Prediction of drug-drug interactions from chemogenomic and gene-gene interactions and analysis of drug-drug interactions

by
AZAT AKHMETOV

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of

the requirements for the degree of

Master of Science

Sabancı University

SPRING, 2013

Prediction of drug-drug interactions from chemogenomic and gene-gene interactions and analysis of drug-drug interactions

Approved by:

Asst. Prof. Murat Çokol (Thesis supervisor)	…….
Prof. O. Uğur Sezerman	…….
Asst. Prof. Erdal Toprak	…….
Prof. Hikmet Budak	…….
Assoc. Prof. Berrin Yanıkoğlu	…….

Date of Approval: …….

© Azat Akhmetov 2013

All Rights Reserved

Prediction of drug-drug interactions from chemogenomic and gene-gene interactions and analysis of drug-drug interactions

Azat Akhmetov

Biological Sciences and Bioengineering, Master's Thesis, 2013

Thesis Supervisor: Asst. Prof. Murat Çokol

Keywords: Chemogenomics, DrugBank, drug interactions, drug interaction prediction, genetic interactions, high throughput screening, biological networks

Abstract

The interactions between multiple drugs administered to an organism concurrently, whether in the form of synergy or antagonism, are of clinical relevance. Moreover, understanding the mechanisms and nature of drug-drug interactions is of great practical and theoretical interest. Work has previously been done on gene-gene and gene-drug interactions, but the prediction and rationalization of drug-drug interactions from this data is not straightforward. We present a strategy for attacking this problem and producing a computational solution. Our approach makes use of published work on large-scale genetic, chemogenomic and drug-drug interactions in order to find compound pairs that are likely to interact synergistically or antagonistically with each other in *S. cerevisiae*. We defined gene sets whose heterozygous deletion confers sensitivity to a drug as 'drug target candidates.' For each drug pair whose interaction is known in *S. cerevisiae*, we found the number of genetic interactions between each drug's 'target candidates.' We examined whether genetic interaction frequency between 'drug target candidates' is different than overall genetic interaction frequency. We attempted to use this as a basis for prediction of drug-drug interactions, and experimentally tested some of the interactions.

Additionally, we have also analyzed the DrugBank database of drug-drug interactions. DrugBank includes data about the interactions of clinically used drugs in human patients, which is supplied in natural language format. We have standardized this data by a process of manual curation, and produced a large dataset of machine-readable human drug-drug interaction data. We also present some analyses performed on this dataset.

İlaç-ilaç etkileşimlerinin kemogenomik ve gen-gen etkileşimlerinden tahmini ve analizi

Biyoloji Bilimleri ve Biyomühendislik, Yüksek Lisans Tezi, 2013

Tez Danışmanı: Yrd. Doç. Murat Çokol

Anahtar Kelimeler: Kemogenomik, DrugBank, ilaç etkileşimleri, ilaç etkileşimi tahmini, gen etkileşimleri, yüksek very taraması, biyolojik ağlar

Özet

Bir organizmaya uygulanan farklı ilaçların beraber kullanıldıklarında gösterdikleri etkileşimler, gerek sinerji gerek antagonizm şeklinde, klinik önem taşımaktadır. Ayrıca, ilaç-ilaç etkileşimlerinin mekanizmasını ve doğasını anlamak büyük hem pratik hem de teorik açıdan önemlidir. Daha önce ilaç-ilaç ve gen-ilaç etkileşimleri üzerinde çalışılmış olsa da, bu verilerden ilaç-ilaç etkileşimlerinin tahmini ve rasyonalizasyonu karmaşık bir problem oluşturmaktadır. Projemizde bu problemin işlemli olarak çözülmesi için bir strateji sunuyoruz. Kullandığımız yöntem yayınlanmış, büyük çaplı genetik, kemogenomic ve ilaç-ilaç etkileşimi verilerini kullanarak *S. cerevisiae*'da sinerjistik ya da antagonistik etkileşimler yapma ihtimali yüksek olan bileşik çiftleri bulmaktadır. İlaçlar için, heterozigot delesyonları ilaca duyarlılık yaratan genleri 'hedef adayları' olarak tanımladık. İlaç çiftlerinin 'hedef adayları' arasındaki genetik etkileşimlerin sıklığına bakarak, bunu genel genetik etkileşim sıklığı ile karşılaştırdık. Bunu temel alarak bilinmeyen ilaç çiftlerinin etkileşimini tahmin etmeye çalıştık ve bazı tahminleri deneysel olarak kontrol ettik.

Ayrıca, DrugBank veritabanında bulunan ilaç-ilaç etkileşimlerini analize ettik. DrugBank insanlarda klinik olarak kullanılan ilaçların etkileşimlerinin doğal dil olarak içermektedir. Bu verileri elle kategorize ve standardize ederek bilgisayarla okunabilen bir formata çevirdik. Bu verilerin de analizini sunuyoruz.

to my parents

Acknowledgements

I express my deepest and sincere gratitude to my graduate advisor Murat Çokol for his mentorship, support and guidance for me throughout my research and my studies. I feel very fortunate to have had a chance to be his student and be mentored by him.

As well, I would like to thank the other members of my thesis committee for taking the time to review my thesis: Erdal Toprak, Uğur Sezerman, Hikmet Budak and Berrin Yanıkoğlu.

I am grateful to Arda Durmaz, Kaan Yılancıoğlu, Melike Çokol and Murodzhon Akhmetov for their cooperation and help.

I would like to acknowledge The Scientific and Technological Research Council of Turkey (TUBITAK) and the Faculty of Science and Engineering for providing the funds which financed my graduate education.

Last but not least, I would like to thank my parents Gulshat and Nail for their continued support, love and patience.

Contents

List of Tables

List of Figures

Chapter 1: Prediction of Drug-Drug Interactions

1. Introduction and Background

This section provides an overview of the various sources of data used in this research project, as well as the species of yeast with which experiments were conducted.

1.1. *Saccharomyces cerevisiae*

Saccharomyces cerevisiae, also known as baker's yeast or budding yeast, is a well-known species of yeast which has been used in various industries (such as brewing and baking) for centuries. It is also one of the most commonly used model organisms in biology.

It is a unicellular eukaryotic organism, which has a doubling time of 1-2 hours [**1**] [**2**]. It can be easily cultured in the laboratory with standard culture media such as YPD [**3**]. This makes it a very convenient organism for studying drug interactions, making it a popular choice in many studies.

1.2. Chemogenomics

The field of chemogenomics encompasses the study of chemicals and their effect on an organism in the context of its genome. With chemogenomics, the emphasis is placed on considering the response of the genome as a whole in order to understand the effects of a chemical [**4**]. Among other things, chemogenomics offers the promise of resolving the targets of drugs which cannot be adequately studied with conventional approaches due to their action involving the participation of many genes in the genome [**5**].

In particular, an important study by Hillenmeyer and colleagues in 2008 bears relevance for our research. In this study, a large variety of conditions (which included many small molecules in addition to nutrient deficient media and different temperature levels) is applied to yeast (*S. cerevisiae*) deletion libraries in order to generate chemogenomic profiles, which quantify the role of the genes in the yeast genome.

The Hillenmeyer study employs two kinds of yeast deletion libraries for separate sets of experiments. There is a set of 5985 heterozygous deletion strains and another set of

4770 homozygous deletion strains. Every deletion strains correspond to yeast lineage with a single open reading frame (ORF) removed from its genome. The strains are identified by the gene which was deleted.

In either case, in order to conduct the chemogenomic profiling experiments, all of the deletion strains (heterozygous or homozygous) yeast strains in a library (heterozygous or homozygous) are pooled together, and grown for 24 hours in culture medium under the condition being tested (e.g. in presence of a drug). A control experiment is performed by growing the pool of strains in a normal yeast culture medium without imposing additional conditions.

All of the deletion strains were constructed in a specific manner, such that a unique, specific 20 base pair sequence of DNA is inserted into the genome of each strain during the deletion process. This short sequence serves as a barcode that can be used to identify individual strains [**4**] [**6**] [**7**] [**8**]. When the experimental condition is applied, the growth rate and fitness of each individual strain is altered in some way, presumably relating to the particular gene deleted from it. At the end of the 24 hour culture period, the relative populations of different deletion mutants can be compared to the control experiment, in order to determine which strains have benefited from the condition and which have become less fit [**4**]. Statistical analysis is performed on numbers of deletion strains under each condition and the end result is a series of z-scores, indicating how the fitness of the strain has been perturbed in that particular condition in multiples of the sample standard deviation. [**4**] There are certain important differences in the interpretation of results for experiments with homozygous and heterozygous deletion libraries.

With homozygous deletion libraries, the assay is referred to as "homozygous profiling" (HOP). The gene in each strain is removed entirely, so that the gene is completely non-functional in the mutant [**9**]. The consequence is that if a condition, or commonly a drug, exerts its influence on the yeast cell by interacting with the product of this gene in some fashion, then it will no longer be able to exert this influence on the deletion mutant, which lacks this target gene. As such, genes which correspond to strains that have become enriched after growing in a condition are thought to be targeted by that condition [**4**]. Ordinarily, the yeast cannot survive the complete removal of any gene. 19% of all yeast genes lead to lethality when removed, another 15% lead to a serious growth defect – the so called essential genes [**10**] [**11**] [**12**]. The Hillenmeyer study itself demonstrated

that for 97% of genes at least one environment exists which makes them essential, even if they are not essential in the standard culture conditions [**4**]. This is part of the reason why the homozygous deletion library is smaller than heterozygous.

With heterozygous deletion libraries, the assay is called "haploinsufficiency profiling" (HOP). In this case, only one of the two copies of a gene is deleted to construct the deletion mutant strain. Therefore, the gene is still active, but may be active at a lower activity level [**9**]. In this case, it is thought that any condition which affects the yeast by impairing the function of the deleted gene will be more effective, since the strain has been sensitized to it by virtue of a lower gene product level [**4**]. Deletion mutants which are most strongly affected by the condition relative to control will therefore correspond to genes most critical for the response to that condition.

1.3. Gene-Gene Interactions

As said above, 19% of the genes in the yeast genome are necessary for survival in standard culture conditions. The remaining genes, when deleted singly, do not preclude the yeast's survival. But when two genes are deleted in the same strain, various effects such as growth defects can emerge, which are referred to as genetic interactions [**13**].

The BioGRID is an online database which stores and makes available physical (protein) and genetic interactions from different species [**14**], including *S. cerevisiae*. Yeast genetic interactions account for the largest group of interactions listed, making up 74% of all genetic interactions [**14**].

A 2010 study by Costanzo and colleagues [**15**] is worth discussing here. The Costanzo study is the largest contributor of interactions to the BioGRID *S. cerevisiae* genetic interaction database, it currently makes up 34% of these genetic interactions. As part of this study, a synthetic genetic array [**8**] was used to create a large number of double mutants (which have deletions in two genes); in total 5.4 million pairs of genes were probed.

Once constructed, these double mutants are cultured and the colony size is tracked [**16**]. After analysis, the data quantifies how fitness has been affected by combining two gene deletions in a single strain. Costanzo and colleagues present a scoring system for measuring the strength of a genetic interaction in terms of the fitness perturbation. These

scores are also available on the BioGRID, however, not all BioGRID interactions have a score since some contributing studies are qualitative in nature.

1.4. Drug Interactions

By analogy to epistatic interactions, it is possible to speak of the interactions between individual drugs [**17**]. Pairs of drugs used concurrently may be thought of as behaving antagonistically, synergistically depending on how much stronger or weaker the outcome is compared to what would be expected from considering the individual effects of each drug. The case where the outcome coincides with the expectation of observing a sum of individual effects, the drugs are considered to not interact, and are said to be independent or additive.

In principle, it is possible to speak of drug effectiveness in terms of any number of given phenotypic changes, not all of which are practical or expedient to measure. Commonly, in order to measure drug effectiveness, growth rate is used. A study which gathered detailed data on mRNA levels during a number of drug interaction experiments, and concluded that about 70% of the variation is explained by a parameter corresponding to growth [**18**]. Therefore, growth appears to be the single most important factor with regard to drug combination treatments; moreover, growth also has immediate clinical relevance.

While it is generally understood that drug antagonism represents a decrease of the effect, and synergy represents an increase, the exact details of the definition of interaction are consequent on the method of measurement used to detect it.

Bliss independence [**19**] is one of the simpler methods, although it is widely used [**20**] [**21**]. With Bliss independence, the inhibition caused by individual drugs is measured as a fraction of the negative control. If after using drug A, r_a of the cells continue to grow, and after using drug B, r_b of the cells continue to grow, then it is expected that when these drugs are used together is $r_a . r_b$ of the cells will continue to grow (a smaller number than either r_a or r_b since typically these are between 0 and 1).

An issue with Bliss independence is that it does not take into account the concentrations of the drugs. Even in absence of any interaction, a nonlinear increase in inhibition may take place, simply because the act of combining the drugs has also increased drug dosage [**17**].

Another method, Loewe additivity [**22**], alleviates the issue of changing concentrations. With Loewe, the baseline of independence is defined in terms of self-self interactions: A drug is assumed to not interact with itself. Therefore, if $[A]$ and $[B]$ are the concentrations of drugs A and B respectively, this pair of drugs is independent if $x.[A] + (1 - x)[B]$ remains constant for different values of x. In other words, on a two dimensional representation of the drug gradients, straight lines of constant inhibitions connect equivalent (in multiples of MIC, the minimum inhibitory concentration) drug dosages. Outcomes above or below this baseline are taken as synergistic or antagonistic. This manifests itself as a "bending" of the isoboles outward or inward towards the origin point (representing the control case where no drug is added). A recent study by Cokol and colleagues [**23**] has implemented a version of the Loewe model in order to quantify drug-drug interactions in a number of experiments.

2. Motivation and Contribution of the Thesis

Our original intention was to ascertain whether it is possible to predict the interactions between pairs of drugs in the yeast *Saccharomyces cerevisiae* based on information regarding the effects of the individual drugs and the known data regarding the functional genomics of this organism. More specifically, we have attempted to utilize two datasets:

- The chemogenomic profile dataset published by Hillenmeyer and colleagues as accompaniment to their 2008 study [**4**], which is primarily used in order to prepare lists of "candidate targets", which are genes that appear to have been important in how the organism responds to a drug.
- The database of genetic interaction data provided by BioGRID.

In addition to these, a set of data with drug pair whose interaction status was previously tested and thus is known, has been used for purposes of calibration and testing of our algorithms.

The first of our inputs, namely the lists of candidate drug targets, is fundamentally easier to generate quickly, as for n drugs there need to be n experiments. The second input of genetic interaction data differs in this regard, in that for n genes, a number of individual tests on the order of n^2 must be performed. However, efficient methods exist for rapid, high throughput querying of genetic interactions. Moreover, the number of genes in an organism tends to be limited (*Saccharomyces cerevisiae* has close to 6000 ORFs [**24**]) further bounding the problem size. And lastly, at this time, a non-trivial fraction of this space has already been explored: BioGRID lists close to 198000 genetic interactions.

Neither of these applies to the task of mapping out drug-drug interactions. The problem of finding the interactions between n drugs scales with n^2, and the individual experiments are far more laborious and error-prone than genetic interaction experiments; dosage of the drug also becomes a more significant issue as opposed to the binary nature of gene knockout studies. The number of chemicals that exists is either unbounded or at least very large, and very few of their combinations (especially in a comparative sense) have been tested. Of those studies which have tested them, substantial heterogeneity exists to the point where data from different studies may be incompatible.

Therefore, the ability of somehow inferring *in silico* whether a given pair drugs will interact based on these other more readily available datasets represents a substantial gain in terms of work required to obtain the valuable data of drug combination effectiveness.

3. Methods and Materials

3.1. Candidate Drug Target Assignment

In order to calculate correspondence scores, it is necessary to have a list of so called "candidate targets" for each drug. Phenomenologically speaking, the candidate targets of a drug are those genes whose deletion strains resulted in the most statistically significant perturbation in survival during competitive growth experiments under the influence of that drug. The implicit assumption our method carries is that such genes, which when absent greatly alter the susceptibility of the yeast cell to the drug, are most important for the action of the drug.

We have defined the methods of assignment of such candidate drugs targets on a drug-by-drug basis. In other words, from one chemogenomic profile corresponding to one drug, it is possible to play any of several methods which we describe below, to extract in each case a set of genes which are the candidate targets. Provided a set of such targets is available for either drug in a pair, these sets are later used to calculate the correspondence score of the pair.

The methods we defined were as follows:

- Rank method: Top n strains were taken as targets for each profile, with the same n for all drugs.
- Cutoff method: Strains above or below a given z-score threshold were taken as targets for each profile, with constant threshold across drugs.
- Inflection method: An attempt is made to determine the point where the distribution of genes z-scores changes, and z-scores above this point are taken as targets.
- Best replicate method: Similar to the cutoff method, but instead of combining the replicates with Stouffer's method, the replicates were combined by taking the maximum or minimum score recorded for each given strain and drug.
- HOP method: For each drug, the targets were assigned to be those genes for which the genetic interaction vector has a Spearman correlation coefficient [**25**] of at least c vs. the discretized chemogenomic profile.

Of these, the first three operate with combined heterozygous (HIP) profiles. The fourth method, as noted, operates with non-combined HIP profiles. The last method operates on combined homozygous (HOP) profiles.

For each of the five methods we attempted, there was one free variable which behaved similar to a threshold, and had to be optimized. This is referred to in this text as the "parameter" of that method. The selection of an appropriate value for this parameter is described elsewhere, in 3.5.1. Calibration of algorithm parameters. The different methods are summarized in Table 3.

3.1.1. Fitness score combination

In the Hillenmeyer dataset, it was often the case that conditions had been repeated, because more than one experiment was conducted for each condition. For our purposes, we required exactly one chemogenomic profile for each unique condition, so for those conditions where more than one profile existed, we combined all the replicates to obtain a consensus profile.

Since the Hillenmeyer dataset provides both p-value and z-score profiles for each experiment, it is possible to combine each of those. For combining n p-values p_i, a method published by Fisher [**26**] can be used:

$$\chi^2 = -2\sum_i ln(p_i)$$

Which gives a chi-square statistic corresponding to the combined p-value, with $2n$ degrees of freedom.

For the combination of n z-scores z_i, a method was presented by Stouffer [**27**]:

$$Z = \frac{\sum_i z_i}{\sqrt{n}}$$

Which gives the combined Z-score for the set of replicates.

3.2. Construction of a Gene-Gene Interaction Network

I order to construct a gene-gene interaction matrix; we have used the database of *S. cerevisiae* genetic interactions in the BioGRID [**14**]. The database contains a mixture of

physical (protein-protein) and genetic interactions. The latter were the group that was pertinent to our purposes.

Both physical and genetic interactions in the BioGRID are further labeled by the class of experiment conducted to detect them. Among genetic interactions, not all classes were equally well represented. Therefore, we designated certain classes as our accepted positive and negative genetic interactions. The basis for selection was mostly prevalence – we ignored classes which were very few in number. The classes of negative and positive genetic interactions included are given in Table 1.

Negative	**Positive**	**Not included**
Negative Genetic	Dosage Rescue	Dosage growth defect
Synthetic Growth Defect	Phenotypic Suppression	Dosage lethality
Synthetic Lethality	Positive Genetic	Phenotypic enhancement
	Synthetic Rescue	Synthetic haploinsufficiency

Table 1: Classes of BioGRID genetic interactions selected to include in our analysis (left two columns) and the classes which were excluded (rightmost column).

The BioGRID data identifies gene pairs as ordered pairs, so the format of the data allows for directionality of the individual interactions. Reasoning that genetic interactions are inherently non-directional (in other words, the distinction between an interaction of gene A with gene B vs. the interaction of gene B with gene A was not considered biologically meaningful) and that the data on directionality is more to do with the experimental setup rather than the underlying biology, we discarded the information regarding this directionality in the gene-gene interaction network we constructed.

Lastly, the list of *S. cerevisiae* genes is subject to revision as new data becomes available. Therefore, some new genes may be added or some old ones may be removed as time passes. In our case, since we intended to combine the gene-gene interaction matrix with the target lists obtained from the chemogenomic data of Hillenmeyer et al., a further question of compatibility arose since not only did the Hillenmeyer study use a list of genes that was valid at the time of the study and since had been revised, but they also did not possess deletion mutant strains for every gene thought to exist in the *S. cere-*

visiae genome at the time. The Hillenmeyer dataset identifies genes primarily by their ORFs, and we have therefore regarded data concerning only those genes the ORFs of which have been mentioned both in the Hillenmeyer and the latest published list of genes from SGD [**24**].

At the end of the procedure, we generated three square binary matrices: One for genetic interactions we took as negative, one for genetic interactions we took as positive, and one with the two matrices combined. For each such matrix M_{ij}, we set $M_{ij} = 1$ if an interaction of gene i with gene j or of gene j with gene i belonging to a relevant class of interactions (listed in Table 1) has been recorded in the BioGRID. Detection of a genetic interaction in a single experiment was therefore sufficient for us to include it in our data. For the matrix containing both negative and positive genetic interactions, the presence of an interaction belonging to a class of either the negative or positive genetic interaction classes from Table 1 was sufficient.

We also ignored self-self interactions, meaning the data where the interaction of gene i with gene i is recorded. The diagonals of our genetic interaction matrices, corresponding to such self-self interactions, were therefore composed solely of zeros.

3.3. Evaluation of Drug-Drug Interaction

We based our method of evaluating drug-drug interaction based on the method described in [**23**]. Briefly, we calculated alpha scores measuring the interaction and then classified scores within a range as drug synergy, drug antagonism, or independence.

For the results of drug-drug interaction experiments conducted in the original Cokol 2011 publication, which served as the training data, the alpha scores were already given in the supplemental information of the paper. For the thresholds, we have used the thresholds originally given in the publication. Specifically, drug interaction experiment outcomes where $\alpha \leq -0.78$ were considered synergistic, experiments where $\alpha \geq 0.68$ were considered antagonistic, and values between these two boundaries ($-0.78 < \alpha < 0.68$) were considered independent [**23**].

For our own experiments, we used the same method of evaluating the isobole length, but adapted it to work with 4-by-4 matrices (where the Cokol 2011 publication used 8-by-8 matrices). As input, the algorithm is given a series of optical density measurements over time for each well in the experiment, with different wells corresponding to differ-

ent drug concentration combinations. Each well's measurements are then collapsed into a single score (by getting the area under the growth curve) and the matrix of growth scores is normalized. On the normalized matrix, the amount of the "bend" of the longest contour line is quantified with a logit function and returned as the α score. The α score is expected to be below 0 for synergistic interactions (where the contour lines bend towards the well with no drug applied) and above 0 (where the contour lines bend in the other direction). Values close to 0, indicating independence, are expected when the contour line is a straight line connecting the two corners of the matrix. For details, see Algorithm 1.

3.4. Calculation of Gene-to-Drug Correspondence Scores

Our intention was to determine whether it is possible to infer drug-drug interactions based on examination of the (a) "candidate drug target" lists derived from chemogenomics data, and (b) a gene-gene interaction network.

We also possessed a set of data for some drug-pairs, which were experimentally obtained from a previous study. These may be regarded as the desired outputs of our own computational work, and as such, can be used for verification of the method ultimately employed.

Thus, in order to link the two input datasets together, we have developed a computational approach which considers drugs in pairs, and calculates a numerical metric which can be taken as a quantification of the likelihood that these two drugs will interact. In this text, for the sake of convenience, this metric will be referred to as the "gene to drug correspondence score".

Algorithm 1: Alpha score calculation [**23**]

Input:

- M: A 4-by-4-by-96 matrix M_{ijk} where each element shows the kth optical density reading of the yeast culture growing in the ith well of the jth row on the multi-well culture plate, with $M_{1,1,k}$ representing the measurements from the well where no drug was applied.

Output: A number (α) quantifying the interaction between the two drugs assayed in this experiment.

1. Create a new matrix $g_{ij} \coloneqq \left(\sum_k M_{ijk}\right) - \min(\{M_{ijk} | 1 \leq k \leq 96\})$ to represent the growth scores.
2. Create a new matrix $n_{ij} \coloneqq g_{ij}/g_{1,1}$ to represent the normalized growth.
3. Find the number $c \coloneqq \max(\{n_{1,4}, n_{4,1}\})$ representing the larger of the two corners of the matrix where the maximum amount of the single drug was applied.
4. Define a sequence $s_i \coloneqq c + 0.01 * i$.
5. Obtain the contour line for each s_i where $s_i \leq 1$ and $i \geq 0$, represented as a series of line segments.
6. Select the longest contiguous line, generating two vectors x and y describing the respective coordinates of each point along this line.
7. Create a new vector $a_i \coloneqq \log(\frac{y_i}{1-y_i}) - \log\left(\frac{x_i}{1-x_i}\right)$.
8. Return $\alpha \coloneqq median(a_i)$.

In order to calculate this correspondence score, we began with the initial assumption that the number of genetic interactions between genes targeted by drugs which themselves interact must be higher. The basis of this approach consists of the following: Both a genetic knockout (which participates in genetic interaction events) and direct or indirect chemical inhibition of a gene product must have the same effect, the blocking of gene action. In the first case, the gene action is blocked because there is no gene to exert its effect, and in the second, the gene action is blocked because the gene product is chemically prevented from exerting its function. Indeed, the possibility of such a "correspondence" between the effects of gene knockouts and drugs has attracted the attention of researchers [**28**] [**29**] [**30**] [**31**]. In our case, we hypothesized that pairs of drugs which interact must interact because in fact, each drug is blocking actions of genes which in turn themselves interact.

The generation of candidate target lists is described elsewhere (see section 3.1. Candidate Drug Target Assignment), and these lists are taken as inputs by the correspondence calculator algorithm. Given a set of genes for either drug (referred to as v and v in Algorithm 2), as a first step, genes which are in both sets are removed (this amounts to taking the set difference in either direction, producing what is referred to in Algorithm 2 as u' and v'). In our current implementations and given the present data, no gene is considered to interact with itself in any case, but removing shared targets obviates the need for that consideration and produces more salient data.

The remaining targets are thus unique for each drug within that pair. Therefore, by necessity the genetic interaction graph between them can only be a bipartite graph, with each drug's targets confined to their own part. From this, the maximum possible number of edges (representing genetic interaction) can be calculated as the product of the numbers of nodes on either side, or the cardinalities of the two sets. This number, referred to as P, is an upper bound. The actual number of edges can be counted by querying of the gene-gene network that we have produced (described in section 3.2. Construction of a Gene-Gene Interaction Network). This is the number G.

Algorithm 2: Calculation of correspondence score

Inputs:

- u, v: Two lists of candidate targets for two drugs.
- G: A genetic interaction matrix where $G_{ij} = 1$ if genes i and j interact and 0 otherwise.

Output: A correspondence score C showing the prevalence of genetic interactions between the candidate targets of the two drugs.

1. Generate unique target lists u' and v':
    a. $u' \coloneqq u \backslash v$ ($a \backslash b$ denotes the set difference of a and b)
    b. $v' \coloneqq v \backslash u$
2. Calculate the potential interactions $P \coloneqq |u'| * |v'|$.
3. If $P = 0$ return $C = 0$.
4. Initialize $R \coloneqq 0$.
5. For each i from 1 to $|u'|$ do:
    a. For each j from 1 to $|v'|$ do:
        i. $g \coloneqq u'_i$
        ii. $h \coloneqq v'_j$
        iii. If $G_{gh} = 1$ increment R by 1.
6. Return $C \coloneqq R/P$.

An obvious measure, which we have elected to employ, is to compare G and P in terms of their proportion. This proportion, with a provision to account for the undefined case of 0 divided by 0, is called the correspondence score. Following the logic of our initial hypothesis of gene-drug correspondence, we expect that this "correspondence score", which is unique to each drug pair, shall have some sort of relation to the interaction status (whether measured in binary, ternary or continuous variable fashion) of that pair which may be exploited in order to predict the interaction status from the correspondence score alone. The formula we have used was:

$$Correspondence = \begin{cases} G/P, & P \neq 0 \\ 0, & P = 0 \end{cases}$$

The provision for the case of $P = 0$ (which would otherwise generate an undefined correspondence score) notably includes:

- Self-self drug interactions (where a drug's interaction with itself is measured)
- Interactions of different drugs which happen to have identical lists of candidate targets
- Interactions of different drugs where the candidate targets of one drug are a strict superset of the other candidate targets of the other drug

According to our hypothesis, in all of these cases we would not predict any interaction to occur based on our concept of correspondence, further justifying the use of a correspondence score of 0 for such cases.

3.5. Predictions

We were able to calculate correspondence and make predictions from different sets of inputs: With the gene-gene interaction matrix, it is possible to use only negative genetic interactions, only positive genetic interactions, or both. Likewise, it is possible to consider drug interactions in terms of synergy vs. no synergy, antagonism vs. no antagonism, and independence vs. interaction. Furthermore, with the candidate target lists, it is possible to generate the lists of targets for each drug with different methods. A list of the variants we have used is given in Table 2.

However, in each case, the inputs are:

- A binary matrix showing which genes are known to interact with each other
- A set of vectors listing the candidate targets of each drug
- A mapping of drug pair to binary value showing whether the drugs in this pair are taken to interact or not

Though we were able to measure the performance of our prediction method across these different inputs, we only experimentally verified the case of predicting drug synergies based on negative genetic interactions.

Target assignment	**Genetic interactions**	**Drug-drug interactions**
Rank	Negative	Synergy
	Positive	Antagonism
	All	Any interaction
Cutoff	Negative	Synergy
	Positive	Antagonism
	All	Any interaction
Inflection	Negative	Synergy
	Positive	Antagonism
	All	Any interaction
Best replicate	Negative	Synergy
	Positive	Antagonism
	All	Any interaction
HOP correlation	Negative	Synergy
	Positive	Antagonism
	All	Any interaction

Table 2: A list of the 15 validation runs presented in this text, according to which target assignment method, genetic interaction set, and drug-drug interaction set were used.

This section describes in detail how predictions were made and verified. To summarize:

1. We calibrated, or trained, our algorithm on the known drug-drug pairs from the Cokol 2011 dataset, excluding self-self experiments, which served as our training set, in order to fine-tune it so that it was best able to detect the difference between the correspondence scores of interacting and non-interacting drug pairs. It was then possible to generate a list of correspondences for unknown drug pairs, and make predictions of interaction based on these.
2. To check the performance of our algorithm for a given set of inputs, we repeatedly hid a subset (the "validation set") of our known drug-drug interaction data, and attempted to predict them. The performance was quantified with receiver operating characteristic (ROC) curves.
3. To experimentally verify the predictions, we again trained the algorithm on the entire set of known drug-drug interactions we had available, generated correspondence scores for all possible pairings of the 343 conditions included in the Hillenmeyer study. We selected a correspondence threshold based on the ROC curve and predicted the drug pairs above this as interacting and the ones below it as non-interacting. Out of this large number of predictions, we selected 12 drugs and experimentally tested their combinations for interaction.

3.5.1. Calibration of algorithm parameters

Each of our methods for candidate drug target assignment was devised such that there was one crucial free variable, which is referred to as the "parameter" of that method. However, calibration was necessary to ascertain the value of that parameter which provides the best prediction performance. For instance, if candidate drug targets are taken to be n most affected deletion strains for a drug, then there remains the task of finding an appropriate value of n.

To attack the problem of selecting an appropriate parameter for each given method, we defined a range containing the probable values of the parameter, made repeated samplings within this range, and selected the one among them which afforded the most predictive power. The exact ranges we have used are given in Table 3.

Method	Parameter represents	Range
Rank	Number of targets per drug	$5, 10, 15, \ldots 495$
Cutoff	Z-score threshold for targets	$10, 9.9, 9.8, \ldots 1.0$
Inflection	Threshold for differential of z-score across genes	$0, 1.10^{-5}, 2.10^{-5}, \ldots\ 88.10^{-5}$
Best replicate	Z-score threshold for targets, replicates combined by taking the maximum of all experiments	$10, 9.9, 9.8, \ldots 1.0$
HOP correlation	Correlation of chemogenomic profile to the genetic interaction matrix column	$0, 0.001, 0.002, \ldots\ 0.1$

Table 3: Ranges of parameters scanned for every target assignment method.

For each given possible value of the parameter, we first divided our training data into lists of drug pairs which were classed as interacting or non-interacting. We then calculated the correspondence scores for each of these pairs. As our hypothesis relied on the correspondence scores of one class being more likely to be larger than the other, we used the MATLAB implementation of the Wilcoxon Rank Sum test [**32**] [**33**] (equivalent to the Mann-Whitney U test, see Algorithm 3) to calculate a p-value corresponding to the degree to which the correspondence scores of the two classes were distinct. The parameter which resulted in the smallest p-value was then assigned as the appropriate value for that set of inputs and candidate drug target assignment method.

For a flowchart describing the overall process of parameter calibration see Figure 1.

Algorithm 3: Mann-Whitney U test [**34**]

Inputs:

- Sets u and v containing the observations for two variables.

Output: A p-value corresponding to the likelihood that the two variables come from the same statistical distribution.

1. Calculate the U statistic by combining the two sets of observation and count the number of times an observation from u is preceded by an observation from v.
2. Calculate $\mu \coloneqq \frac{|u||v|}{2}$ where $|u|$ and $|v|$ are the number of elements in u and v respectively.
3. Calculate $\sigma \coloneqq \sqrt{\frac{|u||v|(|u|+|v|+1)}{12}}$.
4. Calculate $z \coloneqq \frac{U-\mu}{\sigma}$.
5. Return the p-value corresponding to the z-statistic from a standard normal distribution.

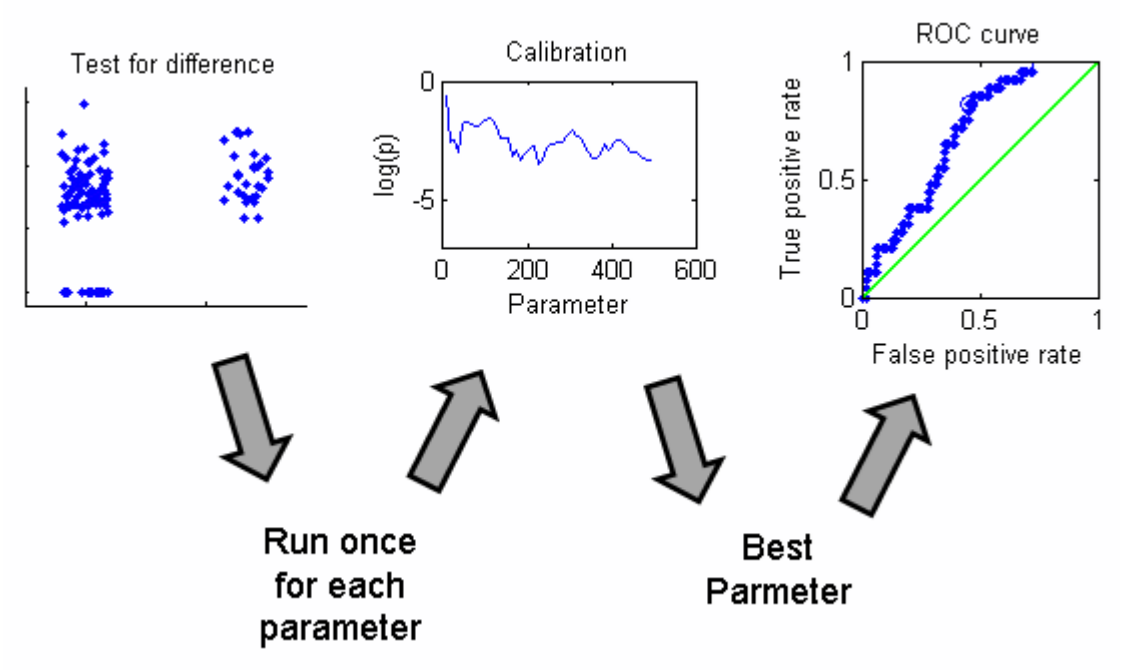

Figure 1: A diagram illustrating the overall process of calibrating the prediction algorithm. The drug pairs in training set are separated into interacting and non-interacting and a Mann-Whitney test is performed to obtain a p-value measuring the likelihood of the two groups being distinct. This is repeated for every value of the parameter that was selected for testing. In the resulting plot of p-values for various values of the parameter, the parameter with the best p-value is chosen as the optimal parameter for that combination of inputs. The predictive power of is further described by an ROC curve.

3.5.2. Validation with training data

In order to obtain a better idea of what the performance of the prediction algorithm is likely to be with a given configuration of inputs, we attempted to create average ROC curves from repeated runs in which the algorithm attempted to predict the interaction status of already known drug pairs.

To accomplish this, we sequestered a randomly selected subset of 30 drug pairs from our training data into the validation set. The algorithm was then trained on the remaining training data, excluding these 30 validation pairs. Therefore the algorithm had no knowledge of these 30 points prior to prediction. We then calculated the correspondences of these 30 pairs, and compared them with the actual interaction status of each pair. We measured the success rate of such a prediction attempt with the area under the ROC curve ("area-under-curve", or AUC – for plotting of AUC curves see Algorithm 4). This process was repeated 20 times, producing 20 ROC curves, producing 20 AUC values.

The average and standard deviation of these were taken as the "average AUC" and "error of the average AUC", respectively. For plotting, the average AUC curve and the error bars were calculated with threshold averaging, using.

3.5.3. Prediction of novel interacting drug pairs

In order to predict the interaction status of unknown pairs, the algorithm was trained on the entire set of drug-drug interaction data we had available from the Cokol 2011 study, excluding self-self experiments. For gene-gene interactions, only the negative genetic interactions were used. For candidate drug target assignment, the top n strains method (with $n = 225$) was used. We classed drug-drug interactions as synergistic vs. non-synergistic.

After calibrating the algorithm, we calculated the correspondence for pairwise combinations of all 343 conditions included in the Hillenmeyer dataset (and for which, therefore, a list of candidate targets could be assigned). In order to actually predict the synergies, we classed correspondence scores above a threshold as synergistic and the remainder as non-synergistic. For the value of this threshold, we chose the correspondence value which corresponded to the point on the ROC curve closest to the top-left corner, which was 0.0072192.

Among the multitude of predictions thus generated, we selected 54 pairs among 12 drugs (not including the self-self pairings, which our algorithm ignored during training and prediction). The drugs were selected (Table 6) after considering different factors such as obtaining a balance of both positive and negative predictions, selecting drugs which were easy to obtain and affordable, and safety of the chemicals.

All the interactions among these 12 drugs including the self-self pairings (which served as controls) were experimentally verified by us.

Algorithm 4: Plotting of the receiver operating characteristic (ROC) curve [**35**]

Inputs:

- s_i: A vector showing the score for the ith event
- r_i: A binary vector showing the actual outcome for the ith event, with 0 showing negative and 1 showing positive outcome

Output: A vector of points (x, y) describing the ROC curve, which can be plotted in two dimensions to produce the ROC plot.

1. Create vector t_i holding the elements of s_i but sorted in ascending order.
2. Generate a sequence q:
    a. For each $j < n$ (where n is the number of elements in s), $q_j \coloneqq \frac{s_j + s_{j+1}}{2}$.
    b. q_0 is set to a number lesser than all elements of s.
    c. q_n is set to a number greater than all elements of s.
3. Calculate the sequences TP_j, FP_j, TN_j, FN_j for each j ($|a|$ indicates number of elements in set a):
    a. $TP_j \coloneqq \left|\{i|s_i > q_j\} \cap \{j|r_i = 1\}\right|$
    b. $FP_j \coloneqq |\{i|s_i > q_j\} \cap \{j|r_i = 0\}|$
    c. $TN_j \coloneqq |\{i|s_i < q_j\} \cap \{j|r_i = 0\}|$
    d. $FN_j \coloneqq |\{i|s_i < q_j\} \cap \{j|r_i = 1\}|$
4. Calculate for each j:
    a. $x_j \coloneqq \frac{FP_j}{FP_j + TN_j}$.
    b. $y_j \coloneqq \frac{TP_j}{TP_j + FN_j}$.
5. Return (x_j, y_j) for every j.

Algorithm 5: Threshold averaging of multiple ROC curves [**35**]

Inputs:

- x_{ij}, y_{ij}: Two matrices containing, respectively, the x and y-coordinates of the ith point on the jth ROC curve.

Output: Two vectors $\bar{x}_i$ and $\bar{y}_i$ containing, respectively, the x and y-coordinates of the average ROC curve as well as two vectors u_i and v_i containing, respectively, the square root of the variance of the x and y-coordinates of point i.

1. For each i:
    a. Set $\bar{x}_i \coloneqq mean(\{x_{ij} | 1 \leq j \leq n\})$.
    b. Set $\bar{y}_i \coloneqq mean(\{y_{ij} | 1 \leq j \leq n\})$.
    c. Set $u_i \coloneqq \sqrt{Var(\{x_{ij} | 1 \leq j \leq n\})}$.
    d. Set $v_i \coloneqq \sqrt{Var(\{y_{ij} | 1 \leq j \leq n\})}$.
2. Return $\bar{x}_i$, $\bar{y}_i$, u_i and v_i.

4. Results and Discussion

4.1. Interaction prediction

4.1.1. Genetic interaction matrix

In the entire BioGRID database (version 3.2.99) there were 322349 total interactions, including physical and genetic. Genetic interactions made up 197634 of these, or 61.3%. We used only the genetic interactions. A breakdown of the count of different interaction listings recorded in the database can be seen in Table 4.

Of these genetic interactions, 147028, or 45.6% of all genetic interactions, fell under the categories we defined as "negative". The largest sub-category of these were interactions specifically marked as "negative genetic" (74.0%), but synthetic growth defect and synthetic lethality interactions were also fairly sized groups (15% and 11% respectively).

The groups we classed as positive genetic interactions amounted to 39828 (12.4% of all genetic interactions) in aggregate. Interactions marked "positive genetic" were more than half of this group (58.6%), with dosage rescue, phenotypic suppression and synthetic rescue making up the rest in roughly similar proportions (12.9%, 14% and 15% respectively).

Our definitions of positive and negative interactions did not cover all classes, so some types of interaction reports were consequently ignored by our analysis. Collectively, we ignored 10578 interactions, or 5.4% of all genetic interactions. This comparatively small group included interaction types which were too few in number or too ambiguous for our purposes.

We observed two kinds of irregularities in the BioGRID database: Firstly, the database is structured such that a query gene and an interaction partner gene are defined for each interaction. For an interaction between genes A and B, potentially there will be another interaction between B and A (due to the manner in which researchers submit their data to the BioGRID, this is may not necessarily be the case, however).

Class	Number	Percent (of total)	Percent (of superset)
Physical interactions	124715	38.7%	38.7%
Genetic interactions	197634	61.3%	61.3%
Positive	39828	12.4%	20.2%
Dosage rescue	5152	1.6%	12.9%
Phenotypic suppression	5587	1.7%	14.0%
Positive genetic	23332	7.2%	58.6%
Synthetic rescue	5957	1.8%	15.0%
Negative	147028	45.6%	74.4%
Negative genetic	108809	33.8%	74.0%
Synthetic growth defect	22046	6.8%	15.0%
Synthetic lethality	16173	5.0%	11.0%
Excluded	10578	3.3%	5.4%
Dosage growth defect	1921	0.6%	18.2%
Dosage lethality	1603	0.5%	15.2%
Phenotypic enhancement	6771	2.1%	64.0%
Synthetic haploinsufficiency	283	0.1%	2.7%
Total	**322349**	**100%**	**-**

Table 4: Counts of genetic interaction entries that were present in the raw BioGRID database for *Saccharomyces cerevisiae*, version 3.2.99.

There were 120509 cases such that a genetic interaction was recorded between A and B, but there was no corresponding interaction recorded between B and A (although in some of these). We assumed that these were sufficient evidence, reasoning that genetic interactions are reciprocal in nature and that there is no reason to expect that an interaction will no longer be observed if the subject and query genes were swapped.

Secondly, we observed 18 genetic interactions where the systematic names of both interacting genes were exactly the same, such that these implied the gene was interacting with itself. These entries did not all come from the same study, but some were submitted by different studies. Our algorithm for constructing the gene-gene interaction matrix

from BioGRID ignored self-self interactions, so these 18 entries were skipped at that stage, but even if that were not the case the correspondence score algorithm would not include self-self genetic interactions in the calculation since it builds lists of unique potential interactors. Because of this, and their very small number, we did not investigate these apparent self-self entries further.

Our final negative genetic interaction matrix contained 101955 interacting pairs between 5025 genes. The positive genetic interaction matrix contained 27048 pairs between the same genes. The combined (positive and negative) genetic interaction matrix contained, as would be expected, their sum of 129003 interacting pairs. These numbers are substantially lower than the numbers of genetic interactions we observed in the BioGRID. Part of the reason is the high amount of redundancy: The counts we gave in Table 4 are numbers of entries in BioGRID. For a given pair of genes (A, B), there could be more than one interaction entry for (A, B) and for (B, A), but all of these would be recorded as a single pair in our constructed matrices, since we are interested in whether any interaction (meeting our criteria) at all has been observed between two genes.

Furthermore, BioGRID references a total of 5782 genes, while we confined our analysis to a list of 5025. Interactions involving 757 genes were therefore present in the BioGRID but were ignored by us since we did not have sufficient data (e.g. chemogenomics) to include them in correspondence calculations.

Matrix	**Interacting pairs**	**Genes**
Negative genetic interactions	101955	5025
Positive genetic interactions	27048	5025
All genetic interactions	129003	5025

Table 5: A breakdown of the number of interacting gene pairs and the number of *S. cerevisiae* genes taken into account by our analyses.

4.1.2. Validation

We have made several validation runs for various combinations of inputs. In every case, 30 of 119 known drug-drug interaction pairs were hidden, and predicted after training on the remaining 89. The process was repeated 20 times.

Each run produced a single calibration curve of p-value vs. parameter, used to select the best parameter. As a representative example, the calibration curves of the negative genetic interaction, rank method, synergy prediction case (which is also the one we examined experimentally) are given in Figure 2. According to these, it appears that a number of sweetspots exist for optimal prediction of synergies: It's possible to obtain good performance if each drug is assigned about 30, 230, 370 and 480 genes as candidate targets. It is also possible that higher numbers could provide good performance, but the task of scanning different parameter values is computationally expensive and we did not query those possibilities.

Each calibration curve therefore reveals which numbers of targets for a drug are appropriate, at least for prediction of drug interactions. The shape of a calibration curve per se does not appear to have any obvious pattern, and it is likely that a multitude of complex factors influences it, since it is ultimately a consequence of the genomic landscape of the organism. However, it is remarkable that although in terms of magnitude there is a fair degree of variation across different training data sets, the overall shape remains fairly robust: The sequence of peaks and valleys is reproduced consistently across different runs, even though a slightly different set of training data is used in every case. From this, it is possible to make the deduction that the calibration curves are specific to the genome (or more precisely the genetic interaction network) of an organism and as such, closely related species which have similar drug interactions should generate similar curves. Unfortunately, we did not have opportunity to test this notion.

Figure 2 gives the calibration curves pertaining to only one configuration of inputs. In total, there were 15 such validation processes, each consisting of 20 runs. Detailed plots for the remainder are not reproduced here for the sake of brevity.

Lastly, if one consults the actual p-values of the curves, it can be observed that in many cases the p-value falls below the p=0.05 mark (or about -1.3 on the base 10 logarithmic scale) which ordinarily indicates statistical significance. However, as the curves are not

adjusted for multiple hypothesis testing, it is likely not correct to draw such a conclusion.

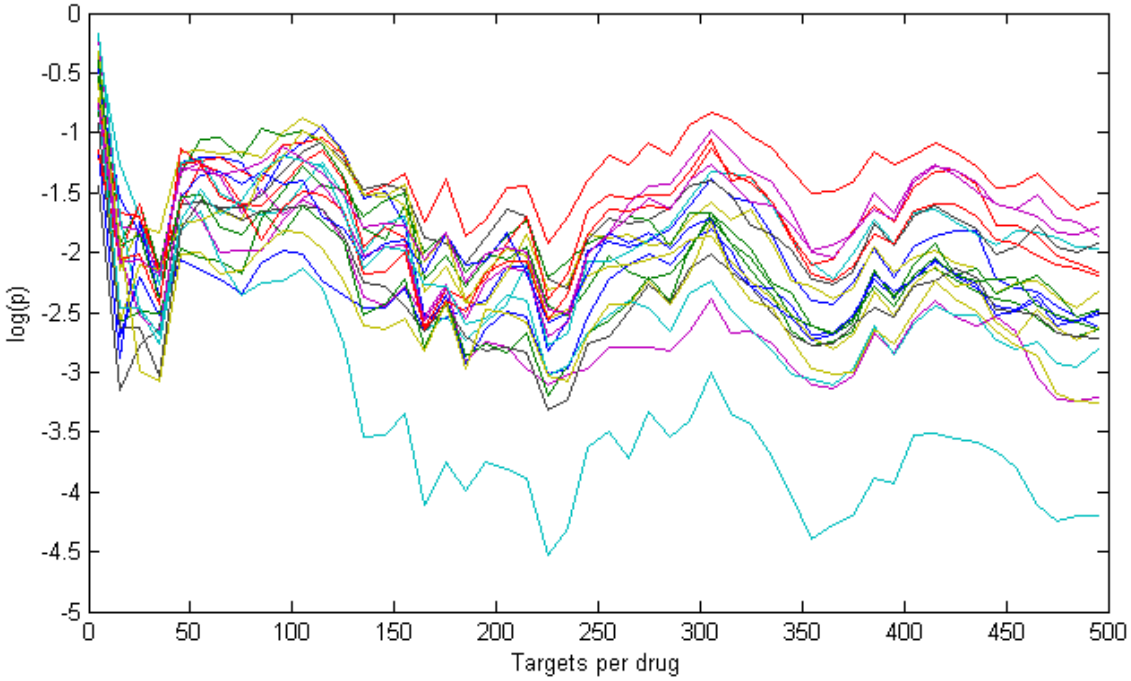

Figure 2: Calibration curves for 20 validation runs, each representing an attempt to train on 87 known pairs. Targets were selected using the rank method, which assigns the top n strains as candidate targets for a drug (x-axis). Negative genetic interactions only were used for correspondence calculations. Every line represents one run. The vertical axis shows the logarithm (base 10) of the Mann-Whitney U test p-value for a difference of distribution between the correspondence scores of synergistic drug pairs vs. non-synergistic pairs.

In each run, the algorithm used the calibration curve to pick the best parameter which optimized the target assignment scheme, and generate an ROC curve for the prediction of the 30 hidden known pairs which represented performance had that optimal target assignment method been used. Since we performed 20 runs, 20 such curves were generated, and they were averaged to obtain an estimate of the overall performance for a given configuration of prediction bases (such as the selection of genetic interactions, and types drug interactions to be predicted, as well as target assignment scheme). A representative ROC curve is given in Figure 3, which is also the ROC curve corresponding to the configuration we tested experimentally.

The ROC curve shows that in general, this particular method produced performance slightly better than a random guess. It also appeared that the performance was comparatively better when a lack of false negatives was valued more than a lack of false positive.

Intuitively, a positive in our setup represents an interaction between to drugs (whether defined as synergy, antagonism, or either) while a negative is the lack of such an interaction. A true positive or true negative would be a drug pair correctly predicted by the algorithm. A false positive would be a pair predicted to interact, which did not turn out to interact in reality. A false negative is a pair which interacts, but was not predicted to interact.

In drug-drug interaction experiments, a crucial difficulty is the large number of potential pairs that need to be tested for interaction, which renders a brute-force approach impractical; every experiment represents non-trivial expenditure of material cost and labor. Therefore, in addition to prediction per se, it is also valuable to be able to simply filter out pairs which are unlikely to interact and thus could be excluded from an experimental interaction screen, increasing interaction hit rate and reducing the per-experiment expense.

In our case, such an "elimination of unlikely pairs" relates to true negative and false negative rates. Specifically, the critical number is the false negative rate, which represents the number of drug pairs the algorithm would wrongly prescribe the exclusion of. Therefore, to an experiment designer interested in eliminating low-value experiments (i.e. drug combinations unlikely to be found to interact even if tested); a low number of false negatives is of particular value. For our algorithm, the optimal tradeoff between false positives and false negatives is a smaller number of the latter, at the expense of the former, suggesting its suitability for such a pre-screening purpose.

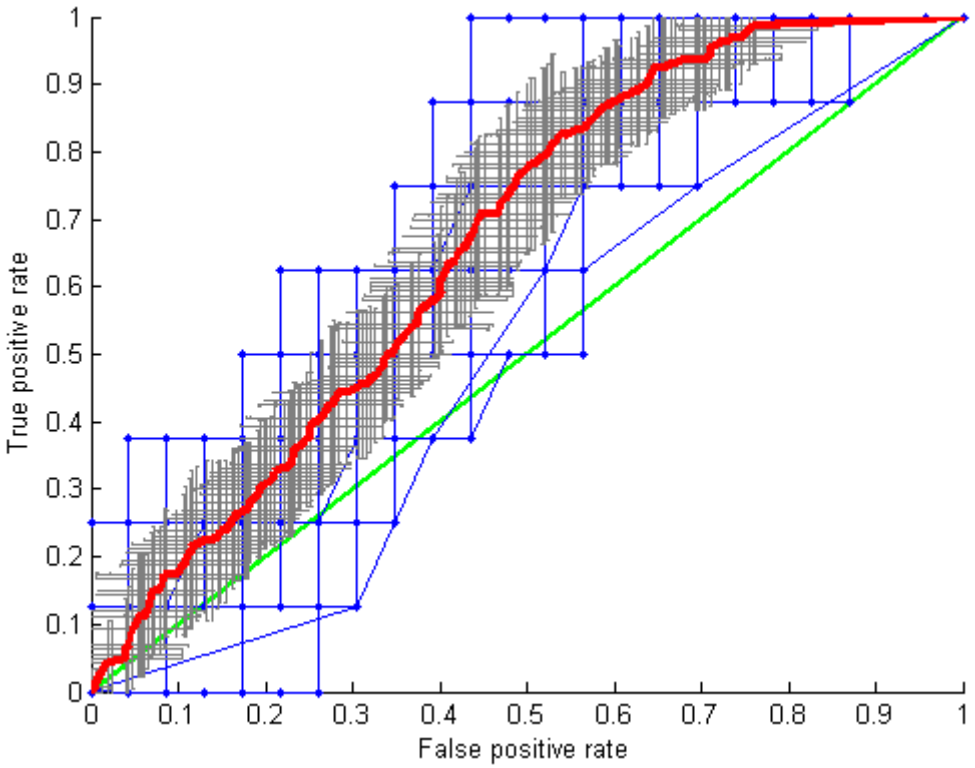

Figure 3: Representative ROC curve, showing the actual average ROC curve for the case where negative genetic interactions were used with the rank method to predict drug synergies. Each blue curve shows the ROC curve of individual validation runs (there are 20 in total but many overlap). The green line shows the random-guess baseline. The red curve tracks the average of all 20 ROC curves, and the grey lines show 1 standard deviation, representing the error, at each point.

In all, we conducted 15 different validations, 3 for every target assignment method: One to predict synergy using negative genetic interactions, one to predict antagonism from positive genetic interactions, and one to predict either kind of interaction from all genetic interactions. The area under curve (AUC) metrics summarizing each such validation are given in Figure 4. As the error bars indicate, in many cases there was not sufficient ground to believe that on average the prediction scheme would fall above the random guess line a significant fraction of the time, even though the mean was above the random guess threshold, however prediction of synergies from negative genetic interactions was the exception to this, possibly due to larger amount, completeness and quality of data for both genetic interactions and drug synergies.

Most often, combining synergies with antagonism and negative genetic interactions with positive reduces predictive power, as would be expected since combining interactions in this manner removes the distinction between positive and negative outcome

from a given interaction. But interestingly, when targets were assigned such that genes were assumed to be targeted by a drug if their deletion produced a similar effect to that of the drug (as measured by correlation between the homozygous deletion chemogenomic profile and the genetic interaction vector of a gene) then combining positive with negative genetic interactions and grouping together synergies with antagonisms (therefore discriminating only interaction vs. no interaction for drug-drug pairs) the mean performance appeared to be better than both synergies with negative interactions and antagonisms with positive genetic interactions.

The HOP method uses data qualitatively different from the haploinsufficiency profiling (HIP) data used by the other 4 methods [**4**]. With HOP experiments, genes are completely deleted from each strain, so the expectation is that strains corresponding to genes targeted by the drug (the outcome of drug action being reduced growth of the strain) will have better fitness in presence of the drug compared to control. The deletion is expected to render a strain resistant to the drug which targets the deleted gene. With HIP experiments, only one copy of the gene is deleted, so it is expected that deletion strains will become more susceptible to the drug and grow less in its presence. This difference could be responsible of the different behavior of our HOP target assignment method compared to the others.

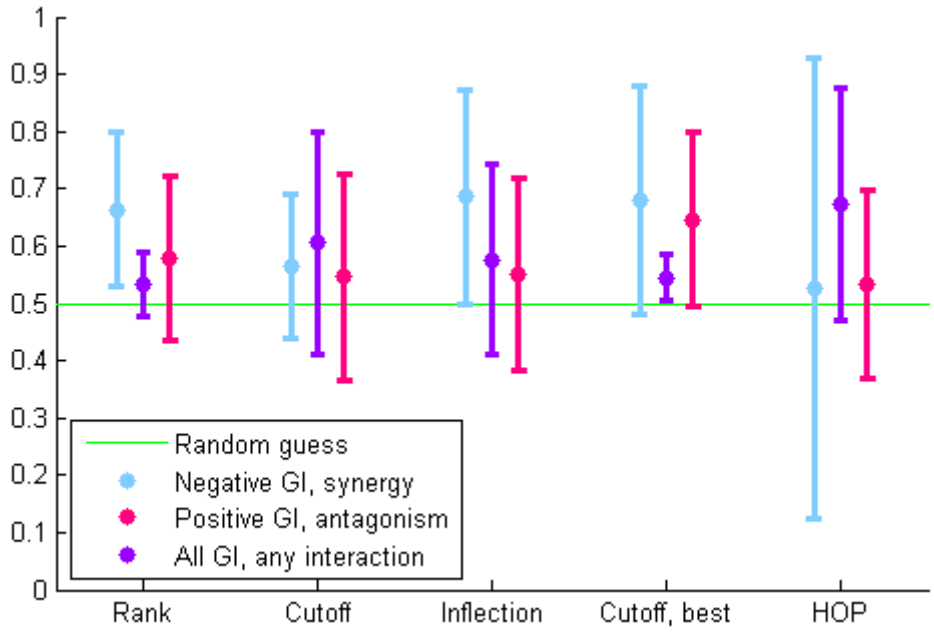

Figure 4: Comparison of prediction performance of various prediction methods as measured by area under curve (AUC) of the ROC curve. Filled circles show mean AUC over 20 validation runs and vertical bars show 95% confidence interval (1.96 standard deviations) for the mean. Cyan bars show the performance when synergy is predicted using negative genetic interactions, purple shows prediction of antagonism from positive genetic interactions, purple shows prediction of all drug-drug interactions from all genetic interactions. Green line shows the AUC for random guessing. Horizontal axis labels show target assignment method: Rank, n targets per drug. Cutoff, targets above determined by a z-score threshold. Inflection, targets determined by second differential of z-score vs. genes. Cutoff, best, similar to cutoff method but replicate experiments are combined by taking the maximum instead of using the Stouffer method. HOP, targets are assigned on basis of high correlation between the chemogenomic profile and the genetic interaction vector of a gene.

4.1.3. Predictions

For predictions of unknown pairs, we have elected to use negative genetic interactions only to predict synergies vs. non-synergistic pairs, using the rank method to assign targets.

During the training phase, the calibration curve given in Figure 5 was produced. According to this, the best parameter for target selection was determined to be $n = 225$ targets for each drug.

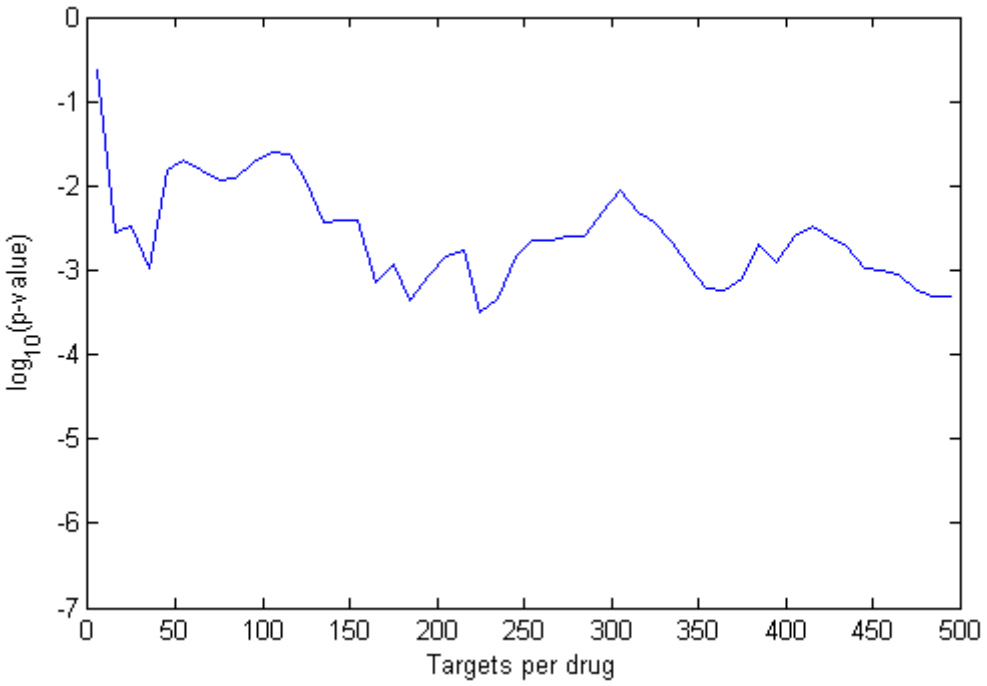

Figure 5: Parameter response curve for the final prediction run. Rank method was used with negative genetic interactions to predict synergies vs. non-synergies.

The ROC curve for the 119 known drugs is given in Figure 6. The area under the curve was 0.71. It can be seen that the curve resembles the average curve from the validation runs given in Figure 3 and that the AUC is slightly higher than the mean AUC given in Figure 4. This is the expected outcome, since the validation sets used to calculate average AUC were smaller, thus there was less training data and mean prediction performance should have been slightly smaller than in this final run of predicting unknown pairs.

Based on the ROC curve, the best prediction threshold was found to be 0.0072, meaning that optimal prediction performance would result if correspondence scores above this were predicted synergistic and those below it were predicted non-synergistic.

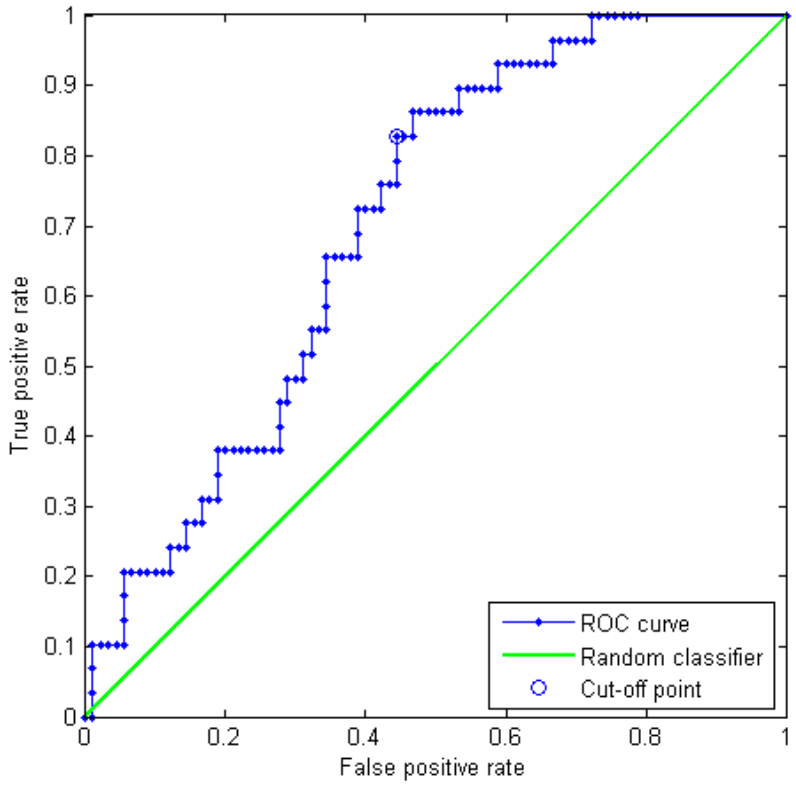

Figure 6: ROC curve for the algorithm after calibrating and training on the entire available dataset of known drug-drug interactions. Green line represents the random guess. Blue circle shows the cut-off point which results in the ROC value closest to the top left corner.

There were a total of 343 conditions given in the Hillenmeyer dataset, and we calculated correspondence scores for each non self-self combination. This generated a total of $\binom{343}{2} = \frac{343.342}{2} = 58653$ correspondence scores.

Figure 7 shows the histogram of all correspondence scores thus calculated. The shape of the curve clearly does not constitute a normal distribution, as measured by the Lilliefors test of normality [**36**] (the p-value for rejecting the hypothesis that the distribution is normal was less than 10^{-3}).

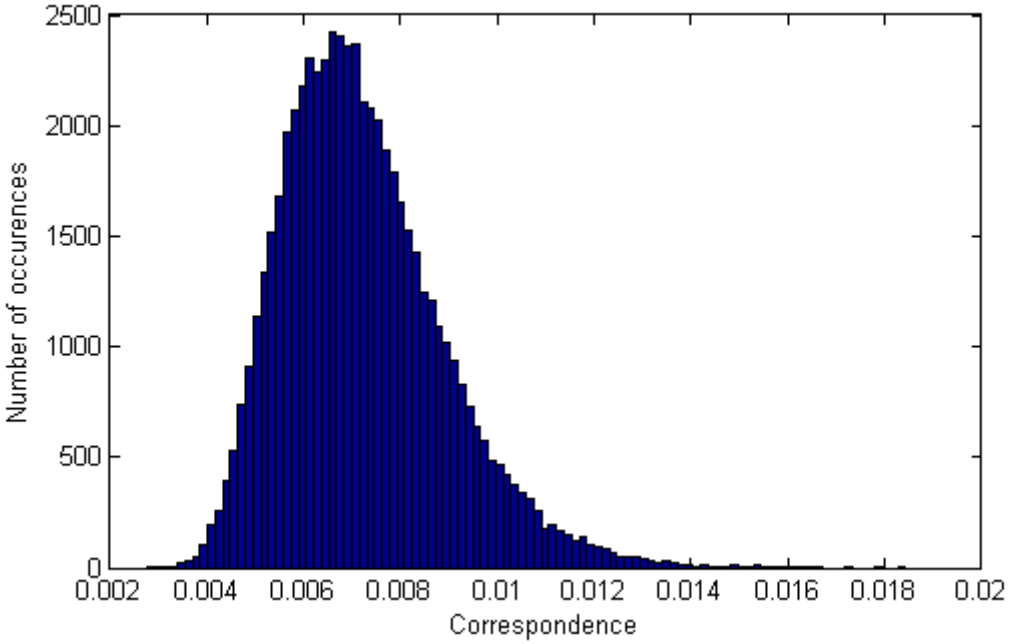

Figure 7: Histogram of correspondence scores calculated for unknown condition pairs covered in the Hillenmeyer dataset, calculated using only negative genetic interactions, and distinguishing only synergies from non-synergies, using the rank target assignment method with 225 targets per drug.

Figure 8 shows a histogram of the base 10 logarithm scores. These were also not normal, as calculated by the Lilliefors test ($p < 10^{-3}$). Although the normal fit appears to follow the data closely, some irregularities and asymmetry is still present even after taking the logarithm.

4.1.4. Experimental verification

For experimental verification, we selected 12 drugs from the Hillenmeyer dataset (see Table 6), and took their correspondence scores (calculated as described in the previous section) to represent predictions. This comprised a total of 66 unique non-self pair correspondence scores. 12 more correspondence scores for self-self pairings were assumed to be 0 as per the definition of the correspondence score.

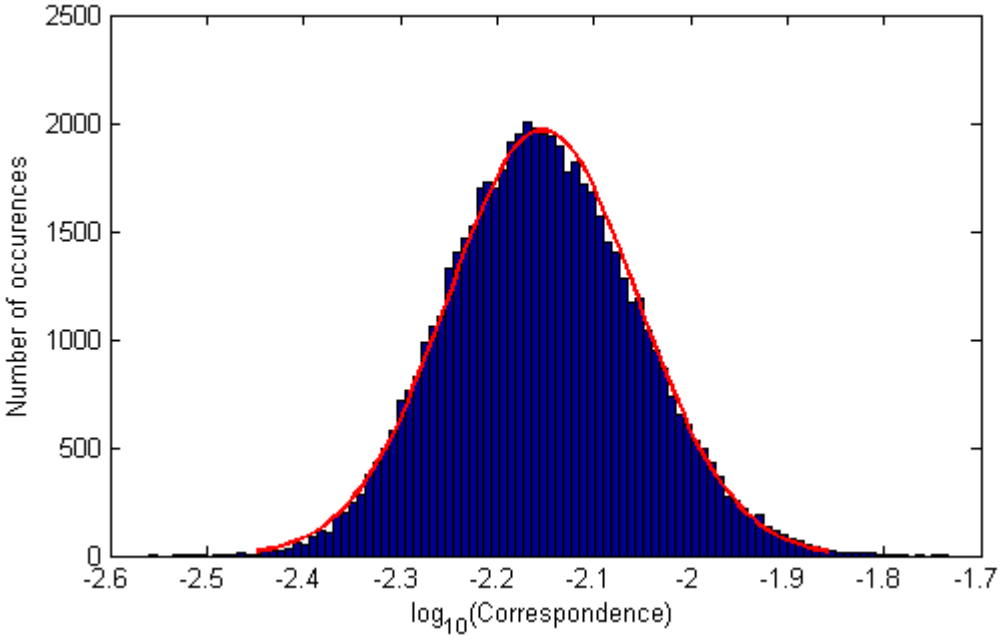

Figure 8: Histogram of the base 10 logarithm of correspondence scores calculated for unknown condition pairs covered in the Hillenmeyer dataset, calculated using only negative genetic interactions, and distinguishing only synergies from non-synergies, using the rank target assignment method with 225 targets per drug. Red line shows normal fit.

18N	1,8-nonadiene
57D	chloroxine
5CQ	cloxyquin
ALV	alverine citrate
BEN	benomyl
BPA	bisphenol
DRO	drofenine hydrochloride
FLA	flufenamic acid
FNP	fenoprofen
RAP	rapamycin
SUL	sulconazole nitrate
TUN	tunicamycin

Table 6: List of drugs that were experimentally verified and their abbreviations.

The correspondence scores calculated for the pairings of these drugs were evenly distributed across the typical range of all correlations. A matrix showing an overview of the scores is given in Figure 9.

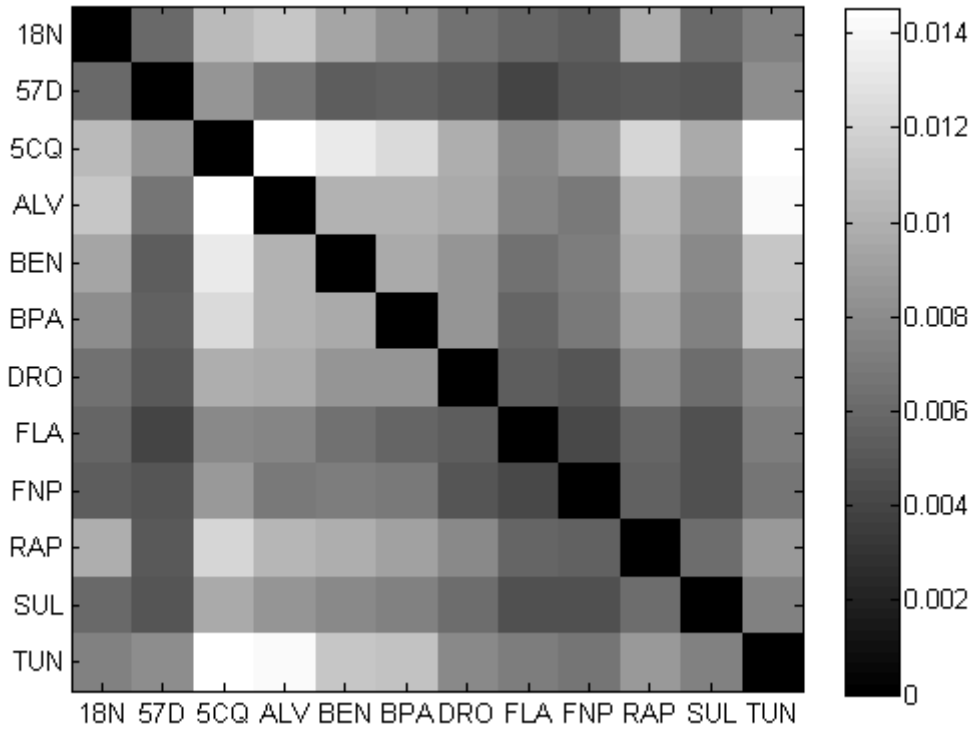

Figure 9: A matrix showing the correspondence score calculated for every pairings of 12 experimentally tested drugs. The matrix is redundant, in reality pairs are unordered and thus the two triangular parts above and below the diagonal are only mirror images of each other.

Upon conducting the drug-drug interaction experiments for all of these pairs, we calculated the alpha scores for each of them to determine the interaction status. The growth scores resulting from the experiment are shown in Figure 10.

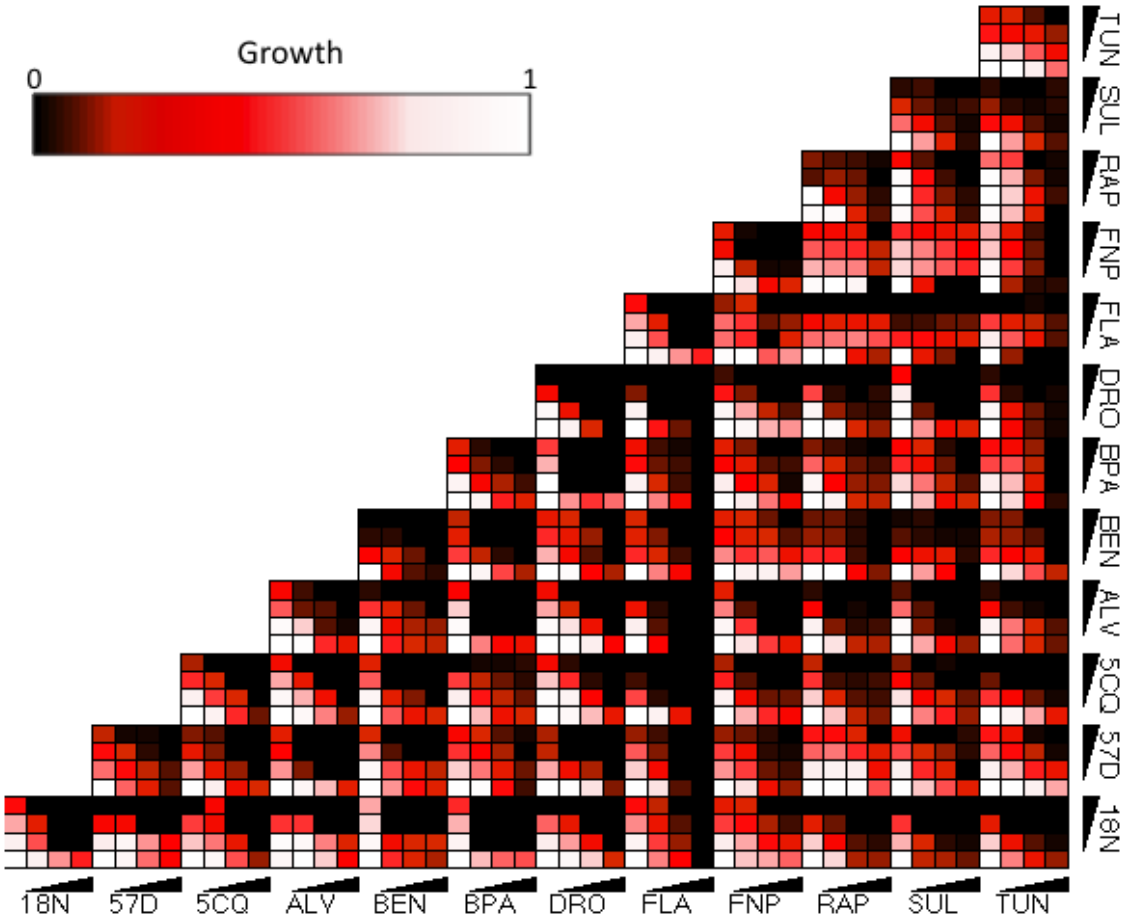

Figure 10: Growth matrices showing the growth for individual drug concentration combinations for tested drug pairs. Colour shows normalized growth score, measured as area under the growth curve. Each 4-by-4 matrix has a concentration gradient of two drugs, one along the horizontal and one along the diagonal. For either drug, experiment was set up such that the bottom left corner contains no drug, while the top row (for vertical drug) and right column (for horizontal drug) contain approximately the minimum inhibitory concentration (MIC) of the respective drug.

We calculated the alpha scores for each of the 4-by-4 growth score matrices as described in the Methods and Materials section. The alpha scores obtained in this way (Figure 11).

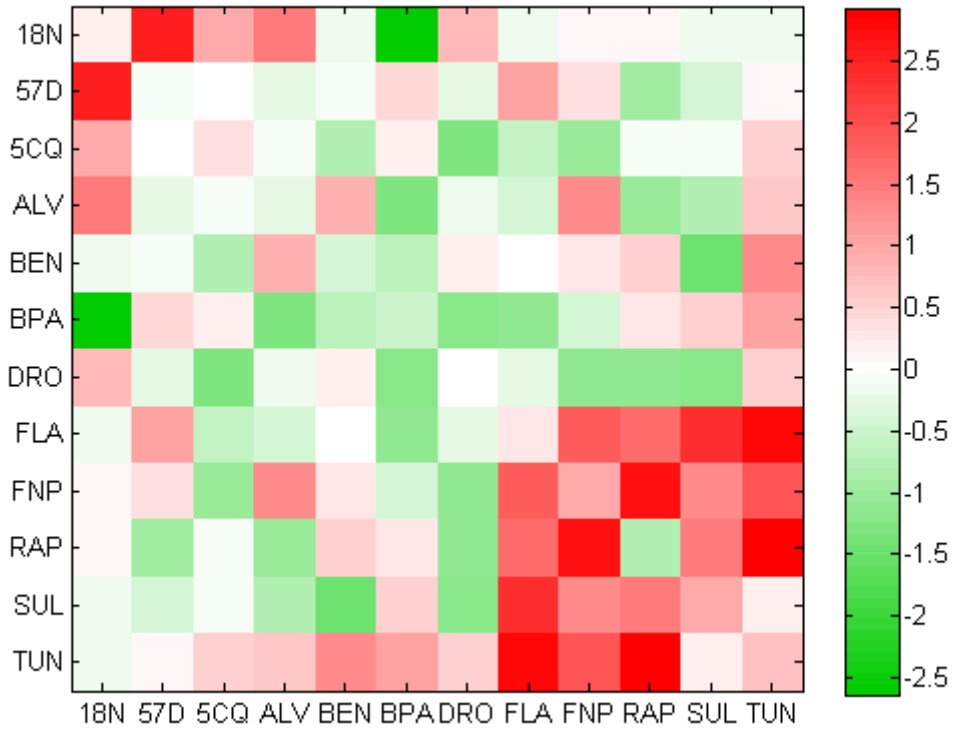

Figure 11: Alpha scores, quantifying the interaction status, of empirically tested drug pairs. Colour shows the alpha score. Values close to 0 (white) indicate independence, low, negative values (green) indicate synergy, high, positive values (red) indicate antagonism. The matrix is redundant, only one alpha score was calculated for each unique pair, therefore the two triangular parts above and below the diagonal are mirror images.

Although our initial calculations appeared to indicate that a fair prediction success rate could be expected, we did not observe a strong relation between the alpha scores and the correspondence scores. Figure 12 shows a plot of alpha and correspondence scores. Self-self pairs are included in this plot and show up as a column to the left, because they are automatically assigned a correspondence score of 0. The two variables appear to have varied in a manner largely independent of each other, contrary to what we had expected. The Pearson correlation coefficient was -0.053, and the Spearman correlation coefficient was -0.069, indicating almost no relation in either case. It should be noted that because synergy is defined to be a negative alpha score, a very good prediction would performance would generate a strong anticorrelation, so in this case, correlation coefficients closer to negative 1 indicate a better performance.

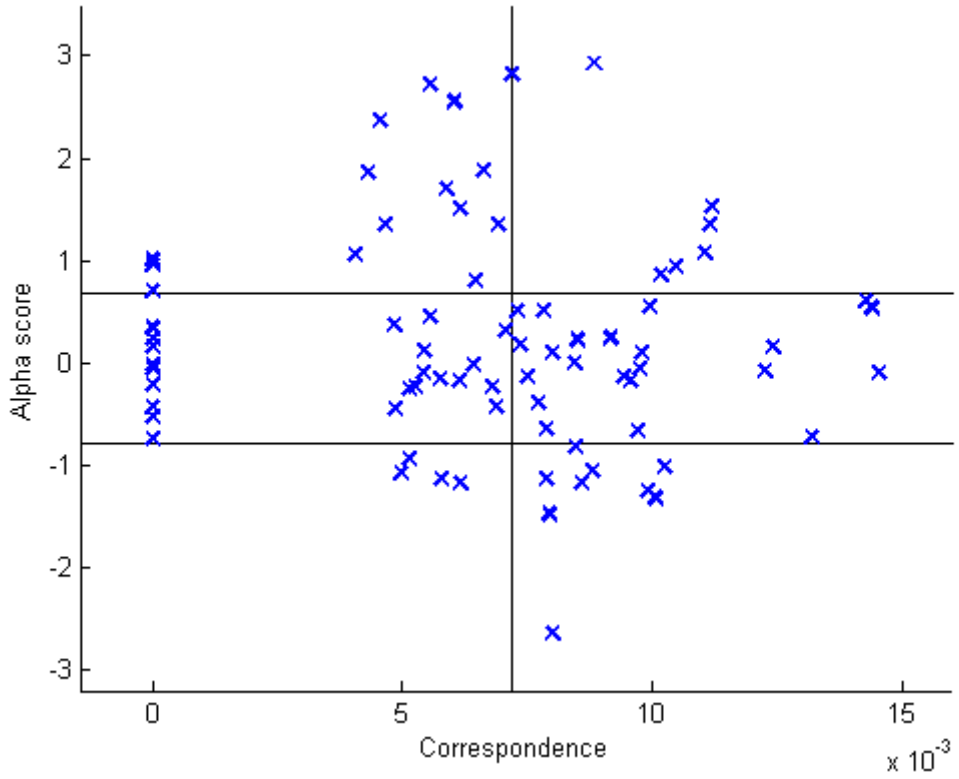

Figure 12: Scatter plot showing alpha score obtained from experiments (vertical axis) vs. calculated correspondence score (horizontal axis). Vertical line shows the threshold of 0.0072, which was expected to be the optimal correspondence threshold for predicting synergy vs. non-synergy. Horizontal lines show the thresholds of $\alpha = -0.78$ and $\alpha = 0.68$ for synergy (negative values) and antagonism (high, positive values) given by the 2011 study by Cokol and colleagues [**23**].

For prediction, our strategy was to separate drug pairs by correspondence scores into pairs predicted to be synergistic, or not synergistic (antagonistic or independent pairs). Thus, by comparing to the threshold of 0.0072, the correspondence score is collapsed into a binary prediction. Pairs with a correspondence scores above the threshold are predicted to be synergistic. Figure 13 shows the effectiveness of this prediction method.

Although, as noted earlier, a very strong relationship is absent, it is nevertheless conspicuous that the median, 25th and 75th percentiles, upper and lower bounds for alpha scores of pairs predicted to be synergistic (i.e. predicted to have a low alpha score) were indeed slightly lower. However, the lower bound no longer follows this trend if the only outlier is not excluded.

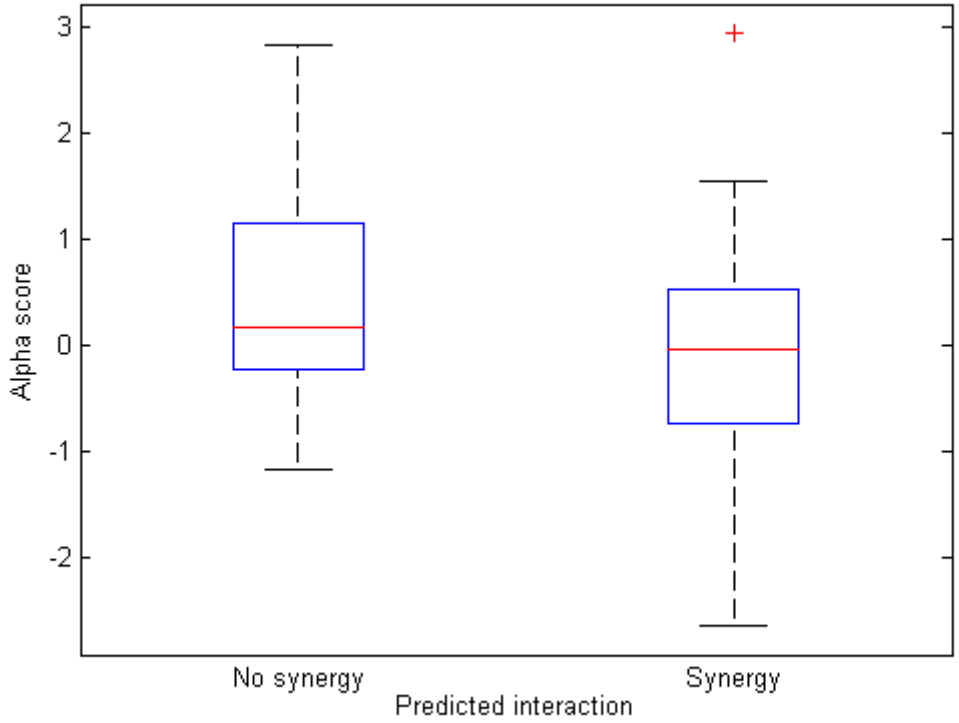

Figure 13: Box plot showing the actual experimental outcome of drug pairs predicted to be synergistic or not synergistic (including independence and antagonism). Red lines show median of alpha scores in either group. Blue boxes show 25th and 75th percentiles. Black lines show the entire range of all data points in the category, with the exception of outliers, which are shown with red crosses.

We also compared the statistical distribution of the correspondence scores within the two categories shown in Figure 13 (predicted to be synergistic vs. the rest). The Mann-Whitney U test p-value corresponding to the null hypothesis of both categories coming from the same distribution was 0.19958. Classically, values less than 0.05 are considered significant, so the alpha scores of our predicted synergistic pairs do not differ from the others at a statistically significant level.

To better compare the results with the validation runs, we constructed an ROC plot showing the true positive and false positive rates of prediction (Figure 14). Overall, performance was poor and the AUC was only 0.51. This is in contrast to the ROC curve obtained after training (Figure 6), which had a much higher AUC of 0.71 and was overall closer to the top left corner.

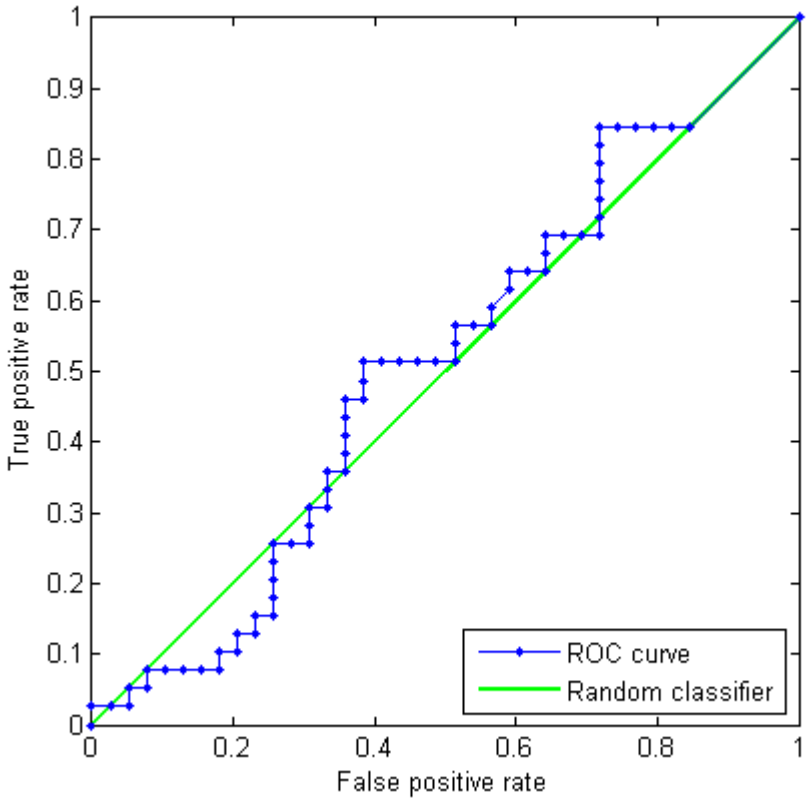

Figure 14: ROC curve for experimentally tested predictions. Green line show random classifier, blue shows actual ROC curve.

5. Conclusions and Future Work

In this project, we have described the implementation of a method to predict drug-drug interactions based on large dataset of gene-gene interactions and chemogenomic profiles.

Although initial tests had shown a high ability to predict interactions based on validation with known data, we did not observe a strong predictive ability after experimentally testing predictions. That said, we nevertheless discovered a large number of novel drug-drug interactions in the process.

A clear avenue for improvement of the prediction algorithm is the refinement of the correspondence calculation function. In the present project, we have only demonstrated a function which considers the immediate neighbors of a gene while evaluating the prevalence of genetic interactions. Although in this current version, genes which are 2 "hops" apart, and genes which are many nodes apart are both considered as not interacting, it is possible that genes in close proximity may be influencing each other through epistatic events as well. To evaluate such a possibility, a more sophisticated graph algorithm would be required to count the connections between the targets of respective drugs.

In addition, we have demonstrated five methods of assigning targets to drugs in this project, and experimentally verified only one. Improvement of these target assignment method would doubtless contribute to increasing performance. With the availability of a large gene-gene interaction network, as well as external pathway databases, it might be possible to employ a more comprehensive method of assigning targeted genes than can be done with the chemogenomic profiles alone.

Lastly, as the datasets used for training the algorithm grow, performance is also likely to improve.

Chapter 2: Standardization of DrugBank

1. Introduction and Background

DrugBank [**37**] is database which contains extensive information regarding the nomenclature, ontology, chemistry, structure, function, action, pharmacology, pharmacokinetics, metabolism and pharmaceutical properties of a large number of drugs, including clinically used, FDA-approved drugs. Information is stored in "DrugCards", which are composed of over 150 fields of data pertaining to a drug.

Among other things, DrugBank includes a list of known interaction partners of each drug, and a natural language description of the phenotype associated with their combination use, collated from a variety of sources. Although this information is very valuable, it is recorded in natural language and is not straightforward to computationally process.

2. Motivation and Contribution

Although DrugBank contains a vast array of valuable information regarding drug-drug interactions, the wholesale analysis of this data is somewhat restricted because it is recorded in natural language, since this presents an additional challenge of parsing the data prior to conducting research on it.

In order to facilitate the study and analysis of drug-drug interactions, we have worked together with a number of volunteers to manually curate and standardize the interaction data recorded in DrugBank. We have verified the accuracy of our curation with various methods, and created a final dataset in which all interactions are categorized according to their features. The dataset we have created is substantially more straightforward to analyze.

During the course of our efforts, we have also composed a hierarchical grouping of different phenotypes, which details their relation to each other.

3. Methods and Materials

3.1. Filtering of DrugBank interactions

As the basis of our research, we used a copy of the DrugBank database last updated on 13 February, 2012.

Before processing and standardizing the interaction information given in DrugBank, we noticed that many interaction descriptions were repeated verbatim throughout the database. Since manually reviewing one of such repeated annotations is sufficient to understand all of them, we were able to reduce the workload substantially by eliminating duplicate annotations from the corpus of data that was manually reviewed. This reduced the number of annotations from 21750 to 6299.

Furthermore, we also noticed that in many cases, the interaction sentences for different drug pairs were syntactically identical, and the only difference was the name of the drugs referenced. We were then able to extract patterns from the corpus of annotations, which mention only anonymized placeholders for the names of drugs, and therefore cover a large number of annotations. The extraction of these patterns is described in detail in section 3.2. Standardization of DrugBank interaction data.

However, after extracting patterns it was further possible to eliminate duplicate patterns, which again reduced the number of entries due for manual review from 6299 to 1897 annotations.

A flowchart showing the amount of data after the two filtering steps is shown in Figure 15.

3.2. Standardization of DrugBank interaction data

The annotations in DrugBank were given on a pairwise basis. For each drug, the known interaction partners of that drug were listed and an annotation text explaining the interaction was present. We elected to first manually process the annotation texts independent from their associated drugs, and then match the standardized data of each annotation to its relevant drug pairs.

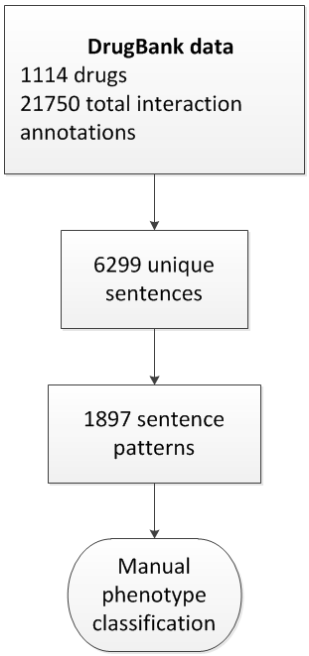

Figure 15: Datasets generated at each stage. The DrugBank database we began with contained 21750 interaction annotations among 1114 drugs. There were 6299 total unique sentences (some annotation texts were repeated several times). After anonymizing the drug names, 1897 syntactical patterns of interaction annotation were obtained. These were manually classified by volunteers.

In order to reduce the workload that would be needed at the manual curation stage, we attempted to reduce the number of annotations in DrugBank by exploiting certain common recurring patterns in the annotations.

The process of extraction of patterns and their manual review is illustrated in Table 4.

3.2.1. Pattern format

In order to extract the patterns, we replaced the drug names referenced in annotation with the strings “Drug_1” and “Drug_2” where applicable. “Drug_1” replaced the name of the drug for which the annotation appeared, and “Drug_2” replaced the name of the drug that was given as the partner.

Raw DrugBank data	Triprolidine -> Ethanol	"Triprolidine may enhance the CNS depressant effects of Ethanol."
	Ethanol -> Triprolidine	"Triprolidine may enhance the CNS depressant effects of Ethanol."
	Prochlorperazine -> Galantamine	"Possible antagonism of action"
	Galantamine -> Prochlorperazine	"Possible antagonism of action"
	Triprolidine -> Donepezil	"Possible antagonism of action"
	Donepezil -> Triptolidine	"Possible antagonism of action"
Duplicates filtered	ID 1: "Triprolidine may enhance the CNS depressant effects of Ethanol." ID 2: "Possible antagonism of action"	
Anonymized syntactical patterns	ID 1: "Drug_1 may enhance the CNS depressant effects of Drug_2." ID 2: "Possible antagonism of action"	
Manual annotations	ID 1: "cns depressant+>" ID 2: "general+"	

Table 7: An example illustrating the data processing operations with representative data from the project. The DrugBank interaction annotations are given on a per-pair basis with redundancy for both directions. The same sentences or syntactical patterns are also often repeated. Filtering duplicates collapses duplicated sentences which arise from reciprocals and the same sentence being reused throughout the database. Anonymization of drug names further collapses sentences which are syntactically identical, and differ only in the names of drugs they mention. Having thus vastly reduced the number of sentences that must be manually curated, we instructed volunteers to assign phenotype tags to each pattern which uses a controlled formal grammar and vocabulary that is easy to parse computationally.

The anonymized annotation patterns we obtained in this way were given to volunteers for manual review and curation. We asked the volunteers to write “tags” for each of the annotation patterns according to the following rules:

- Each tag is a word or phrase representing one of the phenotypes mentioned as arising from the combination usage of the two drugs by the annotation pattern.
- If more than one phenotype is described, one tag for each one is to be written.
- For each tag, a combination of special symbols is to be included to describe metadata regarding the effect observed: The direction in which the phenotype was affected (increase, decrease), the causal relation between the effects of the two drugs (one drug altering the action of another, or reciprocal), and the certainty with which the interaction annotation was phrased (some annotations used certain wording, while others only stated the possibility of interaction). A full list of the symbols is given in Table 8.

We distributed the patterns to our volunteers in such a fashion that each pattern was assigned to exactly 2 different persons. The algorithm for the assignment is given in Algorithm 6.

3.3. Convergence

After the first round of review, we evaluated the tags assigned to each sentence by using a metric based on the Jaccard coefficient [**38**] [**39**]. We refer to this metric as "convergence". The Jaccard coefficient measures the distance between two sets as the ratio between the number of elements in common, versus the total number of elements in the union of the two sets. We used a modified version of the classical definition by adding a provision for division by 0:

$$J = \begin{cases} |A| > 0 \; or \; |B| > 0, & \dfrac{|A \cap B|}{|A \cup B|} \\ |A| = |B| = 0, & 0 \end{cases}$$

In our case, the two sets A and B represented the set of tags assigned by each of the two persons reviewing a given pattern. As a set contains only unique elements, it should be noted here how the equality of elements is determined. We used two definitions of equality:

- Two tags are equal if the phrases used for them (ignoring the symbols) are identical strings. This is referred to as a non-strict comparison.
- Two tags are equal if the previous, non-strict comparison holds, but also the symbols given for the tags are semantically equal (that is to say they convey the same information regardless of where in the string they occur). This is referred to as a strict comparison.

After determining the convergence of all patterns, we repeatedly selected patterns with a convergence less than 1 and reassigned them to the volunteers. This process resulted in the vast majority of the patterns being reviewed with reasonable accuracy (where both persons reviewing it had submitted the same information).

A small number of patterns remained at 0 convergence because the annotation text did not convey any useful interaction information – some annotations in DrugBank were composed solely of recommendations for treatment, and did not specifically state any-

thing regarding the outcome of the combined use of the two drugs, and we ignored such recommendations for our analysis.

3.4. Synonym groups

While performing the analysis of the submissions generated by our volunteers, we noticed that different phrases could sometimes be used to denote the same concept. For instance, synonymous words could be used to describe the same phenotype in different places. Additionally, our participants occasionally made misspellings and typographical errors, leading to an equivalent situation. This generated spurious instances of low convergence, where in fact the tags submitted by either participant matched, and they had simply used different synonyms.

To deal with this issue, we manually reviewed the set of all unique phrases used by the participants, generating a "collective vocabulary" of tag phrases. We then identified sets of phrases which were judged to have the same meaning, thus belonging to a "synonym group". One phrase was chosen to represent each such group of synonyms, referred to as the canonical form.

While calculating convergence, phrases which belonged to synonym group were first replaced by the canonical form, and convergence was calculated afterwards. This allowed us to exclude the effects of synonyms and typographical errors from the analysis.

3.5. Output format

Once we had finished reviewing the patterns to a satisfactory level, they were used to generate the final database of standardized interactions. This took the form of a table, where each row contains the data for a phenotype observed between a pair of drugs, with columns specifying the two participating drugs, the text of the interaction the text of the annotation, the interaction kind, direction, and certainty.

We generated this table in two formats, in the first all data was shown by strings. In the second version, every string in each column was mapped to an integer, to generate an entirely numerical table for better compatibility with other software.

Category	Symbol	Numerical code	Meaning
Interaction kind	+	0	Synergy
	/	1	Additivity
	-	2	Antagonism
	?	3	Unspecified
Certainty	(blank)	0	Uncertain
	*	1	Certain
Direction	(blank)	0	Reciprocal
	>	1	Drug_1 is altering Drug_2
	<	3	Drug_2 is altering Drug_1

Table 8: A list of symbols used to convey metadata regarding tags for DrugBank annotations. There were three categories of symbols, at most one symbol from each category was included in the string representing each tag. The inclusion of the first category, "interaction kind", was mandatory, the other two were optional. If no symbol from a category was found, this was interpreted as a "blank". In versions of our data where each symbol was represented by a number, the numerical codes show the mapping of each symbol.

Algorithm 6: Distribution of DrugBank patterns

Inputs:

- A vector p containing a list of individual patterns that are to be assigned.
- A vector s containing a list of participants who will review the patterns.

Output: A set of vectors u_i where each element lists the elements of p that are assigned to person s_i.

1. $j \coloneqq 1, k \coloneqq 1$.
2. Generate vectors f and g. For each i from 1 to $|p|$ do:
    a. If $k > |s|$, then: $k \coloneqq 1, j \coloneqq j + 1$
    b. If $j > |s|$, then: $j \coloneqq 1$
    c. $f_i \coloneqq j$
    d. $g_i \coloneqq k$
    e. Increment i and j.
3. Obtain a vector q by shuffling p.
4. For each i from 1 to $|q|$:
    a. $j \coloneqq f_i, k \coloneqq g_i$
    b. Assign pattern q_i to participants s_j and s_k: $u_i \coloneqq (s_j, s_k)$

3.6. Verification

In order to verify the validity of our data, we distributed individual tags to a pool of volunteers using Algorithm 6. We developed a program which essentially displayed data pertaining to a single row of a table to the volunteers, and asked them whether the information was correct. A representative image of this program is shown in Figure 16.

By looking at the patterns where both participants judged the entry to be incorrect, we were able to obtain a rough estimate of the correctness of the data we generated. Since the entire dataset was too large to verify in its entirety, we verified only a fraction of randomly sampled entries as a proxy for the entire set.

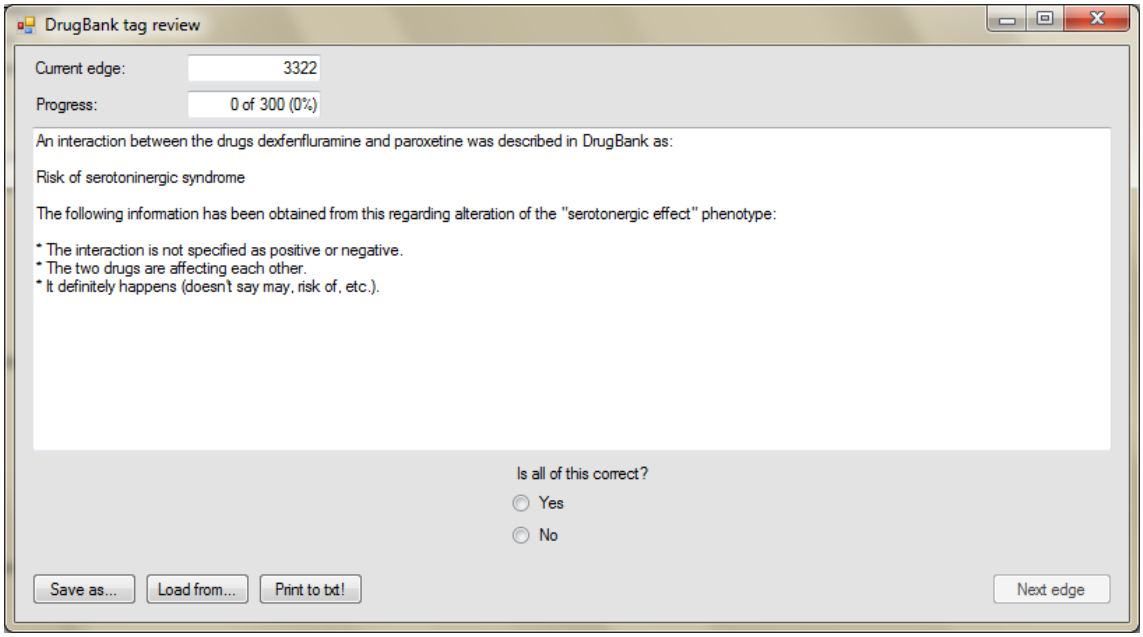

Figure 16: A screen capture of the program used to verify the standardized DrugBank interaction data.

3.7. Construction of a Phenotype Hierarchy

With the finalized set of standardized DrugBank interaction, we reviewed the list of unique phrases used to describe phenotypes for relations between pairs of phrases.

We generated a tree, rooted at the special keyword “general” (used for interaction annotation where it was stated that only the “effect” of a drug is altered). Every node of this tree was connected to a parent which represented a phenotype thought to logically in-

clude that node. In some cases, we added new phrases which grouped together related concepts, such as “Nervous system” to group together phenotypes relating to seizures, dopamine and serotonin imbalances, and similar effects involving the nervous system.

This tree represented a hierarchy of phenotype descriptions, which allowed us to evaluate the relations of different clusters of effects.

4. Results and Discussion

4.1. Tag assignment

We began our stage of manual review with 13 participants. 1897 patterns in total were distributed among them, with around 290 patterns (the numbers were not exactly equal since some divisions did not produce integer results) being assigned to every person. The number of tags assigned to each pattern by each participant is shown Figure 17. It can be seen that most commonly, our participants found only one phenotype for each pattern, slightly less commonly they found 2 phenotypes, and the number of patterns with larger number of phenotypes decreases progressively. A notable number of patterns in each case were assigned no tags. This includes both cases where the annotation in truth did not contain any information relevant to our analysis, but also cases where the participants left the pattern blank (as we instructed them to leave blank any difficult patterns for later review).

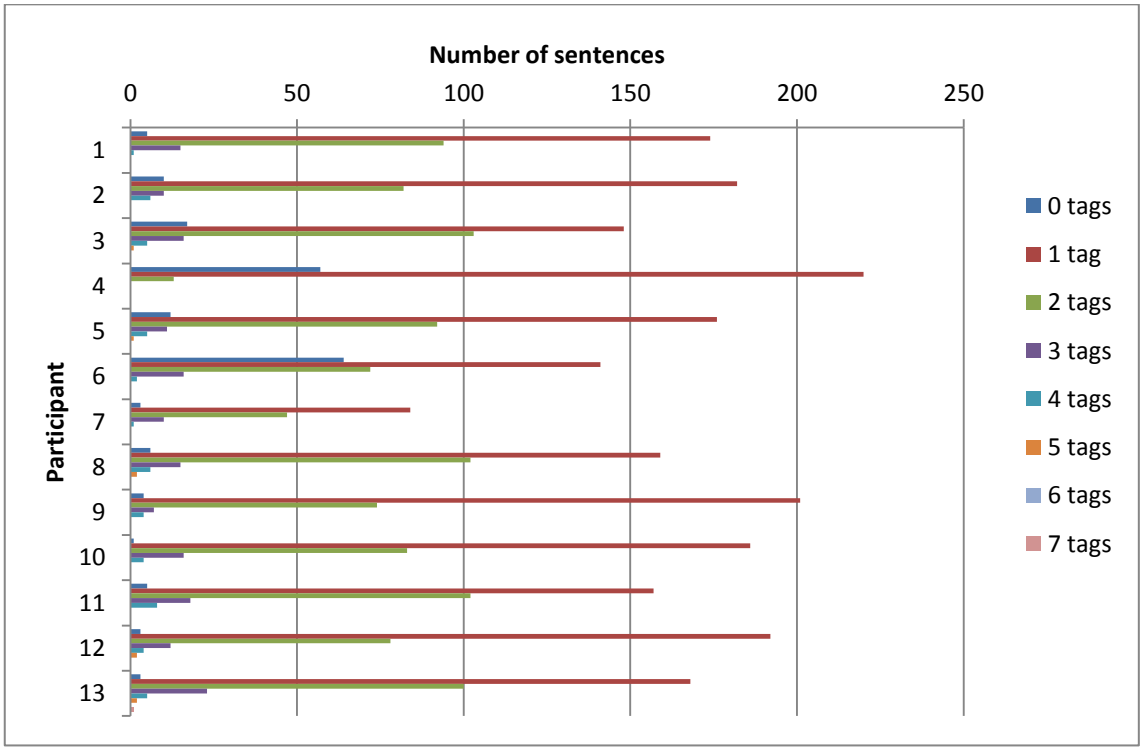

Figure 17: Histograms showing how many tags were assigned to how many patterns by participants at the end of the first pass of manual curation. It can be seen that the distribution of tag number per pattern is roughly uniform between participants.

Figure 18 shows the mean number of tags per pattern between all of the participants. This confirms what can be visually observed from Figure 17: There is an inverse relationship between the number of tags and the number of patterns such that fewer patterns have been found to reference a large number of phenotypes.

It can be expected that if the case of 0 tags is excluded, the number of tags per pattern and the numbers of patterns will follow an exponential function such as $y = ae^{bx}$. The best exponential fit is shown in Figure 19. This function is $y = 461e^{0.95}$ with an R^2 of 0.98 and root-mean-square error of 9.448. Although the fitted function is able to approximate the data, it follows the trend only imperfectly, as can be seen form the datapoint for 2 tags remaining above the fitted line while the remainder fall below it.

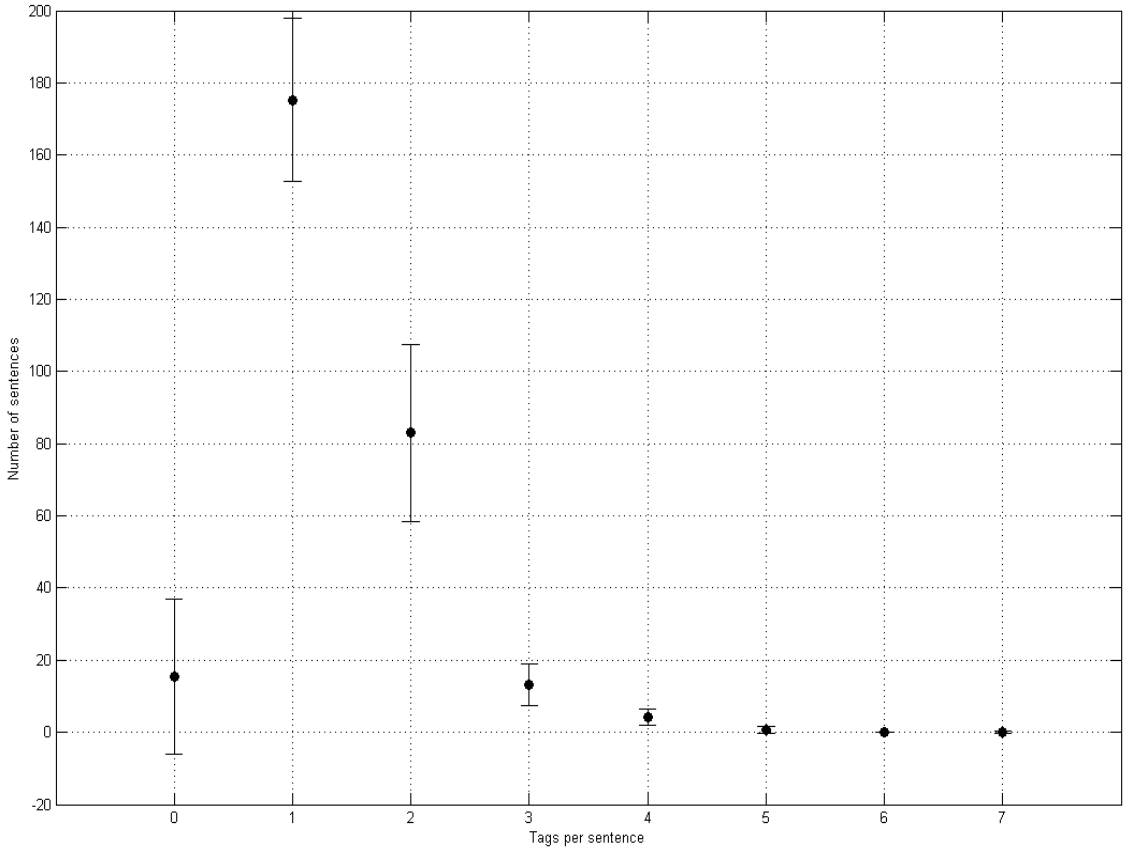

Figure 18: Mean number of tags initially assigned to each pattern by participant.

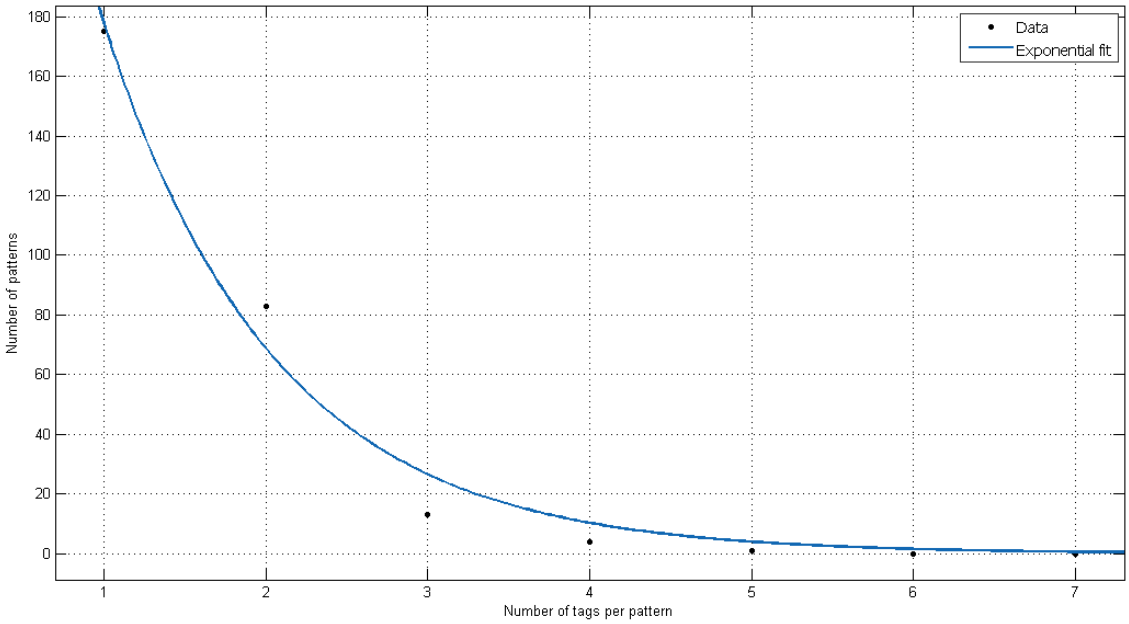

Figure 19: Mean number of tags per pattern with exponential fit.

The progress of convergence after successive rounds of manual review is given in Figure 20. As expected, we observed a gradual decrease in the number of patterns which had a convergence less than 1. After the final round of review, the vast majority of the patterns had a convergence of 1, and the remainder had convergence of 0, being made of annotations that were uninformative and therefore could not be assigned any tags.

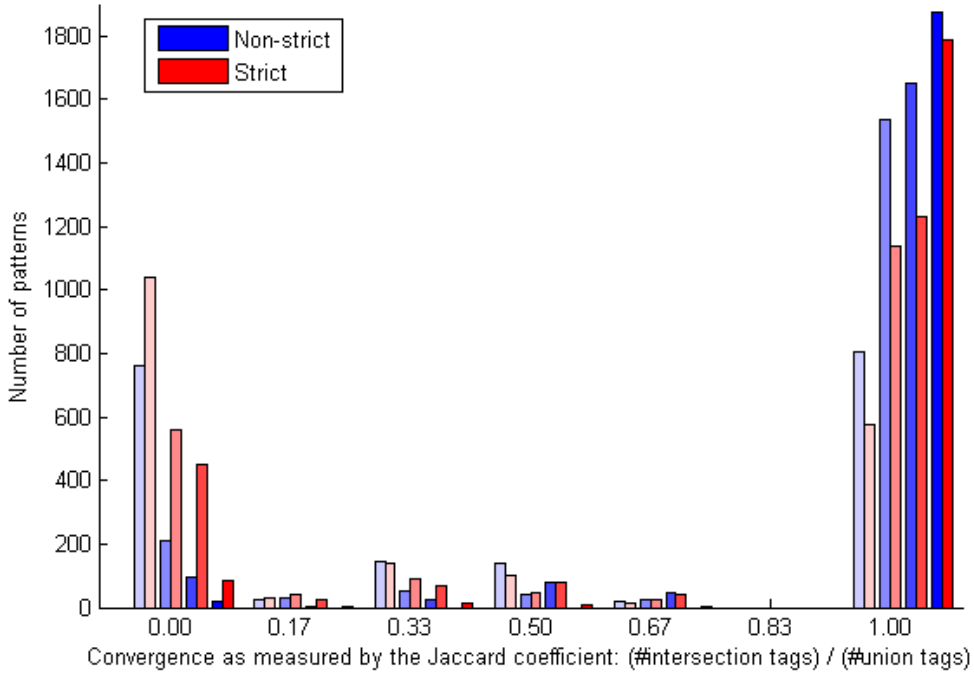

Figure 20: Histogram of tag convergence over time. Convergence of each pattern as measured by the Jaccard similarity coefficient. Blue shows non-strict comparison and Red shows strict comparison. Successively darker shades show the convergence at each successive stage of the manual review process.

4.2. Dataset characteristics

Our final dataset detailed the interactions of 1114 drugs (of which 3 had no interaction phenotypes with any other drug) and as described by a total of 174 unique phenotypes.

According to our standardization scheme, each interaction was also classified with regard to the certainty of the phrasing of the annotation, whether the interaction was reciprocal, and the kind of interaction (increase, decrease, and so on). A breakdown of the number of interaction in each class is given in Table 9.

We saw that 2756 of 14845 interactions (18.6%) were worded in certain terms. The remainder used phrasing such as “possible increase of hypertensive effect” or “risk of serotonin syndrome”, which implied some uncertainty regarding the interaction.

In many cases, an annotation in DrugBank described an interaction in terms of one drug modifying the effect of another drug. For instance, “Drug_1 decreases the clearance of Drug_2” or “Drug_1 enhances the hypotensive effect of Drug_2”. Since it is stated that

one of the drugs is the primary agent of interaction, we recorded such instances as "directional" interactions. In other cases, when no direction was referenced, or the two drugs were said to affect each other (patterns such as "Drug_1 and Drug_2 increase each other's efficacy") we regarded the interaction as undirected, or reciprocal.

Of the interactions listed in our final dataset, the majority, 8594 (57.9%) were directed. However, a large number, 6251 were reciprocal.

Lastly, the third classification involved the nature of the interaction, and whether it increased or decreased a phenotype. In some cases, the annotations explicitly stated that an interaction was synergistic or antagonistic, for these the assignment of interaction kind is straightforward. However, more commonly a phenotype was said to be increased (or amplified, enhanced, and so on) or decreased.

When a phenotype was said to be increased, we assumed that what is meant is a synergistic effect, or in other words that the increase was greater than the simple consequence of additivity. Synergies made up 8009 (54.0%) of the interactions. Likewise, when a phenotype was said to be decreased, we assumed this to mean an antagonistic interaction of the two drugs. These account for another 5316 (35.8%) of the interactions.

A minor, but fairly sizable fraction (1349, 9.1%) of annotations stated that an effect was additive, in these cases, we classified the interaction as additive (essentially, only sentences which actually used the word "additivity" were placed in this group instead of synergy or antagonism).

Certainty	
Certain	2756
Uncertain	12089
Total	14845

Direction	
Directed	8594
Reciprocal	6251
Total	14845

Interaction kind	
Synergistic	8009
Antagonistic	5316
Additive	1349
Unspecified	171
Total	14845

Table 9: The numbers of entries in the final dataset belonging to various classes.

In addition, a small number of annotations (171, 1.2%) only mentioned that the effect was altered when the two drugs were used in combination, but did not specify whether it was increased or decreased. These were placed in the "unspecified" category.

It is possible to regard our whole dataset as the definition of a graph in adjacency list format, with each edge having additional features of phenotype, certainty, direction, and interaction kind. If the dataset is regarded as representing a directed network, it is possible to represent both directed and reciprocal interactions as one way and two way arrows. Subnetworks constructed by taking only the edges corresponding to a single interaction phenotype can be generated, with optional inclusion of non-certain interactions, with color denoting the interaction kind.

In some cases, the analysis of both the entire network and the subnetworks was made substantially more involved by the additional features of directed edges. Therefore, we attempted to evaluate whether it was the possibility of omitting or collapsing direction.

Figure 21 shows the 10 major contributors of directed edges. A very common type of interaction recorded in DrugBank was the modification in various ways of the effective concentration of a drug at the target site, by altering the bioavailability, absorption, clearance, and similar properties (these may be referred to as "pharmacokinetic" effects [**40**]). This is in line with expectations, since drugs which alter such pharmacokinetic properties of other drugs would lead to a directed effect: If Drug_A has some effect (such as decreasing blood pressure) and Drug_B alters absorption of other drugs, the combination would cause less Drug_A to be absorbed and the ultimate observation would be a lessening of the (for instance, hypotensive) effect of Drug_A while no appreciable alteration of Drug_B would be observed.

Interestingly, it appears that toxicity, anticoagulant effect, hypoglycemia, side effects (entries in DrugBank would often refer simply to the "side effects" or "adverse effects" of a drug without specifying) and sympathomimetic properties (which refers to mimicking of the various neurotransmitters which are involved in the functioning of the sympathetic nervous system [**41**] [**42**]) are commonly classed as directed. It is likely that the reason is that such effects are often modified because the function of a drug responsible for them is altered in a pharmacokinetic sense, such as toxicity being decreased as a consequence of increased clearance.

The biggest phenotype, "general", includes the interactions where a general annotation is given stating whether the interaction is synergistic, antagonistic and so on, without specifying the exact phenotype. A large number interactions belong to general to begin with, and it is plausible that many of those would be directed.

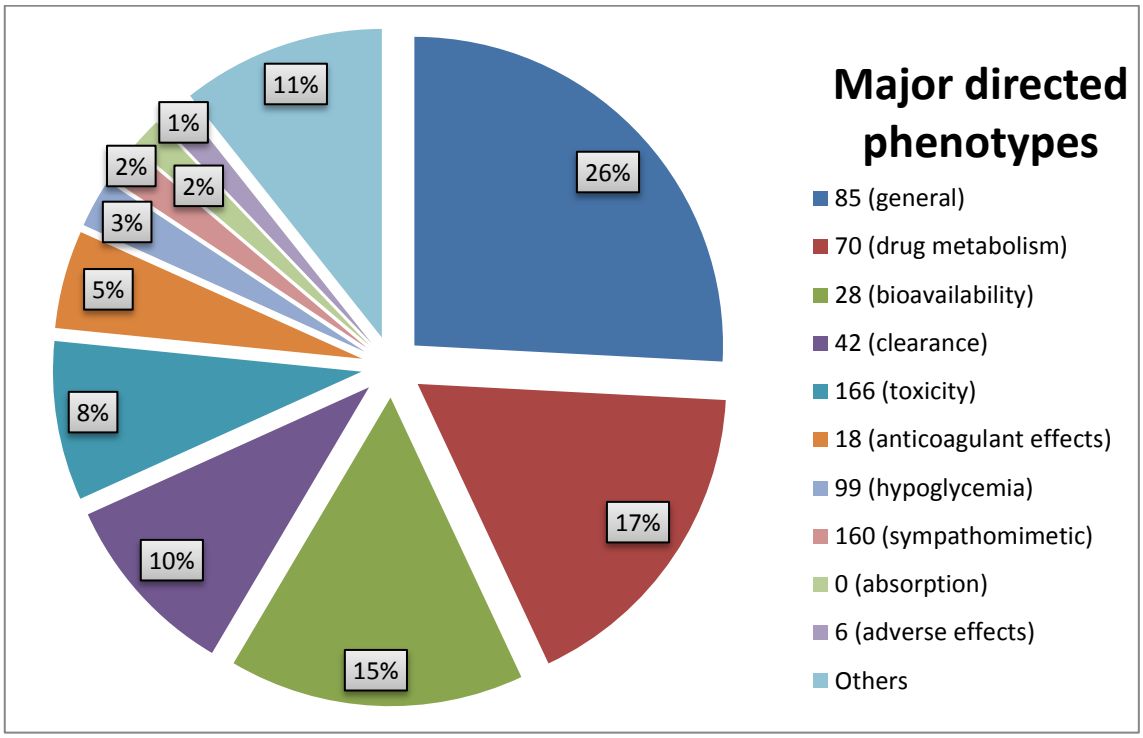

Figure 21: 10 phenotypes which contained the highest number of directed edges. In the legend, the numbers give the numerical ID of the phenotype in parentheses. "Others" is a sum of all remaining phenotypes.

Figure 22 shows the phenotypes which contained the largest number of reciprocated edges. Notably, toxicity appears to contain a high number of reciprocated edges as well as directed edges, in fact, only 62% of toxicity interactions are directed. Most likely, the large number of interactions altering toxicity includes several different mechanisms, some of which are directed and some of which are not. Another phenotype which was also observed to have many directed edges is "general", which, as said above, likely includes a mixture of mechanisms. Most of the other 9 major phenotypes were largely reciprocal (95% or more) with the exception of hypotension (81%) and anticholinergic (70%).

It can be seen that reciprocal interaction was likely for phenotypes relating to the functioning of the cardiovascular system (blood pressure, heart rate, pulse) and various neu-

rological effects. Given the large absolute and relative numbers of reciprocated interactions in these phenotypes, they would be good candidates for collapsing of direction. Since they are already largely reciprocated, removing the directed interactions from them would generate an undirected network without altering the original phenotype subgraph significantly.

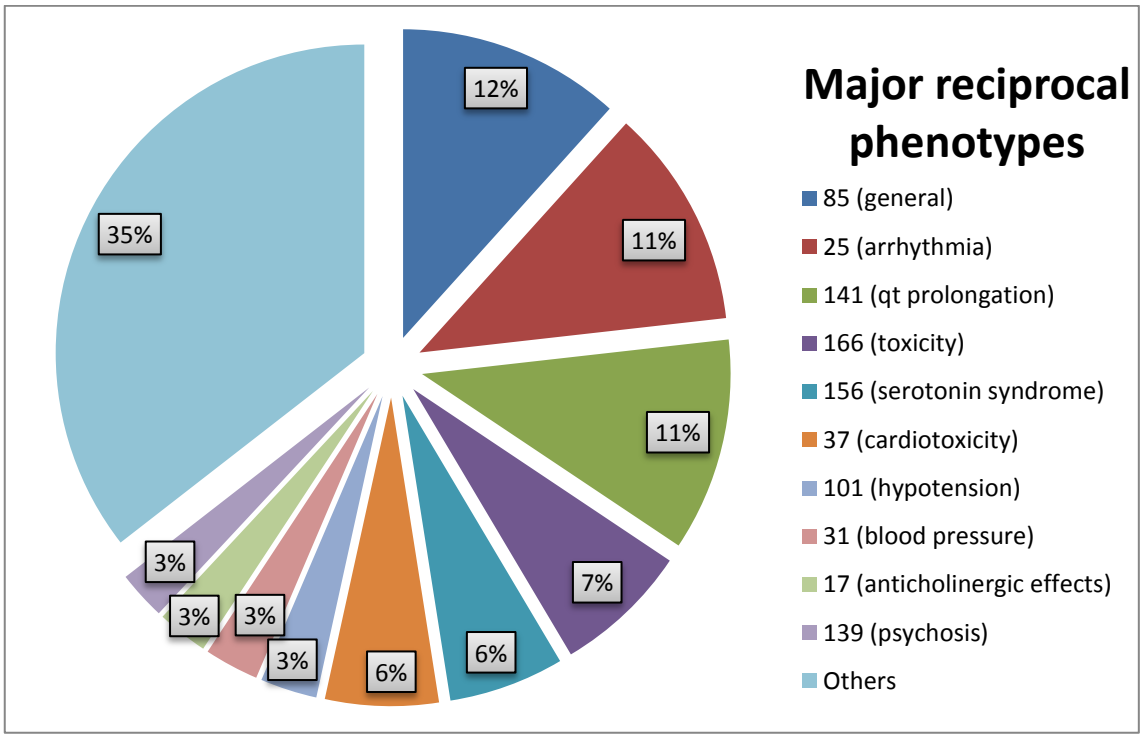

Figure 22: Major phenotypes which have reciprocal interactions, where the combination phenotype is not said to have a direction or specifically alter the function of only one of the drugs. In the legend, numbers show numerical ID codes of the phenotype, and "Others" is a sum of all other phenotypes.

4.3. Phenotype hierarchy

We have also provided a hierarchical tree of the phenotypes included in our data. Figure 23 shows the topology of the phenotype hierarchy.

We have assigned phenotype relations on the basis of inclusion, in the form of a tree. Every phenotype was given exactly one parent phenotype, except for "general", which was the root of the tree. The intention was to assign a phenotype as a parent when the parent phenotype semantically "includes" it.

It is possible to determine relationship based on more than one basis. For instance, in cases where a tissue type, organ or system is mentioned, we were able to easily assign parents based on which organ, tissue or system was part of which. Another common theme was specification of a phenotype by pharmacological concepts (such as bioavailability, toxicity, clearance, absorption, drug metabolism etc.) and these were grouped together by determining which concepts can be taken to be subclasses of the parent.

At times, more than one parent could be assigned to a phenotype, such as for instance the phenotype "cardiotoxicity", which can be thought to belong to "toxicity" as well as "Cardiovascular" (as it affects the heart). In such cases, we attempted to decide which aspect of the phenotype was more fundamental to its manifestation. For cardiotoxicity, for instance, the parent would be "toxicity" because cardiotoxicity is a toxicity-related effect, rather than a cardiovascular system effect.

We added a number of new nodes in order to group related phenotypes as well. In total, 18 such new nodes were introduced. They were signified by a name that starts with capital letters, unlike the normal phenotypes which are always lowercase. These newly introduced nodes do not have any interactions belonging directly to them,

Major hubs in the hierarchy tree were:

- "Pharmacokinetics" (introduced node) – includes phenotypes such as absorption, bioavailability, drug metabolism and clearance.
- "Blood" (introduced node) – includes phenotypes relating to the immune system (such leukopenia and immunosuppression), blood clotting and balance of ions in the blood.
- "Cardiovascular" (introduced node) – includes phenotypes such as blood pressure, blood vessel contraction and dilation, heart failures and arrhythmias.
- "toxicity" – includes various toxicity effects, such as cardiotoxicity, hepatotoxicity, neurotoxicity and nephrotoxicity.
- "Beneficial effect" (introduced node) – includes a number of effects which would generally be considered beneficial and conducive to treatment of patient. Included phenotypes such as antiarrhythmic effects, antiretroviral effects, contraception, antineoplastic effects and cardioprotective effects.

- “Nervous system” (introduced node) – included phenotypes such as neuromuscular effects, serotonin syndrome, cholinergic effects, anticholinergic effects, analgesia, seizures, sedation, psychosis and anorexia.
- “Renal effects” (introduced node) – included phenotypes such as renal impairment, crystalluria and diuretic effects.

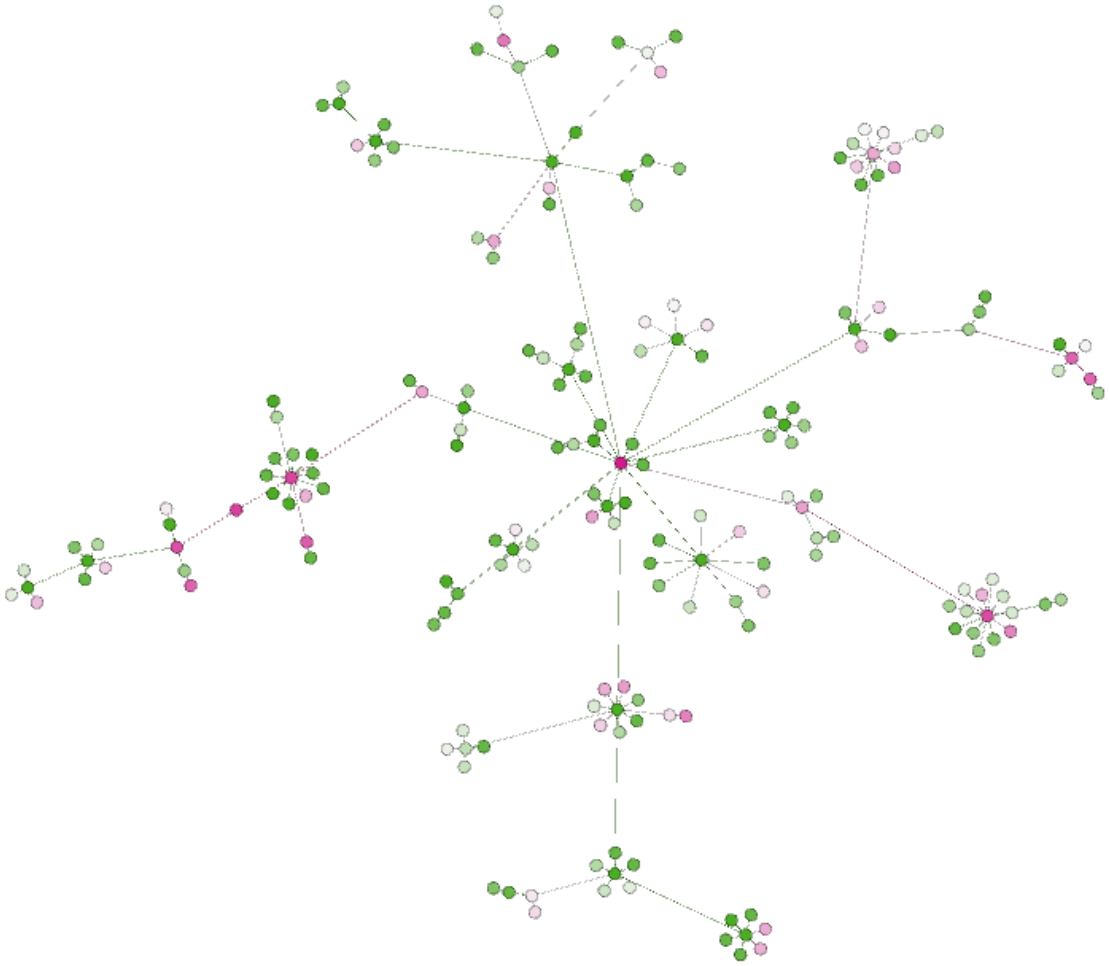

Figure 23: A graph showing the topology of our phenotype hierarchy tree. Each node represents one phenotype, and the colour shows the number of interactions recorded for that phenotype in relative terms (purple is more, green is less). The central purple note represents the “general” phenotype, which all other phenotypes were assumed to belong to.

4.4. Network characteristics

In order to examine part of the interaction dataset in greater detail, we investigated the network properties of the whole network and subnetworks.

For the entire network, we collapsed the initial dataset by removing phenotype data. A network was constructed such that two nodes (each node representing a drug) were connected if there was any interaction at all recorded between them, regardless of which phenotype or interaction kind (meaning synergy, antagonism, additivity or unspecified interaction). The network was then analyzed in two ways: We considered all of the interactions in the network, and only the ones marked as certain.

The characteristics of the network are given in Table 10. We observed that in all, 1111 drugs were involved in some sort of interactions. This is in comparison to 1114 which were initially present in the raw dataset – 3 drugs did have interaction annotations but these annotations were not found to describe any relevant interaction phenotype. Only 710 drugs were connected to other drugs by certain interactions.

	All	**Only certain**
Connected nodes	1111	710
Directed edges	8594	1560
Directed loops	0	0
Two-way arrows	10721	3691
Edges if collapsed to undirected	10865	3723
Components (1+ nodes)	3	3
3-cliques (communities)	11008 (6)	1226 (9)
4-cliques (communities)	7828 (6)	349 (4)
Biggest-clique (communities)	14: 10 (1)	6: 17 (7)

Table 10: Network characteristics of the entire interaction network, after collapsing all phenotypes. The left column shows data for the whole dataset, the right column shows only interactions marked as certain.

We observed that the 1111 drugs made up 3 connected components (such that it was possible to find a path between nodes within a single component, but not nodes in different components). Considering only the certain interactions, we saw that although there were much fewer edges connecting the nodes, when singleton nodes were excluded, the remaining 710 drugs still formed 3 connected components. We surmise that the-

se 401 drugs which did not appear connected in the certain-only network were involved only in interactions which were recorded in an uncertain manner.

We observed 11008 3-cliques which made up 6 communities (where a community is made up of overlapping cliques) in the whole network. In the certain-only network, there were only 1226 3-cliques and they were broken up into 9 communities.

Similarly for 4-cliques, the whole network contained 7828 4-clique in 6 communities, while the certain-only network contained 349 4-cliques in 4 communities.

4.5. Verification

We had also conducted a verification procedure to check the accuracy of our data. In total, we reviewed 1050 of the interactions present in our dataset. These were distributed such that every interaction was seen by exactly two participants, and marked as correct or incorrect.

11 interactions were marked as incorrect by both participants, and upon further review found to be incorrect and were rectified. This makes up 1.05% of all the interactions verified. The interactions were selected at random.

Another 75 interactions (7.14%) were marked incorrect by only one participant. Upon further review, only 25 (2.38%) of these actually needed rectification.

5. Conclusions and Future Work

We have described here the standardization of a large drug-drug interaction database and presented the end result. Our approach can be duplicated for other similar databases, and it can also be used to update the data as the contents of DrugBank itself are revised.

The creation of a standardized, systematic dataset that is conducive to computational processing will, we hope, prove valuable to researcher wishing to investigate the properties of drug-drug interaction networks, especially in the context of humans, where extensive experimentation may not always be viable or practical.

Furthermore, with a standardized version provides a foundation from which more effective record keeping practices can be established, obviating the need for further manual curation in future versions of this database and others.

6. References

1 Boekhout TaRV. Yeasts in food (Woodhead Publishing Series in Food Science, Technology and Nutrition). Woodhead Publishing; 2003.

2 Herskowitz I. Life cycle of the budding yeast Saccharomyces cerevisiae. Microbiological Reviews. 1988;52(4):536.

3 Sherman F. Getting started with yeast. Methods in enzymology. 2002;350:3-41.

4 Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St.Onge RP, Tyers M, Koller D, et al. The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. Science. 2008;320(5874):362-365.

5 Smith AM, Ammar R, Nislow C, Giaever G. A survey of yeast genomic assays for drug and target discovery. Pharmacology & therapeutics. 2010;127(2):156-164.

6 Chua HN, Roth FP. Discovering the Targets of Drugs Via Computational Systems Biology. Journal of Biological Chemistry. 2011 Jul;286(27):23653-23658.

7 Hoon S, M Smith A, Wallace IM, Suresh S, Miranda M, Fung E, Proctor M, Shokat KM, Zhang C, Davis RW, et al. An integrated platform of genomic assays reveals small-molecule bioactivities. Nature Chemical Biology. 2008 Jul;4(8):498-506.

8 Tong AHY. Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. Science. 2001 Dec;294(5550):2364-2368.

9 Hoon S, St.Onge RP, Giaever G, Nislow C. Yeast chemical genomics and drug discovery: an update. Trends in Pharmacological Sciences. 2008 Oct;29(10):499-504.

10 Deutschbauer AM. Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast. Genetics. 2005 Jan;169(4):1915-1925.

11 Winzeler EA. Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. Science. 1999 Aug;285(5429):901-906.

12 Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andrew M Smith B, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002 Jul;418(6896):387-391.

13 Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, et al. Global mapping of the yeast genetic interaction network. science. 2004;303(5659):808-813.

14 Stark C. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006 Jan;34(90001):D535-D539.

15 Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, et al. The Genetic Landscape of a Cell. Science. 2010 Jan;327(5964):425-431.

16 St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nature genetics. 2007;39(2):199-206.

17 Yeh PJ, Hegreness MJ, Aiden AP, Kishony R. Drug interactions and the evolution of antibiotic resistance. Nature Reviews Microbiology. 2009 Jun;7(6):460-466.

18 Bollenbach T, Kishony R. Resolution of Gene Regulatory Conflicts Caused by Combinations of Antibiotics. Molecular Cell. 2011 May;42(4):413-425.

19 Bliss CI. The toxicity of poisons applied jointly. Annals of Applied Biology. 1939 Aug;26(3):585-615.

20 Wolf JB, III EDB, Wade MJ. Epistasis and the Evolutionary Process. Oxford University Press, USA; 2000.

21 Mani R, St.Onge RP, Hartman JL, Giaever G, Roth FP. Defining genetic interaction. Proceedings of the National Academy of Sciences. 2008 Mar;105(9):3461-3466.

22 Loewe S. Die quantitativen Probleme der Pharmakologie. Ergebnisse der Physiologie. 1928 Dec;27(1):47-187.

23 Cokol M, Chua HN, Tasan M, Mutlu B, Weinstein ZB, Suzuki Y, Nergiz ME, Costanzo M, Baryshnikova A, Giaever G, et al. Systematic exploration of synergistic drug pairs. Molecular Systems Biology. 2011;7(544).

24 Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research. 2011 Dec;40(D1):D700-D705.

25 Spearman C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology. 1904;15(1):pp. 72-101.

26 Fisher RA. Statistical methods for research workers. Edinburgh Oliver \& Boyd; 1925.

27 Stouffer SA. American Soldier: Adjustment During Army Life. Sunflower Univ Press; 1949.

28 Roemer T, Boone C. Systems-level antimicrobial drug and drug synergy discovery. Nature Chemical Biology. 2013 Mar;9(4):222-231.

29 Wang YY, Xu KJ, Song J, Zhao XM. Exploring drug combinations in genetic interaction network. BMC Bioinformatics. 2012;13(Suppl 7):S7.

30 Agarwal AK, Tripathi SK, Xu T, Jacob MR, Li XC, Clark AM. Exploring the Molecular Basis of Antifungal Synergies Using Genome-Wide Approaches. Frontiers in Microbiology. 2012;3.

31 Parsons AB, Brost RLL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR, Boone C. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nature Biotechnology. 2003 Dec;22(1):62-69.

32 Gibbons JD, Chakraborti S. Nonparametric Statistical Inference, Fifth Edition (Statistics: Textbooks and Monographs). Chapman and Hall/CRC; 2010.

33 Hollander M, Wolfe DA. Nonparametric Statistical Methods (Wiley Series in Probability and Statistics). Wiley-Interscience; 1999.

34 Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics. 1947 Mar;18(1):50-60.

35 Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006 Jun;27(8):861-874.

36 Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association. 1967;62(318):399-402.

37 Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al. DrugBank 3.0: a comprehensive resource for Omics research on drugs. Nucleic Acids Research. 2010 Dec;39(Database):D1035-D1041.

38 Jaccard P. The distribution of the flora in the alpine zone. New Phytologist. 1912 Feb;11(2):37-50.

39 Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Addison-Wesley; 2005.

40 Huang J, Niu C, Green CD, Yang L, Mei H, Han JDJ. Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network. PLoS Computational Biology. 2013 Mar;9(3):e1002998.

41 Barger G, Dale HH. Chemical structure and sympathomimetic action of amines. The Journal of physiology. 1910;41(1-2):19-59.

42 Katzung B, Masters S, Trevor A. Basic and Clinical Pharmacology, 11th Edition. McGraw-Hill Medical; 2009.