# Drug Target Interaction Prediction with Missing not at Random Labels

Anonymous Author(s)

## ABSTRACT

We introduce a novel probabilistic model, **F**actorization with **M**issing **N**ot at **R**andom **L**abels (FMNRL), for **D**rug-**T**arget **I**nteraction (DTI) prediction. Unlike previous studies which label unknown DTIs as negative samples, we treat the unnown DTIs as missing not at random responses. FMNRL assumes the labels are probabilistically generated from feature vectors of both drugs and targets, and a hidden matrix mapping from the drug features to target features. By associating the possibility of a missing response to the possibility of a negative label, FMNRL can learn the hidden feature space mapping better and thus provide more accurate DTI predictions. Experimental results show that FMNRL outperforms state-of-the-art methods.

## KEYWORDS

Missing Not At Random, Drug Target Interaction Prediction, Probabilistic Factor Models

**ACM Reference Format:**
Anonymous Author(s). 2018. Drug Target Interaction Prediction with Missing not at Random Labels. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

**D**rug-**T**arget **I**nteraction (DTI) is fundamental to drug discovery and design. As biochemical experimental methods for DTI identification are extremely costly and time-consuming, computational DTI prediction methods have received a growing popularity in literature. The mojority of existing computational methods treat DTI prediction as a binary classification task, where known DTIs are labeled as positive and unknown DTIs are labeled as negative [1]. To address the imbalanced problem arised from the binary classification scheme, many research works attempt to extract a subset of reliable negative samples, e.g. by random sampling [4] or by **U**nlabel **L**earning (PU Learning) [2].

Instead of labeling the unknown DTIs as negative, we argue that it is more natural to consider the unknown DTIs, i.e. DTIS that are neither identified in vivo to be positive nor experimentally validated to be negative (non-interacting drug-target pairs), as missing responses (i.e. the labels are missing). Our assumption in this work is that labels are not missing at random. This is an intuitive and reasonable assumption because researchers will use their domain expertise to filter DTIs with a high possiblity to be positive and priorize validations for these DTIs in vivo. For example, if researchers prefer to validate drugs with certain pharmacokinetic interactions, i.e. one drug affects the in vivo absorption, distribution, metabolism, or excretion of the target, then the unknown DTIs are not missing at random because drugs without pharmacokinetic interactions with the target are less likely to be positive and hence more likely to be missing.

### 1.1 Contribution

Our chief contribution in this work is a novel **F**actorization with **M**issing **N**ot at **R**andom **L**abels model (FMNRL). To the best of our knowledge, this is the first time missing not at random theory is applied in DTI identification. The inputs of FMNRL are feature vectors of drugs and targets which are leant and integrated from heterogenous data sources, the partially observed labels (i.e. positive or negative), and the fully observed reponses (i.e. given or missing). The FMNRL model mimics the probabilitic procedures to generate the labels from feature vectors and the responses from labels. Specifically, the labels are related to feature vectors of both drugs and targets, and a hidden matrix mapping from the drug features to target features. The possibility of giving a response is associated to the possibility of a positive label.

We conduct experiments on a large-scale DPI database. Our model achieves significantly better AUPR result and comparable AUROC result to the best of related works [4]. In the biomedical field, AUROC is considered to be a more robust and better assessment than AUROC [4]. Thus our improvment in AUPR is promising.

### 1.2 Related Work

One component of our work (i.e labels are generated by feature vectors learnt and fused from heterogenous information networks) is inspired by a recent work DTINet [4]. However, there are three key differences between our work and DTINet. (1) DTINet is based on deteministic matrix factorization, our work is based on probabilistic factor models. For example, the hidden feature space mapping matrix, labels, and responses are all random variables. This setting enables the FMNRL model to regulate the parameters (i.e. hidden feature space mapping matrix) by introducing approapriate priors and improves performance on sparse data set. (2) DTINet is based on randomly missing responses, i.e. it samples uniformly a set of unknown DTIs as negative sample, while FMNRL is based on missing not at random theories. Statistical theory in [3] shows that applying a model based on missing at random assumptions can lead to biased parameter estimation on data sets with missing not at random entries (3) DTINet adopts only a subset of unknown DTIs to preserve a balanced number of positive and negative samples, while our model uses all information in the data set.

We also want to distinguish our work with another line of research. Usually only positive DTIs are deposited in known databases. Due to the lack of negative samples, recently **P**ositive **U**nlabel **L**earning (PU Learning) is empolyed in DTI identification, e.g. to facilitate negative sample extraction [2]. PU learning does not explicitly associate the status of an instance (i.e. being positive or unlabel) with the value of its hidden label. We also want to mention

*Conference'17, July 2017, Washington, DC, USA*
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

here that, although we experiment with datasets where only positive DTIs are deposited, FMNRL is extendable without difficulty to databases where positive and negative DTIs are available. Thus our model is applicable in more scenarios.

## 2 THE PROPOSED METHOD

We start with the problem definitions and notations in Sec. 2.1. We then describe the proposed model FMNRL in Sec. 2.2. Finally we present the influence algorithm in Sec. 2.3.

### 2.1 Preliminaries

DTI identification is often modeled as a binary classification task. Formally, we are given $P \in \mathcal{R}^{N\times M}$ a set of DTI labels, where $p_{i,j} = 0$ indicates a negative interaction between drug $i$ and target $j$, $p_{i,j} = 1$ indicates a positive DTI, the feature vectors on drug side $X \in \mathcal{R}^{NtimesK}$, where $x_{i,k}$ represents drug $i$'s weight on drug feature $k$, the feature vectors on target side $Y \in \mathcal{R}^{M\times L}$, where $y_{j,l}$ represents target $j$'s weight on target feature $l$. The problem is to predict for a new drug-target pair $< i', j' >$, the possibility of a positive DTI $p(p_{i',j'} = 1)$.

It is worthy to note that many previous works have shown that extracting features $X, Y$ from heterogenous data sources are beneficial. Similar to DTINet [4], we use a compact feature expression learnt from aggregating diffusion probabilities on multiple information networks. For example, on the drug side, we can collect drug side-effect network, drug similarity network and so on. On the target side, we can collect protein desease network, protein protein network and so on. The feature learning phase is beyond the scope of this paper. We refer the readers to [4].

In addition to the features $X, Y$ and labels $P$, we make one essential modification to the problem definition. We assume that the inputs also contain responses $R \in \mathcal{R}^{N\times M}$, where $R_{i,j} = 0$ indicates that an unknown DTI, $R_{i,j} = 1$ indicates a verified DTI (positive or negative). For positive responses $R_{i,j} = 1$, the labels $P_{i,j}$ are observed. For negative responses $R_{i,j} = 1$, the labels are hidden and unknown.

### 2.2 Model

We use a factor model, depicted in Fig. ??. The features $X, Y$ are in different dimensions. To associate the drug features with the target features, we introduce a hidden variable $Z \in \mathcal{R}^{K\times L}$, where $Z_{k,l}$ is a projection that maps the drug feature dimension $k$ to the target feature dimension $l$. We assume that $Z$ is sampled from a Gaussian distribution,

$$\forall k, l, Z_{k,l} \sim \mathcal{N}((0, \sigma^2), \tag{1}$$

where $\sigma^2$ is the variance. We use zero mean to favor sparse feature mapping, i.e. a drug feature $k$ is associated with a few target features.

We then assume that the binary label $P_{i,j}$ is generated from the following process.

$$\forall i, j, p(P_{i,j} = 1) = \frac{1}{1 + \exp(XZY)_{i,j}} \tag{2}$$

The binary response is sampled from a Bernouli distribution. The parameters of the Bernouli distribution are related to the value of each $P_{i,j}$. Therefore we define $\beta_p \in \mathcal{R}^2, p \in \{0,1\}, \forall p, \beta_{p,0} > 0, \beta_{p,1} > 0, \beta_{p,0} + \beta_{p,1} = 1$, we have:

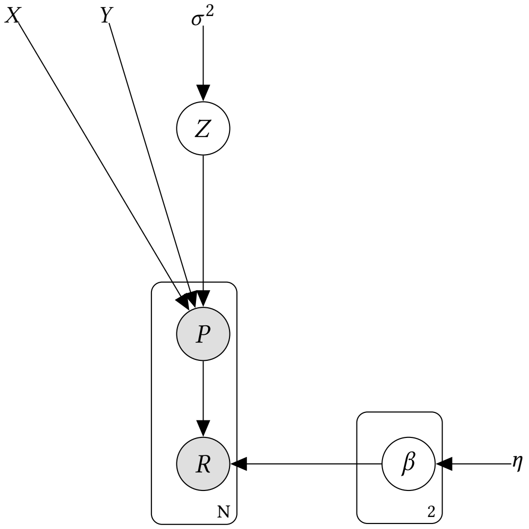

**Figure 1: Graphical Representation of the FMNRL model**

$$\forall p \in \{0,1\}, \beta_p \sim Beta(\eta), \tag{3}$$
$$\forall i, j, R_{i,j} \sim Bern(\beta_{P_{i,j},1}), \tag{4}$$

where $\eta$ is the hyperparameter for the Beta distribution.

### 2.3 Inference

The objective is to maximize the likelihood which consists of two terms. The first term is on partial observations, i.e. $R_{i,j} = 0$ and $P_{i,j}$ unknown. The second terms is on full observations, i.e.$R_{i,j} = 1$ and known $P_{i,j}$.

$$\mathcal{L} = \sum_{R_{i,j}=0} \sum_{P_{i,j}} \log p(R_{i,j}, P_{i,j}|X, Y, \sigma^2, \eta) + \sum_{R_{i,j}=1} \log p(R_{i,j}, P_{i,j}|X, Y, \sigma^2, \eta) \tag{5}$$

Direct optimization for both terms in Equ. 5 are intractable, as they involve integration over continuous hidden variable $Z$. We employ the variational inference algorithm to infer the parameters. To use the mean field variational inference, we must approximate the sigmoid function in Equ. 2 by the exponential distribution. We use the property that any sigmoid function $\sigma(\cdot)$ has a lower bound:

$$\sigma(z) \geq \sigma(\zeta) \exp(z - \zeta)/2 - \lambda(\zeta)(z^2 - \zeta^2), \tag{6}$$

where $\lambda(\zeta) = [\sigma(\zeta) - 1/2]/[2\zeta]$.

As shown in Alg. ??, in each iteration of the inference we alternatively optimize $Z, \beta$ and $\zeta$.

We defining $s_1 = \{(i,j) \in \mathcal{R}^{N\times M} | R_{i,j} = 1\}$, and $s_2 = \{(i,j) \in \mathcal{R}^{N\times M} | R_{i,j} = 0\}$.

We assume:

$$q(P, Z, \beta | R, X, Y, \sigma^2, \eta) = q(P|\theta) q(Z|\mu, \upsilon) q(\beta|\rho) \tag{7}$$

```
input  : P, R, X, Y
output : $\mu, \upsilon, \rho, \theta, \zeta$
initialization;
repeat
    for $Z_{k,l} \in Z$ do
        $\mu_{k,l} \leftarrow \frac{\sum_{(i,j)\in s_2}(\theta_{i,j}-\frac{1}{2})X_{i,k}*Y_{j,l}+\sum_{(i,j)\in s_1}\frac{1}{2}X_{i,k}*Y_{j,l}}{2*(\sum_{i,j}\lambda(\zeta_{i,j})X_{i,k}^2Y_{j,l}^2+\frac{1}{\sigma^2})}$;
        $\upsilon_{k,l} \leftarrow \frac{1}{\sqrt{2*(\sum_{i,j}\lambda(\zeta_{i,j})X_{i,k}^2Y_{j,l}^2+\frac{1}{\sigma^2})}}$;
    end
    for $\beta$ do
        $\rho_{0,0} \leftarrow \eta_{0,0}$;
        $\rho_{0,1} \leftarrow \sum_{(i,j)\in s_2}(1-\theta_{i,j}) + \eta_{0,1}$;
        $\rho_{1,0} \leftarrow |s_1| + \eta_{1,0}$;
        $\rho_{1,1} \leftarrow \sum_{(i,j)\in s_2}\theta_{i,j} + \eta_{1,1}$;
    end
    for $(i,j) \in s_2$ do
        $l_1 = exp(\psi(\rho_{1,1}) - \psi(\rho_{1,0}+\rho_{1,1}) + X_i\mu Y_j^T)$;
        $l_2 = exp(\psi(\rho_{0,1}) - \psi(\rho_{0,0}+\rho_{0,1}))$;
        $\theta_{i,j} \leftarrow \frac{l_1}{l_1+l_2}$;
    end
    for $(i,j) \in s_1 + s_2$ do
        $\zeta_{i,j} \leftarrow \left|X_i\mu Y_j^T\right|$;
    end
until convergence;
```

**Algorithm 1:** FMNRL algorithms

# 3 EXPERIMENT

## 3.1 Experimental Setup

**Datasets.** We use the same datasets as in [4]:, i.e. the drug-target interaction labels are obtained from the latest version of DrugBank (version 3.0); the feature vectors $X$ are extracted three heterogenous networks: the drug-drug interaction network, drug-disease network and the drug-side-effect network; the feature vectors $Y$ are extracted from three heterogenous networks, the protein-disease association network, protein-protein network and . As shown in Tab. **??**, only % of the drug-target interactions are labelled as positive, none is labelled as negative.

**Table 1: Statistics of the dataset**

| Data | #Drugs | #Targets | #Positive | #Negative | #Unknown |
|---|---|---|---|---|---|

**Comparative Methods.** We compare our FMNRL model with 5 state-of-the-art methods: (1) (2) (3) (4) (5)DTINet [4], which is a regression model that learns feature space mapping $Z$ by the loss function $\min_Z \sum_{i,j}(P_{i,j} - (XZY)_{i,j})^2$. We do not change the default settings for all the above comparative methods.

**Evaluation.** The major evaluation metric is (AUPR), which is commonly adopted in bioinformatic studies. AUPR computes... An auxiliary evaluation metric is (AUROC), which computes... As in [4], we perform two cross-validation tests. The first one is on the full data set, i.e. we randomly segment the whole data set to 5 divisions and conduct 5-fold cross-validation. The second one is on the sampled data set. We randomly sample the unknown drug-target interactions and construct a data set in which the ratio of positive and unknown samples is 1 : 10.

## 3.2 Results and Analysis

**DTI Prediction Performance.**

**Stability.**

**Parameter Tunning.** If there's more space, include the effects of parameters $\sigma^2, \eta$.

# 4 CONCLUSION

# REFERENCES

[1] Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. In *Briefings in bioinformatics*, 15(5), 734-747.

[2] Peng L, Zhu W, Liao B, et al. Screening drug-target interactions with positive-unlabeled learning. In *Scientific Reports*, 2017, 7(1): 8087.

[3] R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing Data, 1987.

[4] Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. In *Nature Communications*, 2017, 8(1).