Heliyon 10 (2024) e39728

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Using multiple drug similarity networks to promote adverse drug event detection

Biswajit Padhi [a,1], Ruoqi Liu [a,1], Yuedi Yang [b], Xueqiao Peng [a], Lang Li [c], Pengyue Zhang [b,*], Ping Zhang [a,c,d,**]

[a] Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave, Columbus, OH 43210, USA
[b] Department of Biostatistics and Health Data Science, Indiana University School of Medicine, 410 W. 10th Street HITS 3000, Indianapolis, IN 46202, USA
[c] Department of Biomedical Informatics, The Ohio State University, 1800 Cannon Drive, Columbus, OH 43210, USA
[d] Translational Data Analytics institute, The Ohio State University, 1760 Neil Ave, Columbus, OH 43210, USA

ABSTRACT

The occurrence of an adverse drug event (ADE) has become a serious social concern of public health. Early detection of ADEs can lower the risk of drug safety as well as the expense of the drug. While post-market spontaneous reports of ADEs remain a cornerstone of pharmacovigilance, most existing signal detection algorithms rely on substantial accumulated data, limiting their applicability to early ADE detection when reports are scarce. To address this issue, we propose a label propagation model for generating enhanced drug safety signals using multiple drug features. We first construct multiple drug similarity networks using a range of drug features. We then calculate initial drug safety signals using conventional signal detection algorithms. These original signals are subsequently propagated across each drug similarity network to obtain enhanced drug safety signals. We evaluate our proposed model using two common signal detection algorithms on data from the FDA Adverse Event Reporting System (FAERS). Results demonstrate that enhanced drug safety signals with pre-clinical information outperform the standard safety signal detection algorithms on early ADE detection. In addition, we systematically evaluate the performance of different drug similarities against different types of ADEs. Furthermore, we have developed a web interface (http://drug-drug-sim.aimedlab.net/) to display our multiple drug similarity scores, facilitating access to this valuable resource for drug safety monitoring.

1. Introduction

Adverse drug events (ADEs) refer to unexpected and harmful reactions that occur when medications are used as intended. ADEs have been a persistent challenge in the medical and healthcare communities due to their significant medical and financial burden on patients, and in rare cases, can result in death [1,2]. ADEs become the fourth leading cause of death in the United States, exceeding serious diseases such as pulmonary disease, diabetes and AIDS etc [3]. Each year, over two million major ADEs occur in patients, leading to approximately 100,000 deaths [2]. As a consequence, ADE detection is crucial for drug safety. Early detection of potential ADEs associated with drug candidates during the initial clinical phases of drug development can mitigate risks for patients and reduce hospital costs.

However, pre-marketing clinical studies often fail to fully detect adverse events due to their limited cohort sizes, short durations, and lack of patient diversity [4,5]. As a result, most ADEs are revealed only after the drugs have been marketed. In contrast, post-

* Corresponding author.
** Corresponding author at: Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Ave, Columbus, OH 43210, USA.
E-mail addresses: zhangpe@iu.edu (Pengyue Zhang), zhang.10631@osu.edu (Ping Zhang).
[1] Equal contribution.

https://doi.org/10.1016/j.heliyon.2024.e39728
Received 10 October 2024; Accepted 22 October 2024

Available online 5 November 2024

marketing monitoring, with its larger sample sizes and longer observation periods, allows for a more comprehensive discovery of ADEs.

Spontaneous reporting systems (SRSs) [6] are regulatory agencies-developed mechanisms for monitoring drug safety throughout the post-marketing period. SRSs are gathered from a range of sources, including healthcare professionals, governmental agencies, pharmaceutical firms, medical literature, and direct patient contact. With an abundance of valuable data, SRSs are critical in delivering early warnings about suspected ADEs. A number of signal detection techniques have been proposed for detecting drug safety signals in SRSs. The most widely used methods are Proportional Reporting Rate (PRR) [7] and Reporting Odds Ratio (ROR) [8,9], both of which are based on the ADEs' most frequent statistical analysis. Bayesian approaches are also popular in signal detection files, the most common methods are Bayesian Confidence Propagation Neural Network (BCPNN) [10] and Multi-item Gamma Poisson Shrinker (MGPS) [11]. Drug safety signals (also called signal scores), calculated using these signal detection techniques, are surrogate measures of statistical relationships between drug-ADE pairs, with higher scores indicating stronger associations [6].

Recently, researchers have explored a variety of strategies to improve the detection of ADEs. Most existing methods are designed to generate signals and/or re-rank original signals for drugs with sufficient reports in SRS. For example, Vilar et al. [12–15] predict the adverse effects of pharmacological compounds based on their chemical structure and additional biological features. To address this, one recent study [16] is proposed to generate safety signals for newly approved drugs with few or no safety reports in SRS. Specifically, they develop a label propagation framework that enhances drug safety signals by combining drug chemical structures with data from the FDA Adverse Event Reporting System (FAERS), allowing for the early identification of potential ADEs for newly approved drugs compared to conventional signal detection methods. However, their study primarily focuses on a single source of drug features, which may limit the scalability and comprehensiveness of ADE detection.

To date, a variety of pre-clinical drug property data has been accumulated and is readily accessible. Given that this data from various sources contains a diverse range of properties, it is crucial to comprehensively evaluate different drug features (including chemical, biological, and phenotypic characteristics) for ADE detection. Thus, in this paper, we leverage these multiple drug features to enhance ADE detection. Our approach consists of three main steps. First, we compute the original FAERS-based drug safety signals using conventional signal detection algorithms. Next, we construct drug-drug similarity networks based on a range of drug characteristics, including target sequences, ATC codes, chemical structures, and GO annotations. Finally, we apply a label propagation operation on each similarity network separately to obtain enhanced drug safety signals from the original signal scores.

Overall, our contributions are as follows:

- We propose a label propagation framework utilizing multiple drug similarity networks to generate enhanced drug safety signals for the early detection of ADEs.
- We compare different drug similarity networks for enhancing drug safety signal generation and demonstrate that the proposed framework outperforms existing safety signal detection methods.
- Furthermore, we group ADEs into their respective MedDRA System Organ Classes (SOCs) and compare the performance of the proposed framework on the top 10 most frequent SOCs.
- We perform a case study on two drug-ADE pairs to demonstrate the efficacy of different similarities on different ADE mechanisms.
- We have also developed a website to display the drug similarity scores, providing a valuable resource for drug safety monitoring.

2. Overall framework

Fig. 1 shows the overall framework of this project. The entire workflow is divided into 3 parts. First, the original signal scores for Drug-ADE pairs from the cumulative FAERS dataset are calculated using the MGPS and BCPNN algorithms. Then we created the 6 unique drug similarity networks using various characteristics of the drugs. Finally, the enhanced drug safety signal scores are generated by combining the original scores and similarity networks using a label propagation method.

3. Results

3.1. Dataset

FAERS database. Because the terms used in the FAERS database [17] are determined by the reporter, inaccurate descriptions may be included, we used the standardized FAERS data from 2004 to 2022 [18]. They curated and standardized the FAERS database entries using the preferred term (PT) from the Medical Dictionary for Regulatory Activity (MedDRA) [19]. We extracted 2872 individual drugs and 20,772 ADEs, resulting in a total of 4,207,448 unique drug-ADE combinations. Fig. 2 displays a histogram depicting the number of drugs associated with each ADE. This histogram is divided into 50 bins of size ≈ 48 which indicates ≈ 9200 ADEs drugs are associated with 1 to 48 drugs, ≈ 2600 ADEs are associated with 49 to 96 drugs with subsequent bins representing higher drug counts. The vertical dashed line shows the mean of the distribution that indicates on average 203 drugs are associated with each ADE. Similarly, the histogram in Fig. 3 shows the number of ADEs associated with each drug. In this figure the there are also 50 bins each covering ≈ 246 ADEs. Here, ≈ 1000 drugs are associated with 1 to 246 ADEs, ≈ 350 drugs are associated with 246 to 492 ADEs and so on. the dashed lines shows on average 1465 ADEs are associated with each drug. Both of these histograms show a highly right-skewed distribution, which represents that a small number of drugs are linked to a large amount of ADEs and vice-versa.

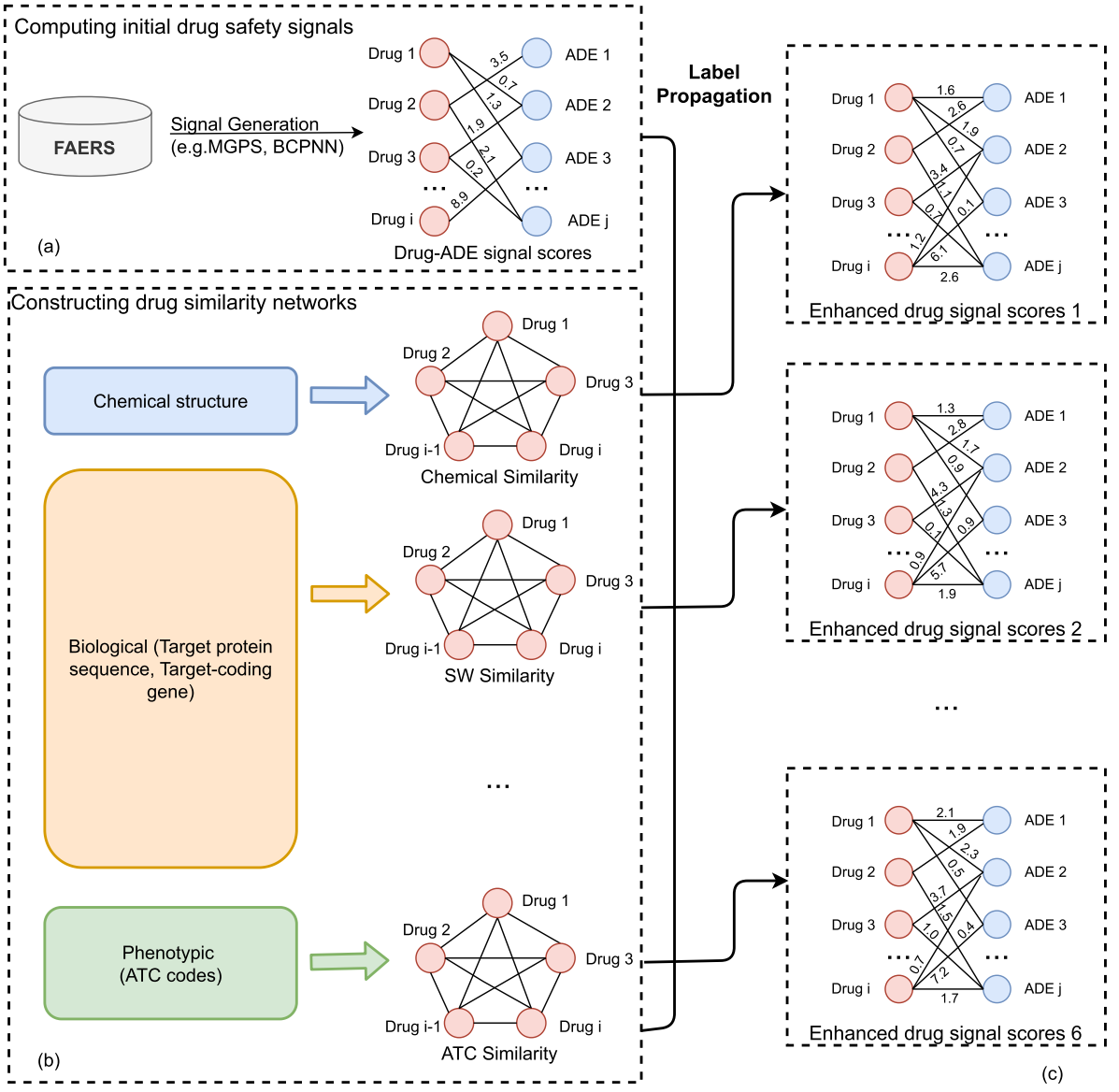

Fig. 1. Overall framework of label propagation using multiple similarity networks. The method includes three parts: (a) Original drug safety signal score computation, (b) Multiple similarity network construction, and (c) Enhanced drug safety signal score generation with similarity networks.

Drugbank database. Drugbank [20] is an open and free drug database that contains a great deal of information on drugs (e.g. target, chemical properties, pharmacology, toxicology). Each drug in this database has its own unique identifier, Drugbank_ID. SMILES annotation is used to represent the chemical structure. The uniprot_ID of the target is retrieved from the Uniprot database. To measure similarities, the chemical structure, sequence of targets, Go words for targets, and ATC code were used as features of the medications.

SIDER database. The side effects were collected from the Side Effect Resource (SIDER) database [21]. The SIDER database currently contains 4251 adverse reactions to 1344 drugs, totaling 152,277 drug-ADE pairs. A total of 643 drugs with all the characteristics (chemical structure, ATC code, sequence, Go terms) were mapped from the Drugbank database to the SIDER database. Both the Preferred Term (PT) and Lowest Level Term (LLT) of MedDRA are used to record ADEs in SIDER. We only selected the Preferred Term(PT) as our ADEs for evaluation.

3.2. Experiment setup

As positive controls, we used known drug-ADE pairs from the SIDER data set and unknown drug-ADE pairs as negative controls. Due to a scarcity of positive samples, we create evaluation datasets that contain the drugs present in both the drug similarity matrix

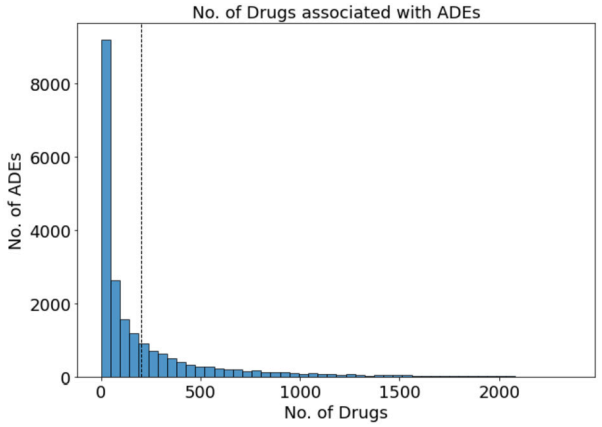

Fig. 2. Frequency of no. of drugs associated with ADEs.

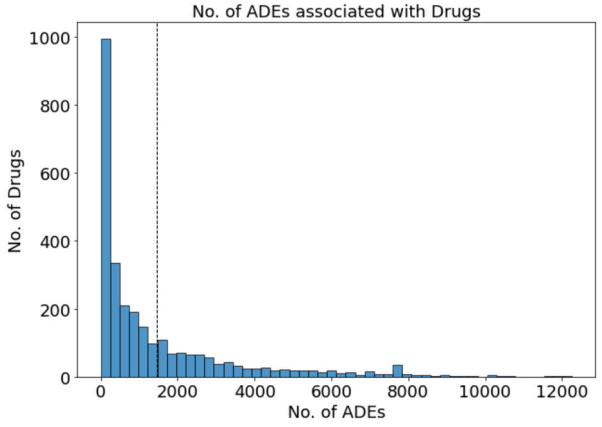

Fig. 3. Frequency of no. of ADEs associated with drugs.

and SIDER, and the top 1000 most frequent ADEs from FAERS data. As the six proposed drug similarity matrices have different numbers of drugs, we generate separate evaluation datasets for each one of them. We evaluate our proposed framework against two baselines (MGPS, BCPNN). We chose MGPS and BCPNN as our baseline because they are not affected by the sampling variance issue and are considered more reliable by the FDA [22]. As evaluation can be viewed as a binary classification job in this work, we used the Area Under the Curve (AUC), the Area Under the Reaction-Recall Curve (AUPR), and the F1 score. The AUC score represents the true positive rate (TPR) and false positive rate (FPR). The TPR and FPR can be computed as follows:

$$\begin{cases} TPR = \dfrac{TruePositive}{TruePositive + FalseNegative} \\ FPR = \dfrac{FalsePositive}{FalsePositive + TrueNegative} \end{cases} \tag{1}$$

AUPR can also be calculated using the precision and recall scores, which illustrates the trade-off between precision and recall at various decision thresholds. Precision may be defined as the probability that the output safety signal score is accurate. The possibility of true safety signals being assessed as outputs may be determined using recall. The definition of precision and recall is shown in Eq. (2).

$$\begin{cases} Precision = \dfrac{TruePositive}{TruePositive + FalsePositive} \\ Recall = \dfrac{TruePositive}{TruePositive + FalseNegative} \end{cases} \tag{2}$$

The F1 score is the harmonic mean of the precision and recall:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Table 1
Comparison between different label propagation framework and corresponding baseline methods using MGPS signal scores on all years reports.

Score	Baseline	LP-ATC	LP-Chem	LP-Seq	LP-GO_BP	LP-GO_CC	LP-GO_MF
AUC	0.743	0.753	0.759	0.759	0.752	0.756	0.755
AUPR	0.226	0.231	0.236	0.242	0.243	0.246	0.246
F1	0.319	0.322	0.330	0.335	0.335	0.339	0.339

Table 2
Comparison between different label propagation framework and corresponding baseline methods using BCPNN signal scores on all years reports.

Score	Baseline	LP-ATC	LP-Chem	LP-Seq	LP-GO_BP	LP-GO_CC	LP-GO_MF
AUC	0.743	0.763	0.786	0.784	0.764	0.776	0.775
AUPR	0.224	0.253	0.280	0.293	0.263	0.290	0.289
F1	0.319	0.332	0.371	0.378	0.350	0.378	0.378

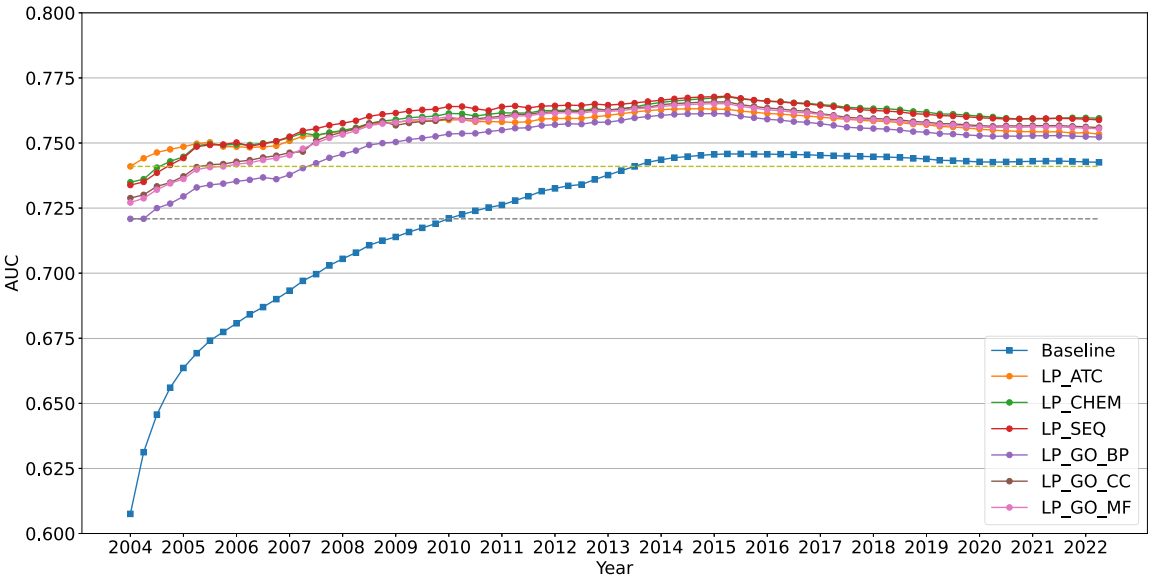

Fig. 4. Comparison of the different methods based on MGPS on yearly cumulative reports.

Furthermore, we divide the datasets into validation (80%) and test (20%) sets. The validation set is used to calculate the optimal value of the model parameter for the Label-Propagation framework. Then we run the best performing model on the test set and compare its result with results obtained by the baseline methods on the same test set.

3.3. Performance assessment

Performance among all ADEs. To evaluate the proposed method, we first compared it to two baseline reference methods (MGPS, and BCPNN) on all-year data (from 2004 to 2022). As shown in Table 1 and 2, BCPNN outperforms MGPS in terms of both AUC and AUPR. On AUC, AUPR, and F1 scores, our strategy surpasses the corresponding baseline methods. Overall, our method can improve AUC score by about $0.02 - 0.04$, and AUPR score by about $0.03 - 0.07$ compared with all the baseline methods. The results show that drug-drug similarity helps enhance safety signals, as similar drugs may cause the same adverse effects. The method incorporates information from similar drugs and improves the original drug safety signal.

Performance among different similarity networks. The results shown in Table 1, demonstrate that all the similarity network has similar results with MGPS signal scores on cumulated all-year data. However, there is more difference among the similarity network results with BCPNN data (Table 2). Here, LP-Chem achieves the highest AUC of 0.759 and 0.786 using MGPS and BCPNN signal scores respectively. LP-Seq also achieves a comparable AUC of 0.759 and 0.784 with the corresponding signal scores.

Additionally, we presented the annual change curve for MGPS and BCPNN using the AUC scores (shown in Fig. 4 and 5). These graphs show that in general, preclinical drug similarities enhance ADE detection - as early as 10 years. The dashed reference lines show how long it takes the baseline to reach the AUC scores obtained by our enhanced safety scores in 2004 Q1, which is the first

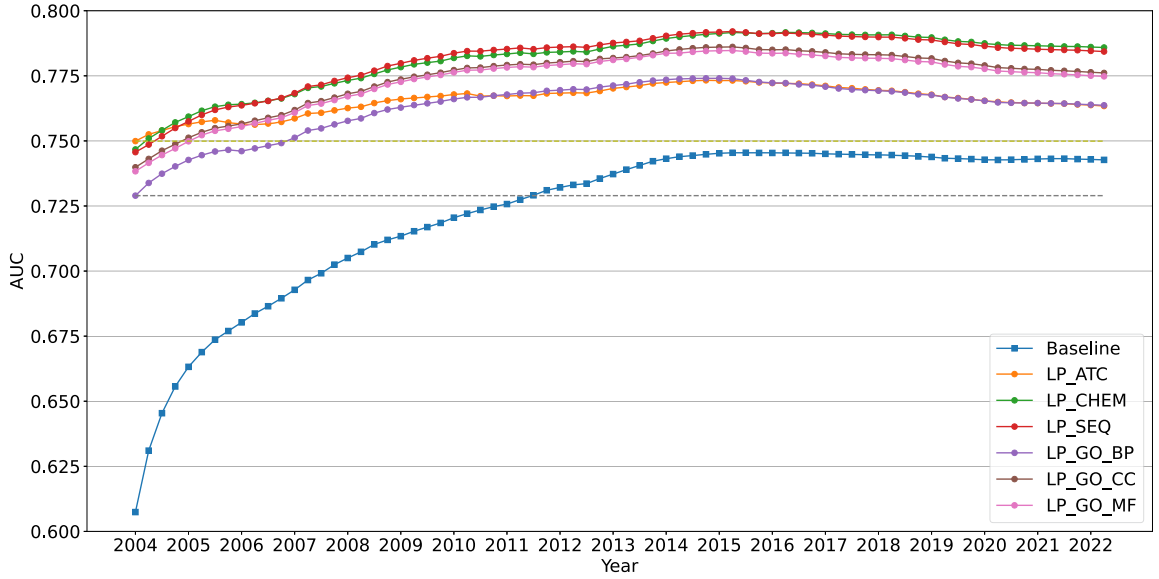

Fig. 5. Comparison of the different methods based on BCPNN on yearly cumulative reports.

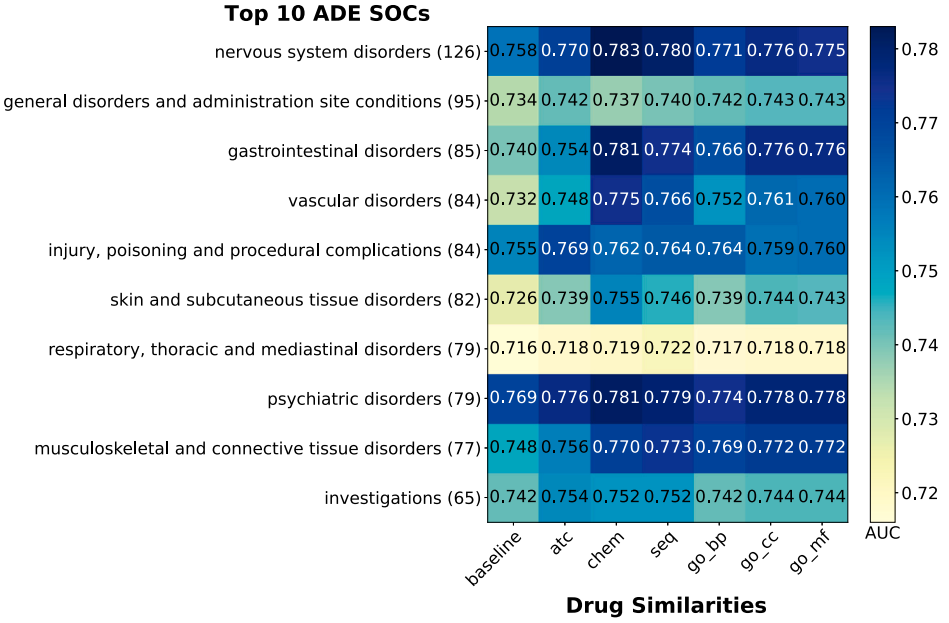

Fig. 6. Comparison of AUC achieved using the different methods based on MGPS on top 10 most frequent MedDRA SOCs. (#) at the end of SOC names in the y-axes denotes the number of unique MedDRA PTs from our filtered SIDER dataset mapped to each SOC.

quarter in our FAERS dataset. The bottom dashed line compares the baseline to the worst performing LP score whereas the top dashed line compares it to the best performing LP-score. In the case of MGPS, the baseline reaches the bottom reference line in 6 years and the top reference line in almost 10 years. In the case of BCPNN, it takes the baseline over 7 years to attain similar performance as the lower reference line, and it never achieves as high AUC as the upper reference line.

Performance among top 10 most frequent MedDRA SOC classes. The MedDRA vocabulary is divided into 27 non-mutually exclusive System Organ Classes (SOCs). These SOCs are further divided into High Level Group Term (HLGT), then into High Level Term (HLT), then Preferred Term (PT), and lastly Low Level Term (LLT). To investigate the performance of our proposed model, we grouped PT ADEs in our result into their corresponding SOCs and calculated AUC for the top 10 most frequent SOCs based on cumulated FAERS data.

The heatmaps in Fig. 6 and Fig. 7 demonstrate that all LP scores outperform baselines in both MGPS and BCPNN methods, with one exception: the SOC “Respiratory, Thoracic and Mediastinal disorders” using BCPNN. For MGPS, for MGPS, the SOC “Nervous System

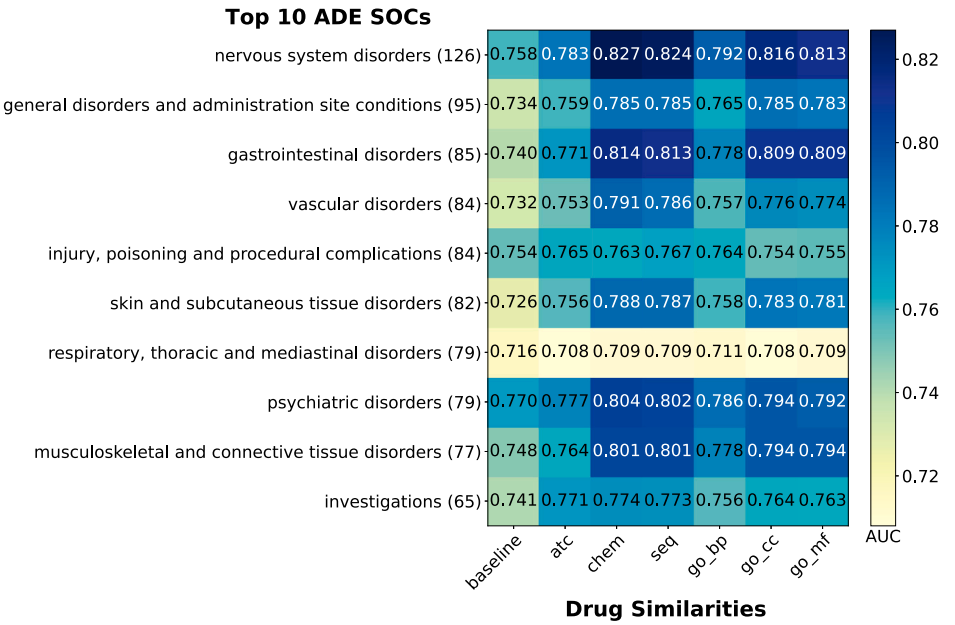

Fig. 7. Comparison of AUC achieved using the different methods based on BCPNN on top 10 most frequent MedDRA SOCs. (#) at the end of SOC names in the y-axes denotes the number of unique MedDRA PTs from our filtered SIDER dataset mapped to each SOC.

Disorders" achieves the highest AUC of 0.783 using LP-Chem. LP-Chem performs best for 6 out of 10 SOCs, while LP-ATC performs best for 3 SOCs. BCPNN shows similar patterns to MGPS, albeit with higher overall AUC scores. The SOC "Nervous System Disorders" again achieves the highest AUC of 0.827 using LP-Chem. LP-Chem performs best for 8 out of 10 SOCs, while LP-Seq performs best for 1 out of 10 SOCs.

These results highlight the consistent superiority of LP-Chem across both methods. "Respiratory, Thoracic and Mediastinal disorders" achieves the lowest AUC among the SOCs for all LP and baseline scores in both methods.

3.4. Website

We designed a website (http://drug-drug-sim.aimedlab.net/) to display our drug similarity scores obtained using the methods described in the methodologies section.

The home page, shown in Fig. 8, presents a user-friendly interface for initiating a drug search. Users can choose to search by either DrugBank ID (DB ID) or Drug Name. After that, users can select the drug of choice in the searchable drop-down list which dynamically updates based on the user input. The search results can be further refined by selecting required similarity matrices from the 6 options, i.e., ATC, Chem, Seq, Go-BP, Go-CC, and Go-MF. Users also have the option to specify the maximum number of results they wish to see.

Fig. 9 displays the result page of the website, which provides a comprehensive overview of the search results. In this example, a search was performed for the drug "DESMOPRESSIN" using the Drug Name as the input type. The search utilized the ATC, Chem, and Seq similarity matrices, with a maximum of 20 results requested. The results are presented in a clear, tabular format representing the selected similarity matrices. Each table is divided into 3 columns that displays the DB ID, Drug Name, and a corresponding Similarity Score. The drugs in the similarity tables are sorted in descending order of their similarity scores. Every DB ID in the result page provides a hyperlink to the corresponding page for that drug on the DrugBank website. Additionally, the question mark icon beside the similarity names provide the formula used to calculate those similarities.

4. Discussion

In this paper, we explore a label propagation based on an approach for early detection of adverse effect caused by drugs using drug similarities and post-market spontaneous adverse effect reports. We calculate six drug similarity metrics using various drug characteristics such as chemical structure, target protein sequence, target coding gene and phenotype. Additionally, we generate baseline drug safety signal scores from the FAERS quarterly reports using MGPS and BCPNN. After that, we employ the proposed label propagation framework to generate the enhanced drug safety signal scores using the drug similarities and baseline similarity scores. Lastly, we compare the performance of our enhanced drug safety signal scores with the baseline. We use SIDER side effects database as our positive controls. As demonstrated in the previous section our proposed method outperforms the baseline scores.

Statistical significance of model performance To demonstrate that the differences between the results of the proposed label propagation methods and the baseline methods are statistically significant, we conduct a paired t-test on the AUC scores of the models and calculate

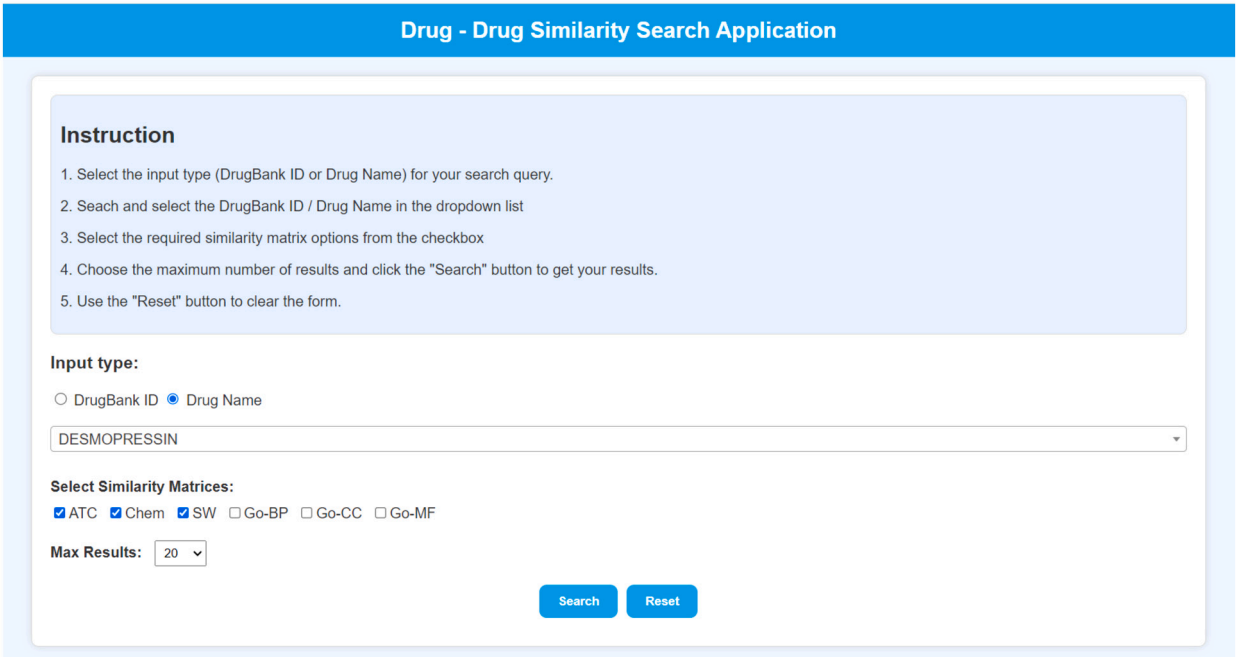

Fig. 8. Landing page of the website.

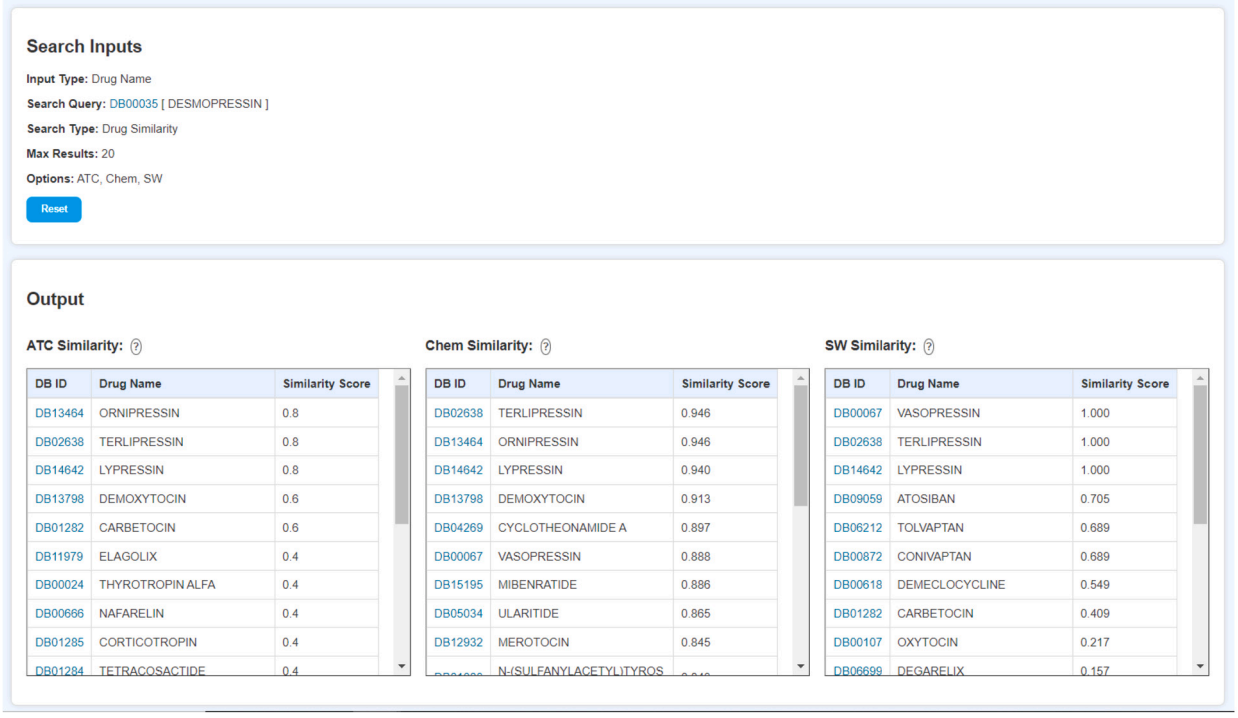

Fig. 9. Result page of the website.

the p-value. Table 3 shows the p-value of the six label propagation models with different drug similarities against their respective baselines. We consider 0.05 to be our threshold for this case. As all the p-values in Table 3 is less than the threshold improvement in performance achieved by the label propagation models are statistically significant.

Case study We would like to discuss two drug-ADE pairs documented by drug labels: pitavastatin - rhabdomyolysis (Fig. 10) and nitroprusside - hypotension (Fig. 11). The mechanism of pitavastatin-induced rhabdomyolysis remains unclear and may involve

Table 3
P-values for AUC comparisons between label propagation models against baseline methods.

Model	MGPS	BCPNN
LP-ATC	6.945×10^{-17}	5.299×10^{-23}
LP-Chem	1.125×10^{-20}	4.576×10^{-38}
LP-Seq	4.466×10^{-21}	1.979×10^{-38}
LP-GO_BP	1.111×10^{-17}	1.185×10^{-26}
LP-GO_CC	6.039×10^{-20}	9.025×10^{-35}
LP-GO_MF	9.950×10^{-20}	3.861×10^{-34}

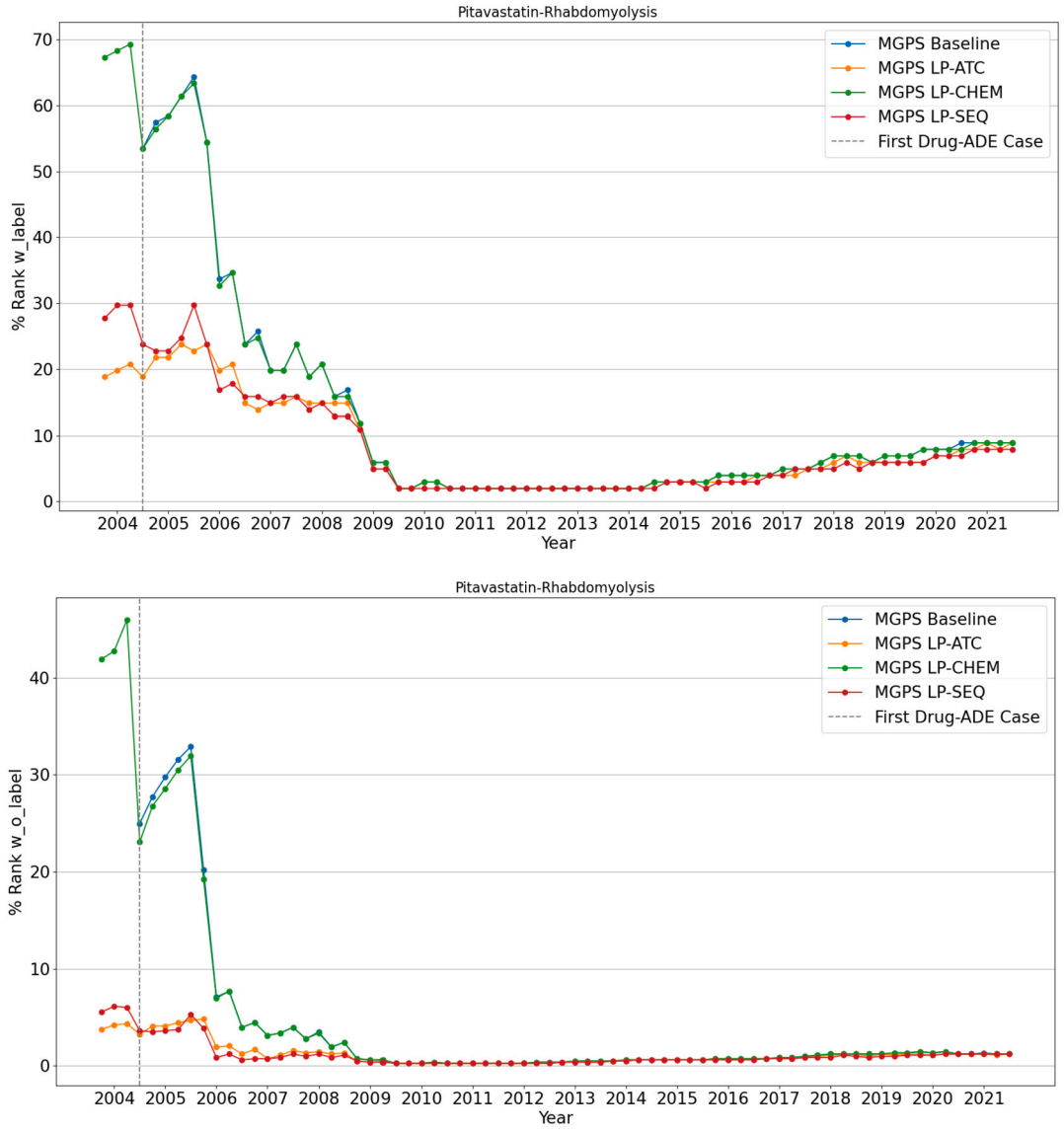

Fig. 10. Quarterly change in the rank of Pitavastatin-Rhabdomyolysis Drug-ADE pair.

complex mechanisms such as mitochondrial dysfunction or autoimmunity [23][24]. Nitroprusside can naturally trigger hypotension, as the drug acts as a blood pressure lowering agent [25]. In our analysis, we identify protein sequence-based similarity has the best performance on prioritizing pitavastatin - rhabdomyolysis and chemical structure-based similarity has the best performance on prioritizing nitroprusside - hypotension. Given these observations, we hypothesize that an ADE with complex mechanisms could be better characterized by protein sequence-based similarity, and an ADE related to the mechanism of action of the corresponding drug could be better characterized by chemical similarity. Further studies are warranted to investigate these hypotheses.

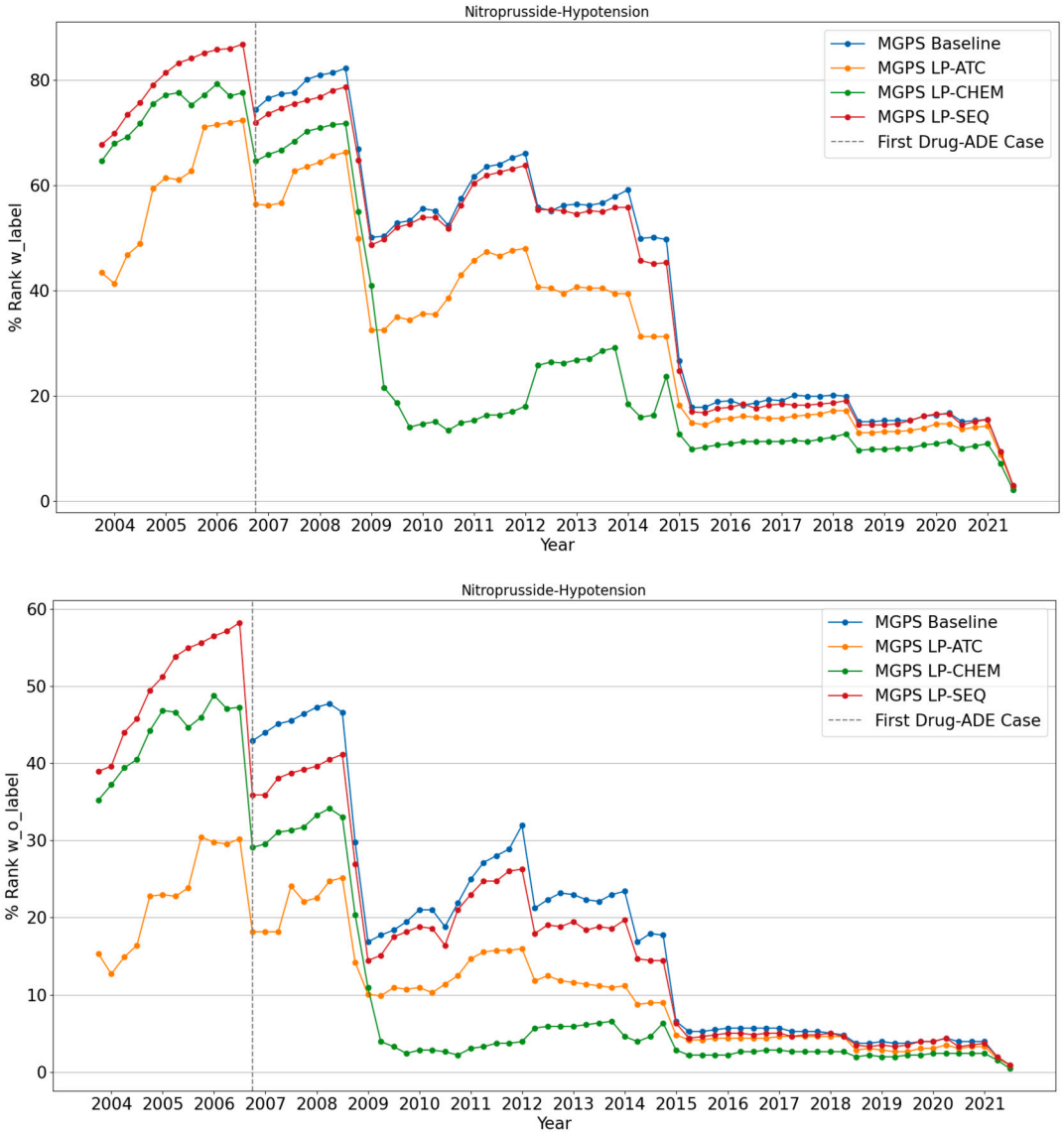

Fig. 11. Quarterly change in the rank of Nitroprusside-Hypotension Drug-ADE pair.

The quarter by quarter change in the ranks of these two drug-ADE pairs among all drugs from SIDER with and without labels are shown in Fig. 10 and Fig. 11. To calculate the rank, all drugs are divided into two categories based on their ground truth labels in relation to the target ADE. If a drug and target ADE pair exists in SIDER, the drug is classified as “with label”; otherwise, it is classified as “without label”. The “rank with label” for a drug-target ADE pair in a given quarter is determined by its position in a descending list of signal scores for all “with label” drug-target ADE pairs present in the FAERS data for that quarter. Similarly, the “rank without label” for a drug-target ADE pair in a given quarter is determined by its position in a descending list of signal scores for all “without label” drug-target ADE pairs present in the FAERS data for that quarter. These ranks are then normalized into a percentage by dividing it by the total number of drugs in their respective lists and multiplying by one hundred. The vertical dashed lines in the figures indicate the quarter when the first case of the specified drug-ADE pair was reported.

5. Methodology

5.1. Drug similarity network construction

We implement several drug similarity networks with different types of biological entities. For drug features, we use molecular structure and ATC code to calculate drug similarity. For drug-related target features, sequence and Go terms are utilized to measure the drug similarity.

Chemical structure similarity. We extract chemical structure information from the DrugBank database and the chemical structure similarity is constructed based on Pubchem fingerprint [26]. PubChem defines 881 chemical substructures, so an element can describe a drug in 881 dimensions, with each substructure having a value of 1 or 0, indicating its presence or absence. The chemical similarity of a drug-drug pair is calculated by the Jaccard Score using the vector form of chemical fingerprints.

$$S_{chem}(a,b) = Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

where A and B signify two distinct drug profiles. In our work, we compute the chemical structure similarity among all the drugs with small molecules from the Drugbank database. Then we got a matrix of size $11168 * 11168$.

ATC code-based similarity. ATC codes organize drugs hierarchically by organ or system on which they work, therapeutic impact, and chemical properties. All FDA-approved drugs have ATC codes that may be derived from the public database Drugbank. In this paper, we define the kth level ATC code based similarity S_k of drug a and drug b as

$$S_k(a,b) = \frac{ATC_k(a) \cap ATC_k(b)}{ATC_k(a) \cup ATC_k(b)} \tag{5}$$

where ATC_k represents all ATC codes from 1st to kth level. Then the similarity score $S_{atc}(i,j)$ is defined as follows:

$$S_{atc}(a,b) = \frac{\sum_{k=1}^{n} S_k(a,b)}{n} \tag{6}$$

where n represents the five levels of ATC codes and ranges from 1 to 5. In this paper, we totally computed similarities among 3224 drugs.

Sequence based similarity. We extract all target proteins for each drug from the Drugbank database. The similarity of drug targeted protein sets A and B for a given pair of drugs is calculated as follows:

$$S_{seq}(A,B) = \frac{\sum_{P_i \in A} \sum_{P_j \in B} \left\{ \begin{array}{c} \mathbb{1}\left[P_i = P_j\right] + \rho_{ij} \times \mathbb{1}\left[\left[P_i \neq P_j\right] \\ \wedge \left[\left(P_i, P_j\right) \notin (A \cap B)\right]\right] \end{array} \right\}}{\sum_{P_i \in A} \sum_{P_j \in B} \left\{ \begin{array}{c} \mathbb{1}\left[P_i = P_j\right] + \mathbb{1}\left[\left[P_i \neq P_j\right] \\ \wedge \left[\left(P_i, P_j\right) \notin (A \cap B)\right]\right] \end{array} \right\}} \tag{7}$$

Where P_i and P_j are the protein sequences from set A and B respectively, ρ_{ij} is the similarity between P_i and P_j and $\mathbb{1}[.]$ denotes the indicator function. ρ is calculated using R *Biostrings* package [27]. We got the similarity network of size $1897 * 1897$.

Go-term based similarity. All pharmacological target-coding genes have their Gene Ontology (GO) annotation [28]. Biological processes (BP), molecular function (MF), and cellular component (CC) are three categories of empirically proven or literature-driven evidence that we employ. Semantic comparison of GO annotations enables comparisons of genes and gene products. In this paper, we use an R package, named GOSemSim [29], to compute the similarity between two drug-targeted genes.

If A and B are two sets of genes targeted by a drug pair D_1 and D_2, respectively. The GO similarity of the drug pair is determined by Equation (8). Here ρ_{ij} is the similarity between genes G_i and G_j which is calculated using GOSemSim. For the three types, we calculate the similarity of 700 drugs.

$$S_{go}(A,B) = \frac{\sum_{G_i \in A} \sum_{G_j \in B} \left\{ \begin{array}{c} \mathbb{1}\left[G_i = G_j\right] + \rho_{ij} \times \mathbb{1}\left[\left[G_i \neq G_j\right] \\ \wedge \left[\left(G_i, G_j\right) \notin (A \cap B)\right]\right] \end{array} \right\}}{\sum_{G_i \in A} \sum_{G_j \in B} \left\{ \begin{array}{c} \mathbb{1}\left[G_i = G_j\right] + \mathbb{1}\left[\left[G_i \neq G_j\right] \\ \wedge \left[\left(G_i, G_j\right) \notin (A \cap B)\right]\right] \end{array} \right\}} \tag{8}$$

5.2. Regenerate drug safety signal scores

Inspired by the label propagation method [30], the original signals are propagated on each drug similarity network. We use MGPS and BCPNN as baseline methods for generating the original safety signals.

The initial label for the drug D_k is the kth row of the original signal scores matrix S, S_k. Then, we used the Bregmanian Bi-Stochastication algorithm to normalize the drug similarity matrix A, ensuring that both the row and column sums are equal to one.

As with the initial label propagation based on the graph's weighted edges, the second step recursively propagates labels from nodes with labels to nodes without labels using W. During each iteration, each node's label information is updated by absorbing labels from neighbors with a probability (γ) and maintaining prior labels with a probability ($1+\gamma$). From step $t-1$ to step t, the revised label information for a drug node i can be used as follows:

$$Y_i^t = \gamma W Y_i^{t-1} + (1-\gamma) S_i \tag{9}$$

where Y_i^t denotes the tth iteration's updated label information for drug node i.

After t iteration, for all nodes, the label information is denoted as:

$$Y^t = (\gamma W)^t S + (1-\gamma) \sum_{i=0}^{t-1} (\gamma W)^i S \tag{10}$$

Apparently, $\sum_{j=0}^{N} A_{i,j} = 1$, the spectral radius $\rho(W) \leq 1$, and $0 < \gamma < 1$. Therefore, we get the updated label information for drugs, which is denoted as:

$$Y = \lim_{t \to \infty} Y^t = (1-\gamma)(I - \gamma W)^{-1} S \tag{11}$$

Note that I is an identity matrix, and S is the matrix for initial label information. Actually, the coverage solution can also be written as follows:

$$J = \gamma \times tr(Y^T (I - W) S) + (1-\gamma)(I - \gamma W)^{-1} \| Y - S \|_F^2 \tag{12}$$

where $tr(\cdot)$ signifies the matrix's trace and $\| \cdot \| F$ denotes the matrix's Frobenius norm. The first term is a smoothness term, which indicates consistency throughout the intrinsic network, resulting in a smooth change in the enhanced signal score for each reference drug-ADE pair. This is consistent with our hypothesis that similar drugs have similar adverse drug reactions (ADEs). The second term is a fitting term that stipulates that the increased signal scores must be near to their initial values. The parameter γ is used to arbitrate between these two contradictory terms. Due to the fact that the formula for Y is convex, we can derive the global optimal solution by fixing the first derivative of J with respect to Y to 0, retaining $Y = (1-\gamma)(I - \gamma W)^{-1} S$.

As for the parameter γ, we run the label propagation with γ from $0.1, 0.2, ..., 0.9$. The final model is built with γ that yields the maximum AUC score. For a detailed comparison of performance for all γ values, refer to Table A1 and Table A2.

CRediT authorship contribution statement

Biswajit Padhi: Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Ruoqi Liu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Writing – original draft, Project administration, Writing – review & editing. **Yuedi Yang:** Data curation, Investigation, Methodology. **Xueqiao Peng:** Data curation, Writing – original draft. **Lang Li:** Conceptualization, Funding acquisition, Investigation, Writing – review & editing. **Pengyue Zhang:** Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – review & editing. **Ping Zhang:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Code availability

The source code for this paper can be downloaded from the GitHub repository at: https://github.com/BiswajitPadhi99/ADE-prediction-using-Drug-Similarities

Additional information

Correspondence and requests for materials should be addressed to PeZ and PiZ.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded in part by the National Institutes of Health (NIH) under award number R01GM141279. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e39728.

Data availability

The quarterly drug-ADE reports used in this work was downloaded from the FAERS website. The side effects data was collected from the SIDER website. Additional drug information was extracted from DrugBank. The six drug similarity matrices used in this paper are available at: https://zenodo.org/records/13270611. We performed further data processing to filter and prepare these datasets for our use.

References

[1] T. Eicher, et al., Metabolomics and multi-omics integration: a survey of computational methods and resources, Metabolites 10 (2020) 202.
[2] J. Lazarou, B.H. Pomeranz, P.N. Corey, Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies, JAMA 279 (1998) 1200–1205.
[3] K.M. Giacomini, et al., When good drugs go bad, Nature 446 (2007) 975–977, https://doi.org/10.1038/446975a.
[4] A.C. Rosen, et al., Impact of dermatologic adverse events on quality of life in 283 cancer patients: a questionnaire study in a dermatology referral clinic, Am. J. Clin. Dermatol. 14 (2013) 327–333.
[5] J. Sultana, P. Cutroneo, G. Trifirò, Clinical and economic burden of adverse drug reactions, J. Pharmacol Pharmacother. 4 (2013) S73.
[6] R. Harpaz, et al., Performance of pharmacovigilance signal-detection algorithms for the fda adverse event reporting system, Clin. Pharmacol. Ther. 93 (2013) 539–546.
[7] S.J. Evans, P.C. Waller, S. Davis, Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports, Pharmacoepidemiol. Drug Saf. 10 (2001) 483–486.
[8] K.J. Rothman, S. Lanes, S.T. Sacks, The reporting odds ratio and its advantages over the proportional reporting ratio, Pharmacoepidemiol. Drug Saf. 13 (2004) 519–523.
[9] P. Waller, E. Van Puijenbroek, A. Egberts, S. Evans, The reporting odds ratio versus the proportional reporting ratio: 'deuce', Pharmacoepidemiol. Drug Saf. 13 (2004) 525–526.
[10] A. Bate, et al., A Bayesian neural network method for adverse drug reaction signal generation, Eur. J. Clin. Pharmacol. 54 (1998) 315–321.
[11] W. DuMouchel, Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system, Am. Stat. 53 (1999) 177–190.
[12] S. Vilar, et al., Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis, J. Am. Med. Inform. Assoc. 18 (2011) i73–i80.
[13] S. Vilar, R. Harpaz, L. Santana, E. Uriarte, C. Friedman, Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis, PLoS ONE 7 (2012) e41471.
[14] S. Vilar, et al., Similarity-based modeling applied to signal detection in pharmacovigilance, CPT: Pharmacometr. Syst. Pharmacol. 3 (2014) 1–9.
[15] S. Vilar, N.P. Tatonetti, G. Hripcsak, 3d pharmacophoric similarity improves multi adverse drug event identification in pharmacovigilance, Sci. Rep. 5 (2015) 1–9.
[16] R. Liu, P. Zhang, Towards early detection of adverse drug reactions: combining pre-clinical drug structures and post-market safety reports, BMC Med. Inform. Decis. Mak. 19 (2019) 1–9.
[17] Fda's adverse event reporting system (faers), [EB/OL], https://open.fda.gov/data/faers/. (Accessed 30 June 2019).
[18] J.M. Banda, et al., A curated and standardized adverse drug event resource to accelerate drug safety research, Sci. Data 3 (2016) 1–11.
[19] E.G. Brown, L. Wood, S. Wood, The medical dictionary for regulatory activities (meddra), Drug Safety 20 (1999) 109–117.
[20] D.S. Wishart, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082.
[21] M. Kuhn, M. Campillos, I. Letunic, L.J. Jensen, P. Bork, A side effect resource to capture phenotypic effects of drugs, Mol. Syst. Biol. 6 (2010) 343.
[22] C. Xiao, Y. Li, I.M. Baytas, J. Zhou, F. Wang, An mcem framework for drug safety signal detection and combination from heterogeneous real world evidence, Sci. Rep. 8 (2018), https://doi.org/10.1038/s41598-018-19979-7.
[23] N. Safitri, M.F. Alaina, D.A.E. Pitaloka, R. Abdulah, A narrative review of statin-induced rhabdomyolysis: molecular mechanism, risk factors, and management, Drug Healthc. Patient Saf. 13 (2021) 211–219, https://doi.org/10.2147/DHPS.S333738.
[24] J. Bouitbir, G.M. Sanvee, M.V. Panajatovic, F. Singh, S. Krähenbühl, Mechanisms of statin-associated skeletal muscle-associated symptoms, Pharmacol. Res. 154 (2020) 104201, https://doi.org/10.1016/j.phrs.2019.03.010.
[25] D.G. Hottinger, D.S. Beebe, T. Kozhimannil, R.C. Prielipp, K.G. Belani, Sodium nitroprusside in 2014: a clinical concepts review, J. Anaesthesiol. Clin. Pharmacol. 30 (2014) 462–471, https://doi.org/10.4103/0970-9185.142799.
[26] Pubchem substructure fingerprint v1.3, [EB/OL], ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. (Accessed 30 June 2019).
[27] H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoy, Biostrings: efficient manipulation of biological strings, https://bioconductor.org/packages/Biostrings, 2024, R package version 2.72.1.
[28] T.G.O. Consortium, et al., The Gene Ontology knowledgebase in 2023, Genetics 224 (2023) iyad031, https://doi.org/10.1093/genetics/iyad031, https://academic.oup.com/genetics/article-pdf/224/1/iyad031/51074934/iyad031.pdf.
[29] G. Yu, et al., Gosemsim: an r package for measuring semantic similarity among go terms and gene products, Bioinformatics 26 (2010) 976–978.
[30] P. Zhang, F. Wang, J. Hu, R. Sorrentino, Label propagation prediction of drug-drug interactions based on clinical side effects, Sci. Rep. 5 (2015) 1–10.