.859** ± 0.03
DrugBank	4,731	1,413	1,050	1,483,650	0.32%	-	-	-	**0.794 ± 0.01**

[a] *The last four columns present the comparison with Yamanishi’s [29;32] and Gonen’s [34] for the same dataset.*

[b] *weighted-average mean and covariances, evaluated using the number of interactions as weights*

In summary, the PMF method is particularly suitable for screening and inferring repurposable drugs or potential side effects from large datasets where computational assessment of structure similarity kernels become prohibitively expensive. In cases where the dataset of known interactions is too small, on the other hand, 2D or 3D similarity metrics provide more accurate assessments.

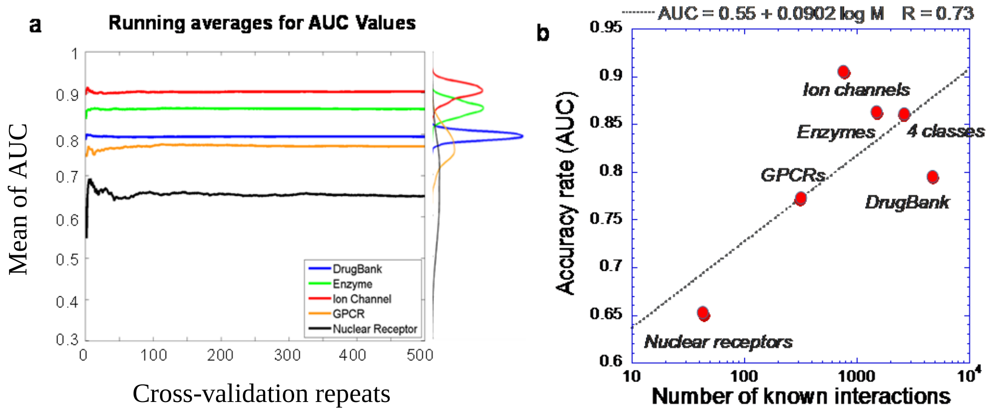


Figure 9: Closer examination of the statistical distribution and robustness of the results obtained for four classes of drug-targets and for the entire DrugBank

Five-fold cross-validation runs were repeated for 500 iterations with different selections of hidden/known subsets to examine the robustness of the results. Panel a displays the running averages obtained for the AUC values as we performed 500 iterations. Results for Enzymes (*green*), Ion Channels (*red*), as well as the entire DrugBank (*blue*) exhibit small fluctuations (see histograms on the *right* ordinate), GPCRs exhibit moderate fluctuations (*orange*); whereas, the nuclear receptors (*black*) show significant variations. The running averages converge after ~ 100 iterations and are robustly maintained to yield values listed in Table 2 (last column). The dependence of the final AUC values on the number of known drug-target interactions in the examined dataset is shown in panel b. A correlation coefficient of 0.73 is obtained, upon logarithmic fitting of the data. When the number of known interactions is sufficiently large (e.g. *N x M* > 500), the occupancy of the full interaction space appears to affect the overall performance. (From [97])

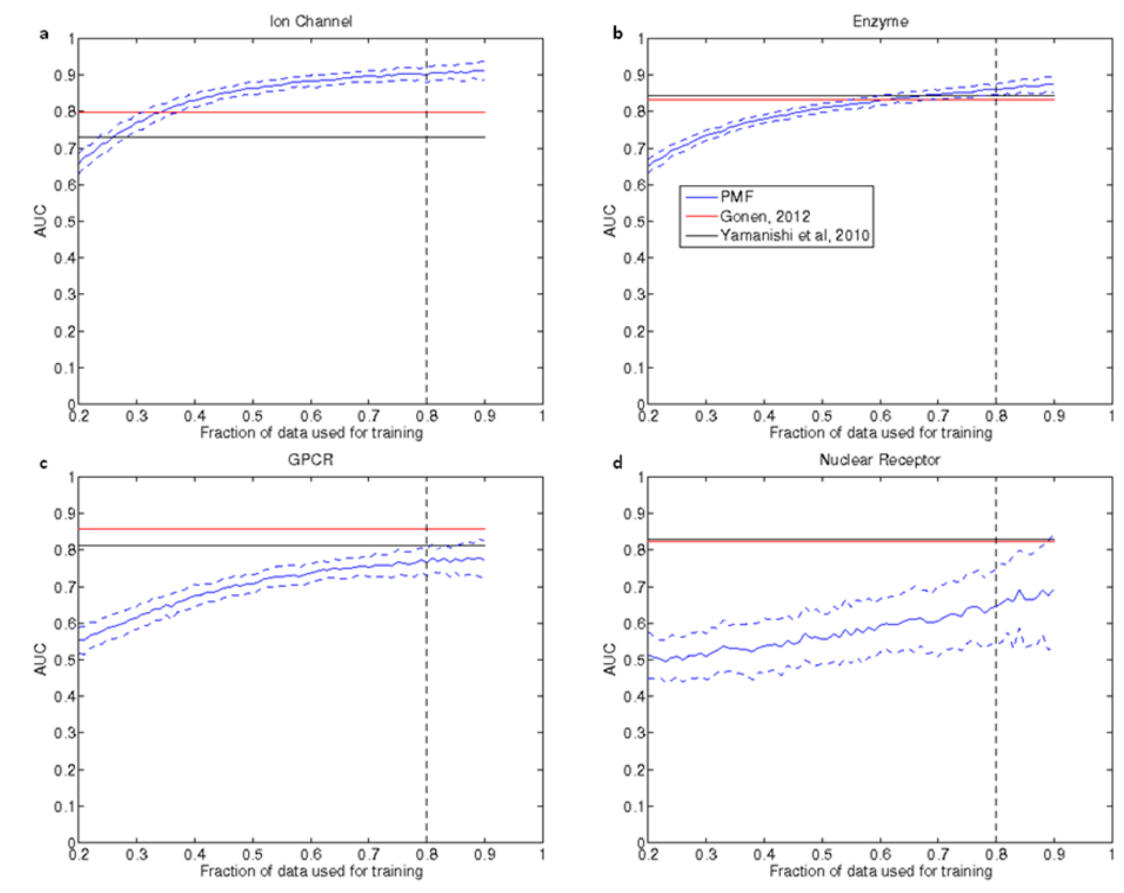


Figure 10: Evaluation of method performance as a function of dataset size on DrugBank v3 data.

We evaluated [97] the method's performance as a function of the training dataset size, displayed for the four subsets of targets listed in Table 2. The fraction of the dataset used for training the model was changed from 20% to 90% and the resulting AUC was recorded. Since there is randomness in assigning data points to the train/test datasets, each step was repeated 100 times. The solid curves show the average and the dotted curves showing one standard deviation above/below the average. The dashed vertical bar indicates the fraction (80%) used for generating the AUC values listed in Table 2. The two horizontal lines indicate the AUC values attained by Gonen (2012) (*red*) and Yamanishi et al (2010) (*black*). (From [97])

2.2.3 LFM Based Predictive Active Learning on Drug-Target Interactions

As a more stringent test, 3,318 (70%) of the known 4,731 interactions in DrugBank v3 were randomly hidden, reducing the average number of interactions per drug from 3.35 to 1. The resulting ‘incomplete’ interaction matrix was then used to predict the hidden interactions, one at a time (rank-ordered by statistical weights $\omega_{ij}(d_i, t_j)$) as described in the methodology section. The outcome was checked in a simulated experiment to assess whether the predicted interaction is a true positive (TP) or a false positive (FP). If the prediction is an existing, but hidden, interaction, the result is considered a TP (or hit), otherwise a FP (or miss). Then the model is updated in line with our AL scheme, and this loop is repeated until the completion of m = 1,000 predictions. At that point, the simulation is halted and the overall performance of the model, or the hit ratio, is evaluated. Note that this method gives us a lower bound for hit ratio because the predictions are labeled as hits only if they are annotated in DrugBank, although they can be true but not yet observed experimentally or annotated in DrugBank.

The results are presented in Figure 11. The figure displays the number of hits as a function of the number of predictions, obtained with three approaches: active learning (*dark blue* curves), passive learning (*dark red* curves) and random (*green*). The approach is able to achieve, on average, 587 hits out of 1,000 predictions via AL, 407 hits, via PL; and the corresponding variances (indicated by the dashed curves) are 35 and 46, respectively. Compared to the random probability of 2.23 hits per 1,000 predictions, the AL result is a 263-fold improvement over random. The improvement of AL over PL is 1.44 fold. The AL improvement over random was reported to be up to 3.19-fold in a previous SVM-based study for predicting the activity of 1,316 drugs against a single target [102]. The same study also reported 1.59-fold improvement between passive and active learners. Closer examination of the results from the top 100 predictions (enlarged in the inset) further shows that hit ratios of 88.0 ± 4.7% and 82.2 ± 6.4% are obtained by the respective AL and PL protocols. The results are obtained with D = 50, which yields optimal results, as can be seen from Figure 12 and Figure 13.

These results permit us to draw two conclusions. First, a hit ratio of 88% is attainable in the top 100 predictions (and 59% in top 1,000) upon adopting a PMF-based AL strategy for identifying hidden/unknown interactions in a sparse (0.32% occupancy) dataset of about 1.5 million potential interactions. Second, the AL method outperforms random by two orders of magnitude and PL by a ratio of 1.5 approximately, in support of AL strategy for predicting new interactions.

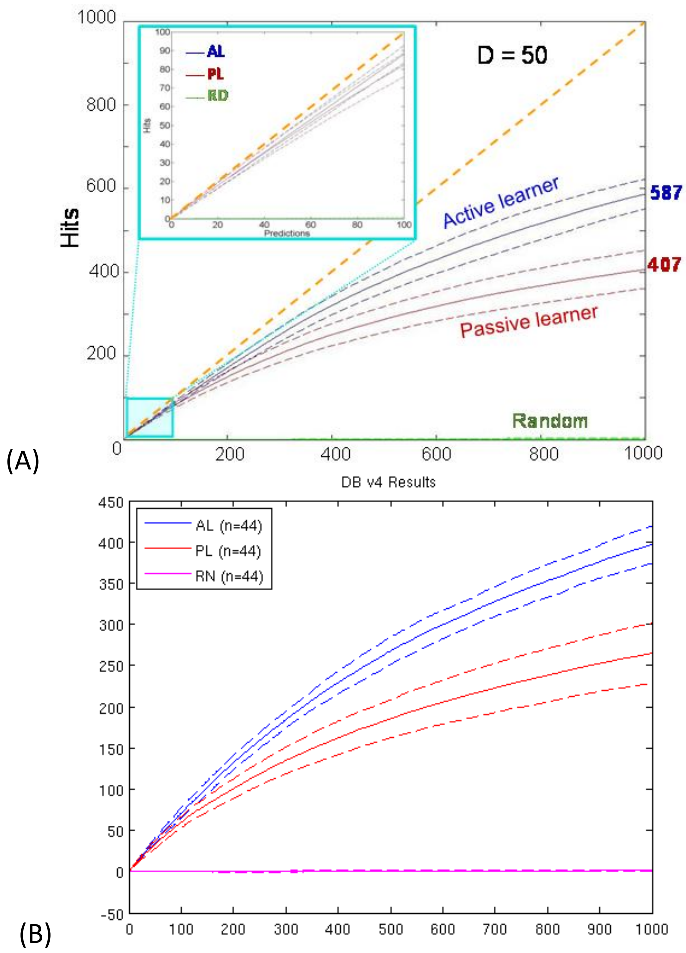


Figure 11: Active learning hidden drug-target interaction prediction performance.

(A) The number of drug-target interactions per drug was reduced from 3.35 (average) to 1 by hiding 70% of known interactions, selected randomly. Simulations were repeated n = 96 times for each of the 1 < m < 1,000 predictions (abscissa) and the number of hits (correctly identified hidden interactions) is plotted for each run, along the ordinate. The *dark blue and dark red* solid curves refer to the average performance obtained by active learning and passive learning protocols, respectively, using D = 50, ε = 3, λ = 0.01, and μ = 0.9 in the adopted PMF algorithm. Dashed curves show the corresponding variances (by one standard deviation) above and below the mean value. The *green* curves (practically overlapping with the abscissa) refer to results from random predictions. The inset shows a close-up of the first 100 predictions. AL reaches an accuracy rate (hit ratio) of 88.0 ± 4.7% and 58.7± 3.5 % in the respective cases of m = 100 and 1,000 predictions. These results are on DrugBank v3 data. (From [97]) (B) The same results reproduced on DrugBank v4 [107] data.

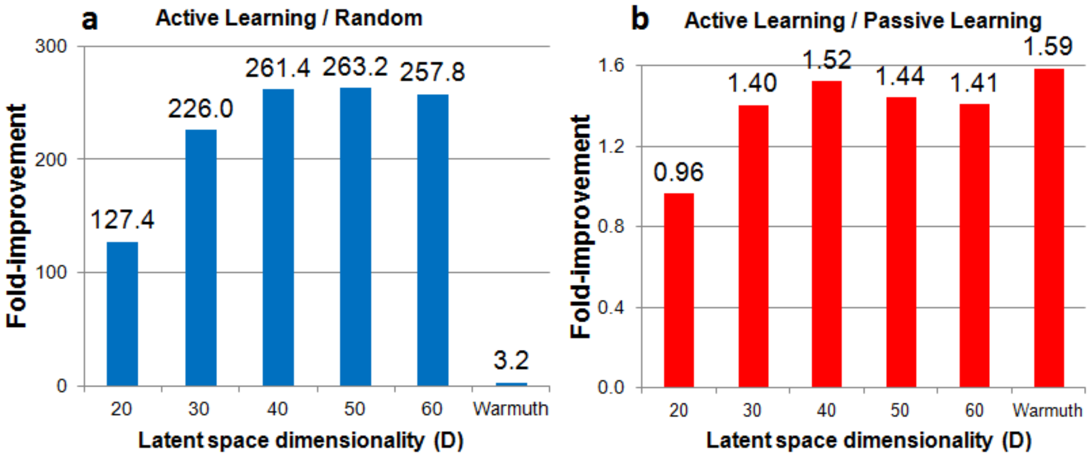


Figure 12: Improvement in prediction accuracy by AL over random and over PL

Random selection of experiments is used as a representative of the performance of the brute force strategies commonly employed in screening based drug discovery efforts. Improvement over random allows the comparison of the various active learning paradigms. Fold-improvement is based on hit ratios obtained at the end of 1,000 predictions, using same parameters as Figure 11. The AL performance levels off at about *D* = 50 in panel a. The last bar in each panel refers to the work of Warmuth et al. (2003). (From [97])

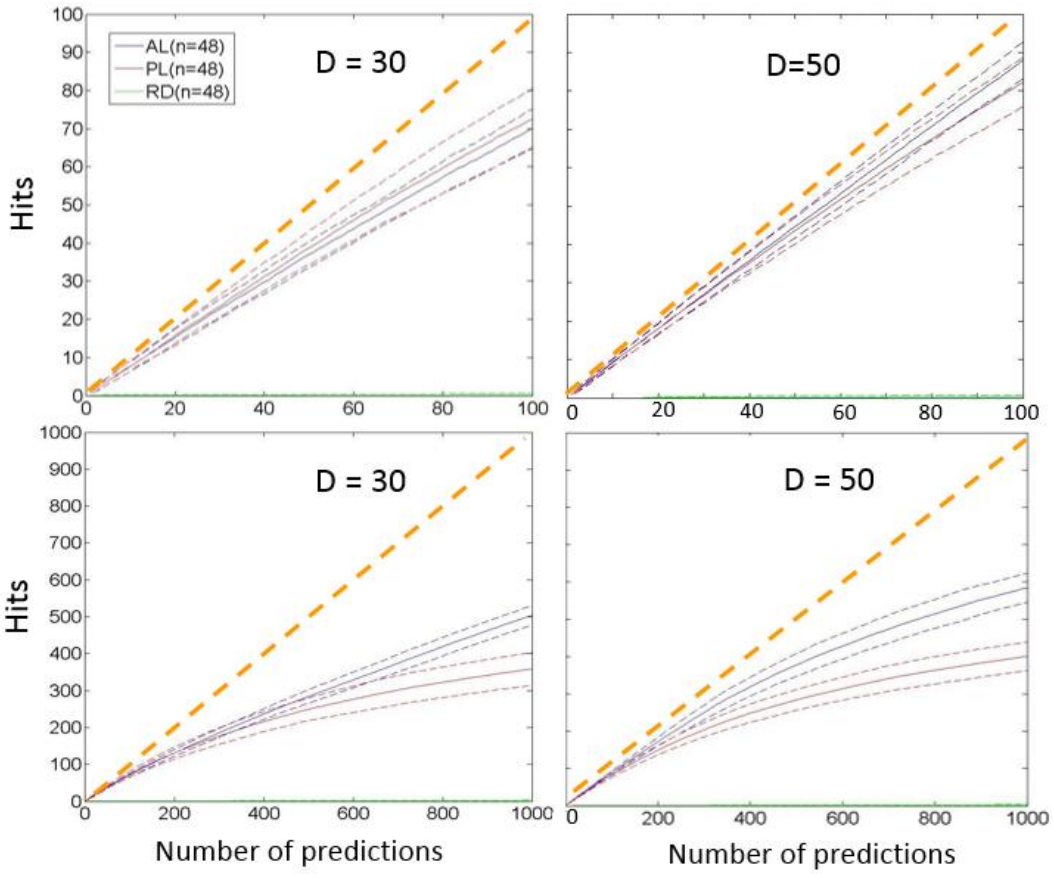


Figure 13: Comparison of the predictive performance of the active and passive learners (AL and PL), and random model as a function of latent variable space dimensionality

In each plot, the ordinate shows the number of hits (accurately predicted hidden interactions) as a function of the number of predicted drug-target associations (abscissa). Blue, red and green solid curves refer to AL, PL, and random results and dashed curves indicate the standard deviation (see caption for Figure 11). The dashed orange line indicates the 100% performance limit for comparative purposes. 70% of the interactions were hidden/removed randomly at the beginning of each simulation, and computations were repeated 48 times with different selections of hidden associations. The upper panels display the results for the top-ranking 100 predictions, and the lower, for the top-ranking 1,000. Overall the AL accuracy rate increases from 50.4% to 58.7%, as we increase the dimensionality from *D* = 30 to 50, for m = 1,000 predictions, and the respective variances are 2.7 and 3.5%. In the case of N = 100 predictions, the AL accuracy rate increase from 70.1% to 88.0%, and the respective variances are 5.1 and 4.7%. (From [97])

2.3 EFFICIENT & ONLINE LATENT FACTOR MODEL BASED DRUG-TARGET INTERACTION PREDICTIONS

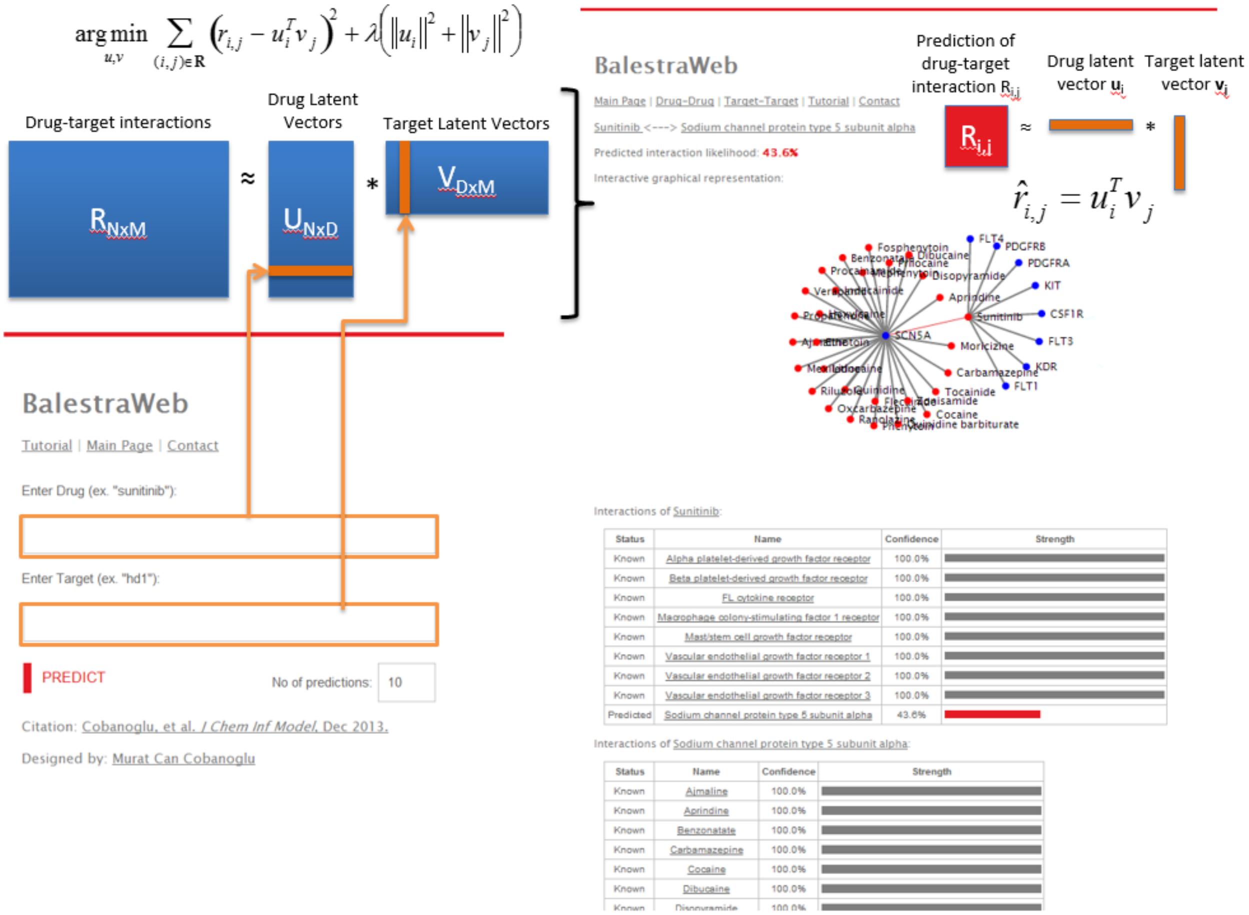
BalestraWeb is built by training a latent factor model, as described in our previous work [97], on approved drugs and their interactions data from DrugBank v3 [99]. To build the latent factor model we use the GraphLab collaborative filtering toolkit implementation [108]. We mapped all the known names, brand names and synonyms of the drugs and targets to the relevant latent factor using a precomputed hash table that allows constant time access and enables maximal efficiency.

The server allows users to submit three types of queries: drug-target interaction, drug-drug similarity and target-target similarity. In the former case (Figure 14), the input is mapped to the corresponding drug latent vector (LV) and target LV, and the dot product of these vectors yields a score for the probabilistic occurrence of the queried drug-target interaction. In the current version, there is an update compared to the original version where this operation is repeated across 128 models the results of which are averaged to reach the reported final value. Alternatively, the user can enter a single type of input, either a drug or a target. If a single drug is entered, the server retrieves the LV for that drug and screens it against the entire set of LVs corresponding to all targets, so as to identify known and newly predicted targets reporting the targets with the maximal predicted interaction scores. Drug-drug and target-target similarity queries provide information on drugs (or targets) similar to the query drug (or target) based on the cosine similarity of their LVs where the average cosine similarity across all models is reported to the user. The use of model averaging enables robust model learning since it prevents the random initialization of the model training stage from playing a role in the reported results.

The output is an interactive graph (which can be downloaded in JSON format) and a table displaying both the known drug-target interactions for the query drug (target) and the top N predicted targets (drugs), rank-ordered by their score. This interactive graph is rendered using scalable vector graphics, which is a widely used tool for displaying graphics on the Web as it is highly efficient in terms of network bandwith use as well as being highly communicative and, if preferred, interactive [109-112]. It enables the transmission of the network using only about 5% of the bandwith that would be required to communicate a static bitmap representation of the same graph, while also enabling interactive use. Furthermore, users can select to view a second layer of interactions beyond the immediate neighbors of the query drug/target in the bipartite network of drugs and targets. The resulting subnet of interactions thus provides a more complete picture of the investigated drug/target in the context of the interactions of their known targets/drugs.

In addition to providing information on the distribution of scores in general, in the tutorial, we provide query-specific histograms in the output files: the distribution of predicted confidence score (for each member of the drug-target pairs) or the histogram of cosine similarities (for each member of the drug-drug or target-target pairs). These histograms facilitate the interpretation of the specific score released for the query pair in the context of the complete distribution of scores for the investigated drug/target, and help make a better assessment of the significance of the outputted score.

BalestraWeb has been significantly updated as of May 7, 2015. Multiple updates/improvements have been made compared to the version published last year. The new BalestraWeb v2 uses the average of 128 models to do all calculations in order to learn a robust model that does not become affected by the random initialization of the latent factor learning process. This is important because during the model learning stage, the optimization landscape is tremendous because the number of parameters to be learned is high (the exact number of parameters to learn are *N*D+M*D*; which is 1313*50 + 1455*50 = 138,400 for DrugBank v4 [107]). Consequently the algorithms that can learn a model in a reasonable time almost always converge to a local optimum. Therefore depending on where in the parameter manifold the random initialization places the model, the converged model (i.e. the learned model) can be different for different initialization seeds. This creates a high variability in the model outputs, which is undesirable. However averaging over a very high number of models all of which have been randomly initiated effectively removes/minimizes this problem because the optima that are frequently reached are all sampled. Instead of picking a single model which yields the first but not necessarily the best fitness, we take the average of all the trained models to minimize overfitting. As it currently exists, there are 11 newly predicted drug-target associations in the current version of BalestraWeb with a predicted interaction score above 90%. These are presented in Table 3. The list of all predictions above the threshold of 70% are reported in Appendix B. The code and all the auxiliary files that run BalestraWeb can be downloaded at http://balestra.csb.pitt.edu/static/balestraweb.zip and the explanation of the contents of this file is provided in Appendix C.


$\arg\min_{u,v} \sum_{(i,j)\in R} (r_{i,j} - u_i^T v_j)^2 + \lambda(\|u_i\|^2 + \|v_j\|^2)$
Drug-target interactions
Drug Latent Vectors
Target Latent Vectors
R_{NxM}
≈
U_{NxD}
*
V_{DxM}
BalestraWeb
Tutorial | Main Page | Contact
Enter Drug (ex. "sunitinib"):
Enter Target (ex. "hd1"):
PREDICT
No of predictions: 10
Citation: Cobanoglu, et al. J Chem Inf Model, Dec 2013.
Designed by: Murat Can Cobanoglu
BalestraWeb
Main Page | Drug-Drug | Target-Target | Tutorial | Contact
Sunitinib <---> Sodium channel protein type 5 subunit alpha
Predicted interaction likelihood: 43.6%
Interactive graphical representation:
Prediction of drug-target interaction $R_{i,j}$
Drug latent vector u_i
Target latent vector v_j
$R_{i,j}$
≈
*
$\hat{r}_{i,j} = u_i^T v_j$
Fosphenytoin
Dibucaine
Benzonatate
Prilocaine
Procainamide
Mephenytoin
Disopyramide
Verapamil
Encainide
Aprindine
Propafenone
Lidocaine
Sunitinib
SCN5A
Ajmaline
Ethotoin
Moricizine
Mexiletine
Tocainide
Carbamazepine
Riluzole
Quinidine
Flecainide
Zonisamide
Oxcarbazepine
Cocaine
Ranolazine
Phenytoin
Quinidine barbiturate
FLT4
PDGFRB
PDGFRA
KIT
CSF1R
FLT3
KDR
FLT1

Interactions of Sunitinib:

Status	Name	Confidence	Strength
Known	Alpha platelet-derived growth factor receptor	100.0%	▬
Known	Beta platelet-derived growth factor receptor	100.0%	▬
Known	FL cytokine receptor	100.0%	▬
Known	Macrophage colony-stimulating factor 1 receptor	100.0%	▬
Known	Mast/stem cell growth factor receptor	100.0%	▬
Known	Vascular endothelial growth factor receptor 1	100.0%	▬
Known	Vascular endothelial growth factor receptor 2	100.0%	▬
Known	Vascular endothelial growth factor receptor 3	100.0%	▬
Predicted	Sodium channel protein type 5 subunit alpha	43.6%	🟥

Interactions of Sodium channel protein type 5 subunit alpha:

Status	Name	Confidence	Strength
Known	Ajmaline	100.0%	▬
Known	Aprindine	100.0%	▬
Known	Benzonatate	100.0%	▬
Known	Carbamazepine	100.0%	▬
Known	Cocaine	100.0%	▬
Known	Dibucaine	100.0%	▬
Known	Disopyramide	100.0%	▬

Figure 14: BalestraWeb architecture and underlying methodology.

The user input (*lower left*) is mapped onto the latent factor vector(s) u_i (for drugs) or v_j (for targets), learned by minimizing squared error regularized by Frobenius norm (see equation at top left). The output (*right*) contains a score R_{ij} representative of predicted interaction confidence along with a graphical representation of the close neighborhood of the query drug (*red dots*) and/or target (*blue dots*) in the drug-target association network, along with a table of known (*grey bars*) and predicted (*red bars*) interactions. Similar features hold for drug-drug and target-target similarity searches and outputs. (From [113]]

Table 3: The top predictions on BalestraWeb v2

The table below shows the drug-target pairs with the highest predicted interaction likelihood scores among all possible BalestraWeb queries. In other words, these are the interactions that BalestraWeb considers most likely based on DrugBank v4. Therefore they represent the top candidates for experimental validation/

Drug ID	Drug Name	Target ID	Target Name	Score
DB00116	Tetrahydrofolic acid	BE0002176	Methylenetetrahydrofolate reductase	1
DB00116	Tetrahydrofolic acid	BE0000331	Serine hydroxymethyltransferase, cytosolic	1
DB00145	Glycine	BE0000331	Serine hydroxymethyltransferase, cytosolic	0.99971
DB00116	Tetrahydrofolic acid	BE0000292	Serine hydroxymethyltransferase, mitochondrial	0.99954
DB00128	L-Aspartic Acid	BE0000277	Calcium-binding mitochondrial carrier protein Aralar2	0.9993
DB00145	Glycine	BE0000292	Serine hydroxymethyltransferase, mitochondrial	0.99912
DB00370	Mirtazapine	BE0000291	5-hydroxytryptamine receptor 1A	0.92736
DB00543	Amoxapine	BE0000572	Alpha-2B adrenergic receptor	0.92068
DB00408	Loxapine	BE0000393	5-hydroxytryptamine receptor 2B	0.91003
DB04946	Iloperidone	BE0000533	5-hydroxytryptamine receptor 2C	0.9008
DB00696	Ergotamine	BE0000342	Alpha-2C adrenergic receptor	0.90057

2.4 METHODOLOGY FOR BUILDING EFFECTIVE MODEL-AVERAGED LATENT FACTOR BASED DRUG-TARGET INTERACTION PREDICTION MODELS

In this section, I outline the steps needed to be taken in order to learn an effective LFM for predicting drug-target interactions, based on my experience during my PhD work.

First, there are a multitude of different algorithms to learn LFM – even though I have started off with discussing PMF, it is important to realize that PMF is only one of the many competing latent factor learning algorithms. To mention a few, there is alternating least squares [114], ALS with parallel coordinate descent [115;116], stochastic gradient descent (SGD) [117], biased methods such as biased stochastic gradient descent [118] as well as many others published in the collaborative filtering literature. All of these models aim to accomplish the same objective at the core: to learn latent factor models that best characterize the nodes of a bipartite network based on their connections. ALS, SGD and PMF are highly similar in their objective functions, effectively minimizing squared loss regularized with the Frobenius norm of the latent variable vectors. The biased methods have a slight difference in that they include global and node-specific bias terms designed to differentiate the nodes that are globally promiscuous from those that are not.

The LFM learning methods have different hyperparameters that control some fundamental aspects of the learning process some of which we have seen above in our discussion of PMF. Specifically important hyperparameters are: the dimensionality of the latent variable space (D), the regularization parameter governing how much to penalize complex models (λ), model update parameter that governs how much to update the model at each iteration (γ) whose larger values yield faster convergence but lower values yield more accurate convergence to local optima. Adaptive approaches where gamma is reduced by a certain rate at each iteration can also be construed. This list is not exhaustive: depending on the type of learning algorithm, there can be other (or fewer) parameters to tune.

There exist Bayesian treatments to these learning algorithms such as the Bayesian Probabilistic Matrix Factorization; BPMF [53] which aim to allow automatic complexity control by essentially integrating over all the hyperparameters. The problem is that after the Bayesian treatment, it is no longer possible to obtain a closed form expression of the gradient of the objective function, which necessitates approximating the posterior directly. This can be achieved through approaches such as Markov Chain Monte Carlo sampling, however sampling the posterior is usually computationally quite expensive, and also makes the method output harder to interpret.

Therefore to build an effective non-Bayesian model, it is important to compare the performance of multiple hyperparameter combinations and algorithms. An example (results from the comparison I did on STITCH v3.1 in order to learn the LFM algorithm/ hyperparameterization to use for our work on HD) can be seen in Figure 15. The full list of results, acquired by testing each method and hyperparameter combination on 16 randomly partitioned train/test sets can be found in Appendix D. As best practice, I would recommend scanning these parameters in log-scale (i.e. 10^{-12}, 10^{-10}, etc) and in the following ranges: for latent variable dimensionality (D) the range 2^4 to 2^7 (which, in log-scale, is simply 4 values: 2^4, 2^5, 2^6, and 2^7); for λ/γ the range 10^{-12} to 10^0 (skipping every other entry, i.e. 10^{-12} then 10^{-10} and so on, might be performed to minimize computational resource use).

If a non-Bayesian algorithm is used, there is a caveat that needs to be taken into account. The random initialization of the parameters that are being learned (note the nomenclature distinction: λ/γ are hyperparameters of the learning process, whereas the actual contents of the latent variable matrices are the parameters that are being learned) can effect the final converged model. Therefore it is often a good idea to use the average of multiple (more than 100) models. Here it is important to take care that only the inner product of the latent variable matrices (i.e. predicted matrices) should be averaged. Averaging the latent variable matrices directly across models is an important mistake to be avoided as it can be algebraically shown to be different from the average of the predictions of the individual models.

In order to test the quality of the models, one commonly used strategy is to evaluate the root mean squared error (RMSE) on held-out test data. I used RMSE for evaluation of the algorithms and hyperparameters to be used in modeling STITCH for the HD project, as shown in Figure 15. It is important to remember that whenever random train/test splits are used, the entire operation that involves this partition must be repeated many times in order to average out the effect of randomness in the train/test split.

Depending on the objective to be accomplished, different performance metrics can also be devised. For example, to learn the model used in the current iteration of BalestraWeb (as of May 7, 2015) I have used rank-based performance evaluation, where the objective maximized is the median rank of the true-but-hidden interactions among all the predicted interaction partners of each drug. It is important to clarify the following point: When evaluating different model learning strategies, I optimize for the model that minimizes the median rank. Once that is completed, I use the best-performing hyperparameters to use the optimal model learning strategy to build the model underlying BalestraWeb. These parameters can be found in the code provided online at http://balestra.csb.pitt.edu/static/balestraweb.zip the contents of which are explained in Appendix C. The scores in Table 3, or anywhere in BalestraWeb for that matter, come from the model that is learned with this optimal strategy. Hence the scores in Table 3, and the rank performance described here are from two entirely different approaches.

Using a rank-based performance evaluation when training BalestraWeb makes the most sense because the most important aspect for the use case of BalestraWeb is to be able to rank the correct but unknown predictions as high as possible in the output predictions. Compared to the hyperparameters used to train the model used in Figure 13, the parameters identified through this scan achieve about 25% higher median rank for true but unknown (i.e. hidden) interactions, achieving a median rank of 20. This is quite impressive: 50% of the true but unknown interactions occur among the top 20 predictions.

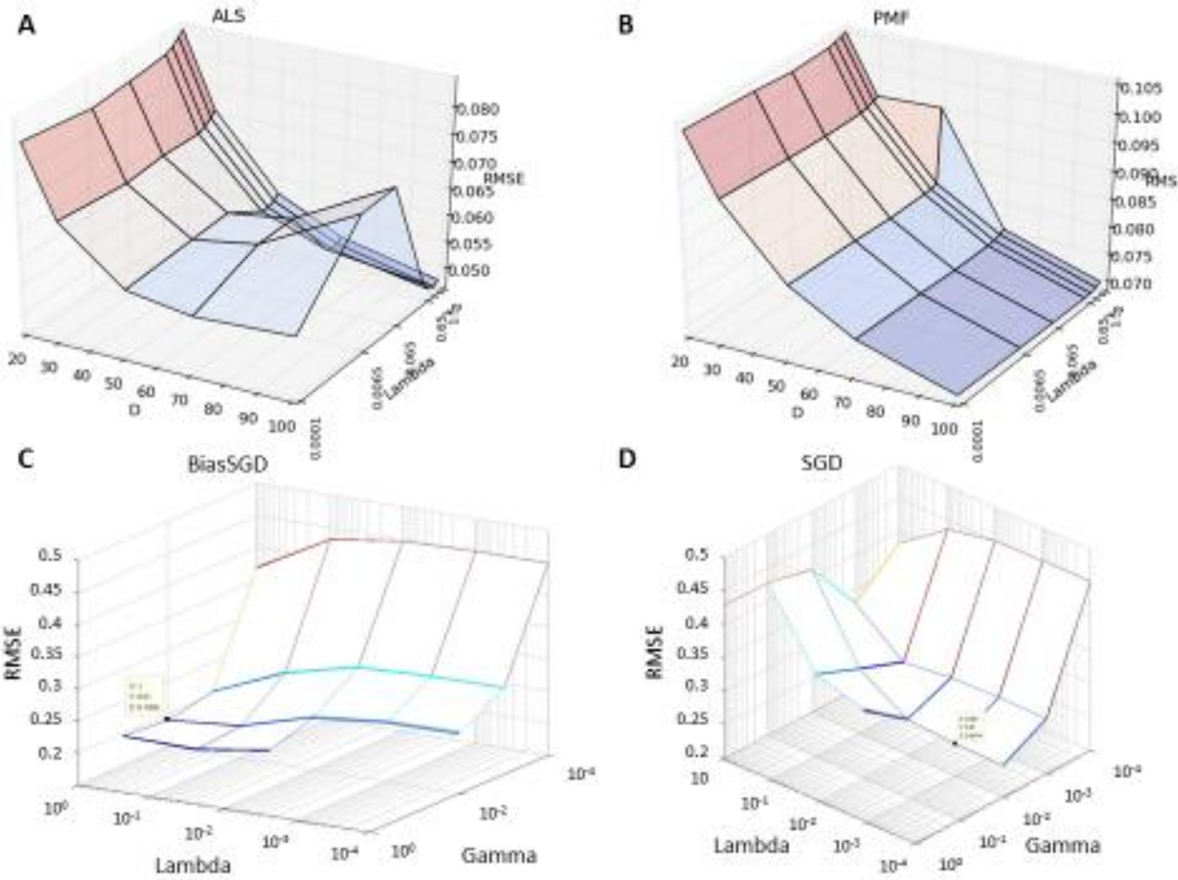


Figure 15: Hyperparameter optimization results with four different latent factor model learning algorithms

The figure shows results from hyperparameter optimization runs where the STITCH v3.1 is used to compare four LFM learning algorithms and their various hyperparameterizations. The panels (a), (b), (c), and (d) respectively show the results with the ALS, PMF, BiasSGD, and SGD algorithms. For the first two (i.e. ALS and PMF) the hyperparameters that are scanned here are the dimensionality of the latent variable space (D) and the regularization hyperparameter that determines how much to penalize complex models (λ). For the last two, the parameters that are scanned are λ as defined before, and γ which is the parameter that controls for how much to update the model at each iteration of the learning process. The vertical axis is always the root mean squared error (RMSE) on the held out test data, averaged over 16 iterations to minimize the effect of randomness in the data train/test split. ALS has performed the best at an RMSE of 4.7% when trained with λ=1.5 and D=100. Therefore I subsequently used ALS trained with these choices of parameter in our HD work to train the LFM on STITCH v3.

2.5 BALESTRATK: PYTHON TOOLKIT FOR DRUG TARGET INTERACTION DATA ACCESS AND INTEGRATION

Programmatically accessing drug-target interaction data requires parsing and constructing data structures amenable for efficient storage and access. Specifically, O(1) time complexity[5] in random access, O(1) name-based lookup operations, O(n) for iterations over all interactions of a particular protein or drug that is called by name are important requirements because these are commonly encountered operations when conducting computational drug repurposing research. Specifically, there are two main databases that I have used as the source of drug-target interactions in my research: DrugBank [99;107] and STITCH [45;119]. These two databases have different uses; DrugBank is a smaller but more richly annotated dataset of interactions with 15,120 links between 7,740 drugs and 4,103 proteins whereas STITCH offers 4,523,609 interactions between 141,799 drugs and 19,488 human proteins. In fact, STITCH is a superset of DrugBank data as it incorporates DrugBank among 14 databases in the latest version of STITCH (v4). Yet for the user who needs to programmatically access a set of specific drugs, the only method is to download these datasets in aggregate format and write purpose-specific scripts that access the relevant information. I have built a Python toolkit that obviates that need by collecting and packaging the code that I have had to write to conduct the research presented in this thesis: BalestraTK. The open source code is hosted on GitHub, which is a commonly used open source code repository and is publicly accessible at: https://github.com/mcc-/balestraTK

[5] O(.): The so-called big-O notation is commonly used in computer science to describe the complexity (often time complexity) of performing an operation in terms of the variable of interest; with constants eliminated from the inside of the parantheses. O(1) represents an operation that takes constant time and hence is the lowest time complexity that can be achieved, O(n) represents an operation that is linear in the number of inputs, O(n^2) is quadratic, etc.

The mode of operation is such that there is a single one-time cost of parsing the data in these databases and constructing the data structures at the very first use of the toolkit, which leads to constant time access in every subsequent use of the toolkit as many times as desired. The data structures that are built are persistent (i.e. they are saved in the disk) and therefore they are used from one instantiation to the other. This means that once the data structures are constructed all subsequent uses also instantiate rapidly.

When using BalestraTK, the user must point the toolkit to the folder where they keep the data downloaded from DrugBank; or in the case of STITCH simply where they intend to keep the data, and the code automatically downloads the required data there. The difference between these databases are due to differences in their license terms: DrugBank prohibits redistribution, thus I require that the user visit the DrugBank website to download the data and then point the toolkit to where the data is stored locally on his/her machine. STITCH license enables the sharing and redistribution of data thus I have coded the toolkit to automatically download the data when the toolkit is first used, if it is not already there. What this means is that if the user already has the STITCH files downloaded, he/she can simply point to the appropriate folder and the toolkit automatically uses them whereas in the absence of one or multiple required files, the toolkit will download the appropriate files to the directory specified by the user.

When using STITCH, one important limitation is that the STITCH interaction file contains only PubChem compound identifiers for chemicals, and Ensembl/UniProt identifiers for proteins. In order to perform most repurposing and/or computational drug discovery efforts the names of both proteins and chemicals, as well as the structures of these chemicals and the sequences, GO identifiers, PFAM identifiers, etc of these proteins are also useful. In recognition of this fact, I automatically download the necessary information and integrate it into the appropriate data structures. DrugBank already contains this information embedded in its data representation in an integrated manner, and I make them accessible programmatically as well. Detailed information as well as usage examples are provided on the public source code repository referenced above.

3.0 COMPUTATIONAL DISCOVERY OF THERAPEUTIC AGENTS AGAINST ALPHA-1 ANTITRYPSIN DEFICIENCY (ATD)

In the following chapter, I will discuss the results of our study towards the identification of novel computational therapeutic agents. The data we analyzed have been acquired from two different studies, and therefore the results in the chapter are organized accordingly. In the first study, we have analyzed the data acquired from a whole genome RNAi knockdown study performed in a *C. elegans* model of ATD [65]. In the second study, we have analyzed the results from a high content screen conducted with the Prestwick library of approved small molecule chemicals.

3.1 METHODOLOGY FOR DRUG REPURPOSING BASED ON MODEL ORGANISM GENE KNOCKDOWN DATA

Model organisms are useful for interrogating different diseases in multiple contexts. Here we present a methodology for using model organism whole genome knockdown (RNAi) data to inform a drug repurposing approach. This methodology has been developed specifically with the intent to apply it to the aim 2, subaim 1 within the context of the A1AD project.

3.1.1 Whole Genome RNAi Knockdown Screen

RNA interference screen was conducted to identify the genetic modulators that affect the accumulation of ATZ [120]. A transgenic *C. elegans* line expressing GFP tagged ATZ in the intestinal cells was derived. The intestinal cells were selected because these cells have the highest biological environment resemblance to the human liver cells. To simplify the identification of transgenic animals, the head muscle promoter myo-2 was tagged with mRFP to serve as co-injection marker. The successful injections led to the derivation of transgenic animals expressing mRFP in the head region and mGFP in the intestinal region.

The transgenic animals were then used to knockdown each of the 16,256 known *C. elegans* genes by using RNAi fed to the worms through bacterial vector. Around 300 worms were exposed to each specific RNAi culture, followed by sorting them into three separate wells (100 worms per well) of a 96-well optical bottom plate. The worms were then anesthetized with 4 mM Levamisole and imaged. In these images, ATZ accumulation manifests high GFP fluorescence and *vice versa.* The GFP signal is normalized with respect to the number of worms that are still alive – which can be quantified through the RFP signal.

One or more plates were processed in separate batches. For every batch, two types of controls were setup: (i) GFP-quenching controls and (ii) empty vector feed controls. The GFP-quenching controls (to be called GFP controls for brevity) allow for the quantification of the baseline signal that is measured even when there is minimal GFP fluorescence. Due to the instrumental variability, this minimal level of signal changes from batch to batch which is why these controls are important. The empty vector feed controls (to be simply called vector controls) are experiments where the bacterial vector was provided with no RNA. The objective of these controls was to determine the fluorescence readout in the absence of any alteration to the disease progression. These controls allow elucidation of the impact of knocking out a particular gene on the disease progression.

Knockouts that result in suppression of fluorescence intensity (i.e. ATZ accumulation) are called ‘suppressor knockouts’. For these types of knockouts, the signal level was similar to the GFP controls. If the gene knockout had no effect on ATZ accumulation, the readout was similar to the vector control readout. These knockouts can be termed ‘no-effect knockouts’. If the gene knockout caused excessive accumulation of ATZ, the fluorescence intensity was excessively high. These are called ‘enhancer knockouts’ and there are no controls that model higher than normal signal. The results of a typical experiment, along with the controls are shown in Figure 16a.

Ideally, if there is ATZ accumulation in the cell (as in vector controls), the intensity of fluorescence should be high and if there are fluorescence quenchers present (as in GFP control experiments), it should be low. However the fluorescence intensity is not spread over a uniform range between different batches, as illustrated in Figure 16b. Therefore it became evident that the raw fluorescence values cannot be compared across batches. Moreover, there were some batches where fluorescence values were inconsistent with the controls, as illustrated in Figure 16c and these data were filtered out from the computational analysis.

3.1.2 The Computational Methodology for Analyzing Whole Genome Knockdown Data

Our computational method consists of two parts: (i) target identification based on suppressor knockouts, (ii) prediction of repurposable drugs inhibiting these targets. Suppressor knockouts identify the genes whose removal alleviates ATZ aggregation. Therefore inhibition of the products of these genes by known drugs should also alleviate ATZ aggregation. The steps in the procedure are enumerated and illustrated in Figure 17.

To identify the suppressor knockouts, we made use of the observation that the GFP-quenching controls model the readout from suppressor knockouts, while the vector controls model the readout of a no-effect knockout. Therefore we collected the control and gene knockout screen data (Step 1) and trained a logistic regression classifier (Step 2) to distinguish the gene knockout as suppressor knockout or not based on the reported fluorescence intensity after knockout. We trained a separate logistic regression classifier for each batch since only values in the same batch can be meaningfully compared.

The logistic regression classifier was used to estimate the probability that a gene knockout was a suppressor knockout (based on the recorded fluorescence signal). If the probability that the gene suppresses ATZ clearance was more than $1 – 10^{-6}$, in other words if the probability of error was less than 10^{-6}, the gene was classified as a suppressor (Step 3). There were 54 genes that were thus identified to cause ATZ accumulation (Step 4). We looked up the sequence of these 54 genes in the WormBase resource using the corresponding gene identifiers [94]. The full list of the 54 genes is provided in Appendix E. For 44 of these genes, WormBase delivered a known sequence (Step 5a).

In parallel, we retrieved the sequences for the targets of all approved drugs in DrugBank [99], comprising step 5b in Figure 17. We then built a database using known drug target sequences, and compared each newly identified ATZ target against this database using the BLAST algorithm [121]. We used an E value cutoff of 1e-6 to select for high sequence similarity (Step 6). We identified three ATZ accumulating genes that were very similar (in sequence) to known drug targets (Step 7). These worm genes, along with the top three most similar drug targets are shown in Figure 17. The list of drugs targeting these human proteins is provided in Table 4. The Tanimoto heatmap showing the level of similarity between these drugs is provided in Table 5.

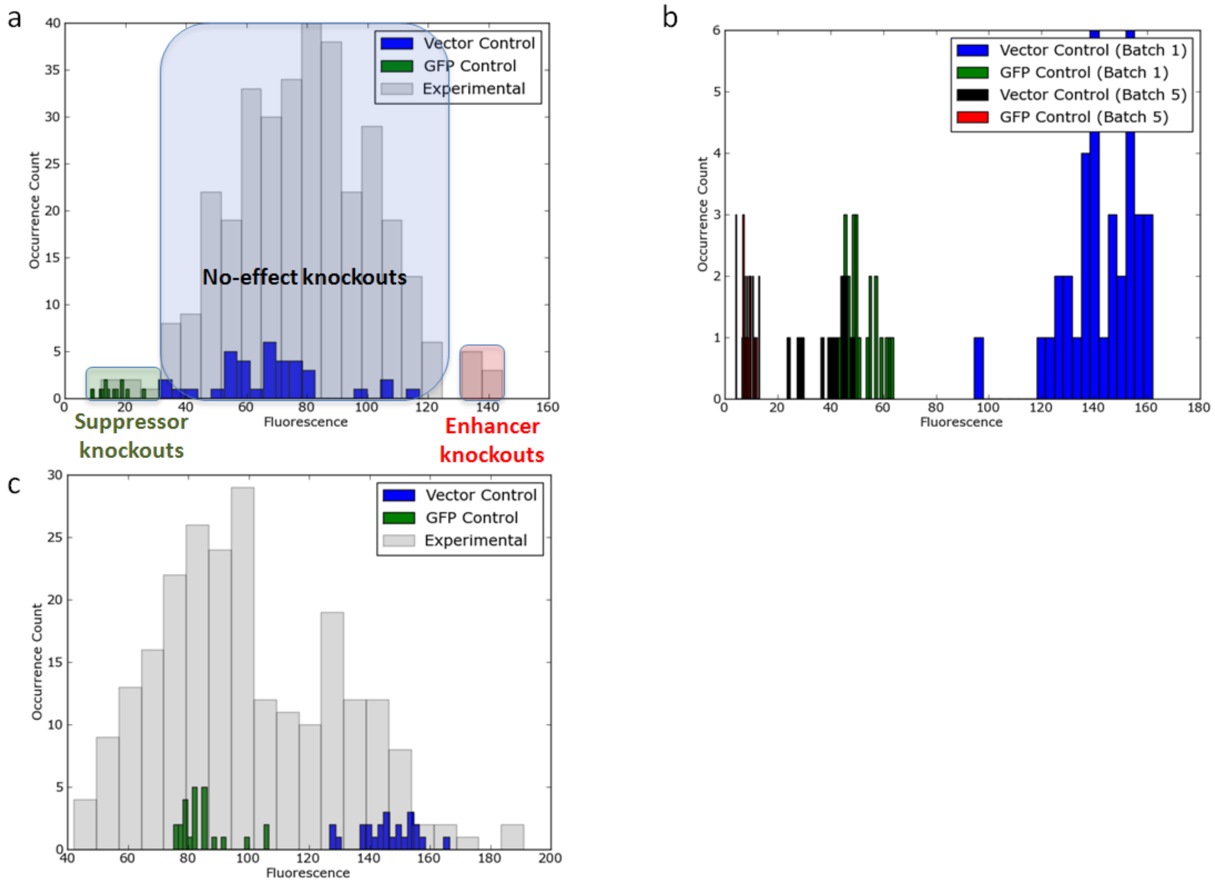
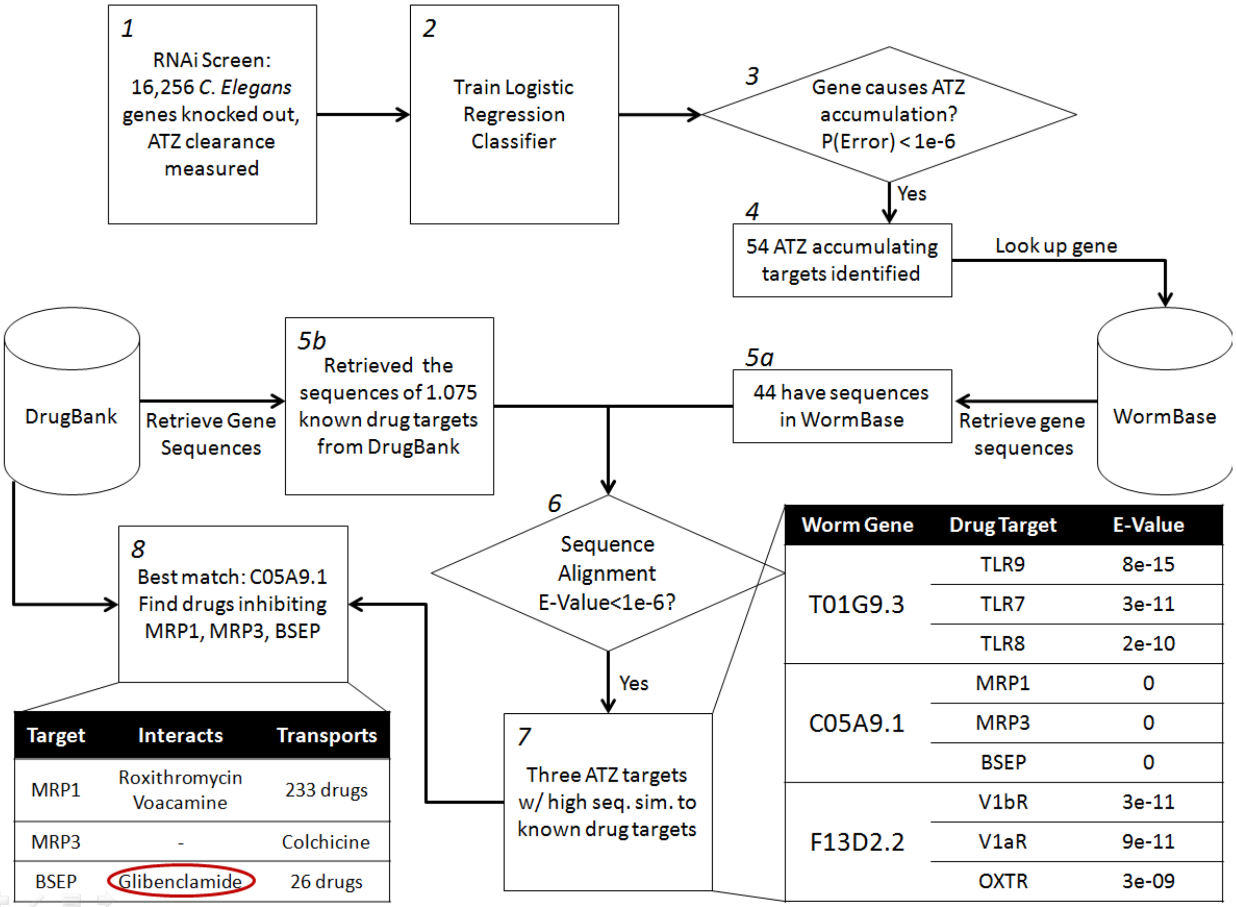


Figure 16: Representative data samples from the RNAi knockdown data show the motivation of our batch-specific classifier based computational workflow

The data clearly shows that data across batches cannot be reliably compared and we have therefore analyzed each batch separately. (a) A typical experiment where the GFP controls model the signal of knockouts reducing ATZ aggregation and vector controls model the effect of no-change knockouts. (b) The range in which fluorescence values are distributed change significantly from one batch to another. (c) For some batches with an inherent measurement flaw, the control fluorescence values were distributed at a much different range than the knockout experiment readouts.


1
RNAi Screen: 16,256 *C. Elegans* genes knocked out, ATZ clearance measured
2
Train Logistic Regression Classifier
3
Gene causes ATZ accumulation? P(Error) < 1e-6
Yes
4
54 ATZ accumulating targets identified
Look up gene
WormBase
Retrieve gene sequences
5a
44 have sequences in WormBase
DrugBank
Retrieve Gene Sequences
5b
Retrieved the sequences of 1.075 known drug targets from DrugBank
6
Sequence Alignment E-Value<1e-6?
Yes
7
Three ATZ targets w/ high seq. sim. to known drug targets
8
Best match: C05A9.1 Find drugs inhibiting MRP1, MRP3, BSEP

Target	Interacts	Transports
MRP1	Roxithromycin Voacamine	233 drugs
MRP3	-	Colchicine
BSEP	Glibenclamide	26 drugs

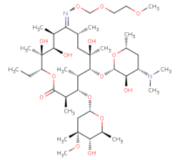
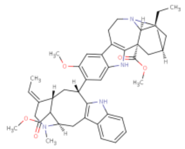
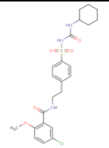
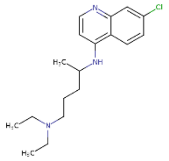
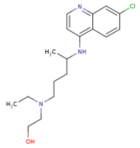
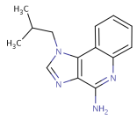
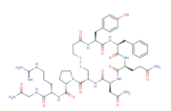
Worm Gene	Drug Target	E-Value
T01G9.3	TLR9	8e-15
	TLR7	3e-11
	TLR8	2e-10
C05A9.1	MRP1	0
	MRP3	0
	BSEP	0
F13D2.2	V1bR	3e-11
	V1aR	9e-11
	OXTR	3e-09

Figure 17: Visual representation of the computational workflow for whole genome knockdown data analysis

The data used for this analysis is the whole genome RNAi knockdown data acquired on a *C. elegans* model of ATD by the Perlmutter lab. The data are initially analyzed using batch-specific machine learning tools to identify the important target genes in every batch, then these genes are mapped to human orthologs. These human proteins are then assessed to identify potential drugs that target the targets of interest. Three such drugs have been identified: Roxithromycin, Voacamine, and Glibenclamide. Glibenclamide has been identified as the best drug candidate among these three drugs because its target of interest, BSEP, is more specific (i.e. less promiscuous) than the target of the other two drugs, MRP1, based on the available data.

Table 4: Whole genome knockdown data analysis results

Computational analysis of the human proteins with high sequence similarity to targets suppressing ATZ clearance, along with the structures of drugs targeting these human proteins are shown. The sequence similarity E-values are also reported.

ATZ clearance suppressor	**Drug Targets with High Similarity**	**Sequence Similarity (E Value)**	**Targeting Drugs**	**Drug Structure**
C05A9.1	Multidrug resistance protein 1	0	Roxithromycin	
			Voacamine	
	Multidrug resistance protein 3	0	–	–
	Bile salt export pump	0	Glyburide (also called glibenclamide)	
T01G9.3	Toll-like Receptor 9	8e-15	Chloroquine	
			Hydroxychloroquine	
	Toll-like Receptor 7	3e-11	Imiquimod	
	Toll-like Receptor 8	2e-10	Imiquimod	Already shown
			Hydroxychloroquine	Already shown
F13D2.2	Vasopressin 1b Receptor	3e-11	Desmopressin	

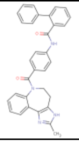
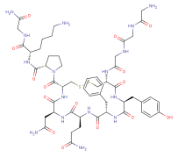
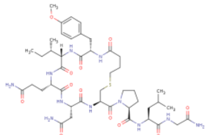
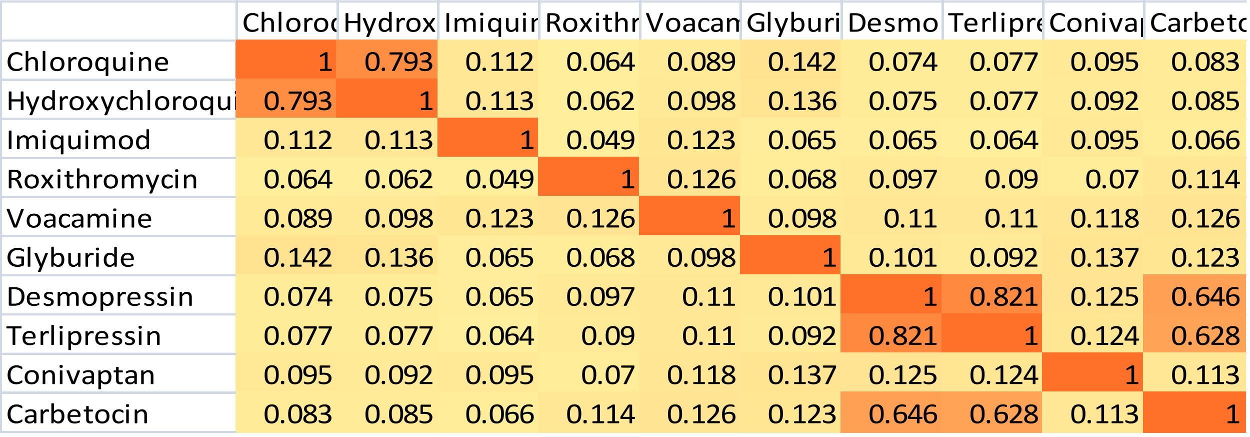
			Conivaptan	
			Terlipressin	
	Vasopressin 1a Receptor	9e-11	Desmopressin	Already shown
			Terlipressin	Already shown
	Oxytocin Receptor	3e-09	Carbetocin	

Table 5: The Tanimoto heatmap showing similarity between the drugs identified for ATD



	Chloroq	Hydrox	Imiquir	Roxithr	Voacan	Glyburi	Desmo	Terlipre	Conivap	Carbeto
Chloroquine	1	0.793	0.112	0.064	0.089	0.142	0.074	0.077	0.095	0.083
Hydroxychloroqui	0.793	1	0.113	0.062	0.098	0.136	0.075	0.077	0.092	0.085
Imiquimod	0.112	0.113	1	0.049	0.123	0.065	0.065	0.064	0.095	0.066
Roxithromycin	0.064	0.062	0.049	1	0.126	0.068	0.097	0.09	0.07	0.114
Voacamine	0.089	0.098	0.123	0.126	1	0.098	0.11	0.11	0.118	0.126
Glyburide	0.142	0.136	0.065	0.068	0.098	1	0.101	0.092	0.137	0.123
Desmopressin	0.074	0.075	0.065	0.097	0.11	0.101	1	0.821	0.125	0.646
Terlipressin	0.077	0.077	0.064	0.09	0.11	0.092	0.821	1	0.124	0.628
Conivaptan	0.095	0.092	0.095	0.07	0.118	0.137	0.125	0.124	1	0.113
Carbetocin	0.083	0.085	0.066	0.114	0.126	0.123	0.646	0.628	0.113	1

3.1.3 Integration of Drug-Target Interaction and Drug Approval Status from Multiple Sources

The NCGC Pharmaceutical Collection (NPC) [122] was used to download the complete list of 7,793 drugs approved for human use (as of 11/20/2012). Of these, 1,426 were matched using their PubChem Compound Identifier (CID) to the chemicals collected in the STITCH chemical-protein interaction database (DB) [45]. STITCH DB currently contains information on 210,169,728 interactions between more than 300,000 chemicals and 2.6 million proteins from 1,133 organisms. These 1,426 approved drugs (represented in both NCGC and STITCH DBs) are reported in STITCH to act as either activators or inhibitors of 5,373 human proteins (targets). Sequence information for 4,022 of these targets could be found among the 205,537 human protein sequences downloaded from the Ensembl DB [123] (as of 11/20/2012).

3.1.4 Mapping Between *H. sapiens* and *C. elegans* Targets

The sequences of these 4,022 human drug targets were screened against the sequences of the worm proteostasis network (PN) modifiers identified in the RNAi screening experiments, using BLASTP [121]. 29 worm (*C. elegans*) orthologs were identified, which represent 29.6% of the initial 98 RNAi hits. Of these, 24 (83%) are reported to be orthologs in Ortholist [95] as well.

3.1.5 Identification of Repurposable Drugs

Having identified the human counterparts of the worm PN modifiers we scanned STITCH to determine whether any known drugs that target these human proteins could be repurposed against ATD. We mapped each of the original RNAi hits to their human orthologs. The human orthologs were chosen to be the target of an approved drug. This analysis of 29 RNAi hits yielded an ensemble of 244 human targets with 525 corresponding approved drugs. Since the Perlmutter lab previously ran the LOPAC library of compounds against ATD, we focused on the compounds that were dissimilar to the chemicals in LOPAC. Therefore we compared each drug to all the chemicals in LOPAC, and filtered out those with the highest similarity compound's name and similarity score. This way, we extracted 30 approved drugs that targeted a human ortholog of a worm PN modifier gene, which were dissimilar to any previously investigated compound.

3.2 GLIBENCLAMIDE AS A NOVEL REPURPOSABLE CANDIDATE AGAINST ATD

All three of the *C. elegans* genes identified as ATD targets (C05A9.1, T01G9.3, and F13D2.2; shown in Figure 17) with high similarity to known drug targets are good leads for directing further experimentation. Yet the objective of the computational analysis was to deliver the single most promising lead. Therefore we concentrated on the gene with the highest sequence similarity to a known drug target: C05A9.1. This gene had an E value of 0 when compared against three out of all the known drug targets, these targets being multi-drug resistance protein 1 (MRP1), MRP3 and the bile salt export pump (BSEP). We looked up the interaction partners of these three proteins in DrugBank (step 8 in Figure 17) and discovered that MRP1 is annotated as a target of roxithromycin and voacamine. BSEP is targeted only by glibenclamide (MRP3 was not annotated as the target of any approved drugs). All three of these drugs are good candidates for experimental testing. However, we prioritized glibenclamide over the other two since glibenclamide's target, BSEP, is annotated to be involved in the transport of 26 drugs while MRP1 is reportedly involved in the transport of 233 drugs. This ten-fold difference in the number of proteins they interact with suggested that BSEP is less promiscuous among the two targets. Therefore we prioritized the experimental testing of the corresponding drug, glibenclamide.

3.3 ADDITIONAL REPURPOSABLE CANDIDATES AGAINST ATD

Our computational approach (described in section 3.1.5) that aims to identify potentially repurposable drugs started with 104 *C. elegans* genes that were confirmed to be PN modifiers, 85 of these were successfully mapped to human PN modifiers using OrthoList [95] and/or WormBase [94], and in turn these human orthologs were mapped to drugs acting on them through STITCH [45] and MetaCore [124] with 12 of the PN modifier *C. elegans* genes being mapped to 48 distinct drugs. The results of the computational analysis were made available through an internet-accessible interactive tree-style visualization framework (at www.ccbb.pitt.edu/faculty/bahar/hitanalysis/) a screenshot of which is presented in Figure 18. Using this interactive visualization framework, one can see that the worm gene 'ageing alteration 1' (abbreviated name: AGE-1, id: B0334.8) for example, is similar to the human protein phosphatidylinositol-4,5-bisphosphate 3-kinase (PIK3CA). The user can click on the B0334.8's name, see the WormBase page corresponding to this gene, and find out that this gene is named AGE-1 and that it is a central component of the *C. elegans* insulin-like signaling pathway. Likewise, more information about the homologue, PIK3CA, can be retrieved by clicking the name. PIK3CA is targeted by three drugs, caffeine, wortmannin and theophylline, according to MetaCore and STITCH. Both of these compounds have been tested by the Perlmutter Lab, with wortmannin showing activity. As illustrated over this single example, one can use the website to interrogate all the experimental results, and accompanying predictions.

Based on the examination of our PN modifier set using comparative analysis to other RNAi screens, pathway analysis and ortholog searches, no PN master gene set emerged with exception of a few known PN modifiers such as AGE-1, inositol-requiring 1 protein kinase related (IRE-1) and abnormal DAuer Formation transcription factor (DAF-16). Rather than further investigate the biologic activity of each new PN modifier, we sought to utilize the gene set as whole to serve as potential drug target list and search for potential compounds that would be effective in decreasing sGFP::ATZ accumulation. The advantage of this approach is 1) prior knowledge of the gene function was not required, just whether the gene functioned as a PN enhancer or inhibitor in order to select an agonist or antagonistic compound, respectively, 2) the low cost and high processivity of screening and validation in *C. elegans*, 3) selection of druggable targets from a gene set based on phenotype, 4) the identification of drugs that could be tested rapidly for efficacy in other types of protein misfolding disorders, and 5) acceleration of the drug discovery process by re-purposing of FDA-approved drugs that also prove to be effective in vertebrate models of misfolded protein disorders.

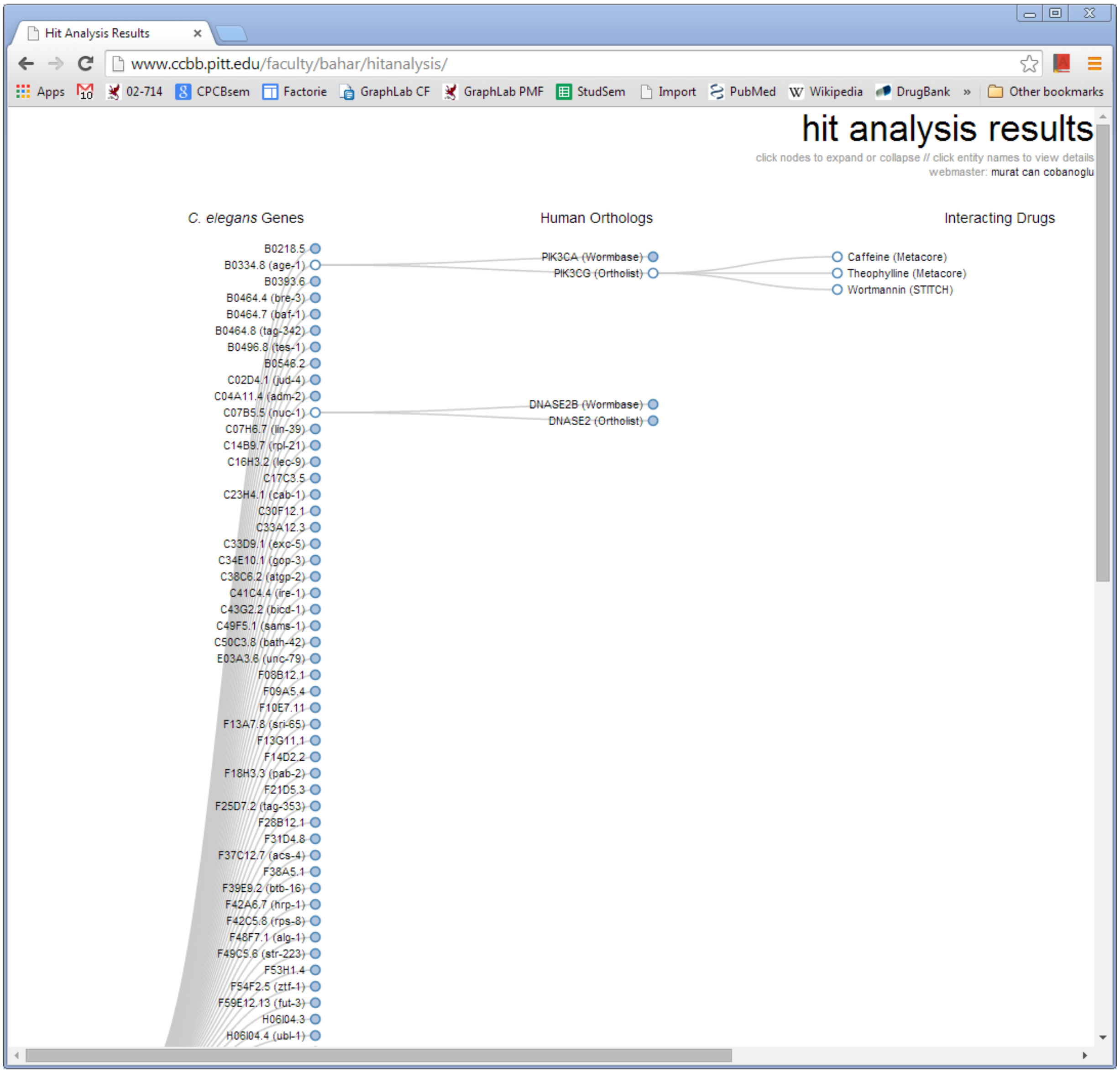


Figure 18: The screenshot of our interactive visualization

I have built a website to visualize the results of the analysis in order to facilitate the interrogation of our results by other scientists. The interactive visualization shows the *C. elegans* genes that were identified to be significant on the left column, the human orthologs that these genes map to in the center column, and the drugs known to be interacting with them on the right-hand side column. The service is accessible over the web at: www.ccbb.pitt.edu/faculty/bahar/hitanalysis/

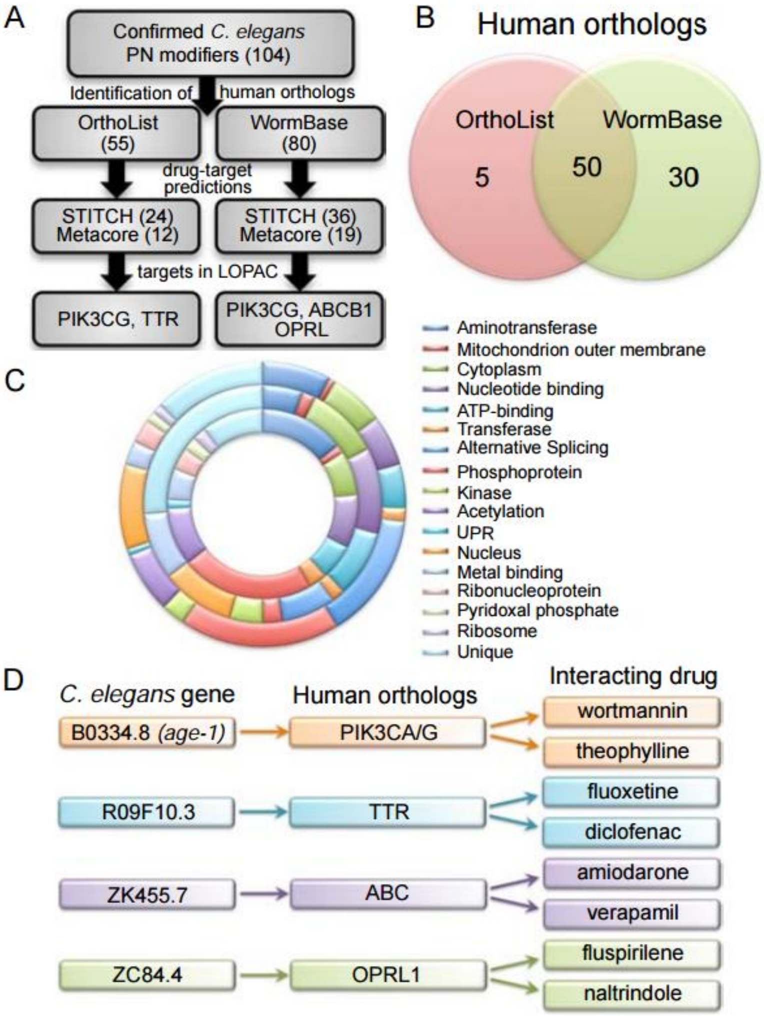


Figure 19: Human orthologs and drug-target interaction prediction

Panel a shows a flow chart summarizing the in silico approach used to identify human drug targets from the 104 C. elegans PN modifiers. Panel b shows a Venn diagram showing the overlap between human orthologs identified by OrthoList and WormBase. Panel c shows DAVID analysis comparing the WormBase (outer ring) and OrthoList (inner ring) assigned orthologs to the original C. elegans protein profile (middle ring). Finally panel d shows the final list of targets and interacting drugs identified using STITCH and MetaCore. (From [65])

In some instances, PN modifiers had multiple (>75) predicted drug interactions. Conversely, we found some drugs to have multiple predicted targets. For example, midostaurin, a synthetic indolocarbazole kinase inhibitor, was predicted to interact with several targets including tyrosine-protein kinase (ABL), vascular endothelial growth factor receptor 2 (VEGFR2), platelet-derived growth factor receptor (PDGFR), RAC-alpha serine/threonine-protein kinase (AKT-1), phosphoinositide 3-kinase (PI3K), mitogen-activated protein kinase 10 (MAPK10), serine/threonine-protein kinase/endoribonuclease (ERN1), receptor-type tyrosine-protein kinase (FLT3) and AMP-activated protein kinase alpha 1 catalytic subunit (PRKAA1). To increase stringency, drugs with multiple or nonspecific target interactions were omitted from further analysis. Moreover, only drug-target interactions predicted by both STITCH and MetaCore were chosen for further investigation. Since some drugs were not readily available due to licensing restrictions or excessive cost, we tested only those compounds that were found in Library Of Pharmacologically Active Compounds (LOPAC). In total, 8 drugs targeting 4 PN modifiers namely, PI3K, Transthyretin (TTR), ATP-binding cassette (ABC) and opiate receptor-like 1 (OPRL- 1) met our criteria for further investigation, as shown in Figure 19. To determine whether any of the 8 compounds were potentially therapeutic, sGFP::ATZ animals were treated for 24 hours and misfolded protein accumulation was measured using the ArrayScan VTI automated imaging machine. Fluphenazine was identified in a previous small molecule screen to reduce sGFP::ATZ accumulation and was included as a positive control [125]. Average results from three independent experiments showed that wortmannin, fluspirilene, fluoxetine and amiodarone significantly decreased sGFP::ATZ accumulation in a dosedependent fashion (12.5-100 µM) compared to the DMSO control (Figure 200, panels A-B). Based on these findings, we selected one of these compounds, fluspirilene, and tested it on a mammalian cell line expressing

ATZ. As shown with *C. elegans*, fluspirilene showed a dose dependent decrease in ATZ accumulation in ATZ-inducible HeLa cell line, HTO/Z (Figure 200, panel C).

We used a genetic approach to obtain insight into drug-target interactions. Wortmannin is a fungal steroid metabolite that inhibits mostly class I and III phosphatidylinositol 3-kinases (PI3Ks) [126]. In a *C. elegans* model of hypoxic injury, 100 µM wortmannin blocks autophagy by inhibition of the class III PI3K, VPS-34 [127]. Since autophagy inhibition enhances sGFP::ATZ accumulation, wortmannin was more likely to inhibit the class I PI3K, AGE-1, which would phenocopy the effects of reduced insulin/insulin-like signaling (IIS), rather than VPS-34 [128]. To determine whether AGE-1 was the target of wortmannin in this model, the Silverman lab first crossed sGFP::ATZ animals with AGE-1(hx546) mutants. As expected, the loss of AGE-1 activity resulted in a marked, but not complete, decrease in ATZ accumulation (Figure 21A). If the effects of wortmannin and AGE-1(hx546) on sGFP::ATZ accumulation were in the same or different pathways, then treatment of sGFP::ATZ; AGE-1(hx546) animals with an effective, but not maximal, dose of wortmannin (Figure 21B) would be expected to have no or an additive effect, respectively. No additive effect was detected (Figure 21C), despite GFP(RNAi) demonstrating that the sGFP::ATZ levels were not below the ArrayScan VTI level of detection (Figure 21B-C). Loss of AGE-1 activity, activates a downstream FOXO transcription factor, DAF- 16, which leads to decreased sGFP::ATZ accumulation (Figure 21A). Thus, if wortmannin inhibits AGE-1, a DAF-16 loss-of-function mutation should suppress the protective effects of the drug. This was the case as sGFP::ATZ; DAF-16(m26) animals were resistant to the effects of the drug, although sGFP::ATZ accumulation could still be modulated with GFP(RNAi) treatment (Figure 21D). Interestingly, the other three compounds isolated via the in silico screen, as well as the fluphenazine positive control, reduced sGFP::ATZ accumulation in sGFP::ATZ; DAF-16(m26) animals (Figure 21E). Taken together, these studies strongly suggested that wortmannin inhibited the class I PI3K, AGE-1 and that the other compounds were active on other target pathways. If this were the case, than combination therapy between wortmannin and one of the other compounds should be feasible. To test this hypothesis, we treated sGFP::ATZ animals with equal amounts of wortmannin and fluphenazine at three different concentrations. In all cases, combination therapy decreased sGFP::ATZ accumulation more than either monotherapy (Figure 21F).

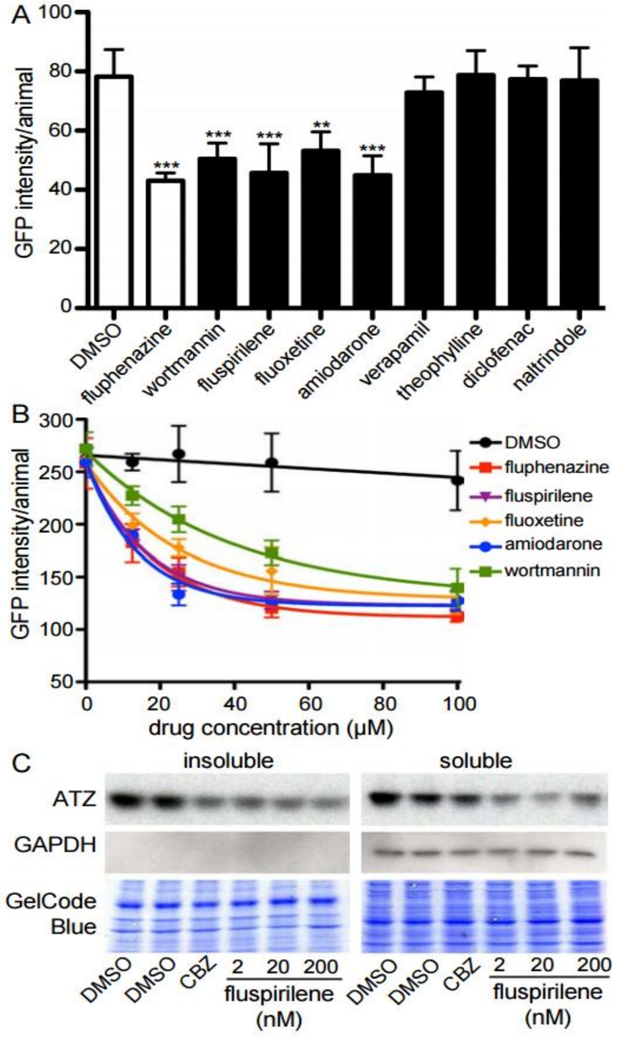


Figure 20: Experimental testing of drugs predicted against ATD

Panel a shows the fluorescence on L4 GFP::ATZ animals that were treated with 100 µM of each drug for 24 h, and analyzed using the ArrayScan VTI. Panel b shows the drug dose response curves. The experiment was repeated 3 times, and a representative experiment shown. The error bars represent the SD of 5 replicate wells (n>150 animals/treatment). Statistical significance was determined by using a Student's t-test (*** P < 0.001, **P < 0.01). Panel c shows the effect of fluspirilene on steady state levels of ATZ in a cell line model of ATZ. HeLa cells engineered to express ATZ (HTO/Z) were treated with DMSO, carbamazepine (CBZ) (positive control) or fluspirilene for 48 h. Lysates were prepared and separated into soluble and insoluble fractions. Samples were analyzed by immunoblotting with antibodies against AT (top), and GAPDH (middle). GAPDH is cytosolic marker and its absence in the insoluble fraction indicates correct fractionation. The blots were also stained with GelCode Blue (bottom) to demonstrate equal sample loading in each well. (From [65])

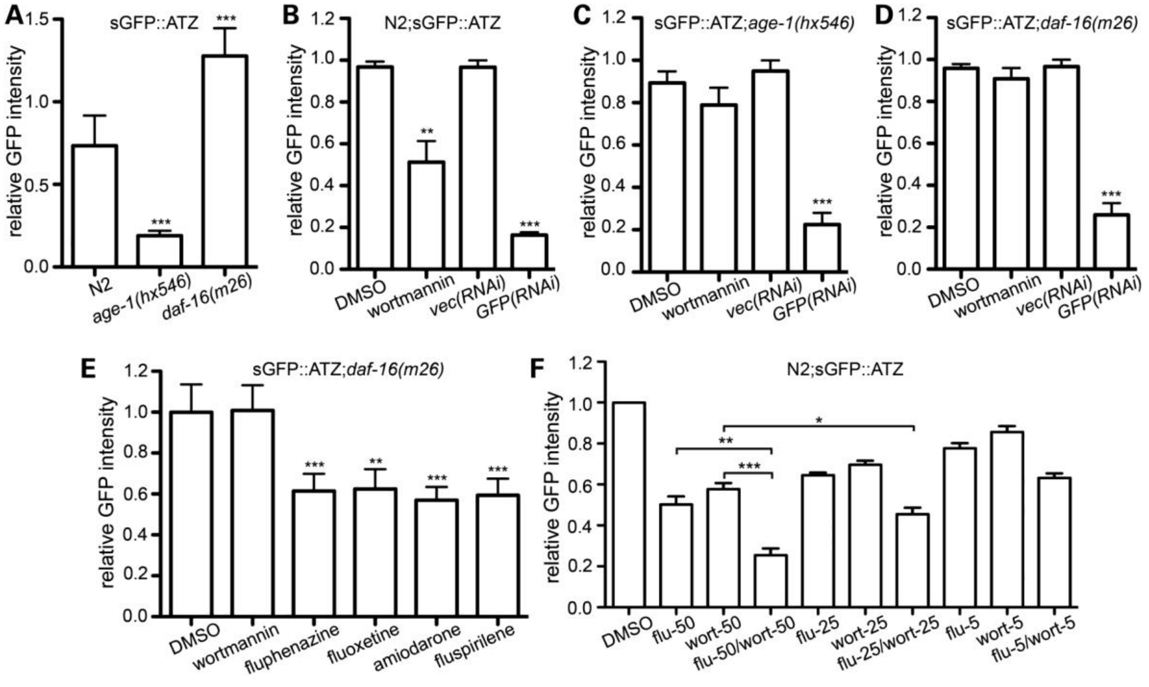


Figure 21: Validating AGE-1 as the target of wortmannin action

Panel a shows the steady state expression levels of sGFP::ATZ in the N2, AGE-1(hx546) and DAF-16(m26) backgrounds. Data is normalized to N2;sGFP::ATZ worms. Panels b through d show the effect of wortmannin on steady state levels of sGFP::ATZ. N2;sGFP::ATZ (B), sGFP::ATZ;age- 1(hx546) (C) and sGFP::ATZ;DAF-16(m26) (D) animals were treated with wortmannin (100 µM) for 24 h and analyzed using the ArrayscanVTI. GFP(RNAi) treatment was included as a control to show that ATZ levels could be further reduced in each line. Note wortmannin reduced the sGFP::ATZ level in the wild-type N2 but not in AGE-1(hx546) or DAF-16(m26) mutant backgrounds. Panel e shows the effect of various drugs on sGFP::ATZ;DAF-16(m26) animals. Of the drugs known to decrease sGFP::ATZ levels in the N2 background, only wortmannin failed to reduce sGFP::ATZ in the DAF-16(m26) background. Panel f shows data from ATZ::GFP animals that were treated with 5,12.5 or 50µM of fluphenazine and wortmannin, either singly or in combination. The data was normalized to the untreated DMSO control within each experiment. All experiments were repeated at least 3 times with n>150 animals/treatment. Error bars represent SD (A-E) or SEM (F). Statistical significance was determined using the Student's t-test. ****P* < 0.001, ***P* < 0.01, **P* < 0.05 (From [65])

3.4 METHODOLOGY FOR HIGH CONTENT SCREENING DATA ANALYSIS AND HIT DIVERSIFICATION

As part of the A1AD project, the Perlmutter, Silverman and Pak labs have screened the Prestwick chemical library (PCL) to test the ability of these drugs to modulate ATZ aggregation using the transgenic *C. elegans* model that was developed as the model system [129;130]. The Prestwick library consists of 1280 drugs approved for human use and mostly off-patent provided in DMSO solution at 10 mM concentration hence ready for rapid screening deployment [96]. The effect on suppressing ATZ aggregation has been quantified using 'B-scores' where lower scores indicate aggregation suppression, and higher scores indicate increased aggregation. In accord with the terminology adopted in our earlier work [65] compounds that significantly suppress ATZ aggregation are called 'inhibitors' (B-score < - 2); those that increase ATZ aggregation are termed 'enhancers' (B-score > 2). Both groups have significant effect, and are collectively called 'actives'. The remaining are called 'no-effect' compounds.

We analyzed the high content screening (HCS) data by a 4-step protocol: (i) chemical-based active diversification; (ii) target-based active diversification, (iii) mapping of drugs to their targets and the pathways of these targets, and the identification of the targets and pathways of active compounds that are significantly enriched. Each of these four steps is described in detail in the following subsections, with the methodology employed for the mapping of drugs to targets/pathways and the enrichment analysis of these targets/pathways (steps (iii) and (iv)) described together because their analysis steps were inseparably connected.

3.4.1 Chemical-Based Active Diversification

The inhibitor chemicals have the desired effect of suppressing ATZ aggregation; hence it is useful to discover other purchasable compounds that could potentially be better therapeutic agents than those found in the Prestwick screen.

First, we decided to identify the chemical descriptors that distinguish the inhibitor compounds from those with no-effect and the enhancers – in other words compounds with B-score < -2 *vs* the rest. We have performed this through training a logistic regression classifier that learns to classify a compound based on its chemical fingerprints as inhibitor or not. To extract fingerprints from the chemical structures we used OpenBabel's python wrapper Pybel and specifically the MACCS[6] fingerprints (of which there are 166) as calculated by Pybel [93;131]. The distribution of weights of this classifier, along with the chemical structures of the highest and lowest coefficients are shown in Figure 22. The highest coefficients represent the chemical features most useful for discriminating actives; while the lowest coefficients, conversely, identify chemical groups that are least discriminative. We used this classifier to classify all 12.8 million purchase-ready compounds in ZINC [132]. We identified the compounds classified as being potential inhibitors against ATD, then clustered them based on chemical composition, selected one representative from each cluster (the closest to the centroid of the cluster), and provided a list of 342 compounds to be tested for activity, and this list can be found in Appendix F. The workflow we adopted is visualized in Figure 23.

[6] MACCS: Molecular ACCess System

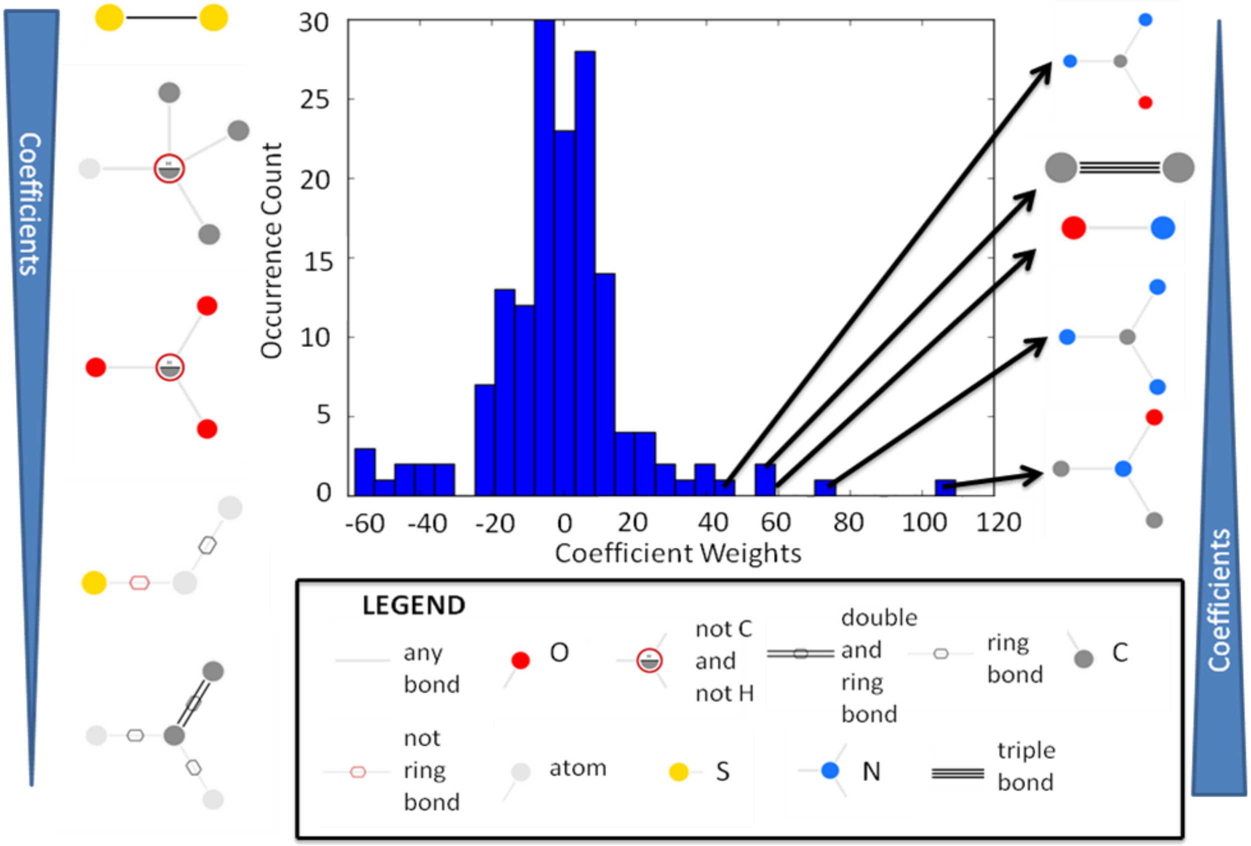


Figure 22: Visual representation of the chemical fingerprint classifier for active identification

The coefficients corresponding to chemical fingerprints that best classify the ATZ aggregation modulator chemicals have high absolute values. To extract fingerprints from the chemical structures we used OpenBabel's python wrapper Pybel and specifically the MACCS fingerprints (of which there are 166) as calculated by Pybel [93;131]. The features with strong positive values (shown on the right hand side) select for chemicals with high activity in ATZ clearance. Conversely, the features with strongly negative values represent features that strongly select for molecules with little or inverse effect in the disease progression. Therefore the chemical has the features with high coefficient values (right hand side), and does not exhibit the chemical features that have low cofficient values (left hand side).

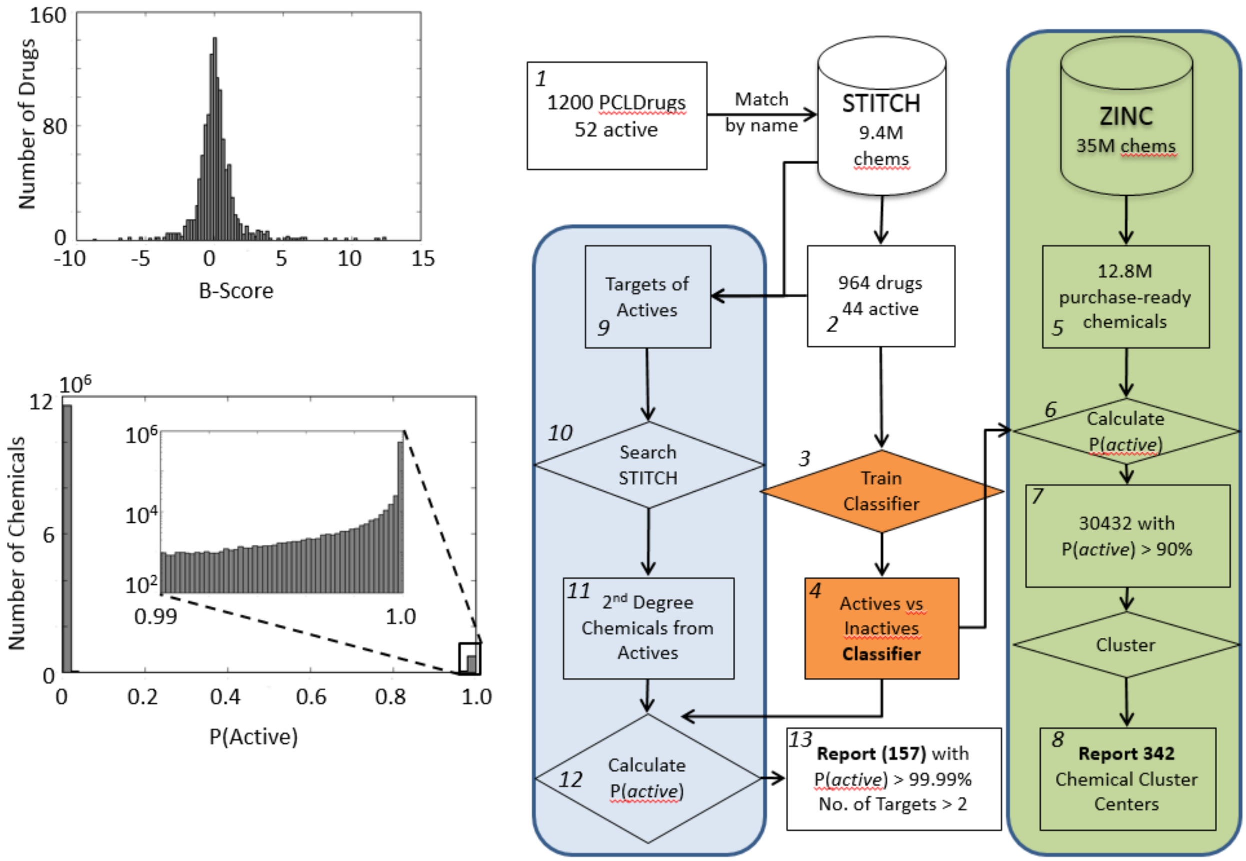


Figure 23: Visual description of the high content screening data analysis and hit diversification workflow. The right hand side panel shows the workflow of the computational process. Specifically, the tested drugs in the PCL (the distribution of the activity, as measured by the B-score, is provided on the top left) are used as input to identify the chemical features that distinguish the 52 actives from the remaining inactives, then these properties are used to search ZINC (the distribution of probability of activity is provided on the bottom left), as well as being used to search STITCH-target sharing compounds. The 157 results of the target based diversification are reported in Appendix F and the 342 results of the chemical based diversification are reported in Appendix G.

3.4.2 Target-Based Active Diversification

Another approach to identify chemicals with the desirable inhibitory effects is to look for other drugs associated with the targets of the inhibitors. To this end, we have identified the interaction partners (i.e. the proteins that are targeted) of the inhibitor compounds from the STITCH dataset v4 [119], then identified all the drugs that potentially interact with these targets (using STITCH), and then used the chemical feature based classifier we trained in the previous step to calculate the probability of being an inhibitor for each of these compounds. We extracted the compounds that were classified as actives by the classifier (based on their probability of being active estimated by the classifier) with molecular weight above 300 (to exclude non-drug-like molecules such as zinc, copper, or mercury in STITCH). There were 157 such compounds. We report them in Appendix G.

3.4.3 Overlap between target-based and chemical-based active diversification

We have reported 342 chemicals through the chemical based active diversification strategy (reported in Appendix F) and 157 chemicals selected through target based active diversification (reported in Appendix G). It is important to investigate the degree of overlap between these two different sets of chemicals. To evaluate if there are any compounds shared in both lists, we computed the Tanimoto similarity and identified that there were no compounds shared.

We next asked if these two sets of chemicals are more similar than would be expected by chance. To this end, we compared the similarity between the two selected sets of compounds to the similarity between two sets of 1200 randomly selected (without replacement; i.e. these two sets are mutually exclusive) compounds from the ZINC purchase ready set library [132] . We compared the resultant distributions using Kullback-Leibler (KL) divergence, which is a measure of the difference between two distributions as described by Baldi and Nasr {Baldi, 2010 585 /id}, where the authors report that a typical molecule has a KL divergence of 0.003 whereas the atypical molecule has KL divergence of 1.075. The KL divergence between these two distributions (i.e. similarity of two sets of random compounds versus similarity of the compounds in two diversification sets) is 0.031, meaning that these two distributions are similar, which in turn means that the two different strategies have produced consistent compounds. The two histograms in Appendix H illustrate that these two sets of chemicals, the result of chemical based diversification (i.e. those in Appendix F) and the result of target based diversification (i.e. those in Appendix G) have as little similarity as to be expected in a two large randomly selected set of chemicals. This validates that our two diversification strategies are indeed necessary since they diversify and select for different compounds.

3.4.4 Target/Pathway Identification Through Enrichment Scores

To calculate enrichment scores, the Prestwick library of compounds were mapped to pathways in two ways: through STITCH [119] targets and through KEGG [133] targets. Prestwick compounds were mapped to the corresponding chemicals in the STITCH database [119], the proteins listed as their interaction partners in STITCH were identified, and these proteins were mapped to the corresponding proteins in KEGG [134] through ENSEMBL [135]. Finally, these targets and the pathways that these targets occur in were identified for each drug.

Prestwick compounds were mapped to the corresponding drug entries in the KEGGdrug database [133]. The targets of these drugs, as well as the pathways of these targets were identified in KEGG and associated with each drug.

In both instances, the KEGG is used as the source of pathway information. However the target-to-drug mapping is more extensive in STITCH with 390,000 chemicals, 2.6M proteins and 1 trillion interactions [119]. KEGG, on the other hand, has 10,103 drugs with no detailed statistics of the interactions. Therefore, we will focus here on the enrichment scores derived from the STITCH, although these were also calculated for the KEGG.

For each drug we define activity as either enhancing (B-score > 2) or suppressing (B-score <-2) ATZ aggregate formation, which is a quantification of the disease phenotype. All drugs which influence a target/pathway of interest are expected to either increase or decrease aggregation, and conversely any drug that increases or decreases aggregation is influencing a pathway/target of interest. To quantitatively identify the targets/pathways of interest, we collected all drugs with suppressive or enhancory effect into a set of active drugs and calculated the enrichment score of a given target or pathway *t* as follows:

$$E(t) = \frac{\sum_{d \in A} I(d,t)/|A|}{\sum_{d} I(d,t)\ /|D|}$$

where *A* is the set of all active drugs, *D* is the set of all drugs, |.| denotes the number of elements in a set, and *I* is the indicator function that is 1 if drug *d* targets *t* (*t* is either target or pathway depending on the enrichment being calculated) and 0 otherwise. In our case |*A*|=104, and |*D*|=966.

We identified the pathways/targets of interest by quantifying the candidate pathway/target set that best separates the actives (B-score < -2 or B-score > 2) from the inactives (-2 < B-Score < 2). In order to achieve this goal, we calculated the reduction in Shannon's entropy after splitting the drugs according to their interaction with each candidate, and computed the enrichment score, picking the best target/pathway recursively. Lower entropy indicates less disorder (higher confidence) in the respective drug activity annotation.

In the Prestwick library (PL), the number of active compounds are less than the inactive compounds. This imbalance was corrected by weighting the actives with $w_a = {(|D| - |A|)}/{|A|}$ such that the total weighted sum of actives equals that of the inactives. The weighted data were used to learn a decision tree by minimizing the entropy using the method described by Quinlan [136]. In our case, this method selects the target that best separates the actives from the inactives by choosing the target that when split accordingly minimizes the information entropy (also called Shannon's entropy) defined as: $H(X) = -\sum_i p(x_i) \ln\left(p(x_i)\right)$ where the probabilities are frequency counts of the members of the two classes (actives/inactives) weighted according to the weighting scheme described above. Intuitively, the method selects those targets that separate the drugs into two sets: actives and inactives. The results of this procedure are shown in Table 6 and Table 7 for the respective datasets KEGG and Stitch. The full names of the proteins listed in Table 6 are as follows: dopamine receptor D2 (DRD2), calcium channel, voltage-dependent, L type, alpha 1C subunit (CACNA1C), 5-hydroxytryptamine (serotonin) receptor 2C, G protein-coupled (HTR2C), angiotensin I converting enzyme (ACE), calcium channel, voltage-dependent, N type, alpha 1B subunit (CACNA1B), adrenoceptor alpha 2B (ADRA2B), prostaglandin-endoperoxide synthase 2 (PTGS2), glutamate receptor, ionotropic, kainate 5 (GRIK5), adenylate cyclase-coupled 5-hydroxytryptamine (serotonin) receptor 7 (HTR7), protein phosphatase 3, catalytic subunit, alpha isozyme (PPP3CA), prostaglandin-endoperoxide synthase 1 (PTGS1), calcium channel, voltage-dependent, T type, alpha 1H subunit (CACNA1H), adrenoceptor alpha 1A (ADRA1A), solute carrier family 6 (neurotransmitter transporter), member 4 (SLC6A4), and 5-hydroxytryptamine (serotonin) receptor 1A (HTR1A). The full names of the proteins listed in Table 7 are as follows: v-rel avian reticuloendotheliosis viral oncogene homolog A (RELA), adrenoceptor alpha 1A (ADRA1A), renin (REN), calcium channel, voltage-dependent, L type, alpha 1S subunit (CACNA1S), cholinergic receptor, muscarinic 1 (CHRM1), ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), and adrenoceptor alpha 1D (ADRA1D). Since the results are based on different datasets, the drugs are annotated with different targets based on the dataset and thus the results vary between the two tables. Adrenoceptors and calcium channels are common in both tables. The results are described and discussed in detail in section 3.5.2.

We also performed this entire procedure for pathways instead of targets after mapping each target to its KEGG pathway. The results were not appropriately high quality because we do not have perfect information on the (i) drug-to-target mapping, (ii) target-to-pathway mapping and when these two get compounded in the drug-to-target-to-pathway mapping the end result was that the compounded errors made it impossible to form a convincingly accurate enrichment analysis. Therefore we did not analyze those results further.

Table 6: The tree structure of the targets of active drugs based on KEGG target information

We trained an entropy-minimization based decision tree to separate active drugs from inactive drugs using the algorithm due to Quinlan [136]. The targets of the drugs identified to be active in the screen were analyzed using an information entropy minimization strategy to build the following tree. The calcium channels, which are overrepresented are highlighted in orange. At each node, the entropy of the class labels (i.e. 'active' or 'inactive') are shown.

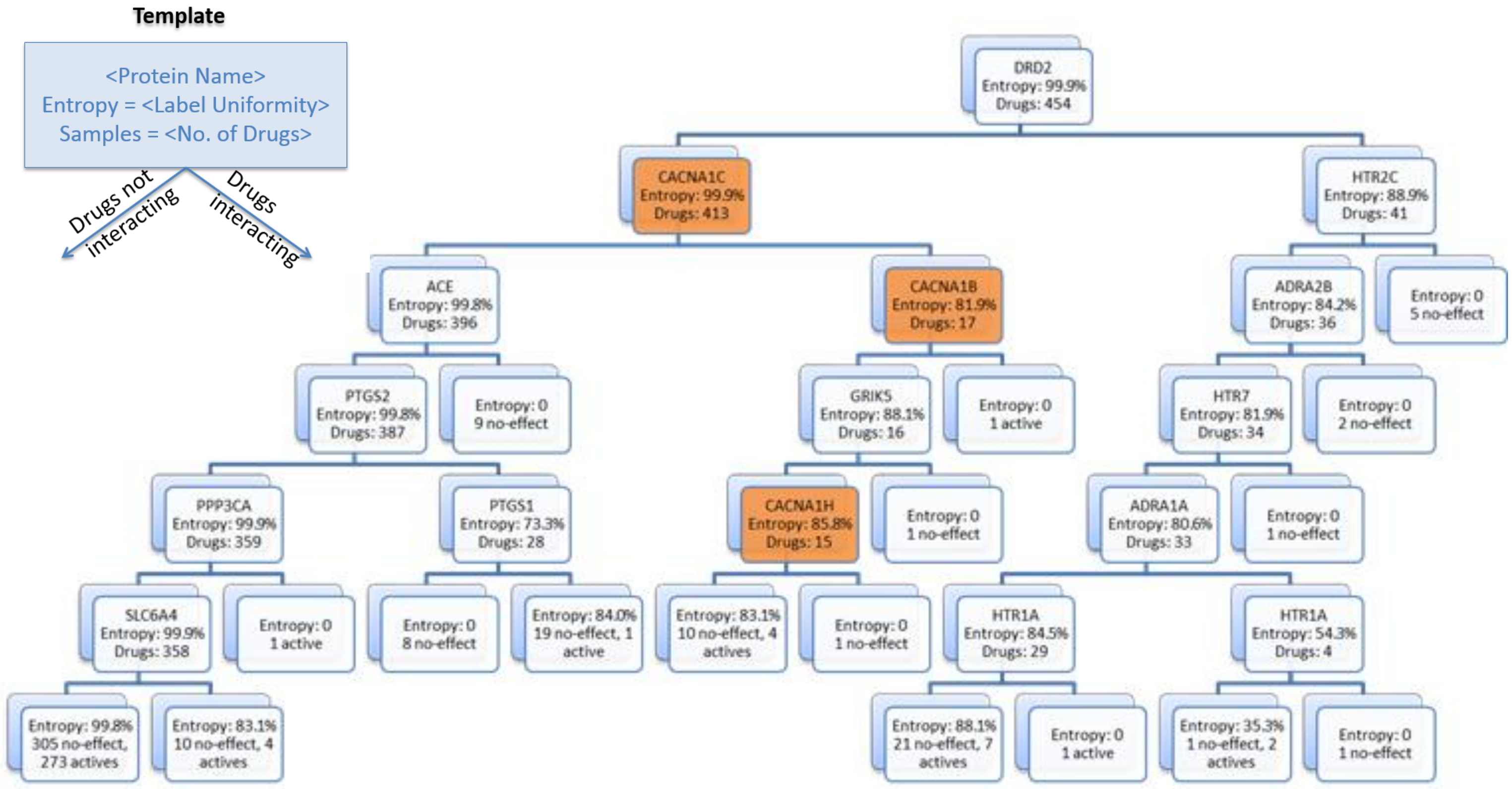
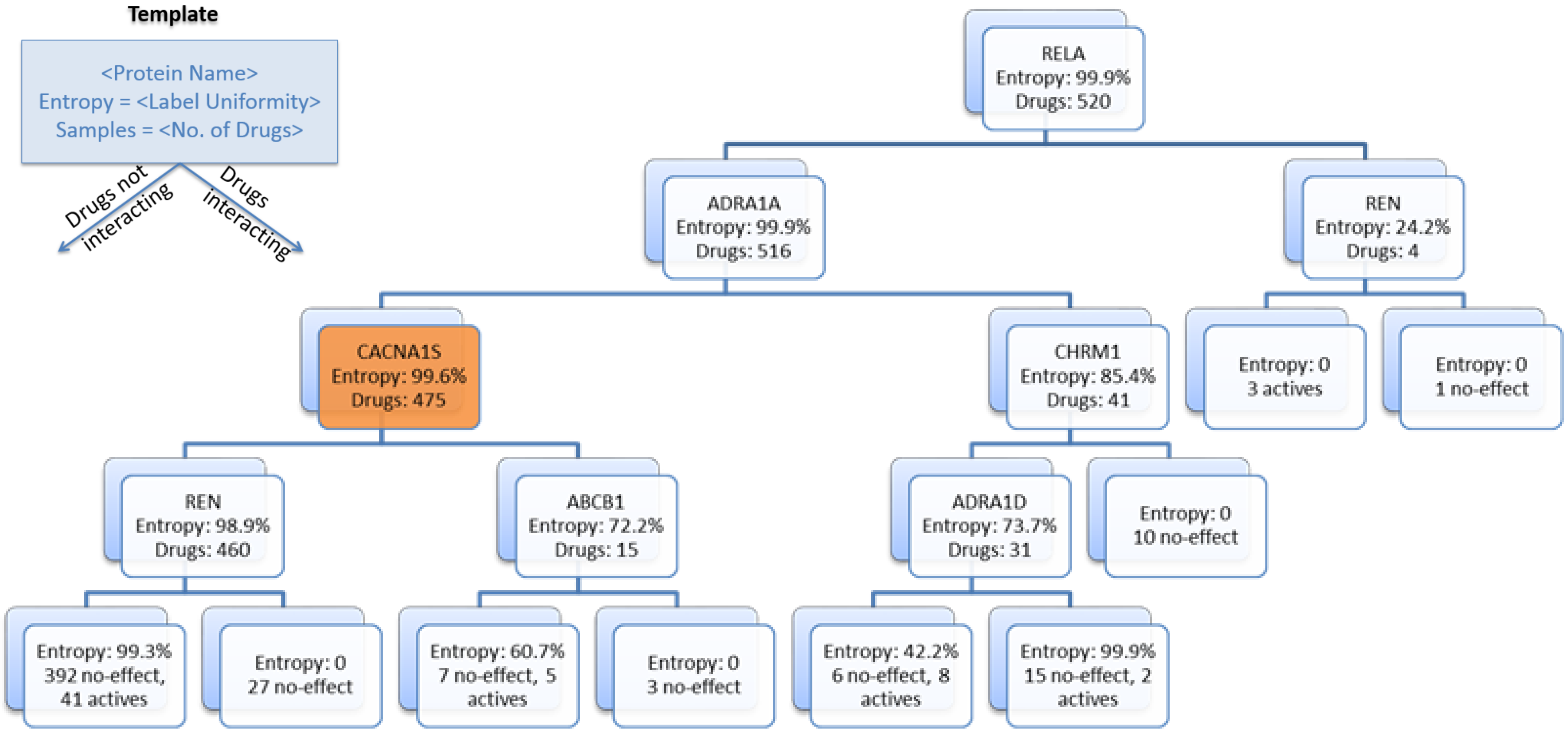

Template
<Protein Name>
Entropy = <Label Uniformity>
Samples = <No. of Drugs>
Drugs not interacting
Drugs interacting
DRD2 Entropy: 99.9% Drugs: 454
CACNA1C Entropy: 99.9% Drugs: 413
HTR2C Entropy: 88.9% Drugs: 41
ACE Entropy: 99.8% Drugs: 396
CACNA1B Entropy: 81.9% Drugs: 17
ADRA2B Entropy: 84.2% Drugs: 36
Entropy: 0 5 no-effect
PTGS2 Entropy: 99.8% Drugs: 387
Entropy: 0 9 no-effect
GRIK5 Entropy: 88.1% Drugs: 16
Entropy: 0 1 active
HTR7 Entropy: 81.9% Drugs: 34
Entropy: 0 2 no-effect
PPP3CA Entropy: 99.9% Drugs: 359
PTGS1 Entropy: 73.3% Drugs: 28
CACNA1H Entropy: 85.8% Drugs: 15
Entropy: 0 1 no-effect
ADRA1A Entropy: 80.6% Drugs: 33
Entropy: 0 1 no-effect
SLC6A4 Entropy: 99.9% Drugs: 358
Entropy: 0 1 active
Entropy: 0 8 no-effect
Entropy: 84.0% 19 no-effect, 1 active
Entropy: 83.1% 10 no-effect, 4 actives
Entropy: 0 1 no-effect
HTR1A Entropy: 84.5% Drugs: 29
HTR1A Entropy: 54.3% Drugs: 4
Entropy: 99.8% 305 no-effect, 273 actives
Entropy: 83.1% 10 no-effect, 4 actives
Entropy: 88.1% 21 no-effect, 7 actives
Entropy: 0 1 active
Entropy: 35.3% 1 no-effect, 2 actives
Entropy: 0 1 no-effect

Table 7: The tree structure of the targets of active drugs based on STITCH target information

We trained an entropy-minimization based decision tree to separate active drugs from inactive drugs using the algorithm due to Quinlan [136]. The targets of the drugs identified to be active in the screen were analyzed using an information entropy minimization strategy to build the following tree. The calcium channels, which are overrepresented are highlighted in orange. At each node, the entropy of the class labels (i.e. 'active' or 'inactive') are shown.



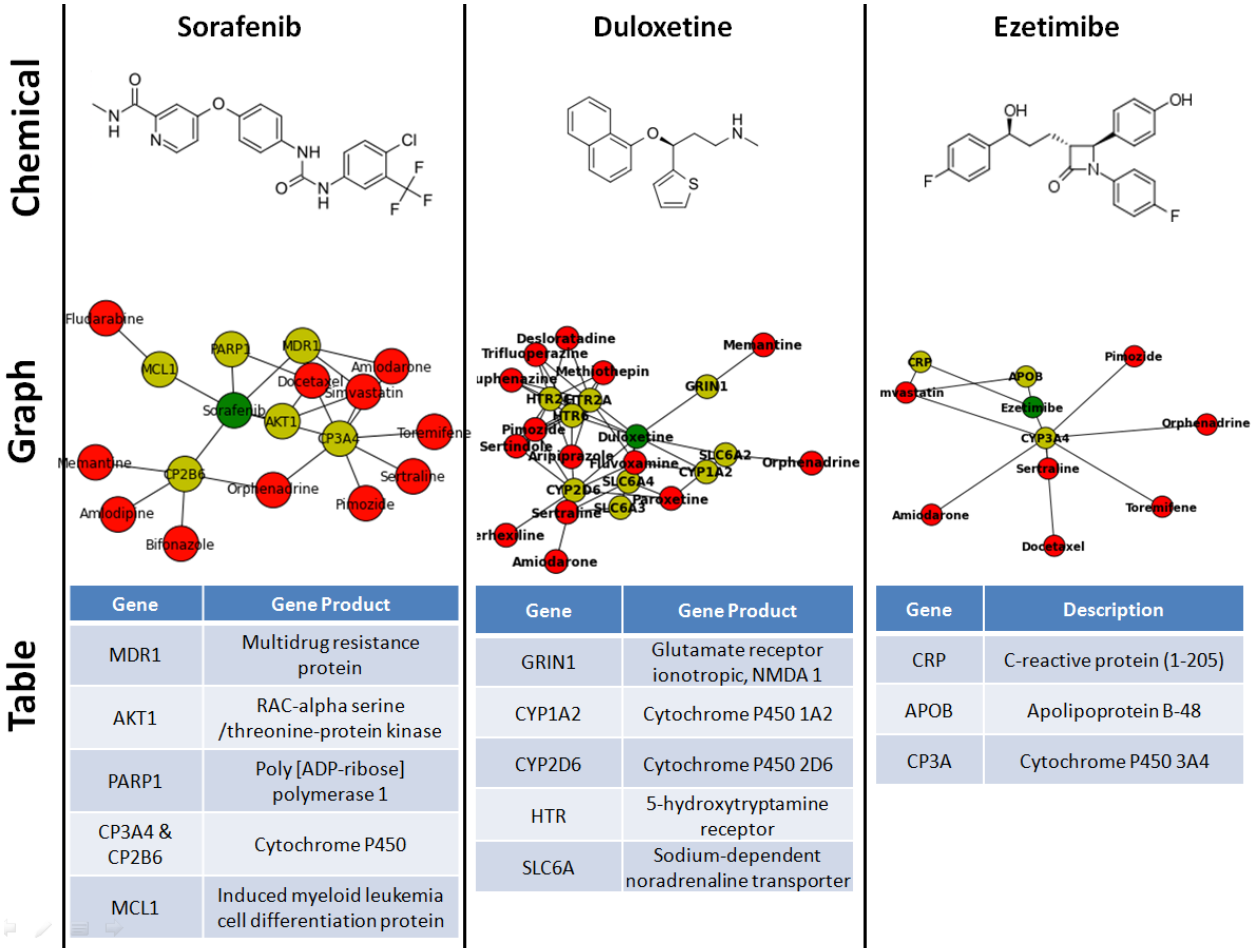
3.5 DIVERSIFICATION OF PROTECTIVE AGENTS AND PROPOSED MECHANISM

The computational analysis techniques that we have described in chapter 3.4 serve different goals, hence our results are divided among these different goals. Specifically, for compound diversification the STITCH-based results offer the best predictions for three reasons: Firstly, STITCH-based predictions offer repurposing possibilities since the predictions can be filtered to select for compounds that have been approved for use in humans. This has the advantage that it would reduce the time and cost of therapy development significantly when compared to the development of a novel chemical. Secondly, as the STITCH-based compounds are also filtered using the chemical structure based active/inactive classifier, the STITCH-based compounds have also been selected to possess chemical structures characteristic of desired activity; in addition to having at least one target in common with a drug already approved. Hence the STITCH-based compound predictions perform diversification of hits in both the proteomic space as well as the chemical space. Finally this method screens from a space of 300,000 chemicals as opposed to the 1,200 that were experimentally tested; therefore there is an advantage in using this computational method as it would not be feasible to brute force experimentally screen such a large chemical space. For these reasons, we have prioritized our STITCH-based target diversification method for discovering potential new actives.

3.5.1 Lead Diversification

We performed three lead diversification predictions using the methodology described above, and then manually screened the results reported in Appendix G to identify the top three potential candidates for ATZ aggregate inhibition to have a feasible number of experimentally testable predictions. Among the compounds in this list, we selected only the compounds that were already FDA approved for human use in order to enable repurposing and a quick translation of the discoveries we make. After sorting the compounds based on the number of targets (since the number of targets are all in the order of 10^1 we took this to indicate mostly how well studied these compounds are) and proceeded down the list one by one, manually analyzing the compounds for multiple criteria.

For each compound, if the compound was not an approved drug, we eliminated it and skipped to the next compound. Then, given the compound is approved, we looked at the targets that each compound shares with known actives, and looked for diversity. The idea here is that we do not want to have three compounds all very similar to each other. For example instead of having two compounds both targeting ATP binding cassette containing proteins, it is preferable to have one of the two target the cholesterol pathway. This way of having multitude of targets provides a way for each tested compound to provide information about another set of targets instead of testing the validity of the same targets multiple times. As such, we can interrogate a larger segment of the chemical/proteomic interaction landscape with fewer experiments and thus maximize the utility gained from the experiments. Finally we filtered out compounds that were already tested and shown to be protective – such as docetaxel, which ranked high in our list but it has already been shown to be protective in our PCL screen experiments. We selected the three best predictions in order to maximize the cost/benefit from a feasible number of follow-up experiments. These three predictions are interesting because they represent non-overlapping mechanisms of actions ranging from antineoplastic to antidepressive to blood cholesterol lowering drugs. Likewise their targets, and hence the targets that they share with the hits of the Prestwick screen are also entirely different allowing them to interrogate the various cellular processes that can be important for protection against ATD. If any of them fails in clinical testing whereas the others show neuroprotective activity, this is useful in enabling us to focus on the specific mechanism of action that is most relevant to ATD among the many different alternatives.



Gene	Gene Product
MDR1	Multidrug resistance protein
AKT1	RAC-alpha serine /threonine-protein kinase
PARP1	Poly [ADP-ribose] polymerase 1
CP3A4 & CP2B6	Cytochrome P450
MCL1	Induced myeloid leukemia cell differentiation protein

Gene	Gene Product
GRIN1	Glutamate receptor ionotropic, NMDA 1
CYP1A2	Cytochrome P450 1A2
CYP2D6	Cytochrome P450 2D6
HTR	5-hydroxytryptamine receptor
SLC6A	Sodium-dependent noradrenaline transporter

Gene	Description
CRP	C-reactive protein (1-205)
APOB	Apolipoprotein B-48
CP3A	Cytochrome P450 3A4

Figure 24: Possible repurposable candidates against ATD.

We show the three repurposable predictions against ATD: antineoplastic sorafenib, antidepressive duloxetine and anti-hyperlipidemic ezetimibe. Sorafenib is approved for use as an antineoplastic in humans against kidney cancer, advanced thryoid carcinoma, and finally advanced primary liver cancer. Duloxetine is approved for use in humans against major depressive disorder and generalized anxiety disorder. Ezetimibe is approved for use in humans to lower blood cholesterol levels by decreasing cholesterol absorption in the small intestine. The chemical structure of the drug is shown on top. The targets that are shared with drugs successful in the screen are shown in the middle as a graph; where the prediction drug is shown in green, the targets are shown in yellow and finally other drugs that were successful in the screen that share a target are shown in red. Finally, the name of the gene products of each gene shown in the graph is presented in the tables at the bottom of each column.

3.5.1.1 Sorafenib

The drug sorafenib has been approved for use as an antineoplastic in humans against primary kidney cancer, advanced primary liver cancer, and advanced thyroid carcinoma [137-141]. It has a well-characterized interaction profile in STITCH, with 62 targets listed in human. Since sorafenib is an antineoplastic drug, it shares targets mainly with other antineoplastics. We have discovered that six of these targets are shared with drugs that are of interest based on our experimental data. Furthermore, sorafenib has been classified as an active based on its chemical structure using our classifier that was trained on the chemical structure of the active drugs; therefore it matches all the characteristic chemical properties of the active chemicals. The interaction partners that sorafenib shares with other experimentally selected drugs are shown in Figure 24 along with other information.

3.5.1.2 Duloxetine

Duloxetine has been approved for use as an antidepressant in humans against major depressive disorder and generalized anxiety disorder [142-146]. Duloxetine is also well-characterized in STITCH, having 18 reported targets. There are 9 targets that are shared with the significantly neuroprotective drugs, which are shown in Figure 24 along with auxiliary information on compound structure and target names. Where sorafenib shared antineoplastic targets, duloxetine shares serotonin and sodium-dependent transporter targets with the ATD-protective drugs. Since the chemical structure based classifier has been applied to duloxetine as well, it clearly contains the chemical features that are important for the recognition of activity. The known activity, and therefore targets of duloxetine are different from those of sorafenib; therefore it represents a good alternative prediction for testing.

3.5.1.3 Ezetimibe

Ezetimibe is used to decrease blood cholesterol levels by decreasing absorption of cholesterol in the small intestines [147-149]. Ezetimibe has 16 targets reported in STITCH, two of them shared with simvastatin – which is also an anti-cholesterol drug, that ezetimibe is commonly co-administered with. Simvastatin has been shown to be highly active in alleviating ATD as it has a experimentally reported B-score of -2.60. Since ezetimibe reportedly shares two of its mechanistic targets (shown in Figure 24) and ezetimibe has also passed the chemical structure based classifier that filters out the drugs with inactive-like chemical features; it is also a good candidate for further experimental validation.

3.5.2 Mechanism Identification

We have analyzed the targets of the drugs that showed protective activity against ATD, as well as the pathways that these targets occurred in, to identify the mechanism of action of the successful compounds. We have focused on the target analyses, and not the pathway analyses; the reason being that in pathways we are operating on two levels of uncertainty: there are drugs which have unannotated targets; likewise there are targets with unannotated involvement in pathways. Missing annotations from both of these compound when looking at drug-to-pathway results presents a significant problem; the correction of which is a database curation work. When considering target enrichment, however, this limitation no longer exists since we have only one mapping (drug-to-target) and while there might be targets missed, the confidence level of the targets we do know, is significantly higher. Consequently we have found the most enriched pathway results to be inconsistent when evaluated with various different methods; whereas targets have given consistent results. Therefore we choose to focus on targets.

Furthermore, we wanted to focus on targets that showed a strong signal for being active in the mechanism. Therefore we wanted the effect to be reproducible. To ensure this, we looked for at least two inhibitor drugs that interact with each target in our enrichments. This ensures that the inhibitor effect of modulating the target is reproduced – by at least one other drug that interacts with this target.

When we compared the results of target enrichment analysis using the score based approaches and the entropy based approaches, with both methods performed with both STITCH-based and KEGG-Drug based data, there was one target that always showed significance: calcium channel.

3.5.2.1 Calcium channels

The calcium channels appear high among the top ten most enriched targets when compared using the enrichment score defined in methods in both the STITCH-based [119] and the KEGG-Drug based [133] approaches, as shown in Table 6 and Table 7. In addition, calcium channels also appear in the entropy minimization based decision trees that were learned on KEGG-Drug and STITCH data. This indicates four possibilities: (i) The direct interaction with calcium channels is responsible for protective effect, (ii) there is an indirect effect of interaction with calcium channels that leads to protection, (iii) there is a target similar to calcium channel, whose interactions are not captured in both of the two different databases that we used (the reason we used two different databases was to reduce this possibility) and that target is causing the response, (iv) there is no relationship between suppression of ATD and calcium channels; this is purely a random occurrence (the reason we used four method/database combinations was to reduce this possibility).

We used multiple databases to reduce the chances that there is an unknown interaction that dominates the activity that is missing from both databases. However, in analyses using both databases the calcium channels appeared high; hence this has a low likelihood. Likewise, the possibility that this is due to chance alone is unlikely when considering the fact that all four method/database combinations indicate that calcium channels are enriched among the targets of the drugs that showed ATD suppressive activity. Further determination of the possibility that calcium channel interaction leads to protection in ATD needs to be tested experimentally to be fully validated.

3.5.2.2 Adrenoceptors

Adrenoceptors appear enriched in the results shown in Table 6 and Table 7, in addition to calcium channels described in the previous subsection. Specifically, adrenoceptor alpha 2B (ADRA2B), adrenoceptor alpha 1A (ADRA1A), and adrenoceptor alpha 1D (ADRA1D) subtypes appear as important distinguishing targets in both tables. As with the calcium channels discussed in the previous subsection, these proteins have been selected due to the impact they have in differentiating the drugs that were active in modulating ATD disease phenotype and inactive drugs with no impact on the disease progression. Future experimental studies are required to confirm (or refute) the role of adrenoceptors in ATD or in other protein conformational disease contexts.

4.0 COMPUTATIONAL AND EXPERIMENTAL DETERMINATION OF NEUROPROTECTIVE THERAPEUTICS AGAINST HUNTINGTON'S DISEASE (HD)

In the following chapter, I will report our results on determining the mechanism of action of a diverse set of compounds which were found to be neuroprotective in a model of Huntington's disease using the LFM methodology we described in Chapter 2, specifically section 2.2.1 where we validated the use of LFM as descriptors. Then I will discuss our computational work for the identification of novel therapeutic candidates for use against Huntington's disease. Finally, I am going to report the follow up experimental work to test those predictions which I have participated in. Therefore the chapter is divided into three sections, with the first describing the mechanism identification work, the second discussing the LFM based predictive work and the third describing the experimental work.

4.1 MECHANISM OF ACTION OF DIVERSE NEUROPROTECTIVES

In order to discover new therapeutic candidates against HD, we have mapped the list of drugs known to be neuroprotectives to their targets in STITCH, identified their overlapping targets, and listed other drugs known to interact with the selected targets while having diverse activity profiles otherwise. The process is intended to generate target-based diversification when a small number of actives are available. Figure 30 presents a schematic description of the computational workflow designed for this process.

First we compiled a list of 24 known broadly neuroprotective drugs, as follows: Most of these drugs were reported by Wang and coworkers [92] after a two-stage screen where they first screened for inhibition of cytochrome c release from isolated mitochondria, and then tested the hits in a secondary assay for neuronal cell death inhibition. The authors tested 1040 drugs from the National Institute of Neurological Disorders and Stroke (NINDS) library, and found 21 drugs that successfully prevented the release of cytochrome c from isolated mitochondria when challenged with calcium and protecting neuronal cells from death. Of these 21 compounds, 15 were found to be effectively inhibiting neuronal cell death in follow-up assays, (with 6 having nanomolar IC50 values, termed Group I, and 9 having micromolar IC50 values, termed Group II) whereas 6 were found to be ineffective (Group III). Taking this information into account, we used the 15 that were effective in neuronal cell death inhibition as well as the cytochrome c release inhibition. We also compiled a list of all the drugs that were in clinical trials due to their neuroprotective effect to form Group IV. The structures, names and the groups of the entire set of compounds discussed here are shown in Figure 25. We collectively annotated the drugs in Groups I, II, and IV as the set of known neuroprotectives to inform our computational approaches.

The compounds in this list are traditionally annotated with a highly diverse set of therapeutic indications with no unifying theme:methazolamide is used for treatment of glaucoma, minocycline is an antibiotic, while azathioprine is used for immunosuppression. Therefore we set out to determine the mechanism of action of these drugs using computational methods. Specifically, we looked at the interaction information available about these drugs in three databases: known interactions in STITCH [45] and DrugBank [99], PMF predictions made as described in our previous work [97] on both of these databases, and SEA predictions [28] on ChemBL data [150]. We analyzed the information about these drugs using three different methods: direct set overlap of their targets, overlap between PMF predicted targets of these drugs, and 3D chemical similarity based search.

4.1.1 Overlap of Known Targets

To assess the mechanism of action of these drugs, we pooled together information on their known targets from two different databases: STITCH v3 [45] and DrugBank v3 [99][7]. Detailed information about our results can be found in **Error! Reference source not found.**. Not every single neuroprotective could be found in both of these databases: we could map 14 of these 24 in DrugBank which is a smaller database with less than 10,000 chemicals; whereas STITCH is a bigger database with data on about 300,000 different chemicals therefore we could identify 22 of the 24 drugs. The drugs that could be identified in each database were subsequently mapped to 50 known targets in DrugBank and 175 known targets in STITCH. The numbers of overlapping targets between all methods (*see below the* descriptions of the methods) across all databases are shown in Figure 27.

We looked at the overlap between these different targets, with the results shown in Figure 26. Specifically the only overlap among the known targets in DrugBank is between drugs melatonin and bepridil which share calmodulin, and minocyclin and doxyclyclin which share

[7] Please note that despite the fact that version 4 of both of these databases are currently available and the default for both of these resources, they have been released in 2014 whereas we conducted this study in 2013 therefore version 3 was the latest version available at the time.

30S ribosomal proteins S4 and S9 as targets. There are no other overlapping targets in DrugBank.

We identified the overlap between all 731 targets in six different ways, specifically the overlap between (i) PMF predicted targets in DrugBank, and 3D predicted targets in DrugBank, (ii) the DrugBank known targets and 3D predicted targets, (iii) STITCH and SEA, (iv) DrugBank and SEA, (v) DrugBank and STITCH, and finally (vi) DrugBank, STITCH and SEA. The results are shown in Figure 28.

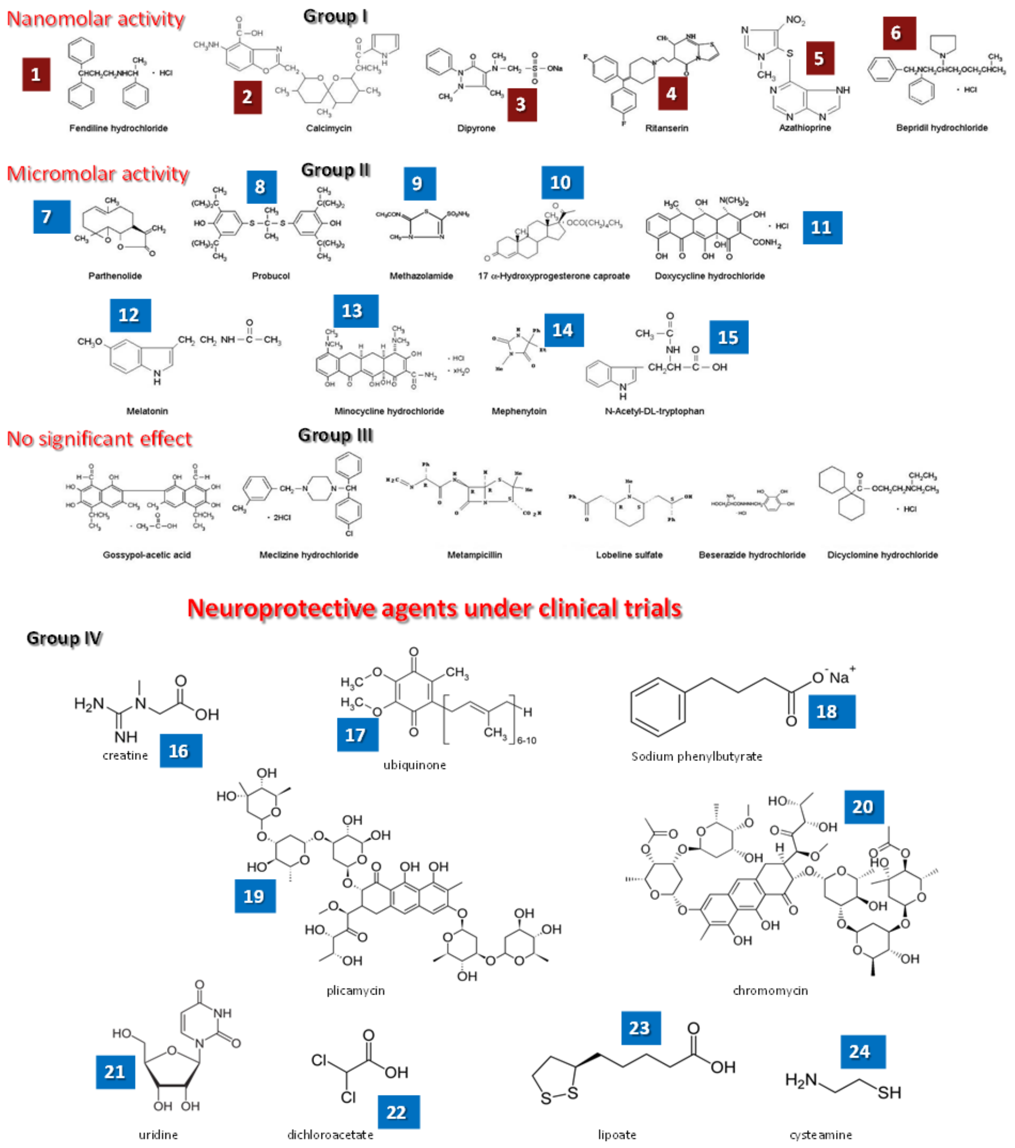
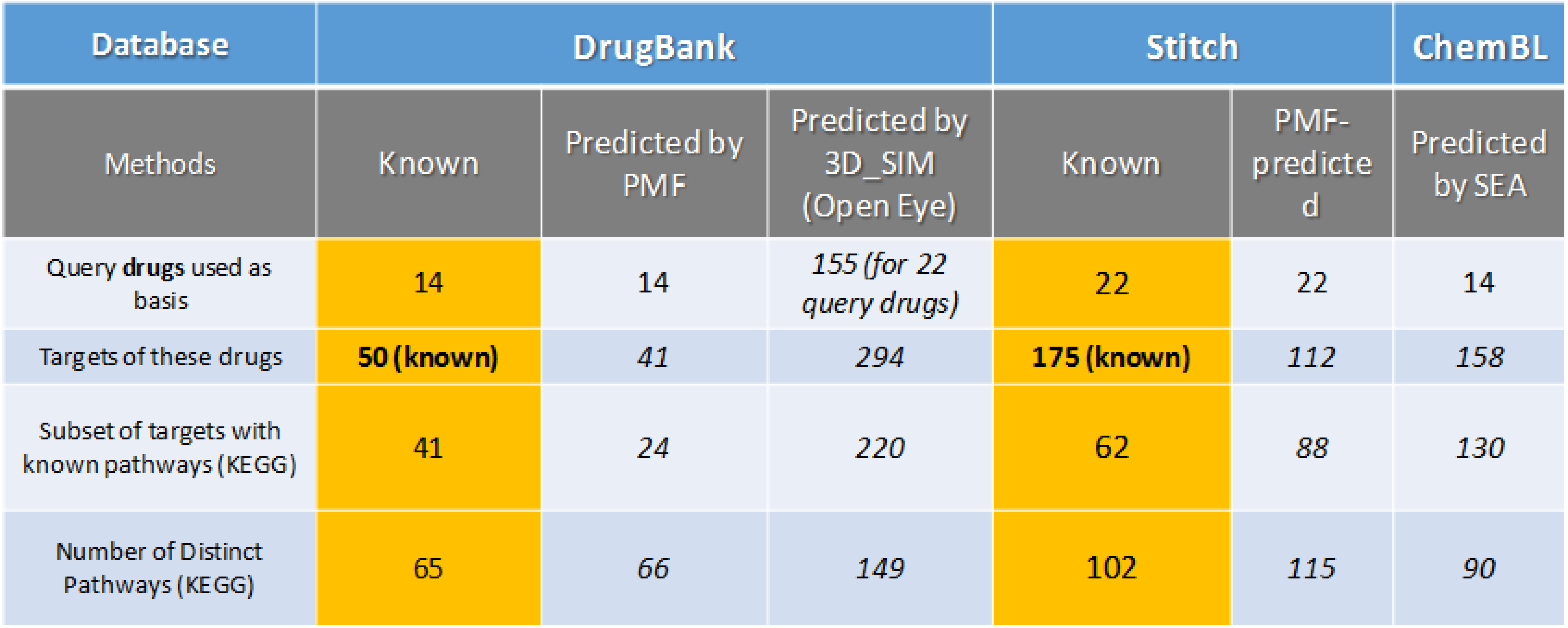


Figure 25: Neuroprotective drugs used to inform computational method

From an NINDS library of 1040 drugs, 21 were found to be inhibitors of cytochrome c release in isolated mitochondria which were then tested for their ability to inhibit neuronal cell death in a HD model cell line and 6 were found to have IC50 values in the nanomolar range (Group I), 9 were found to have IC50 values in the micromolar range (Group II), whereas 6 were shown not to have a significant neuroprotective effect (Group III) in previous work [92]. We also compiled a list of 9 neuroprotective drugs currently in clinical trials (Group IV). We used groups I, II, and IV (24 total) as neuroprotectives to inform the computational method.

Table 8: Summary results of our analysis of target data for the known neuroprotectives

The table below summarizes the findings from our computational assessment of the data available on the targets of the 24 known neuroprotectives. Of these 24 compounds, 14 were identified in DrugBank [99], 22 were identified in STITCH [45], and 14 were amenable for query on the Similarity Ensemble Approach (SEA) server [28].



Database	DrugBank			Stitch		ChemBL
Methods	Known	Predicted by PMF	Predicted by 3D_SIM (Open Eye)	Known	PMF-predicted	Predicted by SEA
Query **drugs** used as basis	14	14	*155 (for 22 query drugs)*	22	22	14
Targets of these drugs	**50 (known)**	*41*	*294*	**175 (known)**	*112*	*158*
Subset of targets with known pathways (KEGG)	41	*24*	*220*	62	*88*	*130*
Number of Distinct Pathways (KEGG)	65	*66*	*149*	102	*115*	*90*

4.1.2 Chemical Similarity Comparison

We performed chemical similarity comparison and looked for the targets of the compounds that are highly similar to the known neuroprotectives. We used the Similarity Ensemble Approach (SEA) to predict 158 ChemBL targets based on chemical similarity, and we also used 3D similarity to identify 294 targets in DrugBank. Specifically, for our 3D similarity calculation we compared the structures of the 22 that we successfully mapped to STITCH compounds to all the chemical structures in DrugBank using the ROCS 3D small molecule structural similarity methods developed by OpenEye™ [151]. Briefly stated, this method represents a given chemical structure with Gaussians that are centered on each atom. There are two different types of Gaussians: 'colorless' for simple steric overlap and 'colored' Gaussians where each color represents a different physico-chemical property (positive charge, negative charge, hydrophobicity, etc). The overlap among these Gaussians allows us to numerically evaluate the similarity between two compounds in terms of their shape as well as their electrostatic properties.

For the 22 drugs that could be mapped to the STITCH database, we identified their chemical structures from the data in STITCH, and used those structures to search for their analogues in DrugBank. We identified a total of 155 drugs that were similar to these 22 chemical structures. These drugs in turn mapped to 294 targets, 220 of which had KEGG pathways. These 220 KEGG pathways were replicated among them, and therefore they matched to 149 unique pathways.

4.1.3 LFM Predictions

We used LFM in predictive function in order to discover the unknown interactions of the drugs that we could map to DrugBank and STITCH. The results reported in Chapter 2 demonstrate that the latent factor models can function remarkably as predictors of drug-target interaction. In this biomedical project where we have a set of drugs with largely unexplained mechanisms of action for their reported neuroprotective activity, it is necessary to identify any potential targets of these drugs that might explain this novel activity. Therefore we trained latent factor models on both DrugBank and STITCH, and used them to predict the unknown interactions of the drugs of interest.

Our results in STITCH show that there was one target, Lysine-specific demethylase (PHF2), which was predicted to be the interaction partner for seven drugs: bepridil, parthelonide, dioxycycline, mephenytoin, N-acetyl DL Tryptophan, ubiquinone and cysteamine. PHF2 is a lysine demethylase that functions only after activation by PKA, acts on both histones and non-histone proteins, and is known to form a complex with and mediate the methylation/demethylation of ARID5B [152]. It is important to note that the interaction between these seven drugs and PHF2 is predicted using the LFM we built, and now reported to be known therefore it would be important to experimentally validate this interaction as a first step to further understanding. However it is important to note that this interaction might point to a key new neuroprotective mechanism when it is considered that in the literature it has been reported that PKA has a role in preventing the induction of apoptosis in astrocytes [153], where it has been shown that an agent that activates the PKA pathway (octadecaneuropeptide) leads to protection from apoptosis. Since PHF2 activity is only possible after PKA activation, and since PKA based apoptotic protection in brain cells has been previously demonstrated in the literature, this target overlap result serves as an interesting precursor for further study.

Based on STITCH data, HDAC was shared as a target by compounds melatonin and N-acetyl DL Tryptophan. This was an important finding because HDAC inhibitors have been previously reported to ameliorate disease phenotype [154-156], transport deficit [157], and motor deficit [158] in Huntington's disease models. These publications indicate that targeting HDAC ameliorates disease phenotype in HD and therefore point out the potential significance of this predictive finding.

4.1.4 Pathway Mapping of Targets

We have analyzed the targets that we identified for their roles in known pathways. Specifically, we mapped each of the drug targets to KEGG pathways and identified the pathways with the highest number of drugs acting on them. Calcium signaling pathway emerged as a significant pathway of interest from this study owing to the multitude of known and predicted targets of the known neuroprotectives in this pathway, in addition to the broad literature support for the role of calcium in HD pathophysiology. This section is dedicated to our findings on this pathway in detail.

There are two known neuroprotectives, bepridil (drug #6) and melatonin (drug #12) that are known to target two proteins in the calcium signaling pathway: voltage-dependent calcium channel subunit α1 (CaV1) which is targeted by bepridil, and calmodulin (CALM) which is reportedly targeted by both bepridil and melatonin. The calcium channel subunit α1 has been reported to be singularly sufficient to conduct Ca^{2+} transfer across the membrane [159]. More generally bepridil is known to be a calcium channel blocker with the latest version of DrugBank (v4.2) reporting interactions between bepridil and a large variety of calcium channel subunits, indicative of its broad calcium channel blocker activity. Traditionally calcium channel blockers are used for antihypertensive function in the clinic, therefore the role of bepridil in HD remains unclear at first, yet a deeper look reveals that this is actually a highly interesting discovery.

The role of calcium channels in ataxia mechanistically caused by an expanded polyglutamine repeat have been established by previous studies: an expanded CAG repeat in human calcium channel subunit α1A has been reported to be the causal mutation for spinocerebellar ataxia type 6 (SCA6) in humans [160]. There are multiple significant similarities between the two diseases. They are both late onset neurodegenerative disorders that manifest in uncontrolled muscle movements the characteristic *chorea* movement in Huntington's disease is highly akin to uncontrolled movement in SCA6. Furthermore expanded polyglutamine repeats have been reported to be causal in multiple other late onset neurodegenerative diseases: spinocerebellar ataxia type 1 (SCA1) [161], spinocerebellar ataxia type 2 (SCA2) [162], spinocerebellar ataxia type 3/Machado-Joseph disease (SCA3/MJD) [163], spinobulbar muscular atrophy (SBMA) [164], dentatorubral-pollidoluysian atrophy/Haw-Rover syndrome (DRPLA/HRS) [165]. These results indicate that polyglutamine repeats are causal for multiple neurodegenerative diseases that have similar clinical presentations to HD and the further finding that one of those polyglutamine expansions has been localized to the calcium channel implicates the role of modulation of calcium. Hence the neuroprotective role of the calcium channel blocker bepridil to be due to its calcium channel interactions is reasonable when evaluated within the context of previous findings.

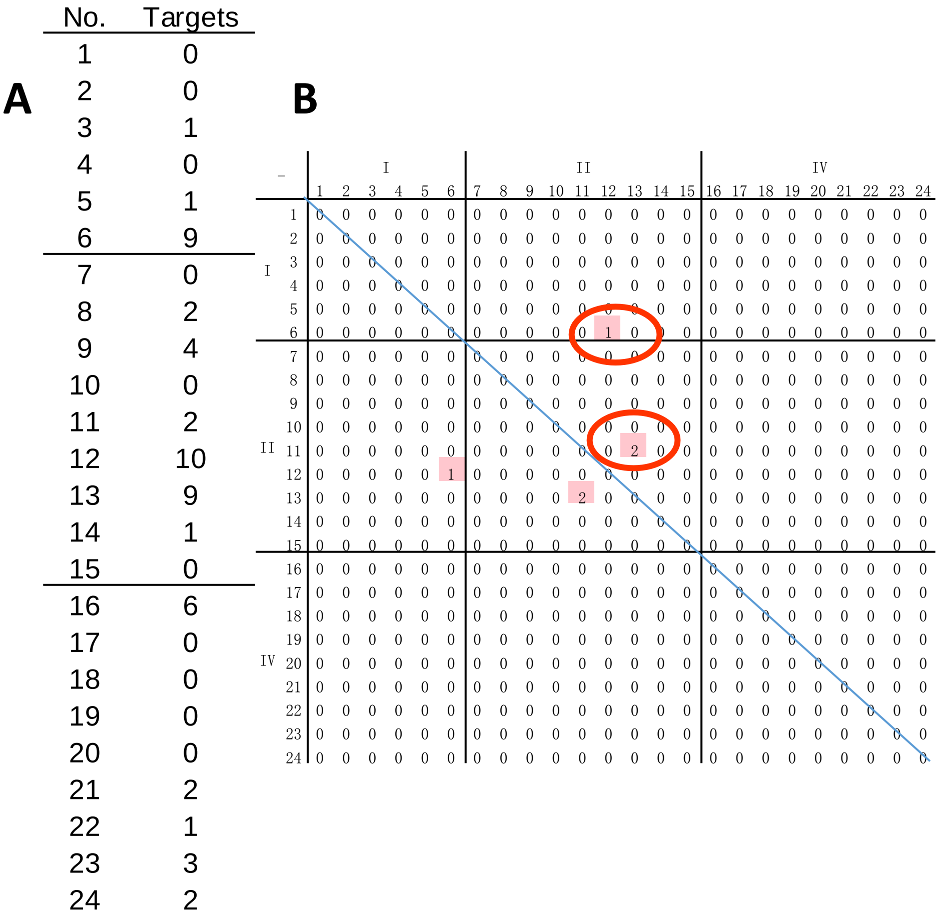
Bepridil is reported to interact with calmodulin in a Ca^{2+}-dependent manner [166], along with melatonin which is also reported to be a calmodulin inhibitor [167]. Furthermore, melatonin’s activity on calmodulin is implicated in its rapid (<1 min) and transient (5-6h) effect of ROS generation in cells [168]. These findings implicate that the modulation of the calcium signaling pathway could be important for modulating Huntington’s disease.

Furthermore, despite having only two drugs and targets implicated in the calcium signaling pathway, we have identified a large set of predicted interactions involving proteins in this pathway. This strong predicted role for calcium in the mechanism of Huntington’s disease is supported by previous findings reported in the literature.

On the tissue level, researchers have discovered that in post-mortem brain specimens of patients who have died from Huntington’s disease there is a substantial loss of neurons containing the calcium-binding protein calbindin 28K [169]. These calbindin containing neurons, as well as the striatal component that they are located in are reported to be particularly reported in Huntington’s disease therefore the observed effect seems to be specific to Huntington’s disease instead of a general response to the neurodegenerative process.

On the cellular level, researchers have identified that mutant huntingtin directly interacts with neuronal mitochondrial membranes and leads to mitochondrial membrane depolarization at lower calcium loads accompanied with lower membrane potential [170]. Furthermore, these mitochondrial calcium abnormalities have been observed months before the presentation of pathological or behavioral abnormalities. The researchers demonstrate that the mitochondria in lymphoblasts from HD patients have a significantly reduced Ca^{2+} retention capacity (on average 64 nmol/mg protein in HD patients versus 146 nmol/mg protein in healthy control). Despite the fact that the exact long-term functional consequences of the mitochondrial Ca^{2+} defect are unknown, the fact remains that calcium related pathways are significantly defective in HD patients.

In summary, the multitude of predicted and known interactions with calcium signaling pathway shown in Figure 29 are grounded when considering the role of calcium widely reported in the literature on Huntington's disease.



A

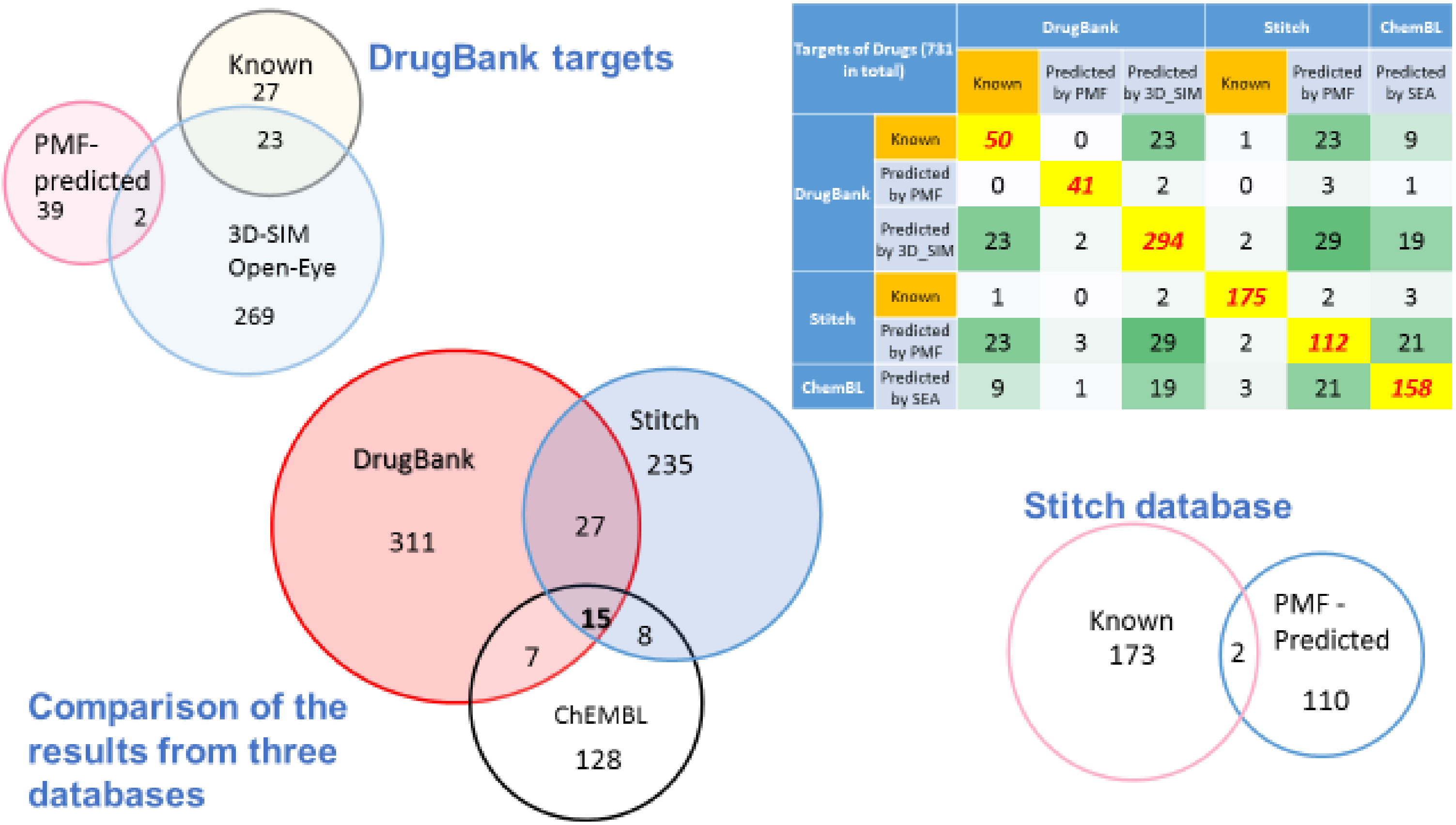
No.	Targets
1	0
2	0
3	1
4	0
5	1
6	9
7	0
8	2
9	4
10	0
11	2
12	10
13	9
14	1
15	0
16	6
17	0
18	0
19	0
20	0
21	2
22	1
23	3
24	2

B

		I						II									IV								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
I	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
II	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IV	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 26: The overlap between the known targets of the neuroprotective drugs in DrugBank

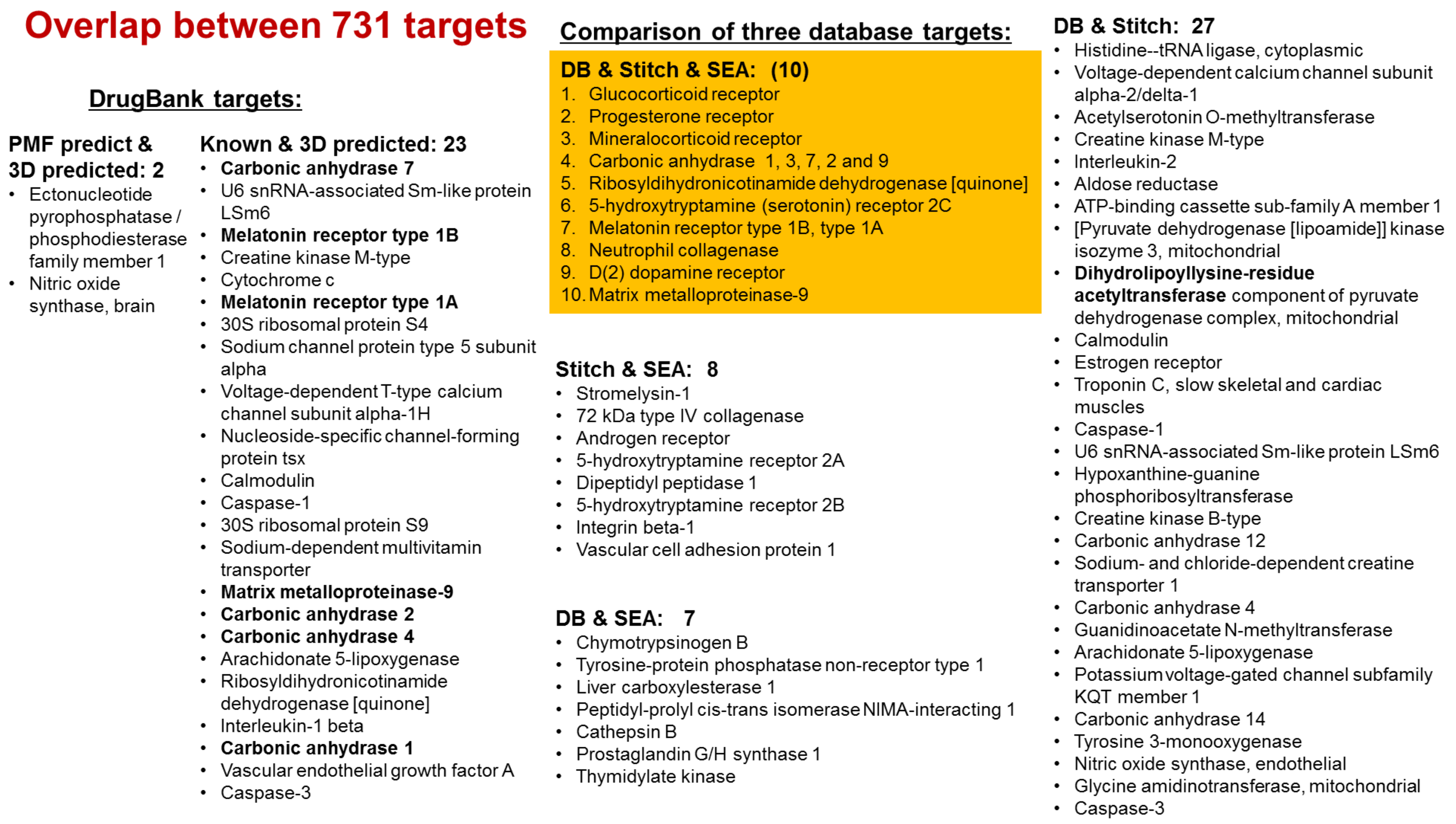
The drugs are separated into groups and indexed as in Figure 25, with the known target information displayed in two ways: Panel a shows the number of targets identified for every single one of the 14 (out of 24) compounds that could be found in DrugBank. This panel shows which 14 of the 24 were mapped to DrugBank as well as the number of targets for each of them. Panel b shows the overlap between the targets of these drugs (and is therefore symmetric). This panel clearly shows that there is very little known target overlap in DrugBank, with only drugs 12 (melatonin) and 6 (bepridil) sharing one target (calmodulin), and drugs 13 (minocycline) and 11 (doxyclycline) sharing two targets (30S ribosomal proteins S4 and S9). There are no overlapping targets in DrugBank other than these.



Targets of Drugs (731 in total)		DrugBank			Stitch		ChemBL
		Known	Predicted by PMF	Predicted by 3D_SIM	Known	Predicted by PMF	Predicted by SEA
DrugBank	Known	50	0	23	1	23	9
	Predicted by PMF	0	41	2	0	3	1
	Predicted by 3D_SIM	23	2	294	2	29	19
Stitch	Known	1	0	2	175	2	3
	Predicted by PMF	23	3	29	2	112	21
ChemBL	Predicted by SEA	9	1	19	3	21	158

Figure 27: The summary visualization of the overlap between the 731 targets identified for neuroprotective drugs

There are a total of 731 targets that were identified in the three databases that we considered using the five methods that we used to process the targets of these drugs. The diagonal terms on the matrix show the number of targets identified using that method, whereas the off-diagonal terms are the overlap of targets among the two specified method/databases (since overlap is symmetric, the matrix is also symmetric). The Venn diagrams show the overlap between the three main databases used, and internally the overlap between the target identification methods within STITCH and DrugBank visually.



Overlap between 731 targets

DrugBank targets:

PMF predict & 3D predicted: 2

- Ectonucleotide pyrophosphatase / phosphodiesterase family member 1
- Nitric oxide synthase, brain

Known & 3D predicted: 23

- **Carbonic anhydrase 7**
- U6 snRNA-associated Sm-like protein LSm6
- **Melatonin receptor type 1B**
- Creatine kinase M-type
- Cytochrome c
- **Melatonin receptor type 1A**
- 30S ribosomal protein S4
- Sodium channel protein type 5 subunit alpha
- Voltage-dependent T-type calcium channel subunit alpha-1H
- Nucleoside-specific channel-forming protein tsx
- Calmodulin
- Caspase-1
- 30S ribosomal protein S9
- Sodium-dependent multivitamin transporter
- **Matrix metalloproteinase-9**
- **Carbonic anhydrase 2**
- **Carbonic anhydrase 4**
- Arachidonate 5-lipoxygenase
- Ribosyldihydronicotinamide dehydrogenase [quinone]
- Interleukin-1 beta
- **Carbonic anhydrase 1**
- Vascular endothelial growth factor A
- Caspase-3

Comparison of three database targets:

DB & Stitch & SEA: (10)

1. Glucocorticoid receptor
2. Progesterone receptor
3. Mineralocorticoid receptor
4. Carbonic anhydrase 1, 3, 7, 2 and 9
5. Ribosyldihydronicotinamide dehydrogenase [quinone]
6. 5-hydroxytryptamine (serotonin) receptor 2C
7. Melatonin receptor type 1B, type 1A
8. Neutrophil collagenase
9. D(2) dopamine receptor
10. Matrix metalloproteinase-9

Stitch & SEA: 8

- Stromelysin-1
- 72 kDa type IV collagenase
- Androgen receptor
- 5-hydroxytryptamine receptor 2A
- Dipeptidyl peptidase 1
- 5-hydroxytryptamine receptor 2B
- Integrin beta-1
- Vascular cell adhesion protein 1

DB & SEA: 7

- Chymotrypsinogen B
- Tyrosine-protein phosphatase non-receptor type 1
- Liver carboxylesterase 1
- Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1
- Cathepsin B
- Prostaglandin G/H synthase 1
- Thymidylate kinase

DB & Stitch: 27

- Histidine--tRNA ligase, cytoplasmic
- Voltage-dependent calcium channel subunit alpha-2/delta-1
- Acetylserotonin O-methyltransferase
- Creatine kinase M-type
- Interleukin-2
- Aldose reductase
- ATP-binding cassette sub-family A member 1
- [Pyruvate dehydrogenase [lipoamide]] kinase isozyme 3, mitochondrial
- **Dihydrolipoyllysine-residue acetyltransferase** component of pyruvate dehydrogenase complex, mitochondrial
- Calmodulin
- Estrogen receptor
- Troponin C, slow skeletal and cardiac muscles
- Caspase-1
- U6 snRNA-associated Sm-like protein LSm6
- Hypoxanthine-guanine phosphoribosyltransferase
- Creatine kinase B-type
- Carbonic anhydrase 12
- Sodium- and chloride-dependent creatine transporter 1
- Carbonic anhydrase 4
- Guanidinoacetate N-methyltransferase
- Arachidonate 5-lipoxygenase
- Potassium voltage-gated channel subfamily KQT member 1
- Carbonic anhydrase 14
- Tyrosine 3-monooxygenase
- Nitric oxide synthase, endothelial
- Glycine amidinotransferase, mitochondrial
- Caspase-3

Figure 28: The 731 targets of the neuroprotective drugs reported simultaneuously in multiple databases

The targets of the neuroprotective compounds that overlap between the various databases are shown based on the overlapping resources. The targets that are reported in multiple data sources and/or by multiple methods are more likely to be correct targets and therefore they are more convincing in their potential to be of significance to neuroprotection.

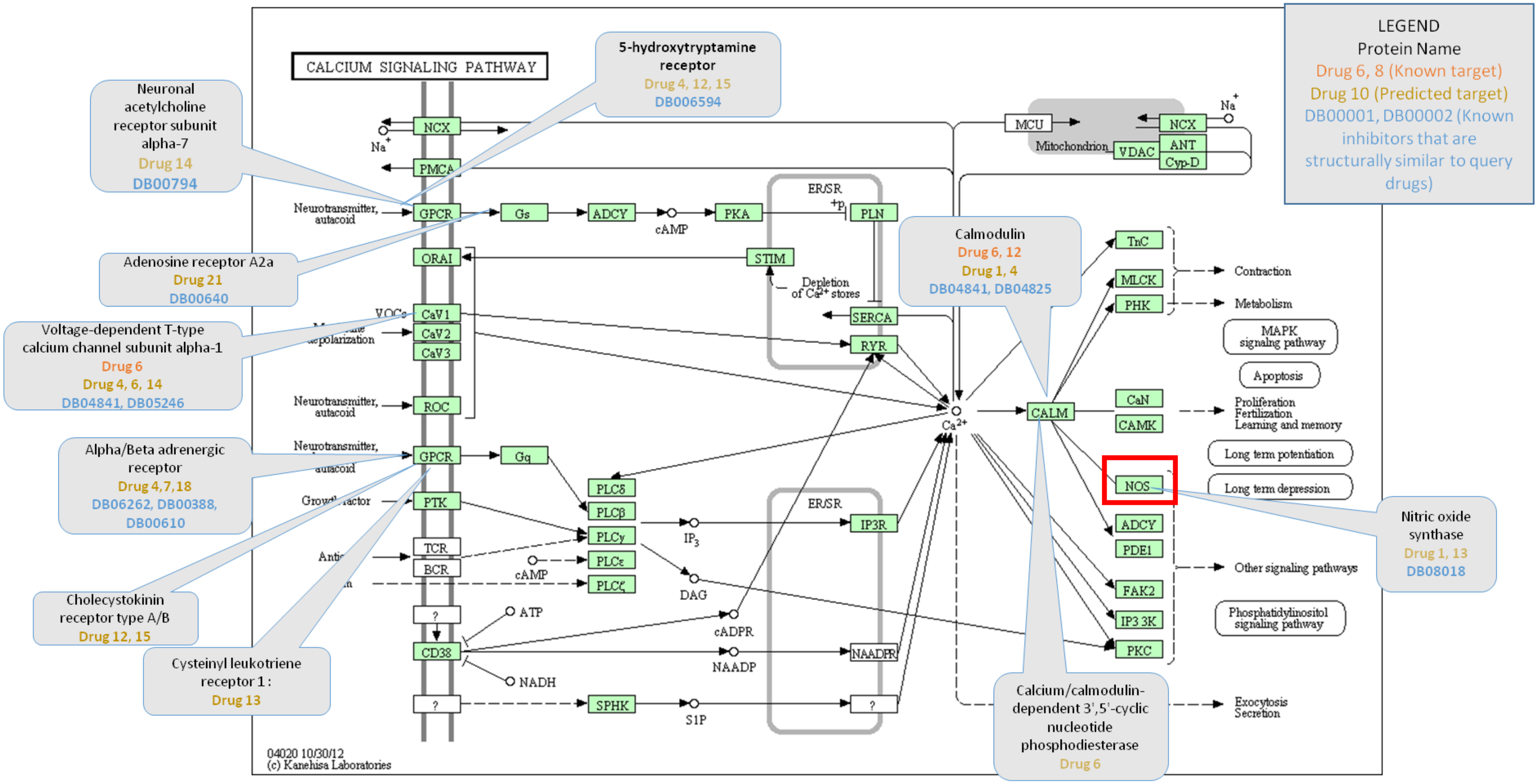


Figure 29: The calcium signaling pathway marked with the targets of neuroprotectives

The figure displays the calcium signaling pathway as reported in KEGG, annotated with the targets of known neuroprotectives. As reported in the legend on the figure, the annotations in grey boxes first report the full name of the target being highlighted with the pointer. The indices of the targeting known neuroprotective drugs are then reported in red if there exist any (where the indices are as in Figure 25). Then the known neuroprotectives that are predicted to interact with the target by any of the prediction methods (SEA or PMF) are shown in yellow. Finally, the last line reports any known inhibitors that are structurally similar to the known neuroprotective (i.e. the ‘query drug’). Nitric oxide synthase is highlighted in red because of its relation to our experimental findings.

4.2 LFM-BASED ACTIVE DIVERSIFICATION

We mapped the drugs in Groups I, II, and IV to form a set of known neuroprotective seed compound set. We then mapped these 24 known neuroprotective drugs to 349 targets in humans in STITCH v3.1 with a 90% cutoff (STITCH lists drug-target interactions with a 0% to 99.9% confidence score, and annotates those above 90% as being very high confidence interactions). Of these 349 targets, 32 were overlapping targets of two or more drugs. These targets were sorted based on the number of known neuroprotectives targeting them, with the target shared by the highest number of known neuroprotectives ranking first. For each target on this list, all the other drugs interacting with it (i.e. any drug not in the original set of 24 neuroprotective drugs but known to interact with a target of interest) were selected and sorted for maximal difference to the known neuroprotectives that were used to select that target. Specifically, we have trained a latent variable model on the drug-target interactions in STITCH using the method described in [97]. The drugs were sorted for maximal distance in the latent variable space, which represents maximal dissimilarity in their interaction profile to the neuroprotectives. The motivation here is that by selecting drugs that are targeting the most frequently shared targets of the known neuroprotectives that are otherwise as dissimilar as possible in their interaction profile, we will achieve two objectives (i) diversify the known neuroprotective drugs, (ii) potentially gain an insight into the mechanism of action. I am going to explain both of these points in more detail.

Among all the drugs that are known to be neuroprotectives, it is highly unlikely that every single one operates on a different mechanism of action. Therefore, it is quite likely that the contrary is true – i.e. that there are a smaller number of mechanisms, shared by a subset of the drugs in our initial seed set. Therefore looking at the list of most frequently shared targets gives us a ranked list of the targets which are most likely to be important for the mechanisms of action. Operating on this shortlist, which reduces the number of likely culprits in the mechanism of action from the set of all targets to 32, we have a higher likelihood of identifying compounds with high neuroprotective activity. Looking at the drugs of these 32 targets that are as dissimilar as possible to the known neuroprotectives is useful in helping us spend our limited experimental resources on as diverse a set of neuroprotective candidates as possible.

Testing the drugs interacting with these targets but are as diverse as possible, simultaneously allows us to gain an insight into the mechanism of action. If there are a multitude of drugs that have been identified to work as neuroprotectives, all of which have been selected based on their interaction with a particular target, then that target is implicated for a key role in achieving neuroprotective activity. The identification of targets through testing of compounds for activity alone is useful because it enables the use of high levels of automation in handling compounds and running large compound library screens to provide useful information about target identification, which is traditionally not as amenable to high throughput methods. Deduce a strong hypothesis about the mechanism of action from running compounds alone, is ground-breaking because what is arguably the least streamlined step of the drug discovery step (target identification) can be done as high-throughput amenable as the screening step because our method enables testing mechanism of action using compounds alone.

Specifically, if compounds that are maximally dissimilar to the known neuroprotectives, but otherwise share the target of interest (shared by as many neuroprotectives as possible), also show the desired activity then this implicates the target. Conversely, if compounds that are maximally similar but do not share the target of interest do not show the desired activity, this further implicates the target as being integral to the mechanism of action. Hence an informed hypothesis about the mechanism of action can be acquired by testing compounds alone, with the added benefit that compounds with the desired activity of interest can be identified in the process.

When performing the analysis, each subsequent drug added to the list of predictions after the first one was chosen to be maximally distant from the previously selected predictions as well as the drugs that supported the hypothesis. This adaptive strategy prevents all the tests from focusing on a set of highly similar compounds, and ensures that the selected compounds all sample diverse parts of the chemical space. The results of this method are shown in Appendix I, where the target selected as the hypothesis is indicated, followed by the support drugs that led to the selection of that target. In the following rows, each row starts with the name of the chemical that is the top hypothesis, then the average similarity to all the PubChem CIDs that mapped to the support drug with the reported name on top of the column are reported. The rightmost column contains the average of the similarity scores to all of the support drugs. For every compound that comes after the first hypothesis compound, the similarity to all the prior compounds are also taken into account, in order to achieve adaptive selection of compounds that are not too similar to each other.

We chose 18 drugs from among all the compounds selected for 6 targets, based on feasibility of acquisition and experimental testing. The drugs we selected and the targets that informed their selection are shown visually in Figure 31. To achieve the objective of forming an informed hypothesis would have required much more experimentation, testing at least 10 compounds for every target of interest. Since we were constrained by experimental feasibility and therefore could not conduct experiments with a satisfactory number of compounds, our objective was only to identify new neuroprotectives. We followed up on the selected compounds with a phenotypic cell based toxicity assay for validation of their neuroprotective activity.

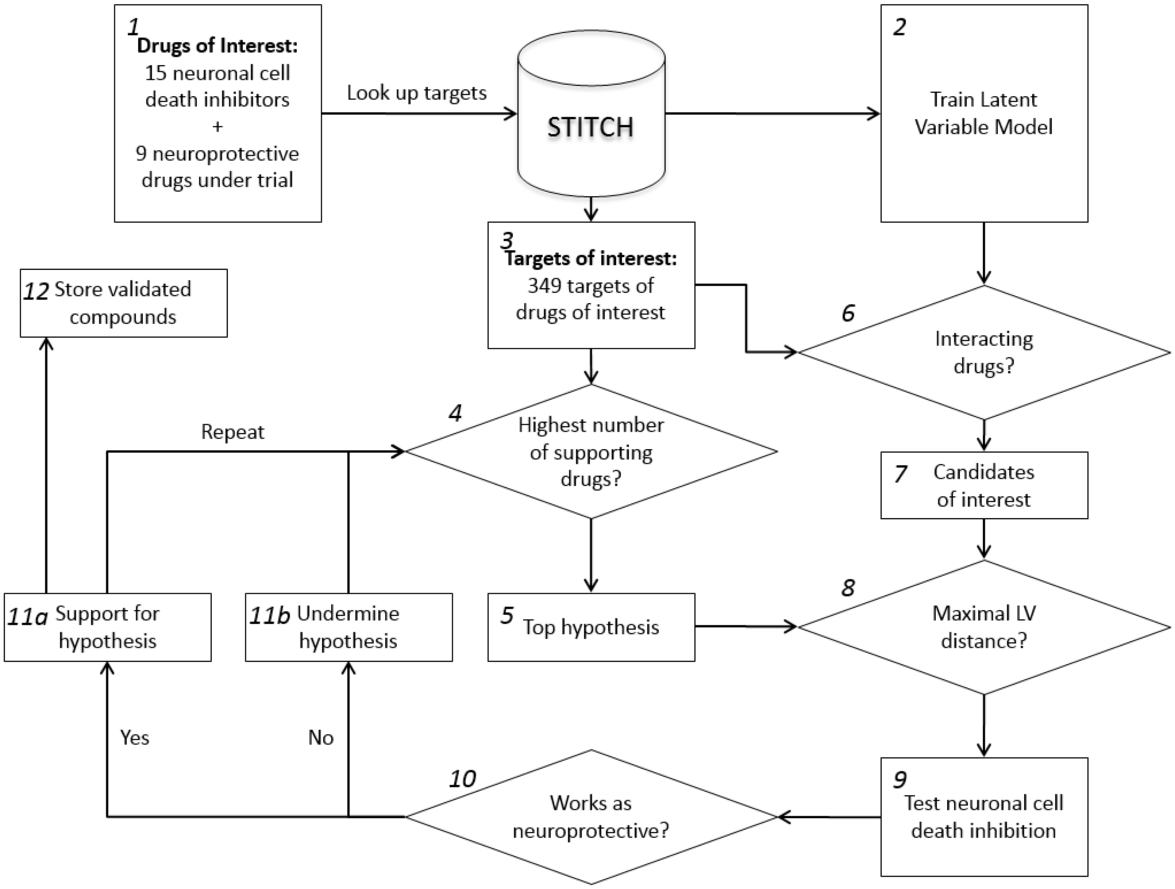

1 Drawings of Interest:
15 neuronal cell death inhibitors
+
9 neuroprotective drugs under trial
Look up targets
STITCH
2 Train Latent Variable Model
3 Targets of interest:
349 targets of drugs of interest
12 Store validated compounds
6 Interacting drugs?
4 Highest number of supporting drugs?
Repeat
7 Candidates of interest
11a Support for hypothesis
11b Undermine hypothesis
5 Top hypothesis
8 Maximal LV distance?
Yes
No
10 Works as neuroprotective?
9 Test neuronal cell death inhibition

Figure 30: Computational workflow for active diversification

The computational workflow described above was used for diversification of compounds with neuroprotective activity, however it is broadly applicable for any desired activity type of interest. (1) The computational workflow starts with known active compounds, in this case we used 15 neuronal cell death inhibitors identified by Wang et al, 2008 and 9 compounds currently in clinical trials. (2) The STITCH interaction dataset is used to train latent variable (LV) model that describes each drug and target's interaction characteristics. (3) The known targets of the drugs of interest are looked up from STITCH. (4,5) The target(s) that interact with the highest number of drugs of interest selected as the top hypothesis. (6,7) The drugs known to interact with the targets of interest are extracted from STITCH, these constitute the repurposable candidates of interest. (8) The candidates of interest are sorted according to maximal LV distance (i.e. maximum dissimilarity) to the drugs of interest that interact with the selected target of interest. (9,10) The top candidate(s) tested for desired activity, in this case neuronal cell death inhibition. (11) The successful results feedback into the algorithm, (12) the successful results are stored.

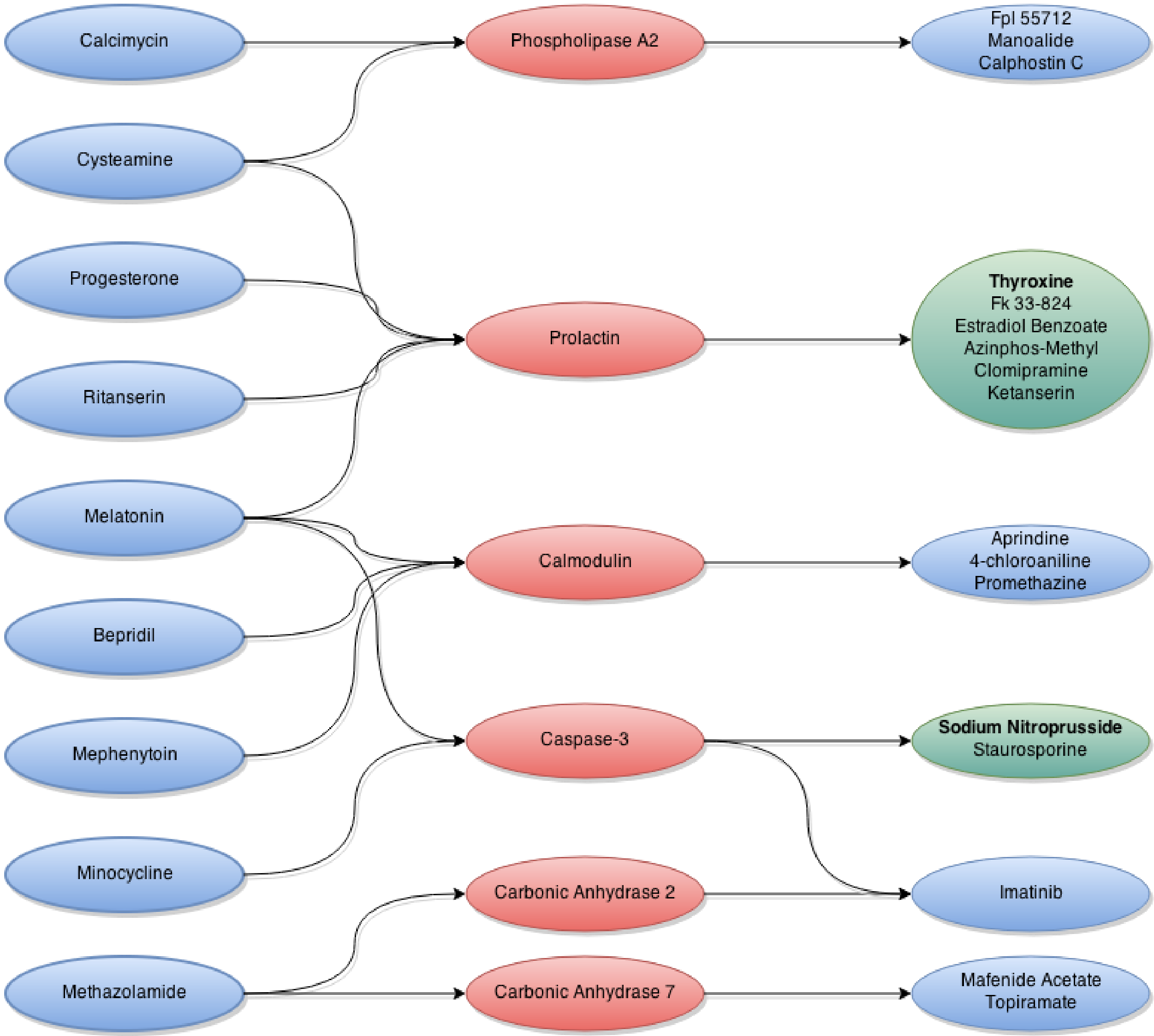


Figure 31: Compounds selected as neuroprotective candidates and the path to their selection.

The drugs in the left column are the neuroprotective drugs that were used to inform the computational method. They are connected to the targets in the middle, shown in red, and multiple connections mean that the target is shared by multiple drugs as indicated. The drugs on the right hand side show the drugs that are known to influence the targets in the middle with very high confidence, that are otherwise dissimilar in their interaction profile to the drugs on the first column and that were feasibly acquired and used for experimental validation. Two of these predictions, thyroxine and sodium nitroprusside showed statistically significant protection; hence they are highlighted in green color. Thyroxine showed protection indistinguishable from the positive control drug, methazolamide at the same dose of 100 µM; whereas sodium nitroprusside showed statistically significantly better protection than methazolamide at 100 µM.

Table 9: The results from the LFM based neuroprotective diversification workflow

The following table represents the LFM based neuroprotective results, specifically the results that led to the identification of Sodium Nitroprusside. The entire set of results are available in Appendix I. First the name of the target shared by the neuroprotectives is reported, i.e. the ‘hypothesis’ that this target is key for neuroprotective activity. In this case, that hypothesis is Caspase 3. Then the supporting neuroprotective drugs that led to the selection of this target are reported: Melatonin, and Minocycline in this case. Then the compounds that are selected based on the LFM-based workflow are listed along with the number of targets they have, their distances to the support drugs, and the average of those distances. It is important to keep in mind that the drugs after the first are selected based on their dissimilarity from not only the support drugs, but also all of the other previously selected compounds hence the average distance to support drugs does not indicate the order in which they were selected.

Hypothesis 3: CASP3_HUMAN (Caspase-3 subunit p12, organism:9606)				
Support:		Melatonin	Minocycline	
Compounds to test hypothesis:				
Drug	Target Count	LV distance to support:		Average:
Sodium Nitroprusside (CID000045469)	14	5.237	5.148	5.193
Imatinib (CID100005291)	48	5.290	5.061	5.175
Staurosporine (CID000044259)	85	4.770	5.399	5.085
Rxb (CID111632008)	1	4.743	4.842	4.793
Tpck (CID000439647)	6	4.405	4.183	4.294
Zoledronic Acid (CID100068740)	83	4.350	4.552	4.451
P-Bromoanisole (CID000007730)	1	4.869	4.550	4.709
Kainate (CID000010255)	69	4.318	4.652	4.485
Nordihydroguaiaretic Acid (CID100004534)	21	4.703	4.503	4.603
Inhibitor 65B (CID005327315)	1	4.725	4.043	4.384
Ptf (CID100013016)	1	4.494	4.654	4.574
Peroxynitrite (CID100104806)	21	4.461	4.515	4.488
Gemcitabine (CID000060749)	28	3.829	4.191	4.010
15-Deoxy-Delta12,14-Prostaglandin J2 (CID100001444)	20	4.433	4.331	4.382
Thapsigargin (CID000446378)	84	3.754	4.319	4.036
Chebi:400985 (CID009851134)	1	4.025	4.017	4.021
Pyrrolidine Isatin Analogue 11F (CID111712912)	1	4.469	3.901	4.185
Inhibitor 64B (CID005327307)	1	4.497	4.524	4.511
Db08213 (CID100001389)	1	4.093	4.043	4.068
3-Morpholinosydnonimine (CID100005219)	4	4.450	3.664	4.057
Pzn (CID005289238)	1	3.724	3.933	3.828
Ac-Devd-Cho (CID100004330)	5	3.880	4.150	4.015
Salidroside (CID100159278)	2	4.093	3.316	3.704
Chebi:461307 (CID111700402)	1	4.030	4.108	4.069
Gsno (CID100003514)	5	3.782	4.016	3.899

4.3 EXPERIMENTAL VALIDATION

We performed experimental testing of the 18 compounds selected from among the results of the computational method (Figure 31). We have identified sodium nitroprusside (SNP) to be significantly neuroprotective, statistically significantly outperforming the positive control compound methazolamide. These experiments were mainly conducted by Hossein Mousavi at the Friedlander Lab under the guidance of Robert Friedlander, with the results replicated in the University of Pittsburgh Drug Discovery Institute by Celeste Reese, Laura Vollmer, Seia Comsa and myself under the guidance of Lans Taylor, Andrew Stern and Mark Schurdak. We have then interrogated the effect of SNP on mitochondrial respiration, with these experiments conducted again by Hossein Mousavi in the Friedlander lab. The results are reported in detail below.

4.3.1 Assessment of Neuronal Cell Death Inhibition for Computationally Selected Compounds

We performed cell toxicity assay (LDH) using STHdh Q111/Q111 (Q111) striatal-derived cell lines, testing each drug at 7 different doses: doses increasing 10-fold from 1nM to 100µM, and an additional dose of 30µM, with a positive control of methazolamide at 100µM (see Appendix J for the entire set of results). The results were from an LDH screen where the readout is fluorescence from a reporter of LDH. LDH is a protein that occurs at a controlled level across all cells and it only leaks out after cell membrane impermeability is comprised, indicating cell death. The LDH viability assay is conducted by observing LDH separately first in the supernatant (indicating the released LDH from cells that have lost membrane integrity) and then in the cell lysate (indicating the LDH still contained in the cells). Q7 and Q111 model cell lines were grown for 24 h in 96-well plates at 5000 cells/well followed by 24 h of treatment. Cells underwent stress conditions by being kept in non-permissive temperature (37C for the Q7 and Q111 cells) and serum free media for 18 hours. Cellular viability was measured Cytotoxicity Detection Kit (LDH) manufactured by Roche.

In these experiments, we identified sodium nitroprusside as a significant neuroprotective. SNP at 50, 100 and 200 uM showed statistically significant neuroprotection against cell death (p-values 0.02, 0.05, and 0.005 respectively) when compared to the vehicle control; with SNP at 100µM showing statistically significantly better protection than the positive control methazolamide at 100µM (p-value 0.007). We next performed cell death assay using SNP in higher concentrations. SNP had the highest protection at 200 µM and was toxic in higher doses (Figure 33, panel a). To verify the results SNP in LDH assay we have performed propidium iodide (PI) uptake based cell death assessment. Similar results were obtained when the neuroprotection assay was reproduced on multiple dates and with both PI and LDH assays (Figure 32). These experimental findings support our *in silico* predictions that SNP would work as neuroprotective in HD.

4.3.2 Sodium Nitroprusside Protection Does Not Impact Mitochondrial Respiration

SNP degrades spontaneously and subsequently generates Fe^{2+}, Cyanide and nitric oxide (NO). NO released from SNP and its role in cardiac and vascular cells has been widely explored over the past six decades owing to the long history of the use of SNP as an antihypertensive [171-174]. However its potential as a neuroprotective has not been reported.

NO has been known to be physiological modulator of mitochondria function and cyclic GMP pathway [175-177]. NO like carbon monoxide (CO) binds to the same site in mitochondria as oxygen. It reversibly reduces the affinity of cytochrome C oxidase to oxygen, however this affect has been observed to be fast and more like a regulatory effect than a blocker effect, unlike CO. To understand whether SNP exerts its neuroprotection via releasing NO and attenuating mitochondrial function, we analyzed different states of mitochondrial respiration in addition of this compound. Glutamate/Malate and Succinate were used as substrates for mitochondrial complex-I and II respectively. SNP had no effect in either states of respiration (Figure 33, panels b & c). There was no change in mitochondrial membrane potential using SNP (Figure 33, panel d). These data suggest that SNP does not attenuate mitochondrial respiration via NO release in isolated mitochondria, and its neuroprotection effect in HD cells is unrelated to mitochondrial direct physiological alteration and it is more likely to be an accumulative effect rather than a direct and canonical effect like mitochondrial complexes blockers.

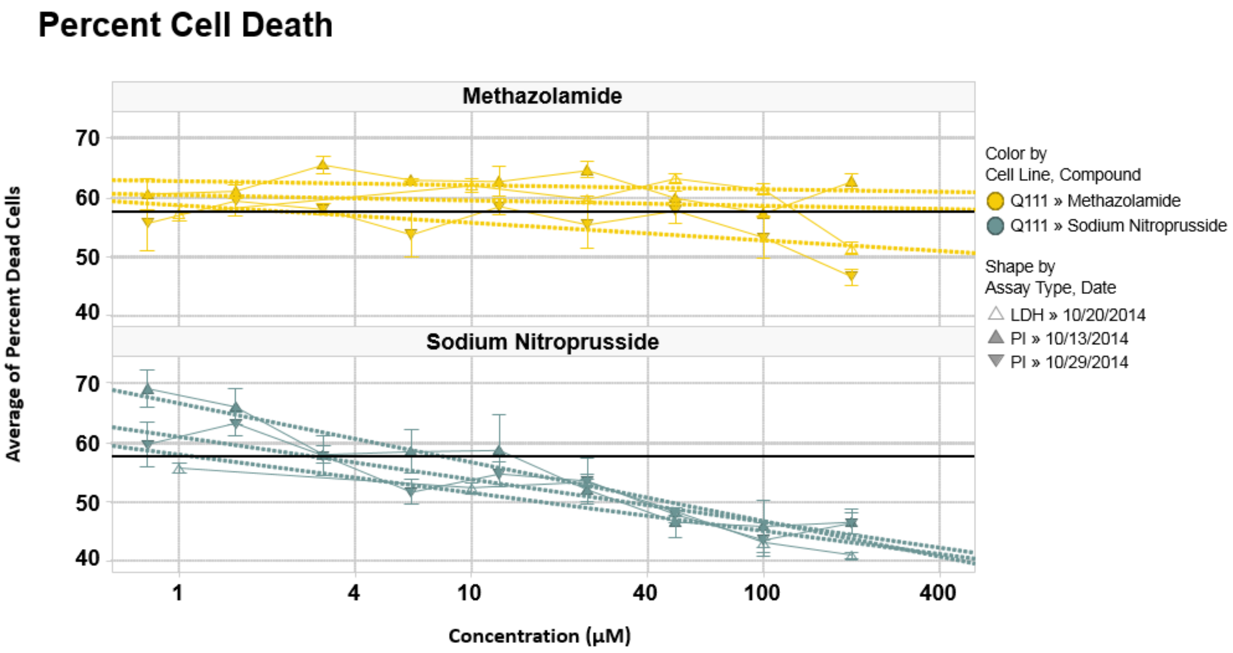


Figure 32: Neuroprotective effect of sodium nitroprusside is repeatedly stronger than methazolamide

The neuroprotective effect of sodium nitroprusside observed to be statistically significantly better than the positive control drug, methazolamide (Appendix J), led us to subsequent experimentation with a different assay. The ordinate shows the percent cell death after insult on Q111 HD model cell lines, whereas the abscissa shows the concentration of the experimented drug. The significantly better neuroprotective effect of sodium nitroprusside reproduced over two different assays, LDH and PI performed on multiple dates.

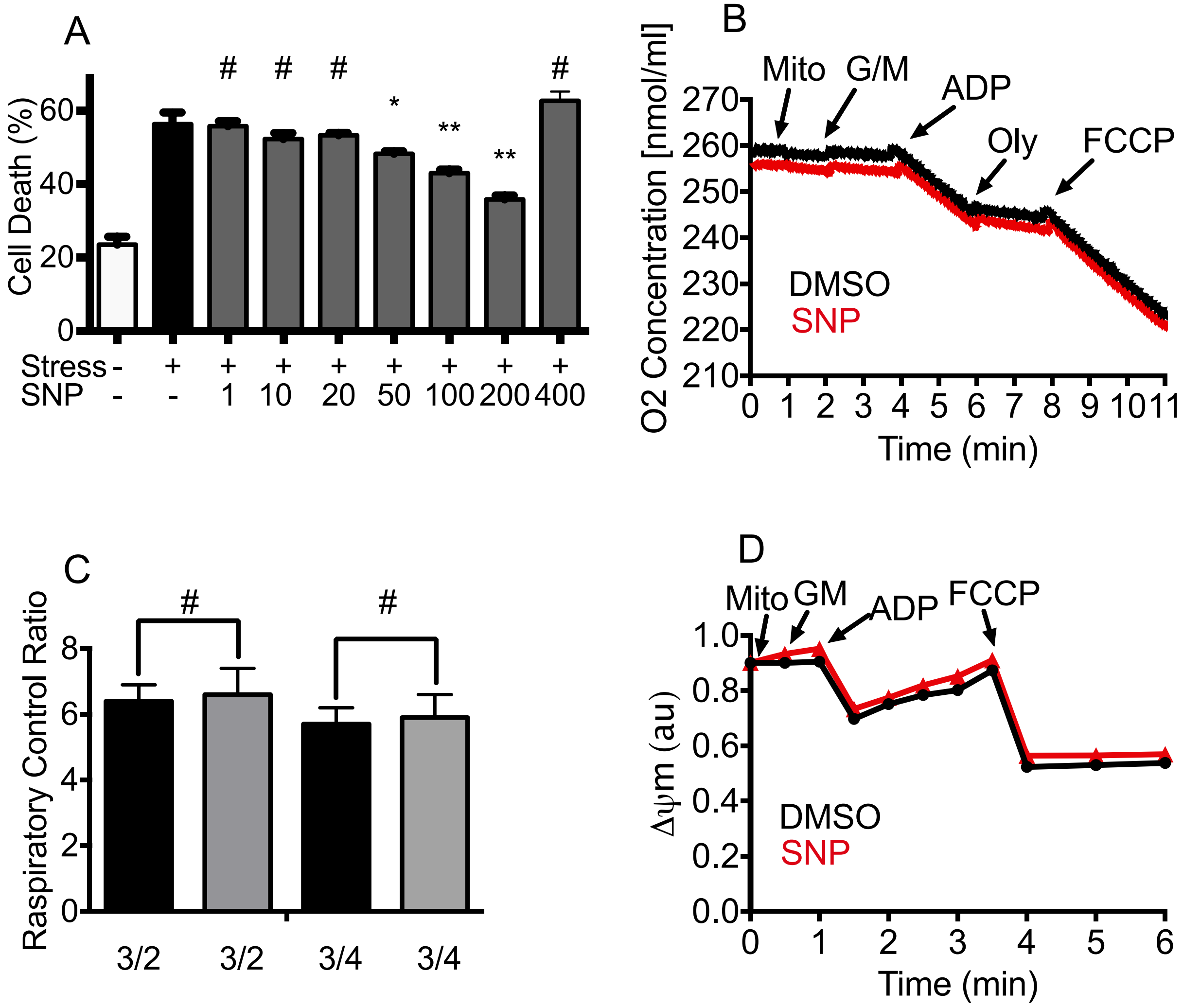


Figure 33: Effect of SNP in Q111 cells and isolated mitochondria

Experimental validation of our prediction based on data collected by our collaborator Hossein Mousavi in Robert Friedlander's lab. (A) HD cells viability in stress condition (Serum deprivation and temperature shift to 37 C) in addition of SNP in different doses. (B) Effect of SNP in isolated brain mitochondria using G/M as substrate for complex-I. (C) Respiratory control ration (RCR) in isolated brain mitochondria with (gray) and without (black) SNP,. (D) Effect of SNP in isolated brain mitochondria membrane potential using TMRM. Traces are representative of 4 or more independent experiments. Mitochondria (Mito), sodium nitrprusside 100 µM (SNP), Glutamate/Malate (GM), Oligomycin (Oly), Carbonyl cyanide 4-(trifluoromethoxy)phenylhydrazone (FCCP), tetramethyl rhodamine methyl ester (TMRM). 3/2 indicates state3/state2 respiration, 3/4 indicates state3/state4 respiration. (*, $p<0.05$, **,$p<0.001$, #, P=NS, ANOVA)

5.0 DISCUSSION

We have discussed methods and results serving multiple different goals and biomedical projects. The LFM-based approach for predicting drug target interactions presented here proposes a solution to the important scientific problem of discovery of unknown interactions (Section 2.1). The methodologies we developed have improved upon the state-of-the-art (Section 2.2). BalestraWeb serves to make the solution we have developed usable by a large number of biomedical researchers all around the world (Section 2.3). Our work on ATD required computational techniques other than LFM, which we designed and then implemented to fit the needs of the particular project at hand with satisfactory results (Chapter 3.0). The data available on HD required the use of LFM, as well as a multitude of other methods that we developed and implemented with again satisfactory results, as we have managed to identify a drug, sodium nitroprusside, that worked better than the state-of-the-art neuroprotective (Chapter 4.0). Due to the extensive and broad implications of the work presented in this dissertation, the discussion is divided into sections structured according to the structure of the work presented.

5.1 LFM APPROACHES FOR ANALYZING DRUG-TARGET INTERACTIONS

Over the last couple of years, there have been a number of computational studies performed to identify targets of existing drugs and drug candidates other than those originally known/proposed to be targeted. A pioneering study is that of Roth, Shoichet and coworkers [26;28] based on compound chemical similarities. Dudley et al focused on inverse correlations between gene expression profiles in the presence of a drug and in a disease state [37]. Yamanishi and his colleagues represented drugs and targets in an integrated ‘pharmacological space’ [29;32]. Gonen used a KBMF method where chemical and genomic similarities were integrated [34]. We proposed a PMF-based AL methodology that can be advantageously used for large datasets.

The applicability of the method to large datasets is worth further attention, given that we will increasingly have access to bigger data such as the STITCH database [45], which will be exploited for repurposable drug identification. The software developed here, made accessible in http://www.csb.pitt.edu/Faculty/bahar/files/, is readily scalable. For very large datasets, which typically have more known interactions, the PMF is able to construct a better model using the plethora of available data; whereas when the number of known interactions is limited, the use of chemical and genomic kernels allows KBMF to outperform PMF. The application of KBMF to large datasets may, however, become challenging, For example, STITCH contains on the order of 10^6 proteins and 10^5 compounds, implying that 10^{12} sequence and 10^{10} chemical similarity comparisons are needed to make predictions. However, the PMF method is independent of chemical, structural or other similarity metrics, and its computation time scales linearly with the number of known interactions; and it proves to perform well on large datasets. The datasets reporting drug-target interactions are constantly improving in quality and quantity, and therefore expected to give even better results when analyzed by an efficient tool. The extension of the method to analyzing big data (with millions of nodes) is foreseeable in the near future. The GraphLab [108] or the GraphChi tool [178] can be used for optimized and parallelized model learning for further performance improvements.

The fact that the PMF is independent of 2D/3D shape comparison methods commonly employed in drug-target pair inferences implies that the derived LVs capture similarities based on the interaction patterns of drugs at the cellular level, even if their molecular structures are dissimilar (see Table 2, Figure 6, Figure 7 and Figure 8). As such, the method may be advantageously used for lead hopping, thus complementing those (e.g. SVM classification algorithms) used in conjunction with 2D or 3D pharmacophoric fingerprints as in the work of Saeh and coworkers [179]. Inasmuch as the currently proposed method does not require structural data for proteins but knowledge of drug-target interactions, it can be advantageously applied to membrane proteins (major drug targets) for which structural data still remain sparse. It can also be used to make predictions across major drug or target classification boundaries. One implication is that the *de novo* predictions are not restricted to major drug or target classification boundaries.

A major utility of the developed tool is the ability to deliver testable hypotheses with regard to repurposable drugs, thus significantly reducing the search space for identifying potent applications of existing drugs (that proved to meet ADMET requirements). The number of experiments that can be efficiently conducted is usually limited, e.g. of the order of 10^2 if not 10^1 for high-confidence assays as opposed to the complete space of ~1.5 million combinations for the dataset used in this study. The fact that the top-ranking predictions exhibit a hit ratio of 59% (for the top 1,000 predictions; or 88% for top 100 predictions) suggest that *de novo* predictions made by the presently introduced method of approach applied to increasingly large datasets are likely to provide useful guidance for experimentally testing, streamlining or prioritizing existing or investigational drugs or new compounds. Another important by-product is the probabilistic assessments on potential side effects, a topic that will become increasingly important with advances in personalized medicine.

Owing to these important considerations, we have built BalestraWeb to make our work easily accessible to biomedical researchers. BalestraWeb provides users the ability to predict the most likely interaction partners of any drug or target beyond those known and compiled in DrugBank. The technology used to build the web server scales linearly with the number of drugs or targets and is therefore easily scalable to larger datasets as they become available. The modular architecture of the software allows us to update the web server to reflect changes as new data become available. Free, fast, and easy-to-use, BalestraWeb enables researchers to help eliminate improbable drug-target interactions and efficiently focus their limited resources on selected drugs.

5.2 COMPUTATIONAL DISCOVERY OF THERAPEUTICS AGAINST ATD

The major advantage of our study was the determination that this type of RNAi screen could be used to rapidly identify potential drug targets using computational approaches, even in the absence of extensive knowledge about target functions, other than their effects on SGFP::ATZ accumulation. We employed two independent, but complementary sources, STITCH and MetaCore, to identify chemical/drug and protein interactions [45;124]. Of the 85 human PN modifiers queried, a total of eight compounds (two directed against each of four targets) were selected as a proof-of principal for this strategy. Remarkably, one compound for each of the four targets showed a dose-dependent decrease sGFP::ATZ accumulation in *C. elegans*. Failure of the other four compounds to have the predicted effects in *C. elegans* were not investigated, but might be due to differences in pharmacokinetics, pharmacodynamics or target-binding site homology between the *C. elegans* and mammalian systems. While the overall success rate of 50% was encouraging, small numbers preclude the calculation of a meaningful positive predictive value. The results do, however, underscore the great potential for combining genome-wide RNAi screens with computational drug-discovery methodologies. The demonstration that one of the compounds, fluspirilene, was also effective in reducing ATZ accumulation in a mammalian cell line lends additional support for further development of this rapid preclinical drug discovery/repurposable strategy.

A second advantage of this drug-discovery strategy was the use of facile genetic techniques in *C. elegans* to determine whether the observed drug effect was due to activity within the predicted target pathway or to an off-target effect. Wortmannin was identified in the screen as a potential inhibitor of the type I PI3K kinase, AGE-1. AGE-1 functions downstream of the sole insulin-like receptor, DAF-2, and inhibition of this IIS pathway suppresses the proteotoxic effects of sGFP::ATZ in this *C. elegans* model, as well as other *C. elegans* models of misfolded protein accumulation [180]. However, wortmannin also inhibits the class III PI3K, VPS-34, which blocks autophagy in *C. elegans* and mammals as well [127]. Since autophagy was an important means of reducing sGFP:: ATZ accumulation [181], suppression of this pathway would be deleterious to these animals. Treatment of the animals with wortmannin decreased sGFP::ATZ accumulation, and this effect was neither enhanced in *AGE-1* mutants nor effective in animals with a mutation in the downstream AGE-1 target gene, *DAF-16*. Taken together, we concluded that the effects of wortmannin at the concentrations used in these animals were via inhibition of AGE-1 and not VPS-34 or some off-target pathway. The large collection of *C. elegans* single gene mutants, combined with a simple quantitative readout system using fluorescent fusion proteins, makes this system ideal for identifying potential drug targets or target pathways after phenotype-based drug screening. While this technology was not meant to replace target identification by the gold-standard of drug–ligand binding measurements *in vitro*, it does provide the rationale for embarking upon more detailed kinetic or structural studies with purified reagents or expensive development of a lead series by exploring structure–activity relationships*in vitro*.

A third advantage of this drug-discovery strategy was the ability to test for the efficacy of combinational therapy. Due to their toxicity, drugs like wortmannin have been largely abandoned as therapeutics in humans. One means to lower toxicity is to use different delivery systems, such as microspheres, to directly deliver lower concentrations of a drug directly to the tissue of interest [182]. Another means to avoid toxicity is to utilize lower concentration of drug by combining it with other therapeutics directed at different targets or target pathways. By using the genetic methods outlined above, we showed that unlike wortmannin, none of the other three candidates exerted their effect via the IIS pathway. This observation was validated by using the *C. elegans* model to show that comparable reductions of sGFP::ATZ accumulation could be achieved at lower doses of wortmannin when it was combined with one of the other drugs. This effect underscores the ability of this experimental system to both identify and test the efficacy of complementary therapeutics. In conclusion, these studies showed that by utilizing the hits from a genome-wide RNAi screen, computational methods could be used to rapidly and strategically develop compounds to prime the preclinical drug-discovery pipeline for rare or neglected diseases lacking effective treatments.

We have also presented two novel computational analysis methods that were designed specifically for the purpose of analyzing the Prestwick dataset screening results in order to discover the mechanism of action of protective drugs as well as the diversification of the known protectives in Section 3.5. There were two goals: (i) to diversify leads, and identify other potential lead compounds; (ii) to identify potential mechanism of action regulators. We have identified sorafenib, duloxetine, and ezetimibe as the potential new hits based on the fact that they share chemical structure and targets with known suppressors. For the mechanism of action, we have identified that calcium channels are a strong candidate based on evaluation of the experimental data using two different databases, with two different evaluation methods. The hypothesis that calcium channels could be relevant to the ATZ aggregation inhibition, and the hypotheses that the three suggested drugs (see Figure 24) could be therapeutically useful needs to be tested experimentally to be validated.

5.3 NEUROPROTECTIVE IDENTIFICATION FOR HD

In this study, we have devised and applied a novel *in silico* method to perform target based hit diversification using previously published information and public databases. We then demonstrated that by testing 18 of the predictions we identified two new neuroprotective repurposable candidates in HD with one of them, SNP, outperforming the other compounds. The exact mechanism with which SNP exerts its neuroprotective effect is not known, however it has been first reported to be an effective antihypertensive in 1928, its potential has been clinically realized between 1951 and 1955, and it became commercially available as an approved antihypertensive for use in the United States in 1974 [171;172]. In addition to being a well established hypotensive agent, it has long been known to increase cGMP levels [183], inhibit cytosolic Ca^{2+} levels [184], and increase nitric oxide (NO) levels [173]. More recently, it has been reported to be effective in prevention of apoptosis of macrophages after hydrogen peroxide insult by preventing activation of caspase-3 and caspase-9 with a 24 hour pretreatment before insult [185]. SNP has been widely reported to be an apoptosis inducer at relatively higher doses, but protective in lower doses: induces apoptosis in mouse C2C12 myoblast cells [186]; induces apoptosis in H9C2 cardiac muscle cells at doses of 2mM or higher [187]; in vascular smooth muscle cells, induces apoptosis at 1.5mM, while pretreatment with 30 µM or higher SNP was found to be protective against high dose SNP induced toxicity [188]; it was found to reduce staurosporine induced caspase activity and apoptosis in cardiomyocytes at doses of 100 µM [189]. These findings demonstrate that SNP impacts caspase activation through its multitude of effects on nitric oxide levels, cGMP levels, and mitochondrial activity. These are potentially driven by its chemical composition as SNP can create NO and cyanide through decomposition, with the latter disrupting mitochondrial activity. Literature also shows that it is important to keep SNP at low doses for safety, and thus our findings that SNP can be a neuroprotective in concentration ranges as low as 30 – 100 µM makes SNP a feasible therapeutic agent candidate for use in HD *in vivo* studies.

The fact that 2 of the 18 tested predictions (thyroxine and SNP) were statistically significantly protective represents an 11% hit rate, which compares against the reported average hit rate of 1.8% hit rate in uninformed compound collection screens conducted by the NIH [122]. In a comprehensive study using the SEA method [27], 27.8% of all the experiments suggested by SEA and later executed were found to be correct interactions. However, we note that: (1) the computational method we developed here was mostly designed to implicate targets regulating neuroprotective activity, this was not a method designed purely to find more neuroprotectives but primarily to identify mechanism of action of neuroprotectives; (2) a significant portion of our predictions have not been tested, in contrast to the fact that almost every strong prediction was experimentally tested in this study. Due to the low number of total compounds tested a reasonably strong assessment of the hit rate cannot be made without conducting significantly higher numbers of experiments. However this preliminary finding might be viewed an encouraging result demonstrating the potential of the computational analysis of publicly available data, combined with the previously available information, in general. There is a potential to form informed hypotheses *in silico* using the compendium of public data resources and this can help generate effective translational therapies by allowing the repositioning of known drugs against new indications with efficiency. Furthermore, considering that these 18 drugs were not selected purely based on the method's suggestions but instead heavily influenced by the feasibility of acquisition suggests that the method could potentially yield an even higher discovery rate.

To summarize, in this study we have demonstrated the viability of target-based active diversification as a computational technique for finding therapeutic agent candidates for repurposable drugs against diseases with no known therapies; and that SNP or its safer derivatives could potentially be used in HD therapy as a neuroprotective agent. More generally, the computational techniques described here can be used for a diverse range of diseases. Specifically, the LFM based methodology described in Section 4.2 can be broadly applied to any disease of interest where some small set of initial hits are identified. Thus, the methodology described herein could be helpful in identifying the mechanism of action of any compounds of interest within any disease as the method is entirely independent of HD. Once the targets of interest are identified, the method could be used to select drugs that act on different targets thus create polypharmacological therapeutic strategies that target multiple different mechanisms to potentially achieve synergistic effect. Furthermore, the patients could be characterized using genomic/transcriptomic profiling and for patients with a signifcant change in one or more of the targets implicated as being related to the mechanism, that information could be used to adjust the therapeutic strategy thereby achieveing patient stratification. Our study also invites attention to the importance of access to big data sources for rapid discoveries of repurposable drugs against myriad untreated diseases.

5.4 CONCLUDING REMARKS

On the broadest level, the work presented in this dissertation aims to demonstrate that computational analysis of experimental data can help build useful testable hypotheses. The importance of experimental work in generating new biomedical data is without question. However guiding the experimental work using both the results from previous experimental results and the large public datasets as appropriate can deliver improved returns. The private data that are acquired within the context of a specific project are important in generating models about the biomedical components of significance for that particular question. The public datasets, on the other hand, provide collections of big data that cannot be compiled by any single research group; therefore they present an important source to explore. The combination of those two datasets, and the development and implementation of computational methods designed specifically for the needs of the biomedical driving project at hand appear to consistently yield plausible hypotheses, both in ATD (as evidenced by the discovery of glibenclamide and four new repurposable drug candidates) and HD (as suggested by the discovery of sodium nitroprusside as a neuroprotective in HD). It is my hope that this work, as well as many other valuable contributions to the field, will help enable the widespread use of computational techniques to guide biomedical assays and thus facilitate efficient use of the resources available. Finally, by combining and packaging the code I have written over the course of my PhD to handle drug-target interaction datasets in BalestraTK, I hope to facilitate any follow-up of the research activity described in my dissertation.

5.5 FUTURE DIRECTIONS

The work described in this dissertation ranges from machine learning (Chapter 2) to big data analysis (Chapters 3 and 4) to experimental validation of predictions (Chapter 4). Consequently the future directions are also quite diverse. For the work described in Chapter 2, it would be important to test the top predictions (>90% confidence) made by BalestraWeb (shown in Table 3) experimentally to validate/refute. As these predictions are made by an average of 128 models which place true but unknown interactions among the top 20 predictions 50% of the time, they come with a good level of confidence. The researchers can always use BalestraWeb to find out the known and most probable interactions of specific drugs/targets that they are interested in. However to test BalestraWeb itself and to significantly increase its use by the community, it is imperative that the top predictions of BalestraWeb be systematically tested – not a subset selected for feasibility.

For the ATD work, the future directions would be to test Ezetimibe, Sorafenib, and Duloxetine in an ATD model. There is substantial reason for thinking that these drugs can be active, based both on chemical models and target-based models as described in Section 3.5. If proven to beactive, it would increase the likelihood of repurposing a known drug candidate against ATD, thus increasing the likelihood of developing an effective therapy to this disease that negatively affects 1 in 1600 to 2000 children.

For the HD project, the LFM- based hypothesized compounds have not been thoroughly tested as originaly intended. The objective when designing the methodology was to give 100 predictions to be tested. Hence, to continue that work in its original spirit I would recommend building a new LFM of STITCH v4 using the strategies I laid out in Section 2.6. In parallel, the set of known neuroprotective drugs shown in Figure 25 could be experimentally tested to find the subset of drugs that work reproducibly in an HD model (i.e. to identify a refined set of neuroprotectives). Then the algorithmic workflow shown in Figure 30 could be adopted with this improved LFM and data to build a new set of hypotheses, from which a second generation of 100 compounds could be tested, by interrogating 10 compounds for each of the top 10 target hypotheses. New neuroprotectives, probably more powerful than sodium nitroprusside, will be discovered. Perhaps even more importantly, if there exists any target that has led to the discovery of multiple new neuroprotectives, this will implicate a specific mechanism of action. This is very important because it will provide us with a mechanistic understanding of neuroprotection with compound-based experimentation alone.

In general, new machine learning based computational methods are always dependent on the input data being of high quality to function accurately. Therefore it is important to make sure that the public datasets are as comprehensive and well-curated as possible. To that end, integrating data available on multiple databases accurately is highly important. The construction of the STITCH database [45;119;190] was such an effort - in fact the name can be seen as referring to ‘stitching’ different data sources together. Likewise there is a recent initiative aiming to bring different databases together drug-drug interactions [191]. It is with this goal that we have envisioned BalestraTK, where the aggregate data on drug-drug, drug-protein and protein-protein interactions can all be potentially accessed over a single API and where a new integrator can be written for each new database, while conserving the foundation data access API. Moreover, BalestraTK provides significant time savings as it makes arbitrary data access a constant time, rapid operation in what are otherwise large datasets. This enables rapid access to data, and if its development is kept-up-to-date with the future versions (it is up-to-date with the most recent versions of the required databases currently available) utilizing the newest data sources with no code change to existing services.

More broadly speaking, I would suggest the biomedical research community at large to make use of computational approaches coupled with the data available at their hand related to their specific project(s) to guide their experimentation. Oftentimes scientists with more biomedical background than computational will try to use computational tools as a ‘black box’ when in fact they are usually made accessible as an open source, and they should ideally be used in an integrated manner. For the computational scientists, my advice is that it is important to produce algorithmic approaches tailored to the data and questions at hand instead of trying to force the use of a specific approach that they have developed. My recommendation would be to think about the method that would make the most sense given the data and the specific problem under investigation and then devise the computational technique to best addresses that need. I think the full potential of computational methods in revolutionizing biomedical research is yet to be realized.

APPENDIX A

THE CONTENTS OF THE ONLINE FILES FOR PMF

The contents of active_passive_learning_code.zip are described in the table below:

Filename	**Description of contents**
al.m	Active learner code (Matlab)
pl.m	Passive learner code (Matlab)
pmfchang.m	PMF code (Matlab)
rl.m	Random learner code (Matlab)
runme.m	The user only needs to run this file in Matlab to execute the code and get the results.
dtdata.mat	The data file (Matlab)
ReadMe.txt	Instructions on running the code.

The contents of denovo.zip:

Filename	**Description of contents**
README.txt	Instructions on running the code
runpreds.m	Code for generating de novo predictions repeatedly (Matlab)
pmfchang_d.m	PMF code (Matlab)
whole.m	Code for calculating prediction results in matrix form (Matlab)
denovo.m	Code for outputting prediction results in txt file (Matlab)
dtdata_5041.txt	Dataset from DrugBank used for *de novo* predictions
drug_inddict.txt	Index directory for all drugs used for *de novo* predictions
trgt_inddict.txt	Index directory for all targets used for *de novo* predictions
translate.py	Code for translating the output of prediction results into interactions with real names (Python)
predictionfolder	Folder that stores all the prediction results when code is executed

APPENDIX B

DRUG TARGET PAIRS FROM BALESTRAWEB WITH PREDICTED INTERACTION SCORE ABOVE 70%

The columns represent from left to right: the DrugBank v4 ID of the drug, the name of the drug, the DrugBank v4 ID of the target, the name of the target, the BalestraWeb predicted interaction score. The interactions are sorted based on the predicted interaction score. In DrugBank v4, there are 1313 approved drugs, which have 1455 targets with 4860 known interactions between them. Among the 1,905,555 unknown interactions between these drugs and targets, the following 589 are predicted by BalestraWeb's LFM based engine to be top (i.e. above the 70% threshold).

Drug ID	Drug Name	Target ID	Target Name	Score
DB00116	Tetrahydrofolic acid	BE0002176	Methylenetetrahydrofolate reductase	1
DB00116	Tetrahydrofolic acid	BE0000331	Serine hydroxymethyltransferase, cytosolic	1
DB00145	Glycine	BE0000331	Serine hydroxymethyltransferase, cytosolic	0.99971
DB00116	Tetrahydrofolic acid	BE0000292	Serine hydroxymethyltransferase, mitochondrial	0.99954
DB00128	L-Aspartic Acid	BE0000277	Calcium-binding mitochondrial carrier protein Aralar2	0.9993
DB00145	Glycine	BE0000292	Serine hydroxymethyltransferase, mitochondrial	0.99912
DB00370	Mirtazapine	BE0000291	5-hydroxytryptamine receptor 1A	0.92736
DB00543	Amoxapine	BE0000572	Alpha-2B adrenergic receptor	0.92068
DB00408	Loxapine	BE0000393	5-hydroxytryptamine receptor 2B	0.91003
DB04946	Iloperidone	BE0000533	5-hydroxytryptamine receptor 2C	0.9008
DB00696	Ergotamine	BE0000342	Alpha-2C adrenergic receptor	0.90057
DB04946	Iloperidone	BE0000289	Alpha-2A adrenergic receptor	0.89727
DB00477	Chlorpromazine	BE0000560	Muscarinic acetylcholine receptor M2	0.89704
DB00334	Olanzapine	BE0000715	Alpha-1D adrenergic receptor	0.89653
DB00363	Clozapine	BE0000145	D(1B) dopamine receptor	0.89481

DB00246	Ziprasidone	BE0000715	Alpha-1D adrenergic receptor	0.88781
DB06148	Mianserin	BE0000797	5-hydroxytryptamine receptor 1B	0.88758
DB00543	Amoxapine	BE0000342	Alpha-2C adrenergic receptor	0.88596
DB06148	Mianserin	BE0000659	5-hydroxytryptamine receptor 1D	0.88587
DB01142	Doxepin	BE0000650	5-hydroxytryptamine receptor 7	0.88378
DB01238	Aripiprazole	BE0000715	Alpha-1D adrenergic receptor	0.88376
DB06148	Mianserin	BE0000501	Alpha-1A adrenergic receptor	0.88306
DB00988	Dopamine	BE0000451	5-hydroxytryptamine receptor 2A	0.88301
DB00726	Trimipramine	BE0000342	Alpha-2C adrenergic receptor	0.88223
DB00408	Loxapine	BE0000715	Alpha-1D adrenergic receptor	0.88034
DB06216	Asenapine	BE0000145	D(1B) dopamine receptor	0.87895
DB00363	Clozapine	BE0000715	Alpha-1D adrenergic receptor	0.8786
DB01142	Doxepin	BE0000020	D(1A) dopamine receptor	0.8778
DB00321	Amitriptyline	BE0000572	Alpha-2B adrenergic receptor	0.87746
DB00370	Mirtazapine	BE0000572	Alpha-2B adrenergic receptor	0.87546
DB06216	Asenapine	BE0000575	Alpha-1B adrenergic receptor	0.87291
DB01142	Doxepin	BE0000581	D(3) dopamine receptor	0.87012
DB00363	Clozapine	BE0000393	5-hydroxytryptamine receptor 2B	0.87007
DB00543	Amoxapine	BE0000575	Alpha-1B adrenergic receptor	0.86803
DB04946	Iloperidone	BE0000572	Alpha-2B adrenergic receptor	0.86684
DB01224	Quetiapine	BE0000393	5-hydroxytryptamine receptor 2B	0.86663
DB01238	Aripiprazole	BE0000393	5-hydroxytryptamine receptor 2B	0.86338
DB04946	Iloperidone	BE0000575	Alpha-1B adrenergic receptor	0.86263
DB00370	Mirtazapine	BE0000575	Alpha-1B adrenergic receptor	0.86254
DB00477	Chlorpromazine	BE0000715	Alpha-1D adrenergic receptor	0.86204
DB06148	Mianserin	BE0000389	D(4) dopamine receptor	0.86199
DB01142	Doxepin	BE0000389	D(4) dopamine receptor	0.86155
DB00726	Trimipramine	BE0000715	Alpha-1D adrenergic receptor	0.86027
DB00370	Mirtazapine	BE0000020	D(1A) dopamine receptor	0.8602
DB06148	Mianserin	BE0000020	D(1A) dopamine receptor	0.85984
DB00246	Ziprasidone	BE0000393	5-hydroxytryptamine receptor 2B	0.8584
DB00321	Amitriptyline	BE0000342	Alpha-2C adrenergic receptor	0.85827
DB00247	Methysergide	BE0000659	5-hydroxytryptamine receptor 1D	0.85811
DB00321	Amitriptyline	BE0000311	5-hydroxytryptamine receptor 3A	0.85767
DB01403	Methotrimeprazine	BE0000291	5-hydroxytryptamine receptor 1A	0.85537
DB01608	Propericiazine	BE0000501	Alpha-1A adrenergic receptor	0.85532
DB00543	Amoxapine	BE0000659	5-hydroxytryptamine receptor 1D	0.85478
DB00543	Amoxapine	BE0004889	D(1B) dopamine receptor	0.85449
DB00543	Amoxapine	BE0000145	D(1B) dopamine receptor	0.854
DB00696	Ergotamine	BE0000581	D(3) dopamine receptor	0.85365
DB00370	Mirtazapine	BE0000389	D(4) dopamine receptor	0.8499
DB00726	Trimipramine	BE0000581	D(3) dopamine receptor	0.84813

DB06216	Asenapine	BE0000659	5-hydroxytryptamine receptor 1D	0.84753
DB00696	Ergotamine	BE0000020	D(1A) dopamine receptor	0.84715
DB00726	Trimipramine	BE0000797	5-hydroxytryptamine receptor 1B	0.84638
DB00321	Amitriptyline	BE0000756	D(2) dopamine receptor	0.84574
DB00734	Risperidone	BE0000715	Alpha-1D adrenergic receptor	0.84545
DB00540	Nortriptyline	BE0000650	5-hydroxytryptamine receptor 7	0.84446
DB00477	Chlorpromazine	BE0000342	Alpha-2C adrenergic receptor	0.84381
DB00988	Dopamine	BE0000533	5-hydroxytryptamine receptor 2C	0.84278
DB00734	Risperidone	BE0000797	5-hydroxytryptamine receptor 1B	0.84266
DB00477	Chlorpromazine	BE0000247	Muscarinic acetylcholine receptor M5	0.84249
DB00477	Chlorpromazine	BE0000393	5-hydroxytryptamine receptor 2B	0.84201
DB00477	Chlorpromazine	BE0000405	Muscarinic acetylcholine receptor M4	0.84182
DB00477	Chlorpromazine	BE0000289	Alpha-2A adrenergic receptor	0.8412
DB01392	Yohimbine	BE0000020	D(1A) dopamine receptor	0.84103
DB00370	Mirtazapine	BE0000797	5-hydroxytryptamine receptor 1B	0.841
DB00934	Maprotiline	BE0000749	Sodium-dependent serotonin transporter	0.84065
DB00734	Risperidone	BE0000145	D(1B) dopamine receptor	0.84029
DB08815	Lurasidone	BE0000572	Alpha-2B adrenergic receptor	0.84022
DB00726	Trimipramine	BE0000020	D(1A) dopamine receptor	0.84005
DB00508	Triflupromazine	BE0000451	5-hydroxytryptamine receptor 2A	0.83931
DB01151	Desipramine	BE0000501	Alpha-1A adrenergic receptor	0.83865
DB00193	Tramadol	BE0000092	Muscarinic acetylcholine receptor M1	0.83824
DB00726	Trimipramine	BE0000650	5-hydroxytryptamine receptor 7	0.83717
DB00458	Imipramine	BE0000289	Alpha-2A adrenergic receptor	0.83601
DB01403	Methotrimeprazine	BE0000650	5-hydroxytryptamine receptor 7	0.83485
DB00656	Trazodone	BE0000756	D(2) dopamine receptor	0.83467
DB00420	Promazine	BE0000342	Alpha-2C adrenergic receptor	0.83463
DB00508	Triflupromazine	BE0000045	Muscarinic acetylcholine receptor M3	0.83372
DB01267	Paliperidone	BE0000715	Alpha-1D adrenergic receptor	0.83331
DB00656	Trazodone	BE0000572	Alpha-2B adrenergic receptor	0.83318
DB00777	Propiomazine	BE0000291	5-hydroxytryptamine receptor 1A	0.83313
DB00734	Risperidone	BE0000650	5-hydroxytryptamine receptor 7	0.83306
DB06148	Mianserin	BE0000575	Alpha-1B adrenergic receptor	0.83264
DB00420	Promazine	BE0000581	D(3) dopamine receptor	0.8319
DB01392	Yohimbine	BE0000389	D(4) dopamine receptor	0.83169
DB00726	Trimipramine	BE0000393	5-hydroxytryptamine receptor 2B	0.83139
DB00714	Apomorphine	BE0000650	5-hydroxytryptamine receptor 7	0.83129
DB00696	Ergotamine	BE0000393	5-hydroxytryptamine receptor 2B	0.83083
DB01142	Doxepin	BE0000797	5-hydroxytryptamine receptor 1B	0.82926
DB00726	Trimipramine	BE0000389	D(4) dopamine receptor	0.82833
DB00420	Promazine	BE0000291	5-hydroxytryptamine receptor 1A	0.82818
DB00696	Ergotamine	BE0000650	5-hydroxytryptamine receptor 7	0.82744

DB00477	Chlorpromazine	BE0000572	Alpha-2B adrenergic receptor	0.82677
DB01142	Doxepin	BE0000659	5-hydroxytryptamine receptor 1D	0.82598
DB01151	Desipramine	BE0000650	5-hydroxytryptamine receptor 7	0.8257
DB00679	Thioridazine	BE0000715	Alpha-1D adrenergic receptor	0.8255
DB00434	Cyproheptadine	BE0000247	Muscarinic acetylcholine receptor M5	0.82539
DB00458	Imipramine	BE0000581	D(3) dopamine receptor	0.82529
DB04946	Iloperidone	BE0000145	D(1B) dopamine receptor	0.82485
DB00247	Methysergide	BE0000756	D(2) dopamine receptor	0.82417
DB00777	Propiomazine	BE0000342	Alpha-2C adrenergic receptor	0.82347
DB05271	Rotigotine	BE0000451	5-hydroxytryptamine receptor 2A	0.82311
DB06216	Asenapine	BE0000715	Alpha-1D adrenergic receptor	0.82294
DB04946	Iloperidone	BE0000797	5-hydroxytryptamine receptor 1B	0.82281
DB00420	Promazine	BE0000572	Alpha-2B adrenergic receptor	0.82277
DB00540	Nortriptyline	BE0000020	D(1A) dopamine receptor	0.82273
DB01069	Promethazine	BE0000575	Alpha-1B adrenergic receptor	0.82257
DB04946	Iloperidone	BE0000659	5-hydroxytryptamine receptor 1D	0.82228
DB08815	Lurasidone	BE0000501	Alpha-1A adrenergic receptor	0.82165
DB01267	Paliperidone	BE0000797	5-hydroxytryptamine receptor 1B	0.82133
DB00370	Mirtazapine	BE0000945	5-hydroxytryptamine receptor 6	0.82128
DB00434	Cyproheptadine	BE0000405	Muscarinic acetylcholine receptor M4	0.82019
DB00420	Promazine	BE0000289	Alpha-2A adrenergic receptor	0.81976
DB04843	Mepenzolate	BE0000560	Muscarinic acetylcholine receptor M2	0.81963
DB01267	Paliperidone	BE0000650	5-hydroxytryptamine receptor 7	0.81953
DB06148	Mianserin	BE0000311	5-hydroxytryptamine receptor 3A	0.81945
DB00193	Tramadol	BE0000560	Muscarinic acetylcholine receptor M2	0.81899
DB00751	Epinastine	BE0000342	Alpha-2C adrenergic receptor	0.8189
DB00734	Risperidone	BE0000393	5-hydroxytryptamine receptor 2B	0.81778
DB00589	Lisuride	BE0000650	5-hydroxytryptamine receptor 7	0.81776
DB00458	Imipramine	BE0000342	Alpha-2C adrenergic receptor	0.81758
DB00370	Mirtazapine	BE0000715	Alpha-1D adrenergic receptor	0.81683
DB00777	Propiomazine	BE0000572	Alpha-2B adrenergic receptor	0.81672
DB00656	Trazodone	BE0000575	Alpha-1B adrenergic receptor	0.81667
DB08815	Lurasidone	BE0000020	D(1A) dopamine receptor	0.81667
DB01622	Thioproperazine	BE0000715	Alpha-1D adrenergic receptor	0.81611
DB00656	Trazodone	BE0000342	Alpha-2C adrenergic receptor	0.81572
DB04946	Iloperidone	BE0000393	5-hydroxytryptamine receptor 2B	0.81479
DB00543	Amoxapine	BE0000045	Muscarinic acetylcholine receptor M3	0.81475
DB00696	Ergotamine	BE0000389	D(4) dopamine receptor	0.81449
DB01392	Yohimbine	BE0000650	5-hydroxytryptamine receptor 7	0.81424
DB05271	Rotigotine	BE0000342	Alpha-2C adrenergic receptor	0.81414
DB00934	Maprotiline	BE0000291	5-hydroxytryptamine receptor 1A	0.81351
DB00726	Trimipramine	BE0000945	5-hydroxytryptamine receptor 6	0.81342

DB00458	Imipramine	BE0000020	D(1A) dopamine receptor	0.81299
DB00777	Propiomazine	BE0000581	D(3) dopamine receptor	0.81216
DB00540	Nortriptyline	BE0000289	Alpha-2A adrenergic receptor	0.81204
DB00458	Imipramine	BE0000389	D(4) dopamine receptor	0.81174
DB00714	Apomorphine	BE0000501	Alpha-1A adrenergic receptor	0.81007
DB00413	Pramipexole	BE0000650	5-hydroxytryptamine receptor 7	0.80998
DB01267	Paliperidone	BE0000145	D(1B) dopamine receptor	0.80994
DB00543	Amoxapine	BE0000560	Muscarinic acetylcholine receptor M2	0.80993
DB00546	Adinazolam	BE0000764	Gamma-aminobutyric acid receptor subunit alpha-6	0.80874
DB01595	Nitrazepam	BE0004797	Gamma-aminobutyric acid receptor subunit theta	0.80873
DB00934	Maprotiline	BE0000020	D(1A) dopamine receptor	0.80861
DB01403	Methotrimeprazine	BE0000659	5-hydroxytryptamine receptor 1D	0.8077
DB01625	Isopropamide	BE0000560	Muscarinic acetylcholine receptor M2	0.80749
DB00543	Amoxapine	BE0000715	Alpha-1D adrenergic receptor	0.80713
DB00546	Adinazolam	BE0000478	Gamma-aminobutyric acid receptor subunit alpha-4	0.80712
DB00925	Phenoxybenzamine	BE0000172	Beta-1 adrenergic receptor	0.80707
DB01608	Propericiazine	BE0000572	Alpha-2B adrenergic receptor	0.80667
DB00248	Cabergoline	BE0000945	5-hydroxytryptamine receptor 6	0.80663
DB00408	Loxapine	BE0000694	Beta-2 adrenergic receptor	0.8066
DB00777	Propiomazine	BE0000289	Alpha-2A adrenergic receptor	0.80646
DB00805	Minaprine	BE0000291	5-hydroxytryptamine receptor 1A	0.80642
DB00268	Ropinirole	BE0000650	5-hydroxytryptamine receptor 7	0.80634
DB08815	Lurasidone	BE0000533	5-hydroxytryptamine receptor 2C	0.80615
DB00934	Maprotiline	BE0000501	Alpha-1A adrenergic receptor	0.80549
DB00751	Epinastine	BE0000575	Alpha-1B adrenergic receptor	0.80536
DB00458	Imipramine	BE0000572	Alpha-2B adrenergic receptor	0.80526
DB01267	Paliperidone	BE0000393	5-hydroxytryptamine receptor 2B	0.80455
DB01614	Acepromazine	BE0000715	Alpha-1D adrenergic receptor	0.80403
DB01594	Cinolazepam	BE0000764	Gamma-aminobutyric acid receptor subunit alpha-6	0.80373
DB00711	Diethylcarbamazine	BE0000262	Prostaglandin G/H synthase 2	0.80317
DB00370	Mirtazapine	BE0000145	D(1B) dopamine receptor	0.80299
DB00370	Mirtazapine	BE0000146	Histamine H4 receptor	0.80296
DB01142	Doxepin	BE0004889	D(1B) dopamine receptor	0.8026
DB00185	Cevimeline	BE0000560	Muscarinic acetylcholine receptor M2	0.80259
DB00458	Imipramine	BE0000393	5-hydroxytryptamine receptor 2B	0.80254
DB01403	Methotrimeprazine	BE0000945	5-hydroxytryptamine receptor 6	0.80251
DB00268	Ropinirole	BE0000501	Alpha-1A adrenergic receptor	0.8023
DB00797	Tolazoline	BE0000575	Alpha-1B adrenergic receptor	0.80203
DB01151	Desipramine	BE0000945	5-hydroxytryptamine receptor 6	0.80197
DB00751	Epinastine	BE0000572	Alpha-2B adrenergic receptor	0.80183
DB01142	Doxepin	BE0000145	D(1B) dopamine receptor	0.80161
DB00696	Ergotamine	BE0000749	Sodium-dependent serotonin transporter	0.80143

DB05271	Rotigotine	BE0000289	Alpha-2A adrenergic receptor	0.80069
DB00751	Epinastine	BE0000756	D(2) dopamine receptor	0.80031
DB00413	Pramipexole	BE0000501	Alpha-1A adrenergic receptor	0.79947
DB00321	Amitriptyline	BE0000393	5-hydroxytryptamine receptor 2B	0.79912
DB00540	Nortriptyline	BE0000389	D(4) dopamine receptor	0.79909
DB00934	Maprotiline	BE0000945	5-hydroxytryptamine receptor 6	0.79908
DB01392	Yohimbine	BE0000145	D(1B) dopamine receptor	0.79866
DB00540	Nortriptyline	BE0000572	Alpha-2B adrenergic receptor	0.79852
DB00424	Hyoscyamine	BE0000247	Muscarinic acetylcholine receptor M5	0.79805
DB00988	Dopamine	BE0000393	5-hydroxytryptamine receptor 2B	0.79799
DB00420	Promazine	BE0000145	D(1B) dopamine receptor	0.7979
DB06148	Mianserin	BE0000145	D(1B) dopamine receptor	0.79694
DB00321	Amitriptyline	BE0000389	D(4) dopamine receptor	0.79676
DB06216	Asenapine	BE0004889	D(1B) dopamine receptor	0.7962
DB00411	Carbachol	BE0000045	Muscarinic acetylcholine receptor M3	0.79617
DB01151	Desipramine	BE0000020	D(1A) dopamine receptor	0.79601
DB00751	Epinastine	BE0000533	5-hydroxytryptamine receptor 2C	0.79573
DB00540	Nortriptyline	BE0000342	Alpha-2C adrenergic receptor	0.79531
DB06288	Amisulpride	BE0000020	D(1A) dopamine receptor	0.79522
DB00321	Amitriptyline	BE0004889	D(1B) dopamine receptor	0.79473
DB01142	Doxepin	BE0000647	Sodium-dependent dopamine transporter	0.79444
DB01069	Promethazine	BE0000020	D(1A) dopamine receptor	0.79418
DB00589	Lisuride	BE0000575	Alpha-1B adrenergic receptor	0.794
DB08815	Lurasidone	BE0000581	D(3) dopamine receptor	0.79388
DB01403	Methotrimeprazine	BE0000393	5-hydroxytryptamine receptor 2B	0.79318
DB00714	Apomorphine	BE0000575	Alpha-1B adrenergic receptor	0.79318
DB04946	Iloperidone	BE0000715	Alpha-1D adrenergic receptor	0.79275
DB05271	Rotigotine	BE0000533	5-hydroxytryptamine receptor 2C	0.79218
DB01594	Cinolazepam	BE0000478	Gamma-aminobutyric acid receptor subunit alpha-4	0.79176
DB00589	Lisuride	BE0000501	Alpha-1A adrenergic receptor	0.79172
DB00413	Pramipexole	BE0000575	Alpha-1B adrenergic receptor	0.79159
DB06148	Mianserin	BE0000715	Alpha-1D adrenergic receptor	0.79142
DB01242	Clomipramine	BE0000756	D(2) dopamine receptor	0.79139
DB00800	Fenoldopam	BE0000756	D(2) dopamine receptor	0.79107
DB00809	Tropicamide	BE0000247	Muscarinic acetylcholine receptor M5	0.79076
DB00751	Epinastine	BE0000291	5-hydroxytryptamine receptor 1A	0.79068
DB00320	Dihydroergotamine	BE0000291	5-hydroxytryptamine receptor 1A	0.79001
DB00246	Ziprasidone	BE0004889	D(1B) dopamine receptor	0.78919
DB00805	Minaprine	BE0000560	Muscarinic acetylcholine receptor M2	0.78892
DB00696	Ergotamine	BE0000145	D(1B) dopamine receptor	0.78874
DB00477	Chlorpromazine	BE0000797	5-hydroxytryptamine receptor 1B	0.78841
DB08815	Lurasidone	BE0000575	Alpha-1B adrenergic receptor	0.78801

DB00457	Prazosin	BE0000342	Alpha-2C adrenergic receptor	0.78759
DB00268	Ropinirole	BE0000575	Alpha-1B adrenergic receptor	0.78621
DB00363	Clozapine	BE0004889	D(1B) dopamine receptor	0.78532
DB01224	Quetiapine	BE0004889	D(1B) dopamine receptor	0.78497
DB01151	Desipramine	BE0000393	5-hydroxytryptamine receptor 2B	0.78423
DB00543	Amoxapine	BE0000247	Muscarinic acetylcholine receptor M5	0.78404
DB00988	Dopamine	BE0000797	5-hydroxytryptamine receptor 1B	0.7836
DB01224	Quetiapine	BE0000146	Histamine H4 receptor	0.78347
DB01622	Thioproperazine	BE0000289	Alpha-2A adrenergic receptor	0.78308
DB00458	Imipramine	BE0000797	5-hydroxytryptamine receptor 1B	0.78289
DB01392	Yohimbine	BE0000501	Alpha-1A adrenergic receptor	0.78281
DB00216	Eletriptan	BE0000451	5-hydroxytryptamine receptor 2A	0.78176
DB06288	Amisulpride	BE0000291	5-hydroxytryptamine receptor 1A	0.78167
DB00321	Amitriptyline	BE0000581	D(3) dopamine receptor	0.78152
DB01403	Methotrimeprazine	BE0000797	5-hydroxytryptamine receptor 1B	0.78141
DB00387	Procyclidine	BE0000247	Muscarinic acetylcholine receptor M5	0.78093
DB00458	Imipramine	BE0000146	Histamine H4 receptor	0.78067
DB00777	Propiomazine	BE0000650	5-hydroxytryptamine receptor 7	0.78066
DB00988	Dopamine	BE0000501	Alpha-1A adrenergic receptor	0.7802
DB00233	Aminosalicylic Acid	BE0000017	Prostaglandin G/H synthase 1	0.77966
DB00777	Propiomazine	BE0000145	D(1B) dopamine receptor	0.77944
DB01151	Desipramine	BE0000581	D(3) dopamine receptor	0.77847
DB01238	Aripiprazole	BE0004889	D(1B) dopamine receptor	0.77845
DB00370	Mirtazapine	BE0000659	5-hydroxytryptamine receptor 1D	0.77844
DB00656	Trazodone	BE0000486	Sodium-dependent noradrenaline transporter	0.77842
DB01608	Propericiazine	BE0000715	Alpha-1D adrenergic receptor	0.77837
DB00988	Dopamine	BE0000575	Alpha-1B adrenergic receptor	0.77808
DB00247	Methysergide	BE0000581	D(3) dopamine receptor	0.77755
DB01200	Bromocriptine	BE0000945	5-hydroxytryptamine receptor 6	0.77748
DB00540	Nortriptyline	BE0004889	D(1B) dopamine receptor	0.77746
DB00805	Minaprine	BE0000501	Alpha-1A adrenergic receptor	0.77737
DB00934	Maprotiline	BE0000581	D(3) dopamine receptor	0.77708
DB00392	Ethopropazine	BE0000045	Muscarinic acetylcholine receptor M3	0.77697
DB00477	Chlorpromazine	BE0000659	5-hydroxytryptamine receptor 1D	0.77673
DB00247	Methysergide	BE0000289	Alpha-2A adrenergic receptor	0.77656
DB00420	Promazine	BE0000650	5-hydroxytryptamine receptor 7	0.77648
DB00805	Minaprine	BE0000442	Histamine H1 receptor	0.77646
DB00726	Trimipramine	BE0000145	D(1B) dopamine receptor	0.77637
DB00540	Nortriptyline	BE0000311	5-hydroxytryptamine receptor 3A	0.77609
DB01614	Acepromazine	BE0000289	Alpha-2A adrenergic receptor	0.77581
DB00246	Ziprasidone	BE0000146	Histamine H4 receptor	0.77579
DB00247	Methysergide	BE0000020	D(1A) dopamine receptor	0.77509

DB00656	Trazodone	BE0000020	D(1A) dopamine receptor	0.77502
DB00800	Fenoldopam	BE0000389	D(4) dopamine receptor	0.77465
DB00320	Dihydroergotamine	BE0000572	Alpha-2B adrenergic receptor	0.77459
DB01142	Doxepin	BE0000311	5-hydroxytryptamine receptor 3A	0.77452
DB01151	Desipramine	BE0000575	Alpha-1B adrenergic receptor	0.77444
DB06216	Asenapine	BE0000146	Histamine H4 receptor	0.77382
DB00543	Amoxapine	BE0000405	Muscarinic acetylcholine receptor M4	0.77374
DB00804	Dicyclomine	BE0000045	Muscarinic acetylcholine receptor M3	0.77368
DB00988	Dopamine	BE0000945	5-hydroxytryptamine receptor 6	0.77261
DB00777	Propiomazine	BE0000945	5-hydroxytryptamine receptor 6	0.77242
DB00842	Oxazepam	BE0000736	Translocator protein	0.77206
DB01628	Etoricoxib	BE0000017	Prostaglandin G/H synthase 1	0.77196
DB00216	Eletriptan	BE0000533	5-hydroxytryptamine receptor 2C	0.77182
DB01069	Promethazine	BE0000533	5-hydroxytryptamine receptor 2C	0.77181
DB00540	Nortriptyline	BE0000581	D(3) dopamine receptor	0.77139
DB01622	Thioproperazine	BE0000533	5-hydroxytryptamine receptor 2C	0.77102
DB05271	Rotigotine	BE0000797	5-hydroxytryptamine receptor 1B	0.77092
DB00449	Dipivefrin	BE0000172	Beta-1 adrenergic receptor	0.77078
DB01148	Flavoxate	BE0000045	Muscarinic acetylcholine receptor M3	0.77012
DB01238	Aripiprazole	BE0000146	Histamine H4 receptor	0.76975
DB00420	Promazine	BE0000945	5-hydroxytryptamine receptor 6	0.7692
DB00810	Biperiden	BE0000560	Muscarinic acetylcholine receptor M2	0.7687
DB00508	Triflupromazine	BE0000533	5-hydroxytryptamine receptor 2C	0.76869
DB00540	Nortriptyline	BE0000659	5-hydroxytryptamine receptor 1D	0.76857
DB00247	Methysergide	BE0000389	D(4) dopamine receptor	0.76852
DB00679	Thioridazine	BE0000289	Alpha-2A adrenergic receptor	0.76836
DB01614	Acepromazine	BE0000533	5-hydroxytryptamine receptor 2C	0.76826
DB01608	Propericiazine	BE0000342	Alpha-2C adrenergic receptor	0.76781
DB00434	Cyproheptadine	BE0000756	D(2) dopamine receptor	0.76776
DB00988	Dopamine	BE0000659	5-hydroxytryptamine receptor 1D	0.76733
DB01069	Promethazine	BE0000715	Alpha-1D adrenergic receptor	0.76684
DB00805	Minaprine	BE0000045	Muscarinic acetylcholine receptor M3	0.76665
DB00988	Dopamine	BE0000342	Alpha-2C adrenergic receptor	0.76588
DB08815	Lurasidone	BE0000797	5-hydroxytryptamine receptor 1B	0.76536
DB01151	Desipramine	BE0000389	D(4) dopamine receptor	0.76473
DB00543	Amoxapine	BE0000476	5-hydroxytryptamine receptor 1E	0.76443
DB00458	Imipramine	BE0000311	5-hydroxytryptamine receptor 3A	0.76438
DB00458	Imipramine	BE0000659	5-hydroxytryptamine receptor 1D	0.76417
DB00656	Trazodone	BE0000715	Alpha-1D adrenergic receptor	0.764
DB01594	Cinolazepam	BE0003597	Gamma-aminobutyric acid receptor subunit theta	0.76386
DB00193	Tramadol	BE0000311	5-hydroxytryptamine receptor 3A	0.76331
DB00321	Amitriptyline	BE0000020	D(1A) dopamine receptor	0.76327

DB01625	Isopropamide	BE0000092	Muscarinic acetylcholine receptor M1	0.76323
DB01235	L-DOPA	BE0000451	5-hydroxytryptamine receptor 2A	0.76281
DB08815	Lurasidone	BE0000393	5-hydroxytryptamine receptor 2B	0.76272
DB06288	Amisulpride	BE0000533	5-hydroxytryptamine receptor 2C	0.76256
DB00734	Risperidone	BE0000945	5-hydroxytryptamine receptor 6	0.76252
DB00656	Trazodone	BE0000393	5-hydroxytryptamine receptor 2B	0.76237
DB00280	Disopyramide	BE0000405	Muscarinic acetylcholine receptor M4	0.76236
DB00247	Methysergide	BE0000342	Alpha-2C adrenergic receptor	0.76218
DB00805	Minaprine	BE0000486	Sodium-dependent noradrenaline transporter	0.76196
DB00800	Fenoldopam	BE0000451	5-hydroxytryptamine receptor 2A	0.76186
DB00988	Dopamine	BE0000442	Histamine H1 receptor	0.76181
DB06216	Asenapine	BE0000476	5-hydroxytryptamine receptor 1E	0.76179
DB00800	Fenoldopam	BE0000291	5-hydroxytryptamine receptor 1A	0.76161
DB00797	Tolazoline	BE0000715	Alpha-1D adrenergic receptor	0.76148
DB01622	Thioproperazine	BE0000342	Alpha-2C adrenergic receptor	0.76143
DB05271	Rotigotine	BE0000659	5-hydroxytryptamine receptor 1D	0.76019
DB00248	Cabergoline	BE0000476	5-hydroxytryptamine receptor 1E	0.75958
DB01614	Acepromazine	BE0000342	Alpha-2C adrenergic receptor	0.75916
DB05271	Rotigotine	BE0000393	5-hydroxytryptamine receptor 2B	0.75902
DB00696	Ergotamine	BE0000945	5-hydroxytryptamine receptor 6	0.75868
DB00805	Minaprine	BE0000581	D(3) dopamine receptor	0.75863
DB01151	Desipramine	BE0000647	Sodium-dependent dopamine transporter	0.75839
DB00800	Fenoldopam	BE0000533	5-hydroxytryptamine receptor 2C	0.75809
DB01392	Yohimbine	BE0000575	Alpha-1B adrenergic receptor	0.75723
DB00320	Dihydroergotamine	BE0000342	Alpha-2C adrenergic receptor	0.75719
DB00656	Trazodone	BE0000650	5-hydroxytryptamine receptor 7	0.75697
DB00477	Chlorpromazine	BE0000749	Sodium-dependent serotonin transporter	0.75667
DB01221	Ketamine	BE0000749	Sodium-dependent serotonin transporter	0.75662
DB00540	Nortriptyline	BE0000393	5-hydroxytryptamine receptor 2B	0.75638
DB00321	Amitriptyline	BE0000145	D(1B) dopamine receptor	0.75576
DB00458	Imipramine	BE0000145	D(1B) dopamine receptor	0.75569
DB00540	Nortriptyline	BE0000647	Sodium-dependent dopamine transporter	0.75557
DB00696	Ergotamine	BE0000442	Histamine H1 receptor	0.75549
DB00934	Maprotiline	BE0000389	D(4) dopamine receptor	0.75546
DB01100	Pimozide	BE0000020	D(1A) dopamine receptor	0.75398
DB00777	Propiomazine	BE0000393	5-hydroxytryptamine receptor 2B	0.75327
DB00477	Chlorpromazine	BE0000311	5-hydroxytryptamine receptor 3A	0.75324
DB01142	Doxepin	BE0000476	5-hydroxytryptamine receptor 1E	0.75315
DB08815	Lurasidone	BE0000389	D(4) dopamine receptor	0.75271
DB00964	Apraclonidine	BE0000342	Alpha-2C adrenergic receptor	0.75244
DB08815	Lurasidone	BE0000659	5-hydroxytryptamine receptor 1D	0.7515
DB01151	Desipramine	BE0004889	D(1B) dopamine receptor	0.75144

DB06148	Mianserin	BE0000476	5-hydroxytryptamine receptor 1E	0.75138
DB00540	Nortriptyline	BE0000797	5-hydroxytryptamine receptor 1B	0.75133
DB00193	Tramadol	BE0000405	Muscarinic acetylcholine receptor M4	0.75102
DB00751	Epinastine	BE0000020	D(1A) dopamine receptor	0.75065
DB01242	Clomipramine	BE0000291	5-hydroxytryptamine receptor 1A	0.75035
DB01614	Acepromazine	BE0000572	Alpha-2B adrenergic receptor	0.75021
DB08801	Dimetindene	BE0000092	Muscarinic acetylcholine receptor M1	0.75019
DB00751	Epinastine	BE0000715	Alpha-1D adrenergic receptor	0.75014
DB01235	L-DOPA	BE0000291	5-hydroxytryptamine receptor 1A	0.74995
DB06216	Asenapine	BE0000311	5-hydroxytryptamine receptor 3A	0.74954
DB00934	Maprotiline	BE0000575	Alpha-1B adrenergic receptor	0.74951
DB00248	Cabergoline	BE0000442	Histamine H1 receptor	0.7495
DB01618	Molindone	BE0000020	D(1A) dopamine receptor	0.74912
DB00715	Paroxetine	BE0000756	D(2) dopamine receptor	0.7488
DB00321	Amitriptyline	BE0000647	Sodium-dependent dopamine transporter	0.74879
DB00477	Chlorpromazine	BE0000476	5-hydroxytryptamine receptor 1E	0.74862
DB04946	Iloperidone	BE0004889	D(1B) dopamine receptor	0.74854
DB00540	Nortriptyline	BE0000146	Histamine H4 receptor	0.74844
DB00546	Adinazolam	BE0003597	Gamma-aminobutyric acid receptor subunit theta	0.74813
DB00988	Dopamine	BE0000289	Alpha-2A adrenergic receptor	0.74774
DB05271	Rotigotine	BE0000575	Alpha-1B adrenergic receptor	0.74764
DB00508	Triflupromazine	BE0000581	D(3) dopamine receptor	0.7476
DB01151	Desipramine	BE0000289	Alpha-2A adrenergic receptor	0.74758
DB00319	Piperacillin	BE0004290	Penicillin-binding protein 1A	0.7475
DB01151	Desipramine	BE0000715	Alpha-1D adrenergic receptor	0.74748
DB00216	Eletriptan	BE0000289	Alpha-2A adrenergic receptor	0.74746
DB00934	Maprotiline	BE0000393	5-hydroxytryptamine receptor 2B	0.74702
DB00656	Trazodone	BE0000797	5-hydroxytryptamine receptor 1B	0.74685
DB00215	Citalopram	BE0000486	Sodium-dependent noradrenaline transporter	0.74685
DB00477	Chlorpromazine	BE0000486	Sodium-dependent noradrenaline transporter	0.74672
DB00988	Dopamine	BE0000572	Alpha-2B adrenergic receptor	0.74661
DB00800	Fenoldopam	BE0000581	D(3) dopamine receptor	0.74632
DB00540	Nortriptyline	BE0000145	D(1B) dopamine receptor	0.74632
DB04855	Dronedarone	BE0000694	Beta-2 adrenergic receptor	0.74625
DB00247	Methysergide	BE0000572	Alpha-2B adrenergic receptor	0.74623
DB00482	Celecoxib	BE0000017	Prostaglandin G/H synthase 1	0.74591
DB08801	Dimetindene	BE0000045	Muscarinic acetylcholine receptor M3	0.74591
DB01622	Thioproperazine	BE0000572	Alpha-2B adrenergic receptor	0.74585
DB01608	Propericiazine	BE0000756	D(2) dopamine receptor	0.74525
DB00715	Paroxetine	BE0000442	Histamine H1 receptor	0.74491
DB00656	Trazodone	BE0000581	D(3) dopamine receptor	0.74395
DB00280	Disopyramide	BE0000247	Muscarinic acetylcholine receptor M5	0.74387

DB00751	Epinastine	BE0000581	D(3) dopamine receptor	0.74386
DB00193	Tramadol	BE0000756	D(2) dopamine receptor	0.74357
DB01594	Cinolazepam	BE0004797	Gamma-aminobutyric acid receptor subunit theta	0.74306
DB00508	Triflupromazine	BE0000291	5-hydroxytryptamine receptor 1A	0.74305
DB00679	Thioridazine	BE0000581	D(3) dopamine receptor	0.74292
DB01151	Desipramine	BE0000797	5-hydroxytryptamine receptor 1B	0.74233
DB01267	Paliperidone	BE0000945	5-hydroxytryptamine receptor 6	0.74225
DB01151	Desipramine	BE0000311	5-hydroxytryptamine receptor 3A	0.74218
DB00679	Thioridazine	BE0000342	Alpha-2C adrenergic receptor	0.74201
DB00546	Adinazolam	BE0004797	Gamma-aminobutyric acid receptor subunit theta	0.742
DB00321	Amitriptyline	BE0000476	5-hydroxytryptamine receptor 1E	0.74193
DB00589	Lisuride	BE0000715	Alpha-1D adrenergic receptor	0.74169
DB01403	Methotrimeprazine	BE0000311	5-hydroxytryptamine receptor 3A	0.74164
DB01595	Nitrazepam	BE0000736	Translocator protein	0.74161
DB00805	Minaprine	BE0000289	Alpha-2A adrenergic receptor	0.74096
DB00715	Paroxetine	BE0000533	5-hydroxytryptamine receptor 2C	0.74085
DB00964	Apraclonidine	BE0000575	Alpha-1B adrenergic receptor	0.74067
DB00508	Triflupromazine	BE0000247	Muscarinic acetylcholine receptor M5	0.73998
DB01625	Isopropamide	BE0000247	Muscarinic acetylcholine receptor M5	0.73959
DB00679	Thioridazine	BE0000533	5-hydroxytryptamine receptor 2C	0.73944
DB08815	Lurasidone	BE0000715	Alpha-1D adrenergic receptor	0.73915
DB01403	Methotrimeprazine	BE0004889	D(1B) dopamine receptor	0.73912
DB00653	Magnesium Sulfate	BE0002359	Voltage-dependent L-type calcium channel subunit alpha-1D	0.73903
DB00420	Promazine	BE0000393	5-hydroxytryptamine receptor 2B	0.73859
DB00679	Thioridazine	BE0000572	Alpha-2B adrenergic receptor	0.73813
DB08810	Cinitapride	BE0000756	D(2) dopamine receptor	0.738
DB01200	Bromocriptine	BE0000442	Histamine H1 receptor	0.73784
DB00622	Nicardipine	BE0002355	Voltage-dependent L-type calcium channel subunit alpha-1S	0.73774
DB00934	Maprotiline	BE0000647	Sodium-dependent dopamine transporter	0.73747
DB00805	Minaprine	BE0000389	D(4) dopamine receptor	0.7374
DB00216	Eletriptan	BE0000756	D(2) dopamine receptor	0.73732
DB00193	Tramadol	BE0000247	Muscarinic acetylcholine receptor M5	0.73697
DB00714	Apomorphine	BE0000715	Alpha-1D adrenergic receptor	0.73645
DB00679	Thioridazine	BE0000291	5-hydroxytryptamine receptor 1A	0.73641
DB01018	Guanfacine	BE0000342	Alpha-2C adrenergic receptor	0.73622
DB00714	Apomorphine	BE0000945	5-hydroxytryptamine receptor 6	0.73617
DB01100	Pimozide	BE0000451	5-hydroxytryptamine receptor 2A	0.73616
DB01622	Thioproperazine	BE0000389	D(4) dopamine receptor	0.73611
DB01337	Pancuronium	BE0000092	Muscarinic acetylcholine receptor M1	0.73609
DB01226	Mivacurium	BE0000092	Muscarinic acetylcholine receptor M1	0.73607
DB01622	Thioproperazine	BE0000581	D(3) dopamine receptor	0.73582
DB00215	Citalopram	BE0000560	Muscarinic acetylcholine receptor M2	0.73577

DB00726	Trimipramine	BE0000146	Histamine H4 receptor	0.73567
DB00777	Propiomazine	BE0000797	5-hydroxytryptamine receptor 1B	0.7356
DB00413	Pramipexole	BE0000715	Alpha-1D adrenergic receptor	0.73552
DB01049	Ergoloid mesylate	BE0000451	5-hydroxytryptamine receptor 2A	0.7351
DB01200	Bromocriptine	BE0004889	D(1B) dopamine receptor	0.73507
DB00193	Tramadol	BE0000442	Histamine H1 receptor	0.73493
DB01618	Molindone	BE0000533	5-hydroxytryptamine receptor 2C	0.73467
DB01338	Pipecuronium	BE0000092	Muscarinic acetylcholine receptor M1	0.73387
DB00777	Propiomazine	BE0000659	5-hydroxytryptamine receptor 1D	0.73357
DB01235	L-DOPA	BE0000533	5-hydroxytryptamine receptor 2C	0.73336
DB01618	Molindone	BE0000560	Muscarinic acetylcholine receptor M2	0.73277
DB01069	Promethazine	BE0000289	Alpha-2A adrenergic receptor	0.73209
DB00268	Ropinirole	BE0000715	Alpha-1D adrenergic receptor	0.7319
DB00247	Methysergide	BE0000145	D(1B) dopamine receptor	0.73142
DB00247	Methysergide	BE0000945	5-hydroxytryptamine receptor 6	0.73126
DB00248	Cabergoline	BE0000311	5-hydroxytryptamine receptor 3A	0.73105
DB00502	Haloperidol	BE0000291	5-hydroxytryptamine receptor 1A	0.73094
DB05271	Rotigotine	BE0000650	5-hydroxytryptamine receptor 7	0.73086
DB00408	Loxapine	BE0004863	Alpha-1D adrenergic receptor	0.73066
DB00508	Triflupromazine	BE0000389	D(4) dopamine receptor	0.73056
DB00420	Promazine	BE0000659	5-hydroxytryptamine receptor 1D	0.73044
DB01614	Acepromazine	BE0000389	D(4) dopamine receptor	0.73043
DB00508	Triflupromazine	BE0000405	Muscarinic acetylcholine receptor M4	0.73033
DB00589	Lisuride	BE0000945	5-hydroxytryptamine receptor 6	0.73
DB01233	Metoclopramide	BE0000451	5-hydroxytryptamine receptor 2A	0.72978
DB00502	Haloperidol	BE0000533	5-hydroxytryptamine receptor 2C	0.72945
DB00805	Minaprine	BE0000650	5-hydroxytryptamine receptor 7	0.72942
DB00934	Maprotiline	BE0000289	Alpha-2A adrenergic receptor	0.72935
DB01621	Pipotiazine	BE0000533	5-hydroxytryptamine receptor 2C	0.7292
DB00656	Trazodone	BE0000389	D(4) dopamine receptor	0.72905
DB00988	Dopamine	BE0000311	5-hydroxytryptamine receptor 3A	0.72896
DB00831	Trifluoperazine	BE0000575	Alpha-1B adrenergic receptor	0.72874
DB01151	Desipramine	BE0000342	Alpha-2C adrenergic receptor	0.7279
DB00268	Ropinirole	BE0000945	5-hydroxytryptamine receptor 6	0.72778
DB01242	Clomipramine	BE0000442	Histamine H1 receptor	0.72753
DB00656	Trazodone	BE0000659	5-hydroxytryptamine receptor 1D	0.72709
DB00726	Trimipramine	BE0000092	Muscarinic acetylcholine receptor M1	0.72708
DB00589	Lisuride	BE0000442	Histamine H1 receptor	0.72613
DB01233	Metoclopramide	BE0000560	Muscarinic acetylcholine receptor M2	0.72598
DB05271	Rotigotine	BE0000501	Alpha-1A adrenergic receptor	0.7259
DB01614	Acepromazine	BE0000581	D(3) dopamine receptor	0.72528
DB00350	Minoxidil	BE0000262	Prostaglandin G/H synthase 2	0.72524

DB00800	Fenoldopam	BE0000659	5-hydroxytryptamine receptor 1D	0.72508
DB00546	Adinazolam	BE0000736	Translocator protein	0.72471
DB00934	Maprotiline	BE0004889	D(1B) dopamine receptor	0.72399
DB00413	Pramipexole	BE0000945	5-hydroxytryptamine receptor 6	0.72394
DB00216	Eletriptan	BE0000581	D(3) dopamine receptor	0.72355
DB00652	Pentazocine	BE0000420	Delta-type opioid receptor	0.72335
DB06288	Amisulpride	BE0000389	D(4) dopamine receptor	0.72303
DB01608	Propericiazine	BE0000451	5-hydroxytryptamine receptor 2A	0.72288
DB00320	Dihydroergotamine	BE0000533	5-hydroxytryptamine receptor 2C	0.72264
DB00679	Thioridazine	BE0000389	D(4) dopamine receptor	0.72262
DB00714	Apomorphine	BE0000442	Histamine H1 receptor	0.72256
DB00248	Cabergoline	BE0000146	Histamine H4 receptor	0.72239
DB00745	Modafinil	BE0000501	Alpha-1A adrenergic receptor	0.72212
DB00462	Methylscopolamine bromide	BE0000405	Muscarinic acetylcholine receptor M4	0.7217
DB00517	Anisotropine Methylbromide	BE0000405	Muscarinic acetylcholine receptor M4	0.72097
DB00831	Trifluoperazine	BE0000020	D(1A) dopamine receptor	0.72089
DB01618	Molindone	BE0000045	Muscarinic acetylcholine receptor M3	0.72058
DB00216	Eletriptan	BE0000342	Alpha-2C adrenergic receptor	0.72041
DB01062	Oxybutynin	BE0000405	Muscarinic acetylcholine receptor M4	0.72036
DB00434	Cyproheptadine	BE0000291	5-hydroxytryptamine receptor 1A	0.72031
DB00215	Citalopram	BE0000045	Muscarinic acetylcholine receptor M3	0.72023
DB00696	Ergotamine	BE0000647	Sodium-dependent dopamine transporter	0.71961
DB00187	Esmolol	BE0000694	Beta-2 adrenergic receptor	0.71893
DB00246	Ziprasidone	BE0000112	Histamine H2 receptor	0.71878
DB00629	Guanabenz	BE0000342	Alpha-2C adrenergic receptor	0.71872
DB01594	Cinolazepam	BE0000736	Translocator protein	0.71781
DB01175	Escitalopram	BE0000575	Alpha-1B adrenergic receptor	0.71748
DB00986	Glycopyrrolate	BE0000405	Muscarinic acetylcholine receptor M4	0.71741
DB01242	Clomipramine	BE0000650	5-hydroxytryptamine receptor 7	0.71674
DB00934	Maprotiline	BE0000342	Alpha-2C adrenergic receptor	0.71661
DB00934	Maprotiline	BE0000311	5-hydroxytryptamine receptor 3A	0.71654
DB00420	Promazine	BE0000797	5-hydroxytryptamine receptor 1B	0.71633
DB01621	Pipotiazine	BE0000581	D(3) dopamine receptor	0.71614
DB00247	Methysergide	BE0000501	Alpha-1A adrenergic receptor	0.71581
DB00193	Tramadol	BE0000451	5-hydroxytryptamine receptor 2A	0.71578
DB01409	Tiotropium	BE0000405	Muscarinic acetylcholine receptor M4	0.71555
DB06204	Tapentadol	BE0000647	Sodium-dependent dopamine transporter	0.71528
DB00458	Imipramine	BE0004863	Alpha-1D adrenergic receptor	0.71526
DB00216	Eletriptan	BE0000572	Alpha-2B adrenergic receptor	0.7151
DB00543	Amoxapine	BE0000112	Histamine H2 receptor	0.71469
DB01200	Bromocriptine	BE0000476	5-hydroxytryptamine receptor 1E	0.7146
DB01255	Lisdexamfetamine	BE0000501	Alpha-1A adrenergic receptor	0.71453

DB00568	Cinnarizine	BE0002354	Voltage-dependent L-type calcium channel subunit beta 2	0.71439
DB06711	Naphazoline	BE0000572	Alpha-2B adrenergic receptor	0.71436
DB01069	Promethazine	BE0000581	D(3) dopamine receptor	0.71416
DB00568	Cinnarizine	BE0000451	5-hydroxytryptamine receptor 2A	0.71411
DB00728	Rocuronium	BE0000045	Muscarinic acetylcholine receptor M3	0.71402
DB00953	Rizatriptan	BE0000291	5-hydroxytryptamine receptor 1A	0.7139
DB01104	Sertraline	BE0000486	Sodium-dependent noradrenaline transporter	0.71389
DB00216	Eletriptan	BE0000020	D(1A) dopamine receptor	0.71378
DB00383	Oxyphencyclimine	BE0000405	Muscarinic acetylcholine receptor M4	0.7136
DB00751	Epinastine	BE0000393	5-hydroxytryptamine receptor 2B	0.7135
DB01085	Pilocarpine	BE0000405	Muscarinic acetylcholine receptor M4	0.71339
DB00656	Trazodone	BE0000945	5-hydroxytryptamine receptor 6	0.71308
DB01156	Bupropion	BE0000749	Sodium-dependent serotonin transporter	0.71307
DB01100	Pimozide	BE0000501	Alpha-1A adrenergic receptor	0.71283
DB00477	Chlorpromazine	BE0000647	Sodium-dependent dopamine transporter	0.71282
DB01403	Methotrimeprazine	BE0000476	5-hydroxytryptamine receptor 1E	0.71276
DB01175	Escitalopram	BE0000045	Muscarinic acetylcholine receptor M3	0.71242
DB00734	Risperidone	BE0004889	D(1B) dopamine receptor	0.71186
DB01233	Metoclopramide	BE0000045	Muscarinic acetylcholine receptor M3	0.71165
DB01392	Yohimbine	BE0000715	Alpha-1D adrenergic receptor	0.71158
DB00751	Epinastine	BE0000945	5-hydroxytryptamine receptor 6	0.7114
DB01175	Escitalopram	BE0000560	Muscarinic acetylcholine receptor M2	0.71138
DB00502	Haloperidol	BE0000389	D(4) dopamine receptor	0.71136
DB00964	Apraclonidine	BE0000715	Alpha-1D adrenergic receptor	0.71131
DB00490	Buspirone	BE0000451	5-hydroxytryptamine receptor 2A	0.71109
DB01151	Desipramine	BE0000572	Alpha-2B adrenergic receptor	0.71082
DB00334	Olanzapine	BE0004864	Alpha-2C adrenergic receptor	0.71082
DB01135	Doxacurium chloride	BE0000045	Muscarinic acetylcholine receptor M3	0.71054
DB04946	Iloperidone	BE0000146	Histamine H4 receptor	0.7105
DB01151	Desipramine	BE0000145	D(1B) dopamine receptor	0.71039
DB00751	Epinastine	BE0000389	D(4) dopamine receptor	0.71034
DB06216	Asenapine	BE0000560	Muscarinic acetylcholine receptor M2	0.70979
DB00726	Trimipramine	BE0000560	Muscarinic acetylcholine receptor M2	0.70941
DB06709	Methacholine	BE0000560	Muscarinic acetylcholine receptor M2	0.70915
DB00216	Eletriptan	BE0000389	D(4) dopamine receptor	0.70892
DB08815	Lurasidone	BE0000145	D(1B) dopamine receptor	0.70879
DB00363	Clozapine	BE0000112	Histamine H2 receptor	0.70815
DB08815	Lurasidone	BE0000945	5-hydroxytryptamine receptor 6	0.70797
DB00202	Succinylcholine	BE0000405	Muscarinic acetylcholine receptor M4	0.70734
DB06288	Amisulpride	BE0000393	5-hydroxytryptamine receptor 2B	0.70724
DB04946	Iloperidone	BE0000476	5-hydroxytryptamine receptor 1E	0.70714
DB01119	Diazoxide	BE0000535	Carbonic anhydrase 4	0.70698

DB01151	Desipramine	BE0000146	Histamine H4 receptor	0.70584
DB01069	Promethazine	BE0000291	5-hydroxytryptamine receptor 1A	0.70581
DB01392	Yohimbine	BE0000945	5-hydroxytryptamine receptor 6	0.70539
DB00332	Ipratropium bromide	BE0000405	Muscarinic acetylcholine receptor M4	0.70527
DB00370	Mirtazapine	BE0000112	Histamine H2 receptor	0.70469
DB00193	Tramadol	BE0000647	Sodium-dependent dopamine transporter	0.7046
DB00904	Ondansetron	BE0000659	5-hydroxytryptamine receptor 1D	0.70453
DB01587	Ketazolam	BE0000523	Gamma-aminobutyric acid receptor subunit alpha-3	0.70431
DB00450	Droperidol	BE0000575	Alpha-1B adrenergic receptor	0.70421
DB06216	Asenapine	BE0000045	Muscarinic acetylcholine receptor M3	0.70377
DB00714	Apomorphine	BE0000476	5-hydroxytryptamine receptor 1E	0.70366
DB00934	Maprotiline	BE0000146	Histamine H4 receptor	0.70365
DB01118	Amiodarone	BE0000694	Beta-2 adrenergic receptor	0.70334
DB00715	Paroxetine	BE0000647	Sodium-dependent dopamine transporter	0.70323
DB00215	Citalopram	BE0000575	Alpha-1B adrenergic receptor	0.70311
DB08910	Pomalidomide	BE0000017	Prostaglandin G/H synthase 1	0.70299
DB01364	Ephedrine	BE0000749	Sodium-dependent serotonin transporter	0.70291
DB01622	Thioproperazine	BE0000650	5-hydroxytryptamine receptor 7	0.70277
DB01614	Acepromazine	BE0000650	5-hydroxytryptamine receptor 7	0.70273
DB06216	Asenapine	BE0000247	Muscarinic acetylcholine receptor M5	0.70233
DB00805	Minaprine	BE0000247	Muscarinic acetylcholine receptor M5	0.702
DB00805	Minaprine	BE0000342	Alpha-2C adrenergic receptor	0.70184
DB00726	Trimipramine	BE0000247	Muscarinic acetylcholine receptor M5	0.7014
DB00320	Dihydroergotamine	BE0000451	5-hydroxytryptamine receptor 2A	0.70088
DB00805	Minaprine	BE0000575	Alpha-1B adrenergic receptor	0.70087
DB00934	Maprotiline	BE0000715	Alpha-1D adrenergic receptor	0.70086
DB00413	Pramipexole	BE0000442	Histamine H1 receptor	0.70055
DB00751	Epinastine	BE0000659	5-hydroxytryptamine receptor 1D	0.70036

APPENDIX C

THE DESCRIPTION OF THE CONTENTS OF BALESTRAWEB.ZIP

The code and auxiliary files to run BalestraWeb are accessible online at http://balestra.csb.pitt.edu/static/balestraweb.zip and the contents of this file are explained below:

Name	**Description of contents**
balestraweb.py	The code that runs BalestraWeb (Python)
cabinet	Contains the data files that BalestraWeb uses in Python shelve format
html	Contains the HTML files that BalestraWeb serves to the users
static	Contains the static files (i.e. the figures in the tutorial, BalestraWeb icon, etc) that BalestraWeb serves to the users.
helpers	Contains the code to generate BalestraWeb data. To be used as follows: 1) Run learn_multi_model.m (Matlab) 2) Run getDrugBank.py (Python) Doing the above re-generates all the files in the ‘cabinet’ folder.

APPENDIX D

LFM METHOD AND HYPERPARAMETER SEARCH RESULTS

To decide on the optimal approach to build the LFM of STITCH v3 that I used in the HD project, I have conducted a search of the best performing method and hyperparameter combination on STITCH v3 data by partitioning the data 16 times into training, testing sets allocating 90% of the interactions for training, 10% for testing using GraphLab PowerGraph software. The RMSE over these 16 iterations are averaged in the 'RMSE_mean' column, and the standard deviation of these 16 iteration results are provided in the 'RMSE_std' column. The parameter text starts with the shorthand name of the method and then underscore ('_') character is used to delimit the parameters. The full list of results can be downloaded here:

http://balestra.csb.pitt.edu/static/all_results.xlsx

APPENDIX E

TABLE OF *C. ELEGANS* GENES THAT CAUSE REDUCED ATZ AGGREGATION PHENOTYPE UPON RNA INTERFERENCE KNOCKDOWN

Batch ID	Batch Date	Inhibition target genes
1	10-27	**T01G9.3**,*F30A10.7*
3	11-10	R06C1.6,Y53C10A.10,T09E11.9,T15D6.8
4	11-11	*Y65B4B_10.d*
5	11-19	W10D9.5,C08G5.1
6	11-20	C16C4.11
7	11-21	W10G11.12
8	12-11	T12F5.3,C50F2.5,R12E2.13
9	12-12	F33D11.9
11	12-20	*R05G9.c*,C18H9.6,C18H9.7,C18H9.8,T14B4.5,F18A1.4,T05H10.4
12	12-21	W01C9.2,ZK1321.4
17	4-14	C16C10.3
19	4-17	ZK1098.6,C48B4.12a,T05G5.9,C05B5.5,C05B5.6,M04D8.3,M04D8.4, M04D8.5,T20G5.6
20	4-19	*Y119D3_456.a*
21	4-20	*C45G7.6*
22	4-21	F36H12.15
23	4-22	K02B2.6,*T13A10.8*,C06G3.9,C34D4.1
24	4-23	*C49H3.2*,C49H3.5,C49H3.6,*C49H3.7*
27	6-11	R09E12.4,*R09E12.5*,*R09E12.7*,R13D11.3
32	6-18	ZK262.9
35	6-21	**C05A9.1**,C05C9.3,**F13D2.2**,F08B12.1

APPENDIX F

LIST OF CHEMICALS PREDICTED TO REDUCE AGGREATION OF ATZ THROUGH THE ALPHA-1 ANTITRYPSIN HIGH CONTENT SCREENING RESULT ANALYSIS CHEMICAL HIT DIVERSIFICATION METHOD

The columns contain the following information from the left to right:

- ZINC_ID: ZINC compound identifier
- MWT: Molecular weight
- LogP: Partition coefficient
- Desolv_apolar: Apolar desolvation energy (kcal/mol)
- Desolv_polar: Polar desolvation energy (kcal/mol)
- HBD: Number of hydrogen bond donors
- HBA: Number of hydrogen bond acceptors
- tPSA: Topological polar surface area (suspected to be in Å^{2}, exact specification not found in ZINC documentation)
- Charge: Net charge of the molecule
- NRB: Number of rotatable bonds
- Cluster size: Total number of chemicals in the same cluster as this chemical

ZINC_ID	MWT	LogP	Desolv_apolar	Desolv_polar	HBD	HBA	tPSA	Charge	NRB	Cluster
ZINC75662250	297.707	4.24	7.85	-7.72	1	1	12	0	1	797
ZINC67287455	445.516	3.04	7.63	-11.21	2	9	106	0	10	643
ZINC00641264	365.433	3.77	6.3	-13.72	3	6	79	0	4	588
ZINC20283449	438.503	3.78	7.12	-11.36	2	7	74	0	7	585
ZINC04482400	304.346	2.93	-0.56	-8.42	2	6	76	0	6	529
ZINC72190830	251.251	2.1	6.57	-4.16	2	1	26	0	3	521
ZINC04625454	228.23	1.84	5.86	-9.12	2	4	56	0	1	426
ZINC43568601	222.218	2.31	4.27	-29.6	4	3	52	1	2	426
ZINC83664119	218.255	1.64	4.98	-36.24	3	3	42	1	1	425

ZINC00073782	316.401	4.25	0.41	-7.33	2	5	67	0	5	409
ZINC20906829	396.53	3.89	9.43	-43.05	1	4	20	1	4	403
ZINC00716297	360.366	2.26	-2.69	-10.48	2	8	95	0	5	395
ZINC76216131	354.381	3.71	9.04	-7.71	2	5	67	0	5	374
ZINC38276461	194.273	2.8	5.97	-39.59	3	1	28	1	1	370
ZINC34940780	250.362	3.47	7.47	-48.25	3	1	28	1	2	357
ZINC00041497	292.291	1.33	1.64	-11.1	3	7	97	0	4	354
ZINC19271873	200.664	2.85	5.98	-38.34	2	1	17	1	1	350
ZINC65516145	344.333	1.52	5.5	-16.04	1	5	51	0	5	347
ZINC75688609	297.251	3.65	4.71	-7.02	2	2	32	0	1	346
ZINC43669139	345.374	2.77	5.1	-13.07	3	6	79	0	5	344
ZINC23358157	393.641	3.51	8.96	-77.07	3	4	32	2	7	340
ZINC01430262	338.72	4.14	2.65	-14.14	2	4	56	0	3	339
ZINC00060635	344.411	3.96	7.36	-8.96	2	6	77	0	5	337
ZINC75961665	239.289	1.87	5.23	-39.96	2	2	20	1	1	332
ZINC01815570	552.515	5.82	5.51	-27.97	1	5	66	0	8	331
ZINC22765711	352.406	2.41	4.57	-11.82	1	4	42	0	5	331
ZINC01257249	215.251	2.23	6	-34.34	4	2	52	1	2	327
ZINC02414084	514.544	4.98	3.35	-29.47	1	6	75	0	8	323
ZINC67803975	370.44	1	7.69	-45.47	2	7	85	1	4	322
ZINC75644572	253.297	0.78	-1.7	-37.14	5	4	69	1	1	322
ZINC00143298	300.365	3.62	7.67	-9.25	2	4	97	0	1	316
ZINC36222011	252.341	3.81	9.77	-24.04	3	3	45	1	1	314
ZINC22799364	428.578	2.07	9.7	-42.44	1	6	46	1	5	311
ZINC12346325	393.443	2.89	-2.56	-18.54	2	7	80	0	6	310
ZINC20560246	431.536	4.46	10.94	-9.23	2	6	71	0	7	306
ZINC67898612	320.416	2.94	7.84	-32.61	2	4	42	1	4	305
ZINC57992309	348.364	2.77	7.28	-45.35	3	4	46	1	6	303
ZINC75778695	280.351	2.69	5.89	-39.41	3	4	51	1	2	301
ZINC05342165	387.475	3.69	7.03	-41.93	3	5	67	1	5	299
ZINC13353751	359.404	5.58	10.29	-11.31	3	4	57	0	3	293
ZINC75688714	257.255	2.37	4.3	-3.53	1	2	21	0	1	290
ZINC00193919	338.363	3.73	7.42	-9.31	2	6	77	0	5	289
ZINC19952371	228.319	3.88	6.77	-28.87	4	3	52	1	2	288
ZINC04898579	380.596	6.51	0.36	-35.47	3	2	36	1	7	287
ZINC76216342	304.346	3.11	5.37	-9.51	2	6	77	0	6	280
ZINC20450997	339.45	3.42	6.08	-32.95	2	3	28	1	5	279
ZINC01426657	374.461	2.99	7.98	-43.7	3	7	81	1	7	277
ZINC20389152	281.37	2.82	8.28	-36.71	1	2	8	1	3	271
ZINC20213933	335.36	1.13	1.83	-11.76	2	8	93	0	6	269
ZINC29538725	380.366	3.67	5.77	-12.77	1	5	51	0	6	267
ZINC95076866	361.391	3.42	11.83	-40.26	1	4	35	1	4	265
ZINC82741878	254.398	0.55	2.02	-41.38	2	4	32	1	3	263
ZINC12345825	365.408	3.43	-1.67	-15.66	2	5	61	0	5	262
ZINC31169347	425.598	3.73	9.01	-10.56	2	5	54	0	7	260
ZINC57478401	407.514	4.34	9.51	-8.69	2	6	71	0	9	259
ZINC40442615	182.218	2.16	2.76	-39.08	3	2	37	1	1	257
ZINC12593869	424.576	4.88	13.45	-93.72	2	4	27	2	9	251
ZINC82914685	204.253	0.97	3.61	-48.12	3	4	55	1	1	251
ZINC72290854	320.457	2.53	7.53	-42.98	3	5	55	1	8	250

ZINC72292602	314.769	2.62	3.01	-9.44	3	4	61	0	4	250
ZINC07538358	378.35	3.06	9.1	-17.6	1	5	59	0	6	249
ZINC75739987	277.698	1.75	3.34	-6.54	2	3	41	0	4	246
ZINC24205259	320.36	2.77	7.44	-21.77	1	8	86	0	3	245
ZINC44256447	400.454	3.29	5.4	-20.63	4	7	91	0	7	241
ZINC83370295	243.277	1.81	3.36	-43.67	2	3	29	1	3	240
ZINC75340352	312.329	2.64	2.9	-14.24	4	7	99	0	3	238
ZINC48270992	411.526	3.06	8.29	-47.88	4	7	84	1	7	234
ZINC00709748	462.496	5.16	1.1	-14.08	2	5	59	0	4	232
ZINC12995259	436.577	5.03	10.08	-13.16	1	5	51	0	6	229
ZINC19911229	372.375	3.89	7.7	-11.6	1	6	69	0	5	229
ZINC07405191	362.376	2.55	2.22	-18.7	1	5	58	0	6	227
ZINC52095677	255.308	3.94	5.28	-5.59	1	2	21	0	3	227
ZINC04854740	349.39	2.42	-2.7	-16.18	2	6	70	0	4	226
ZINC58006085	386.398	3.25	8.5	-23.95	1	5	59	0	7	224
ZINC04692860	371.452	5	11.74	-17.49	1	4	55	0	4	222
ZINC36222030	268.34	3.41	8.61	-26.23	3	4	54	1	2	222
ZINC05035242	369.403	4.78	11.92	-13.66	1	5	55	0	4	220
ZINC19725286	269.343	2.17	7.3	-47.39	4	2	43	1	2	220
ZINC36117077	536.713	6.72	16.68	-13.1	0	5	50	0	10	219
ZINC84638394	346.452	2.42	3.8	-7.73	3	5	71	0	5	218
ZINC75936661	274.254	0.86	5.24	-42.72	4	5	82	1	2	217
ZINC00715242	370.352	2.95	3.05	-9.03	3	5	78	0	4	216
ZINC85808113	409.848	5.79	9.57	-11.67	3	5	66	0	4	216
ZINC00236831	277.388	3.35	-0.7	-36.16	1	2	7	1	3	215
ZINC58860282	377.847	3.94	7.93	-11.36	2	5	54	0	5	215
ZINC76004451	243.232	2.78	6.86	-7.67	0	3	31	0	2	215
ZINC05791507	276.361	2.28	6.33	-15.35	2	4	50	0	3	214
ZINC38050594	233.299	0.87	4.48	-48.12	5	6	94	1	2	213
ZINC43080933	346.451	2.1	6.9	-44.97	3	6	64	1	4	213
ZINC71799351	322.311	2.7	2.93	-12.79	3	5	71	0	6	213
ZINC75688542	271.282	2.75	5.05	-2.85	1	2	21	0	1	213
ZINC05512248	317.291	4.34	7.28	-9.78	1	4	51	0	4	212
ZINC06646225	386.791	2.64	6.5	-8.32	1	7	77	0	4	212
ZINC89873258	299.33	1.54	1.7	-18.73	3	6	79	0	5	211
ZINC20284055	401.531	3.13	8.59	-38.19	3	7	75	1	8	210
ZINC95985360	310.369	2.46	2.98	-14.96	3	5	71	0	5	209
ZINC23549858	203.309	1.96	6.23	-38.64	2	2	20	1	1	208
ZINC42872162	218.284	0.32	4.01	-31.29	5	5	81	1	3	208
ZINC65396788	336.548	1.88	8.84	-173.28	4	5	42	3	5	207
ZINC16137956	263.405	3.51	7.38	-33.56	2	3	37	1	1	206
ZINC01262488	406.526	5.19	9.63	-11.58	2	5	60	0	3	205
ZINC52267717	390.479	4	8.09	-45.2	3	6	64	1	8	205
ZINC67689658	492.576	3.74	7.26	-10.97	2	9	92	0	8	204
ZINC75873243	245.277	2.27	5.29	-34.87	4	3	61	1	3	204
ZINC76236380	348.402	3.99	8.66	-12.87	2	5	67	0	3	204
ZINC38041437	211.353	0.46	4.19	-81.04	4	3	33	2	2	202
ZINC23372937	277.388	1.23	3.35	-41.39	2	4	37	1	4	201
ZINC04368757	296.434	4.42	-0.98	-33.96	3	2	36	1	2	199
ZINC04473477	383.377	3.23	6.21	-12.51	2	5	101	0	2	199

ZINC02086895	370.476	4.64	-1.44	-36.78	3	4	59	1	4	197
ZINC11569834	256.304	2.56	-1	-50.02	3	3	45	1	3	195
ZINC67898809	401.49	2.98	7.39	-43.21	2	6	65	1	5	195
ZINC12114675	393.45	1.66	1.9	-71.4	1	8	79	1	5	193
ZINC72021342	443.465	5.21	11.54	-17.13	1	4	48	0	7	193
ZINC74327156	281.212	3.85	7.35	-10.19	1	3	42	0	2	193
ZINC85559648	358.394	2.1	2.82	-13.1	3	7	89	0	6	193
ZINC38073354	260.332	4	8.14	-40.93	2	2	26	1	4	191
ZINC56443193	416.517	3.58	8.56	-54.03	3	6	64	1	8	191
ZINC82507215	291.459	3.06	6.95	-41.24	2	3	29	1	5	190
ZINC82529272	174.227	1.39	3.43	-7.83	3	3	52	0	1	188
ZINC15836890	484.518	5.45	11.9	-25.97	2	5	67	0	5	187
ZINC19326510	380.482	5.28	15.18	-111.17	2	2	9	2	5	187
ZINC32905567	269.3	2.95	7.65	-20.62	1	4	55	0	4	186
ZINC52451500	409.453	3	9.93	-44.62	1	6	52	1	11	186
ZINC05201736	403.438	2.82	-2.81	-21.61	2	7	79	0	3	185
ZINC00182971	342.42	3.71	-1.78	-8.49	2	5	67	0	5	184
ZINC19838722	255.451	2.08	7.34	-34.72	1	2	8	1	2	184
ZINC67803167	371.464	3.69	9.37	-39.56	2	5	55	1	4	184
ZINC00720129	432.545	4.5	-0.26	-13.11	2	5	59	0	4	182
ZINC75693137	265.259	2.78	4.43	-6.88	1	3	34	0	3	182
ZINC83352758	224.299	2.88	5.69	-37.92	2	2	26	1	1	182
ZINC22766897	320.438	2.65	6.23	-44.72	3	5	55	1	6	180
ZINC19332662	429.564	3.34	9.44	-53.35	1	7	61	1	6	179
ZINC20573365	398.487	1.07	6.09	-55.52	4	8	101	1	6	179
ZINC62667439	275.376	0.12	2.74	-25.76	3	5	53	1	4	178
ZINC00641398	374.128	4.48	-0.49	-8.28	2	3	41	0	2	177
ZINC79002839	290.363	1.72	8.69	-24.47	1	5	59	0	7	175
ZINC00710766	399.537	4.11	11.38	-16.13	1	4	55	0	4	173
ZINC08609006	476.552	4.09	9.78	-16.54	3	7	83	0	8	173
ZINC33009209	444.535	4.95	8.69	-18.78	3	7	83	0	6	173
ZINC24831739	353.777	4.94	7.94	-16.08	1	3	38	0	4	172
ZINC24839382	480.473	4.08	9.7	-16.57	2	10	129	0	11	172
ZINC60974503	428.646	3.67	12.09	-90.67	4	5	50	2	7	172
ZINC00727574	413.517	5.07	1.64	-11.55	1	4	47	0	3	171
ZINC00823629	407.514	4.65	0.05	-9.02	2	6	70	0	9	171
ZINC06051337	328.462	4.1	-2.48	-12.48	2	3	41	0	3	170
ZINC20773830	349.498	3.58	10.79	-44.06	1	3	17	1	3	170
ZINC45946335	220.271	1.64	5.7	-34.24	3	3	42	1	5	170
ZINC13636088	368.452	5.77	12.71	-12.65	2	2	24	0	2	169
ZINC43042690	197.277	1.25	2.53	-41.77	3	2	31	1	3	168
ZINC35451500	461.585	3.94	14.59	-108.16	4	6	68	2	8	167
ZINC95426014	344.435	1.69	5.26	-40.06	4	6	79	1	6	167
ZINC32629276	295.378	2.58	7.05	-42.41	3	4	57	1	3	164
ZINC76217380	343.21	4.34	7.04	-7.98	2	5	67	0	5	164
ZINC13038176	426.542	3.51	6.72	-9.83	2	7	74	0	7	163
ZINC19550215	402.486	2.72	5.31	-49.33	3	5	63	1	7	163
ZINC62665393	271.315	2.55	7.06	-35.16	2	3	35	1	4	162
ZINC04479596	343.358	2.35	-1.52	-17.49	2	6	70	0	4	161
ZINC19334644	332.512	2.13	4.43	-34.91	2	4	31	1	5	161

ZINC00374241	321.323	3.86	2.68	-11.53	1	4	47	0	5	160
ZINC83379541	249.378	2.84	5.78	-40.16	2	3	29	1	4	159
ZINC76121475	207.321	1.08	5.42	-94.78	4	3	47	2	1	158
ZINC39199636	336.377	4.04	9.7	-41.71	2	2	25	1	4	157
ZINC82753682	222.243	0.95	1.48	-12.86	5	4	72	0	3	157
ZINC00072989	284.315	2.19	-1.09	-15.44	2	5	59	0	3	156
ZINC05164248	401.571	4.83	-2.52	-40.18	3	5	62	1	5	156
ZINC02646442	316.748	2.47	6.6	-10.41	1	6	69	0	5	155
ZINC09482238	354.4	4.58	1.07	-12.57	2	3	41	0	1	154
ZINC77094259	248.371	1.54	6.22	-46.11	3	2	37	1	3	154
ZINC31094802	342.414	3.36	7.14	-10.85	2	4	50	0	5	153
ZINC00038297	346.385	3.9	-1.52	-41.78	3	4	59	1	4	152
ZINC20465714	350.439	0.91	1.92	-44.58	5	7	95	1	5	152
ZINC72190711	225.597	0.95	1.12	-5.73	3	2	46	0	2	151
ZINC13154929	294.398	4.23	7.07	-4.52	2	3	45	0	2	150
ZINC06738904	361.424	5.26	11.6	-16.22	1	5	55	0	3	149
ZINC19785836	235.351	2.29	7.17	-33.06	3	3	40	1	5	149
ZINC31912905	315.505	3.99	10.83	-104.62	4	3	46	2	5	149
ZINC52003244	240.347	2.41	2.52	-9.73	3	4	61	0	2	149
ZINC77403636	247.362	2.39	5.32	-43.03	2	3	29	1	2	149
ZINC07406143	370.38	3.66	1.83	-18.56	1	6	71	0	5	148
ZINC12522205	324.38	4.31	7.89	-12.85	2	3	41	0	2	148
ZINC19326582	396.481	4.91	13.97	-128.43	2	3	18	2	6	148
ZINC27579123	399.521	4.52	8.53	-50.26	2	3	29	1	6	147
ZINC44709555	378.444	2.71	8.47	-16.27	0	6	59	0	4	146
ZINC06529588	260.268	3.11	0.28	-11.38	2	4	50	0	4	145
ZINC12776552	375.371	3.01	8.14	-14.78	0	5	48	0	4	145
ZINC82869007	176.219	1.63	3.35	-5.59	2	3	37	0	3	145
ZINC75738600	269.291	2.75	5.23	-7.97	1	3	30	0	4	143
ZINC76217090	318.373	3.29	6.16	-9.33	2	6	77	0	6	142
ZINC19894425	321.352	1.02	3.57	-24.58	1	6	63	0	5	141
ZINC36446804	390.48	4.69	8.82	-11.23	1	3	32	0	4	140
ZINC70270920	283.318	3.74	6.56	-12.11	1	3	38	0	5	140
ZINC72305915	409.958	3.98	6.46	-5.98	3	6	74	0	4	139
ZINC02547536	209.122	0.82	1.11	-30.17	4	3	58	1	2	138
ZINC43193442	403.572	3.35	8.34	-44.43	3	6	58	1	8	138
ZINC00718165	406.938	5.31	-1.46	-10.48	2	3	41	0	2	136
ZINC65497500	244.322	0.99	5.06	-34.34	3	5	58	1	2	136
ZINC90744993	334.322	2.52	3.67	-14.74	3	5	71	0	4	136
ZINC16946471	193.246	3.09	5.16	-4.21	2	1	26	0	1	133
ZINC83690790	260.317	1.12	4.12	-41.37	3	5	60	1	3	132
ZINC15708530	420.559	0.93	5.02	-53.62	4	8	95	1	5	130
ZINC02316333	403.529	3.26	-3.65	-18.23	4	6	93	0	6	129
ZINC32624081	466.362	4.53	6.7	-17.1	3	6	92	0	5	129
ZINC12496088	397.812	6.07	10.62	-12	1	4	39	0	3	128
ZINC16940266	204.293	2.03	3.24	-40.23	3	2	37	1	1	127
ZINC01426909	488.584	5.1	12.66	-15.22	2	5	67	0	5	126
ZINC13033715	505.574	2.73	11.4	-48.52	2	9	99	1	7	126
ZINC19697575	445.585	2.34	7	-13.64	1	7	71	0	7	126
ZINC82505211	220.336	2.53	6.67	-41.15	2	2	26	1	1	126

ZINC83727166	225.262	1.92	3.57	-37.64	3	2	29	1	1	126
ZINC00106521	292.697	3.85	2.28	-13.26	1	4	51	0	4	125
ZINC06603031	411.408	3.11	-1.88	-12.37	4	6	93	0	7	125
ZINC70636341	315.316	2.66	4.22	-9.25	2	3	49	0	5	125
ZINC72002875	440.539	4.02	11.9	-52.52	2	6	55	1	4	125
ZINC09464150	448.947	5.28	1.08	-17.7	1	5	58	0	7	124
ZINC83322304	225.381	2.09	4.56	-39.85	2	2	20	1	2	124
ZINC03626489	399.34	2.99	3.31	-13.2	2	5	75	0	6	123
ZINC04387912	402.419	3.01	-6.63	-17.05	4	5	81	0	4	123
ZINC67801214	285.408	3.84	5.34	-44.01	2	5	52	1	5	123
ZINC09113497	461.558	5.47	13.38	-19.63	1	6	78	0	6	122
ZINC20283200	424.521	3.67	9.62	-59.42	4	7	93	1	11	121
ZINC31932998	310.421	3.95	11.61	-44.06	1	4	31	1	6	121
ZINC62713739	303.332	2.93	8.36	-32.27	2	3	35	1	4	121
ZINC82368377	219.283	3.05	4.83	-36.19	4	2	43	1	1	121
ZINC02528993	253.223	4.14	1.49	-3.13	2	2	35	0	3	120
ZINC75869391	223.295	0.16	3.83	-90.15	6	3	68	2	2	119
ZINC76254880	299.331	2.3	6.25	-45.22	2	5	81	-1	3	119
ZINC71506668	195.217	0.16	1.13	-57.92	4	3	57	1	1	118
ZINC00181081	336.395	5.42	10.32	-12.32	0	6	57	0	1	117
ZINC23359737	438.926	2.98	7.91	-36.99	3	5	55	1	9	117
ZINC26387440	354.815	2.73	5	-20	1	6	76	0	5	116
ZINC82795137	285.386	2.69	6.27	-49.61	3	2	31	1	2	116
ZINC01686012	196.273	3.12	-0.5	-40.05	2	1	16	1	2	115
ZINC07940530	366.464	2.51	-3.13	-17.76	1	6	75	0	6	115
ZINC15734349	409.49	3.07	7.5	-14.53	2	8	81	0	3	115
ZINC40007529	256.305	2.7	7.18	-11.72	2	4	50	0	5	115
ZINC08672544	460.528	3.92	2.69	-48.79	1	7	64	1	7	114
ZINC13497967	480.605	5.72	12.84	-13.4	1	5	59	0	8	114
ZINC82729967	214.292	2.35	5.78	-31.87	3	3	43	1	2	114
ZINC77090631	296.312	4.16	7.83	-43.84	3	2	41	1	3	113
ZINC11570917	267.396	3.75	0.27	-39.14	2	2	20	1	2	112
ZINC12346198	351.381	2.96	-1.6	-16.57	2	5	61	0	3	112
ZINC20719888	393.414	3.46	9.49	-46.52	3	7	89	1	5	110
ZINC28434156	368.477	3.99	9.92	-18.5	1	5	59	0	5	110
ZINC35774882	217.358	0.04	0.8	-45.86	3	3	40	1	2	110
ZINC19841532	301.863	2.71	6.18	-41.71	1	3	17	1	3	109
ZINC55226540	413.905	4.3	8.41	-8.53	3	6	79	0	8	109
ZINC05071760	367.412	2.73	6.73	-13.83	3	6	83	0	3	108
ZINC32795799	290.363	2.15	7.85	-22.21	1	5	59	0	6	108
ZINC03840100	452.511	1.37	-9.9	-26.08	7	9	153	0	7	107
ZINC44016120	355.369	3.28	6.24	-12.5	3	6	83	0	4	107
ZINC00058753	289.335	2.1	-0.72	-8.68	2	6	70	0	4	106
ZINC02520974	471.548	4.39	13.87	-16.9	2	4	70	0	3	104
ZINC04966738	398.776	3.05	-0.36	-6.78	0	7	68	0	4	104
ZINC00184514	242.238	1.11	2.92	-9.65	2	6	76	0	1	103
ZINC15955536	462.556	4.71	13.04	-9.41	0	4	33	0	4	103
ZINC75848180	178.175	-1.25	0.03	-32.13	4	6	88	1	2	103
ZINC00228756	246.313	4.19	8.08	-6.37	2	2	24	0	1	102
ZINC02050214	437.225	4.23	1.97	-40.92	3	6	81	1	7	102

ZINC04479606	369.421	2.64	5.86	-18.53	2	7	80	0	6	102
ZINC00092578	223.272	1.36	0.62	-6.01	3	4	62	0	2	101
ZINC33009242	432.499	5.05	9.46	-18.41	3	6	73	0	5	101
ZINC50182129	195.31	0.6	4.11	-88.21	5	3	58	2	2	101
ZINC01060004	404.47	4.01	11.26	-14.31	2	7	85	0	6	100
ZINC20213814	385.489	3.53	6.94	-10.14	3	6	79	0	8	97
ZINC24926041	394.374	1.85	7.31	-14.98	1	7	85	0	6	97
ZINC00087289	340.437	3.29	8.94	-15.99	0	6	69	0	3	96
ZINC04373552	339.395	2.65	-1.94	-15.49	2	6	70	0	4	96
ZINC75688453	257.279	2.17	4.45	-36.32	3	2	29	1	1	96
ZINC95347952	257.299	4.4	7.68	-2.01	1	1	12	0	4	96
ZINC50843460	211.191	2.25	2.84	-31.47	3	3	49	1	2	95
ZINC10337067	433.854	3.93	11.48	-44.05	2	7	82	1	2	94
ZINC02361911	293.482	4.81	-1.73	-9.64	1	1	12	0	0	92
ZINC19834162	367.348	4.22	8.34	-18.07	1	6	74	0	7	92
ZINC35526454	355.369	3.67	7.42	-10.99	3	6	83	0	6	92
ZINC76039205	286.281	3.68	7.63	-11.91	0	3	30	0	3	92
ZINC01126331	368.742	4.23	2.42	-17.7	1	4	50	0	4	91
ZINC12156665	335.399	2.5	1.36	-50.33	2	5	65	1	6	90
ZINC13220287	383.527	3.62	9.83	-16.41	4	5	69	0	5	90
ZINC40747107	225.4	2.87	4.81	-34.79	2	2	20	1	4	90
ZINC75486190	248.35	2.71	5.1	-27.33	4	4	61	1	3	90
ZINC08039586	388.469	2.69	-2.13	-51.21	3	7	81	1	4	89
ZINC10336776	440.483	4.42	11	-38.15	2	8	88	1	4	89
ZINC18700207	335.4	4.98	9.37	-9.25	1	2	29	0	4	89
ZINC35561767	358.463	4.09	9.09	-10.34	1	5	51	0	2	89
ZINC02066671	514.973	7.22	-0.29	-6.9	2	2	41	0	3	88
ZINC15017282	358.488	4.03	10.22	-9.94	2	4	50	0	5	87
ZINC95080731	185.23	1.55	3.06	-87.35	4	3	54	2	0	87
ZINC20371077	304.396	2.71	6.36	-14.55	2	4	58	0	2	86
ZINC72270271	285.294	1.22	1.28	-9.52	3	5	65	0	4	86
ZINC05783964	301.346	1.33	-3.55	-11.45	2	6	70	0	2	85
ZINC13471085	303.314	1.27	1.49	-16.46	3	6	96	0	3	84
ZINC04908497	353.487	1.27	5.12	-45.85	3	7	75	1	4	83
ZINC19725288	281.379	2.06	6.62	-46.76	4	3	53	1	3	82
ZINC75660792	243.733	2.45	5.77	-29.68	2	2	16	1	2	82
ZINC12085435	409.461	2.69	-3.49	-17.29	2	6	71	0	2	81
ZINC08435175	685.698	7.68	6.1	-11.32	0	6	72	0	8	80
ZINC00083131	347.321	4.3	-2.21	-10.94	3	6	79	0	5	77
ZINC02758246	458.602	6.38	15.41	-19.05	1	5	59	0	7	77
ZINC26513997	212.293	1.59	3.84	-5.84	2	2	29	0	1	77
ZINC04857304	386.61	4.35	-2.36	-45.05	3	3	28	1	6	76
ZINC07652999	405.878	5.17	1.19	-13.24	1	4	55	0	6	76
ZINC08788961	482.621	3.35	-0.32	-14.49	2	7	101	0	6	76
ZINC18197292	485.462	4.26	6.62	-24.36	3	7	96	0	5	76
ZINC27496201	337.678	4.37	6.23	-5.84	2	3	41	0	5	75
ZINC05248322	353.731	4.25	4.35	-7.99	0	4	39	0	4	73
ZINC20432021	431.479	5.78	10.72	-6	0	3	22	0	5	73
ZINC35287371	331.358	5.53	6.06	-52.83	3	4	55	1	5	73
ZINC47778114	330.408	1.44	6.05	-43.53	3	6	68	1	5	72

ZINC48077023	402.49	3.63	8.47	-52.63	3	6	64	1	8	70
ZINC58277988	416.449	4.11	8.4	-15.94	2	5	67	0	7	70
ZINC06596199	313.36	4.42	0.27	-14.25	0	4	43	0	3	66
ZINC11932808	352.48	1.25	-0.42	-49.05	1	6	54	1	4	63
ZINC18061181	308.337	2.42	4.04	-9.58	3	5	82	0	1	63
ZINC09390019	374.418	3.04	-2.41	-13.27	1	7	88	0	5	62
ZINC36042418	469.466	5.52	14.2	-8.97	0	4	68	0	4	61
ZINC71781936	260.243	2.56	6.36	-11.33	1	3	34	0	2	61
ZINC43544551	320.389	2.43	8.09	-19.47	1	6	68	0	6	59
ZINC02052707	363.335	4.01	3.06	-11.06	1	4	51	0	5	58
ZINC40106525	416.543	3.66	8.24	-15.13	1	6	76	0	7	58
ZINC08007395	326.418	2.31	-3.8	-14.15	1	6	75	0	5	53
ZINC05684707	252.277	0.21	-1.72	-15.43	3	5	94	0	1	52
ZINC48341845	344.401	4.14	6.51	-42.16	4	4	58	1	3	52
ZINC83050266	227.328	1.02	3.88	-47.92	3	4	57	1	2	52
ZINC13010330	308.37	3.17	7.27	-15.91	1	6	68	0	4	49
ZINC12357259	265.313	1.73	-2.87	-10.02	4	6	97	0	5	48
ZINC71781001	258.195	1.77	3.87	-11.16	1	3	47	0	2	48
ZINC13407933	473.366	3.92	3.55	-7.7	3	7	105	0	9	45
ZINC00519469	277.344	2.12	-3.23	-34.16	3	5	55	1	2	44
ZINC23182402	468.667	5.3	11.43	-12.45	2	4	31	0	8	44
ZINC95959351	310.438	3.06	7.9	-12.26	2	5	67	0	4	44
ZINC06738456	290.338	3.05	0.02	-20.6	1	4	50	0	3	43
ZINC04577875	303.053	3.54	2.47	-35.28	2	1	16	1	2	42
ZINC09283216	449.464	1.79	-4.01	-20.32	1	10	115	0	6	35
ZINC12668089	234.299	-0.1	5.14	-7.37	2	4	58	0	0	34
ZINC39083420	188.163	-0.15	0.03	-45.44	1	7	105	-1	2	34
ZINC01397478	434.829	4.44	0.14	-41.83	2	6	73	1	6	31
ZINC96008252	231.248	1.5	2.62	-9.25	0	4	47	0	2	31
ZINC88613616	246.287	0.36	-0.8	-28.71	4	7	97	1	4	23
ZINC19808440	436.527	3.69	13.03	-11.69	1	5	59	0	8	20
ZINC03240785	473.401	2.31	-8.21	-23.68	4	8	127	0	5	18
ZINC12412671	417.509	4.99	1.25	-18.41	1	6	72	0	5	10

APPENDIX G

LIST OF CHEMICALS DEDUCED FROM THE ALPHA-1 ANTITRYPSIN HIGH CONTENT SCREENING RESULT ANALYSIS TARGET-BASED HIT DIVERSIFICATION METHOD

The first column (from the left) represents the PubChem compound identifier (CID) of each compound, the second column reports the molecular weight of the compound, the third column reports the name of the chemical, finally the fourth column reports the number of targets in STITCH v4.

Chemical ID	Molecular Weight	Chemical Name	No of Targets
CID00216239	464.82495	sorafenib	62
CID00005002	383.5071	quetiapine	58
CID00005376	371.51456	AC1L1K7T	51
CID00060837	677.1848	irinotecan	21
CID00003143	807.87922	Docetaxel trihydrate	20
CID00091577	466.69514	AC1L3MCS	20
CID00060834	333.8755	duloxetine	18
CID00150311	409.425246	ezetimibe	16
CID09936728	527.61104	CHEMBL91636	14
CID00002689	334.33889	CGS 12066B	14
CID00001238	344.90144	octoclothepin	14
CID00017011	507.43949	Depixol	13
CID00004609	397.29176	oxaliplatin	13
CID00002781	343.89024	NSC293370	13
CID00068595	934.15842	maduramicin	12
CID00005268	379.4522	spiroxatrine	12
CID00001224	361.51974	AC1Q7BEJ	11
CID00004691	329.365403	AC-680	10
CID00060662	568.550603	mibefradil	10
CID09849669	523.66364	SR-973	10

CID00027991	1069.21696	DDAVP	10
CID05311065	1069.21696	desmopressin	9
CID00003404	318.33465	(Z) Fluvoxamine	9
CID00037459	361.51974	butaclamol	9
CID11653679	374.879463	CHEBI:590001	8
CID10318916	402.41948	CHEBI:250218	8
CID09802436	424.425	CHEMBL56837	8
CID10319235	407.50536	SureCN4172086	8
CID03069135	314.397123	Brn 4530212	8
CID10150649	391.46136	CHEBI:447271	8
CID00447475	328.38712	1o5a	8
CID00133038	316.369943	fluorocarazolol	8
CID11037377	429.46628	CHEBI:286771	8
CID00445843	362.83218	1o5e	8
CID11486446	401.901503	4-[(1R,5S)-3-(4-chlorophenyl)-3-hydroxy-8-azabicyclo[3.2.1]oct-8-yl]-1-(4-fluorophenyl)butan-1-one	7
CID03086153	340.46574	SureCN11076489	7
CID00005203	306.22958	Q740	7
CID00001746	303.14269	uPA inhibitor	6
CID09957375	494.1111	SureCN6399366	6
CID10025307	423.37558	SureCN6731065	6
CID00181743	339.38504	Thalictruberine	6
CID10237550	393.43418	CHEBI:447270	6
CID00060830	472.41628	tiotropium	6
CID11553459	416.3944	SureCN4479393	6
CID00127044	427.27665	CHEMBL2112942	6
CID09905731	334.4531	CHEBI:495666	6
CID02728531	382.3273	RH02255	6
CID11544156	313.432203	SureCN4850678	6
CID10335148	303.40084	CHEMBL2112912	6
CID11501540	368.49246	CHEBI:429856	6
CID09924938	358.3026	CHEMBL82093	6
CID11374008	438.45158	CHEMBL58577	6
CID10156375	376.4268	Sultam Hydroxamate 15a	6
CID00068770	363.49432	talinolol	6
CID00071240	408.432373	tefludazine	6
CID11257884	393.39266	CHEBI:400792	6
CID00068186	383.459143	spiramide	6
CID10174078	472.6648	SureCN5209016	6
CID11526445	518.5793	CHEMBL606904	6
CID00154058	398.92568	solifenacin	6
CID00060864	373.87316	AC1L1U2R	6
CID11508116	325.40152	CHEBI:434654	6
CID11590800	464.95413	SureCN4933186	6
CID00065257	375.77238	PMBs	6
CID10047100	409.349	SureCN6723167	6
CID03052780	361.8673	SureCN10982044	5
CID00003075	415.52578	NSC759576	5

CID05497171	453.461323	z-VAD-fmk	5
CID11202065	462.58386	CHEMBL2112985	5
CID00644185	453.461323	z-Val-Ala-Asp-fmk	5
CID00108220	385.23997	beta-CIT	5
CID00128564	379.45066	SC44463	5
CID00004323	379.45066	AC1L1HWN	5
CID00027287	624.0064	zinc protoporphyrin	5
CID00122190	413.29313	RTI-121	5
CID00003377	406.510726	AC1L1TJF	5
CID09842753	378.46414	CHEMBL606963	4
CID11690966	451.404263	SureCN1035715	4
CID10474144	367.456443	1-(4-fluorophenyl)-4-[(1R,5R)-3-hydroxy-3-phenyl-8-azabicyclo[3.2.1]oct-8-yl]butan-1-one	4
CID09852146	626.627863	Z-IETD-FMK	4
CID00005567	409.417133	trifluperidol	4
CID00122197	431.283593	FP-CIT	4
CID10074155	486.67146	SureCN9006340	4
CID09820163	358.43628	SureCN7554088	4
CID00128054	315.4729	N 0734	4
CID11441732	427.41098	L023686	4
CID10788465	424.526006	CHEBI:299920	4
CID00002384	366.37722	AC1L1DK5	4
CID10409701	417.257013	MCL-301	4
CID11603959	411.49406	CHEBI:435901	4
CID11559589	362.48788	CHEBI:409256	4
CID11532153	434.529023	CHEBI:593352	4
CID05289508	661.8603	DB03005	4
CID11620908	547.95661	3:00 PM	4
CID00208917	375.438683	Elopiprazole	4
CID09888555	423.57096	SureCN5507438	4
CID09797476	328.23356	SureCN3423153	4
CID11690910	448.52105	SureCN1034132	4
CID09796407	302.12359	Org 12962	4
CID11505592	644.659626	CHEBI:440657	4
CID10598649	428.60894	CHEBI:205753	4
CID11699469	537.69176	CHEBI:449977	4
CID11676381	432.490003	CHEBI:593224	4
CID11554489	464.95413	SureCN5180132	4
CID00128918	417.89433	SR 57746A	4
CID00005516	405.95962	AC1L1KIN	4
CID00066004	302.41454	alniditan	4
CID02737388	311.2494	1-(2-diphenyl)piperazine	4
CID10131344	325.40152	CHEBI:434692	4
CID10447533	300.36243	SureCN5996022	4
CID09910222	424.425	CHEMBL61193	4
CID11134191	453.587066	CHEBI:327101	4
CID03929516	343.410226	difluorobenztropine	4
CID11539598	446.516583	CHEBI:593418	4

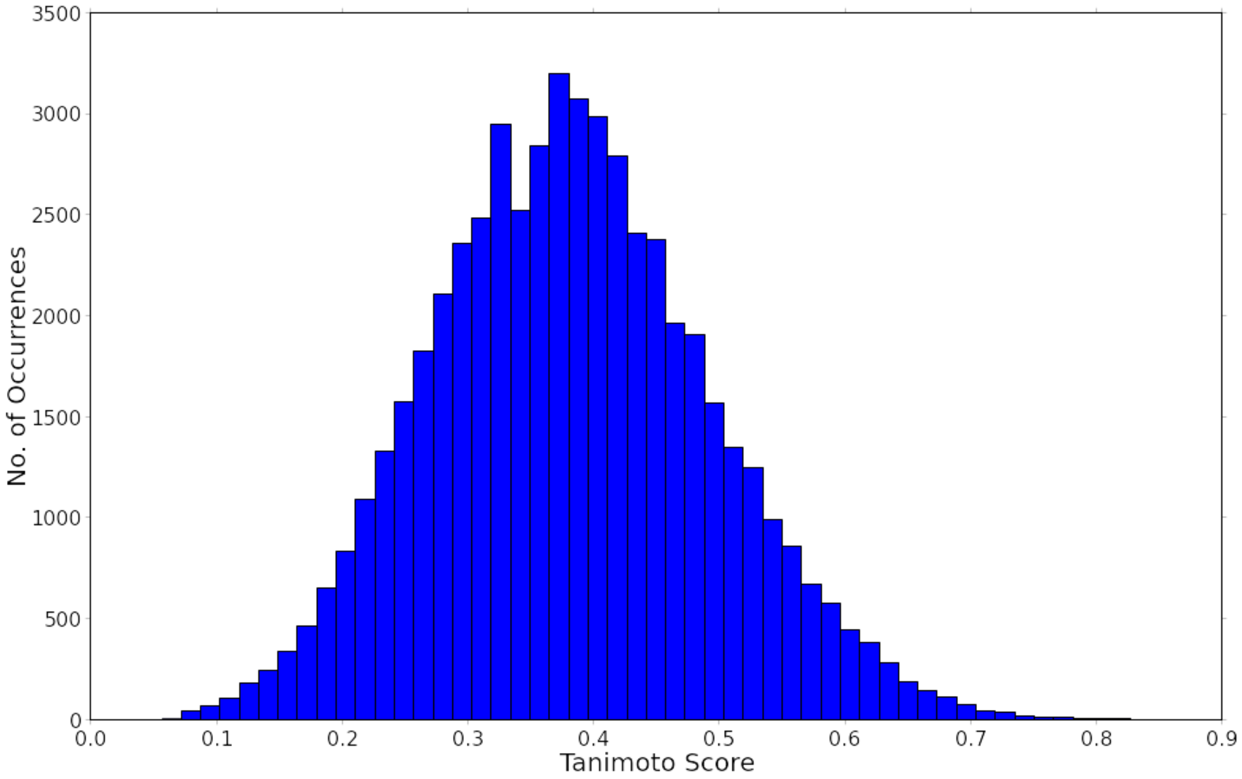
CID11233292	346.77513	SureCN3562639	4
CID05289507	661.8603	DB02226	4
CID09839392	302.41278	SureCN7967298	4
CID09821217	380.40403	CP-122721	4
CID10024324	406.54044	CHEBI:187762	4
CID10163178	485.938403	SureCN231072	4
CID10024183	404.46826	SureCN6930132	4
CID10472143	335.4427	PDSP2_001209	4
CID10836499	432.572823	CHEMBL67024	4
CID10184653	485.938403	afatinib	4
CID11058664	410.31563	CHEBI:128185	4
CID10054373	603.79136	SureCN5650667	4
CID11006894	673.704146	CHEBI:287332	4
CID00002386	334.37842	bis(5-amidino-2-benzimidazolyl)methane	4
CID11668034	380.478343	CHEBI:433105	4
CID11695960	348.43816	CHEBI:430708	4
CID09600423	525.60154	t - 87	3
CID09998835	364.27678	methyl (3S)-3-[4-[(Z)-2-bromovinyl]phenyl]-8-methyl-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID00131993	371.21339	N-Nor-cit	3
CID09952054	385.23997	CTK8G8335	3
CID09800811	395.296358	CHEMBL1214004	3
CID11278435	313.794843	CHEMBL1812750	3
CID10545894	351.457043	1-(4-fluorophenyl)-4-[(1R,5S)-3-phenyl-8-azabicyclo[3.2.1]oct-8-yl]butan-1-one	3
CID11441438	416.53372	Sultam Hydroxamate 23c	3
CID00148193	489.39578	NSC702818	3
CID09947229	364.6945	CHEMBL87031	3
CID10112621	416.53372	(3S)-2-[4-(4-tert-butylphenyl)benzyl]-1,1-diketo-thiazinane-3-carbohydroxamic acid	3
CID09846169	442.296828	Methyl (2S,3S)-8-[(E)-4-fluorobut-2-enyl]-3-(4-iodophenyl)-8-azabicyclo[3.2.1]octane-2-carboxylate	3
CID00159324	489.39578	tipifarnib	3
CID11500578	311.37494	methyl (1R,3S,4S,5S)-3-[4-(2-furyl)phenyl]-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID10410301	428.270248	2-fluoranylethyl 3-[4-[(Z)-2-iodanylethenyl]phenyl]-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID05281881	434.51761	flupenthixol	3
CID10765852	447.73819	methyl (1S,3S,4S,5R)-8-(3-chloropropyl)-3-(4-iodophenyl)-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID11233397	350.2502	methyl (1R,3R,4R,5S)-3-[4-[(Z)-2-bromovinyl]phenyl]-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID11371257	341.46712	CHEBI:400816	3

CID00115430	323.9006	AC1Q3F7T	3
CID09800928	397.25067	methyl 3-[4-[(Z)-2-iodovinyl]phenyl]-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID10574379	425.30383	CHEMBL1945246	3
CID09884800	350.845358	CHEMBL1214003	3
CID11282852	472.40493	3-(8,8-dimethyl-8-azoniabicyclo[3.2.1]oct-3-yl)-2,2-diphenyl-propanenitrile	3
CID10404382	330.425077	methyl 8-[(E)-4-fluorobut-2-enyl]-3-(p-tolyl)-8-azabicyclo[3.2.1]octane-4-carboxylate	3
CID10363398	397.25067	CHEBI:114309	3
CID03366356	525.60154	AC1MOB6C	3

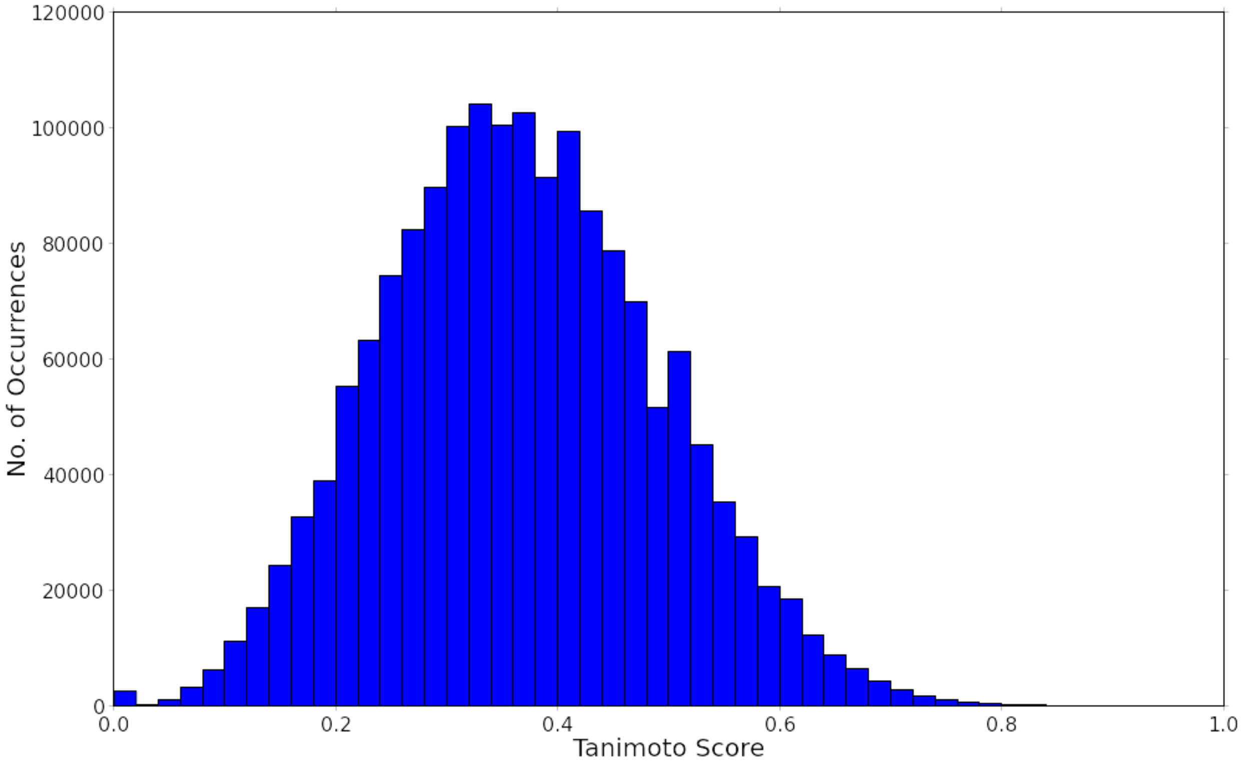
APPENDIX H

COMPARISON OF RESULTS FROM CHEMICAL AND TARGET BASED DIVERSIFICATION OF PRESTWICK LIBRARY SCREEN HITS

The similarity between the chemical based diversification results shown in Appendix F and target based diversification results shown in Appendix G:



The similarity between two groups of 1278 chemicals randomly selected from the ZINC purchase-ready compounds library (which has 12.8 million chemicals):



The similarity distributions are highly similar with a Kullback-Leibler divergence of 0.031 which means that the sets of chemical based and target based diversification results are as similar as would be expected by chance alone. This means that our diversification strategies do actually diversify different sets of chemicals as originally intended.

APPENDIX I

RESULTS FROM COMPUTATIONAL DIVERSIFICATION ANALYSIS OF DRUGS WITH KNOWN NEUROPROTECTIVE ACTIVITY

The following table shows the results of the computational analysis used for neuroprotective diversification and mechanism of action identification. The drugs that were selected for experimental follow-up are shown in yellow, and the drugs that successfully worked as neuroprotectives in these experiments (sodium nitroprusside and thyroxine) by showing statistically significant neuroprotection are shown in green.

Adaptive Compound Selection by Maximal Distance to Support and Previously Selected Compounds						
Hypothesis 1: PRL_HUMAN (Prolactin, organism:9606)						
Support:		Cysteamine	Melatonin	Ritanserin	Progesterone	
Compounds to test hypothesis:						
Drug	Target Count	LV distance to support:				Average:
Fk 33-824 (CID000047470)	3	5.705	5.107	6.317	5.925	5.763
Estradiol Benzoate (CID100003262)	6	6.058	5.406	5.360	6.112	5.734
Thyroxine (CID100000853)	95	5.147	4.891	5.680	4.816	5.134
Azinphos-Methyl (CID000002268)	4	5.066	4.512	6.028	5.470	5.269
Clomipramine (CID100002801)	22	5.069	4.628	5.570	4.844	5.028
Metergoline (CID100004090)	26	5.447	4.619	5.729	5.385	5.295
M-Chlorophenylpiperazine (CID100001355)	18	5.236	4.875	5.623	5.256	5.247
Nalmefene (CID100004422)	7	5.061	4.459	5.820	4.838	5.044
Spiperone (CID100005265)	40	5.035	4.565	5.104	4.698	4.850
Domperidone (CID100003151)	9	5.163	4.462	5.686	4.978	5.072
Chlorpromazine (CID000002726)	48	4.863	4.475	5.335	4.418	4.773
Aripiprazole (CID000060795)	27	4.906	4.260	4.515	4.648	4.582
Ergot (CID100003250)	2	5.060	4.527	4.533	5.071	4.798
Nomifensine (CID100004528)	9	5.010	4.618	5.245	4.962	4.959

8-Br-Camp (CID000032014)	21	4.698	3.721	5.281	4.708	4.602
Ici 182,780 (CID100104741)	18	4.912	4.175	5.293	4.858	4.810
Quinpirole (CID100001257)	22	4.947	4.583	5.393	4.737	4.915
Metyrapone (CID100004174)	15	5.100	4.255	5.687	4.974	5.004
Perphenazine (CID000004748)	19	4.756	3.714	5.066	4.720	4.564
Ketanserin (CID000003822)	30	4.610	4.299	4.691	4.403	4.501
8-Br-Camp (CID100001912)	14	4.995	3.906	5.109	4.872	4.721
Bromocriptine (CID000031100)	47	4.555	3.926	5.025	4.891	4.599
Clomiphene Citrate (CID100002800)	8	4.744	4.289	4.773	4.818	4.656
Ritanserin (CID100005074)	24	4.854	4.023	4.925	4.644	4.612
Ergovaline (CID000104843)	1	4.770	4.135	5.162	4.288	4.589
Hypothesis 2: CALM_HUMAN (Calmodulin, organism:9606)						
Support:		Bepridil	Melatonin	Mephenytoin		
Compounds to test hypothesis:						
Drug	Target Count	LV distance to support:			Average:	
Aprindine (CID100002218)	1	5.438	5.129	5.438	5.335	
4-Chloroaniline (CID000007812)	1	4.826	4.530	4.826	4.727	
Compound 48/80 (CID000104735)	7	4.789	4.673	4.789	4.750	
Promethazine (CID000004927)	5	4.911	4.623	4.911	4.815	
Trifluoperazine (CID100005566)	32	4.433	4.420	4.433	4.429	
Cgs 9343B (CID100065909)	2	4.571	4.229	4.571	4.457	
Nifedipine (CID100004485)	35	4.916	4.465	4.916	4.765	
Phenothiazine (CID100007108)	3	4.734	4.227	4.734	4.565	
Ww7 (CID000005681)	6	4.395	4.466	4.395	4.419	
Verapamil (CID000002520)	40	4.240	3.963	4.240	4.148	
Trifluoperazine (CID000005566)	33	4.150	3.918	4.150	4.073	
Diltiazem (CID100003075)	10	4.435	4.139	4.435	4.336	
Nicardipine (CID100004473)	8	4.677	4.223	4.677	4.526	
Genistein (CID005280961)	97	4.155	4.033	4.155	4.114	
B8509-035 (CID024847739)	1	4.185	3.825	4.185	4.065	
Ww7 (CID100005681)	5	4.339	3.911	4.339	4.197	
Dibucaine (CID100003025)	3	4.540	4.210	4.540	4.430	
Pimozide (CID100016362)	26	4.233	3.804	4.233	4.090	
Loperamide (CID100003954)	82	4.257	3.977	4.257	4.163	
Compound 48/80 (CID100104735)	7	4.029	4.073	4.029	4.043	
Bepridil (CID100002351)	5	3.987	3.795	3.987	3.923	
Fluphenazine (CID100003372)	9	3.989	3.709	3.989	3.895	
Kar-2 (CID100157684)	2	3.849	3.876	3.849	3.858	
Phenothiazine (CID000007108)	3	4.141	3.517	4.141	3.933	
Verapamil (CID100002520)	40	4.112	3.774	4.112	4.000	
Hypothesis 3: CASP3_HUMAN (Caspase-3 subunit p12, organism:9606)						
Support:		Melatonin	Minocycline			
Compounds to test hypothesis:						
Drug	Target Count	LV distance to support:		Average:		
Sodium Nitroprusside (CID000045469)	14	5.237	5.148	5.193		
Imatinib (CID100005291)	48	5.290	5.061	5.175		
Staurosporine (CID000044259)	85	4.770	5.399	5.085		

Rxb (CID111632008)	1	4.743	4.842	4.793		
Tpck (CID000439647)	6	4.405	4.183	4.294		
Zoledronic Acid (CID100068740)	83	4.350	4.552	4.451		
P-Bromoanisole (CID000007730)	1	4.869	4.550	4.709		
Kainate (CID000010255)	69	4.318	4.652	4.485		
Nordihydroguaiaretic Acid (CID100004534)	21	4.703	4.503	4.603		
Inhibitor 65B (CID005327315)	1	4.725	4.043	4.384		
Ptf (CID100013016)	1	4.494	4.654	4.574		
Peroxynitrite (CID100104806)	21	4.461	4.515	4.488		
Gemcitabine (CID000060749)	28	3.829	4.191	4.010		
15-Deoxy-Delta12,14-Prostaglandin J2 (CID100001444)	20	4.433	4.331	4.382		
Thapsigargin (CID000446378)	84	3.754	4.319	4.036		
Chebi:400985 (CID009851134)	1	4.025	4.017	4.021		
Pyrrolidine Isatin Analogue 11F (CID111712912)	1	4.469	3.901	4.185		
Inhibitor 64B (CID005327307)	1	4.497	4.524	4.511		
Db08213 (CID100001389)	1	4.093	4.043	4.068		
3-Morpholinosydnonimine (CID100005219)	4	4.450	3.664	4.057		
Pzn (CID005289238)	1	3.724	3.933	3.828		
Ac-Devd-Cho (CID100004330)	5	3.880	4.150	4.015		
Salidroside (CID100159278)	2	4.093	3.316	3.704		
Chebi:461307 (CID111700402)	1	4.030	4.108	4.069		
Gsno (CID100003514)	5	3.782	4.016	3.899		
Hypothesis 3: PA21B_HUMAN (Phospholipase A2, organism:9606)						
Support:		Cysteamine	Calcimycin			
Compounds to test hypothesis:						
Drug	Target Count	LV distance to support:		Average:		
Fpl 55712 (CID000105007)	3	5.731	5.973	5.852		
1-Acyl-Sn-Glycero-3-Phosphocholines (CID124798684)	68	5.454	5.299	5.377		
A23187 (CID100001959)	33	5.600	5.478	5.539		
Manoalide (CID006437368)	4	5.199	4.847	5.023		
Calphostin C (CID100002533)	9	5.052	5.613	5.332		
Compound 48/80 (CID000104735)	7	5.003	5.471	5.237		
Ochnaflavone (CID105492110)	3	5.458	5.308	5.383		
Aristolochic Acid (CID000002236)	3	4.586	5.574	5.080		
Nordihydroguaiaretic Acid (CID100004534)	21	5.173	5.228	5.200		
Ochnaflavone (CID005492110)	3	4.995	4.746	4.871		
Verapamil (CID000002520)	40	4.359	4.685	4.522		
Phosphatidic Acid (CID100447791)	18	4.605	4.992	4.798		
Chloroquine (CID000002719)	8	4.605	4.999	4.802		
P-Bromophenacyl Bromide (CID000007454)	3	4.469	4.895	4.682		
Platelet-Activating Factor (CID100461545)	10	4.535	4.433	4.484		
Heparin (CID000008784)	77	4.709	4.845	4.777		
Compound 48/80 (CID100104735)	7	4.817	4.759	4.788		
Diacylglycerol (CID006026790)	2	4.578	4.790	4.684		
Verapamil (CID100002520)	40	4.371	5.014	4.693		
P-Bromophenacyl Bromide (CID100007454)	3	4.498	4.451	4.474		
5,8,11,14-Eicosatetraynoic Acid (CID100001780)	3	4.158	4.039	4.098		

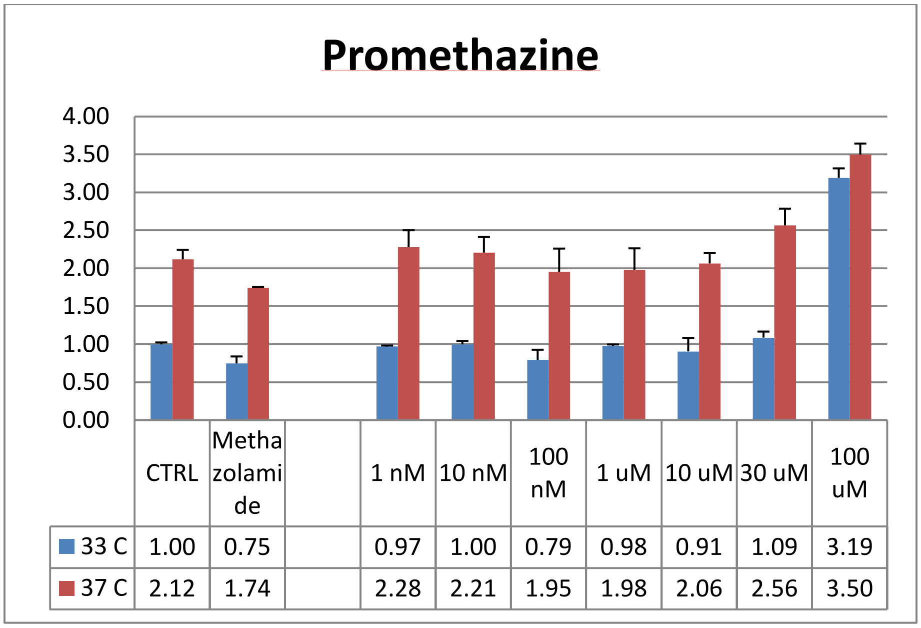
Phosphatidylglycerol (CID045109789)	4	4.559	4.717	4.638		
Aristolochic Acid (CID100002236)	3	4.172	4.750	4.461		
Calphostin C (CID000002533)	14	4.328	4.643	4.486		
Phenidone (CID000007090)	3	4.446	4.610	4.528		
Hypothesis 4: CAH2_HUMAN (Carbonic anhydrase 2, organism:9606)						
Support:		Methazolamide				
Compounds to test hypothesis:						
Drug	Target Count	LV distance to support:	Average:			
Chebi:178579 (CID044296104)	1	6.250	6.250			
J71 (CID046916276)	1	5.633	5.633			
Chembl35532 (CID010915515)	2	5.829	5.829			
Imatinib (CID100005291)	48	5.308	5.308			
Chebi:333101 (CID010843175)	2	5.825	5.825			
Chebi:333241 (CID105067385)	1	5.408	5.408			
N-(3-Chloro-7-Indolyl)-1,4-Benzenedisulfonamide (CID000216468)	12	5.473	5.473			
Chebi:720036 (CID046197893)	2	5.477	5.477			
2H-Thieno[3,2-E]-1,2-Thiazine-6-Sulfonamide 1,1-Dioxide 18 (CID019434092)	1	5.085	5.085			
3,5-Dichlorosulfanilamide (CID100089607)	2	4.813	4.813			
Chebi:415002 (CID144397294)	2	5.386	5.386			
Chebi:332796 (CID010832697)	1	5.736	5.736			
2H-Thieno[3,2-E]-1,2-Thiazine-6-Sulfonamide 1,1-Dioxide 4 (CID019434096)	1	4.909	4.909			
Chebi:385160 (CID010625038)	3	5.487	5.487			
Th0 (CID112563346)	1	5.433	5.433			
Chebi:149899 (CID104094683)	1	4.885	4.885			
Chebi:301123 (CID110519868)	3	5.683	5.683			
Zinc00097317 (CID000708535)	3	4.496	4.496			
Chembl97425 (CID011269105)	1	5.117	5.117			
Chebi:223584 (CID010430595)	4	4.934	4.934			
Chebi:427355 (CID111696964)	4	5.100	5.100			
Nsc402851 (CID000345312)	4	4.703	4.703			
Subsporin C (CID100151723)	5	5.182	5.182			
Chebi:332454 (CID010833817)	2	5.484	5.484			
Hydroxamate 21 (CID006916013)	3	5.401	5.401			
Hypothesis 4: CAH7_HUMAN (Carbonic anhydrase 7, organism:9606)						
Support:		Methazolamide				
Compounds to test hypothesis:						
Drug	Target Count	LV distance to support:	Average:			
Indanesulfonamide Derivative 11C (CID011640067)	4	4.961	4.961			
Indanesulfonamide Derivative 12C (CID011718391)	6	4.649	4.649			
6-Hydrogen-2-Benzothiazolesulfonamide (CID100067944)	3	4.404	4.404			
Mafenide Acetate (CID000003998)	5	4.562	4.562			
Metolazone (CID000004170)	3	4.160	4.160			
Methazolamide (CID100001798)	87	4.481	4.481			
Benzolamide (CID000018794)	14	4.551	4.551			
4-Carboxybenzenesulfonamide (CID000008739)	6	4.116	4.116			

Topiramate (CID100005514)	18	4.306	4.306			
Chlorthalidone (CID000002732)	6	4.402	4.402			
2-Ethylamido-5-Sulfonamidoindane (CID011543564)	4	4.700	4.700			
Indanesulfonamide Derivative 6 (CID011414131)	4	4.347	4.347			
2-Aminoindane-5-Sulfonic Acid (CID044395769)	3	4.009	4.009			
Chebi:595853 (CID042609905)	5	4.250	4.250			
Chlorthalidone (CID100002732)	6	4.409	4.409			
Bumetanide (CID000002471)	13	4.139	4.139			
3Cc (CID111537386)	3	4.430	4.430			
5-Amino-1,3,4-Thiadiazole-2-Sulfonamide (CID100084724)	7	4.210	4.210			
Molport-002-472-850 (CID005172475)	7	4.344	4.344			
667-Coumate (CID105287541)	6	3.904	3.904			
Chebi:494255 (CID117748220)	2	4.288	4.288			
2-Nonylamido-5-Sulfonamidoindane (CID011660633)	4	4.090	4.090			
Dorzolamide (CID100003154)	83	3.800	3.800			
Dichlorphenamide (CID100003038)	39	4.227	4.227			
2-Ethylamido-5-Sulfonamidoindane (CID111543564)	4	3.654	3.654			

APPENDIX J

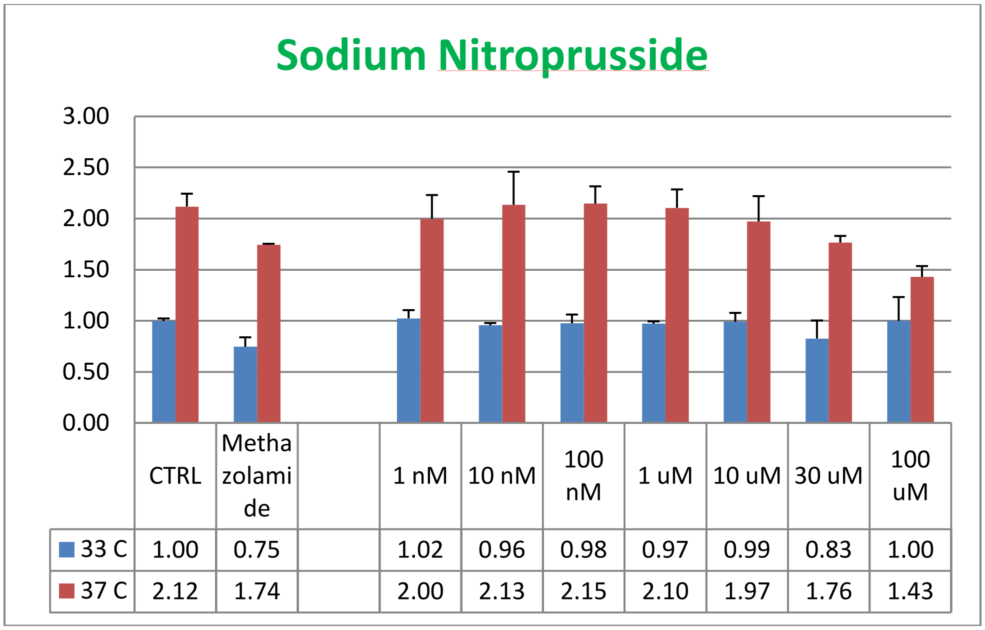
EXPERIMENTAL TESTING OF NEUROPROTECTIVE ACTIVITY OF DRUGS WITH PREDICTED NEUROPROTECTIVE ACTIVITY

The following table shows the results of experimental validation of the compounds with predicted neuroprotective activity. The experiments were performed by Dr. Hossein Mousavi in the Friedlander lab. The results are from an LDH screen therefore higher values indicate more cell death and *vice versa* for lower values. The arbitrary units of fluorescence are normalized to control cells at 33C (i.e. normal culture) conditions.



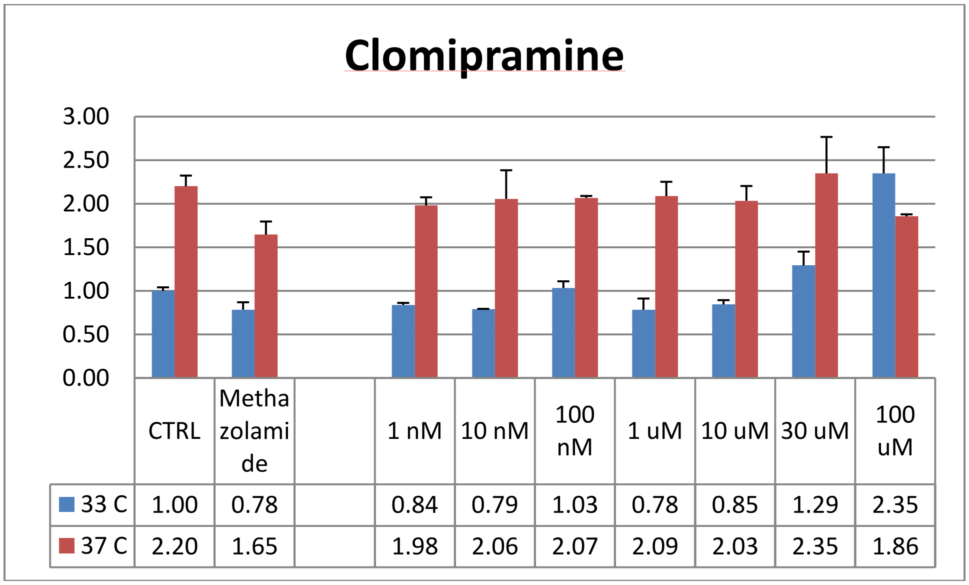
Promethazine

	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.75		0.97	1.00	0.79	0.98	0.91	1.09	3.19
🟥 37 C	2.12	1.74		2.28	2.21	1.95	1.98	2.06	2.56	3.50



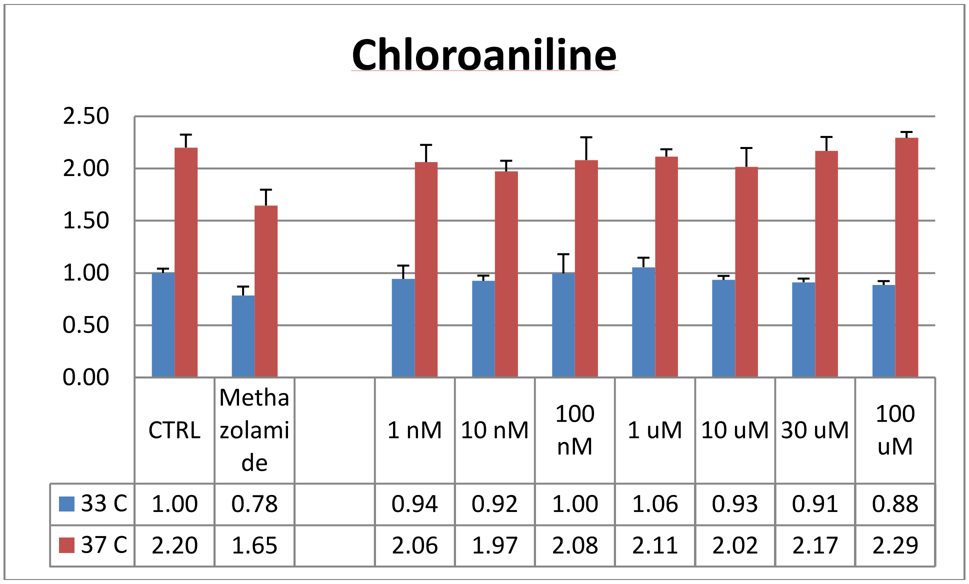
Sodium Nitroprusside

	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.75		1.02	0.96	0.98	0.97	0.99	0.83	1.00
🟥 37 C	2.12	1.74		2.00	2.13	2.15	2.10	1.97	1.76	1.43

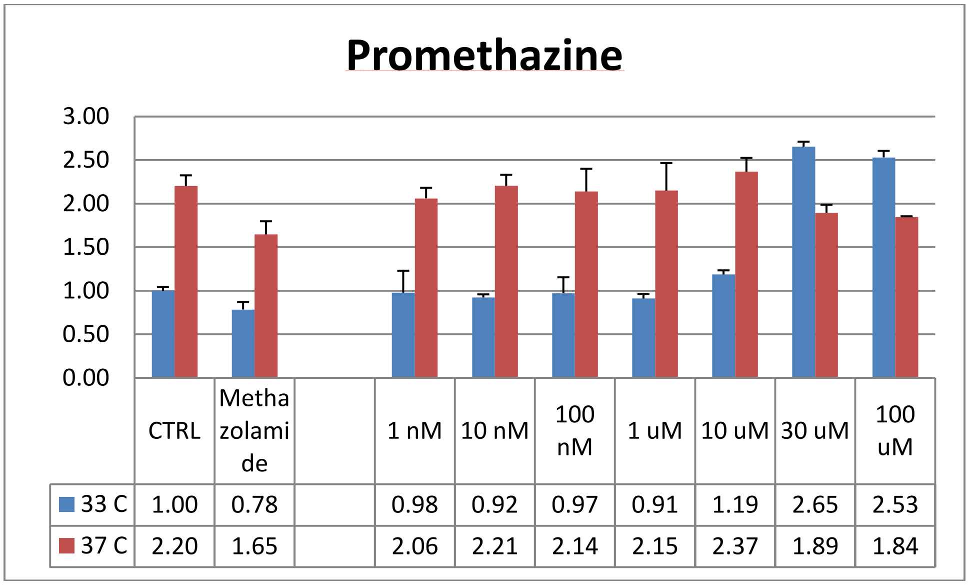


Clomipramine

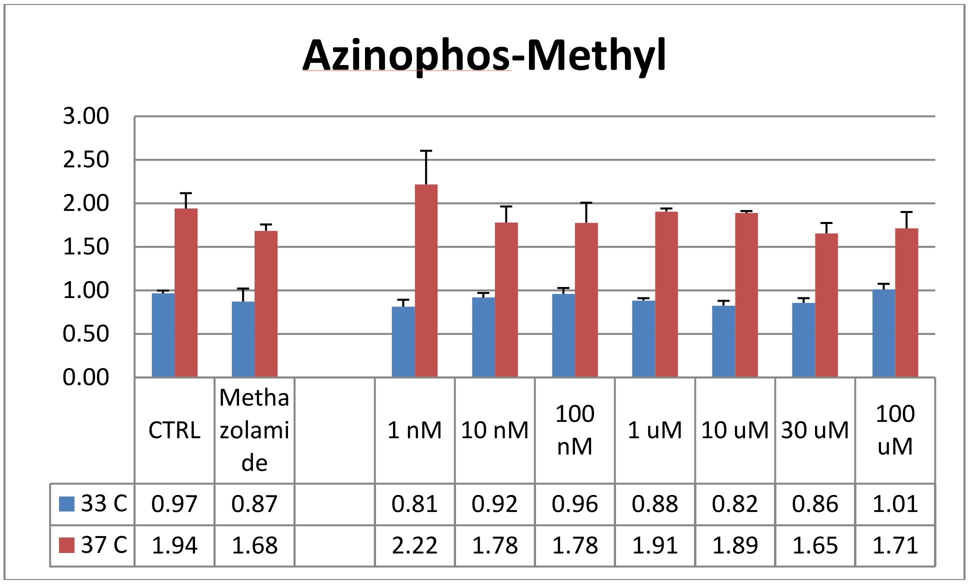
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.78		0.84	0.79	1.03	0.78	0.85	1.29	2.35
🟥 37 C	2.20	1.65		1.98	2.06	2.07	2.09	2.03	2.35	1.86



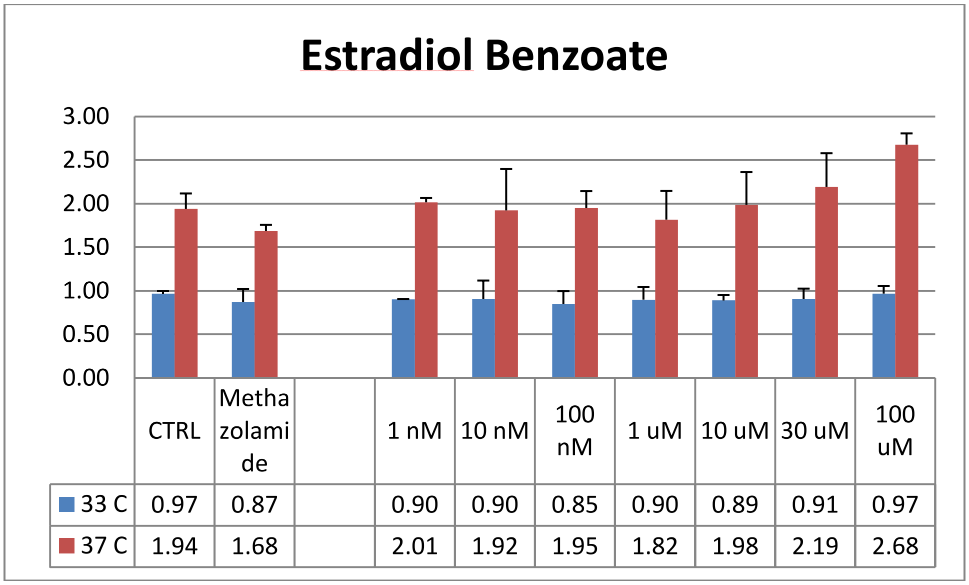
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.78		0.94	0.92	1.00	1.06	0.93	0.91	0.88
🟥 37 C	2.20	1.65		2.06	1.97	2.08	2.11	2.02	2.17	2.29



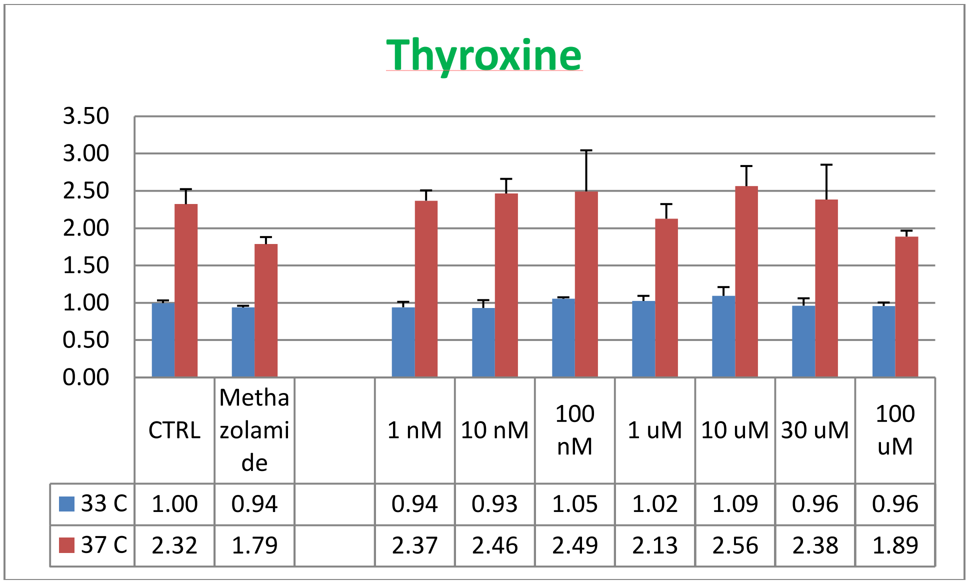
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.78		0.98	0.92	0.97	0.91	1.19	2.65	2.53
🟥 37 C	2.20	1.65		2.06	2.21	2.14	2.15	2.37	1.89	1.84



	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	0.97	0.87		0.81	0.92	0.96	0.88	0.82	0.86	1.01
🟥 37 C	1.94	1.68		2.22	1.78	1.78	1.91	1.89	1.65	1.71

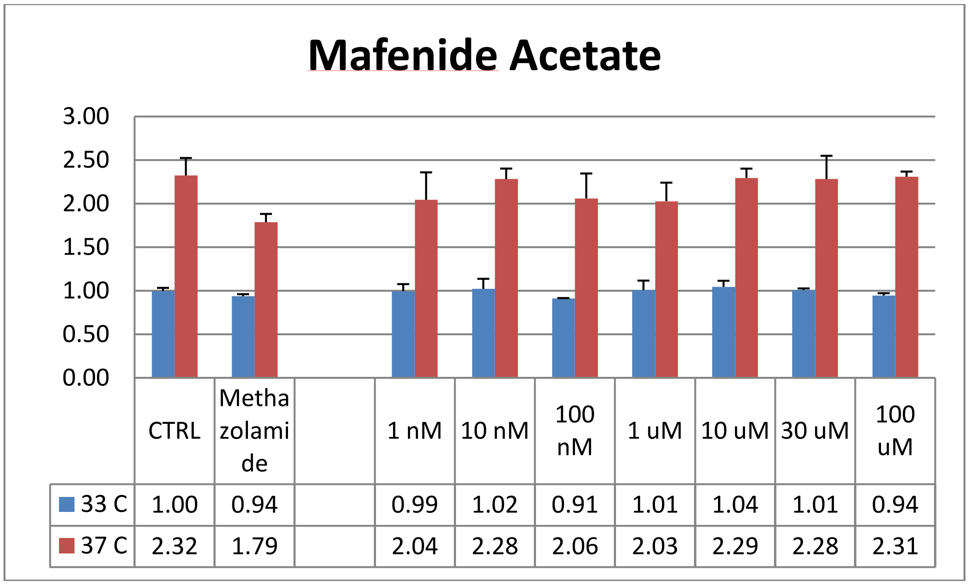


	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	0.97	0.87		0.90	0.90	0.85	0.90	0.89	0.91	0.97
🟥 37 C	1.94	1.68		2.01	1.92	1.95	1.82	1.98	2.19	2.68



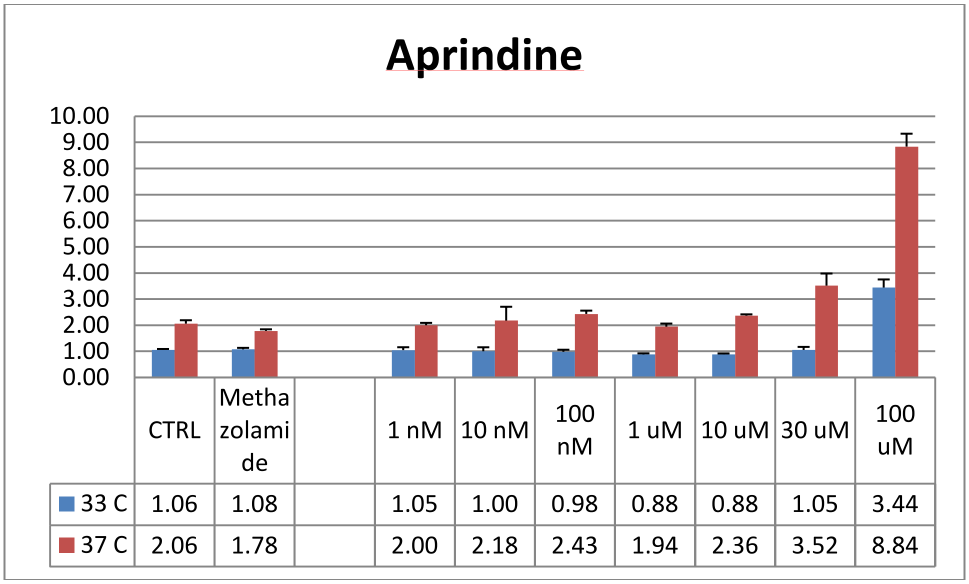
Thyroxine

	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.94		0.94	0.93	1.05	1.02	1.09	0.96	0.96
🟥 37 C	2.32	1.79		2.37	2.46	2.49	2.13	2.56	2.38	1.89

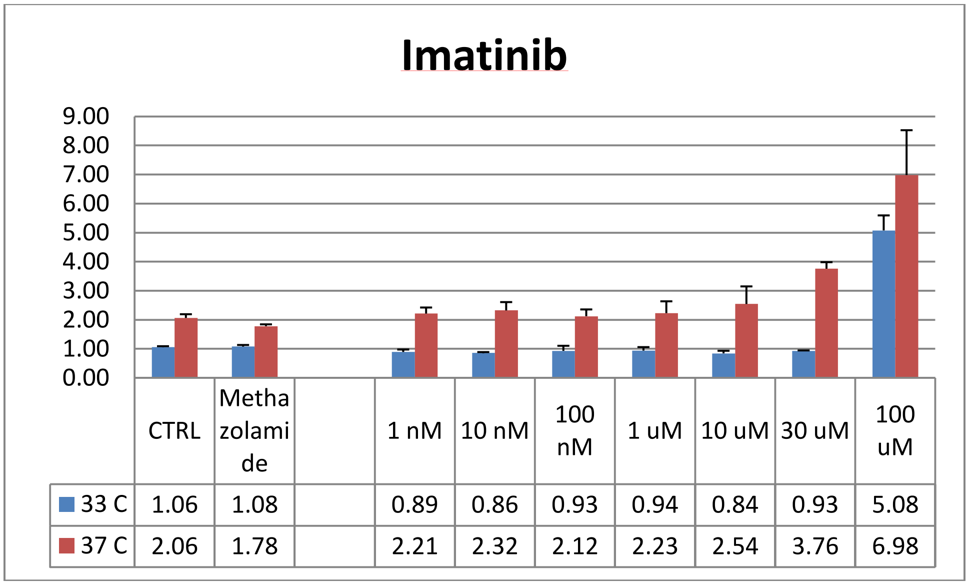


Mafenide Acetate

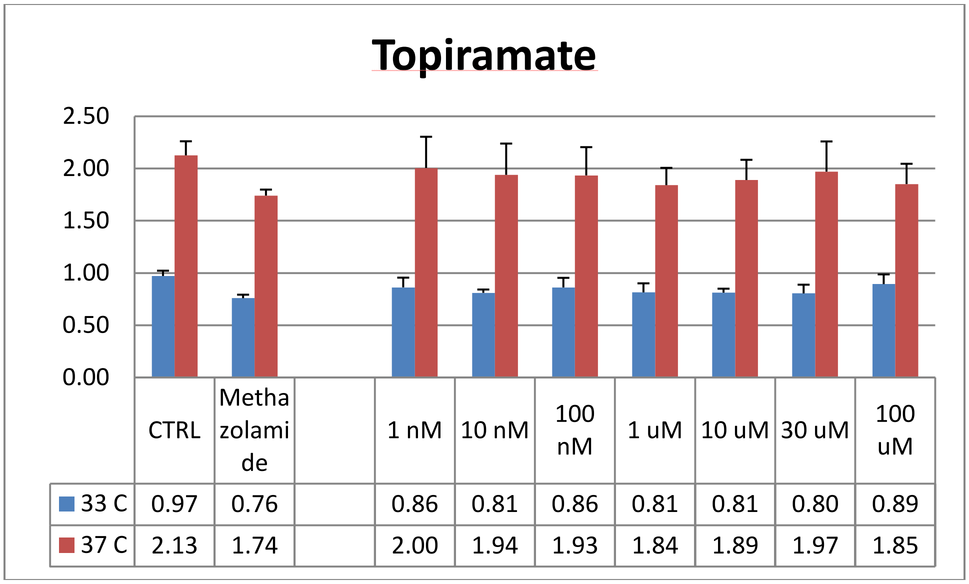
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.94		0.99	1.02	0.91	1.01	1.04	1.01	0.94
🟥 37 C	2.32	1.79		2.04	2.28	2.06	2.03	2.29	2.28	2.31



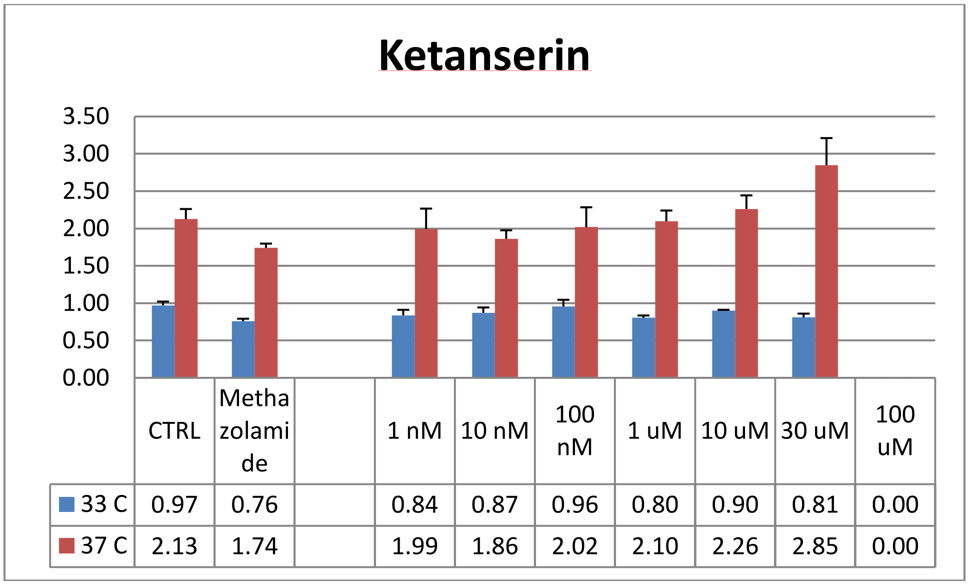
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.06	1.08		1.05	1.00	0.98	0.88	0.88	1.05	3.44
🟥 37 C	2.06	1.78		2.00	2.18	2.43	1.94	2.36	3.52	8.84



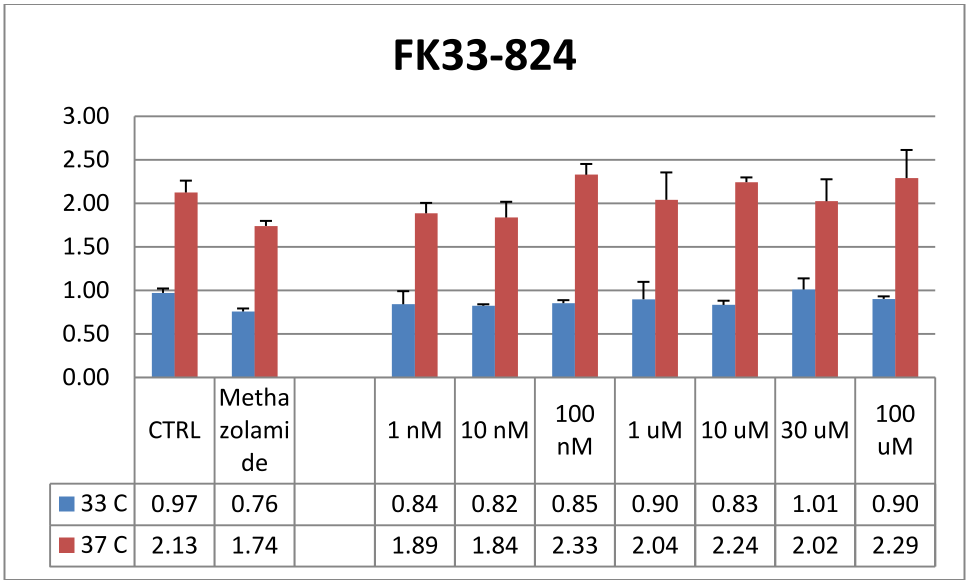
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.06	1.08		0.89	0.86	0.93	0.94	0.84	0.93	5.08
🟥 37 C	2.06	1.78		2.21	2.32	2.12	2.23	2.54	3.76	6.98



	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	0.97	0.76		0.86	0.81	0.86	0.81	0.81	0.80	0.89
🟥 37 C	2.13	1.74		2.00	1.94	1.93	1.84	1.89	1.97	1.85

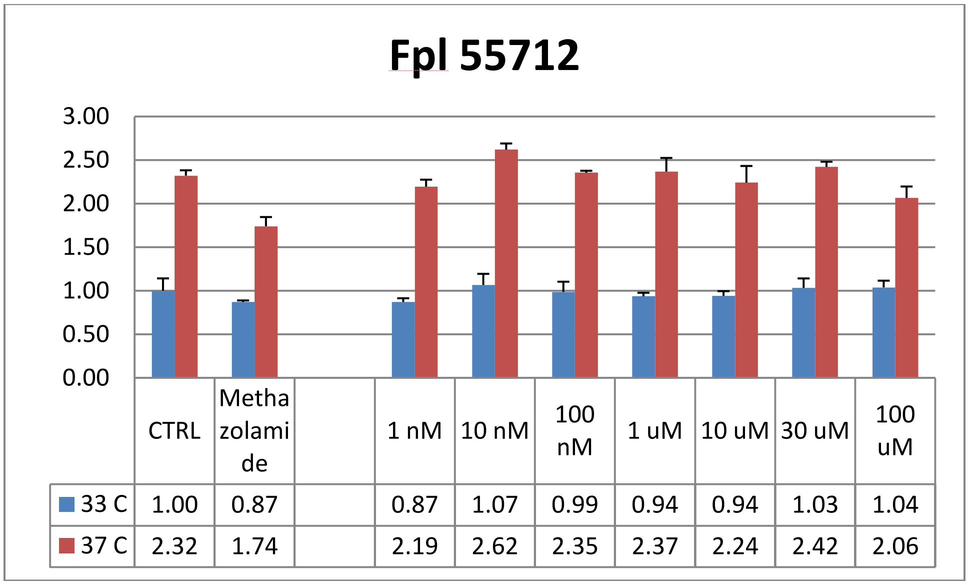


	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	0.97	0.76		0.84	0.87	0.96	0.80	0.90	0.81	0.00
🟥 37 C	2.13	1.74		1.99	1.86	2.02	2.10	2.26	2.85	0.00



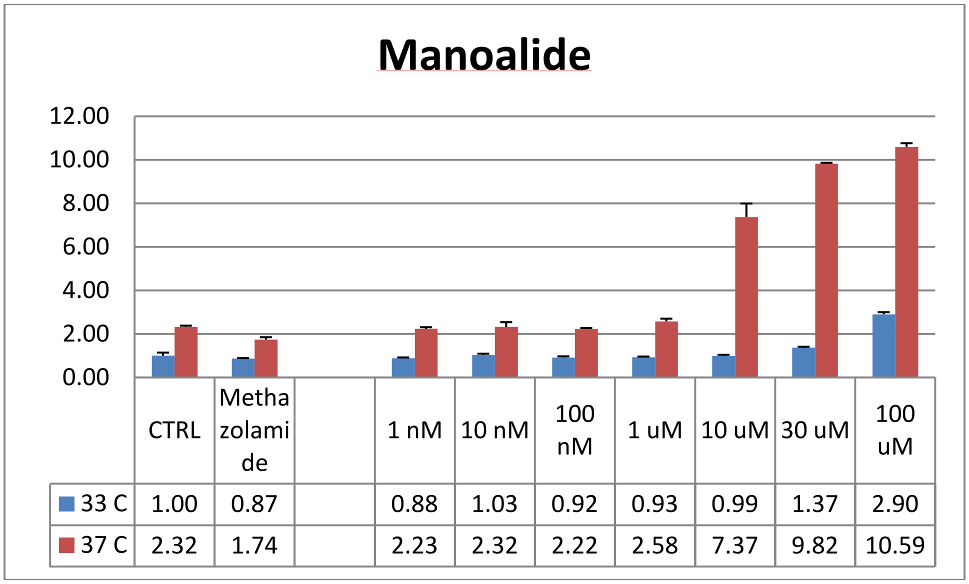
FK33-824

	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	0.97	0.76		0.84	0.82	0.85	0.90	0.83	1.01	0.90
🟥 37 C	2.13	1.74		1.89	1.84	2.33	2.04	2.24	2.02	2.29

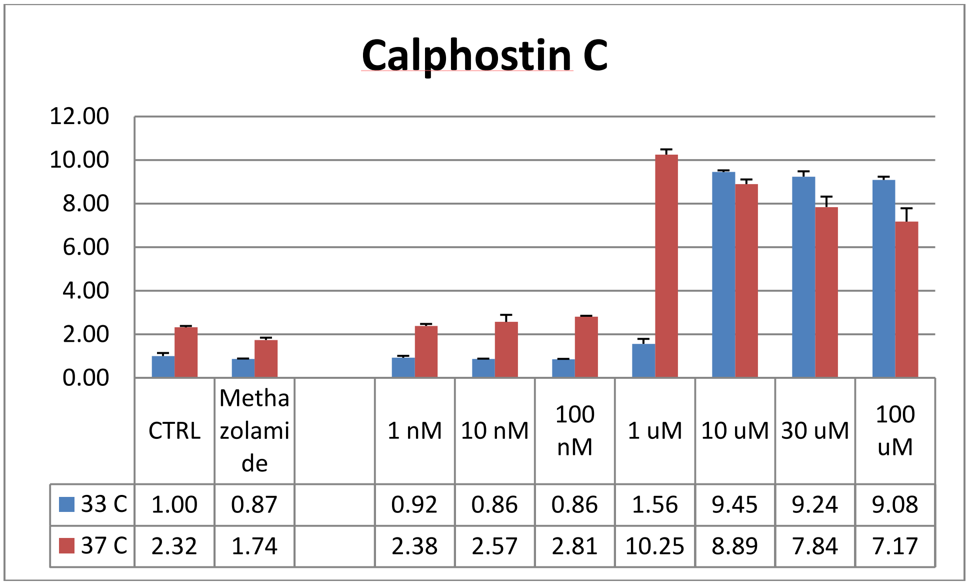


Fpl 55712

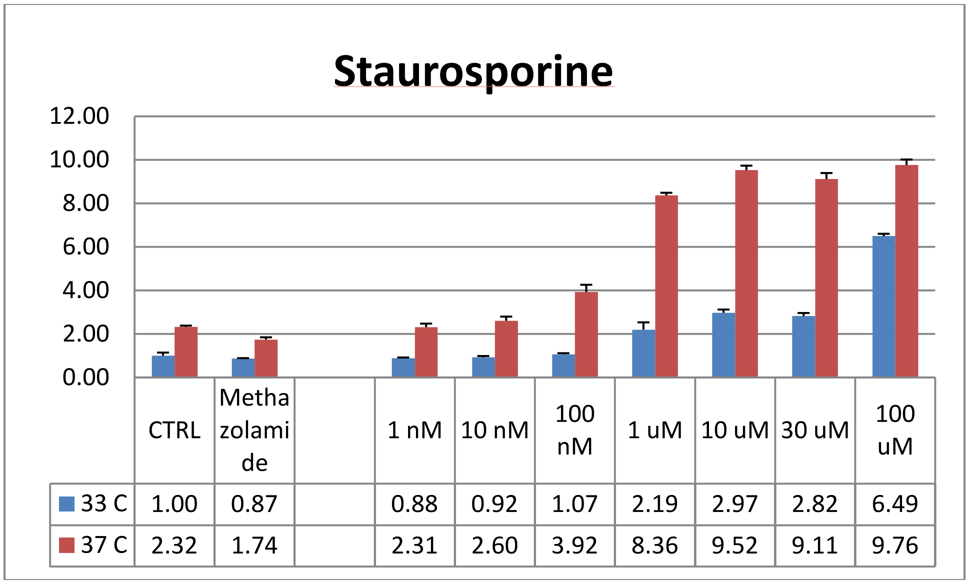
	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.87		0.87	1.07	0.99	0.94	0.94	1.03	1.04
🟥 37 C	2.32	1.74		2.19	2.62	2.35	2.37	2.24	2.42	2.06



	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.87		0.88	1.03	0.92	0.93	0.99	1.37	2.90
🟥 37 C	2.32	1.74		2.23	2.32	2.22	2.58	7.37	9.82	10.59



	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.87		0.92	0.86	0.86	1.56	9.45	9.24	9.08
🟥 37 C	2.32	1.74		2.38	2.57	2.81	10.25	8.89	7.84	7.17



	CTRL	Methazolamide		1 nM	10 nM	100 nM	1 uM	10 uM	30 uM	100 uM
🟦 33 C	1.00	0.87		0.88	0.92	1.07	2.19	2.97	2.82	6.49
🟥 37 C	2.32	1.74		2.31	2.60	3.92	8.36	9.52	9.11	9.76

BIBLIOGRAPHY

[1] K. I. Kaitin and J. DiMasi, "Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000-2009," *Clinical pharmacology and therapeutics*, vol. 89, no. 2, pp. 183-188, Mar.2011.

[2] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nat Rev Drug Discovery*, vol. 9, no. 3, pp. 203-214, Mar.2010.

[3] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington, "Diagnosing the decline in pharmaceutical R&D efficiency," *Nat Rev Drug Discovery*, vol. 11, no. 3, pp. 191-200, Mar.2012.

[4] D. F. Horrobin, "Innovation in the pharmaceutical industry," *J. R. Soc. Med.*, vol. 93, pp. 341-345, 2000.

[5] A. Mullard, "2011 FDA drug approvals," *Nat Rev Drug Discovery*, vol. 11, no. February, pp. 91-94, 2012.

[6] L. M. Jarvis, "Biotechs Beat Pharma In First-Half Earnings," *Chem. Eng. News*, vol. 92, no. 32, pp. 16-17, Aug.2014.

[7] E. David, T. Tramontin, and R. Zemmel, "Pharmaceutical R&D: the road to positive returns," *Nat Rev Drug Discovery*, vol. 8, no. 8, pp. 609-610, Aug.2009.

[8] A. L. Hopkins, J. S. Mason, and J. P. Overington, "Can we rationally design promiscuous drugs?," *Curr. Opin. Struct. Biol.*, vol. 16, no. 1, pp. 127-136, Feb.2006.

[9] A. L. Hopkins, "Network pharmacology," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1110-1111, 2007.

[10] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nat. Chem. Biol.*, vol. 4, no. 11, pp. 682-690, 2008.

[11] P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends Pharmacol. Sci.*, vol. 26, no. 4, pp. 178-182, Apr.2005.

[12] O. Keskin, a. Gursoy, B. Ma, and R. Nussinov, "Towards drugs targeting multiple proteins in a systems biology approach," *Current topics in medicinal chemistry*, vol. 7, no. 10, pp. 943-951, Jan.2007.

[13] T. Korcsmaros, M. S. Szalay, C. Bode, I. A. Kovacs, and P. Csermely, "How to design multi-target drugs : Target search options in cellular networks," *Discovery*, pp. 1-10, 2007.

[14] G. R. Zimmermann, J. Leh+ír, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," *Drug Discovery Today*, vol. 12, no. 1-2, pp. 34-42, Jan.2007.

[15] S. K. Mencher and L. G. Wang, "Promiscuous drugs compared to selective drugs (promiscuity can be a virtue)," *BMC clinical pharmacology*, vol. 5, no. 1, p. 3, Jan.2005.

[16] P. Csermely, T. Korcsmaros, H. J. M. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review," *Pharmacol. Ther.*, no. 0 2013.

[17] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, D. Koller, R. B. Altman, R. W. Davis, C. Nislow, and G. Giaever, "The chemical genomic portrait of yeast: uncovering a phenotype for all genes," *Science*, vol. 320, no. 5874, pp. 362-365, Apr.2008.

[18] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101-113, Mar.2004.

[19] D. Cook, D. Brown, R. Alexander, R. March, P. Morgan, G. Satterthwaite, and M. N. Pangalos, "Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework," *Nat Rev Drug Discovery*, vol. 13, no. 6, pp. 419-431, 2014.

[20] J. Arrowsmith, "A decade of change," *Nat Rev Drug Discovery*, vol. 11, no. 1, pp. 17-18, 2012.

[21] S. H. Yook, Z. N. Oltvai, and A. L. Barabasi, "Functional and topological characterization of protein interaction networks," *PROTEOMICS*, vol. 4, pp. 928-942, Apr.2004.

[22] P. K. Sorger, S. R. B. Allerheiligen, D. R. Abernethy, R. B. Altman, K. L. R. Brouwer, A. Califano, Z. David, D. Argenio, R. Iyengar, W. J. Jusko, R. Lalonde, D. A.

Lauffenburger, B. Shoichet, J. L. Stevens, S. Subramaniam, P. V. D. Graaf, and R. Ward, "Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms," *NIH White Paper*, pp. 1-47, 2011.

[23] B. P. Zambrowicz and A. T. Sands, "Modeling drug action in the mouse with knockouts and RNA interference," *Drug Discovery Today*, vol. 3, no. 5, pp. 198-207, Oct.2004.

[24] S. I. Berger and R. Iyengar, "Network analyses in systems pharmacology," *Bioinformatics*, vol. 25, no. 19, pp. 2466-2472, Oct.2009.

[25] J. M. Berg, M. E. Rogers, and P. M. Lyster, "Systems biology and pharmacology," *Clin. Pharmacol. Ther.*, vol. 88, no. 1, pp. 17-19, July2010.

[26] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nat. Biotechnol.*, vol. 25, no. 2, pp. 197-206, 2007.

[27] E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Cote, B. K. Shoichet, and L. Urban, "Large-scale prediction and testing of drug activity on side-effect targets," *Nature*, vol. 486, no. 7403, pp. 361-367, June2012.

[28] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. Glennon, J. +. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, pp. 175-181, Nov.2009.

[29] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, p. i232-i240, 2008.

[30] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics (Oxford, England)*, vol. 27, no. 21, pp. 3036-3043, Nov.2011.

[31] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug-target interactions using bipartite local models," *Bioinformatics*, vol. 25, no. 18, pp. 2397-2403, Sept.2009.

[32] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, p. i246-i254, June2010.

[33] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining drug and gene similarity measures for drug-target elucidation," *J Comput Biol*, vol. 18, no. 2, pp. 133-145, Feb.2011.

[34] M. Gonen, "Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304-2310, June2012.

[35] S. L. Swann, S. P. Brown, S. W. Muchmore, H. Patel, P. Merta, J. Locklear, and P. J. Hajduk, "A unified, probabilistic framework for structure- and ligand-based virtual screening," *J. Med. Chem.*, vol. 54, no. 5, pp. 1223-1232, Mar.2011.

[36] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. Armstrong, S. J. Haggarty, P. Clemons, R. Wei, S. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929-1935, Sept.2006.

[37] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte, "Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease," *Sci. Transl. Med.*, vol. 3, no. 96, pp. 1-6, 2011.

[38] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte, "Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data," *Sci. Transl. Med.*, vol. 3, no. 96, p. 96ra77, 2011.

[39] A. P. Chiang and A. J. Butte, "Systematic Evaluation of Drug–Disease Relationships to Identify Leads for Novel Drug Uses," *Clin. Pharmacol. Ther.*, vol. 86, no. 5, pp. 507-510, 2009.

[40] T. Cheng, Q. Li, Y. Wang, and S. H. Bryant, "Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining," *J. Chem. Inf. Model.*, vol. 51, no. 9, pp. 2440-2448, 2011.

[41] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, pp. 1-9, 2011.

[42] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, and P. E. Bourne, "Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis," *PLoS Comput Biol*, vol. 5, no. 7, p. e1000423, 2009.

[43] Y. Y. Li, J. An, and S. J. M. Jones, "A computational approach to finding novel targets for existing drugs," *PLoS Comput Biol*, vol. 7, no. 9, p. e1002139, 2011.

[44] L. Xie and P. E. Bourne, "Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 14, pp. 5441-5446, Apr.2008.

[45] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 3: zooming in on protein-chemical interactions," *Nucleic. Acids. Res.*, vol. 40, no. Database issue, p. D876-D880, Jan.2012.

[46] A. Schuffenhauer, J. Zimmermann, R. Stoop, J. J. van der Vyver, S. Lecchini, and E. Jacoby, "An Ontology for Pharmaceutical Ligands and Its Application for in Silico Screening and Library Design," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 4, pp. 947-955, July2002.

[47] D. Butina, "Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets," *J. Chem. Inf. Comput. Sci.*, vol. 39, no. 4, pp. 747-750, 1999.

[48] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403-410, Oct.1990.

[49] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. Michael Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25-29, 2000.

[50] M. McGann, "FRED and HYBRID docking performance on standardized datasets," *J. Comput. Aided Mol. Des.*, vol. 26, no. 8, pp. 897-906, 2012.

[51] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin, "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *J. Med. Chem.*, vol. 47, no. 7, pp. 1739-1749, 2004.

[52] S. Riniker and G. A. Landrum, "Open-source platform to benchmark fingerprints for ligand-based virtual screening," *Journal of Cheminformatics*, vol. 5, no. 1, p. 26, 2013.

[53] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proceedings of the 25th international conference on Machine learning* Helsinki, Finland: ACM, 2008, pp. 880-887.

[54] W. Du and O. Elemento, "Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies," *Oncogene*, 2014.

[55] J. P. Bai and D. R. Abernethy, "Systems pharmacology to predict drug toxicity: integration across levels of biological organization*," *Annu. Rev. Pharmacol. Toxicol.*, vol. 53, pp. 451-473, 2013.

[56] T. Barrett and R. Edgar, "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis," *Methods Enzymol.*, vol. 411, no. 2005, pp. 352-369, Jan.2006.

[57] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic. Acids. Res.*, vol. 36, no. Database issue, p. D901-D906, Jan.2008.

[58] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic. Acids. Res.*, vol. 28, no. 1, pp. 27-30, 2000.

[59] S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, R. Schneider, R. Skoblo, R. B. Russell, P. E. Bourne, P. Bork, and R. Preissner, "SuperTarget and Matador: resources for exploring drug-target relationships," *Nucleic. Acids. Res.*, vol. 36, no. Database issue, p. D919-D922, 2008.

[60] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic. Acids. Res.*, vol. 30, no. 1, pp. 52-55, 2002.

[61] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic. Acids. Res.*, vol. 32, no. Database issue, p. D267-D270, 2004.

[62] R. Abagyan, M. Totrov, and D. Kuznetsov, "ICM – a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation," *J. Comput. Chem.*, vol. 15, no. 5, pp. 488-506, 1994.

[63] M. Rastegar-Mojarad, Z. Ye, J. M. Kolesar, S. J. Hebbring, and S. M. Lin, "Opportunities for drug repositioning from phenome-wide association studies," *Nat. Biotechnol.*, vol. 33, no. 4, pp. 342-345, 2015.

[64] F. J. de Serres, "Worldwide racial and ethnic distribution of alpha1-antitrypsin deficiency: summary of an analysis of published genetic epidemiologic surveys," *Chest*, vol. 122, no. 5, pp. 1818-1829, 2002.

[65] L. P. O'Reilly, O. S. Long, M. C. Cobanoglu, J. A. Benson, C. J. Luke, M. T. Miedel, P. Hale, D. H. Perlmutter, I. Bahar, G. A. Silverman, and S. C. Pak, "A genome-wide RNAi screen identifies potential drug targets in a C. elegans model of alpha1-antitrypsin deficiency," *Hum. Mol. Genet.*, 2014.

[66] S. Eriksson and C. Larsson, "Purification and partial characterization of PAS-positive inclusion bodies from the liver in alpha1-antitrypsin deficiency," *N. Engl. J. Med.*, vol. 292, no. 4, pp. 176-180, 1975.

[67] J. O. Jeppsson, C. Larsson, and S. Eriksson, "Characterization of alpha1-antitrypsin in the inclusion bodies from the liver in alpha1-antitrypsin deficiency," *N. Engl. J. Med.*, vol. 293, no. 12, pp. 576-579, 1975.

[68] R. W. Carrell and D. A. Lomas, "Alpha1-antitrypsin deficiency: a model for conformational diseases," *N. Engl. J. Med.*, vol. 346, no. 1, pp. 45-53, 2002.

[69] G. A. Silverman, J. C. Whisstock, S. P. Bottomley, J. A. Huntington, D. Kaiserman, C. J. Luke, S. C. Pak, J. M. Reichhart, and P. I. Bird, "Serpins flex their muscle I. Putting the clamps on proteolysis in diverse biological systems," *J. Biol. Chem.*, vol. 285, no. 32, pp. 24299-24305, 2010.

[70] R. Huber and R. W. Carrell, "Implications of the three-dimensional structure of alpha 1-antitrypsin for structure and function of serpins," vol. 28, no. 23, pp. 8951-8966, 1989.

[71] N. C. Perera, O. Schilling, H. Kittel, W. Back, E. Kremmer, and D. E. Jenne, "NSP4, an elastase-related protease in human neutrophils with arginine specificity," vol. 109, no. 16, pp. 6229-6234, 2012.

[72] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic. Acids. Res.*, vol. 28, no. 1, pp. 235-242, 2000.

[73] S. Ye, A. L. Cech, R. Belmares, R. C. Bergstrom, Y. Tong, D. R. Corey, M. R. Kanost, and E. J. Goldsmith, "The structure of a Michaelis serpin–protease complex," *Nat. Struct. Biol.*, vol. 8, no. 11, pp. 979-983, 2001.

[74] J. A. Huntington, R. J. Read, and R. W. Carrell, "Structure of a serpin-protease complex shows inhibition by deformation," *Nature*, vol. 407, no. 6806, pp. 923-926, Oct.2000.

[75] D. H. Perlmutter and G. A. Silverman, "Hepatic fibrosis and carcinogenesis in alpha1-antitrypsin deficiency: a prototype for chronic tissue damage in gain-of-function disorders," *Cold Spring Habor Perspectives in Biology*, vol. 3, no. 3 2011.

[76] M. Yamasaki, T. J. Sendall, M. C. Pearce, J. C. Whisstock, and J. A. Huntington, "Molecular basis of alpha1-antitrypsin deficiency revealed by the structure of a domain-swapped trimer," *EMBO reports*, vol. 12, no. 10, pp. 1011-1017, 2011.

[77] D. H. Perlmutter, "Pathogenesis of chronic liver injury and hepatocellular carcinoma in alpha-1-antitrypsin deficiency," *Pediatr. Res.*, vol. 60, no. 2, pp. 233-238, 2006.

[78] D. H. Perlmutter, "Alpha-1-antitrypsin deficiency: importance of proteasomal and autophagic degradative pathways in disposal of liver disease-associated protein aggregates," *Annu. Rev. Med.*, vol. 62, pp. 333-345, Jan.2011.

[79] S. Eriksson, J. Carlson, and R. Velez, "Risk of cirrhosis and primary liver cancer in alpha 1-antitrypsin deficiency," vol. 314, no. 12, pp. 736-739, 1986.

[80] R. G. Crystal, "Alpha 1-antitrypsin deficiency, emphysema, and liver disease. Genetic basis and strategies for therapy," vol. 85, no. 5, pp. 1343-1352, 1990.

[81] A. Janoff, "Elastases and emphysema. Current assessment of the protease-antiprotease hypothesis," vol. 132, no. 2, pp. 417-433, 1985.

[82] E. Janus, N. Phillips, and R. Carrell, "Smoking, lung function, and alpha1-antitrypsin deficiency," *Lancet*, vol. 325, no. 8421, pp. 152-154, 1985.

[83] D. Rudnick and D. H. Perlmutter, "Alpha-1-antitrypsin deficiency: a new paradigm for hepatocellular carcinoma in genetic liver disease," *Hepatology*, vol. 42, no. 3, pp. 514-521, Sept.2005.

[84] D. A. Rudnick, Y. Liao, J.-K. An, L. J. Muglia, D. H. Perlmutter, and J. H. Teckman, "Analyses of hepatocellular proliferation in a mouse model of alpha-1-antitrypsin deficiency," *Hepatology*, vol. 39, no. 4, pp. 1048-1055, 2004.

[85] T. Sveger, "Liver disease in alpha1-antitrypsin deficiency detected by screening of 200,000 infants," *N. Engl. J. Med.*, vol. 294, no. 24, pp. 1316-1321, 1976.

[86] M. E. MacDonald, C. M. Ambrose, M. P. Duyao, R. H. Myers, C. Lin, L. Srinidhi, G. Barnes, S. A. Taylor, M. James, and N. Groot, "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes," *Cell*, vol. 72, no. 6, pp. 971-983, 1993.

[87] C. A. Ross and S. J. Tabrizi, "Huntington's disease: from molecular pathogenesis to clinical treatment," *Lancet neurology*, vol. 10, no. 1, pp. 83-98, Jan.2011.

[88] R. M. Friedlander, "Apoptosis and caspases in neurodegenerative diseases," *N. Engl. J. Med.*, vol. 348, no. 14, pp. 1365-1375, 2003.

[89] P. H. Reddy, M. Williams, V. Charles, L. Garrett, L. Pike-Buchanan, W. O. Whetsell, G. Miller, and D. A. Tagle, "Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA," *Nat. Genet.*, vol. 20, no. 2, pp. 198-202, 1998.

[90] L. Mangiarini, K. Sathasivam, M. Seller, B. Cozens, A. Harper, C. Hetherington, M. Lawton, Y. Trottier, H. Lehrach, S. W. Davies, and G. P. Bates, "Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice," *Cell*, vol. 87, no. 3, pp. 493-506, 1996.

[91] J. Heemskerk, A. J. Tobin, and L. J. Bain, "Teaching old drugs new tricks," *Trends Neurosci.*, vol. 25, no. 10, pp. 494-496, 2002.

[92] X. Wang, S. Zhu, Z. Pei, M. Drozda, I. G. Stavrovskaya, S. J. Del Signore, K. Cormier, E. M. Shimony, H. Wang, R. J. Ferrante, B. S. Kristal, and R. M. Friedlander, "Inhibitors of cytochrome c release with therapeutic potential for Huntington's disease," *J. Neurosci.*, vol. 28, no. 38, pp. 9473-9485, Sept.2008.

[93] N. O'Boyle, C. Morley, and G. Hutchison, "Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit," *Chemistry Central Journal*, vol. 2, no. 1, p. 5, 2008.

[94] T. W. Harris, I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan, W. J. Chen, N. De La Cruz, P. Davis, M. Duesbury, R. Fang, J. Fernandes, M. Han, R. Kishore, R. Lee, H. M.

M++ller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E. M. Schwarz, M. A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, K. Yook, R. Durbin, L. D. Stein, J. Spieth, and P. W. Sternberg, "WormBase: a comprehensive resource for nematode research," *Nucleic. Acids. Res.*, vol. 38, no. Database issue, p. D463-D467, Jan.2010.

[95] D. D. Shaye and I. Greenwald, "OrthoList: a compendium of C. elegans genes with human orthologs," *PloS one*, vol. 6, no. 5, p. e20085, Jan.2011.

[96] C. R. Chong and D. J. Sullivan, "New uses for old drugs," *Nature*, vol. 448, no. 7154, pp. 645-646, 2007.

[97] M. C. Cobanoglu, C. Liu, F. Hu, Z. N. Oltvai, and I. Bahar, "Predicting Drug-Target Interactions Using Probabilistic Matrix Factorization," *J. Chem. Inf. Model.*, vol. 53, no. 12, pp. 3399-3409, Dec.2013.

[98] M. C. Cobanoglu, Z. N. Oltvai, D. L. Taylor, and I. Bahar, "BalestraWeb: efficient online evaluation of drug-target interactions," *Bioinformatics*, p. btu599, 2014.

[99] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic. Acids. Res.*, vol. 39, no. Database issue, p. D1035-D1041, Jan.2011.

[100] M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabasi, and M. Vidal, "Drug–target network," *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1119-1126, 2007.

[101] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," *Adv Neural Inf Process Syst*, 2008.

[102] M. K. Warmuth, J. Liao, G. R+ñtsch, M. Mathieson, S. Putta, and C. Lemmen, "Active learning with support vector machines in the drug discovery process," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 667-673, 2003.

[103] M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431-432, Feb.2011.

[104] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE. T. Automat. Contr.*, vol. 19, no. 6, pp. 716-723, 1974.

[105] E. R. Yera, A. E. Cleves, and A. N. Jain, "Chemical Structural Novelty: On-Targets and Off-Targets," *J. Med. Chem.*, vol. 54, no. 19, pp. 6771-6785, 2011.

[106] P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, and M. T. Stahl, "Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database," *J. Chem. Inf. Model.*, vol. 50, no. 4, pp. 572-584, Apr.2010.

[107] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic. Acids. Res.*, vol. 42, no. Database issue, p. D1091-D1097, Jan.2014.

[108] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein, "Graphlab: A new framework for parallel machine learning," 2010.

[109] D. Nolan and D. T. Lang, "Scalable Vector Graphics," in *XML and Web Technologies for Data Sciences with R* Springer, 2014, pp. 537-580.

[110] D. Lane, "Scalable vector graphics," *AMC*, vol. 10, p. 12, 2007.

[111] A. Quint, "Scalable vector graphics," *IEEE Multimedia*, vol. 10, no. 3, pp. 99-102, 2003.

[112] J. Ferraiolo, F. Jun, and D. Jackson, *Scalable vector graphics (SVG) 1.0 specification* iuniverse, 2000.

[113] M. C. Cobanoglu, Z. N. Oltvai, D. L. Taylor, and I. Bahar, "BalestraWeb: efficient online evaluation of drug-target interactions," *Bioinformatics*, vol. 31, pp. 131-133, 2015.

[114] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," 2008.

[115] H. F. Yu, C. J. Hsieh, S. Si, and I. Dhillon, "Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems," IEEE, 2012, pp. 765-774.

[116] H. F. Yu, C. J. Hsieh, S. Si, and I. S. Dhillon, "Parallel matrix factorization for recommender systems," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 793-819, 2014.

[117] G. Takacs, I. Pilaszy, B. Nemeth, and D. Tikk, "Scalable collaborative filtering approaches for large recommender systems," *The Journal of Machine Learning Research*, vol. 10, pp. 623-656, 2009.

[118] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *IEEE Computer*, vol. 42, no. 8, pp. 30-37, 2009.

[119] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 4: integration of protein-chemical interactions with user data," *Nucleic. Acids. Res.*, vol. 42, no. Database issue, p. D401-D407, Jan.2014.

[120] O. S. Long, "Genetic Modifiers that Affect the Accumulation of the Mutant Protein Alpha-1 Antitrypsin-Z." 2011.

[121] S. F. Altschul, T. L. Madden, a. a. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic. Acids. Res.*, vol. 25, no. 17, pp. 3389-3402, Sept.1997.

[122] R. Huang, N. Southall, Y. Wang, A. Yasgar, P. Shinn, A. Jadhav, D. T. Nguyen, and C. P. Austin, "The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics," *Sci. Transl. Med.*, vol. 3, no. 80, p. 80ps16, 2011.

[123] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, and S. Fitzgerald, "Ensembl 2012," *Nucleic. Acids. Res.*, p. gkr991, 2011.

[124] Y. Nikolsky, S. Ekins, T. Nikolskaya, and A. Bugrim, "A novel method for generation of signature networks as biomarkers from complex high throughput data," *Toxicol. Lett.*, vol. 158, no. 1, pp. 20-29, July2005.

[125] S. J. Gosai, J. H. Kwak, C. J. Luke, O. S. Long, D. E. King, K. J. Kovatch, P. A. Johnston, T. Y. Shun, J. S. Lazo, and D. H. Perlmutter, "Automated high-content live animal drug screening using C. elegans expressing the aggregation prone serpin alpha1-antitrypsin Z," *PloS one*, vol. 5, no. 11, p. e15460, 2010.

[126] Z. A. Knight, "Small molecule inhibitors of the PI3-kinase family," in *Phosphoinositide 3-kinase in Health and Disease* Springer, 2011, pp. 263-278.

[127] V. Samokhvalov, B. A. Scott, and C. M. Crowder, "Autophagy protects against hypoxic injury in C. elegans," *Autophagy*, vol. 4, no. 8, pp. 1034-1041, 2008.

[128] K. D. Kimura, H. A. Tissenbaum, Y. Liu, and G. Ruvkun, "daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans," *Science*, vol. 277, no. 5328, pp. 942-946, 1997.

[129] L. P. O'Reilly, J. A. Benson, E. E. Cummings, D. H. Perlmutter, G. A. Silverman, and S. C. Pak, "Worming our way to novel drug discovery with the Caenorhabditis elegans proteostasis network, stress response and insulin-signaling pathways," *Expert Opinion on Drug Discovery*, vol. 9, no. 9, pp. 1021-1032, 2014.

[130] O. S. Long, J. A. Benson, J. H. Kwak, C. J. Luke, S. J. Gosai, L. P. O'Reilly, Y. Wang, J. Li, A. C. Vetica, and M. T. Miedel, "A C. elegans model of human alpha1-antitrypsin deficiency links components of the RNAi pathway to misfolded protein turnover," *Hum. Mol. Genet.*, p. ddu235, 2014.

[131] N. M. OLBoyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, pp. 33-47, 2011.

[132] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: A free tool to discover chemistry for biology," *J. Chem. Inf. Model.*, vol. 52, pp. 1757-1768, 2012.

[133] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic. Acids. Res.*, vol. 42, no. D1, p. D199-D205, 2014.

[134] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic. Acids. Res.*, vol. 42, no. Database issue, p. D199-D205, Jan.2014.

[135] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. a. Gir+¦n, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. K+ñh+ñri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. J. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2015," *Nucleic. Acids. Res.*, pp. 1-8, Oct.2014.

[136] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[137] E. T. Lam, M. D. Ringel, R. T. Kloos, T. W. Prior, M. V. Knopp, J. Liang, S. Sammet, N. C. Hall, P. E. Wakely, and V. V. Vasko, "Phase II clinical trial of sorafenib in metastatic medullary thyroid cancer," *J. Clin. Oncol.*, vol. 28, no. 14, pp. 2323-2330, 2010.

[138] R. T. Kloos, M. D. Ringel, M. V. Knopp, N. C. Hall, M. King, R. Stevens, J. Liang, P. E. Wakely, V. V. Vasko, and M. Saji, "Phase II trial of sorafenib in metastatic thyroid cancer," *J. Clin. Oncol.*, vol. 27, no. 10, pp. 1675-1684, 2009.

[139] J. M. Llovet, S. Ricci, V. Mazzaferro, P. Hilgard, E. Gane, J. F. d. r. Blanc, A. C. de Oliveira, A. Santoro, J. L. Raoul, and A. Forner, "Sorafenib in advanced hepatocellular carcinoma," *N. Engl. J. Med.*, vol. 359, no. 4, pp. 378-390, 2008.

[140] V. Gupta-Abramson, A. B. Troxel, A. Nellore, K. Puttaswamy, M. Redlinger, K. Ransone, S. J. Mandel, K. T. Flaherty, L. A. Loevner, and P. J. O'Dwyer, "Phase II trial of sorafenib in advanced thyroid cancer," *J. Clin. Oncol.*, vol. 26, no. 29, pp. 4714-4719, 2008.

[141] B. Escudier, T. Eisen, W. M. Stadler, C. Szczylik, S. p. Oudard, M. Siebels, S. Negrier, C. Chevreau, E. Solska, and A. A. Desai, "Sorafenib in advanced clear-cell renal-cell carcinoma," *N. Engl. J. Med.*, vol. 356, no. 2, pp. 125-134, 2007.

[142] M. Rynn, J. Russell, J. Erickson, M. J. Detke, S. Ball, J. Dinkel, K. Rickels, and J. Raskin, "Efficacy and safety of duloxetine in the treatment of generalized anxiety disorder: a flexible-dose, progressive-titration, placebo-controlled trial," *Depress. Anxiety*, vol. 25, no. 3, pp. 182-189, 2008.

[143] J. Hartford, S. Kornstein, M. Liebowitz, T. Pigott, J. Russell, M. Detke, D. Walker, S. Ball, E. Dunayevich, and J. Dinkel, "Duloxetine as an SNRI treatment for generalized anxiety disorder: results from a placebo and active-controlled trial," *Int. Clin. Psychopharmacol.*, vol. 22, no. 3, pp. 167-174, 2007.

[144] D. L. Dunner, D. J. Goldstein, C. Mallinckrodt, Y. Lu, and M. J. Detke, "Duloxetine in treatment of anxiety symptoms associated with depression," *Depress. Anxiety*, vol. 18, no. 2, pp. 53-61, 2003.

[145] D. J. Goldstein, C. Mallinckrodt, Y. Lu, and M. A. Demitrack, "Duloxetine in the treatment of major depressive disorder: a double-blind clinical trial," *J. Clin. Psychiatry*, vol. 63, no. 3, pp. 225-231, 2002.

[146] M. J. Detke, Y. Lu, D. J. Goldstein, J. R. Hayes, and M. A. Demitrack, "Duloxetine, 60 mg once daily, for major depressive disorder: a randomized double-blind placebo-controlled trial," *J. Clin. Psychiatry*, vol. 63, no. 4, pp. 308-315, 2002.

[147] B. Kerzner, J. Corbelli, S. Sharp, L. J. Lipka, L. Melani, A. LeBeaut, R. Suresh, P. Mukhopadhyay, E. P. Veltri, and "Ezetimibe Study Group"., "Efficacy and safety of ezetimibe coadministered with lovastatin in primary hypercholesterolemia," *Am. J. Cardiol.*, vol. 91, no. 4, pp. 418-424, 2003.

[148] C. A. Dujovne, M. P. Ettinger, J. F. McNeer, L. J. Lipka, A. P. LeBeaut, R. Suresh, B. Yang, E. P. Veltri, and "Ezetimibe Study Group", "Efficacy and safety of a potent new selective cholesterol absorption inhibitor, ezetimibe, in patients with primary hypercholesterolemia," *Am. J. Cardiol.*, vol. 90, no. 10, pp. 1092-1097, 2002.

[149] C. Gagne, H. E. Bays, S. R. Weiss, P. Mata, K. Quinto, M. Melino, M. Cho, T. A. Musliner, B. Gumbiner, and "Ezetimibe Study Group"., "Efficacy and safety of ezetimibe added to ongoing statin therapy for treatment of patients with primary hypercholesterolemia," *Am. J. Cardiol.*, vol. 90, no. 10, pp. 1084-1091, 2002.

[150] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, and B. Al-Lazikani, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic. Acids. Res.*, vol. 40, no. D1, p. D1100-D1107, 2012.

[151] P. C. D. Hawkins, a. G. Skillman, and A. Nicholls, "Comparison of shape-matching and docking as virtual screening tools," *J. Med. Chem.*, vol. 50, no. 1, pp. 74-82, Jan.2007.

[152] R. Apweiler, C. O'onovan, M. Magrane, Y. am-Faruque, R. Antunes, B. Bely, M. Bingley, L. Bower, B. Bursteinas, and G. Chavali, "Reorganizing the protein space at

the Universal Protein Resource (UniProt)," *Nucleic. Acids. Res.*, vol. 40, no. Database issue, p. D71-D75, 2012.

[153] Y. Hamdi, H. Kaddour, D. Vaudry, S. Bahdoudi, S. Douiri, J. +. Leprince, H. Castel, H. Vaudry, M. C. Tonon, and M. Amri, "The octadecaneuropeptide ODN protects astrocytes against hydrogen peroxide-induced apoptosis via a PKA/MAPK-dependent mechanism," *PloS one*, vol. 7, no. 8, p. e42498, 2012.

[154] E. A. Thomas, G. Coppola, P. A. Desplats, B. Tang, E. Soragni, R. Burnett, F. Gao, K. M. Fitzgerald, J. F. Borok, and D. Herman, "The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice," *Proceedings of the National Academy of Sciences*, vol. 105, no. 40, pp. 15564-15569, 2008.

[155] J. Pallos, L. Bodai, T. Lukacsovich, J. M. Purcell, J. S. Steffan, L. M. Thompson, and J. L. Marsh, "Inhibition of specific HDACs and sirtuins suppresses pathogenesis in a Drosophila model of HuntingtonΓÇÖs disease," *Hum. Mol. Genet.*, vol. 17, no. 23, pp. 3767-3775, 2008.

[156] R. J. Ferrante, J. K. Kubilus, J. Lee, H. Ryu, A. Beesen, B. Zucker, K. Smith, N. W. Kowall, R. R. Ratan, and R. Luthi-Carter, "Histone deacetylase inhibition by sodium butyrate chemotherapy ameliorates the neurodegenerative phenotype in Huntington's disease mice," *The Journal of neuroscience*, vol. 23, no. 28, pp. 9418-9427, 2003.

[157] J. P. Dompierre, J. D. Godin, B. n. d. C. Charrin, F. P. Cordelieres, S. J. King, S. Humbert, and F. d. r. Saudou, "Histone deacetylase 6 inhibition compensates for the transport deficit in Huntington's disease by increasing tubulin acetylation," *The Journal of neuroscience*, vol. 27, no. 13, pp. 3571-3583, 2007.

[158] E. Hockly, V. M. Richon, B. Woodman, D. L. Smith, X. Zhou, E. Rosa, K. Sathasivam, S. Ghazi-Noori, A. Mahal, and P. A. Lowden, "Suberoylanilide hydroxamic acid, a histone deacetylase inhibitor, ameliorates motor deficits in a mouse model of Huntington's disease," *Proceedings of the National Academy of Sciences*, vol. 100, no. 4, pp. 2041-2046, 2003.

[159] S. W. Jones, "Calcium channels: unanswered questions," *J. Bioenerg. Biomembr.*, vol. 35, no. 6, pp. 461-475, 2003.

[160] O. Zhuchenko, J. Bailey, P. Bonnen, T. Ashizawa, D. W. Stockton, C. Amos, W. B. Dobyns, S. H. Subramony, H. Y. Zoghbi, and C. C. Lee, "Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the +¦1A-voltage-dependent calcium channel," *Nat. Genet.*, vol. 15, no. 1, pp. 62-69, 1997.

[161] H. T. Orr, M. y. Chung, S. Banfi, T. J. Kwiatkowski, A. Servadio, A. L. Beaudet, A. E. McCall, L. A. Duvick, L. P. Ranum, and H. Y. Zoghbi, "Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1," *Nat. Genet.*, vol. 4, no. 3, pp. 221-226, 1993.

[162] S. M. Pulst, A. Nechiporuk, T. Nechiporuk, S. Gispert, X. N. Chen, I. Lopes-Cendes, S. Pearlman, S. Starkman, G. Orozco-Diaz, and A. Lunkes, "Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2," *Nat. Genet.*, vol. 14, no. 3, pp. 269-276, 1996.

[163] Y. Kawaguchi, T. Okamoto, M. Taniwakiz, and M. Aizawa, "CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1," *Nat. Genet.*, vol. 8 1994.

[164] A. R. La Spada, E. M. Wilson, D. B. Lubahn, A. E. Harding, and K. H. Fischbeck, "Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy," *Nature*, vol. 352, no. 6330, pp. 77-79, 1991.

[165] R. Koide, T. Ikeuchi, O. Onodera, H. Tanaka, S. Igarashi, K. Endo, H. Takahashi, R. Kondo, A. Ishikawa, and T. Hayashi, "Unstable expansion of CAG repeat in hereditary dentatorubral–pallidoluysian atrophy (DRPLA)," *Nat. Genet.*, vol. 6, no. 1, pp. 9-13, 1994.

[166] J. M. Lamers, P. D. Verdouw, and J. Mas-Oliva, "The effects of felodipine and bepridil on calcium-stimulated calmodulin binding and calcium pumping ATPase of cardiac sarcolemma before and after removal of endogenous calmodulin," *Mol. Cell. Biochem.*, vol. 78, no. 2, pp. 169-176, 1987.

[167] B. del Rio, J. M. G. Pedrero, C. Martinez-Campa, P. Zuazua, P. S. Lazo, and S. Ramos, "Melatonin, an endogenous-specific inhibitor of estrogen receptor alpha via calmodulin," *J. Biol. Chem.*, vol. 279, no. 37, pp. 38294-38302, 2004.

[168] F. Radogna, L. Paternoster, M. De Nicola, C. Cerella, S. Ammendola, A. Bedini, G. Tarzia, K. Aquilano, M. Ciriolo, and L. Ghibelli, "Rapid and transient stimulation of intracellular reactive oxygen species by melatonin in normal and tumor leukocytes," *Toxicol. Appl. Pharmacol.*, vol. 239, no. 1, pp. 37-45, 2009.

[169] A. Seto-Ohshima, E. Lawson, P. C. Emson, C. Q. Mountjoy, and L. H. Carrasco, "Loss of matrix calcium-binding protein-containing neurons in Huntington's disease," *The Lancet*, vol. 331, no. 8597, pp. 1252-1255, 1988.

[170] A. V. Panov, C. A. Gutekunst, B. R. Leavitt, M. R. Hayden, J. R. Burke, W. J. Strittmatter, and J. T. Greenamyre, "Early mitochondrial calcium defects in Huntington's disease are a direct effect of polyglutamines," *Nat. Neurosci.*, vol. 5, no. 8, pp. 731-736, 2002.

[171] I. H. Page, A. C. Corcoran, H. P. Dustan, and T. Koppanyi, "Cardiovascular Actions of Sodium Nitroprusside in Animals and Hypertensive Patients," *Ciculation*, vol. 11, pp. 188-198, 1955.

[172] J. H. Tinker and J. D. Michenfelder, "Sodium nitroprusside: pharmacology, toxicology and therapeutics," *Anesthesiology*, vol. 45 1976.

[173] E. A. Kowaluk, P. Seth, and H. L. Fung, "Metabolic activation of sodium nitroprusside to nitric oxide in vascular smooth muscle," *J. Pharmacol. Exp. Ther.*, vol. 262, pp. 916-922, 1992.

[174] Y. Yoshioka, a. Yamamuro, and S. Maeda, "Nitric oxide at a low concentration protects murine macrophage RAW264 cells against nitric oxide-induced death via cGMP signaling pathway," *Br. J. Pharmacol.*, vol. 139, pp. 28-34, May2003.

[175] I. D. G. Duarte, B. B. Lorenzetti, and S. H. Ferreira, "Peripheral analgesia and activation of the nitric oxide-cyclic GMP pathway," *Eur. J. Pharmacol.*, vol. 186, no. 2, pp. 289-293, 1990.

[176] F. Murad, "The nitric oxide-cyclic GMP signal transduction system for intracellular and intercellular communication," *Recent Prog. Horm. Res.*, vol. 49, pp. 239-248, 1993.

[177] E. Southam and J. Garthwaite, "The nitric oxide-cyclic GMP signalling pathway in rat brain," *Neuropharmacology*, vol. 32, no. 11, pp. 1267-1277, 1993.

[178] A. Kyrola, G. Blelloch, and C. Guestrin, "GraphChi: Large-scale graph computation on just a PC," in *Proceedings of the 10th conference on Symposium on Operating Systems Design & Implementation* 2012.

[179] J. C. Saeh, P. D. Lyne, B. K. Takasaki, and D. A. Cosgrove, "Lead Hopping Using SVM and 3D Pharmacophore Fingerprints," *J. Chem. Inf. Model.*, pp. 1122-1133, 2005.

[180] E. Cohen, J. Bieschke, R. M. Perciavalle, J. W. Kelly, and A. Dillin, "Opposing activities protect against age-onset proteotoxicity," vol. 313, no. 5793, pp. 1604-1610, 2006.

[181] T. Hidvegi, M. Ewing, P. Hale, C. Dippold, C. Beckett, C. Kemp, N. Maurice, A. Mukherjee, C. Goldbach, S. Watkins, G. Michalopoulos, and D. H. Perlmutter, "An autophagy-enhancing drug promotes degradation of mutant alpha1-antitrypsin Z and reduces hepatic fibrosis," *Science*, vol. 329, no. 5988, pp. 229-232, July2010.

[182] S. Karve, M. E. Werner, R. Sukumar, N. D. Cummings, J. A. Copp, E. C. Wang, C. Li, M. Sethi, R. C. Chen, and M. E. Pacold, "Revival of the abandoned therapeutic wortmannin by nanoparticle drug delivery," *Proceedings of the National Academy of Sciences*, vol. 109, no. 21, pp. 8230-8235, 2012.

[183] K.-D. Schultz, K. Schultz, and G. Schultz, "Sodium nitroprusside and other smooth muscle-relaxants increase cyclic GMP levels in rat ductus deferens," *Nature*, vol. 265, pp. 750-751, Feb.1977.

[184] Hideaki Karaki, Koichi Sato, Hiroshi Ozaki, and Kazuyasu Murakami, "Effects of sodium nitroprusside on cytosolic calcium level in vascular smooth muscle," *Eur. J. Pharmacol.*, vol. 156, pp. 259-266, Nov.1988.

[185] Y. Yoshioka, T. Kitao, T. Kishino, a. Yamamuro, and S. Maeda, "Nitric Oxide Protects Macrophages from Hydrogen Peroxide-Induced Apoptosis by Inducing the Formation of Catalase," *The Journal of Immunology*, vol. 176, no. 8, pp. 4675-4681, Apr.2006.

[186] M. H. Lee, M. H. Jang, E. K. Kim, S. W. Han, S. Y. Cho, and C. J. Kim, "Nitric oxide induces apoptosis in mouse C2C12 myoblast cells," *Journal of Pharmacological Sciences*, vol. 97, pp. 369-376, 2005.

[187] H. j. Chae, H. s. So, S. w. Chae, J. s. Park, M. s. Kim, J. m. Oh, Y. t. Chung, S. h. Yang, E. t. Jeong, H. m. Kim, R. k. Park, and H. R. Kim, "Sodium nitroprusside induces apoptosis of H9C2 cardiac muscle cells in a c-Jun N-terminal kinase-dependent manner," *International Immunopharmacology*, vol. 1, no. 5, pp. 967-978, May2001.

[188] H. J. Kwak, K. M. Park, S. Lee, H. J. Lim, S. H. Go, S. M. Eom, and H. Y. Park, "Preconditioning with low concentration NO attenuates subsequent NO-induced apoptosis in vascular smooth muscle cells via HO-1-dependent mitochondrial death pathway," *Toxicol. Appl. Pharmacol.*, vol. 217, no. 2, pp. 176-184, Dec.2006.

[189] C. Stefanelli, C. Pignatti, B. Tantini, I. Stanic, F. Bonavita, C. Muscari, C. Guarnieri, C. Clo, and C. M. Caldarera, "Nitric oxide can function as either a killer molecule or an antiapoptotic effector in cardiomyocytes," *Biochim. Biophys. Acta*, vol. 1450, pp. 406-413, 1999.

[190] M. Kuhn, D. Szklarczyk, A. Franceschini, M. Campillos, C. von Mering, L. J. Jensen, A. Beyer, and P. Bork, "STITCH 2: an interaction network database for small molecules and proteins," *Nucleic. Acids. Res.*, vol. 38, no. Database issue, p. D552-D556, Jan.2009.

[191] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, and M. Rastegar-Mojarad, "Toward a complete dataset of drugΓÇôdrug interaction information from publicly available sources," *Journal of biomedical informatics*, vol. 55, pp. 206-217, 2015.