InterHG: an Interpretable and Accurate Model for Hypothesis Generation

Haoyu Wang*, Xuan Wang†, Yaqing Wang*, Guangxu Xun‡, Kishlay Jha‡,Jing Gao*
*Purdue University,†University of Illinois at Urbana-Champaign,‡University of Virginia
{wang5346, wang5075, jinggao}@purdue.edu, xwang174@illinois.edu, {gx5bt, kj6ww}@virginia.edu

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669740

***Abstract*—Hypothesis generation, which tries to identify implicit associations between two concepts, has attracted much attention due to its ability of linking key concepts scattered in different articles and enriching plausible new hypotheses. Among existing approaches for hypothesis generation, matrix factorization based methods have achieved start-of-the-art performance. However, matrix factorization based methods suffer from the following limitations: 1) Bridge concepts are determined only as a post-hoc analysis of matrix factorization results; 2) The embeddings of concepts by matrix factorization cannot be explained, and thus it is hard to understand whether the concepts are linked in a semantically meaningful way. To overcome these limitations, we propose an interpretable and accurate hypothesis generation model (InterHG), which improves both accuracy and interpretability compared with existing methods. First, we propose to explicitly model the relationship between bridge concepts and given concept pairs, and conduct tensor factorization to identify link concepts. This reduces information loss and improves accuracy compared with post-hoc approaches. Second, we leverage the description of categories in the tensor factorization, which can output concept embedding as a weighted combination of known categories. With this meaningful embedding representation, medical researchers are able to check the correctness of the suggested link concepts for a given concept pair. We conduct experiments based on MeSH terms (a controlled vocabulary of biomedical concepts) extracted from MEDLINE corpus and category information obtained from UMLS (a comprehensive biomedical concept database). Results demonstrate that the proposed InterHG is highly accurate and produces meaningful embeddings for explanations.**

***Index Terms*—Hypothesis generation, Interpretation, Biomedical domain**

I. Introduction

Medical informatics [1] has become a prosperous field which aims to analyze the vast amounts of medical information such as medical literature and electronic health records. One important task that can advance medical discovery and benefit medical research is hypothesis generation based on existing medical literature. Given two medical concepts, one may want to find their implicit connections that are hidden in the vast medical corpus. An example is shown in Fig1, in which concept_i and concept_j are a given concept pair, and the goal of hypothesis generation is to find concepts in the middle that can bridge the given concept pair. Due to the vast volume of medical articles, it is impossible for medical researchers to query and evaluate such hypotheses manually. For example, given a concept pair *hypertension* and *diabetes*, there are 91,457 articles that mention both of them and numerous candidate concepts that may connect them. This challenge motivates the study on hypothesis generation, which automatically ranks bridge concepts based on their associations to the target concept pair shown in the medical corpus. The hypothesis generation tool thus can assist medical researchers in evaluating the probability of linking two concepts and exploring their connections further via the suggested paths.

Among existing approaches for hypothesis generation, matrix factorization (MF) based methods have shown start-of-the-art performance. In general, matrix factorization based methods adopt the following procedure: 1) Concepts are projected into a low-dimensional space based on the concept co-occurrence matrix via matrix factorization; 2) Given a pair of concepts, bridge concepts are ranked based on similarity between bridge concepts and target concepts in the embdedded space. It was shown that matrix factorization based approaches have achieved the best performance [2], [3] in the experiments that were conducted on MEDLINE[1], a major bibliographical database. In these experiments, Medical Subject Headings (MeSH) terms associated with each article are considered as co-occurred concepts.

However, MF methods suffer from two major limitations. First, the ranking of the bridge concepts is conducted as a post-hoc analysis of the matrix factorization results. The post-hoc analysis does not affect the matrix factorization process, and thus the matrix factorization results may not be optimal in modeling the implicit associations between target concepts. Ideally, this association should be modeled and optimized directly. Second, even though existing hypothesis generation methods can return the top bridge concepts for medical researchers to evaluate their hypothesis of the link between target concepts, it does not provide an explanation on why the concepts are ranked in this way. The ranking is decided based on the similarity defined in the embedded space achieved by matrix factorization, but it is difficult to interpret the meaning of the embedding. In fact, interpretability is crucial in medical informatics [4], [5], but it is missing in current hypothesis generation methods. If we look at MF methods used in broad domains including link predictions and recommendations, some efforts have been made to improve their interpretability. However, these methods cannot be used to interpret hypothesis generation. When explaining hypothesis generation results. It is important to leverage existing medical knowledge, such

[1]https://www.nlm.nih.gov/bsd/medline.html

as the rich ontology and descriptions of medical categories. It is impossible to incorporate such medical knowledge into post-hoc interpretation models [6]–[9]. Some approaches have been developed to leverage external knowledge graphs for the interpretation of link prediction [10]–[12], but they cannot be applied to hypothesis generation tasks because: 1) In biomedical knowledge graphs, a lot of concepts only have one or two neighbours, so it is nearly impossible to learn meaningful interpretation for those concepts. (2) Some popular concepts may connect to thousands of neighbours and it is hard to select appropriate paths for the interpretation.

To overcome these limitations of existing methods, we propose an **inter**pretable **h**ypothesis **g**eneration model (InterHG). We directly model the chance of a bridge concept connecting to the two target concepts as an entry in a tensor. Via tensor factorization, the proposed model is able to directly output the ranked list of bridge concepts given any pair of target concepts. We propose effective strategies to make the tensor factorization process efficient and scalable. To enable reasonable interpretation, we propose to leverage the descriptions of categories available in MEDLINE, which are maintained by subject-matter experts. From these descriptions, we can identify the relationship between categories and concepts, and such information is incorporated into the tensor factorization objective such that concept embeddings can be represented as a weighted combination of categories. By checking these weights, medical researchers are able to verify the correctness of the bridge concept embedding and the plausibility of the hypothesis formed by the target concept pair.

The contributions in this paper are summarized as following four points. 1) We propose InterHG, a novel method that greatly improves both the accuracy and interpretability of hypothesis generation. 2) We propose to model hypothesis generation as a tensor factorization task to directly optimize the output bridge concepts. We further propose effective strategies to reduce the complexity of the tensor factorization solution. 3) To the best of our knowledge, our work is the first to introduce interpretability into the design of a hypothesis generation model. We leverage category descriptions as external knowledge and enable a reasonable interpretation of the output concept embedding. 4) We conduct qualitative study to show the interpretability of the proposed InterHG model. We also conduct quantitative experiments which show that InterHG has higher accuracy compared with state-of-the-art algorithms.

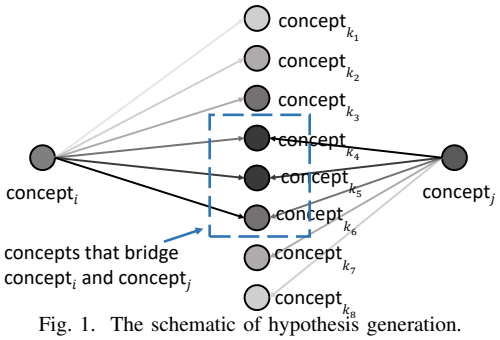

Fig. 1. The schematic of hypothesis generation.

II. METHODOLOGY

A. Preliminaries

In this section, we get started with the preliminaries including problem formulation and one state-of-the-art hypothesis generation method. Throughout the paper, we denote vectors by boldfaced lower- case letters and matrices by boldfaced uppercase letters. All vectors are considered as column vectors.

1) Problem Formulation: Suppose there is a concept set $C = \{concept_1, concept_2, \cdots, concept_n\}$, and we know the number of times the concepts c_i and c_j co-occurs within a document over the corpus $\mathcal{D}$ at time t. $\boldsymbol{R}$ denotes the co-occurrence matrix, where $\boldsymbol{R}_{ij} = \#(c_i, c_j)$ meaning the times c_i and c_j co-occurs at time t.

Hypothesis Generation: Given co-occurrence matrix $\boldsymbol{R}$ and two concepts c_i and c_j at time t, the goal is to predict the top-k concepts most likely co-occur with the given two concepts at time $t+1$.

Interpretation for Hypothesis Generation: To complete the hypothesis generation task, designing an interpretable model to generate top-k concepts, and the reason why generates those concepts can be understood by **medical researchers**.

2) Matrix Factorization for Hypothesis Generation: The method, a state-of-the-art model in [3], is inspired by the word embedding model GloVe [13], which describes the association between two given concepts and one other bridge term via their co-occurrence probabilities. Intuitively, it predicts the logarithmic co-occurrence times via matrix factorization and then computes cosine similarity between concepts' embeddings to generate hypotheses. Formally, it predicts the logarithmic co-occurrence times according to the following equation:

$$\log(\boldsymbol{R}_{ik}) = \boldsymbol{u}_i^T \boldsymbol{v}_k + b_i + \bar{b}_k \tag{1}$$

where $\boldsymbol{u}_i$ is the ith concept's embedding when the concept works as a target term and $\boldsymbol{v}_k$ is the kth concept's embedding when the concept works as a bridge term, and b_i, $\bar{b}_k$ are bias terms. Therefore, the martix factorization-based loss function can be written as

$$\begin{aligned}\arg\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{(i,j)\in\mathcal{D}^+} & f(\boldsymbol{R}_{ij})(\log(\boldsymbol{R}_{ij}) - \boldsymbol{u}_i^T \boldsymbol{v}_j + b_i + \bar{b}_j)^2 \\ & + \lambda(||\boldsymbol{U}||_F^2 + ||\boldsymbol{V}||_F^2)\end{aligned} \tag{2}$$

where $\mathcal{D}^+ = \{(i,j)|\boldsymbol{R}_{ij} > 0\}$ and $f(x) = (x/x_{max})^\alpha$.

However, observing its loss function, we can find that it tends to model hypothesis generation via a circuitous approach, which complements the co-occurrence matrix $\boldsymbol{R}$ firstly and then completes hypothesis generation task. Therefore, the loss function is not related to our task directly, which may result in the following problems: making the task more difficult to solve. Take binary classification as an example. If we can design a perfect and ideal regression model, it can handle the classification task in theory, but designing a perfect regression model is extremely difficult. Similarly, if the completion task is solved greatly, hypothesis generation task can be solved as well. However, it increases the difficulty of the task.

B. Tensor-based Loss Function for Hypothesis Generation

Recall the matrix factorization-based method in section II-A2. It solves the task via two stages and increases the task's difficulty. To overcome the problem, we design a tensor-based loss function to model the problem directly:

$$\begin{aligned}\mathcal{L}_{Tensor} &= \mathcal{L}_{reconstruction} + \lambda\mathcal{L}_{reg} \\ &= \sum_{(i,j,k)\in\mathcal{D}^+_{tensor}} (\mathcal{T}_{ijk} - \boldsymbol{u}_i^T\boldsymbol{u}_k - \boldsymbol{u}_j^T\boldsymbol{u}_k)^2 + \lambda||\boldsymbol{U}||_F^2\end{aligned}$$

where $\mathcal{T}$ is a tensor defined as $\mathcal{T}_{ijk} = \frac{\boldsymbol{R}_{ik}+\boldsymbol{R}_{jk}}{\sum_{p=1}^n \boldsymbol{R}_{pk}}$, and $\mathcal{D}^+_{tensor} = \{(i,j,k)|\boldsymbol{R}_{ik}+\boldsymbol{R}_{jk} > 0\}$. **Intuitively, the tensor $\mathcal{T}$ describes the probability that the concept c_k both has connection with concept c_i and c_j.** Therefore, predicting the value of $\mathcal{T}_{ijk}$ is clearly targeted at hypothesis generation task. However, different from other tensor factorization-based tasks like tag recommendation, the scale of $\mathcal{T}$ here is much greater than that in other task, resulting in computational bottleneck. Proposition 1 illustrates the fact.

Proposition 1: Denote the number of non-zero elements in $\boldsymbol{R}$ is $||\boldsymbol{R}||_0$. Then the number of non-zero elements in $\mathcal{T}$ satisfies the following inequation:

$$||\mathcal{T}||_0 \geq n||\boldsymbol{R}||_0^2$$

Proof 1: We use o_i denoting the number of non-zero elements in the ith row of $\boldsymbol{R}$. Then $||\boldsymbol{R}||_0$ can be represented by o_i as follows: $||\boldsymbol{R}||_0 = \sum_{i=1}^n o_i$. And $||\mathcal{T}||_0$ satisfies: $||\mathcal{T}||_0 \geq \sum_{j=1}^n\sum_{i=1}^n o_i = n||\boldsymbol{R}||_0^2$.

Usually, the medical dataset has more than 20 thousand words corpus and the value of $||\boldsymbol{R}||_0$ is over 10 million. Thus, according to Theorem 1, the scale of $\mathcal{T}$ is too large to be computed efficiently. We propose the following two methods to address this bottleneck.

Only retain the top-k probability. Consider a concept c_k and every concept pair (c_i, c_j) where $(i,j) \in \{(i,j)|i \neq j, i \neq k, j \neq k, \boldsymbol{R}_{ik} > 0, \boldsymbol{R}_{jk} > 0\}$. The triple (i,j,k)'s corresponding value in $\mathcal{T}$ is $\mathcal{T}_{ijk}$. For some chosen k, we find the distribution of $\mathcal{T}_{:,:,k}$ is close to the long-tailed distribution, few high probability data and much more low probability data. Therefore, to trade off computational complexity and information loss, we can sort $\mathcal{T}_{:,:,k}$ retain the top-k values. Then $||\mathcal{T}||_0$, the number of non-zero elements in $\mathcal{T}$, is $\mathcal{O}(nk)$, which is an acceptable scale of the problem. We name this method InterHG-Tensor for short.

Optimize a probability co-occurrence matrix factorization loss function. Recall the reconstruction loss term $(\mathcal{T}_{ijk} - \boldsymbol{u}_i^T\boldsymbol{u}_k - \boldsymbol{u}_j^T\boldsymbol{u}_k)^2 = (\frac{\boldsymbol{R}_{ik}+\boldsymbol{R}_{jk}}{\sum_{p=1}^n \boldsymbol{R}_{pk}} - \boldsymbol{u}_i^T\boldsymbol{u}_k - \boldsymbol{u}_j^T\boldsymbol{u}_k)^2$. Intuitively, if we can enable $\boldsymbol{u}_i^T\boldsymbol{u}_k$ to be close to $\frac{\boldsymbol{R}_{ik}}{\sum_{p=1}^n \boldsymbol{R}_{pk}}$ and enable $\boldsymbol{u}_j^T\boldsymbol{u}_k$ to be close to $\frac{\boldsymbol{R}_{jk}}{\sum_{p=1}^n \boldsymbol{R}_{pk}}$, the reconstruction loss can be close to zero as well. To prove the correctness of our intuitiveness, we propose Proposition 2 below. Thus, optimizing the $\mathcal{L}_{reconstruction}$ can be converted to optimize a probability co-occurrence matrix factorization loss

$$\mathcal{L}_{Prob-MF} = \sum_{(i,j)\in\mathcal{D}^+} (\frac{\boldsymbol{R}_{ij}}{\sum_{p=1}^n \boldsymbol{R}_{pj}} - \boldsymbol{u}_i^T\boldsymbol{u}_j)^2 + \mathcal{L}_{reg}$$

The scale of probability co-occurrence matrix factorization is equal to $||\boldsymbol{R}||_0$, which is easy to be implemented. We name this method InterHG-PMF for short.

Proposition 2: Optimize the probability co-occurrence matrix factorization loss function is equal to optimize the upper bound of the tensor-based loss function.

Proof 2:

$$\begin{aligned}&\mathcal{L}_{reconstruction} \\ &= \sum_{(i,j,k)\in\mathcal{D}^+_{tensor}} (\frac{\boldsymbol{R}_{ik}}{\sum_{p=1}^n \boldsymbol{R}_{pk}} - \boldsymbol{u}_i^T\boldsymbol{u}_k + \frac{\boldsymbol{R}_{jk}}{\sum_{p=1}^n \boldsymbol{R}_{pk}} - \boldsymbol{u}_j^T\boldsymbol{u}_k)^2 \\ &\leq 2 \sum_{(i,j,k)\in\mathcal{D}^+_{tensor}} \underbrace{(\frac{\boldsymbol{R}_{ik}}{\sum_{p=1}^n \boldsymbol{R}_{pk}} - \boldsymbol{u}_i^T\boldsymbol{u}_k)^2 + (\frac{\boldsymbol{R}_{jk}}{\sum_{p=1}^n \boldsymbol{R}_{pk}} - \boldsymbol{u}_j^T\boldsymbol{u}_k)^2}_{\text{probability co-occurrence matrix loss function}}\end{aligned}$$

Therefore, the original problem can be converted to the probability co-occurrence matrix factorization problem.

C. Interpretable Model for Hypothesis Generation

To design a task-specific interpretable model, there are two questions worth considering: (1) Which source or information can be used to interpret hypothesis generation? (2) How can we guarantee that our interpretation are reasons of model decisions?

Because medical researchers are the target population of our interpretable model, it is reasonable to assume medical knowledge can be understood by them. Thus, considering the particularity of hypothesis generation, a task belonging to medical domain, we design an interpretable model fusing other medical knowledge to explain learned embeddings. Since the embeddings of concepts in Section II-B are not interpretable, we design a model representing concepts via medical knowledge explicitly, which makes them transparent and can make sure they are related to decisions directly.

In medical domain, there is a knowledge source consisting of categorical information and their definitions are provided by the subject matter experts. Most concepts can be classified into one or several categories in the knowledge source. We take the definition of category "Nucleotide Sequence" in the knowledge source as an example, where '__' means the concept occurs in corpus $\mathcal{D}$:

***Nucleotide Sequence**: "The sequence of purines and pyrimidines in nucleic acids and polynucleotides. Included here are nucleotide-rich regions, conserved sequence, and DNA transforming region."*

Thus, considering categories in the knowledge source as an "dictionary", we assume that concepts in corpus $\mathcal{D}$ can be represented by the combination of dictionary words with weight. Then our goal is to learn how to represent concepts via categories, and interpreting hypothesis generation can be solved by analyzing the weights of categories of given

concepts. Besides, definitions from experts in the knowledge source are very precise and the underlined words can summarize definitions briefly, so we leverage the underlined words defining the category and aim to infer the representation of the category. Because learning the representation of categories and concepts will influence each other, we design a co-training framework to learn their representation jointly. We show the formulation of the framework as follows.

Suppose the representation of concepts is $\boldsymbol{U} = [\boldsymbol{u}_1^T; \boldsymbol{u}_2^T; \ldots; \boldsymbol{u}_n^T] \in \mathbb{R}^{n \times k}$ and the representation of categories is $\boldsymbol{D} = [\boldsymbol{d}_1^T; \boldsymbol{d}_2^T; \ldots; \boldsymbol{d}_m^T] \in \mathbb{R}^{m \times k}$. For the ith category (e.g. Nucleotide Sequence), the concepts defining it are $\{c_{i_1}, c_{i_2}, \ldots, c_{i_{k_i}}\}$ (e.g. purines, pyrimidines, nucleic acids, etc), which is denoted as set $Def(i)$. As our assumption, the representation of c_i is the weighted combination of $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m$, which can be formulated as $\boldsymbol{u}_i = \boldsymbol{D}^T \boldsymbol{\theta}_i$ where $\boldsymbol{\theta}_i$ is a m-dimension column vector. To enhance its interpretation, we hope $\boldsymbol{\theta}_i$ can model probability such as $(\boldsymbol{\theta}_i)_k = P(\boldsymbol{d}_k | \boldsymbol{u}_i)$. We leverage a similarity function to measure the similarity between $\boldsymbol{d}_k$ and $\boldsymbol{u}_i$ to define the probability:

$$(\boldsymbol{\theta}_i)_k = \frac{e^{s(\boldsymbol{u}_i, \boldsymbol{d}_k)/\tau}}{\sum_{j=1}^{m} e^{s(\boldsymbol{u}_i, \boldsymbol{d}_k)/\tau}}$$

where $s(\cdot)$ is a similarity function which can be specific in a very general form (e.g. Euclidean distance, Mahalanobis Distance, scaled inner product, etc.), and τ is a hyper-parameter controlling the smoothness of $(\boldsymbol{\theta}_i)_k$.

Similarly, the representation of dictionary words can be formulated a $\boldsymbol{d}_i \approx \boldsymbol{U}_{Def(i)}^T \boldsymbol{\beta}_i = \sum_{k \in Def(i)} \boldsymbol{u}_k (\boldsymbol{\beta}_i)_k$, where $(\boldsymbol{\beta}_i)_k$ is defined as $(\boldsymbol{\beta}_i)_k = \frac{e^{s(\boldsymbol{d}_i, \boldsymbol{u}_k)/\tau}}{\sum_{l \in Def(i)} e^{s(\boldsymbol{d}_i, \boldsymbol{u}_l)/\tau}}$. To enable $\boldsymbol{d}_i$ to be close to $\boldsymbol{U}_{Def(i)}^T \boldsymbol{\beta}_i$, we leverage the square error to punish their gap:

$$\mathcal{L}_{Dictionary} = \sum_{i=1}^{m} ||\boldsymbol{d}_i - \boldsymbol{U}_{Def(i)}^T \text{Softmax}(\boldsymbol{U}_{Def(i)} \boldsymbol{d}_i / \tau)||_2^2$$

Here, we take $s(\cdot)/\tau$ as scaled inner product. And as suggestion in [6], we impose orthogonal constraint on $\boldsymbol{D}$. Therefore, to summarize, the whole loss function can be formulated as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{U}, \boldsymbol{D}} \mathcal{L} &= \mathcal{L}_{reconstruction} + \lambda \mathcal{L}_{reg} + \alpha \mathcal{L}_{Dictionary} \\ \text{s.t.} \boldsymbol{D}\boldsymbol{D}^T &= \boldsymbol{I}_{m \times m} \end{aligned} \tag{3}$$

D. Optimization

Because of the orthogonal constraint, it is difficult to optimize $\boldsymbol{U}$ and $\boldsymbol{D}$ directly. Thus, we relax the constraint as a soft regularizer firstly, and then the optimization problem can be solved by any popular gradient-based solvers like SGD, Adam, etc. Specifically, we convert optimization problem Eqn 3 to an unconstrained optimization problem for $\boldsymbol{U}$ and $\boldsymbol{D}$:

$$\arg\min_{\boldsymbol{U}, \boldsymbol{D}} \mathcal{L}_{reconstruction} + \lambda \mathcal{L}_{reg} + \alpha \mathcal{L}_{Dictionary} + \gamma ||\boldsymbol{D}\boldsymbol{D}^T - \boldsymbol{I}||_F^2.$$

And then we optimize $\boldsymbol{U}, \boldsymbol{D}$ respectively via Adam.

III. Experiment

In this section, we evaluate our proposed interpretable hypothesis generation framework with the aim of answering the following research questions. 1) Does our proposed framework outperform state-of-the-art baselines? 2) Is the output of the proposed framework interpretable?

A. Experiment Settings

1) Dataset.: Following [2], [3], [14], we choose MEDLINE, one of the largest and most popular available scientific repositories, as the data source of our experiments. For each article, it associates with its unique identifier (PMID), title, abstract, publication date and Medical Subject Headings (MeSH) terms. According to several previous studies [15], using whole concepts from titles and abstracts will introduce extra noise, and MeSH terms are accurate and high-quality enough for this task. Therefore, we conduct our investigation by setting MeSH terms as our unit of analysis following [2], [3], [14], [16], [17]. Fortunately, the category information of MeSH terms can be obtained from UMLS[2]. So the category information and dictionary words' definitions are from UMLS. In this paper, we use corpus in 2015 from MEDLINE as training data, and corpus in 2016 from MEDLINE as testing data.

To evaluate the performance of the proposed model, following studies [2], [3], [14], [18] in this area, the "golden dataset" is chosen as test cases in the following experiments. We enumerate test cases as follows:

1) Fish-oil (FO) and Raynaud's Disease (RD) (1985)
2) Magnesium (MG) and Migraine Disorder (MIG) (1988)
3) Somatomedin C (SMC) and Arginine (ARG) (1994)
4) Indomethacin (INN) and Alzheimer Disease (AD) (1989)
5) Schizophrenia (SZ) and Calcium-Independent Phospholipase A2 (CI-PA2) (1997)

However, because, in 2015's articles, SMC does not appear, we take place SMC with "Somatomedin (SM)".

2) Comparison Methods.: To evaluate the performance of our proposed interpretable model, we compared it with classical and state-of-the-art baselines suggested in [2], [19]: 1) **Jaccard** [20], a popular technique for link prediction, captures the association between two concepts via the ratio of co-occurrence of them. 2) **Preferential Attachment** [21], another classical link prediction technique, measures the association between two concepts via the sum of the occurrence of the two concepts. 3) **AMW** (average minimum weight) is one of the best algorithms for literature-based discovery [18]. 4) **Arrowsmith** [18], a famous and popular algorithm for literature-based discovery, measures the similarity between two concepts via the number of concepts co-occur with any concept of the two concepts. 5) **Word2Vec** [22] is a state-of-the-art method to learn words' representation in an embedding space via model words' co-occurrence relationship. It includes two forms: Skip-grams model (SG) and continuous-bag-of-words model (CBOW), and we use SG model in our paper. We implement it via **Gensim** library [23]. 6) **Matrix Factorization (MF)**, a state-of-the-art algorithm for link prediction,

[2]https://semanticnetwork.nlm.nih.gov

TABLE I
EVALUATION ON FO-RD, MG-MIG, SM-ARG, INN-AD, AND SC-CI,PA2.

	FO-RD			MG-MIG			SM-ARG			INN-AD			SC-CI,PA2		
	SP@20	SP@50	SP@100	SP@20	SP@50	SP@100	SP@20	SP@50	SP@100	SP@20	SP@50	SP@100	SP@20	SP@50	SP@100
Jaccard	-0.4992	-0.3334	-0.1981	-0.5485	-0.3163	-0.3761	-0.6073	-0.5600	-0.6288	-0.5598	-0.5343	-0.3615	0.0760	0.1452	0.3202
PA	-0.8135	-0.8029	-0.7684	-0.6645	-0.5520	-0.5907	-0.6914	-0.8716	-0.8813	-0.5750	-0.5840	-0.4392	-0.6023	-0.3178	-0.3282
AMW	**0.7925**	**0.6340**	0.6950	-0.6614	0.2322	-0.4233	0.0355	0.3068	0.1050	-0.3325	-0.3826	-0.1624	**0.4326**	**0.3508**	**0.4093**
Arrowsmith	-0.7908	0.7932	0.7804	-0.5935	-0.8710	-0.8813	-0.5935	-0.8710	-0.8733	-0.4795	-0.5880	-0.4474	-0.5915	-0.3227	-0.3578
Word2vec	-0.1038	0.0489	0.1900	-0.1648	0.1695	0.1674	**0.0792**	0.1376	0.3170	-0.1123	-0.0719	0.0766	0.1196	0.0300	0.2142
MF	-0.0421	0.0612	0.1047	0.1166	0.0037	0.1865	0.1878	0.0666	0.0820	0.0897	-0.2451	0.1132	0.1114	0.0586	0.2547
InterHG-PMF	0.5654	0.5064	0.6849	0.0783	**0.4249**	0.3670	-0.0550	0.3324	**0.5465**	**0.4868**	**0.3124**	**0.3432**	0.2333	0.1451	0.3891
InterHG-Tensor	0.5293	0.5828	**0.7042**	**0.1723**	0.3914	**0.3915**	0.0181	**0.3730**	0.5364	0.2298	0.2188	0.2701	0.3243	0.1409	0.3388

achieves great success in hypothesis generation task, according to [2], [3]. It is worth mentioning that, [2], [3] all consider a dynamic system, but we consider static situation here. Thus, to be fair, we only leverage static MF as our baseline.

3) Evaluation Metric.: Hypothesis generation is to generate top-k concepts most likely to bridge a given concept pair, so **Spearman's rank correlation (SP)** is used to evaluate results. Besides, due to lacking of standard ground truth set, following previous work [2], [3], [14], [18], we also generate a ground truth set on testing set according to ranking the following scores: $gt(c_k) = \frac{\#(c_k, c_i) + \#(c_k, c_j)}{\sum_l \#(c_k, c_l)}$, where $\#(c_k, c_i)$ denotes the number of times concept c_k and concept c_i co-occurs. Besides, the predicted intermediary concepts c_k are ranked via F_1-cosine similarity score: $2\frac{cos(\boldsymbol{u}_k, \boldsymbol{u}_i) * cos(\boldsymbol{u}_k, \boldsymbol{u}_j)}{cos(\boldsymbol{u}_k, \boldsymbol{u}_i) + cos(\boldsymbol{u}_k, \boldsymbol{u}_j)}$, where $cos(\boldsymbol{u}_k, \boldsymbol{u}_i)$ is the cosine similarity between $\boldsymbol{u}_k$ and $\boldsymbol{u}_i$.

TABLE II
TOP-6 CATEGORY FOR INTERPRETING CONCEPTS.

Concept	Category	
FO	vitamin	chemical viewed functionally
	food	archaeon
	research activity	finding
RD	mammal	daily or recreational activity
	clinical drug	organ or tissue function
	mental or behavioral dysfunction	anatomical structure
INN	pathologic function	manufactured object
	cell or molecular dysfunction	physiologic function
	patient or disabled group	clinical drug
AD	finding	laboratory or test result
	indicator, reagent, or diagnostic aid	activity
	research activity	organism function

B. Comparison with Baselines

In this section, we report the performance of baselines and the proposed InterHG in Table I to answer the first question. From them, we have following important findings.

First, *InterHG outperforms the most state-of-the-art baselines greatly.* According to the five tables, InterHG-PMF and InterHG-Tensor have better performance than baselines on most cases, especially on INN-AD. Although it does not outperform AMW on some cases, the gap between them is very small. Thus, the proposed InterHG is both interpretable and effective.

Second, *Co-training interpretation module and tensor-based loss function is effective.* In [6], they interpret Word2vec by category information via learning a transformation matrix. However, in their work, interpretation and learning process are independent, which does not take full advantage of the category information for model accuracy. Thus, that model has the same performance as MF. In contrast to it, the InterHG train the interpretation module and prediction module together, making full use of the mapping between categories and definitions to improve model accuracy. And this point is is reflected in the performance of InterHG-PMF, InterHG-Tensor and MF in tables.

Third, *Both InterHG-PMF and InterHG-Tensor have great performance, but they have their own advantages and disadvantages.* Recalling the meaning of InterHG-PMF and InterHG-Tensor, InterHG-PMF is the probability co-occurrence matrix factorization method and InterHG-Tensor is the method retaining the top-k probability. According to the tables, it is obvious that both the two strategies are effective. However, InterHG-PMF optimizes an upper bound of $\mathcal{L}_{Tensor}$, so there may still have a gap with $\mathcal{L}_{Tensor}$. For InterHG-Tensor, it rejects data that is in the tail of data's distribution. Although it make the algorithm efficient enough, it also can not fit those rejected data. Therefore, they perform best in most cases but perform not as good as AMW or Word2Vec in few cases.

C. Interpretation for Hypothesis Generation

In order to interpret the embedding of concepts learned by the proposed InterHG, we show top six categories with the greatest weight value in $\boldsymbol{\beta}_i$ of concepts in "golden dataset", without loss of generality. One category is more likely to correspond to the concept if it has larger weight. The results is reported in Table II sorted by weight in descending order. According to the learned categories, we can interpret how the model learning concepts' representation. Besides, it helps biomedical researchers check the correctness of representation–if the learned categories are reasonable, the representation is more reliable. We invite experts providing interpretation as follows:

1) Fish oil usually contains some [vitamin] A and D. It is a [chemical viewed functionally] for healthcare. It is also a kind of [food] supplement. However, it is hard to related fish oil to [archaeon], [research activity] or [experimental model of disease], etc.

2) Raynaud's disease is a disease in [mammal] (human). It affects [daily or recreational activity] (skin turn white and blue) by impairing [organ or tissue function] (disorder of the blood vessels) in [anatomical structure] (blood vessels). However, it is hard to relate Raynaud's disease to [mental or behavioral dysfunction], [virus] or [idea or concept], etc.

3) Indomethacin is a [manufactured object] (drug) used to relieve [pathologic function] (pain, swelling, and joint stiffness) caused by [cell or molecular dysfunction] (arthritis, gout, bursitis, and tendonitis). It is also used to relieve pain from various other conditions. This medication is known as a [clinical drug] (nonsteroidal anti-inflammatory drug (NSAID)). It works by blocking your body's [physiologic function] (production of certain natural substances that cause inflammation). However, it is hard to relate indomethacin to [organization], [research activity] or [experimental model of disease], etc.

4) Alzheimer disease is the [indicator, reagent, or diagnostic aid] (most common cause of) dementia, a general term for memory loss and other cognitive abilities serious enough to interfere with [activity] (daily life). However, it is hard to relate alzheimer disease to [finding], [laboratory or test result] or [research activity], etc.

From experts' interpretation, it can be observed that the ten concepts can be interpreted via learned categories. Besides, most of concepts can be interpreted via the top three categories. It demonstrates that the proposed InterHG is interpretable and the interpretation is reliable.

IV. Conclusions

Hypothesis generation is a vital task in medical informatics that enables medical researchers to verify the implicit connections between two target concepts. However, the limitations of existing matrix factorization based approaches (i.e., the indirect modeling of bridge concept associations with target concepts and the lack of interpretability) prohibit their usage in real practice. Towards conquering these limitations, we propose a novel hypothesis generation model called InterHG that can output accurate and interpretable results. We proposed to model hypothesis generation as a tensor factorization tasks so that the association between bridge and target concepts is modeled directly. To reduce its computational complexity, we proposed two effective strategies, i.e., InterHG-Tensor and InterHG-PMF. Furthermore, we proposed to incorporate a regularizer based on the known category-concept relationship into the objective function so that the learned concept embeddings can be interpreted as a set of category weights. Such output allows medical researchers to verify the effectiveness of concept embedding and the plausibility of the connection between target concepts. Our experiments on MEDLINE data demonstrate that the proposed InterHG model achieved high accuracy and meaningful interpretable hypothesis generation results.

Acknowledgement

This work is supported in part by the US National Science Foundation under grant NSF IIS-1747614 and NSF IIS-2141037. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.

[2] K. Jha, G. Xun, Y. Wang, and A. Zhang, "Hypothesis generation from text based on co-evolution of biomedical concepts," in *Proceedings of KDD'19*. ACM, 2019, pp. 843–851.

[3] G. Xun, K. Jha, V. Gopalakrishnan, Y. Li, and A. Zhang, "Generating medical hypotheses based on evolutionary medical concepts," in *Proceedings of ICDM'17*. IEEE, 2017, pp. 535–544.

[4] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of KDD'17*, 2017, pp. 1903–1911.

[5] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[6] K. Jha, Y. Wang, G. Xun, and A. Zhang, "Interpretable word embeddings for medical domain," in *Proceedings of ICDM'18*. IEEE, 2018, pp. 1061–1066.

[7] G. Peake and J. Wang, "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems," in *Proceedings of KDD'18*. ACM, 2018, pp. 2060–2069.

[8] N. Liu, X. Huang, J. Li, and X. Hu, "On interpretation of network embedding via taxonomy induction," in *Proceedings of KDD'18*. ACM, 2018, pp. 1812–1820.

[9] W. Cheng, Y. Shen, L. Huang, and Y. Zhu, "Incorporating interpretability into latent factor models via fast influence analysis," in *Proceedings of KDD'19*. ACM, 2019, pp. 885–893.

[10] W. Ma, M. Zhang, Y. Cao, W. Jin, C. Wang, Y. Liu, S. Ma, and X. Ren, "Jointly learning explainable rules for recommendation with knowledge graph," in *Proceedings of WWW'19*. ACM, 2019, pp. 1210–1221.

[11] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, "Explainable reasoning over knowledge graphs for recommendation," in *Proceedings of AAAI'19*, vol. 33, 2019, pp. 5329–5336.

[12] Z. Chen, X. Wang, X. Xie, T. Wu, G. Bu, Y. Wang, and E. Chen, "Co-attentive multi-task learning for explainable recommendation," in *Proceedings of IJCAI'19*. AAAI Press, 2019, pp. 2137–2143.

[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of EMNLP'14*, 2014, pp. 1532–1543.

[14] K. Jha, G. Xun, Y. Wang, V. Gopalakrishnan, and A. Zhang, "Conceptsbridges: Uncovering conceptual bridges based on biomedical concept evolution," in *Proceedings of KDD'18*. ACM, 2018, pp. 1599–1607.

[15] M. Yetisgen-Yildiz and W. Pratt, "Using statistical and knowledge-based approaches for literature-based discovery," *Journal of Biomedical informatics*, vol. 39, no. 6, pp. 600–611, 2006.

[16] X. Hu, X. Zhang, I. Yoo, X. Wang, and J. Feng, "Mining hidden connections among biomedical concepts from disjoint biomedical literature sets through semantic-based association rule," *International Journal of Intelligent Systems*, vol. 25, no. 2, pp. 207–223, 2010.

[17] P. Srinivasan and B. Libbus, "Mining medline for implicit links between dietary substances and diseases," *Bioinformatics*, vol. 20, no. suppl_1, pp. i290–i296, 2004.

[18] M. Yetisgen-Yildiz and W. Pratt, "A new evaluation methodology for literature-based discovery systems," *Journal of biomedical informatics*, vol. 42, no. 4, pp. 633–643, 2009.

[19] J. Lever, S. Gakkhar, M. Gottlieb, T. Rashnavadi, S. Lin, C. Siu, M. Smith, M. R. Jones, M. Krzywinski, and S. J. Jones, "A collaborative filtering-based approach to biomedical knowledge discovery," *Bioinformatics*, vol. 34, no. 4, pp. 652–659, 2018.

[20] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of IMECS'13*, vol. 1, no. 6, 2013, pp. 380–384.

[21] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[23] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.