DSIPHER: Drug SIde effect Prediction HEuRistic

Aashli Pathni[*]
Biological Sciences Graduate Program
apathni@umd.edu

Alexander Wikner[*]
Institute for Research in Electronics and Applied Physics
awikner1@umd.edu

Gowthami Somepalli[*]
Department of Computer Science
gowthami@umd.edu

Kamal Gupta[*]
Department of Computer Science
kampta@umd.edu

1 Introduction

Most of the drugs available in the market currently have been reported to cause one or more adverse drug reactions (ADRs, henceforth referred to as side effects). These side effects may or may not be reported on drug labels, package inserts or in publicly available documents. In the US alone, ADRs are the fourth leading cause of death every year [1]. Most drug discovery pipelines are shutdown in later stages when a serious side-effect has been found, resulting in wastage of human and financial capital. Even with extensive clinical trials, we might not be able to obtain a comprehensive list of side effects because sometimes the side effects are circumstantial, either the person has a phenotype which gets exacerbated due to the drug or the drug might be interacting with another medication taken by the patient and showing side-effects which might not happen otherwise.

Due to the availability of ample data aggregated in online databases, in silico side-effect prediction is being attempted in recent years. Early approaches have focused on drug substructure [2] or drug targets' position in signaling to predict the potential side-effects [3]. More recent papers [4] are using additional Chemical, Biological and Phenotypic features like chemical substructures, type of protein targets and information on other side-effects. Liu et al. [4] have built a binary classifier for each side-effect, considering all the known drug, side-effect combinations as positives and unknown ones as negatives. Zhang et al 2015. [5] have addressed this as a multi-label learning problem with ensemble learning done on kNN classification. Zhang et al. 2016 have used linear neighborhood similarity metric along with two different optimization techniques, similarity matrix integration method (LNSM-SMI) and cost minimization integration method (LNSM-CMI) to perform the kNN. One of the recent papers Zheng et al 2019 [6] proposes a new approach to sample negative drugs for a given side-effect. This method suggests to find the drugs structurally most dissimilar to the drugs causing a given side-effect, and take them as pseudo-negatives.

The aim of this project is to improve the prediction of side effect occurrence in response to drug treatments. We specifically aimed to predict whether a side effect would occur for a given drug. We hypothesized that given information about the set of proteins a given drug interacts with, we can predict the side-effect occurrence. We further speculate that these predictions can be improved given additional information about target proteins, like protein sequence or protein-protein interactions, and drug structural and functional information.

[*] Procrastinated equally on the project

Preprint. Under review.

In the process of predicting drug side effect occurrence, we assume that the databases we used to obtain protein-protein and drug-protein interaction information are complete and accurate. Additionally, we assumed that drug side effect interaction information is precise, that is, the reported side effect is caused by the drug alone and not due to a pre-existing condition in the patient or due to spurious reporting.

2 Our Approach

2.1 Target-Side Effect Distance Correlation

We first examine the simple hypothesis that if a drug targets similar proteins, then it will have similar side effects. This hypothesis was motivated by the observation in Cheng et. al. [7] that drugs that interacted with similar protein targets were more likely to cause additional side effects when used together, thus indicating a correlation between targeting of specific proteins and the occurrence of side effects. To investigate this hypothesis, we chose to look at the correlation between the "distance" between drugs in protein target space and in side effect space. We propose that if drugs are farther away from one another in proteins target space then they will be farther away from one another in side effect space, and thus have less similar side effects. We chose to explore this hypothesis using reduced versions of the the SIDER and OFFSIDES drug-side-effect data sets and the STITCH drug-protein interaction data set obtained from Zitnik et. al. [8].

We define a drug's representation in side effect space, $\mathbf{d}_{se}$ in the following way,

$$\mathbf{d}_{i,se} = \sum_{j=1}^{N} \delta_{ij}\mathbf{e}_j, \quad \delta_{ij} = 1 \text{ if drug } i \text{ causes side effect } j \text{ and is 0 otherwise} \tag{1}$$

Here, N is the number of distinct possible side effects and $\mathbf{e}_j$ is the j^{th} N-dimensional unit vector with 1 in the j^{th} and 0's everywhere else. Since the number of possible side effects present in our data set is very large, we apply principle component analysis (PCA) to the 639 drug-side-effect representations. From this, we obtain a lower-dimensional side effect space representation for each drug,

$$\tilde{\mathbf{d}}_{i,se} = \sum_{j=1}^{M} \alpha_{ij}\mathbf{p}_j \tag{2}$$

Here, M is the number of principle component vectors used to represent each drug, $\mathbf{p}_j$ is the j^{th} principle component, and α_{ij} is the projection of the $\mathbf{d}_i$ into the space spanned by $\mathbf{p}_j$. We choose M to be such that 90% of the variance in the original representations is contained in the reduced representations. This reduces the dimensionality of the space from $N = 10,184$ to $M = 436$. Given this representation, we calculate the Euclidean distance between two drug's in side effect space,

$$D_{ij,se} = ||\tilde{\mathbf{d}}_{i,se} - \tilde{\mathbf{d}}_{j,se})||_{L^2} \tag{3}$$

We used two different methods to determine the distance between drugs in protein target space. The first is exactly the same as the previously described method. In this case, we apply PCA to 1,774 drug-target representations that contain information on interactions (either 1 or 0 for present/not present) with 7,795 proteins. This reduces the dimensionality to only 26. We then define the drug-target space distance as:

$$D_{ij,t} = ||\tilde{\mathbf{d}}_{i,t} - \tilde{\mathbf{d}}_{j,t})||_{L^2} \tag{4}$$

The second method uses Eq. 2 in Cheng et. al. to determine the distance between drugs A and B given the average graph distance between proteins in the underlying protein-protein interaction network [7],

$$\mathrm{s_{AB}} \equiv \langle \mathrm{d_{AB}} \rangle - \frac{\langle \mathrm{d_{AA}} \rangle + \langle \mathrm{d_{BB}} \rangle}{2} \tag{5}$$

Having determined drug-drug distances using each of these methods, we compute the Spearman correlation coefficient between the distance metrics between the 283 drugs contained in both the drug-side-effect and drug-target data sets. Using the PCA method in drug-target space, we found a Spearman correlation of $\mathrm{R} = 0.12$, indicating that the drug-drug distances are positively correlates as we expect, but only very weakly so. Figure 1 demonstrates that for a given set of drugs with similar distances between one another in protein target space, their distances in side effect space take on a very wide range of values.

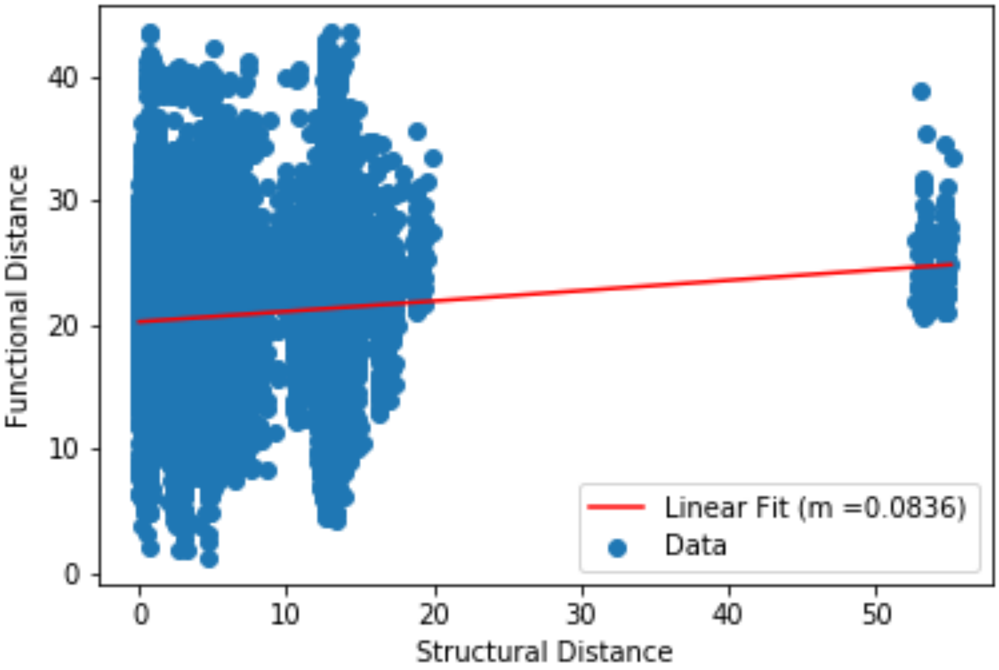

Figure 1: Plot of drug-drug distances in structural (protein target) and functional (side effect) space computed using equations 4 and 3 over 283 common drugs. A linear fit to the data results in a line with slope very close to zero, indicating that these values are not strongly correlated.

Moreover, Figure 2 shows that, if we zoom in to a small range of protein target distance values, we observe significant clustering in protein target space without observing similar clustering in side effect space. This indicates that while similar protein targets might be causing similar side effects, this phenomena likely does not depend on the large scale overlap between drugs and the proteins they target but on specific proteins. If we use Eq. 5 to determine distance in protein target space, we obtain similar results to what we show here. These results indicate that predicting what side effect a drug might cause will require a significantly more complex model.

2.2 Transformer Model

As discussed in Section 2.1, although drug interactions with protein give an indication of possible side effects that the drug may cause, a simple linear model such as PCA is not sufficient to discover features within drug-protein interactions that correlate well with drug-sideeffects. Intuitively we would want a model that use the information of all the proteins that a given drug interacts with, and use this information to predict side-effects that a drug may cause. Such a model needs to (1) be able to handle variable number of proteins different drugs interact with and (2) also be invariant to the order of these proteins. While LSTMs [9] are commonly used deep learning models used for tasks that take variable length sequence as input, they are not invariant to the order of input elements. We instead propose to use a modified version of Transformer [10] for our purpose.

Transformers [10] are models that take a sequence of discrete values as input and generate another sequence of discrete values as output. The inputs are typically modeled with a lookup table where each input in the sequence is mapped to a low-dimensional embedding. Let $S = \{s_i\}_{i=1}^{N}$ be the

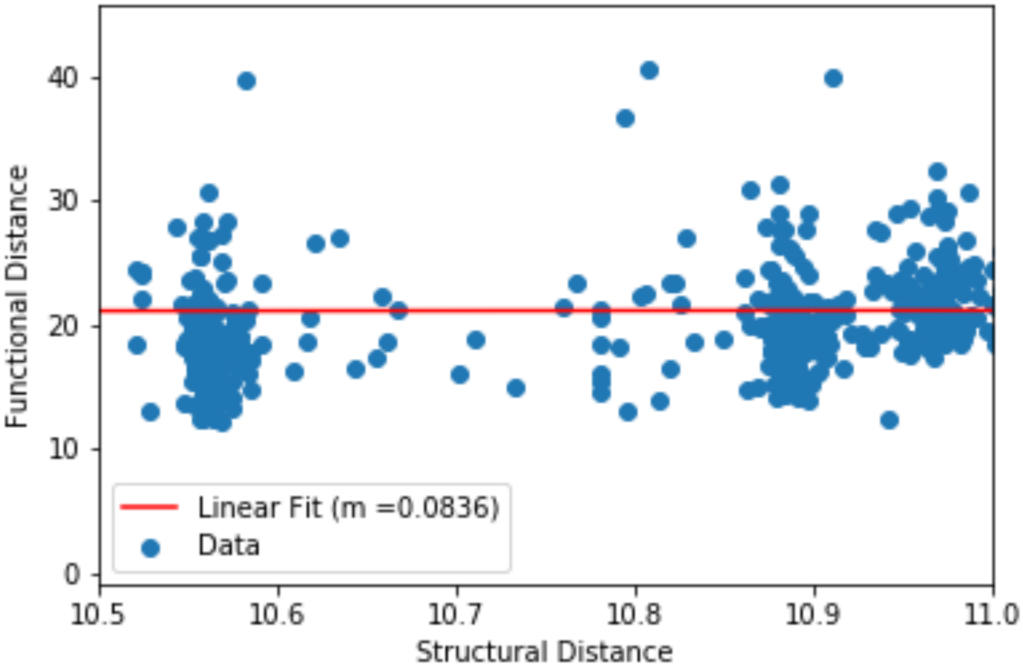

Figure 2: Zoomed in version of Figure 1. This plot shows significant clustering in target space with little clustering in side effect space.

vocabulary of discrete values in input sequences and $T = \{t_i\}_{i=1}^{M}$ be the vocabulary of discrete values in output sequences. We denote by $\phi_k = \phi(s_k)$ and $\theta_k = \theta(t_k)$ the lookup tables that map these discrete values to an embedding space. Given an input sequence of length n, $(s_1, s_2, ..., s_n)\,; s_i \in S$ and a target sequence of length m, $(t_1, t_2, ..., t_m)\,; t_i \in T$, the Transformer models the k^{th} value of the output sequence, as:

$$\begin{aligned} p(t_k) &= f_{\text{dec}}\left(f_{\text{enc}}(\phi_{1:n}), \theta_{1:k-1}\right) \\ p(t_1) &= f_{\text{dec}}\left(f_{\text{enc}}(\phi_{1:n}), \theta(t_{\text{bos}})\right) \\ p(t_{\text{eos}}) &= f_{\text{dec}}\left(f_{\text{enc}}(\phi_{1:n}), \theta_{1:m}\right) \end{aligned} \tag{6}$$

where t_{bos} and t_{eos} are two special tokens to represent beginning and end of sequence. Each of encoder f_{enc} and decoder f_{dec} have stacked self-attention based pointwise fully connected architectures. At the end of final decoder block, f_{dec} has a softmax layer to compute the probabilities of output tokens and minimize the cross-entropy loss.

We next talk about how we use this Transformer model for predicting drug side effects. Let $S = \{s_i\}_{i=1}^{L}, D = \{d_i\}_{i=1}^{M}, T = \{t_i\}_{i=1}^{N}$ be set of all the side effects, drugs and proteins respectively. The task of predicting side effects of a drug can be thought of as predicting for a given drug d_i and side-effect s_j pair, probability that d_i will call s_j. We represent this drug-sideeffect pair as a sequence $\{\langle\text{bos}\rangle, s_j, t_{d_{i,1}}, t_{d_{i,2}}, \ldots, t_{d_{i,k}}, \langle\text{eos}\rangle\}$, where $\{t_{d_{i,1}}, t_{d_{i,2}}, \ldots, t_{d_{i,k}}\}$ are set of k protein targets that given drug d_i interacts with. Our modified Transformer takes this input sequence and learns an encoding for each of the elements of input sequence. We removed the decoder as well as sinusoidal position encoding proposed in the original Transformer model to make the model invariant to the order of input sequence and predict only a single output. Figure 3 shows the final architecture of our model for a sample input.

2.3 Initializing Protein Embeddings

Protein Embeddings from Protein-Protein Interactions One method for initializing protein embeddings is to represent proteins using the representation obtained from factoring the N by N

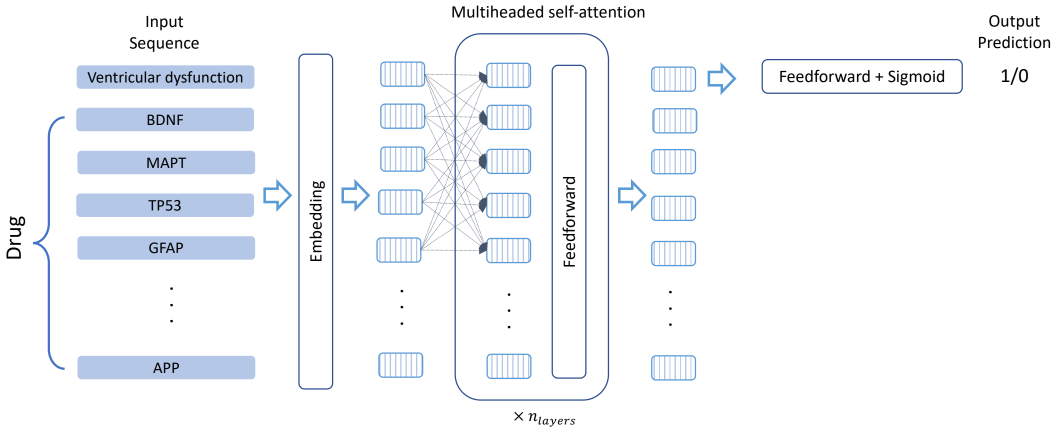

Figure 3: The architecture for our modified Transformer model depicted for a toy example. It takes a sequence of side-effect and set of proteins as input and predicts whether a drug interacting with this set of proteins will cause the side-effect or not. We pad each sequence with a special ⟨bos⟩ token in the beginning and ⟨eos⟩ token in the end.

symmetric protein-protein interaction matrix $\mathbf{R}$ from Szklarczyk et. al [11]. Since most protein-protein interactions are unmeasured, this reduces to a sparse matrix factorization problem that we solve using Probabilistic Matrix Factorization (PMF) implemented using the SurPRISE package [12, 13]. This method factors the full sparse matrix into the following form:

$$\hat{\mathbf{R}} = \mu + \mathbf{Q}^T\mathbf{P} \tag{7}$$

Here, μ is the mean matrix value and $\mathbf{Q}$ and $\mathbf{P}$ are the left and right D by N protein representation matrices. D is the representation dimension set at the beginning of the algorithm. PMF uses stochastic gradient descent to find $\mathbf{Q}$ and $\mathbf{P}$ that minimize the following cost function,

$$\sum_{r_{ij}\in\mathbf{R}} (r_{ij} - \hat{r}_{ij})^2 + \lambda(||\mathbf{Q}||_{L^2}^2 + ||\mathbf{P}||_{L^2}^2) \tag{8}$$

Here, λ is an L^2 regularization term. One can then extract protein representations from the columns of either $\mathbf{Q}$ or $\mathbf{P}$. For our purposes, we choose to extract protein representations from $\mathbf{P}$ and use those to initialize the embeddings in our model.

Protein Embeddings from Protein Sequences Another method for initializing embeddings is to obtain protein feature vectors from the protein's amino acid sequence. To do so, we use the UniRep mLSTM model [14]. This deep learning model has been trained using 24 million protein sequences to predict the next amino acid in a given sequence. In the process, UniRep learns a D-dimensional representation of the protein structure that can be obtained by averaging each of the hidden states of the mLSTM model given an input protein sequence. For our purposes, we use UniRep's 64 variable model to obtain protein embeddings from these averaged hidden states.

3 Experiments

3.1 Datasets

3.1.1 Protein-protein interactions: STRING

The STRING ('Search Tool for the Retrieval of Interacting Genes/Proteins')database records known and predicted protein-protein interactions. The reported interactions are obtained from genomic context predictions, experimental validation of protein-protein interactions, protein co-expression data, automated text mining and other databases [11].

STRING contains data for 19,353 human proteins with 11,759,455 interactions between them. Evidence for protein-protein interactions comes from the following associations:

- Conserved neighborhood: Proteins translated from genes that co-occur in the same neighborhood across different genomes
- Fusion: Fusion proteins resulting from gene fusion events
- Co-occurrence: Proteins with similar phylogenetic distributions of orthologs in a given organism
- Co-expression: Proteins with consistently similar expression patterns across different conditions
- Experiments: Experimentally reported protein interactions using biochemical, biophysical and genetic methods
- Textmining: Proteins frequently mentioned together in the same sentence, abstract or article in full-text articles
- Databases: Manually curated protein interactions reported in pathway databases

3.1.2 Drug-protein interactions: STITCH

STITCH ('Search Tool for Interacting Chemicals') aggregates drug-protein interaction information available across manually curated databases, pathway databases and experimentally validated databases [15].

STITCH contains drug-protein interaction data for 19,195 human proteins and 787,039 drugs. Drugs can interact with one of the following protein types:

- Target: The target protein to which the drug binds, resulting in a modification of target protein function. Drugs usually have either an agonistic or antagonistic effect on protein targets.
- Enzyme: The proteins involved in catalysis of chemical reactions in which the drug participates.
- Transporter: The cell membrane-bound proteins which regulate drug entry into the cell.
- Carrier: The proteins that bind to the drug and shuttle it to transporter proteins. Drug carriers may be used to increase the efficiency of drug delivery to the desired target cell or tissue.

There exists an overlap of 15,291 proteins between STITCH and STRING.

3.1.3 Drug-side effects: SIDER and OFFSIDES

SIDER SIDER contains drug-induced side effects reported in public documents and package inserts. SIDER contains 139,756 drug-side effect associations over 1,430 drugs and 5,868 side effects [16, 17].

OFFSIDES OFFSIDES compiles off-label drug-induced side effects reported by physicians and patients using the FDA Adverse Event Reporting System (FAERS). OFFSIDES contains 9,505,200 drug-side effect associations between 3,394 drugs and 17,552 side effects [18].

We compiled drug side effect data from SIDER and OFFSIDES for the purposes of our project. There exists an overlap of 2,424 drugs between the compiled drug side effect dataset and STITCH.

3.2 Baselines

Liu et al [4] have built and evaluated using a five-fold cross-validation on 832 drugs.1385 binary classifiers have been built one for each of the side-effects. One issue with the method is, in each of the side-effect classifiers, all the drugs which are not found to be causing the particular side-effect are taken as negatives. This is not a very valid assumption. Unknown relationships do not have necessarily have to be negative. Another issue is, in each of the classifiers, other side-effect information is taken as features. This is a form of information leakage and knowledge of other side-effects might give a complete picture of the occurrence of a side-effect.

The method proposed in Zheng et al [6] builds a similar classifier, with a difference in negative sampling. In this method, they propose to find the drugs dissimilar to drugs causing the side-effect in question and label them as pseudo-negatives for that side-effect. This is a novel approach, which we are planning to implement in our future experiments. Even this method has the issue of information leakage because they are using other side-effect information to capture the negative samples.

As you can see in Figure 4, knowledge of other side-effects will giveaway the presence of side-effects for a drug. Another issue with the current set of models is, they cannot predict side-effects for completely new off-the-market drugs since they need phenotypical feature information(i.e. presence of other side-effects).

In addition to the above two baseline methods, we have built a simple baseline neural network model. We have taken the known drug, side-affects as positives and rest as negatives and one-hot representation of all the proteins interacting with the drug, and side-effect represented as one-hot encoding of side-effect vector. Each drug, side-effect, feature vector is approximately 8000 dim vector. Using the protein, side-effect features and labels, we have built a 4 layer fully-connected neural network with 120 units in each layer optimized over binary cross-entropy loss.

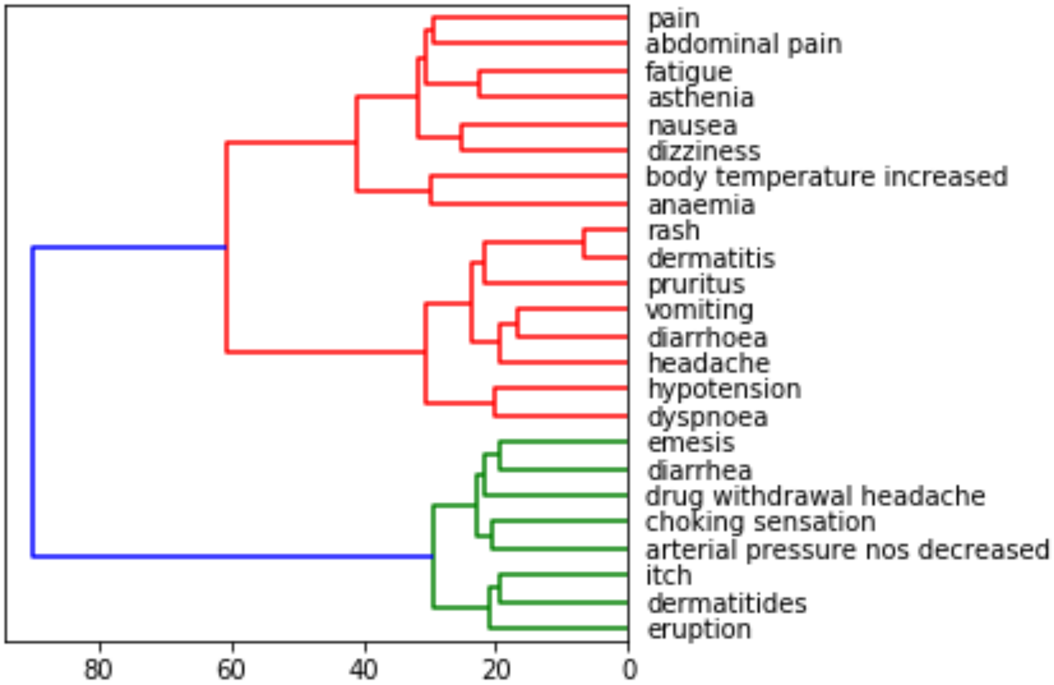

Figure 4: We have done the Hierarchichal clustering of top 30 side-affects, and as you can see, many side-effects are highly correlated with other side-effects.

3.3 Training procedure

Unlike Liu et al. [4], we don't want to use any information about drugs other than their protein interactions to make a prediction about their side-effects. To this end, we split our drugs in training and validation set in the 90-10 ratio. For all the drugs in our training and validation data, we consider positives training examples as those drug-side effect pairs which were reported in either SIDER or OFFSIDES datasets.

While positives examples are easy to obtain from empirical evidence, getting negatives is a bit of a challenge. In our work, for the drugs in the validation dataset, we sample random drug side-effect pairs that were not present in the positive set. We keep the number of negative examples same as the number of positive examples.

For sampling negatives in the training dataset, we try two different strategies. In the first strategy, for each positive drug-side effect pair, we take $n_{negative}$ random drug-side effect pairs that were not present in the positive set as negatives. In the second strategy, we generate $n_{negative}$ samples by

Table 1: Comparison

Side effect	Drugs	Negative sampling	AuPRC	AuROC
one-hot	binary	Random drug-se	0.5249	-
one-hot	(incremental) PCA	Random drug-se	-	-
learned (random init)	learned (random init)	Random protein	0.5332	-
learned (random init)	learned (from PPI), 32D	Random drug-se	0.5465	-
learned (random init)	learned (from PPI), 100D	Random drug-se	**0.5545**	-
learned (random init)	learned (from AA Sequence), 64D	Random drug-se	0.5417	-
learned (random init)	learned (random init)	Random drug-se	0.5539	-

randomly replacing 10% of the proteins that a drug interacts with, with a random protein from the database. We report results for both strategies in Section 3.4

3.4 Evaluation

3.4.1 Quantitative

We use Area under Precision-Recall Curve (AuPRC) on the validation set as the primary evaluation metric. Similar to the area under receiver operating characteristic curve (AuROC), AuPRC summarizes the precision-recall curve with a single number and is not based on various thresholds. Both AuPRC and AuROC are typically robust to huge data imbalances. We chose AuPRC over AuROC because a curve dominates in ROC space if and only if it dominates in PR space [19]. Table 1 shows performance of various models on our datasets.

3.4.2 Qualitative

This section outlines a few drug side effect combinations with previously published associations and predicted by our model with high confidence levels.

Omeprazole is a proton pump inhibitor that decreases the amount of acid produced in the stomach and is used to treat symptoms of gastroesophageal reflux disease (GERD). Proton pump inhibitors have been shown to cause kidney stones [20], which can develop into bladder stones [21]. Our model predicts that Omeprazole can cause bladder stones with 70% confidence.

Imatinib is a small molecule kinase inhibitor used in the treatment of multiple cancers. The anti-cancer effect of Imatinib has been shown to be, in part, due to its immunomodulatory and immunosuppressant function [22]. In accordance with this, our model predicts that Imatinib can cause a reduction in neutrophil count with 98% confidence.

Flumazenil is a drug used to reverse sedation in a patient after a medical procedure. Adverse events due to Flumazenil use include tachycardia, a type of arrhythmia [23]. Our model predicts tachycardia as a side effect associated with Flumazenil with 98% confidence.

Ebastine is an antihistamine used to treat allergic reactions such as hay fever. Ebastine reduces the production of GM-CSF (Granulocyte-macrophage colony-stimulating factor), which stimulates stem cells to produce granulocytes [24]. Our model predicts that Ebastine causes a decrease in granulocyte count with 98% confidence.

The suggested model has a potential to pin-point to which protein (of the ones drug is interacting with) is probably responsible for causation of a given side-effect. You can see in Figure 5,ADRA2C, which is found to be important in eyes has been given higher weight in later layers of neural network for a glaucoma side-effect instance.

4 Summary

Conclusions One of the remaining open questions is, how our model compares to other methods. We can check the performance of our model with other methods by using the same training and validation set presented by Liu et al. [4]. This dataset has around 800 drugs and 1300 side-effects.

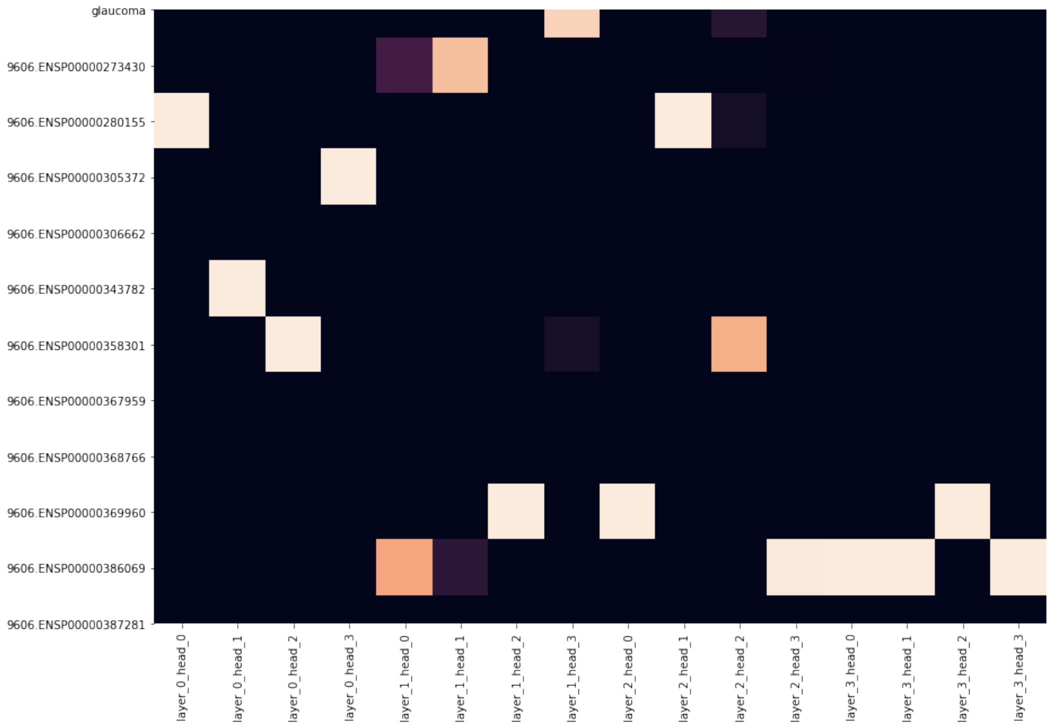

Figure 5: In this figure, you can see the attention visualization for one of the drugs and respective side-effect glaucoma. The protein represented by 9606.ENSP000000386069 is ADRA2C - Adrenoceptor Alpha 2C, a type of neurotransmitter. According to Genecards, this gene is highly expressed in certain types of eye cells. As you can see in the later layers on the NN, more emphasis is focused on this protein, compared to others, which makes sense because Glaucoma is an eye disease.

Over the course of our experiments, we found that information about drug-protein interactions and protein-protein interactions is useful but not sufficient to be able to predict side-effect occurrence for a drug with high accuracy. Exploration of our datasets revealed a few synonyms and spelling discrepancies in the side effect vocabulary that could be affecting the model. For instance, vomiting and emesis are synonyms of each other, diarrhea and diarrhoea are two spelling variations of the same word.

Future Work Future directions for the improvement of our model would involve including additional information such as intended drug effect, tissue-specific protein expression data [25], information about drug structure and mechanism of action [26] and Gene Ontology annotations for genes expressing target proteins [27].

Other methods of improving the model would include removing synonyms and spelling discrepancies in side effect vocabulary and initializing side effect embeddings from word models instead of random initialization.

"It is possible to commit no mistakes and still lose. That is not a weakness; that is life."

- Captain Jean-Luc Picard

References

[1] Janet Sultana, Paola Cutroneo, and Gianluca Trifirò. Clinical and economic burden of adverse drug reactions. *Journal of pharmacology & pharmacotherapeutics*, 4(Suppl1):S73, 2013.

[2] Anton F Fliri, William T Loging, Peter F Thadeio, and Robert A Volkmann. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nature chemical biology*, 1(7):389, 2005.

[3] MA Yildirim, KI Goh, ME Cusick, and AL Barabasi. Vidal marc. drug-target network. *Nat Biotechnol*, 25(10):1119–1126, 2007.

[4] Mei Liu, Yonghui Wu, Yukun Chen, Jingchun Sun, Zhongming Zhao, Xue-wen Chen, Michael Edwin Matheny, and Hua Xu. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association*, 19(e1):e28–e35, 2012.

[5] Wen Zhang, Feng Liu, Longqiang Luo, and Jingxia Zhang. Predicting drug side effects by multi-label learning and ensemble learning. *BMC bioinformatics*, 16(1):365, 2015.

[6] Yi Zheng, Hui Peng, Shameek Ghosh, Chaowang Lan, and Jinyan Li. Inverse similarity and reliable negative samples for drug side-effect prediction. *BMC bioinformatics*, 19(13):554, 2019.

[7] Feixiong Cheng, István A Kovács, and Albert-László Barabási. Network-based prediction of drug combinations. *Nature communications*, 10(1):1197, 2019.

[8] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[11] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2018.

[12] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[13] Nicolas Hug. Surprise, a Python library for recommender systems. `http://surpriselib.com`, 2017.

[14] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[15] Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2015.

[16] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2015.

[17] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1), 2010.

[18] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science translational medicine*, 4(125):125ra31–125ra31, 2012.

[19] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

[20] Tigran Makunts, Isaac V Cohen, Linda Awdishu, and Ruben Abagyan. Analysis of postmarketing safety data for proton-pump inhibitors reveals increased propensity for renal injury, electrolyte abnormalities, and nephrolithiasis. *Scientific reports*, 9(1):2282, 2019.

[21] Charles YC Pak. Kidney stones. *The lancet*, 351(9118):1797–1801, 1998.

[22] D Wolf, H Tilg, H Rumpold, G Gastl, and AM Wolf. The kinase inhibitor imatinib-an immunosuppressive drug? *Current cancer drug targets*, 7(3):251–258, 2007.

[23] Elisabeth I Penninga, Niels Graudal, Morten Bækbo Ladekarl, and Gesche Jürgens. Adverse events associated with flumazenil treatment for the management of suspected benzodiazepine intoxication–a systematic review with meta-analyses of randomised trials. *Basic & clinical pharmacology & toxicology*, 118(1):37–44, 2016.

[24] Alison Campbell, François-Bernard Michel, Clotilde Bremard-Oury, Louis Crampette, and Jean Bousquet. Overview of allergic mechanisms. *Drugs*, 52(1):15–19, 1996.

[25] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[26] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2017.

[27] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.