



**University of
Sunderland**




McGarry, Kenneth, Graham, Yitka, McDonald, Sharon and Rashid, Anuam (2018) RESKO: Repositioning drugs by using side effects and knowledge from ontologies. *Knowledge-Based Systems*, 160. pp. 34-48. ISSN 0950-7051

Downloaded from: <http://sure.sunderland.ac.uk/id/eprint/9701/>

Usage guidelines

Please refer to the usage guidelines at <http://sure.sunderland.ac.uk/policies.html> or alternatively contact sure@sunderland.ac.uk.

RESKO: Repositioning drugs by using side effects and knowledge from ontologies

Ken McGarry ^b, Yitka Graham ^a, Sharon McDonald ^b, Anuam Rashid^a

^aFaculty of Health Sciences and Well-being,
University of Sunderland, City Campus,
Sunderland, SR1 3SD, UK

^bFaculty of Technology,
University of Sunderland, St Peters Campus,
Sunderland, SR6 0DD, UK

Abstract

The objective of drug repositioning is to apply existing drugs to different diseases or medical conditions than the original target, and thus alleviate to a certain extent the time and cost expended in drug development. Our system RESKO, REpositioning drugs using Side Effects and Knowledge from Ontologies, identifies drugs with similar side-effects which are potential candidates for use elsewhere, the supposition is that similar side-effects may be caused by drugs targeting similar proteins and pathways. RESKO, integrates drug chemical data, protein interaction and ontological knowledge. The novel aspects of our system include a high level of biological knowledge through the use of pathway and biological ontology integration. This provides a explanation facility lacking in most of the existing methods and improves the repositioning process. We evaluate the shared side effects from the eight conventional Alzheimer drugs, from which sixty-seven candidate drugs based on a side-effect commonality were identified. The top 25 drugs on the list were further investigated in depth for their suitability to be repositioned, the literature revealed that many of the candidate drugs appear to have been trialed for Alzheimer's disease. Thus verifying the accuracy of our system, we also compare our technique with several competing systems found in the literature.

Keywords: side-effects; graph theory; pattern matching; protein targets; ontologies

1. Introduction

In this paper we demonstrate how adverse drug side-effects can be used to identify potential candidates for drug repositioning for a variety of diseases. Drug repurposing or repositioning involves using existing pharmaceutical products for diseases or problems they were not specifically designed for. There are many advantages since off-the-shelf drugs have undergone extensive testing and their toxicological properties are well known, therefore the costs are greatly reduced and also time to product delivery [31]. Thus it is more economical to re-purpose an existing drug than develop one from scratch [16]. Difficulties in drug development arise because diseases are often complex with multi-factorial components such as interactions between genes, proteins and the environment [7]. Furthermore, drugs that are highly selective in terms of their targets are very rare. Many patients when taking a drug will experience unwanted side-effects as the medication may also also interact to varying degrees with non-target proteins [52]. However, this feature can be used to search for drugs with similar side-effects that might target the defective biological functions more effectively than conventional

drugs. Since there is wealth of freely available drug and protein databases, drug repositioning is an ideal application area for knowledge based systems and computational statistics.

However, not all of the drug repositioning discoveries are through computational intelligence techniques. Interestingly, there are many examples where unanticipated side-effects have proven to be beneficial to patients suffering from unrelated problems to the original purpose of the drug thus allowing the drugs to be re-deployed [53]. The most often cited example is Sildenafil, a drug developed by Pfizer which was intended to treat heart problems by allowing better blood flow. It was discovered to have a particular side-effect on the male participants, it was later marketed as Viagra, the drug now has annual sales in excess of \$1.6 Billion [1]. Other notable drugs such as the infamous Thalidomide that caused birth defects in the 1950s, has been redeployed to treat leprosy and multiple myeloma [45].

A deeper understanding of the causes of disease is necessary, in particular knowledge of the genetic differences between individuals will eventually lead to improved treatments [40]. This has only recently been made possible by the development of advanced genomic and proteomic techniques which are able to provide detailed and accurate data on individual cellular processes [12]. We are

Email address: ken.mcgarry@sunderland.ac.uk (Ken McGarry



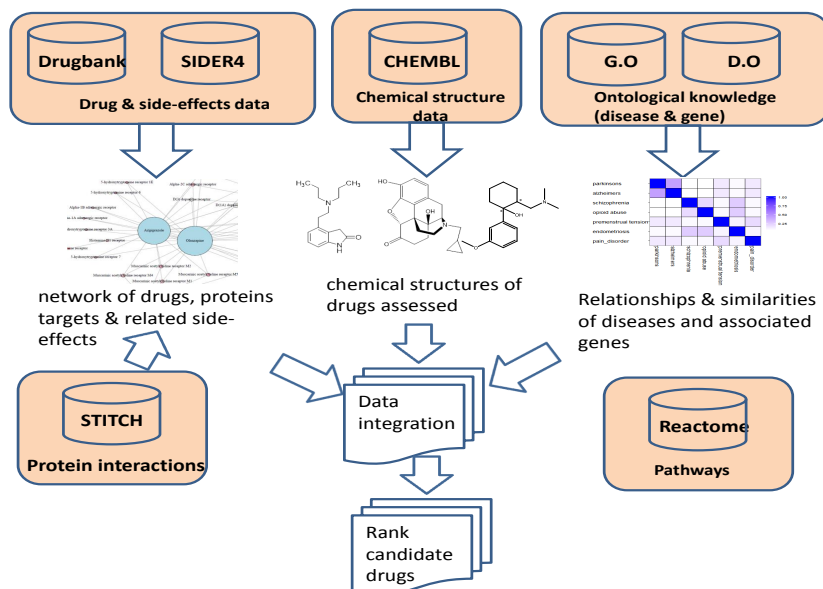


Figure 1: Overview of system operation, showing database sources, data flow and statistical analysis

now able to determine which genes (and proteins) interact together and form related functional groups that modify the behavior and ultimately the health of the cell [36]. Recent bioinformatic studies on cooperating modular groups of genes have suggested that diseases themselves are networked together [18]. The concept of the human disease network or *diseasome* is relatively new and is now starting to be explored as means of developing new drug products to tackle and combat diseases [21].

Our objective is to identify drug repurposing candidates for Alzheimer’s disease, the approach taken in this paper is to view the problem as one of identifying and assessing side-effects in known drugs. The system we developed is shown in Figure 1, we obtain and examine the side-effects of the eight most common Alzheimer’s (e.g. Donepezil, Galantamine, Rivastigmine, Citalpram, Risperidone, Memantine, Escitalopram and Aripiprazole) drugs found in the Side Effect Resource (SIDER4) database [27]. The drugs were compared against their chemical properties, on-targets and off-targets which provided further information necessary to prune down the list of potential candidate drugs.

The other sources of information include the gene ontology and the disease ontology. These provide expert level descriptions of biological functions, structures and the genes involved in the pathways targeted by the drugs and represent higher level knowledge. Furthermore, they provide an indication of similarity between diseases and links between shared genes and pathways. In this paper we extend previous work by incorporating chemical structure information from the Chemical European Molecular Biology Laboratory (ChEMBL) database [9] and also knowledge from the two main biological ontologies (gene and

disease ontologies) to enhance biological understanding. Integrating ontologies within a larger system has been successfully implemented by researchers in the past [44, 29].

The main technical challenge of our work is to integrate the disparate sources of data and knowledge in a principled way we use matrix algebra and a modified version of Zitnik’s data fusion scheme using matrix factorization, originally developed for integrating molecular data to discover disease-to-disease associations [60]. In Figure 2 each of the sources of data (chemical structure, protein interaction, gene ontology, diseases ontology and pathway information) are in matrix format, the ontology information is converted into separate matrices annotating the proteins, the overall relation matrix used for creating the complex network is very sparse. The equations and processes are explained in greater detail in the methods section.

The remainder of this paper is structured as follows. Section two reviews the related work and places the novelty of our methods into context; section three describes our methodology, the data and the algorithm’s developed. Section four highlights the results with special emphasis on Alzheimer’s disease and we compare our results with competing systems using other diseases. Finally, section five presents the conclusions along with future work.

2. Related work

In recent years a variety of computational methods have been applied to drug repositioning problems, we discuss the features and limitations of the most important systems. The PREDICTING Drug IndiCaTions (PREDICT) system tackles the issue by building drug-to-drug similarities and disease-to-disease indications for personalized medicine [19]. The main limitation of PREDICT

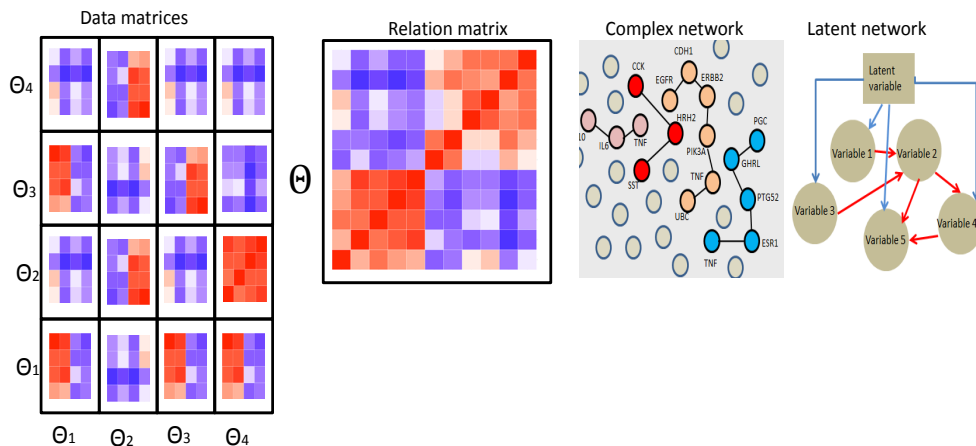


Figure 2: Matrix combination and transformation. The four main data types (protein interactions, chemical structure, ontology enrichment and pathway data (represented by $\Theta_1 \dots \Theta_4$) are integrated using matrix factorization. The large, sparse relation matrix is used to build complex networks. The latent network is derived from the adjacency matrix holding the complex network connectivity.

is that it lacks drug-to-target information which is necessary for understanding rational drug design and hence improves accuracy. The Connectivity MAP (CMAP) system was developed by compiling the data available from Affymetrix genechips [12]. CMAP integrates the *eXtreme Sum* (XSum) scoring algorithm which can identify drug-interaction pairs, however the method lacks reliable validation data as some diseases had worse than random performance. A different approach was taken by the Guilt By Association (GBA) method [13]. GBA system mapped the diseases to one logical set and mapped the drugs to another set, the degree of overlap between the two sets was determined systematically. Only a small number of drug candidates could be verified and the false positive rate appears rather high and the system is potentially biased with better results for older drugs. The Similarity-based Large-margin learning of Multiple Sources (SLAMS) method integrates several sources of data such as chemical structure and protein targets, chemical data is integrated through the creation of binary fingerprints [59]. SLAMS appears to outperform many competing repositioning but the authors conclude it may have a high false positive rate.

Other methods consider using graph network theory to integrate diverse data types and provide statistical information. Detecting motifs of bi-cliques extracted from a drug-target-disease network can be achieved by integrating disease-to-chemical associations [14]. Motifs are statistically significant patterns of connections that represent a biologically relevant function or activity. The bi-cliques produced appear to have a high false positive rate through using only the complex networks structure. Another approach, the Heterogeneous Graph Based Inference (HGBI) method implemented a complex network analysis [49]. The complex network consisted of a triple layer of interconnected nodes using protein interaction and drug target data. The HGBI system was applied to diseases with no known drugs and was able to perform well by

ranking drugs that have received attention in the clinical trials literature. Computational complexity may limit the HGBI system with the addition of further layers.

The multiple Drug information sources and multiple Disease information sources to facilitate drug Repositioning tasks (DRR) method is notable because it generates a list of proteins to be targeted by specific drugs using an omics based approach [57]. DDR used the Online Mendelian Inheritance In Man (OMIM) database to provide information on gene mutations and their frequencies to generate a candidate list of 524 Alzheimer’s disease implicated proteins. Some were known from the literature and some were speculative. DDR is probably more accurate than most repositioning systems but lacks biological knowledge. The similarity Measures and BI-Random Walk (MBiRW) method uses an iterative procedure integrating a random walk to measure similarity of drugs and diseases [34, 35]. MBiRW incorporates novel similarity measures and is well validated against gold standard data but lacks target information and biologically relevant information. Other work on complex networks has revealed further insights into why drug similarity does not always account for identifying useful drugs for repositioning[20]. It is often the case that network evaluation of many diseases and drug pairs reveals how diseases and their drugs are clustered into neighborhoods of proximity, though it does not have the infrastructure to explain why. Other researchers has concentrated on applying various combinations of methods [33, 41, 53, 50, 32].

The novel aspects of our work include a strong level of biological knowledge through the use of pathway and biological information ontology integration. This provides a level of understanding lacking in most of the existing methods and improves the repositioning process. Integration of biological knowledge enables the users to relate drugs to their side-effects and on/off protein targets. The disparate sources of data are integrated through the use of matrix

factorization. We validate our work through the analysis of our top scoring drugs against evidence from the drug repurposing literature and compare our method against several of the most similar computational techniques. We use various statistical measures to assess and rank our drug candidates.

3. Methods

In the following discussions, reference should be made to Figure 3 for clarification of dataflow and processing. The stages required to produce a list of candidate drugs that exhibit potential to be re-purposed are:

- Determine disease of interest
- Search Drugbank for main conventional drug treatments
- Search SIDER4 for side-effects associated with each drug
- Obtain list of joint side-effects common to all treatments
- Use list to search SIDER4 for any drug sharing at least 50% of these side-effects
- Obtain chemical structures, protein interactions, biological pathways for these drugs
- Annotate proteins with gene and disease ontology terms
- Build matrices of annotations, chemical structures and proteins and integrate matrices
- Build a complex network, calculate statistics
- Convert graph into latent network, determine effects of latent variables
- Calculate ranking for each candidate drug

3.1. Data bases and sources of information

The initial starting point for using our system is to obtain the Unified Medical Language System (UMLS) code for the disease of interest. UMLS codes provide a unique ID for each disease and sub-disease, and is used to interrogate the DrugBank database to provide a list of known, conventional treatments. Drugbank contains the majority of drugs that are currently prescribed, or have been withdrawn or are at the clinical trial stage [28]. This resource is widely used by those developing drugs, chemists, pharmacologists and others involved in pharmaceuticals research. Every drug is listed with its main targets, known off-targets along with chemical structure and other important characteristics.

Having obtained a list of conventional drugs used to treat the disease of interest, the SIDER4 database is accessed to give a list of side-effects for each drug [26].

SIDER4, an important repository of hundreds of drugs and thousands of their known side-effects on humans. This freely available database contains more than 1,430 drugs and 58,880 side effects. The information on the drugs is collected, from the national registries and charity organisations [27]. In addition to side-effects for the 1,430 drugs SIDER4 also has information on the frequency of their relative occurrence in patients. It should be noted that SIDER4 contains many more side-effects per drug than would normally appear in the packet-insert that is included in every drug prescription. We did not use any of the adverse effect ontologies such as Ontology Adverse Effects [22]. Although these ontologies could provide a deeper understanding of the relationship between treatments and side-effects, the pattern matching algorithm we used in algorithm 1 is sufficient for our purposes.

We removed from SIDER4, 10% of the most common side-effects that occur with most drugs such as Nausea, Dermatitis, Rash, Vomiting, Headache, Dizziness etc as per the work of Atias [3]. In total, 573 side-effect types were removed. Further preprocessing replaced lengthy drug/compound names with their drugbank identifiers. For example (10ALPHA,13ALPHA,14BETA,17ALPHA)-17-HYDROXYANDROST-4-EN-3-ONE becomes DB07768.

Our basic proposition is that shared side-effects between the drugs used to treat a disease can be combined into a search pattern to interrogate the SIDER4 database again to identify candidate drugs causing similar side-effects. The assumption is that these drugs are targeting similar biochemical pathways (inadvertently) and therefore may be candidates for re-purposing to the original disease. The side-effects pertaining to the nine drugs commonly used to treat Alzheimer’s Disease (AD), such as Donepezil, Galantamine and Rivastigmine were obtained from the SIDER4 database.

Once a list of new, candidate drugs has been generated and sorted according to their percentage side-effect similarity, other databases can be interrogated for chemical structure, protein targets and biological pathways relating to these drugs. These results are further enhanced by enrichment from gene ontology and disease ontology for biological plausibility.

The chemical structure of the candidate drugs was downloaded from ChEMBL, this database contains chemical compounds represented as text data called the Simplified Molecular Input Line Entry System (SMILES) format [51]. We created a series of fingerprints for each drug based on atom-pair arrangements of 2048 bits in size. We consider the chemical structure of the candidate drugs useful information to assess their suitability for repositioning.

The candidate drugs have known on-target and off-target proteins, this knowledge is augmented by accessing protein-to-protein interactions and drug/chemical to protein interactions found in the search tool for interactions of chemicals (STITCH) database [38]. The Search Tool for the Retrieval of Interacting Genes (STRING) database

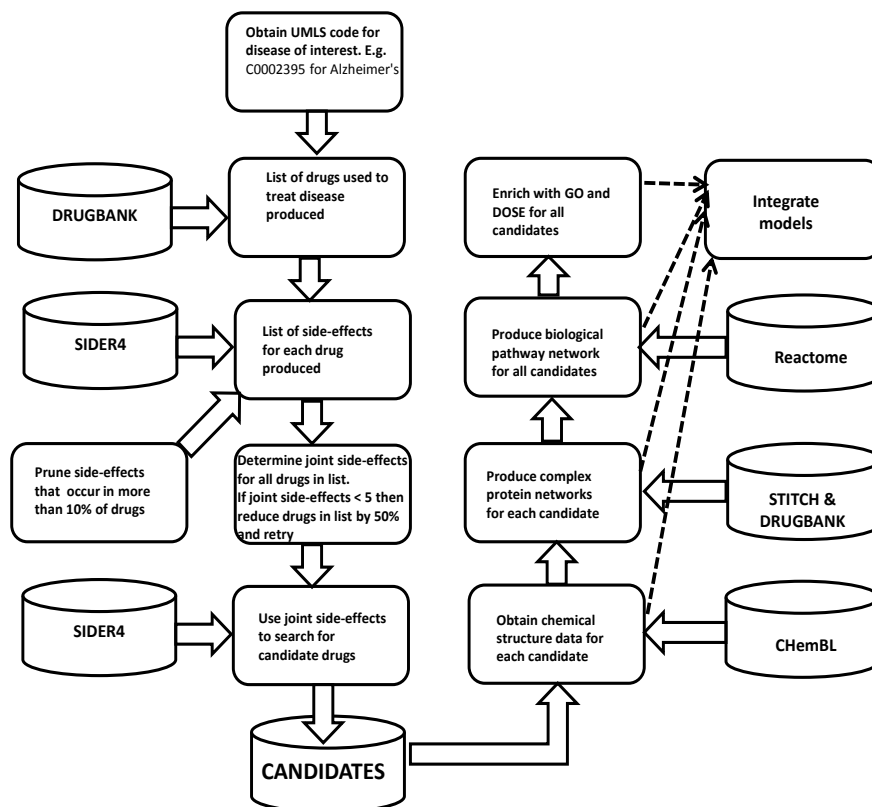


Figure 3: Detailed step by step process diagram

makes use of over six million known protein-to-protein interactions discovered by text mining and annotation by human experts [46]. Additionally, a further 30,000 associations are predicted based on similar protein characteristics using predictive data mining. These databases contains protein to chemical mappings and augments the information held in Drugbank which may not contain all the potential drug to protein interactions.

The Reactome database is accessed to provide indications of which biological pathways are involved with the on-target and off-target protein interactions affected by the candidate drugs [54]. Deeper insights can be gained into the biological functions by allowing users of the system to relate side-effects to drugs and then to the pathways. Individual proteins must not be regarded in isolation but seen as a system of interacting components. The associated pathways are incorporated into the overall drug ranking measure.

3.2. Searching for side-effect similarity

The work described in this paper extends the complex network and pattern matching algorithms previously developed by the authors [37] to include a novel data integration method that assesses the candidate drugs, their chemical characteristics, associated proteins and biological pathways as described in section 3.7.

Referring to algorithm 1, lines 2-7 perform the initialization of key values: to obtain the UMLS code for the disease of interest, for the minimum number of side-effects for a viable search to be conducted (set at ≥ 3), the number of conventional drugs currently treating this disease is set to zero and finally a percentage value of side-effects similarity for a candidate drug to be considered useful (set at 50%).

Lines 8-9 obtain the drugs currently treating the disease of interest. Lines 10-17 ensure that each conventional drug is annotated with a list of more than three or more side-effects, if any drug has fewer than three side-effects it is discarded. The critical processing is performed by line 18 which creates a list of the side-effects common to all of the conventional drugs. This list of common side-effects is used by lines 19-25 to search SIDER4 and return any candidate drug that has at least one side-effect in the search list. The new list of repositioned candidate drugs is sorted by lines 26-27, in descending order and only those with greater than 50% similarity are returned from the function call.

Empirically, it was determined that conventional drugs with fewer than three side-effects are generally not useful as they tend to have either highly specific or overly common side-effects. They cannot provide a rich enough search pattern and these are pruned from the list of drugs. The common side-effects for the remaining conventional

Algorithm 1 Identification of drugs with similar side-effects

```
1: procedure SEARCHSIDEFFECTS(DRUGBANK, SIDER4, UMLS) ▷ Databases used
2:   do initialize
3:   umls ← get UMLS code for disease ▷ e.g. C0002395 for Alzheimer's
4:   SEthreshold = 3 ▷ A drug must have 3 or more side-effects to be useful
5:   drugcount = 0 ▷ setup for N drugs for the disease
6:   SIMcutoff = 50 ▷ > 50 percent similarity required
7:   end initialize
8:   DrugList ← DRUGBANK[umls] ▷ Search DRUGBANK for drugs treating this disease
9:   drugcount ← length(DrugList)
10:  for i ≤ drugcount do
11:    ise ← SIDER4[i] ▷ For each drug get associated side-effects
12:    secount[i] ← length(ise) ▷ count side-effects
13:    if secount[i] > SEthreshold then
14:      DrugList[i] ← DrugList[i] ▷ Update DrugList with usable drugs
15:      SE[i] ← ise ▷ Save Drug side-effects
16:    end if
17:  end for
18:  JSE ∈ SE[i] ▷ Get Joint Side-Effects for Current drugs
19:  for j ≤ DrugList[i] do
20:    ReposList[j] ← SIDER4 ∩ JSE ▷ Search SIDER4 for any new drug with these side-effects
21:    Drugpercentage[j] ← CalcPercentageSimilarity(ReposList[j]) ▷
22:    if Drugpercentage[j] > SIMcutoff then
23:      ReposList[j] ← ReposList[j] ▷ Create ReposList with candidate drugs
24:    end if
25:  end for
26:  SortReposList[j]descendingorder
27:  return ReposList, SE, SEcommon ▷ Return drugs with > 50 percent similarity and their side-effects
28: end procedure
```

drugs are determined, this is used to search SIDER4 again for any candidate drug having at least one side-effect in common.

3.3. Graph theory

We use graph theoretic methods to build networks of interacting proteins such as the drug on-targets and off-targets (where known). These networks provide various statistical measures describing the relationships between the interacting proteins and the candidate drugs, with some tentative indications of why a particular side-effect occurs. We incorporate these statistical measures into the overall integration mechanism, which contributes in ranking the drugs in terms of usefulness for repositioning. Graph creation and inferencing is usually through matrix algebra, edge lists are converted into connectivity matrices, we can define a graph $G = (V, E)$ where the nodes also called vertices V containing links called edges E . In our application, we determine the relevance of protein connectivity patterns using criteria from the graph theoretic centrality statistics [17, 6].

We use a number of commonly used graph-based measures to evaluate the protein networks. The closeness statistic (CC) provides a measure of how near each node is to every other node in the network. Some nodes may be more prominent than others due to their topology. The distance $d(v_i, v_j)$ from a given protein i to another j is generally characterized as the number of links in the shortest path between them, N is the number of all proteins (nodes) in the network. We define the closeness centrality of protein i :

$$CC(v_i) = \frac{N-1}{\sum_j d(v_i, v_j)} \quad (1)$$

The clustering coefficient (C_i) provides a local measure of modularity of the network in terms of shared components and interactions for each vertex. Thus a randomly connected network would have a very different coefficient to a biologically cohesive network [5].

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \quad (2)$$

Where L_i defines the number of links between k_i neighbors of vertex i , k_i is the degree (number of connections) of node i . The coefficient C_i is bounded between 0 and 1.

The clustering coefficient for the entire graph is simply the average of the local clustering values $\langle C \rangle$ of each vertex C_i for all nodes $i = 1, \dots, N$.

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i \quad (3)$$

3.4. Gene ontology

Gene ontology (GO) provides useful biological information and it is recognized as the *de facto* standard for gene product annotation [29]. For our purposes, we produce a binary matrix of proteins annotated with the presence or absence of terms. This enables an assessment to be made regarding the biological plausibility of the interacting proteins to observe the extent they actually cooperate in viable biological functions rather than random or spurious associations. For each protein associated with the candidate drugs enrichment was performed using gene ontology (GO), the enrichment is based on similarity measures using information content techniques. The content of which is calculated by taking the negative log probability of the terms t appearing in the database. A number of measures can be used to rank semantic similarity, we used the Wang

criterion as it reflects the biological plausibility better than other measures because of the way semantic similarity of the GO terms are calculated, using both the locations of the terms in the GO graph and their relations with their ancestor terms [48].

$$Wang(A, B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)} \quad (4)$$

For the Wang equation, where $S_{A(t)}$ represents the S-value of GO term t related to term A and $S_{B(t)}$ is the S-value of GO term t related to term B , $SV(A)$ and $SV(B)$ are the semantic values of GO term A and B .

3.5. Drug chemical fingerprints

The chemical data held in the large SDF file is used to create a unique fingerprint for each drug. We used the Tanimoto similarity coefficient was to compare the chemical structure of our 77 drugs based on pairwise binary fingerprints [11]. The Tanimoto coefficient generally gives better results than other metrics. Equation 5 shows the generally accepted form of the Tanimoto coefficient.

$$T(N_a, N_b) = \frac{N_c}{N_a + N_b - N_c} \quad (5)$$

where: N_a is the number of unique fragments in the first compound and N_b is the number of unique fragments in the second compound; N_c is the shared amount of overlapping fragments. Thus we would expect to see drugs with similar chemical structures to have similar properties and medicinal effects.

3.6. Disease ontology

We integrated information from the Disease Ontology (DO) to annotate the drug targets and other proteins to obtain a deeper interpretation of the biological processes and structures [30, 43]. The DO database contains knowledge on 8,043 inherited, developmental and acquired human diseases. Through enrichment analysis, the R package DOSim is able to explore the biological meaning of related genes in terms of structure, function and hierarchy. The concepts in DOSim are organized into a directed acyclic graph (DAG) similar to a tree structure, the concepts are linked by 'is-a' relationships. The lower the term or concept is positioned in the hierarchy then the more specific the term is, higher-up terms describe higher level or more abstract concepts. In order to identify biological themes the Hypergeometric model is used to assess the frequency of those semantic terms in a list and to determine whether the number of selected genes associated with disease is larger than expected than by chance alone.

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \quad (6)$$

Where N is the total number of genes present in the distribution, M is the number of genes within that distribution that are annotated (either through primary links or secondary links) to the node of interest, n is the size of the list of genes of interest and k is the number of genes within that list related to the node.

3.7. Data integration

Individually, each method of data analysis provides important information in a specific area, however, our systems strength comes from the integration of these disparate sources of knowledge in a principled way [8].

$$\| \mathbf{R} \approx \mathbf{G} \mathbf{S} \mathbf{G}^T \|^2 + \sum_{t=1}^4 tr(\mathbf{G}^{T(t)} \mathbf{G}) \quad (7)$$

where: \mathbf{R} is the relation matrix used to construct the complex network, \mathbf{G} identifies the drugs, \mathbf{S} is the weighting factor for the drugs, \mathbf{G}^T is the transpose, $\Theta^{(t)}$ represents the matrices containing the variables for PPI, drug structure, GO, DO and pathways.

$$\Theta^{(t)} = \begin{pmatrix} \Theta_1 & 0 & 0 & 0 \\ 0 & \Theta_2 & 0 & 0 \\ 0 & 0 & \Theta_3 & 0 \\ 0 & 0 & 0 & \Theta_4 \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 0 & \mathbf{R}_{12} & \mathbf{R}_{13} & \mathbf{R}_{14} \\ \mathbf{R}_{21} & 0 & 0 & 0 \\ \mathbf{R}_{31} & 0 & 0 & 0 \\ \mathbf{R}_{41} & 0 & 0 & 0 \end{pmatrix}$$

Similarly, the \mathbf{G} and \mathbf{S} block matrix factors are structured and calculated by:

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 & 0 & 0 & 0 \\ 0 & \mathbf{G}_2 & 0 & 0 \\ 0 & 0 & \mathbf{G}_3 & 0 \\ 0 & 0 & 0 & \mathbf{G}_4 \end{pmatrix}, \mathbf{S} = \begin{pmatrix} 0 & \mathbf{S}_{12} & \mathbf{S}_{13} & \mathbf{S}_{14} \\ \mathbf{S}_{21} & 0 & 0 & 0 \\ \mathbf{S}_{31} & 0 & 0 & 0 \\ \mathbf{S}_{41} & 0 & 0 & 0 \end{pmatrix}$$

3.8. Model validation strategies

The large, integrated matrix network containing the protein to protein interaction network, the chemical data, pathway and ontologies is converted into a latent network model. The advantages for doing so include the ability to model latent variables in the form of vertex classes and thus perform classification metrics such as accuracy, true positive rate, false positive rate and allows for plotting of ROC curves and precision-recall curves. The modeling of variables that are generally unobserved but are likely to play a major role in vertex connectivity is motivated by the assumption that lack of covariate information allows the vertices's to be exchangeable in a principled way [24]. The disadvantage is that there are many possible configurations of networks and pair-specific effects. They are both implemented and validated through Bayesian methods using Monte Carlo Markov Chain techniques (MCMC) to simulate from the posterior distributions the model specification [25].

In particular we use 10-fold cross-validation to test the goodness-of-fit, here the model is randomly divided into ten subsets of equal size and in each cross-validation trial one subset is taken for the test set while remaining nine folds comprise the training set [10].

We define the latent variables U_i as a set of vectors $(U_i, \dots, U_{iQ})^T$, these formed from the adjacency matrix \mathbf{U} from the complex network of protein-protein interactions, the pathways and ontologies. We build several such matrices to highlight different complexities of model (i.e. PPI only, PPI+Pathway, PPI+Pathway+Ontology, all data sources).

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \mathbf{X}, u_1, \dots, u_{N_v}) = \prod_{i < j} P_{ij} (1 - P_{ij})^{1 - y_{ij}} \quad (8)$$

$$\begin{array}{c} \mathbf{U} \text{ covariates/adjacency} \\ \begin{pmatrix} U_{11} & U_{12} & U_{13} & U_{14} \\ U_{21} & U_{22} & U_{23} & U_{24} \\ U_{31} & U_{32} & U_{33} & U_{34} \\ U_{41} & U_{42} & U_{43} & U_{44} \end{pmatrix} \end{array} \times \begin{array}{c} \mathbf{\Lambda} \text{ Latent variables} \\ \begin{pmatrix} U_{\{\{11\}\}} & 0 & 0 & 0 \\ U_{\{\{21\}\}} & U_{\{\{22\}\}} & 0 & 0 \\ U_{\{\{31\}\}} & U_{\{\{32\}\}} & U_{\{\{33\}\}} & 0 \\ U_{\{\{41\}\}} & U_{\{\{42\}\}} & U_{\{\{43\}\}} & U_{\{\{44\}\}} \end{pmatrix} \end{array} = \begin{array}{c} \widehat{\mathbf{U}} \widehat{\mathbf{\Lambda}} \widehat{\mathbf{U}}^T \text{ Eigenvalues} \\ \begin{pmatrix} U_{\{11\}} & 0 & 0 & 0 \\ 0 & U_{\{22\}} & 0 & 0 \\ 0 & 0 & U_{\{33\}} & 0 \\ 0 & 0 & 0 & U_{\{44\}} \end{pmatrix} \end{array} \quad (9)$$

In equation 9 the notation of the three matrices where : U_{11} indicates in \mathbf{U} the simple index notation to access its elements, however in $\mathbf{\Lambda}$ implies a symmetric matrix and is expressed by multiset index notation $U_{\{\{11\}\}}$. In matrix $\widehat{\mathbf{U}} \widehat{\mathbf{\Lambda}} \widehat{\mathbf{U}}^T$ the elements are eigenvalues can be accessed by $U_{\{11\}}$. The cross-validated eigenmodels enable predictions to be made on edge status to generate metrics for goodness of fit such as accuracy and precision to be assessed.

3.9. Ranking candidate drugs

The final data analysis phase is to rank and assess the list of candidate drugs, this is a two stage process. First, each candidate drug is individually compared with the conventional drugs using the generalized Jaccard coefficient (equation 10) which can manage continuous values.

$$\text{Score}_{ij} = \frac{|F(D_i) \cap F(D_j)|}{|F(D_i) \cup F(D_j)|} \quad (10)$$

Where: $F(D_j)$ are the variables of interest, such as protein interactions, pathways, ontology annotation and chemical similarity shared between the candidate drugs and the conventional drugs $F(D_i)$. This produces a matrix containing individual values for each drug. However, these need to be modified by the overall strength and effect of the variables, similar to coefficients in a regression equation. The second stage calculates the actual score given to each candidate drug by equation 11 which takes the difference for each candidate/conventional pair multiplied by the $\widehat{\mathbf{U}} \widehat{\mathbf{U}}^T$ value for that variable.

The MCMC eigenmodel [23] and shown in equation 8, simulates models from the relevant posterior priors, describing the influence of pairs of covariates X_{ij} . A series of transformations convert the elements of a diagonal matrix giving the relative importance of each latent variable (U_i). The product $\widehat{\mathbf{U}} \widehat{\mathbf{U}}^T$ holds the eigenvectors of the pairwise latent effects. In equation 8 where \mathbf{Y} denotes class labels, \mathbf{X} contains the covariates (e.g complex network connectivity), the effect of the latent variables is contained in $U_1 \dots U_{N_v}$, the probabilities for each vertex pair are represented by p_{ij} . The process of generating the eigenmodels is validated by 10-fold cross-validation.

$$\begin{aligned} \text{DrugScore}_i = & \text{protein-interactions}_{ij} U_{\{11\}} + \\ & \text{pathways}_{ij} U_{\{22\}} + \\ & \text{GO-similarity}_{ij} U_{\{33\}} + \\ & \text{Chem-similarity}_{ij} U_{\{44\}} + \\ & \text{DO-similarity}_{ij} U_{\{55\}} \end{aligned} \quad (11)$$

The diagonal of the matrix $\widehat{\mathbf{U}} \widehat{\mathbf{U}}^T$ contains the eigenvalues which we use as coefficients for the latent variables, these act as weights for the importance of the latent variables and are multiplied by the score index calculated from the Jaccard matrix. The candidates are then reevaluated using these scores and re-ranked according to their potential for repurposing.

3.10. Hardware and software platform

We implemented the system using the R language with the RStudio programming environment, on an Intel Xenon 64-bit CPU, using dual processors (3.2GHz) with six cores, and 128 GB of RAM. R is very extendable using packages written by other researchers [42]. We used the following R packages: GOSim [56], DOSim [55], ChemmineR[4], eigenmodel [23]. Our R code and data files are freely available to all researchers on GitHub for download: <https://github.com/kennmcgarry/DrugSideEffects>

4. Results

The first stage is to obtain the UMLS code for our disease of interest (Alzheimer's), the code (C0002395) is used to automatically search drugbank for the list of drugs used

to treat Alzheimer’s disease, these are presented in table 1. We removed DB07701 as it is a new drug and lacks important information required for our method such as an Anatomical Therapeutic Chemical (ATC) code, commercial drug name and lacks literature based evidence. The remaining list of eight conventional treatments are used to search SIDER4 and returns all the side-effects associated with each drug. We then combine the side-effects jointly for the drugs.

Table 1: List of current treatments for Alzheimer’s and known side-effects. DB07701 is an experimental drug and is excluded from our experiments.

	Treatment	Side-effects
1	Citalopram	545
2	Galantamine	50
3	Risperidone	402
4	Donepezil	152
5	Rivastigmine	230
6	Memantine	208
7	Escitalopram	545
8	Aripiprazole	544
9	DB07701	152

Figure 4 shows a Venn diagram illustrating the common side-effects, because of graphical limitations only five of the nine Alzheimer’s drugs can be shown. The diagram indicates 10 common side-effects, however, using all nine drugs gives only eight common side-effects. The decision was made to tackle the problem by using only the side-effects common to all eight drugs which meant that eight side-effects were available for this study. These shared side-effects were used to predict similar interactions which provided a list of 77 drugs, the criteria being a 50% cut-off point of shared side-effects. We only examined the first 25 drugs with the highest ranks. Table 2 lists all eight side-effects.

We decided to use the overlaps between the nine drugs as this allowed for 8 side-effects to be available for study, this simple approach proved to be quite powerful. Initial work on the side-effects and known-targets can be found in [37].

Table 2: List of shared side-effects common to all nine drugs

	side-effect
1	Apathy
2	Aphasia
3	Flat affect
4	Hypokinesia
5	Libido increased
6	Muscle contractions involuntary
7	Paranoia
8	Transient ischaemic attack

The side-effects are typically what would be expected of drugs targeting the central nervous system. The drugs are used to either to slow the progress of dementia or to alleviate the worst symptoms of depression and anger. Algorithm 1, produced the following list of drugs shown in table 3, we have displayed the top 25 drugs based on side-effect similarity of greater than 50%. According to the

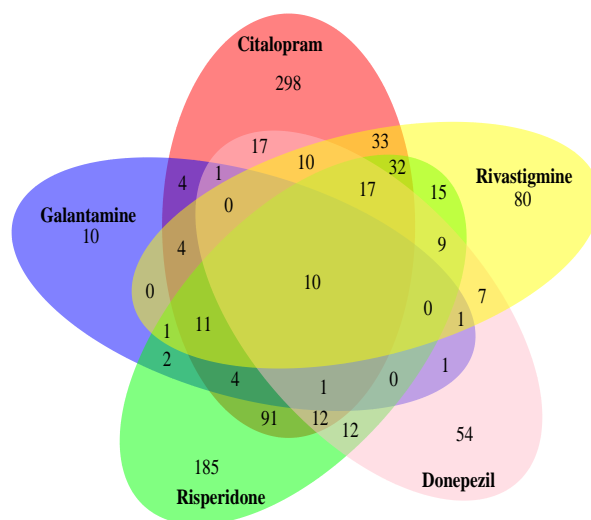


Figure 4: Venn diagram showing side-effects (shared and unique) between five of the nine main Alzheimer’s drugs. 10 side-effects are common to all five drugs, using nine drugs we obtain eight side-effects.

Anatomical Therapeutic Chemical (ATC) codes, the majority of the drugs are ‘N’, that is to say they are intended to target problems relating to the Nervous system.

Some of the drugs have been used to treat the symptoms of Alzheimer’s disease such as depression, anxiety etc. Some have actually been used with a view to halt the degenerative process. The candidate drugs at the lower end of our similarity scale such as Tramadol, an analgesic drug, affects the peripheral and central nervous system [39]. Tramadol, an opioid is mainly used to treat pain, in terms of Alzheimer’s, patients are advised to avoid it, as it appears to worsen the condition. This is as it has anti-cholinergic side effects. It causes serious side effects like serotonin syndrome, this is due to the influx of the neurotransmitter serotonin and then it inhibits the reuptake of serotonin, which can cause toxic levels. This could be a possible reason for tramadol worsening the condition of Alzheimer sufferers. Side effects like confusion and agitation worsen for Alzheimer patients on tramadol, the effects seem to be severe in the elderly [47].

Zolpidem actually appears to be implicated with causing dementia when underlying diseases, such as hypertension and diabetes after controlling for potential confounders, such as age, sex, coronary artery disease, diabetes and anti-hypertension drugs are taken into account. In table 4 for 10 of our candidate drugs we have presented for each drug up to five of the original diseases and symptoms they were designed to treat. On average each drug was used to treat 21 diseases/indications, the minimum was one disease and the maximum value was 88 diseases.

4.1. Analysis of the chemical structures and protein targets

A computational model was built using the chemical data from drugbank, this indicates the connectivity pat-

Table 3: Drugs with > 50% similarity. Where * may actually cause/exacerbate dementia related problems in some cases.

	Drugname	NoSideEffects	Similarity(%)	drugbankid	atc_codes	Repositioned?
1	Ropinirole	418.00	100.00	DB00268	N04BC04	Y
2	Bupropion	288.00	87.50	DB01156	N06AX12	Y
3	Pramipexole	398.00	87.50	DB00413	N04BC05	Y
4	Quetiapine	188.00	87.50	DB01224	N05AH04	Y
5	Selegiline	216.00	87.50	DB01037	N04BD01	Y
6	Sertraline	208.00	87.50	DB01104	N06AB06	Y
7	Topiramate	301.00	87.50	DB00273	N03AX11	Y
8	Venlafaxine	338.00	87.50	DB00285	N06AX16	Y
9	Gabapentin	272.00	75.00	DB00996	N03AX12	Y
10	Lamotrigine	152.00	75.00	DB00555	N03AX09	Y
11	Mirtazapine	153.00	75.00	DB00370	N06AX11	Y*
12	Oxcarbazepine	123.00	75.00	DB00776	N03AF02	Y
13	Pregabalin	561.00	75.00	DB00230	N03AX16	N*
14	Cevimeline	241.00	62.50	DB00185	N07AX03	N
15	Clomipramine	221.00	62.50	DB01242	N06AA04	Y
16	Fluvoxamine	221.00	62.50	DB00176	N06AB08	Y
17	Mycophenolate	262.00	62.50	DB00688	L04AA06	N
18	Paroxetine	362.00	62.50	DB00715	N06AB05	N*
19	Pergolide	156.00	62.50	DB01186	N04BC02	N
20	Rasagiline	131.00	62.50	DB01367	N04BD02	Y
21	Riluzole	196.00	62.50	DB00740	N07XX02	Y
22	Tiagabine	102.00	62.50	DB00906	N03AG06	N
23	Tolcapone	101.00	62.50	DB00323	N04BX01	N
24	Tramadol	373.00	62.50	DB00193	N02AX02	N
25	Zolpidem	129.00	62.50	DB00425	N05CF02	Y*

Table 4: Top 10 candidate drugs and the problems/symptoms they were designed to treat

Drug	Disease/indications
Ropinirole	Hypotension, Muscle rigidity, Parkinson’s disease, Restless legs syndrome, Tachycardia
Bupropion	Mental disorder, Depression, Dysthymic disorder, Fatigue, Asthenia
Pramipexole	Parkinsonism, Parkinson’s disease, Restless legs syndrome
Quetiapine	Mental disorder, Bipolar disorder, Bipolar I disorder, Psychotic disorder, Depression
Selegiline	Fatigue, Asthenia, Feeling guilty, Major depression, Parkinson’s disease
Sertraline	Agoraphobia, Anger, Anxiety, Depression, Dysthymic disorder
Topiramate	Bipolar disorder, Cluster headache, Migraine, Epilepsy, Partial seizures
Venlafaxine	Agoraphobia, Anxiety, Chest pain, Depersonalisation, Depression
Gabapentin	Ataxia, Diabetes mellitus, Dizziness, Epilepsy, Erythema multiforme, Fatigue
Lamotrigine	Bipolar disorder, Dysmenorrhoea, Epilepsy, Grand mal convulsion, Parkinson’s disease

terns linking drug to target receptors. We downloaded the relevant structure data files (SDF) for the 77 drugs (69 candidate drugs and 8 current drugs) using their drugbank identifiers, The atom-pairs are converted into molecular fingerprints and these are clustered.

We created a chemical fingerprint for each of the drugs, we chose 2048 atom-pairs although it is possible to use a structure containing 4096 most common atom pairs in DrugBank. The pairwise distances are calculated for all 77 entries (69 candidate drugs and 8 current drugs) between the given fingerprints and then fit a Beta distribution to the resulting Tanimoto scores, conditioned on the number of set bits in each fingerprint. The highest matching drug is similarity is Pergolide with a 35% commonality. The majority of drugs have between 32% and 25% chemical similarity. Since these all these candidate drugs had similar side-effects and since their chemical composition is quite different, we can conclude that chemical similarity alone should not be used to seek out drugs for re-purposing.

Using hierarchical clustering on the drug similarity matrix provides a little bit more information. Through a process trial and error we deduced that there are 10 clusters. The validity of clusters in terms of composition is more or less confirmed by the similarity matrix. However, analysis of the chemical similarity matrix does not provide a full

explanation and further integration of data is needed.

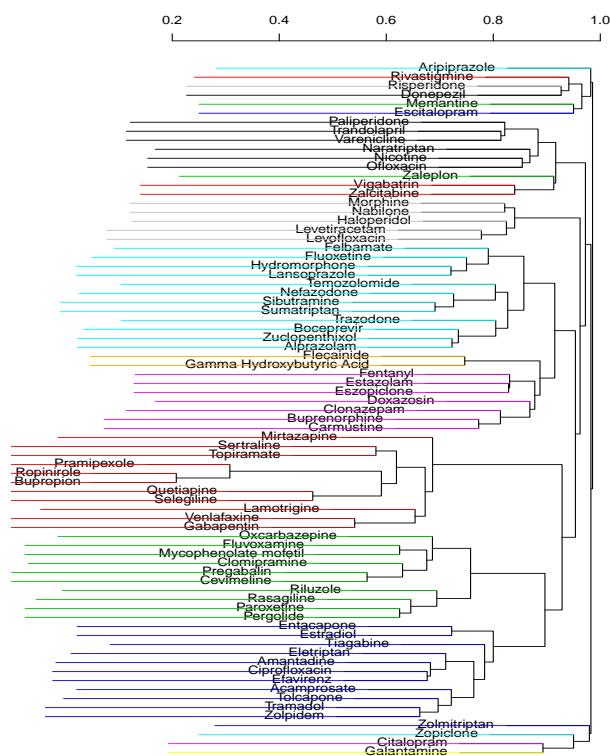
The silhouette plot in Figure 5b indicates that the majority of the drug clusters are reasonably good in terms of quality of fit. Values near one (unity) indicates that the observation is well placed in its cluster; values near zero show that it’s likely that an observation should belong to some other cluster.

- 0.71-1.0 - A strong structure has been found
- 0.51-0.70 - A reasonable structure has been found
- 0.26-0.50 - The structure is weak and could be artificial
- < 0.25 No substantial structure has been found

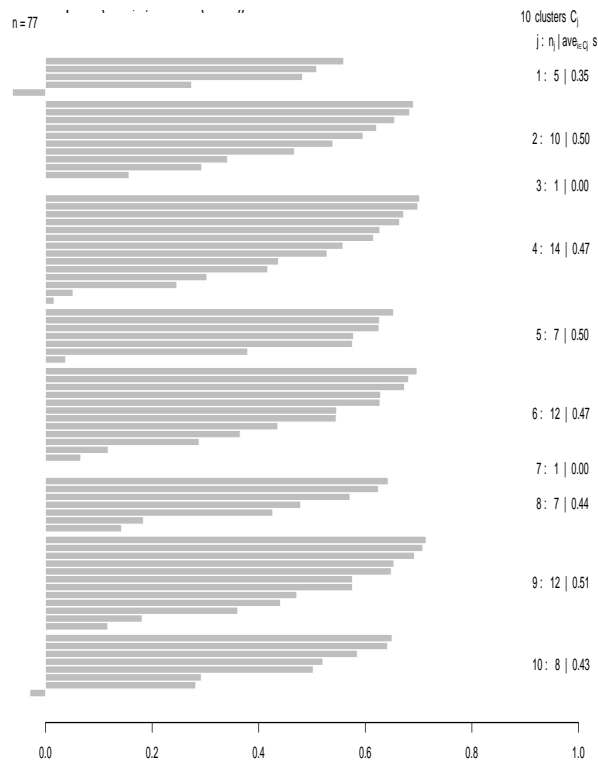
The chemical information must be collated with higher level biological knowledge and known protein interactions for a deeper and more accurate indication of the usefulness of the candidate drugs. However, we examine the next set of results from a drug centric viewpoint.

4.2. Complex network analysis of drug to protein connectivity patterns and biological pathways

Drug to protein target interaction patterns represent important biological functions and signaling mechanisms.



(a) Dendrogram of clusters



(b) Silhouette plot determining goodness of fit of the clusters.

Figure 5: Clustering the 69 candidate drugs and eight conventional treatment drugs ($n=77$), 10 clusters are identified based on the chemical similarity matrix. The silhouette plot indicates goodness of fit

The drugs designed to combat diseases are targeted at specific proteins but unfortunately also affect off-targets and create side-effects. Furthermore, proteins tend to operate together in modules that perform specific functions and that proteins can belong to several modules. It is therefore essential to build up a complete profile of the drugs and their interaction partners.

Diagrams are of limited use when displaying connectivity patterns of complex networks, even the small networks can easily become incomprehensible and visually are of limited use. The network statistics based on the values of hubness, closeness and betweenness provide a deeper understanding of the connectivity patterns than a visual assessment. The protein targets and known protein interactions of the 25 drugs were identified, downloaded from the relevant database (STITCH) and complex network analysis was applied to the subsequent network structures. We created 25 separate protein-to-drug interaction networks and calculated the relevant statistics for each and then combined the 25 networks into one large interaction network and recalculated the statistics.

The statistics for the top 25 drug networks is shown in table 5, every network has individual values for each protein (not shown) in terms of hubness and betweenness. The number of edges, number nodes, the modularity, and average path will be identical for each protein in a given network. It should be stated that these networks are both

off and on-targets and the data is from STITCH. In Figure 8 we display the drug network patterns, where drugs are denoted by triangle and proteins by circles. Protein targets colored in yellow are common to two or more drugs. The protein targets that have common drugs implies they affect common pathways, this will be the reason for the common or shared side-effects.

The pathways presented in table 6 show 20 of the 50 pathways identified, these play a role in a number of important cellular and metabolic processes, when this pathway annotation is integrated with the other drug-to-protein networks we are able to discern a number of overlapping functions and processes. Observing the enrichment process we can see that proteins implicated with the same disease (and related diseases) are more predisposed to interact with each other rather than interact with other proteins. These proteins also have common GO terms and are likely to be located in the same tissues. Proximity, in protein-protein networks as borne out by the network statistics is an important factor. Disease genes appear to high degree but low clustering coefficients.

Although drugs act at the level of protein/enzyme interactions, their overall effect is to modify the cells processes or signaling mechanisms that comprise important biological pathways. A biological pathway can be described as a collection of ordered, discrete actions among molecules that can breakdown proteins or assemble them

Table 5: Complex network statistics for the 25 candidate drug networks (individually), sorted by betweenness

	modularity	avepath	nedges	nverts	transit	degree	diameter	clos	between	dens	hubness
pramipexole	0.245	1.9	72.0	23	0.6	2.000	3.000	0.022	0.000	0.285	0.126
quetiapine	0.055	1.4	131.0	21	0.7	2.000	3.000	0.023	0.000	0.624	0.109
bupropion	0.231	1.8	72.0	21	0.7	11.000	3.000	0.034	10.725	0.343	0.406
pregabalin	0.030	1.5	13.0	8	0.6	7.000	2.000	0.143	13.333	0.464	1.000
rasagiline	0.114	1.5	33.0	12	0.6	11.000	2.000	0.091	17.817	0.500	1.000
tolcapone	0.079	1.6	14.0	9	0.4	8.000	2.000	0.125	19.000	0.389	1.000
cevimeline	0.086	1.8	46.0	15	0.7	6.000	4.000	0.045	27.179	0.438	0.511
tiagabine	0.166	1.8	13.0	11	0.2	10.000	2.000	0.100	40.500	0.236	1.000
oxcarbazepine	0.221	1.7	24.0	14	0.3	13.000	2.000	0.077	61.500	0.264	1.000
tramadol	0.257	1.8	21.0	14	0.3	13.000	2.000	0.077	68.000	0.231	1.000
zolpidem	0.035	1.4	160.0	24	1.0	23.000	2.000	0.043	116.000	0.580	1.000
selegiline	0.220	1.9	152.0	36	0.4	26.000	4.000	0.023	190.501	0.241	1.000
ropinirole	0.361	1.7	77.0	25	0.6	24.000	2.000	0.042	209.833	0.257	1.000
paroxetine	0.319	1.8	127.0	33	0.5	32.000	2.000	0.031	301.926	0.241	1.000
mycophenolate	0.424	1.8	88.0	30	0.5	29.000	2.000	0.034	325.793	0.202	1.000
pergolide	0.358	1.8	123.0	33	0.6	32.000	2.000	0.031	352.400	0.233	1.000
riluzole	0.333	1.8	72.0	31	0.3	30.000	2.000	0.033	357.500	0.155	1.000
gabapentin	0.367	1.8	78.0	32	0.3	31.000	2.000	0.032	387.317	0.157	1.000
topiramate	0.367	1.9	68.0	34	0.2	33.000	2.000	0.030	472.843	0.121	1.000
lamotrigine	0.427	1.8	127.0	40	0.4	39.000	2.000	0.026	590.368	0.163	1.000
fluvoxamine	0.363	1.9	81.0	39	0.2	38.000	2.000	0.026	622.633	0.109	1.000
venlafaxine	0.230	1.9	82.0	39	0.2	38.000	2.000	0.026	627.945	0.111	1.000
clomipramin	0.396	1.8	183.0	49	0.4	48.000	2.000	0.021	879.129	0.156	1.000
sertraline	0.356	1.9	153.0	49	0.4	48.000	2.000	0.021	912.696	0.130	1.000
mirtazapine	0.442	1.8	214.0	51	0.5	50.000	2.000	0.020	943.046	0.168	1.000

Table 6: Shared pathways based on common proteins targeted drugs, all p-values associated with gene/ratio are statistically significant at <0.05

ID	Description	GeneRatio
HSA-375280	Amine ligand-binding receptors	29/170
HSA-112315	Transmission across Chemical Synapses	39/170
HSA-211859	Biological oxidations	38/170
HSA-112316	Neuronal System	43/170
HSA-211981	Xenobiotics	16/170
HSA-211945	Phase 1 - Functionalization of compounds	25/170
HSA-112314	Neurotransmitter Receptor Transmission	26/170
HSA-373076	Class A/1 (Rhodopsin-like receptors)	35/170
HSA-211897	Cytochrome P450 - arranged by substrate	19/170
HSA-975298	Ligand-gated ion channel transport	13/170
HSA-390666	Serotonin receptors	10/170
HSA-629594	Highly calcium permeable receptors	10/170
HSA-500792	GPCR ligand binding	37/170
HSA-425407	SLC-mediated transmembrane transport	29/170
HSA-181431	Acetylcholine Binding Events	10/170

in the cell. Such a pathway can trigger the assembly of new molecules by turning genes on and off. There are several types of pathways that control metabolism, signaling cascades, transport and otherwise generally control the behavior and actions of the cell in response to stimuli. The normal healthy body requires the coordination of many biological pathways, the majority of diseases involve the malfunction of proteins that cooperate in pathways. In fact 30% of drugs operate on one type of protein called G-protein-coupled receptors (GPCRS).

4.3. Disease similarity analysis using the gene and disease ontology

For each of the 77 drugs we note the disease/symptoms each one is targeted against and produce a correlation matrix from the terms annotating each disease from the disease ontology. Observing the correlation values (using Wang similarity measure) we can see the terms are reasonably similar between Parkinson’s and Alzheimer’s disease. In Figure 7 the main types of diseases treated by the 77

drugs are shown. The similarity measure is constrained to lie between 0 and 1, with 1 being the highest level of similarity. Ignoring the diagonal, the highest measure is between psychosis and schizophrenia (0.64) and between Parkinson’s and Alzheimer’s (0.44), depression and psychosis (0.39), followed by epilepsy and Alzheimer’s (0.35). Estimation of correlational strength is based on Cohen’s measures where very high 0.9 - 1.0; high 0.7 - 0.9; moderate 0.5 - 0.7 ; low 0.3 - 0.5.

Overall, we find 48 DO terms used to describe the diseases in a hierarchy. Through extensive cross-mapping of DO terms to standard clinical and medical terminologies such as the Medical Subject Headings (MeSH), the International Classification of Diseases (ICD), Systematized Nomenclature of Medicine (SNOMED) and Online Mendelian Inheritance in Man (OMIM), can reflect the current knowledge of human diseases and their associations with phenotype, environment and genetics. In table 7 the 48 terms assigned to the 10 diseases are shown, in addition to exact matches the Wang measure also assesses the semantic similarity.

The protein networks can be further enhanced by integrating known biological knowledge from gene ontology [2, 15]. This information illuminates biological pathways, signaling mechanisms, genes, associated diseases that the proteins are involved in and indicates where the drug of interest is active. The information is important since the drug will often interact with off-targets, identifying these interactions will often lead to an explanation of why a drug has particular side-effects.

In table 8 a sample of GO terms assigned to the proteins affected by the candidate drugs is presented. The terms are as we would expect given they relate mainly to the central nervous systems and neuronal functioning. The proteins are a mixture of planned targets and off-targets, that is to say the drugs have inadvertently targeted some

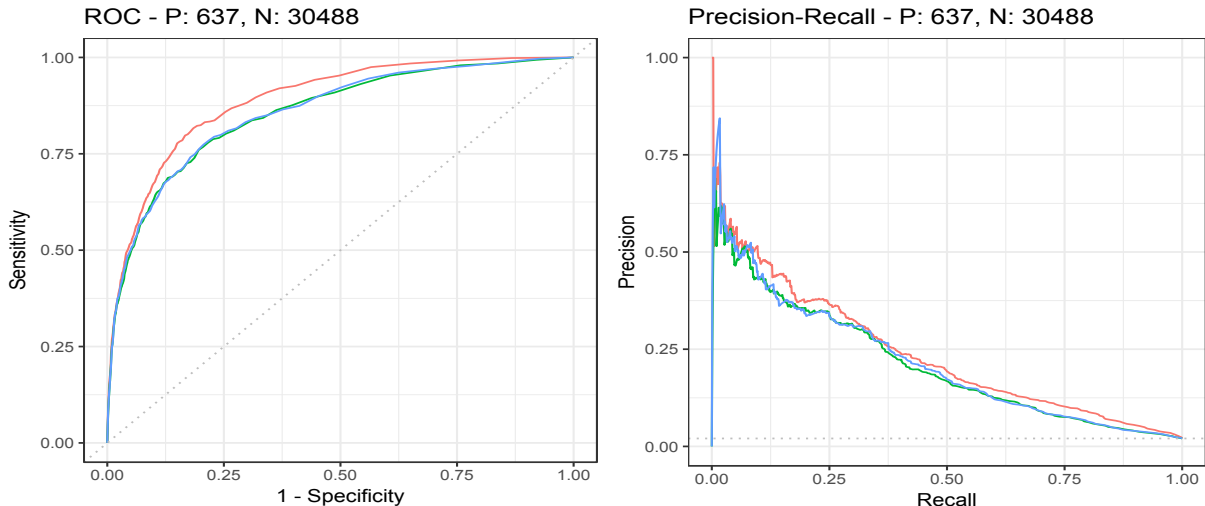


Figure 6: The ROC and PR curves for the goodness of fit of the complex latent network model. The red curve represents the combined protein-interactions+ontology+pathways. The blue curve is the protein-interactions + pathways, the green curve is the protein-interactions alone

proteins. This is part of the reason why side-effects occur.

4.4. Data integration, model accuracy and drug scoring

The sources of data are merged in a principled way using the matrix factorization scheme described in equation 7. Based on the linear equation the sources of data were each allocated a coefficient (similar to regression) and a matrix was produced for each drug and latent variable in the network. The diagonal contains the combined eigenvalue for each drug, the Jaccard index was calculated for variable also and used as a weight or coefficient. This was used to reevaluate the initial list of candidate drugs as determined by side-effect similarity alone. Referring to table 9, the initial ranking (table 3) determined by side-effect similarity alone has changed somewhat. Ropinirole is still the top scoring drug but the positions of the majority of the drugs has changed and the last three drugs were not on the original list of 25 having moved up the list.

As a measure of confidence, we validated our findings from the latent network [25]. In the left hand-side of Figure 6 we plot the ROC curves for the latent network. We use eigenmodels with 5-fold cross-validation using MCMC to simulate the posterior distributions. We take the typical values of 10,000 for burn-in and 1,000 for inferencing. This allows us to model the effects of using: 1. protein and drug interactions alone, 2. protein interactions + pathways, 3. protein interactions + pathways + level 2 ontology. The Area under the Curve (AUC) for the three models are 75%, 88% and 93% respectively. Increasing the amount of information provides the complex network with greater, internal robustness.

Generally, all complex networks can be validated like classifiers either between networks (comparing different networks) or within a single network (predicting internal

connections) [25]. The latent network is validated by classifying the connections as belonging to individual nodes in the graph. Several subsets are sampled through 5-fold cross-validation, at each step the non-sampled section of the network is predicted and class accuracy determined and displayed using receiver operating characteristic curves (ROC) and precision-recall (PR) curves. The ROC is based on evaluating the tradeoffs between specificity and sensitivity. Specificity is the probability of predicting that a link exists between two nodes given the correct situation is indeed a link, whereas sensitivity is the probability of predicting no link exists between two nodes given the true state is indeed no such link exists.

In Figure 6 the precision-recall (PR) curves are presented, these diagrams indicate the precision values for corresponding sensitivity (recall) values. The PR plot presents a global evaluation of the network model. Should a high area under the curve exist, this represents both high recall and high precision. High precision implies a low false positive rate, and high recall suggests a low false negative rate. Large values for both criteria, indicate that the classifier is providing good results (high precision), as well as returning a majority of all positive results (high recall).

4.5. Comparison with other drug repositioning techniques

In table 10 we compare our results with three competing methods from the literature, Zhang [59], Wang [49] and Gottlieb [19]. With the exception of Wang, all the authors have provided a ranking.

Our highest ranking drug is Ropinirole and does not appear in the other systems. Selegiline appears as the top drug for Zhang’s system and is ranked as 4th top drug for our system. The highest ranked drug for the Wang

Table 7: 48 DO terms assigned to the diseases

1	degenerative disease	25	syndrome
2	rheumatism	26	disease
3	psychotic disorder	27	delirium, dementia, cognitive disorder
4	Movement disorder	28	cerebral degeneration
5	organic brain syndrome	29	organic psychosis
6	central nervous system disease	30	Tauopathies
7	neurodegenerative disorder	31	dementia
8	Parkinsonian disorder	32	endogenous depression
9	body system disease	33	episodic mood disorder
10	musculoskeletal system disease	34	mental depression
11	nervous system disease	35	Autonomic nervous system disorder
12	peripheral nervous system disease	36	peptic ulcer
13	hereditary degenerative disease of central nervous system	37	gastrointestinal system disease
14	neuromuscular disease	38	disease by infectious agent
15	neuropathy	39	opportunistic mycosis
16	disease of biological process	40	lentivirus infectious disease
17	disease of behavior	41	viral infectious disease
18	disease of anatomical entity	42	AIDS-related opportunistic infectious disease
19	Muscle, ligament and fascia disorder	43	RNA virus infectious disease
20	myopathy	44	HIV infectious disease
21	soft tissue disease	45	retroviridae infectious disease
22	tissue disease	46	fungal infectious disease
23	brain disease	47	reproductive system disease
24	basal ganglia disease	48	female reproductive system disorder

Alzheimers	1	0.44	0.12	0.06	0.12	0.06	0.08	0.35	0.35	0.45	0.35
Parkinsons	0.44	1	0.12	0.06	0.12	0.06	0.08	0.35	0.35	0.45	0.35
Mental disorder	0.12	0.12	1	0.66	0.22	0.66	0.79	0.13	0.13	0.15	0.13
Depression	0.06	0.06	0.66	1	0.15	0.64	0.49	0.07	0.07	0.09	0.07
Fatigue	0.12	0.12	0.22	0.15	1	0.15	0.18	0.13	0.13	0.15	0.13
Bipolar disorder	0.06	0.06	0.66	0.64	0.15	1	0.49	0.07	0.07	0.09	0.07
Anxiety	0.08	0.08	0.79	0.49	0.18	0.49	1	0.09	0.09	0.11	0.09
Migraine	0.35	0.35	0.13	0.07	0.13	0.07	0.09	1	0.66	0.53	0.66
Epilepsy	0.35	0.35	0.13	0.07	0.13	0.07	0.09	0.66	1	0.53	0.66
Restless legs	0.45	0.45	0.15	0.09	0.15	0.09	0.11	0.53	0.53	1	0.53
Hypotension	0.35	0.35	0.13	0.07	0.13	0.07	0.09	0.66	0.66	0.53	1
Alzheimers											
Parkinsons											
Mental disorder											
Depression											
Fatigue											
Bipolar disorder											
Anxiety											
Migraine											
Epilepsy											
Restless legs											
Hypotension											

Figure 7: Disease ontology similarity matrix, the correlations are based ‘off the diagonal’ and take on values between 0 and 1.

system is Lorazepam, which does not appear in any of the other systems. The top ranked drug for Gottlieb system is Amantadine, this is ranked by our system as 34th and by Zhang’s system as 3rd ranked. Interestingly, Carbidopa was filtered out of our system since we did not have access to its ATC code - Zhang’s system places this drug as ranked 2nd and Gottlieb’s system places it as 8th. However, upon replacing it back in our database and re-running the tests (ignoring chemical structure and lack of ATC code) Carbidopa is picked up by our system with

Table 8: Example of GO Biological Process terms related to the on-target proteins for all candidate drugs. All p-values associated with gene/ratio are statistically significant at <0.05

GO: ID	Description	GeneRatio
375280	Amine ligand-binding receptors	30/156
211981	Xenobiotics	15/156
112315	Transmission across Chemical Synapses	36/156
112316	Neuronal System	39/156
211945	Functionalization of compounds	23/156
211859	Biological oxidations	29/156
211897	Cytochrome P450 - arranged by substrate type	19/156
112314	Neurotransmitter Receptor Binding	26/156
373076	Class A/1 (Rhodopsin-like receptors)	36/156
975298	Ligand-gated ion channel transport	13/156

a raw side-effect similarity score of 28%. The only other occurrence of drugs appearing in our system as Wang’s is Pergolide. An earlier version of our system had a greater similarity with Wang’s candidate drugs, however when we filtered out drugs with many common side-effects, and focused on fewer, more specific side-effects, Olanzapine and other drugs disappeared from our list.

4.6. Analyzing other diseases

In order to be assured of the robustness of our approach we tackle several other diseases using our system. Referring to the related work we analyze three other diseases that have been investigated for the potential to re-purpose current drugs. These results are displayed in table 11.

We examined Rheumatoid Arthritis and compared our results with Zhang [58], who obtained different results to us but we both had identified Imatinib as a potential candidate that had received attention from a clinical trial. The next disease was Systemic Lupus Erythematosus, which we compared with Zhang’s other work [59]. Our third scoring drug is in fact a class of drugs (Conjugated Estrogens) of which Leflunomide is an example identified by Zhang, this appears in our lists but with a similarity of 33% and would not technically be considered. We compared the results of

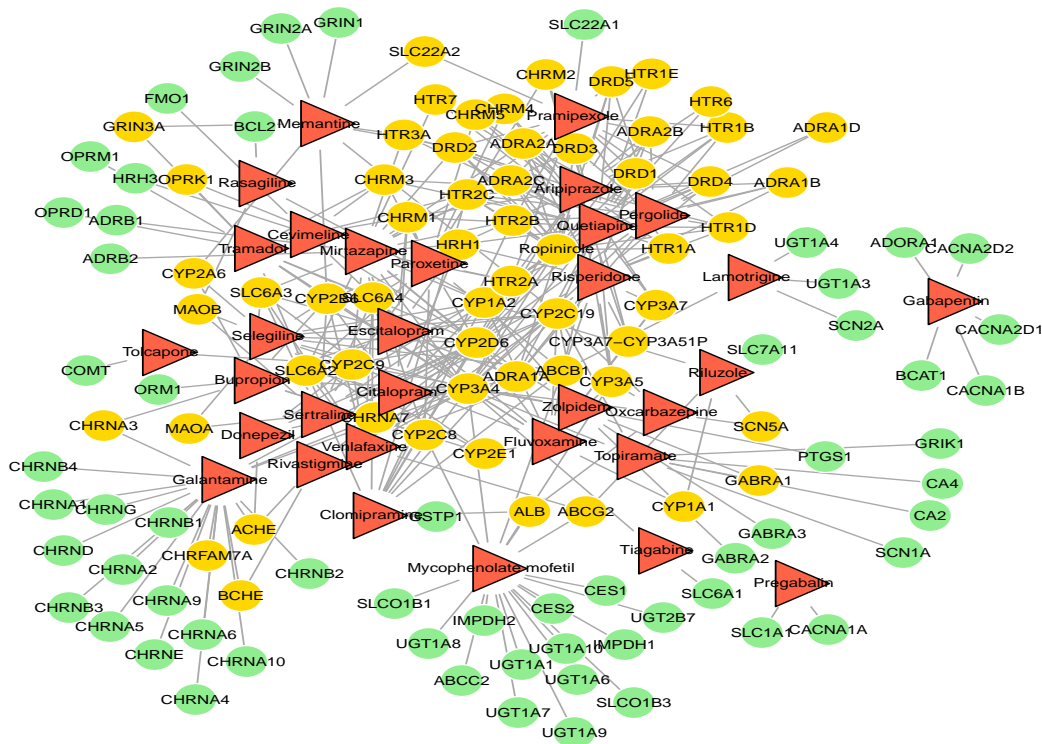


Figure 8: Drug network view of proteins and diseases, drugs are denoted by triangle and proteins by circles. Protein targets colored in yellow are common to two or more drugs

Table 9: Final ranked list of candidate drugs for Alzheimer’s disease

drugbank_id	name	atc_codes	score	
1	DB00268	Ropinirole	N04BC04	0.71
2	DB00413	Pramipexole	N04BC05	0.70
3	DB00273	Topiramate	N03AX11	0.70
4	DB01037	Selegiline	N04BD01	0.70
5	DB01156	Bupropion	N06AX12	0.68
6	DB00285	Venlafaxine	N06AX16	0.67
7	DB01104	Sertraline	N06AB06	0.64
8	DB01224	Quetiapine	N05AH04	0.56
9	DB00230	Pregabalin	N03AX16	0.56
10	DB00776	Oxcarbazepine	N03AF02	0.56
11	DB00555	Lamotrigine	N03AX09	0.45
12	DB00996	Gabapentin	N03AX12	0.45
13	DB00370	Mirtazapine	N06AX11	0.45
14	DB00323	Tolcapone	N04BX01	0.32
15	DB01367	Rasagiline	N04BD02	0.32
16	DB00906	Tiagabine	N03AG06	0.32
17	DB00185	Cevimeline	N07AX03	0.31
18	DB00740	Riluzole	N07XX02	0.31
19	DB00425	Zolpidem	N05CF02	0.28
20	DB00193	Tramadol	N02AX02	0.28
21	DB01186	Pergolide	N04BC02	0.27
22	DB01242	Clomipramine	N06AA04	0.26
23	DB00688	Mycophenolate mofetil	L04AA06	0.25
24	DB00176	Fluvoxamine	N06AB08	0.23
25	DB00715	Paroxetine	N06AB05	0.23

Non Small Cell Lung Cancer (NSCLC) with those of Wang [49]. Virtually all of our top ranked drugs had been the attention of multiple clinical trials. Out of Wang’s top drugs; Cisplatin, Carboplatin, Methotrexate and Temozolomide, only Cisplatin and Temozolomide were identified by our system at 33% similarity.

5. Discussion

Empirically, by using a smaller number of highly specific side-effects we discovered a more valid (based on evaluating the literature) list of drugs for repurposing. There is however, a trade-off between the number of current drugs used in the initial search and the potential for returning either many or very few side-effects. Our system requires diseases with at least two drugs as a certain ‘critical mass’ of common side-effects (for the disease) need to be present.

Furthermore, if the initial side-effect database is not pre-processed by removal of common side-effects then many useless candidate drugs will be returned in the search, as common side-effects such as dizziness and nausea are experienced by almost all drugs and thus are unhelpful. On the other hand, using more than 10-15 drugs in the initial side-effect search is unlikely to lead to side-effects common to all. Our top scoring drug (Ropinirole) is used to treat Parkinson’s, restless legs and shaking conditions. It is a dopamine receptor agonist drug and there is research to suggest it has a neuro-protective effect. We used the research literature to verify if the candidate drugs identified by our system had any potential for repositioning. Several, in addition to Ropinirole have been the subject of clinical trials but the results have been mixed.

We conducted several literature based comparisons between our system and the related methods. The differences uncovered between the various drug re-purposing systems

Table 10: Comparison of various methods for top ranked drugs

McGarry	Zhang(2014)	Wang(2014)	Gottlieb(2011)
Ropinirole (0.81)	Selegiline (0.70)	Lorazepam	Amantadine (0.99)
Pramipexole (0.66)	Carbidopa (0.69)	Alprazolam	Ipratropium bromide(0.97)
Topiramate (0.66)	Amantadine (0.68)	Clonazepam	Divalproex Sodium (0.95)
Selegiline (0.65)	Procyclidine (0.68)	Diazepam	Procyclidine (0.92)
Bupropion (0.64)	Valproic Acid (0.67)	Escitalpram	Scopolamine (0.92)
Venlafaxine (0.60)	Metformin (0.65)	Ziprasidone	Trihexyphenidyl (0.91)
Sertraline (0.58)	Bexarotene (0.64)	Risperidone	Benzatropine (0.90)
Quetiapine (0.57)	Neostigmine (0.63)	Pergolide	Carbidopa (0.88)
Pregabalin (0.56)	Galantamine (0.63)	Olanzapine	Neostigmine (0.88)
Oxcarbazepine (0.56)	Nilvadipine (0.61)	Bromocriptine	Scopolamine (0.86)

Table 11: Other diseases, details of trials from <https://clinicaltrials.gov/>

Disease	Candidate drugs	Reposition evidence
Rheumatoid arthritis (RA) (C0003873)	Alosetron	None
	Cevimeline	None
	Citalopram	NCT01154647 proposed in 2010 for RA
	Escitalopram	None
Systemic Lupus Erythematosus (SLE) (C0024141)	Imatinib	NCT00154336 proposed 2005
	Aripiprazole	None
	Carbamazepine	None
	Conjugated Estrogens	NCT00392093 proposed in 2006 for SLE
	Conjugated Estrogens	NCT00006133 proposed in 2000 for SLE
Non Small Cell Lung Cancer (NSCLC) (C0007131)	Conjugated Estrogens	NCT00000419 proposed in 1999 for SLE
	Bortezomib	NCT00714246 proposed in 2008 for NSCLC
	Carfilzomib	NCT01941316 proposed in 2013 for NSCLC
	Lenalidomid	None
	Pazopanib	NCT00367679 proposed in 2013 for NSCLC
Abiraterone	NCT01884285 proposed in 2013 for NSCLC	

simply reflects the different methods of data integration and types of data used. Although, no methods produced exactly the same list of candidates, their top scoring candidates were drugs that had been the focus of clinical trials. Indeed, the various algorithms must be considered complementary in terms of candidate identification.

The use of ontology’s has provided our system with knowledge at a higher level than protein interactions and chemical structures. The information provided by the ontology’s help reduce the false positive rate (FPR), as seen by the ROC/PR, without this information the FPR tends to increase with unsuitable drugs being identified as potential candidates.

6. Conclusions

Drug repositioning is a viable and useful process to augment and complement the work already accomplished in drug development. We have presented a novel method that can successfully identify useful candidates for drug repositioning by integrating high level knowledge from several ontology’s, this ability is currently lacking in most competing methods. We have validated its predictive ability by cross-validation and have used several disease case studies. For the case study investigating Alzheimer’s disease, the majority of the 25 candidate drugs exhibited the potential to be repurposed as a treatment for Alzheimer’s. A few have been tested clinically and the results confirm they can reduce the level of new plaque formation. About three or four candidate drugs are actually implicated in causing/exacerbating dementia. Clearly, these drugs are targeting the correct pathways/mechanisms but their drug to protein interactions are harmful. Other drugs examined in

research and clinical trials, exhibited the potential to improve behavioral psychosis, agitation and other behavioral symptoms associated with Alzheimer’s.

Some of the drugs we examined are in early research stages or have only been tested in animal models, or have just entered clinical trials. Our methods have a number of limitations we hope to address in future work. For example, the process of detecting common side-effects is achieved through an iterative process of randomly splitting the list of current treatment drugs by 50% if no common side-effects are found. An optimization solution using a search algorithm or meta-heuristic would improve this process. A likely avenue for future work would be to model and predict potential side-effects for new and experimental drugs using additional resources such as the Ontology for Adverse Effects. We also intend to improve the expert system front-end, providing a more conventional explanation facility to describe the joint pathways and shared target proteins. Despite the limitations, the majority of the research studies examined supported our hypothesis of using side-effects to identify candidate drugs for the first stage in drug repurposing.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments that have improved this paper. We would also like to thank Daniel Himmelstein for making his Python code available to parse the SIDER4.1 data.

References

- [1] T. Ashburn and K. B. Thorl. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3:673–683, 2004.
- [2] M. Ashburner. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
- [3] N. Atias and R. Sharan. An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology*, 18(3):207–218, 2011.
- [4] T. Backman, Y. Cao, and T. Girke. ChemmineR: A compound mining framework for R. *Bioinformatics*, 24(15):1733–4, 2008.
- [5] A. Barabasi. *Network Science*. Cambridge University Press, 1st edition, 2016.
- [6] A. Barabasi and Z. Oltvai. Network biology: understanding the cell’s functional organization. *Nat Rev Genet*, 5:101–113, 2004.
- [7] F. Barrenas, S. Chavali, P. Holme, R. Mobini, and M. Benson. Network properties of complex human disease genes identified through genome-wide association studies. *PLoS ONE*, 4(11):e8090, 11 2009.
- [8] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, Kauai, Hawaii, USA, 2002.
- [9] A. Bento, A. Gaulton, A. Hersey, L. Bellis, J. Chambers, M. Davies, F. Kruger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. Overington. The ChEMBL bioactivity database: an update. *Nucleic Acids Research*, 42:1083–1090, 2014.
- [10] K. Chen and J. Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, (5), 2016.
- [11] Xin Chen, , and Charles H. Reynolds. Performance of similarity measures in 2d fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of Chemical Information and Computer Sciences*, 42(6):1407–1414, 2002. PMID: 12444738.
- [12] J. Cheng, L. Yang, V. Kumar, and P. Agarwal. Systematic evaluation of connectivity map for disease indications. *Genome Medicine*, 6(12), 2014.
- [13] A. Chiang and A. Butte. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical pharmacology and therapeutics*, 86(5):507–510, 2009.
- [14] S. Daminelli, J. Haupt, M. Reimann, and M. Schroeder. Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network. *Integrative Biology*, 4:778–788, 2012.
- [15] M. Deng, Z. Tu, F. Sun, and T. Chen. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6):859–902, 2004.
- [16] J. Dudley, D. Tarangini, and A. Butte. Exploiting drug-disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, 12(4):303–311, 2011.
- [17] L. Freeman. Centrality in social networks I: Conceptual clarification. *Social Networks*, 1:215–239, 1979.
- [18] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.
- [19] A. Gottlieb, G. Stein, E. Ruppini, and R. Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(496), 2011.
- [20] E. Guney. Reproducible drug repurposing: When similarity does not suffice. In *Proceedings of the 18th Pacific Symposium on Biocomputing*, pages 132–143, Kohala Coast, Hawaii, USA, 2017. World Scientific.
- [21] D. He, Z. Liu, and L. Chen. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics*, 12, 2011.
- [22] Y. He, S. Sarntivijai, Y. Lin, Z. Xiang, A. Guo, S. Zhang, D. Jagannathan, L. Toldo, C. Tao, and B. Smith. OAE: The ontology of adverse events. *Journal of Biomedical Semantics*, 5(29), 2014.
- [23] P. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):207–218, 2009.
- [24] D. Hoover. Row-columns exchangeability and a generalized model for exchangeability. In G. Koch and F. Spizzichino, editors, *Proceedings of the International Conference on Exchangeability in Probability and Statistics*, pages 281–291, Rome, 1982. North-Holland Publishing.
- [25] E. Kolaczyk and G. Csardi. *Statistical Analysis of Network Data with R*. Springer, 2014.
- [26] M. Kuhn, M. Al Banchaouchi, M. Campillos, L. Jensen, C. Gross, A. Gavin, and P. Bork. Systematic identification of proteins that elicit drug side effects. *Molecular Systems Biology*, 9(1), 2013.
- [27] M. Kuhn, M. Campillos, I. Letunic, L.J.Jensen, and P. Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, 6(343), 2010.
- [28] V. Law, C. Knox, and Y. Djoumbou et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42:D1091–D1097, 2014.
- [29] X. Lei, J. Zhao, H. Fujitab, and A. Zhang. Predicting essential proteins based on rna-seq, subcellular localization and go annotation datasets. *Knowledge Based Systems*, 8(4):136–148, 2018.
- [30] J. Li, B. Gong, X. Chen, T. Liu, G. Wu, F. Zhang, C. Li, X. Li, S. Rao, and X. Li. Dosim: An R package for similarity between diseases based on disease ontology. *BMC Bioinformatics*, 12(1):266, 2011.
- [31] R. Liu, N. Singh, G. Tawa, A. Wallqvist, and J. Reifman. Exploiting large-scale drug-protein interaction information for computational drug repurposing. *BMC Bioinformatics*, 15(1):210, 2014.
- [32] T. Liu and R. Altman. Relating essential proteins to drug side-effects using canonical component analysis: A structure-based approach. *Journal of Chemical Information and Modeling*, 55(7):1483–1494, 2015. PMID: 26121262.
- [33] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics*, pages 1–9, 2018.
- [34] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F. Wu, and Y. Pan. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016.
- [35] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *bioRxiv*, 2017.
- [36] K. McGarry, J. Chambers, and G. Oatley. Graph based analysis of protein interaction for diabetes research. *Artificial Intelligence in Medicine*, 41(2):129–144, 2007.
- [37] K. McGarry, A. Rashid, and H. Smith. Computational methods for drug repositioning. *Drug Target Review*, 3:31–33, 2016.
- [38] K. Michael, D. Szklarczyk, A. Franceschini, C. von Mering, L. Jensen, Lars Juhl, and P. Bork. Stitch 3: zooming in on protein–chemical interactions. *Nucleic Acids Research*, 40(D1):D876–D880, 2012.
- [39] K. Minami, J. Ogata, and Y. Uezono. What is the main mechanism of tramadol? *Naunyn-Schmiedeberg’s Archives of Pharmacology*, 388(10):999–1007, 2015.
- [40] J. Overington, B. ALLazikani, and A. Hopkins. How many drug targets are there? *Nature Reviews Drug Discovery*, 5:993–996, 2006.
- [41] A. Peyvandipour, N. Saberian, A. Shafi, M. Donato, and S. Draghici. A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, pages 1–9, 2018.
- [42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [43] L. Schriml, C. Arze, S. Nadendla, Y. Chang, M. Mazaitis, V. Felix, G. Feng, and W. Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40:D940

– D946, 2012.

- [44] N. Segura, S. Sanchez, E. Garcia-Barriocanal, and M. Prieto. An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the gene ontology. *Knowledge Based Systems*, 24(1):119–133, 2011.
- [45] S. Singhal and J. Mehta. Antitumor activity of thalidomide in refractory multiple myeloma. *New England Journal of Medicine*, 341:1565–1571, 1999.
- [46] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguéz, T. Doerks, M. Stark, J. Muller, P. Bork, L. Jensen, and C. Mering. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568, 2011.
- [47] M. Vazzana, T. Andreani, J. Fanguero, C. Faggio, C. Silva, A. Santini, M.L. Garcia, A.M. Silva, and E.B. Souto. Tramadol hydrochloride: Pharmacokinetics, pharmacodynamics, adverse side effects, co-administration of drugs and new drug delivery systems. *Biomedicine and Pharmacotherapy*, 70:234 – 238, 2015.
- [48] J. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23(10):1274–1281, 2007.
- [49] W. Wang, S. Yang, X. Zhang, and J. Li. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, 30(20):2923–2930, 2014.
- [50] X. Wang, B. Thijssen, and H. Yu. Target essentiality and centrality characterize drug side effects. *PLoS Computational Biology*, 9(7):1–8, 2013.
- [51] D Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–6, 1988.
- [52] J. Yang, Z. Li, X. Fan, and Y. Cheng. Drug disease association and drug-repositioning predictions in complex diseases using causal inference probabilistic matrix factorization. *Journal of Chemical Information and Modeling*, 54(9):2562–2569, 2014.
- [53] H. Ye, Q. Liu, and J. Wei. Construction of drug network based on side effects and its application for drug repositioning. *PLoS ONE*, 9(2), 2014.
- [54] G. Yu and Q. He. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.*, 12:477–479, 2016.
- [55] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wang, and S. Wang. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.
- [56] G. Yu, G. Yan, and Q. He. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2015.
- [57] M. Zhang, G. Schmitt-Ulms, C. Sato, Z. Xi, Y. Zhang, and Y. Zhou. Drug repositioning for alzheimer’s disease based on systematic ‘omics’ data mining. *PLoS ONE*, 11(12), 2016.
- [58] P. Zhang, P. Agarwal, and Z. Obradovic. Computational drug repositioning by ranking and integrating multiple data sources. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 579–594. Springer, 2013.
- [59] P. Zhang, F. Wang, and J. Hu. Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In *AMIA Annu Symp Proc*, pages 1258–1267. AMIA, 2014.
- [60] M. Zitnik and B. Zupan. Data fusion by matrix factorization. *IEEE Transactions on pattern analysis and machine intelligence*, 37(1):41–53, 2015.