Detecting Adverse Drug Events using Protein Sequence-Structure Similarity Networks

Saminur Islam
Comp. Sci & Electrical Engineering
West Virginia University

Ahmed Abbasi
Mendoza College of Business
University of Notre Dame

Nitin Agarwal
Department of Information Science
University of Arkansas at Little Rock

Wanhong Zheng
Behavioral Medicine & Psychiatry
West Virginia University School of Medicine

Gianfranco Doretto
Comp. Sci. & Electrical Engineering
West Virginia University

Donald A. Adjeroh
Comp. Sci. & Electrical Engineering
West Virginia University

***Abstract*—Adverse drug events represent a key challenge in public health, especially with respect to drug safety profiling and drug surveillance. Drug-drug interactions represent one of the most popular types of adverse drug events. Most computational approaches to this problem have used different types of data, such as drug chemical structure, information about protein targets, side effects, pathways, etc to predict potential interactions between drugs. In this work, we study the question of whether using just genetic information about the drugs can provide significant information about the potential safety profile for a given drug. We propose a novel neural network model to predict adverse drug events using only data about the protein sequence and protein structure associated with the drug targets. We compare the results with those from the state-of-the-art methods on this problem. Our results show that the proposed method is quite competitive, at times outperforming the state-of-the-art.**

***Index Terms*—Drug safety, Drug-Drug Interaction (DDI), Neural Networks, Protein Similarity Network**

I. Introduction

There are more opportunities today than ever before to learn about our health status and to take better care of our health. It has become very important to know about the drugs we take, especially given the increasing number of medications that are being consumed concurrently by an individual. With this increased potential for polypharmacy, there is a corresponding increase in the chance for adverse events involving medications. One key example of adverse drug events is the problem of drug-drug interactions. The sheer number of people taking more than one medication in a given day has made the issue of drug-drug interaction a major public health problem. This is the case, not only in the United States but across different countries globally.

Recent advances in biomedical research have generated a large volume of drug-related data. To effectively handle this enormous amount of data, many initiatives have been introduced to help researchers make sense of the massive data sets. As a result, various drug knowledge bases have been constructed. Among these, Drug Bank, SIDER, PDB, STITCH, SMILES are some of the popular data sources. These knowledge bases record various types of information about drugs, including information about genetic sequences, protein structures, drug side-effects, chemical structures, drug indications, etc. Thus, several approaches have been proffered to utilize the information from these different sources to predict potential interactions between drugs [11], [12], [16]–[18]. DrugBank is one of the most credible databases of known DDIs [6]–[8], and contains more than 300,000 DDIs. However, DDI's number of interactions is less than 1% of the total possible drug pairs that exist in DrugBank. Basically, DDI's are known as the unwanted side effects resulting from the concurrent consumption of two or more drugs [1]–[3]. When a doctor prescribes several drugs simultaneously for a patient, this may cause irreparable side effects. The effects of drugs on each other may lead to other illnesses or even death. These side effects are particularly noticeable in the elderly, or in persons with challenging diseases, such as cancer patients, who take many different drugs daily [4], [5]. Given the relevance of DDI's in an individual's health, and to public health in general, there is a critical need for more accurate and effective computational methods for understanding DDIs and how to predict them.

In this work, we study the problem of DDI prediction from the lens of genetic materials about the drugs. The paper is organized as follows: In the next section, we briefly discuss related work. Section III presents our methodology. Section IV reports on our experiments and results. Discussions are presented in Section V, and Section VI concludes the paper.

II. Background and Related Work

Most existing approaches for DDI prediction are based on different properties of the drug compound, such as its chemical structure, side effects, drug-target relationship, and many more. DDIs can be identified within Vivo models using high-throughput screening [9]. However, the price of such procedures is relatively high, and testing large numbers of drug combinations is not practical [10]. To reduce the number of possible drug combinations, numerous computational approaches have been proposed [11]–[18]. In some of these computational approaches, drug-target networks are constructed, and DDIs are detected by measuring the strength of network connections [15], or by identifying drug pairs that share drug

targets or drug pathways, for instance, using the random walk algorithm [16].

Other major categories of these computational approaches are based on the structural and side effect similarities of drug pairs. For example, Gottlieb et al. proposed the Inferring Drug Interactions (INDI) method, which predicts novel DDIs from chemical and side effect similarities of known DDIs [11]. Vilar et al. used similarities of fingerprints, target genes, and side effects of drug pairs [12], [13]. Cheng et al. constructed features from the Simplified Molecular-Input Line-Entry System (SMILES) data and side effect similarity of drug pairs and applied support vector machines to predict DDIs [14]. Zhang et al. constructed a network of drugs based on structural and side effect similarities and applied a label propagation algorithm to identify DDIs [15]. Recently, Ryu et al. proposed DeepDDI, a computational framework that calculates structural similarity profiles (SSP) of DDIs, reduces features using principal component analysis (PCA), and feeds them to a feed-forward deep neural network [18]. The platform generated 86 labeled pharmacological DDI effects, so DeepDDI [23] is basically a multi-classification (multi-label classification) model.

Vilar et al. developed a model to predict DDIs based on Interaction Profile Fingerprint (IPF) [19]. Quite simply, the interaction probability matrix was computed by multiplying the DDI matrix by the IPF matrix. Afterward, Lu et al. proposed a computational framework by applying matrix perturbation, based on the hypothesis that by randomly removing edges from the DDI network, the eigenvectors of the adjacency matrix of the network should not change significantly [20]. Unfortunately, these two methods employ no other data about drugs, except known DDIs.

More recently, a new family of similarity-driven methods has followed the assumption that similar drugs should have almost similar interactions. Vilar et al. [21] presented a neighbor recommender method by utilizing substructure similarity of drugs. Relying on Vilar's framework, Zhang et al. constructed a weighted similarity network that is labeled based on interaction with each of the drugs [22] and applied an integrative label propagation method using a random walk model on the network to estimate potential DDIs. This prediction framework only considered three types of similarities for predicting DDI via label propagation, namely substructure-based, side effect-based, and offside effect-based label propagation models [22]. Concerning the hypothesis that each type of drug data may assist in disclosing the patterns of interactions, a new inclination toward ensemble methods has emerged.

In this study, we develop a novel DDI prediction method utilizing the protein sequence data from DrugBank and protein structure data from Protein Data bank to calculate different similarity measures to create the similarity matrices for each feature attribute and then use the generated feature matrices to create a single network fusion to measure the potential for interaction between two drugs via the help of a neural network architecture based on multilayer perceptrons. Finally, the results will be compared with existing state-of-the-art machine learning algorithms as well as the recently proposed NDD algorithm by Narjes Rohani & Changiz Eslahchi [24]. The main novelty of our approach is the focus on only genetic materials (protein sequence and protein structures) associated with the drug targets in developing our prediction model. To our knowledge, this is the first attempt at investigating potential DDI prediction by utilizing only information about the protein sequence and structure to generate the feature space fed to the neural network.

III. Methodology

We developed a novel neural network model for the prediction of DDIs. The key idea in our approach is the assumption is that two drugs that have a similar pattern of similarity with other drugs are likely to have a similar pattern of interacting partners. To capture the patterns of similarity between drugs, we use information about the protein sequences and protein structures associated with the protein targets for a given drug.

A. Why protein Sequence and Protein structure information is important in DDI prediction?

DDI is a change in the way a drug acts in the body when taken with certain other drugs when taken with certain medical conditions. Drug interactions can make a drug more or less effective, or induce unexpected side effects in the body.

The primary mechanisms of DDI's are based on pharmacokinetics (PK) and pharmacodynamics (PD). Other than that there are drug structural similarity-based DDI prediction approaches that are being used [37].In pharmacokinetics, the study is mainly on what the body does to the drug. on the other hand, Pharmacodynamics describes the intensity of a drug effect concerning its concentration in body fluid, usually at the site of drug action. Now the question is why people never thought about utilizing the genetic information for DDI prediction? Despite several kinds of research on how genes influence drug response, genetic data has yet to be used to predict DDI. Most drugs do not work the same for everyone at metabolic levels. It can be computationally too expensive to predict side effects in terms of negativity or positivity (called adverse drug reactions). Due to the different genetic makeup of the human body, protein information is overlooked to develop computational methods that predict the potential for interactions between a pair of drugs.

We know that drug development is a complex time-consuming process. It needs years of testing before making it available for people. It is very challenging for chemists to find a promising molecule that could able to bind with a target protein. It is clear, drug molecules and protein targets have a critical relationship. Furthermore, we can see from the protein sequence and structural information that proteins are typically composed of independent evolutionary units. Those evolutionary units provide different functional clues to bind with drugs. This type of binding could occur on multiple evolutionary units. In addition, protein sequence/structure is crucial for recognizing the comparable functionality of proteins. So, data on protein sequence and protein structures could

direct to a possible solution for the DDI prediction problem. That's what we investigated in our work here.

Thus, we construct similarity matrices between drugs based on the protein sequences and protein secondary structures and combine these into one protein sequence-structure similarity matrix using network fusion. Fig. 2 shows a schematic diagram of the general proposed framework. To calculate the similarity matrices we have used cosine distance, Levenshtein distance, Jensen Shannon (JS) divergence, and Euclidean Distance as the similarity measure between a pair of drugs.

B. Distance Matrices

To estimate the how similarity between drugs, we compute distance measures and sometimes similarity measures) between drugs based on their protein sequences and structure. We used four such measures as described below.

1) Cosine Similarity (CS): Cosine similarity metric finds the normalized dot product of the two attributes. By determining the cosine similarity, we would effectively try to find the cosine of the angle between the two objects, when represented as vectors. The cosine of 0° is 1, and it is less than 1 for any other angle.

$$CS(A,B) = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two perpendicular vectors (at 90°) have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors.

2) Levenshtein Distance (L): The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions or deletions, but no substitutions) required to change one word into other. It is closely related to string alignments. The Levenshtein distance between two strings a, b (of lengths $|a|$ and $|b|$, respectively) is given by $L_{a,b}(|a|, |b|)$

$$L_{a,b}(i,j) = \begin{cases} max(i,j) & \text{if min(i,j) = 0} \\ min \begin{cases} L(i-1,j)+1 \\ L(i,j-1)+1 \\ L(i-1,j-1)+1 \end{cases} & \text{otherwise} \end{cases} \quad (2)$$

Essentially, $L_{a,b}(i,j)$ is the distance between the first i character of a and the first j character of b.

3) Jensen Shannon (JS) divergence(JSD): The Jensen–Shannon divergence is a method of measuring the similarity between two probability distributions. It is based on the Kullback–Leibler (KL) divergence, with some notable (and useful) differences, including that it is symmetric, and it always has a finite value. The square root of the Jensen–Shannon divergence is a metric often referred to as the Jensen-Shannon distance. The JS divergence is the average KL divergence of X and Y from their mixture distribution, M:

$$JS(X||Y) = \frac{1}{2} D(X||M) + \frac{1}{2} D(Y||M) \quad (3)$$

where $M = \frac{X+Y}{2}$.

C. Protein Sequence and Structure Similarity Matrices

In our work, we use similarity matrices, rather than distance matrices. Thus, for each distance measure, we convert the values into similarity measurement.

Each protein structure could have multiple chains. Moreover, each drug active ingredient could have multiple protein targets. Thus, we could compute the similarity between two drugs (or drug active ingredients) based on the protein chains associated with the respective protein targets for the drugs. Using the protein structure information from the Protein Data Bank (PDB), and the sequence information from DrugBank, we compute the similarity matrix between drugs, using the four distance/ similarity measures described earlier. For each similarity measure, we record four types of information, viz:

- Minimum Similarity
- Maximum Similarity
- Average Similarity(AS)
- Exponential Weighted Average Similarity(EWAS)

1) Protein Sequence Similarity Matrices: For protein sequences information we used two different approach to generate similarity matrices.

- Protein Sequence Similarity Matrix:- In this approach, Sequence information has been used directly to compute the Similarity matrix. we can compute the Levenshtein distance directly. To compute the cosine, and JS we will first compute the k-mer profiles for each sequence, and then compute the similarity measure based on the profile. To generate the k-mer profiles, we use the suffix array data structure [26].
- Protein Alignment Similarity Matrix:- Here, the sequence information had been direction used to calculate the alignment between a pair of protein sequences. There are two types of alignment normally used for calculating alignment between a pair of protein.
  - Global:- Global alignment is a type of alignment to find sequence alignment by taking entire sequence into consideration.
  - Local:- Local type is a sequence alignment by looking into the subset of the given sequences as well.

  There is one sequence alignment process which is very popular known as Point Accepted Mutation (PAM) which is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the process of natural selection. A PAM matrix is a matrix where each column and row represent one of the twenty standard amino acids.

In bioinformatics, PAM matrices are regularly used as substitution matrices to score sequence alignments for proteins. If we multiply the PAM matrix 250 times against itself we will find PAM250 matrix which frequently used to score the aligned peptide sequences of proteins with known homology and determining the accepted point mutation [38].

So, using the protein target mapping data, we will use the PAM250 matrix to calculate alignments for pairwise protein target sequences. We will get the alignment score and the align length from the function and the aligned sequences for the pairs. The similarity calculation from the alignment would be

$$S_{pp} = \frac{m-d}{m} \times 100\% \tag{4}$$

where S_{pp} represent the protein alignment similarity between protein pairs. m represent the matching score from the alignment and d represent the gap penalty between the two sequence.

2) Protein Structure Similarity Matrices: For protein structure, we first covert the protein 3D structure into a protein string (pString) representation following [35]. The resulting pString is then treated like a sequence of information for structure. The only difference with protein sequence is that each protein structure will have multiple chain(sequence) of information.

We generalized the similarity calculation which will represent the similarity values between two drug active ingredients (DAIs). We already know that each DAI could have multiple protein targets. Also, though each protein target has just one sequence, it could have multiple chains for its 3D structure. Thus, for a given DAI, we capture its protein structure information as follows:

$$[R_1^1, R_2^1 ... R_{K_1}^1, R_1^2, R_2^2 ... R_{K_2}^2 ... R_1^M, R_2^M, ... R_{K_M}^M]$$

where $R_i's$ represent the the protein targets, M denotes the total number of protein targets in the DAI, and $k_1, k_2 ... k_M$ represent the number of chains on each protein target.

Now, we can use this generalized DAI representation for similarity calculation between two DAI's. If two DAI's have N and M protein targets and their number of chains are $k_1, K_2, ...k_N$ and $L_1, L_2, ...L_M$ respectively, then the possible number of comparisons would be:

$$P_c = K_1L_1 + K_1L_2 + ...K_2L_1 + ... + K_ML_N \tag{5}$$

After P_c comparisons at the chain level, we will obtain a vector of similarity values for the two DAI's. We used the vector to calculate the minimum, maximum, average and exponential weighted average similarity between the two DAI's. The exponential weighted average is computed as follows:

$$w_i = \frac{e^{s_i}}{\sum_i (e^{s_i})} \tag{6}$$

Where w_i represents weights and s_i represents a similarity value. We use the flow diagram in Fig. 1 for better visualization of the computation, where we showed the calculation of similarity values between a pair of DAI's. The figure shows the calculation of minimum, maximum, average, and exponential weighted average can be performed between two DAIs, namely, A and B. For the protein sequence dataset, instead of multiple chains, kmers were calculated on single sequence information. The total number of comparisons between two DAI's would be the multiplication of the number of protein targets on each DAI. The remaining process for similarity values calculation was the same as protein 3D structures.

D. Protein Sequence-Structure Similarity Network

As noted above, using the protein sequence, we generate different similarity matrices. These are combined into one similarity matrix (sequence-based similarity network) using network fusion techniques. Similarly, we use the protein structure to generate different similarity matrices, which are also combined into one structure-based similarity network. Each of these networks can be used independently to analyze potential DDIs between drugs or drug active ingredients. To improve the overall performance, we then integrate the sequence-based similarity network with the protein-based network into one overall similarity network. The result is the protein sequence-structure similarity network (PS3N). Our network integration is based on the technique of Similarity Network Fusion(SNF) [25]. SNF is an approach for combining multiple data sources into a single graph representing sample relationships. The procedure works by constructing networks of these samples for each data source that represent how similar each sample is to all the others and then fusing the networks together. The similarity network generation and fusion process use a K-nearest neighbors procedure to down-weight weaker relationships between samples. However, weak relationships that are consistent across data sources are retained during the fusion process. The generated integrated network forms the basis for our analysis of adverse drug events, such as drug-drug interactions.

E. Neural Network Model

The model we are proposing for our problem is solely dependent on the datasets we are using. That means the neural network will have a dependency on the number of drugs we have been using for a dataset. We didn't use more than 4 hidden layers in our neural network model. We use Rectified Linear activation function (ReLU) as the activation function where the dropout rate for each layer would vary from 0.3 to 0.5. Each of the hidden layers is followed by a dropout layer to avoid over-fitting problems during the training of the model. The output of each neuron in a layer is a nonlinear function f of all nodes in the previous layer. f is the ReLU, which is defined as the positive part of its arguments $f(x) = x^* = \max\{x, 0\}$ The final output layer is calculated using the sigmoid function: $Sigmoid(x) = \frac{1}{1+e^{-x}}$

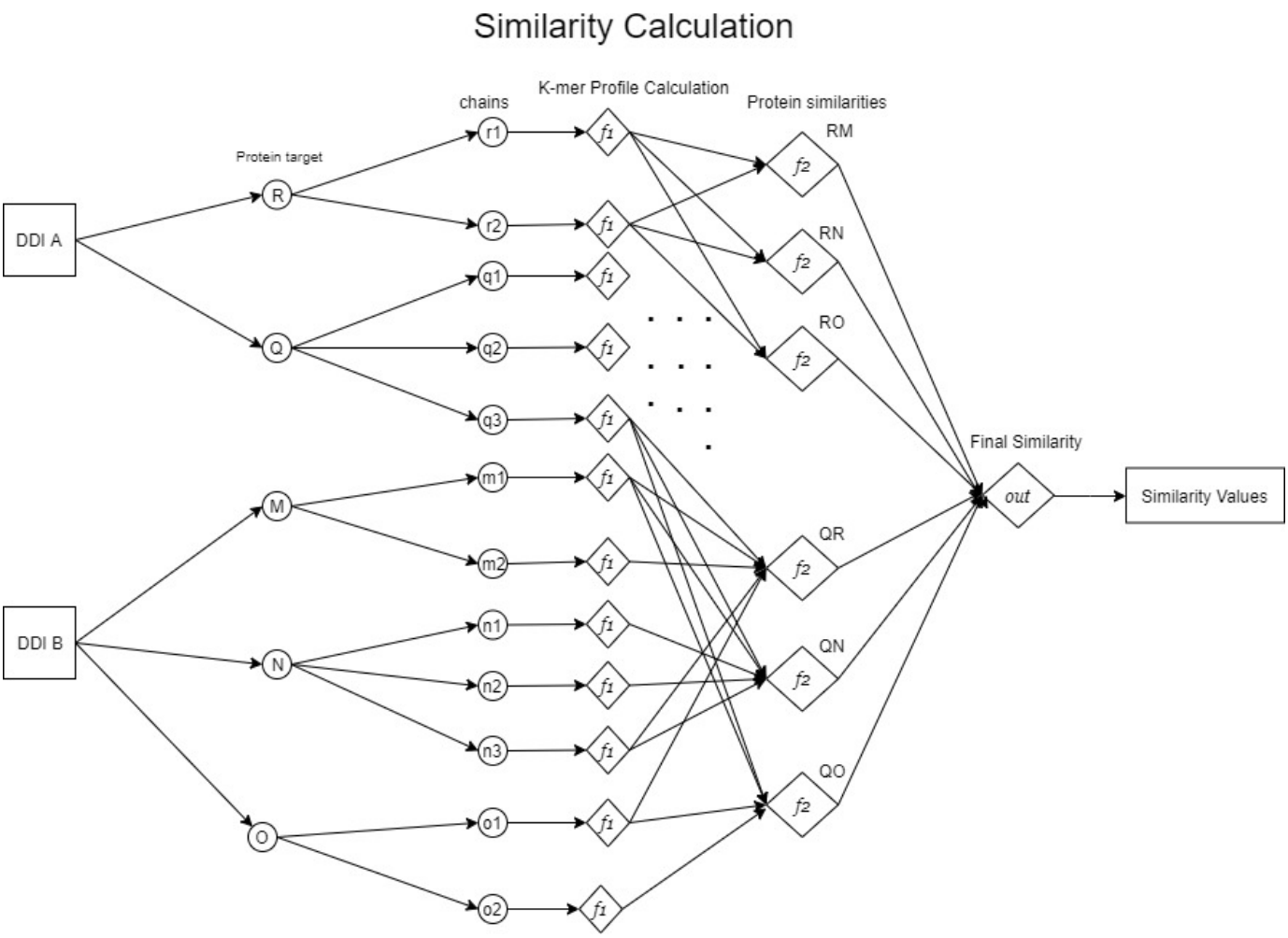

Fig. 1. Similarity calculation process between two drug active ingredients (DAIs). Here we give an example of two drugs having 2 and 4 protein targets, respectively. Each of the proteins can have a different number of chain sequences. At the first function sign, we calculate the 3-mers for the chains from each protein. Then we considered the protein similarity calculation based on the distance four metrics (cosine, Euclidean, JS divergence, Levenshtein). Finally, the minimum, maximum, average, and weighted average similarity values between the DAIs is computed using the output similarity values.

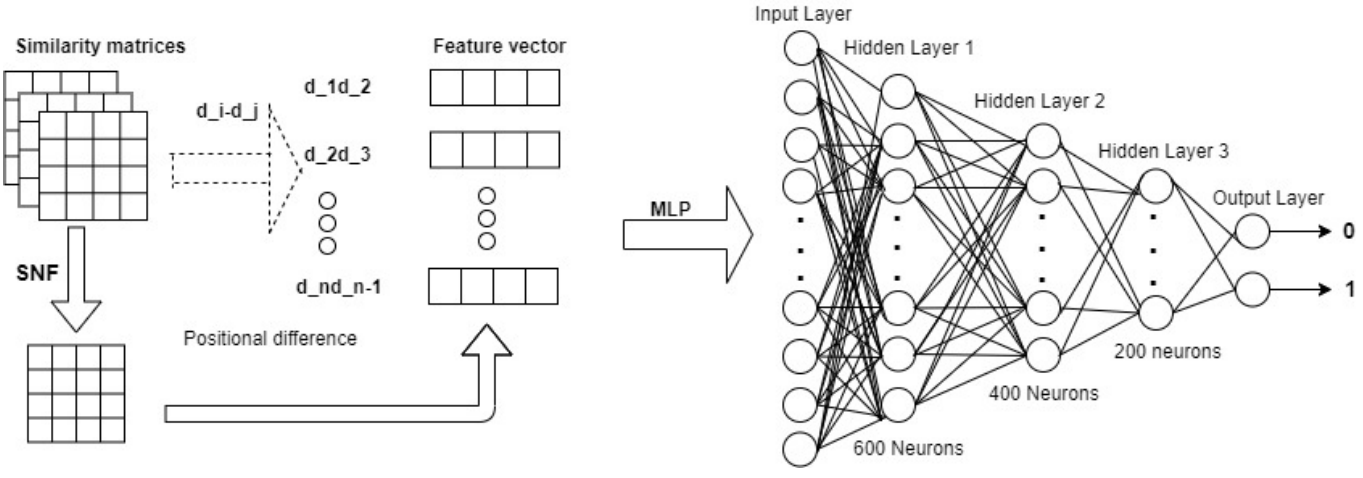

Fig. 2. Proposed Protein Sequence-Structure Similarity Network (PS3N) model for predicting adverse drug events. Using the method of Similarity Network Fusion (SNF) we create a single $N \times N$ fusion matrix for N drugs. From the fusion matrix, we compute the feature vectors for each pair of drugs. In this way we will have possible $\binom{N}{2}$ rows, and each row will have N columns as features. These feature vectors are then fed into a multi-layer perceptron model, where the first hidden layer would discarded for protein sequences.

For each layer, we used Xavier weight initialization, batch size of 100, and 20 - 50 epochs, with binary cross-entropy and stochastic gradient descent (SGD) for optimization. The momentum parameter was set at 0.9.

F. Performance Evaluation

To evaluate the performance of the proposed method, we compared it with machine learning approaches such as KNN (K Nearest Neighbor), RF (Random Forest), Logistic Regression, LDA (Linear Discriminant Analysis), and Support Vector Machine. We also compared our results with state of the art methods proposed in [24], [32], [33], [34]. We evaluated the competitiveness of our models using performance metrics like Precision, Recall, F1, Area under Curve (AUC), and AUPR.

We note that if the interaction between two drugs is assigned to zero, it simply implies that no evidence of their interaction has been found yet. The two may still interact, but the features we have we have used so far are not able to detect that.

IV. Experiments & Results

For our training experiment, we split each dataset into training, validation,and test sets according to a $70\% - 10\% - 20\%$ random split. For each dataset, networks were trained on the training set for a total of 100 epochs with a batch size of 100 for the proposed neural network method.

A. DataSets

In this work, we use two different protein datasets of which one is the protein sequence data with a relationship with drugs. This sequence dataset is extracted from the DrugBank. The second is the protein structure dataset retrieved from RCSB Protein Data Bank (https://www.rcsb.org). Protein chains are extracted from each PDB file using biopython libraries. We combined these for a dataset of 905 drugs (active ingredients) in DrugBank with information on both protein structure and protein sequences.

B. Results

First, we evaluated our model on the single feature matrices to identify the contribution of specific features to the performance of the model. We used the average and exponential weighted average similarity measures to generate the similarity matrices. Table I shows the performance of the proposed model using protein sequences. Table II shows results using protein structures.

We then compared our model performance with other state-of-the-art methods using the datasets defined by us. The constructed protein sequence and structure similarity matrices from the datasets we used proved a clear performance improvement. In Table III shows that our PS3N outperforms all the state-of-the-art methods. In the NDD method our dataset did not work well. So, we used datasets from NDD to evaluate the competitiveness with other state-of-the-art. However, the Table showed generally improved results for all learning models,and the impact of the data imbalance is evident, especially considering the recall and F-measures.

TABLE I
PERFORMANCE OF PS3N USING SIMILARITY MATRICES BASED ON PROTEIN SEQUENCES

Feature name	Precision	Recall	F-measure	AUC	Accuracy
L AS	0.9199	0.9419	0.9308	0.9673	0.9081
JSD AS	0.8837	0.8667	0.8751	0.9181	0.8377
C AS	0.8799	0.9093	0.8943	0.9315	0.8590
L EWAS	0.9499	0.9638	0.9568	0.9832	0.9429
JSD EWAS	0.9406	0.8835	0.9112	0.9559	0.8870
C EWAS	0.9523	0.9632	0.9578	0.9833	0.9443

*L = Levenshtein, JSD = JS Divergence, C = Cosine, AS = Average Similarity, EWAS = Exponential Weighted Average Similarity

TABLE II
PERFORMANCE OF PS3N USING SIMILARITY MATRICES BASED ON PROTEIN STRUCTURES.

Feature name	Precision	Recall	F-measure	AUC	Accuracy
JSDAS	0.8796	0.9279	0.90313	0.9171	0.8564
CAS	0.9037	0.9469	0.9248	0.9499	0.8889
JSD EWAS	0.9762	0.9650	0.9706	0.9895	0.9578
C EWAS	0.9743	0.9738	0.9741	0.9910	0.9627

*L = Levenshtein, JSD = JS Divergence, C = Cosine, AS = Average Similarity, EWAS = Exponential Weighted Average Similarity

On Table IV and V we compare our results with the existing state-of-the-art algorithms and found significant improvements in terms of AUC and Precision, Recall. We considered DS1 and DS2 datasets which were used in NDD to compare the performance of existing methodologies [24], From the two datasets we could generate the protein sequence and structure metrics for a subset of drugs. We used the newly generated feature space in our model to check the performance and showed significant improvement in both cases. In Table IV, we can see that the PS3N showed better performance compare to the others. However, it showed similar results on the datasets based on sequence, structure, or both information. This also holds in Table V which was created from the DS2 dataset.

C. Impact of algorithmic parameters

Table VI shows the impact of different hyperparameters on the performance of the proposed model. From the table, Adam Optimizer with a learning rate of 0.01 produced the best overall result. SGD Optimizer for learning rate 0.05, 0.01 and 0.10 showed almost similar accuracy level as we got for Adam optimizer. In our proposed neural network model, the

TABLE III
RESULTS USING COMBINED PROTEIN SEQUENCE AND PROTEIN STRUCTURE SIMILARITY MATRICES.

Method	Precision	Recall	F-measure	AUC	Accuracy
PS3N	0.9800	0.9818	0.9809	0.9946	0.9725
RF	0.7812	0.7241	0.7516	0.8089	0.8354
SVM	0.5507	0.2076	0.3015	0.5711	0.7320
LR	0.5169	0.1586	0.2427	0.5506	0.7243
LDA	0.5302	0.1740	0.2620	0.5572	0.7269
KNN	0.5470	0.6196	0.5810	0.7107	0.7510
Decision Tree	0.7134	0.7008	0.7071	0.7961	0.8382
NDD	0.5646	0.1927	0.2874	0.7366	0.7311

TABLE IV
PERFORMANCE COMPARISON OF ALL METHODS ON DS1 FROM NDD [24]. WE OBTAINED INFORMATION ON 469 DRUGS FOR PROTEIN SEQUENCES, AND ON 414 DRUGS FOR PROTEIN STRUCTURE. THE FIRST SIX ROWS ARE FROM [24] TO COMPARE THE RESULTS FROM OUR MODEL

Method	AUC	AUPR	F-measure	Recall	Precision
Substructure-based label propagation model [33]	0.937	0.901	0.804	0.797	0.811
Side-effect-based label propagation model [33]	0.936	0.903	0.806	0.793	0.820
Offside-effect-based label propagation model [33]	0.937	0.904	0.809	0.795	0.823
Vilar's substructure-based model [32]	0.936	0.902	0.804	0.797	0.812
Classifier ensemble method [34]	0.956	0.928	0.836	0.827	0.843
Weighted average ensemble method [34]	0.948	0.919	0.831	0.835	0.826
NDD [24]	0.954	0.922	0.835	0.836	0.833
PS3N (Protein Sequence)	**0.974**	0.948	0.916	0.925	**0.906**
PS3N (Protein Structure)	0.972	**0.949**	0.917	0.932	0.903
PS3N (Sequence + Structure)	0.972	0.948	0.917	**0.931**	0.903

TABLE V
PERFORMANCE COMPARISON OF ALL METHODS ON THE DS2 DATASET FROM NDD [24]. WE OBTAINED INFORMATION ON 585 DRUGS FOR PROTEIN SEQUENCES, AND ON 504 DRUGS FOR PROTEIN STRUCTURE. THE FIRST SIX ROWS ARE FROM [24] TO COMPARE THE RESULTS FROM OUR MODEL

Method	AUC	AUPR	F-measure	Recall	Precision
Substructure-based label propagation model [33]	0.788	0.208	0.294	0.537	0.197
Vilar's substructure-based model [32]	0.810	0.244	0.312	0.479	0.232
Classifier ensemble method [34]	0.936	0.487	0.553	0.689	0.462
Weighted average ensemble method [34]	0.646	0.440	0.15	0.226	0.118
NDD [24]	0.994	0.890	0.825	0.804	0.847
PS3N (Protein Sequence)	**0.998**	**0.975**	0.978	0.987	**0.972**
PS3N (Protein Structure)	0.997	**0.975**	0.978	**0.992**	0.964
PS3N (Sequence + Structure)	0.997	0.970	0.977	0.987	0.970

TABLE VI
RESULTS OF PS3N WITH VARIATION ON THE ALGORITHMIC PARAMETERS.

Optimizer	Learning rate	Accuracy
Adam optimizer	0.05	0.7200
Adam optimizer	0.10	0.7213
Adam optimizer	**0.01**	**0.9710**
SGD	0.05	0.9646
SGD	0.10	0.9644
SGD	0.01	0.9637
RMSProp	0.01	0.7210

number of hidden layers will vary baased on the number of drug active ingredients (DAI's) on the datasets. Normally, for protein sequence dataset, it will not more than 4. For Protein structure or the combination of both, it will be between 3 to 5.

V. DISCUSSION

The main objective of this work is to propose a new computational model for DDI prediction utilizing the genetic information. Our work has given a promising direction for addressing DDI prediction problems. We showed different ways of creating the feature space to identify the interaction between a pair of drugs. Roughly, we identified drugs with information on protein structures, and drugs with information on the protein sequence. We created the labeled the feature space by utilizing the interaction information available in DrugBank. The combination of the structure and sequence information resulted in 904 drugs. Our proposed methodology to address the DDI prediction creates a new dimension to the existing techniques. Unlike in the previous methodologies, we considered protein sequence and structure similarity networks for the first time to predict drug interactions. In addition, our similarity network computation technique allows extracting important protein features in terms of different distance measures.

The major drawback of our work is the lack of availability of protein structure level and sequence level information of the same drugs. As we mainly focused on Drugbank and Protein data bank. It was a challenge to find the commonality between the two different datasets. Moreover, the datasets have more unknown interactions than known interaction information. So classification could be an imbalance if don't consider cutoff of the unknowns. However, the time and space complexity for feature space generation is significant which might need to address in the future.

VI. CONCLUSION

In this study, we proposed a novel drug-drug interaction detection mechanism in which the proposed model is divided into three major chunks. One is contributed to creating the similarity profile from the drug bank and protein data bank. The second is the creation of the integrated similarity network (PS3N) about drugs, using information about both their protein sequences and protein structures. The third component is how information from the integrated network is used to develop a deep neural network model for improved prediction of the potential drug interactions. We compare the results produced using the proposed PS3N in deep learning framework with results from other recent machine learning-based approaches. The comparisons showed that our proposed methodology is quite competitive with respect to the state-of-the-art, and at

times outperforming the state-of-the-art methods. Though the computational complexity is high for the pre-processing, there are still opportunities to improve the performance of the model and also improve the datasets as well.

In our proposed methodology, we showed a new approach to dealing with the DDI prediction problem, by exploit only genetic information about the drug protein targets, in particular, information about their protein sequence and protein structure. Potential future work will be to study how the general approach could be extended to other adverse drug events, beyond DDIs. Another would be to see if the general approach can be adapted to use other types of feature attributes about the medications, or about interacting drugs.

REFERENCES

[1] Lazarou, J., Pomeranz, B. H. and Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama 279, 1200–1205 (1998).

[2] Prueksaritanont, T. et al. Drug–drug interaction studies: regulatory guidance and an industry perspective. The AAPS journal 15,629–645 (2013

[3] Kusuhara, H. How far should we go? Perspective of drug-drug interaction studies in drug development. Drug metabolism pharmacokinetics 29, 227–228 (2014).

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Wishart, D. S. et al. Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic acids research 46, D1074–D1082 (2017). 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Knox, C. et al. Drugbank 3.0: a comprehensive resource for omics research on drugs. Nucleic acids research 39, D1035–D1041 (2010)

[8] Law, V.et al. Drugbank 4.0: shedding new light on drug metabolism. Nucleic acids research 42, D1091–D1097 (2013)

[9] Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, Zhang C, Schnell C, Yang G, Zhang Y. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat Med. 2015; 21(11):1318.

[10] Fang H-B, Chen X, Pei X-Y, Grant S, Tan M. Experimental design and statistical analysis for three-drug combination studies. Stat Methods Med Res. 2017;26(3):1261–80.

[11] Gottlieb A, Stein GY, Oron Y, Ruppin E, Sharan R. INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol. 2012;8(1):592.

[12] Vilar S, Uriarte E, Santana L, Tatonetti NP, Friedman C. Detection of drugdrug interactions by modeling interaction profile fingerprints. PLoS One. 2013;8(3):e58321.

[13] Vilar S, Uriarte E, Santana L, Lorberbaum T, Hripcsak G, Friedman C, Tatonetti NP. Similarity-based modeling in large-scale prediction of drug-drug interactions. Nat Protoc. 2014;9(9):2147.

[14] Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Inform Assoc. 2014;21(e2):e278–86.

[15] Zhang P, Wang F, Hu J, Sorrentino R. Label propagation prediction of drugdrug interactions based on clinical side effects. Sci Rep. 2015;5:12339.

[16] Huang J, Niu C, Green CD, Yang L, Mei H, Han J-DJ. Systematic prediction of pharmacodynamic drug-drug interactions through protein-proteininteraction network. PLoS Comput Biol. 2013;9(3):e1002998.

[17] Park K, Kim D, Ha S, Lee D. Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks. PLoS One. 2015;10(10):e0140816.

[18] Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. Proc Natl Acad Sci. 2018;115(18):E4304–11.

[19] Vilar, S., Uriarte, E., Santana, L., Tatonetti, N. P. & Friedman, C. Detection of drug-drug interactions by modelling interaction profile fingerprints. PloS one 8, e58321 (2013).

[20] Lü, L., Pan, L., Zhou, T., Zhang, Y.-C. & Stanley, H. E. Toward link predictability of complex networks. Proc. Natl. Acad. Sci. 112, 2325–2330 (2015).

[21] Vilar, S. et al. Drug-drug interaction through molecular structure similarity analysis. J. Am. Med. Informatics Assoc. 19, 1066–1074 (2012).

[22] Zhang, P., Wang, F., Hu, J. & Sorrentino, R. Label propagation prediction of drug-drug interactions based on clinical side effects. Sci. reports 5, 12339 (2015).

[23] Geonhee Lee, Chihyun Park and Jaegyoon Ahn. Novel deep learning model for more accurate prediction of drug-drug interaction effects. BMC Bioinformatics(2019) 20:415

[24] Narjes Rohani & Changiz Esahchi. Drug-Drug interaction predicting by neural network Using integrated Similarity. Scientific Reports volume 9, Article number: 13645 (2019).

[25] Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. Nat. methods 11, 333 (2014).

[26] Donald Adjeroh, Timothy C.Bell, Amar Mukherjee. The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. Springer, 2008.

[27] Breiman, L. Random forests. Mach. learning 45, 5–32 (2001).

[28] Mitchell, T. M. Logistic regression. Mach. learning 10, 701 (2005).

[29] Freund, Y., Schapire, R. & Abe, N. A short introduction to boosting. Journal-Japanese Soc. For Artif. Intell. 14, 1612 (1999).

[30] Izenman, A. J. Linear discriminant analysis. In Modern multivariate statistical techniques, 237–280 (Springer, 2013).

[31] Lachenbruch, P. A. & Goldstein, M. Discriminant analysis. Biom. 69–85 (1979).

[32] Vilar, S., Uriarte, E., Santana, L., Tatonetti, N. P. & Friedman, C. Detection of drug-drug interactions by modelling interaction profle fngerprints. PloS one 8, e58321 (2013).

[33] Zhang, P., Wang, F., Hu, J. & Sorrentino, R. Label propagation prediction of drug-drug interactions based on clinical side efects. Sci. reports 5, 12339 (2015).

[34] Zhang, W. et al. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. BMC bioinformatics 18, 18 (2017)

[35] Jun Tan and Donald A. Adjeroh. Protein family structure signature for multidomain proteins. International Journal of data Mining and Bioinformatics 20(4)pages285-302(2018)

[36] Peterson, L. E. K-nearest neighbor. Scholarpedia 4, 1883 (2009).

[37] Takako Takeda, Ming Hao, teijun Cheng, Stephen H. Bryant and Yanli Wang. Predicting drug-drug interactions through drug structural similarities and interaction networks incorporating pharmacokinetics and pharmacodynamics knowledge. Journal of Cheminformatics, Article number 16 (2017).

[38] W. A Pearson, Rapid and Sensitive Sequence Comparison with FASTP and FASTA, in Methods in Enzymology, ed. R. Doolittle (ISBN 0-12-182084-X, Academic Press, San Diego) 183(1990)63-98.