Predicting Drug-Drug Interactions Through Large-Scale Similarity-Based Link Prediction

Achille Fokoue(✉), Mohammad Sadoghi, Oktie Hassanzadeh, and Ping Zhang

IBM T.J. Watson Research Center, Yorktown Heights, USA
achille@us.ibm.com

Abstract. Drug-Drug Interactions (DDIs) are a major cause of preventable adverse drug reactions (ADRs), causing a significant burden on the patients' health and the healthcare system. It is widely known that clinical studies cannot sufficiently and accurately identify DDIs for new drugs before they are made available on the market. In addition, existing public and proprietary sources of DDI information are known to be incomplete and/or inaccurate and so not reliable. As a result, there is an emerging body of research on in-silico prediction of drug-drug interactions. We present Tiresias, a framework that takes in various sources of drug-related data and knowledge as inputs, and provides DDI predictions as outputs. The process starts with semantic integration of the input data that results in a knowledge graph describing drug attributes and relationships with various related entities such as enzymes, chemical structures, and pathways. The knowledge graph is then used to compute several similarity measures between all the drugs in a scalable and distributed framework. The resulting similarity metrics are used to build features for a large-scale logistic regression model to predict potential DDIs. We highlight the novelty of our proposed approach and perform thorough evaluation of the quality of the predictions. The results show the effectiveness of Tiresias in both predicting new interactions among existing drugs and among newly developed and existing drugs.

1 Introduction

Adverse drug reactions (ADRs) are the 4^{th} leading cause of deaths in United States surpassing complex diseases such as diabetes, pneumonia, and AIDS [8]. ADR risk increases significantly when taking multiple drugs simultaneously, which is often common in the elderly population and for managing chronic diseases. In fact, 3 to 5 % of all in-hospital medication errors are due to "preventable" drug-drug interactions (DDIs) [8]. Unfortunately, most ADRs are not revealed in clinical trials with relatively small sizes (at most tens of thousands of participants) due to the rare toxicity of some drugs and the large number of drug combinations that would need to be tested to detect potential DDIs. As a result, the only practical way to explore the large space of drug combinations in search of interacting drugs is through in-silico DDI predictions.

Recently, there has been a growing interest in computationally predicating potential DDIs [11,14,17–21]. Similar to content-based recommender systems,

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 774–789, 2016.
DOI: 10.1007/978-3-319-34129-3_47

the core idea of the predominant similarity-based approach [11,17–21] is to predict the existence of an interaction between a candidate pair of drugs by comparing it against known interacting pairs of drugs. Finding known interacting drugs that are very similar to a candidate pair provides supporting evidence in favor of the existence of a DDI between the two candidate drugs.

In this paper, we introduce Tiresias, a framework that takes in various sources of drug-related data and knowledge as input, and provides as output DDI predictions. In Tiresias, we extend the basic similarity-based DDI prediction framework while addressing the following four significant challenges and shortcomings that are mostly overlooked by prior work:

1. Important Use Case of Newly Developed Drugs: Prior work either (1) are fundamentally unable to make predictions for newly developed drugs (i.e., drugs for which no or very limited information about interacting drugs is available) [18] or (2) could conceptually predict drugs interacting with a new drug, but have not been tested for this scenario [11,17]. Similarity-based approaches (e.g. [11,17]) can clearly be applied to drugs without any known interacting drugs. However, in commonly carried 10-fold cross validation evaluation, prior work using similarity-based approaches have hidden drug-drug interaction associations and not drugs. Thus, the large majority of drugs used at testing are also known during the training phase, which is an inappropriate evaluation strategy to simulate the introduction of a newly developed drug. In our experimental evaluation, we show that the prediction quality of the basic similarity-based approaches drops noticeably when instead of hiding drug-drug associations, we hide drugs. We also show that techniques developed in Tiresias significantly improve the prediction quality for new drugs not seen at training.

2. Skewed Distribution of Interacting Drug Pairs: Contrary to most prior work [11,17,21], we do not assume *a priori* a balanced distribution of interacting drug pairs at training or at testing. There is no reason to believe that the prevalence of pairs of interacting drugs in the set of all the drug pairs is close to 50 % (often assumed in past studies). In fact, in Sect. 6, we present a methodology to estimate a lower and upper bound on the true prevalence of interacting drug pairs in the set of all drug pairs. We show that the true prevalence of DDIs is between 10 % and 30 %.

3. Appropriate Evaluation Metrics and Methodology for Skewed Distribution: Existing work [11,17,21] use mainly the area under the R.O.C curves (AUROC) as the evaluation metric to assess the quality of predictions and often justify their decision to rely on a balanced testing dataset because of the valid observation that AUROC is not too sensitive to the ratio of positive to negative examples. However, as shown in [7] and reinforced in our experimental evaluation section, AUROC is not appropriate for skewed distribution. Metrics designed specifically for skewed distribution such as precision & recall, F-score, or area under Precision-Recall curve (AUPR) should be used instead. Unfortunately, when prior work use these metrics, they do so on a balanced testing data set, which results in artificially high values (e.g., for a trivial classifier that report all pairs of drugs as interacting, recall is 1, precision 0.5 and f-score 0.67).

As shown in our evaluation, on unbalanced testing dataset (with prevalence of drug-drug interacting ranging from 10 % to 30 %), the basic similarity-based prediction produces excellent AUROC values, but mediocre F-score or AUPR.

4. Variety of Data Sources Considered, and Incompleteness of Similarity Measures: Existing techniques have relied on a limited number of data sources (primarily DrugBank) for creating drug similarity measures or drug features. However, in this paper, we exploit information originating from multiple linked data sources (e.g., DrugBank, UMLS, the Comparative Toxicogenomics Database (CTD), Uniprot) to create various drug similarity measures. This poses unique data integration challenges. In particular, since various data sources provide only partial information about a subset of drugs of interest, the resulting drug similarity measures exhibit varying levels of incompleteness. This incompleteness of similarity measures, which has been for the most part overlooked by prior work, is already an issue even when a single data source such as DrugBank is used because not all the attributes needed by a given similarity measure are available for all drugs. Without any additional machine learning features, the learning algorithm cannot distinguish between a low similarity value between two drugs due to incomplete data about at least one of the drugs or real dissimilarity between them. To address this important shortcoming, which affects prediction quality as measured by F-score and AUPR, we introduce a new class of features, called calibration features that captures the relative completeness of the drug-drug similarity measures.

In summary, in this paper, we make the four key contributions. First, we introduce a first of kind semantic integration of a comprehensive set of structured and unstructured data sources including, e.g., DrugBank, UMLS, and CTD (cf. Sect. 3) to construct a knowledge graph. Second, we develop new drug-drug similarity measures based on various properties of drugs including metabolic and signaling pathways, drug mechanism of action and physiological effects (cf. Sect. 5). Third, we build a large-scale and distributed linear regression learning model (in Apache Spark) to predict the existence of DDIs while efficiently coping with skewed distribution of DDIs and data incompleteness through a combination of case control sampling for rare events (cf. Sect. 6) and a new class of calibration features (cf. Sect. 5). Finally, we conduct extensive evaluations with real data to achieve DDI prediction with an average F-Score of 0.74 (vs. 0.65 for the baseline) and area under PR curve of 0.82 (vs. 0.78 for the baseline) using standard 10-fold cross validation for the newly developed drugs scenario (for the existing drug scenario: F-Score of 0.85 vs 0.75 and AUPR of 0.92 vs 0.87). Additionally, we introduce a novel retrospective analysis to demonstrate the effectiveness of our approach to predict correct, but yet unknown DDIs. Up to 68 % of all DDIs found after 2011 were correctly predicted using only DDIs known in 2011 as positive examples in training (cf. Sect. 7).

2 Background: Similarity-Based DDI Predictions

Similar to content-based recommender systems, the core idea of similarity-based approaches [11,17,21] is to predict the existence of an interaction between a candidate pair of drugs by comparing it against known interacting pairs of drugs. These approaches first define a variety of drug similarity measures to compare drugs. A drug similarity measure sim is a function that takes as input two drugs and returns a real number between 0 (no similarity between the two drugs) and 1 (perfect match between the two drugs) indicating the similarity between the two drugs. SIM denotes the set of all drug similarity measures. Entities of interest for drug-drug interaction prediction are not single drugs, but rather pair of drugs. Thus, drug similarity measures in SIM need to be extended to produce drug-drug similarity measures that compare two pairs of drugs (e.g., a pair of candidate drugs against an already known interacting pair of drugs). Given two drug similarity measures sim_1 and sim_2 in SIM, we can define a new drug-drug similarity measure, denoted $sim_1 \otimes sim_2$, that takes as input a two pairs of drugs (a_1, a_2) and (b_1, b_2) and returns the similarity between the two pairs of drugs computed as follows:

$$sim_1 \otimes sim_2((a_1, a_2), (b_1, b_2)) = avg(sim_1(a_1, b_1), sim_2(a_2, b_2))$$

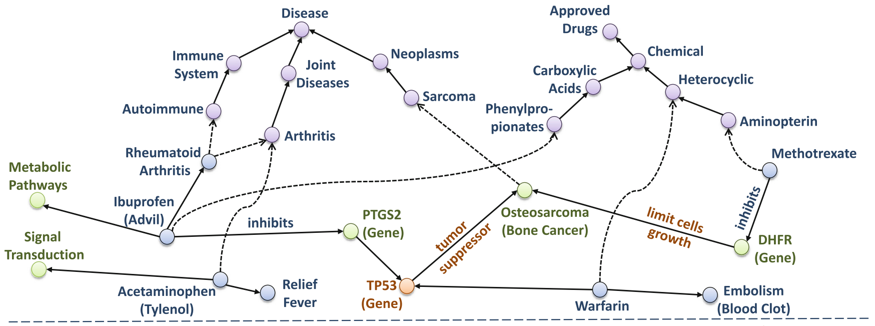
where avg is an average or mean function such as the geometric mean or the harmonic mean. In other words, the first drug similarity measure (sim_1) is used to compare the first element of each pair and the second drug similarity measure (sim_2) is used to compare the second element of each pair. Finally, the results of the two comparisons are combined using, for example, harmonic or geometric mean. The set of all drug-drug similarity measures thus defined by combining drug similarity measures in SIM is denoted $SIM^2 = \{sim_1 \otimes sim_2 | sim_1 \in SIM \wedge sim_2 \in SIM\}$.

Given a set $KDDI$ of known drug-drug interactions, a drug-drug similarity measure $sim_1 \otimes sim_2 \in SIM^2$, and a candidate drug pair (d_1, d_2), the prediction based solely on $sim_1 \otimes sim_2$ that d_1 and d_2 interacts, denoted $predict[sim_1 \otimes sim_2, KDDI](d_1, d_2)$, is computed as the arithmetic mean of the similarity values between (d_1, d_2) and the top-k most similar known interacting drug pairs to (d_1, d_2): $amean(top_k\{sim_1 \otimes sim_2((d_1, d_2), (x, y)) | (x, y) \in KDDI - \{(d_1, d_2)\}\})$ where $amean$ is the arithmetic mean, and, in most cases, k is equal to 1. The power of similarity-based approaches stems from not relying on a single similarity based prediction, but from combining all the individual independent predictions $predict[sim_1 \otimes sim_2, KDDI]$ for all $sim_1 \otimes sim_2 \in KDDI$ into a single score that indicates the level of confidence in the existence of a drug-drug interaction. This combination is typically done through machine learning (e.g., logistic regression): the training is performed using $KDDI$ as the ground truth and, given a drug pair (d_1, d_2), its feature vector consists of $predict[sim_1 \otimes sim_2, KDDI](d_1, d_2)$ for all $sim_1 \otimes sim_2 \in KDDI$.

3 Addressing Data Integration Challenges

One of the salient feature of our Tiresias framework is to leverage many available sources on the Web. More importantly, there is a crucial need to connect these disparate sources in order to create a knowledge graph that is continuously being enriched as ingesting more sources. Notably the life science community has already recognized the importance of the data integration and taken the first step to employ the Linked Open Data methodology for connecting identical entities across different sources. However, most of the existing linkages in the scientific domain are often done statically, which results in many outdated or even non-existent links overtime. Therefore, even when the data is presumably linked, we are forced to verify these links. Furthermore, there are number of fundamental challenges that must be addressed to construct a unified view of the data with rich interconnectedness and semantics — a knowledge graph. For example, we employ entity resolution methodology either through syntactical disambiguation (e.g., cosine similarity, edit distance, or language model techniques [4]) or through semantic analysis by examining the conceptual property of entities [2]. These techniques are not only essential to identify similar entities but also instrumental in designing and capturing similarities among entities in order to engineer features necessary to enable DDIs prediction.

We first begin forming our knowledge graph by ingesting data from variety of sources (including XML, relational, and CSV formats) from the Web. As partially shown in Fig. 1, our data comes from variety of sources such as *DrugBank* [13] that offers data about known drugs and diseases, *Comparative Toxicogenomics Database* [6] that provides information about gene interaction, *Uniprot* [1] that provides details about the functions and structure of genes, *BioGRID* database that collects genetic and protein interactions [5], *Unified Medical Language System* that one is the largest repository of biomedical vocabularies including *NCBI* taxonomy, *Gene Ontology (GO)*, the *Medical Subject Headings (MeSH)* [2], and the *National Drug File - Reference Terminology*



DrugBank: Bioinformatics & Cheminformatics Resource

Drug Name	Drug Targets (Genes)	Symptomatic Treatment
Ibuprofen	PTGS2	Rheumatoid Arthritis
Acetaminophen	PTGS2	Relief Fever
Methotrexate	DHFR	Antineoplastic Anti-metabolite
Warfarin	TP53	Embolism (Blood Clot)

CTD: Comparative Toxicogenomics Database

Gene	Interaction
PTGS2	TP53 (Gene)

Gene	Disease
TP53	Osteosarcoma

Chemical	Pathways	Linked Data Source
Ibuprofen	Metabolic Pathways	KEGG
Acetaminophen	Signal Transduction	Reactome
Methotrexate	Immune System	Reactome

Uniprot: Universal Protein Resource

Gene	Function
TP53	Tumor Suppressor
DHFR	Limits Cell Growth

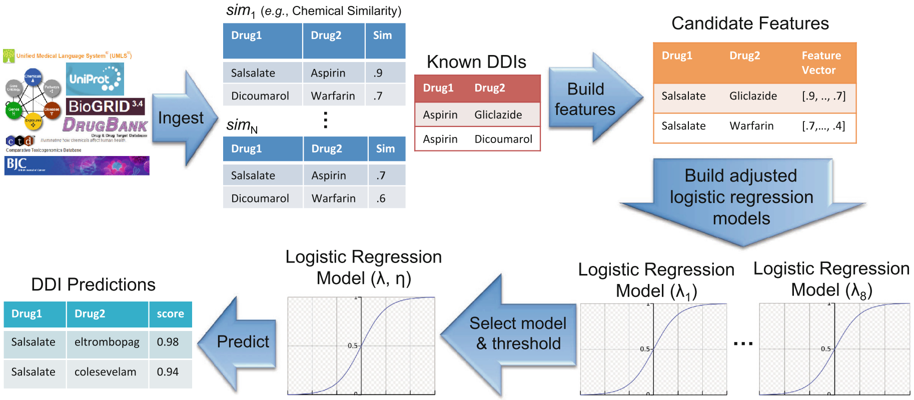
Fig. 1. Semantic curation and linkage of data from variety of sources on the Web.

(NDF-RT) that classifies drug with a multi-category reference models such as cellular or molecular interactions and therapeutic categories [3].

As part of our knowledge graph curation task, we identify which attributes or columns refer to which real world entities (i.e., data instances). Therefore, our constructed knowledge graph possess a clear notion of what the entities are, and what relations exist for each instance in order to capture the data interconnectedness. These may be relations to other entities, or the relations of the attributes of the entity to data values. As an example, in our ingested and curated data, we have a table for *Drug*, and have the columns *Name*, *Targets*, *Symptomatic Treatment*. Our knowledge graph has an identifier for a real world drug *Methotrexate*, and captures its attributes such as *Molecular Structure* or *Mechanism of Actions*, as well as relations to other entities including *Genes* that *Methotrexate* targets (e.g., *DHFR*), and subsequently, *Conditions* that it treats such as *Osteosarcoma (bone cancer)* that are reachable through its target genes, as demonstrated in Fig. 1. Constructing a rich knowledge graph is a necessary step before building our predication model as discussed next.

4 Overview

The overview of our similarity-based DDI prediction approach is illustrated in Fig. 2. It consists of five key phases (the arrows in Fig. 2).


Unified Medical Language System (UMLS)
UniProt
BioGRID 3.4
DrugBank
ctd
Comparative Toxicogenomics Database
BJC
Ingest
Build features
Build adjusted logistic regression models
Logistic Regression Model (λ1)
Logistic Regression Model (λ8)
...
Select model & threshold
Logistic Regression Model (λ, η)
Predict
0.5

sim_1 (*e.g.*, Chemical Similarity)

Drug1	Drug2	Sim
Salsalate	Aspirin	.9
Dicoumarol	Warfarin	.7

sim_N

Drug1	Drug2	Sim
Salsalate	Aspirin	.7
Dicoumarol	Warfarin	.6

Known DDIs

Drug1	Drug2
Aspirin	Gliclazide
Aspirin	Dicoumarol

Candidate Features

Drug1	Drug2	Feature Vector
Salsalate	Gliclazide	[.9, .., .7]
Salsalate	Warfarin	[.7,..., .4]

DDI Predictions

Drug1	Drug2	score
Salsalate	eltrombopag	0.98
Salsalate	colesevelam	0.94

Fig. 2. Overview of similarity-based DDI prediction approach (Color figure online)

Ingestion Phase: In this phase, data originating from multiple sources are ingested and integrated to create various drug similarity measures (represented as blue tables in Fig. 2) and a known DDIs table. Similarity measures are not necessarily complete in the sense that some drug pairs may be missing from the similarity tables displayed in Fig. 2. The known DDIs table, denoted *KDDI*, contains the set of 12,104 drug pairs already known to interact in DrugBank.

In the 10-fold cross validation of our approach, $KDDI$ is randomly split into 3 disjoint subsets: $KDDI_{train}$, $KDDI_{val}$, and $KDDI_{test}$ representing the set of positive examples respectively used in the training, validation and testing (or prediction) phases. Contrary to most prior work, which partition $KDDI$ on the DDI associations instead of on drugs, our partitioning simulates the scenario of the introduction of newly developed drugs for which no interacting drugs are known. In particular, each pair (d_1, d_2) in $KDDI_{test}$ is such that either d_1 or d_2 does not appear in $KDDI_{train}$ or KDD_{val}.

Feature Construction Phase: Given a pair of drugs (d_1, d_2), we construct its machine learning feature vector derived from the drug similarity measures and the set of DDIs known at training. Like previous similarity-based approaches, for a drug candidate pair (d_1, d_2) and a drug-drug similarity measure $sim_1 \otimes sim_2 \in SIM^2$, we create a feature that indicates the similarity value of the known pair of interacting drugs most similar to (d_1, d_2) (see Sect. 5.2). Unlike prior work, we introduce new calibration features to address the issue of the incompleteness of the similarity measures and to provide more information about the distribution of the similarity values between a drug candidate pair and all known interacting drug pairs - not just the maximum value (see Sect. 5.3).

Model Generation Phase: As a result of relying on more data sources, using more similarity measures, and introducing new calibration features, we have significantly more features (1014) than prior work (e.g., [11] uses only 49 features). Thus, there is an increased risk of overfitting that we address by performing L_2-model regularization. Since the optimal regularization parameter is not known a-priori, in the model generation phase, we build 8 different logistic regression models using 8 different regularization values. To address issues related to the skewed distribution of DDIs (for an assumed prevalence DDIs lower than 17 %), we make some adjustments to logistic regression (see Sect. 6).

Model Validation Phase: The goals of this phase are twofold. First, in this phase, we select the best of the eight models (i.e., the best regularization parameter value) built in the model generation phase by choosing the model producing the best F-score on the validation data. Second, we also select the optimal threshold as the threshold at which the best F-score is obtained on the validation data evaluated on the selected model.

Prediction Phase: Let f denote the logistic function selected in the model validation phase and η the confidence threshold selected in the same phase. In the prediction phase, for each candidate drug pair (d_1, d_2), we first get its feature vector v computed in the feature construction phase. $f(v)$ then indicates the probability that the two drugs d_1 and d_2 interact, and the pair (d_1, d_2) is labeled as interacting iff. $f(v) \geq \eta$.

5 Feature Engineering

In this section, we describe the drug similarity measures used to compare drugs and how various machine learning features are generated from them.

5.1 Drug Similarity and Drug-Drug Similarity Measures

Due to space limitation, we describe here only 4 of the 13 similarity measures used to compare two drugs. The other similarity metrics are presented in detail in [9], including physiological effect based similarity, side effect based similarity, two metabolizing enzyme based similarities, three drug target based similarities, chemical structure similarity, MeSH based similarity.

Chemical-Protein Interactome (CPI) Profile Based Similarity: The Chemical-Protein Interactome (CPI) profile of a drug d, denoted $cpi(d)$, is a vector indicating how well its chemical structure docks or binds with about 611 human Protein Data Bank (PDB) structures associated with DDIs [14]. The CPI profile based similarity of two drugs d_1 and d_2 is computed as the cosine similarity between the mean-centered versions of vectors $cpi(d_1)$ and $cpi(d_2)$.

Mechanism of Action Based Similarity: For a drug d, we collect all its mechanisms of action obtained from NDF-RT. To discount popular terms, Inverse Document Frequency (IDF) is used to assign more weight to relatively rare mechanism of actions: $IDF(t, Drugs) = log\frac{|Drugs|+1}{DF(t,Drugs)+1}$ where $Drugs$ is the set of all drugs, t is a mechanism of action, and $DF(t, Drugs)$ is the number of drugs with the mechanism of action t. The IDF-weighted mechanism of action vector of a drug d is a vector $moa(d)$ whose components are mechanisms of action. The value of a component t of $moa(d)$, denoted $moa(d)[t]$, is zero if t is not a known mechanism of action of d; otherwise, it is $IDF(t, Drugs)$. The mechanism of action based similarity measure of two drugs d_1 and d_2 is the cosine similarity of the vectors $moa(d_1)$ and $moa(d_2)$.

Pathways Based Similarity: Information about pathways affected by drugs is obtained from CTD database. The pathways based similarity of two drugs is defined as the cosine similarity between the IDF-weighted pathways vectors of the two drugs, which are computed in a similar way as IDF-weighted mechanism of action vectors.

Anatomical Therapeutic Chemical (ATC) Classification System Based Similarity: ATC [15] is a classification of the active ingredients of drugs according to the organs that they affect as well as their chemical, pharmacological and therapeutic characteristics. The classification consists of multiple trees representing different organs or systems affected by drugs, and different therapeutical and chemical properties of drugs. The ATC codes associated with each drug are obtained from DrugBank. For a given drug, we collect all its ATC code from DrugBank to build a ATC code vector (the most specific ATC codes associated with the drug -i.e., leaves of the classification tree- and also all the ancestor codes are included). The ATC similarity of two drugs is defined as the cosine similarity between the IDF-weighted ATC code vectors of the two drugs, which are computed in a similar way as IDF-weighted mechanism of action vectors.

The set of all drug similarity measures is denoted SIM. As explained in the background Sect. 2, drug similarity measures in SIM need to be extended to produce drug-drug similarity measures that compare two pairs of drugs. SIM^2 denotes the set of all drug-drug similarity measures derived from SIM.

5.2 Top-k Similarity-Based Features

Like previous similarity-based approaches, for a given drug candidate pair (d_1, d_2), a set KDD_{train} of DDIs known at training, and a drug-drug similarity measure $sim_1 \otimes sim_2 \in SIM^2$, we create a new similarity-based feature, denoted $abs_{sim_1 \otimes sim_2}$ and computed as the similarity value between (d_1, d_2) and the most similar known interacting drug pair to (d_1, d_2). In other words,

$$abs_{sim_1 \otimes sim_2}(d_1, d_2) = max(D_{sim_1 \otimes sim_2}(d_1, d_2))$$

where $D_{sim_1 \otimes sim_2}(d_1, d_2)$ is the set of all the similarity values between (d_1, d_2) and all known DDIs:

$$D_{sim_1 \otimes sim_2}(d_1, d_2) = \{sim_1 \otimes sim_2((d_1, d_2), (x, y)) | (x, y) \in KDDI_{train} - \{(d_1, d_2)\}\}$$

5.3 Calibration Features

Calibration of Top-k Similarity-Based Features: For a drug candidate pair (d_1, d_2), a high value of the similarity-based feature $abs_{sim_1 \otimes sim_2}(d_1, d_2)$ is a clear indication of the presence of at least one known interacting drug pair very similar to (d_1, d_2) according to the drug-drug similarity measure $sim_1 \otimes sim_2$. However, this feature value provides to the machine learning algorithm only a limited view of the distribution $D_{sim_1 \otimes sim_2}(d_1, d_2)$ of all the similarity values between (d_1, d_2) and all known DDIs.

For example, with only access to $max(D_{sim_1 \otimes sim_2}(d_1, d_2))$, there is no way to differentiate between a case where that maximum value is a significant outlier (i.e., many standard deviation away from the mean of $D_{sim_1 \otimes sim_2}(d_1, d_2)$) and the case where it is not too far from the mean value of $D_{sim_1 \otimes sim_2}(d_1, d_2)$. Since it would be impractical to have a feature for each data point in D (overfitting and scalability issues), we instead summarize the distribution $D_{sim_1 \otimes sim_2}(d_1, d_2)$ by introducing the following features to capture its mean and standard deviation:

$$avg_{sim_1 \otimes sim_2}(d_1, d_2) = mean(D_{sim_1 \otimes sim_2}(d_1, d_2))$$

$$std_{sim_1 \otimes sim_2}(d_1, d_2) = stdev(D_{sim_1 \otimes sim_2}(d_1, d_2))$$

To calibrate the absolute maximum value computed by $abs_{sim_1 \otimes sim_2}(d_1, d_2)$, we introduce a calibration feature, denoted $rel_{sim_1 \otimes sim_2}$, that corresponds to the z-score of the maximum similarity value of the candidate and a known DDI (i.e., it indicates the number of standard deviations from the mean):

$$rel_{sim_1 \otimes sim_2}(d_1, d_2) = \frac{abs_{sim_1 \otimes sim_2}(d_1, d_2) - avg_{sim_1 \otimes sim_2}(d_1, d_2)}{std_{sim_1 \otimes sim_2}(d_1, d_2)}$$

Finally, for a candidate pair (d_1, d_2), we add a boolean feature, denoted $con_{sim_1 \otimes sim_2}(d_1, d_2)$, that indicates whether the most similar known interacting drug pair contains d_1 or d_2.

Calibration of Drug-Drug Similarity Measures: Features described so far capture similarity values between a drug candidate pair and known DDIs. As such, a high feature value for a given candidate pair (d_1, d_2) does not necessarily indicate that the two drugs are likely to interact. For example, it could be the case that, for a given drug-drug similarity measure, (d_1, d_2) is actually very similar to most drug pairs (whether or not they are known to interact). Likewise, a low feature value does not necessarily indicate a reduced likelihood of drug-drug interaction if (d_1, d_2) has a very low similarity value with respect to most drug pairs (whether or not they are known to interact). In particular, such a low overall similarity between (d_1, d_2) and most drug pairs is often due to the incompleteness of the similarity measures considered. For a drug-drug similarity measure $sim_1 \otimes sim_2 \in SIM^2$ and a candidate pair (d_1, d_2), we introduce a new calibration feature, denoted $base_{sim_1 \otimes sim_2}$, to serve as a baseline measurement of the average similarity measure between the candidate pair (d_1, d_2) and any other pair of drugs (whether or not it is known to interact). The exact expression of $base_{sim_1 \otimes sim_2}(d_1, d_2)$ is as follows:

$$\frac{\displaystyle\sum_{(x,y)\neq(d_1,d_2)\wedge x\neq y} sim_1 \otimes sim_2((d_1, d_2), (x, y))}{|Drugs|(|Drugs| - 1)/2 - 1}$$

The evaluation of this expression is quadratic in the number of drugs $|Drugs|$, which results in a significant runtime performance degradation without any noticeable gain in the quality of the predictions as compared to the following approximation of $base_{sim_1 \otimes sim_2}$ (with a linear time complexity):

$$base_{sim_1 \otimes sim_2}(d_1, d_2) \approx hm(\frac{\displaystyle\sum_{x\neq d_1} sim_1(d_1, x)}{|Drugs| - 1}, \frac{\displaystyle\sum_{y\neq d_2} sim_2(d_2, y)}{|Drugs| - 1})$$

where hm denotes the harmonic mean. In other words, $base_{sim_1 \otimes sim_2}(d_1, d_2)$ is approximated as the harmonic mean of (1) the arithmetic mean of the similarity between d_1 and all other drugs computed using sim_1, and (2) the arithmetic mean of the similarity between d_2 and all other drugs computed using sim_2.

6 Dealing with Unbalanced Data

In evaluating any machine learning system, the testing data should ideally be representative of the real data. In particular, the fraction of positive examples in the testing data should be as close as possible to the prevalence or fraction of DDIs in the set of all pairs of drugs. Although the ratio of positive to negative examples in the testing has limited impact on the area under the ROC curves, as shown in the experimental evaluation, it has significant impact on other key quality metrics more appropriate for skewed distributions (e.g., F-score and area under precision-recall curves). Unfortunately, the exact prevalence of DDIs in the set of all drugs pairs is unknown. Here, we first provide upper and lower bounds on the true prevalence of DDIs in the set of all drug pairs. Then, we discuss logistic regression adjustments to deal with the skewed distribution of DDIs.

Upper Bound: FDA Adverse Event Reporting System (FAERS) is a database that contains information on adverse events submitted to FDA. It is designed to support FDA's post-marketing safety surveillance program for drugs. Mined from FAERS, TWOSIDES [16] is a dataset containing only side effects caused by the combination of drugs rather than by any single drug. Used as the set of known DDIs, TWOSIDES [16] contains many false positives as some DDIs are observed from FAERS, but without rigorous clinical validation. Thus, we use TWOSIDES to estimate the upper bound of the DDI prevalence. There are 645 drugs and 63,473 distinct pairwise DDIs in the dataset. Thus, the upper bound of the DDI prevalence is about 30 %.

Lower Bound: We use a DDI data set from Gottlieb et al. [11] to estimate the lower bound of the DDI prevalence. The data set is extracted from DrugBank [13] and the http://drugs.com website (excluding DDIs tagged as minor). DDIs from this data set are extracted from drug's package inserts (accurate but far from complete). Thus, there are some false negatives in such a data set. There are 1,227 drugs and 74,104 distinct pairwise DDIs in the dataset. Thus the lower bound of the DDI prevalence is about 10 %.

Modified Logistic Regression to Handle Unbalanced Data: For a given assumed low prevalence of DDIs τ_a, it is often advantageous to train our logistic regression classifier on a training set with a higher fraction τ_t of positive examples and to later adjust the model parameters accordingly. The main motivation for this *case-control sampling* approach for rare events [12] is to improve runtime performance of the model building phase since, for the same number of positive examples, the higher fraction τ_t of positive examples yields a smaller total number of examples at training. Furthermore, for an assumed prevalence $\tau_a \leq 0.17$, the quality of the predictions is only marginally affected by the use of a training set with a ratio of one positive example to 5 negative examples (i.e., $\tau_t \sim 0.17$)

A logistic regression model with parameters $\beta_0, \beta_1, \ldots, \beta_n$ trained on a training sample with prevalence of positive examples of τ_t instead of τ_a is then converted into the final model with parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n$ by correcting the intercept $\hat{\beta}_0$ as indicated in [12]:

$$\hat{\beta}_0 = \beta_0 + log\frac{\tau_a}{1-\tau_a} - log\frac{\tau_t}{1-\tau_t}$$

The other parameters are unchanged: $\hat{\beta}_i = \beta_i$ for $i \geq 1$.

We have tried more advanced adjustments for rare events discussed in [12] (e.g., weighted logistic regression and ReLogit), but the overall improvement of the quality of our predictions was only marginal.

7 Evaluation

To assess the quality of our predictions, we perform two types of experiments. First, a 10-fold cross validation is performed to assess F-Score and area under Precision-Recall (AUPR) curve. Second, a retrospective analysis shows the ability of our Tiresias framework to discover valid, but yet unknown DDIs.

7.1 10-Fold Cross Validation Evaluation

Data Partitioning: In the 10-fold cross validation of our approach, to simulate the introduction of a newly developed drug for which no interacting drugs are known, 10 % of the drugs appearing as the first element of a pair of drugs in the set $KDDI$ of all known drug pairs are hidden, rather than hiding 10 % of the drug-drug relations as done in [11,17,18]. Since the DDI relation is symmetric, we consider, without loss of generality, only drug candidate pairs (d_1, d_2) where the canonical name of d_1 is less than or equal to the canonical name of d_2 according to the lexicographic order (i.e., $d_1 \leq d_2$). $KDDI$ is randomly split into 3 disjoint subsets: $KDDI_{train}$, $KDDI_{val}$, and $KDDI_{test}$ representing the set of positive examples respectively used in the training, validation and testing (or prediction) phases, and containing respectively about 8/10th, 1/10th and 1/10th of $KDDI$ pairs. In particular, each pair (d_1, d_2) in $KDDI_{test}$ is such that either d_1 or d_2 does not appear in $KDDI_{train}$ or KDD_{val} (more on this partitioning in Sect. 7.1 of [9]). The training data set consists of (1) known interacting drugs in $KDDI_{train}$ as positive examples, and (2) randomly generated pairs of drugs (d_1, d_2) not already known to interact (i.e., not in $KDDI$) such that the drugs d_1 and d_2 appear in $KDDI_{train}$ (as negative examples). The validation data set consists of (1) the known interacting drug pairs in $KDDI_{val}$ as positive examples, and (2) negative examples that are randomly generated pairs of drugs (d_1, d_2) not already known to interact (i.e., not in $KDDI$) such that d_1 is the first drug in at least one pair in $KDDI_{val}$ (i.e., a drug only seen at validation but not at training) and d_2 appears (as first or second element) in at least on pair in $KDDI_{train}$ (i.e., d_2 is known at training). The testing data set consists of (1) the known interacting drug pairs in $KDDI_{test}$ as positive examples, and (2) negative examples that are randomly generated pairs of drugs (d_1, d_2) not already known to interact (i.e., not in $KDDI$) such that d_1 is the first drug in at least one pair in $KDDI_{test}$ (i.e., a drug only seen at testing but not at training or validation) and d_2 appears (as first or second element) in at least on pair in $KDDI_{train} \cup KDDI_{val}$ (i.e., d_2 is known at training or at validation).

Results: Contrary to prior work, in our evaluation, the ratio of positive examples to randomly generated negative examples is not 1 to 1. Instead, the assumed prevalence of DDIs at training and validation is the same and is in the set {10 %, 20 %, 30 %, 50 %}. For a given DDI prevalence at training and validation, we evaluate the quality of our predictions on testing data sets with varying prevalence of DDIs (ranging from 10 % to 30 %). 50 % DDI prevalence at training and validation is used here to assess the quality of prior work (which rely on a balanced distribution of positive and negative examples at training) when the testing data is unbalanced. For a given assumed DDI prevalence at training/validation and a DDI prevalence at testing, to get robust results and show the effectiveness of our calibration-based features, we perform not one, but five 10-fold cross validations with all the features described in Sect. 5 (solid lines in Figs. 3 and 4) and five 10-fold cross validations without calibration features (dotted lines in Figs. 3 and 4). Results reported on Figs. 3 and 4 represent average over the five 10-fold cross validations.

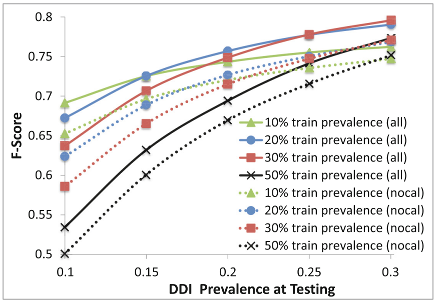


Fig. 3. F-Score for new drugs

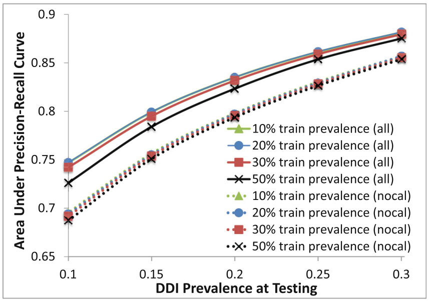


Fig. 4. AUPR for new drugs

The key results from our evaluation are as follows:

- Regardless of the DDI prevalence used at training and validation (provided that it is between 10 % to 30 % -i.e., the lower and upper bound of the true prevalence of DDIs), our approach using calibration features (solid lines in Figs. 3 and 4) and unbalanced training/validation data (non-black lines) significantly outperforms the baseline representing prior similarity-based approaches (e.g., [11]) that rely on balanced training data without calibration features (the dotted black line with crosses as markers). For an assumed DDI prevalence at training ranging from 10 % to 30 %, the average F-score (resp. AUPR) over testing data with prevalence between from 10 % to 30 % varies from 0.73 to 0.74 (resp. 0.821 to 0.825) when all features are used. However, when the training is done on balanced data without calibration features, the average F-score (resp. AUPR) over testing data with prevalence between from 10 % to 30 % is 0.65 (resp. 0.78)[1]. The difference with the baseline is higher the skewer the testing data distribution is.
- For a fixed DDI prevalence at training/validation, using calibration features is always better in terms of F-Score and AUPR (solid vs. dotted lines in Figs. 3 and 4)

[1] Precision (resp. recall) varies from 0.84 to 0.70 (resp. 0.66 to 0.78) with calibration. Precision (resp. recall) is at 0.54 (resp. 0.84) on balanced training w/o calibration.

- As pointed out in prior work, the area under ROC curves (AUROC) is not affected by the prevalence of DDI at training/validation or testing. It remains constant at about 0.92 with calibration features and 0.90 without them.
- Finally, no similarity metric by itself has good predictive power (ATC similarity is the best with 0.58 F-Score and 0.56 AUPR), and removing any given similarity metric has limited impact on the quality of the predictions (the highest decrease was by 1 % in F-Score & AUPR w/o ATC similarity).

Note that we also perform 10-fold cross validation evaluations hiding drug-drug associations instead of drugs. Results presented in [9] show that, even when predictions are made only on drugs with some known interacting drugs, the combination of unbalanced training/validation data and calibration features remains superior to the baseline (F-Score of 0.85 vs 0.75 and AUPR of 0.92 vs 0.87).

7.2 Retrospective Analysis

We perform a retrospective evaluation using as the set of known DDIs ($KDDI$) only pairs of interacting drugs present in an earlier version of DrugBank (January 2011). Figure 5 shows the fraction of the total of 713 DDIs added to DrugBank between January 2011 and December 2014 that our approach can discover based only on DDIs known in January 2011 for different DDI prevalence at training/validation. Figure 5 shows that we can correctly predict up to 68 % of the DDI discovered after January 2011, which demonstrates the ability of Tiresias to discover valid, but yet unknown DDIs.

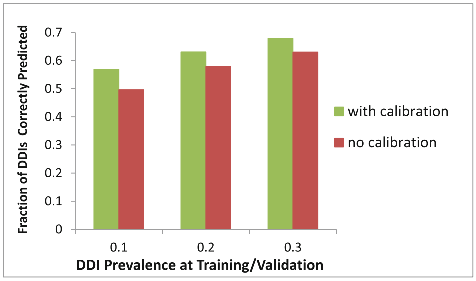


Fig. 5. Retrospective evaluation

8 Conclusion

In this paper, we presented Tiresias, a computational framework that predicts DDIs through large-scale similarity-based link prediction. Experimental results clearly show the effectiveness of Tiresias in both predicting new interactions among existing drugs and among newly developed and existing drugs. The predictions provided by Tiresias will help clinicians to avoid hazardous DDIs in their prescriptions and will aid pharmaceutical companies to design large-scale clinical trial by assessing potentially hazardous drug combinations. We have designed a Web interface and a set of APIs to assist with such use cases [10]. We are currently extending Tiresias to perform link prediction among other entity types in our knowledge graph, turning it into a generic large-scale link prediction system.

References

1. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al.: Uniprot: the universal protein knowledgebase. Nucleic Acids Res. **32**(Suppl. 1), D115–D119 (2004)
2. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(Suppl. 1), D267–D270 (2004)
3. Brown, S.H., Elkin, P.L., Rosenbloom, S., Husser, C., Bauer, B., Lincoln, M., Carter, J., Erlbaum, M., Tuttle, M.: VA national drug file reference terminology: a cross-institutional content coverage study. Medinfo **11**(Pt. 1), 477–481 (2004)
4. Chandel, A., Hassanzadeh, O., Koudas, N., Sadoghi, M., Srivastava, D.: Benchmarking declarative approximate selection predicates. In: ACM SIGMOD International Conference on Management of Data, SIGMOD 2007, pp. 353–364 (2007)
5. Chatr-aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al.: The BioGRID interaction database: 2015 update. Nucleic Acids Res. **43**, D470–D478 (2014). doi:10.1093/nar/gku1204
6. Davis, A.P., Murphy, C.G., Saraceni-Richards, C.A., Rosenstein, M.C., Wiegers, T.C., Mattingly, C.J.: Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res. **37**(Suppl. 1), D786–D792 (2009)
7. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006)
8. Flockhart, D.A., Honig, P., Yasuda, S.U., Rosebraugh, C.: Preventable adverse drug reactions: A focus on drug interactions. Centers for Education & Research on Therapeutics
9. Fokoue, A., Sadoghi, M., Hassanzadeh, O., Zhang, P.: Predicting drug-drug interactions through large-scale similarity-based link prediction. http://researcher.watson.ibm.com/researcher/files/us-achille/adrTechreport.pdf
10. Fokoue, A., Hassanzadeh, O., Sadoghi, M., Zhang, P.: Predicting drug-drug interactions through similarity-based link prediction over web data. In: Proceedings of the 25th International Conference on World Wide Web, WWW 2016. ACM (2016)
11. Gottlieb, A., Stein, G.Y., Oron, Y., Ruppin, E., Sharan, R.: Indi: a computational framework for inferring drug interactions and their associated recommendations. Mol. Syst. Biol. **8**(1), 592 (2012)
12. King, G., Zeng, L.: Logistic regression in rare events data. Polit. Anal. **9**(2), 137–163 (2001)
13. 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 Res. **39**(Suppl. 1), D1035–D1041 (2011)

14. Luo, H., Zhang, P., Huang, H., Huang, J., Kao, E., Shi, L., He, L., Yang, L.: Ddi-cpi, a server that predicts drug-drug interactions through implementing the chemical-protein interactome. Nucleic Acids Res. **42**, W46–W52 (2014). doi:10.1093/nar/gku433
15. Skrbo, A., Begović, B., Skrbo, S.: Classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes. Medicinski arhiv **58**(1 Suppl. 2), 138–141 (2003)
16. Tatonetti, N.P., Patrick, P.Y., Daneshjou, R., Altman, R.B.: Data-driven prediction of drug effects and interactions. Sci. Transl. Med. **4**(125), 125ra31 (2012)
17. Vilar, S., Uriarte, E., Santana, L., Lorberbaum, T., Hripcsak, G., Friedman, C., Tatonetti, N.P.: Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat. Protoc. **9**(9), 2147–2163 (2014)
18. Vilar, S., Uriarte, E., Santana, L., Tatonetti, N.P., Friedman, C.: Detection of drug-drug interactions by modeling interaction profile fingerprints. PLoS ONE **8**(3), e58321 (2013)
19. Zhang, P., Agarwal, P., Obradovic, Z.: Computational drug repositioning by ranking and integrating multiple data sources. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part III. LNCS, vol. 8190, pp. 579–594. Springer, Heidelberg (2013)
20. Zhang, P., Wang, F., Hu, J., Sorrentino, R.: Towards personalized medicine: leveraging patient similarity and drug similarity analytics. AMIA Summits Transl. Sci. Proc. **2014**, 132 (2014)
21. Zhang, P., Wang, F., Hu, J., Sorrentino, R.: Label propagation prediction of drug-drug interactions based on clinical side effects. Scientific reports 5 (2015)