2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669658

A Hybrid-scales Graph Contrastive learning Framework for Discovering Regularities in Traditional Chinese Medicine Formula

Yingpei Wu*†, Zecheng Yin†, Kaiyuan Zhou†, Ruofei Wang†,
Yun Yang‡§, Zepeng Yin¶, Chunyang Ruan‖, Yanchun Zhang**†✉
* School of Computer Science, Fudan University, Shanghai, China
†Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
‡Oncology Institute of Traditional Chinese Medicine,
Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai 200071, China
§ Department of Oncology, Shanghai Municipal Hospital of Traditional Chinese Medicine,
Shanghai University of Traditional Chinese Medicine, Shanghai 200071, China
¶Qingdao West Coast New Area No.2 Traditional Chinese Medicine Hospital, Qingdao, China
‖SSE-ARD, Shanghai Enflame Technology Company
**Institute of Sustainable Industries and Liveable Cities, Victoria University, Melbourne, Australia
ypwu16@fudan.edu.cn, yinzecheng@e.gzhu.edu.cn, yanchun.zhang@vu.edu.au

***Abstract*—Discovering regularities in Traditional Chinese Medicine (TCM) formula has been a hot topic in assisting TCM clinical treatment and poly-pharmacology research. Several machine learning methods, like topic model, auto-encoder, and GNNs, have been proposed for discovering regularities in TCM. However, they are often limited by specific data challenges (e.g., complex relations with rich TCM knowledge, sparsity and ambiguity, expensive data labeling, etc.) in TCM formulae. Addressing these challenges, we first establish a TCM Attributed Heterogeneous Information Network (TAHIN) for modeling massive formulae, which can assemble various types of additional information and capture their relations. Based on the TAHIN, we further propose a novel hybrid-scales graph contrastive learning framework to learn high-quality node representations in a whole unsupervised manner which can be helpful for various tasks of discovering regularities such as herb classification and herb similarity search, etc. Extensive experiments demonstrate the effectiveness and interpretability of our method. Our source code and datasets are available at https://github.com/Yonggie/HsCTRD.**

***Index Terms*—Discovering regularities, Traditional Chinese medicine, Attributed heterogeneous information network, Hybrid-scales graph contrastive learning.**

I. Introduction

TRADITIONAL Chinese Medicine (TCM) is one of the oldest medicine systems. As the principal treatment means, the TCM formula is based on the idea of syndrome differentiation and a holistic view. During TCM diagnostic process, physicians should first carefully distinguish the hidden patterns under individual symptoms. According to these observed patterns and the characteristics of different herbs, physicians then formulate the therapeutic strategies and pick or design a suitable formula for individuals.

Diagnosis and prescription processes, however, are very complicated. In order to maximize the therapeutic efficacy and reduce the side effects, physicians have to not only consider the accuracy of herbs but also carefully analyze the regularities of compatibility of herbs, which involve some subjective judgments to some extent. This makes it difficult for junior practitioners with limited clinical experience to quickly make an accurate and ideal diagnosis when faced with a complex condition. Therefore, a study of discovering regularities in formula compatibility is needed to help TCM practitioners with formula preparation and aid pharmaceutical companies in selecting herbal combinations for finding new formulae.

Fortunately, machine learning and data mining can be applied to address some pain points in TCM. Yao et al. [1] constructed a topic model to characterize the generative process of prescriptions in TCM. Taking prediction task as a non-negative matrix tri-factorization problem, Zhu et al. [2] proposed a supervised learning framework for potential incompatible herb pair prediction. Hu et al. [3] adopted a classic Convolutional Neural Network (CNN) framework to model unstructured texts in medical records for syndrome differentiation of Yin and Yang deficiency in TCM. Based on graph convolution networks (GCNs), Jin et al. [4] proposed a method that simulates the implicit syndrome induction process for herb recommendation. Nevertheless, discovering regularities in short-text-like formulae is non-trivial due to the following challenges.

(1) Multiple relations with rich TCM knowledge. In fact, there are numerous interaction relations in the scenario of diagnosis, which must cogitate rich TCM knowledge. For example, doctors should consider suitable herbs for symptoms and examine their compatibility to make them more effective and have fewer side effects. This procedure requires full consideration of the herbs' function and finding the cause hidden behind the symptoms. Managing such complex structural

【方剂名】 桂枝汤
[Formula name] Cinnamon Twig Decoction

【出处】《伤寒论》
[Source] *Treatise on Cold Damage*

【组成】 君药: 桂枝 (9g)，臣药: 白芍 (9g)，使药: 炙甘草 (6g)，佐药: 生姜 (9g)，佐药: 大枣 (3g)
[Composition herbs] Emperor: Cinnamomi Ramulus (9g), Minister: Paeoniae Radix Alba (9g), Courier: Glycyrrhizae Radix et Rhizoma (6g), Assistant: Zingiberis Rhizoma Recens (9g), Assistant: Jujubae Fructus (3g)

【主治】恶风发热，汗出，头痛，鼻鸣干呕，苔白不渴，脉浮缓或浮弱者。
[Indication symptoms] aversion to wind, fever, perspiration, headache, congested and noisy nose, and dry retching, white coating of the tongue, no thirst, superficial and moderate pulse or superficial and weak pulse.

Fig. 1. An example of TCM formula.

information and preserving the various feature information is a critical problem needing to be solved.

(2) Sparsity and ambiguity. Unlike paragraphs or documents, the short-text-like formulae that lack enough contextual information are usually semantically sparse and ambiguous [5]. Specifically, as shown in Fig. 1, the formula only lists the symptoms and herbs without describing the herbs' functions and their relationship in detail. Thus, sparsity and ambiguity make it difficult for Natural Language Processing (NLP) methods to carry out text mining and knowledge discovery.

(3) Expensive Data Labeling. For TCM, data labeling is usually expensive and time-consuming, and for some special herbs, the label is tough to acquire (such as Tibetan and Miao Medicine, etc.). It would be the ideal method that can achieve competitive performance even without any label.

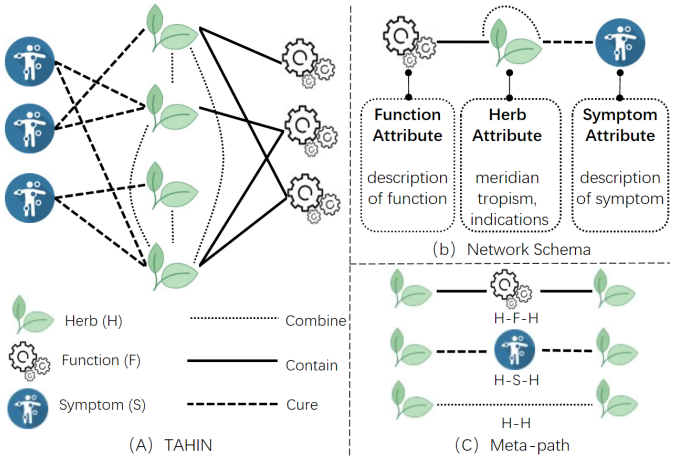

Fig. 2. A toy example of TAHIN and associated demonstration of network schema and meta-path.

For exploiting the multiple relations and feature information, we present a TCM Attributed Heterogeneous Information Network (TAHIN) to model the rich semantics and multiple relations among different types of entities (i.e., herbs, symptoms) derived from TCM formulae. To handle the sparsity and ambiguity, we further incorporate the external attributed features into the TAHIN, such as herb's meridian tropism, indications, and functions, etc. Fig. 2 (b) shows the network schema of the TAHIN in the scenario of TCM formula, which clearly illustrates the entities and their interactions. Several attempts have been made to mine attributed heterogeneous information network and shown remarkable performance in various kinds of applications [6]–[8]. However, they are carefully constructed for particular tasks and fall into the category of supervised learning method, so they cannot be directly applied to the TAHIN and the TCM regularities discovering problem in an unsupervised scenario.

In this paper, we model discovering regularities in formula compatibility as a graph contrastive representation learning problem and propose a novel Hybrid-scales Contrastive learning based TCM Regularities Discovering model, called HsCTRD. Unlike Deep Graph Infomax (DGI) [9] and Deep Multiplex Graph Infomax (DMGI) [10], our model presents a new mutual information maximization paradigm. The basic idea of HsCTRD is to learn representations by maximizing mutual information (MI) [11] between elements by the hybrid scales (i.e., cross-scale and same-scale) contrasting. Cross-scale MI, similar to the one used in DGI and DMGI, is estimated between node and graph, while the same-scale MI is calculated between node and node. Simultaneously maximizing these two kinds of MI can help our model capture information on different scales in TAHIN from several perspectives. Based on this scenario, we further propose the *semantic fusion component* to jointly integrate the embeddings of multiple subnetworks in an unsupervised manner, so as to facilitate them to mutually help each other learn high-quality embeddings useful for various downstream tasks, e.g., herb classification and herb similarity search, etc.

In summary, the contributions of our work are summarised as follows:

- We present a novel TAHIN for modeling the rich semantics and multiple relations of herbs, symptoms, and functions extracted from TCM formulae and literature.
- Based on the TAHIN, we propose the innovative HsCTRD to effectively encode node attributes and structure information in a self-supervised manner. To our knowledge, this is the first study to combine discovering regularities in TCM formula compatibility with MI maximization.
- Different from DGI and DMGI, our proposed model can learn nodes' representations from hybrid scales, which can be more beneficial in capturing local and global information. Further, the proposed *semantic fusion component* can fuse these representations also in a self-supervised manner which can be used to better discover TCM regularities.
- Extensive experiments on two real-world datasets illustrate the best performance of the proposed HsCTRD compared with the state-of-the-art models.

II. Related Works

A. Discovering Knowledge in Medicine

In recent years, deep learning-based medicine knowledge discovery has been a hot and challenging research topic. Li et. al [12] considered the TCM prescription generation task as a dialogue generation problem and developed a Seq2Seq model to automatically generate prescriptions. Wang et al. [13] problematic discovering relationship between symptoms and herbs as a machine translation problem and proposed

a transformer-based model to translate symptoms to herbs. Li et.al [14] presented a herb knowledge enhanced Seq2Seq model to generate prescription. One of the deficiencies among these methods is that the NLP methods are difficult to handle short-text-like prescription, making it difficult to capture the semantic information in the TCM prescriptions. Besides Seq2Seq models, Chen et. al [15] proposed a soft clustering method based on Heterogeneous Information Network (HIN) to explore the categories of formulas. Based on heterogeneous entity networks, Wan et.al [16] proposed a semi-supervised learning algorithm to extract relations from TCM literature. Ruan et al. [17] constructed a bipartite graph embedding based model to discover patterns of TCM prescriptions. Jin et al. [4] developed several GCNs to learn the symptom, herb, and syndrome embeddings from three bipartite graphs, and then simulated the implicit syndrome induction process for herb recommendation. Although graph-based approaches are effective in discovering TCM knowledge, their performance was limited by using only structure information and ignoring node attributes. Furthermore, most of the above methods require external guidance, i.e., annotated labels, which limits their applicability. Yang et al. [18] introduced a kernelized multitask learning model that learns and transfers information from the clinical data of other patients as collaborative information in order to rate a different list of drugs and adverse drug reactions for different patients. Yang et al. [19] summarized how to utilize machine learning approaches for accelerating drug repurposing in TCM and western medicine.

B. Graph Contrastive Learning

Lately, contrastive methods adopted for computer vision have also been adapted to graph domain and achieved great success. Velickovic et al. proposed DGI [9] which extended deep InfoMax [20] to graphs and contrasted node-local patches against global graph representations. Following this line of study, Sun et al. developed InfoGraph [21] which learned graph representations by maximizing the MI between the graph representations and sub-structural representations. GMI [22] extended the idea of traditional MI computations from vector space to the graph domain where maximized the MI from node features and topological structure. For multiplex graph, Park et al. presented DMGI [10] which divided the original graph into several homogeneous ones and followed the infomax objective of DGI for each relation graph. Inspired by SimCLR [23] Zhu et al. presented GRACE [24] which generate two graph views by corruption and learn node representations by maximizing the agreement of node representations in these two views. Motivated by the above models, our proposed HsCTRD is tailored for TAHIN representation learning.

III. Preliminaries

Definition 1. TCM Attributed Heterogeneous Information Network (TAHIN). As a special type of Heterogeneous Information Network (HIN), the TAHIN is defined as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, X\}$ consisting of an node set $\mathcal{V}$, a link set $\mathcal{E}$ and an attribute information matrix $X \in \mathbb{R}^{|\mathcal{V}| \times d}$. There are two mapping functions on TAHIN, one of which is node type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{O}$, and the other is link type mapping function $\psi : \mathcal{E} \rightarrow \mathcal{R}$. $\mathcal{O}$ and $\mathcal{R}$ denote the sets of predefined node and link types, where $|\mathcal{O}| + |\mathcal{R}| > 2$. Specifically, Fig. 2 (a) shows an example of TAHIN. In the TAHIN, we define three types of nodes corresponding to herb, function, symptom, and three types of links denoting various types of relations between them. Herb-Herb edges denote these herbs are co-occurrence in a formula, which indicates the compatibility between them. Herb-Symptom edges reflect the semantics of the herb's treatment or relief of the specific symptoms. Herb-Function edges express that the herb contains a certain function.

Definition 2. TCM Meta-path. A TCM meta-path ρ is defined as a path in the form of $A_1 \xrightarrow{r_1} A_2 \xrightarrow{r_2} \cdots \xrightarrow{r_l} A_{l+1}$, which defines a composite relation $\rho = r_1 \circ r_2 \circ \cdots \circ r_l$ between node type A_1 and A_{l+1} where $\circ$ denotes the composition operator on relations. For example, Fig. 2 (c) illustrates three meta-paths extracted from TAHIN in Fig. 2 (a). Herb-Function-Herb (*HFH*) describes that two herbs have similar effects and further offers clues to finding alternative herbs, and Herb-Symptom-Herb (*HSH*) denotes that both herbs are effective for a specific symptom; while another meta-path Herb-Herb (*HH*) indicates the herbs that co-occur in a formula which imply underlying laws of the herbal compatibility. Because of combining multiple relations, meta-path contains complex semantics, which can capture high-order structure information.

Definition 3. Unsupervised TAHIN embedding. Given the TAHIN $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, X\}$, the task of unsupervised TAHIN embedding is to learn a d'-dimensional vector representation $h_i \in \mathbb{R}^{d'}$ for each node $v_i \in \mathcal{V}$ in a self-supervised manner. It is noted that the learned embeddings can be used for discovering regularities in formula compatibility task such as herb classifcation and herb similarity search, etc.

IV. Proposed Method

A. Overview of the framework

The overview of HsCTRD is shown in Fig. 3. In HsCTRD, (i) we first map the constructed TAHIN to a multi-view network that consists of three single-view attributed networks encoding the relatedness over herbs guided by three designed meta-paths. Then, for each network, (ii) we employ two augmentation strategies, edge perturbation and node feature disturbance, to produce a correlative augmented graph. Taking these graphs as input, (iii) Graph Neural Network (GNN) encoder f computes the representations for graph's nodes. Based on these representations, (iv) we conduct hybrid-scales contrastive learning on them. Furthermore, (v) we propose the *semantic fusion component* to jointly integrate the embeddings of multi-view networks in an unsupervised manner. We will introduce the proposed method for each component in detail below.

B. Multi-view Network Node Embedding

In our task, meta-path is a straightforward means to connect herbs via different relationships among different nodes in

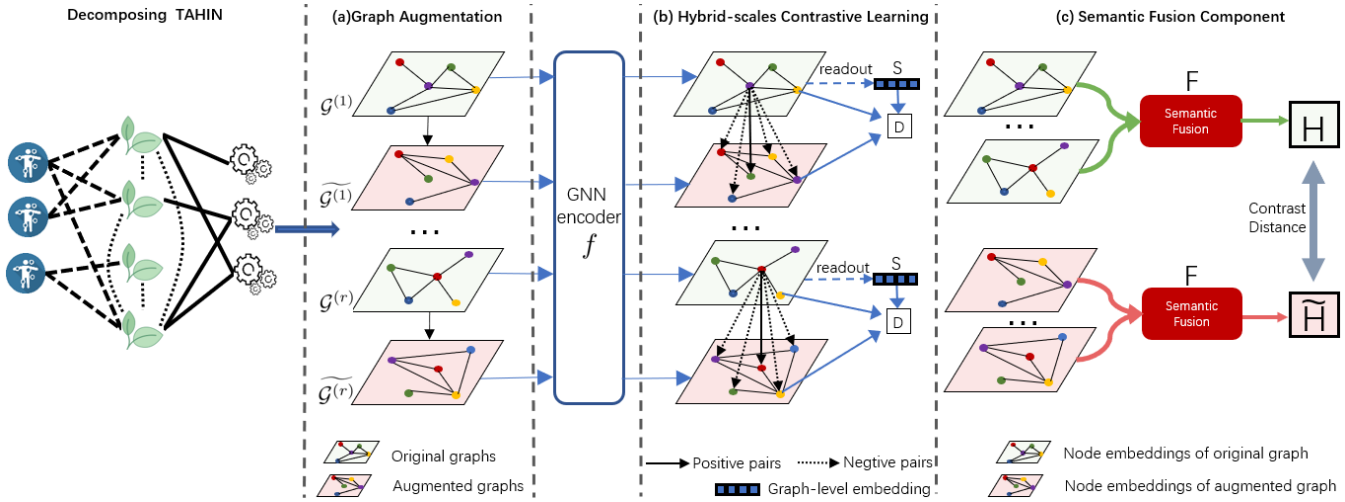

Fig. 3. An overview of our proposed HsCTRD framework.

TAHIN, and enable us to depict the relatedness over herbs in a comprehensive way. Thus, we convert the constructed TAHIN to a multi-view network that consists of three single-view attributed networks encoding the relatedness over herbs guided by three designed meta-paths (i.e., HFH, HSH, HH). We first give the formal definition of multi-view network. A multi-view network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^{(r)}_{r\in\mathcal{R}}, X\}$ consists of a set $\mathcal{V}$ of nodes and a set $\mathcal{R}$ of views, where $\mathcal{E}^{(r)}$ includes all edges in view $r \in \mathcal{R}$. Given the TAHIN $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, X\}$ and R meta-paths, we build a multi-view network with R single-view attributed graph $\mathcal{G}^{(r)} = \{\mathcal{V}, \mathcal{E}^{(r)}, X\}$ for the r-th view guided by the meta-path $\rho_r(r = \{1, ..., R\})$. In our case, in each single-view attributed graph $\mathcal{G}^{(r)}$, each node denotes a herb, and an edge between two herbs means if these two herbs can be connected under a certain meta-path (i.e., HFH, HSH, HH). These graphs describe different interactions between herbs, which can reflect different views of herb latent representations.

In order to capture the structure and attribute information, we employ a one-layer GCN [25] as the encoder f to learn embeddings of each single-view attributed graph.

$$H^{(r)} = f(X, A^{(r)}) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}^{(r)}\hat{D}^{-\frac{1}{2}}XW\right) \tag{1}$$

where $\hat{A}^{(r)} = A^{(r)} + I$ is the adjacency matrix with self-loops of the r-th view attributed graph and $\hat{D}$ is degree matrix where $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$, W is a trainable weight matrix of the encoder f, and σ is the nonlinear activation function, such as the $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$. $H^{(r)} = \{h_1^{(r)}, h_2^{(r)}, ..., h_n^{(r)}\}$ represents higher-level representations $h_i^{(r)} \in \mathbb{R}^{d'}$ of each node v_i in the r-th single-view attributed graph.

C. Graph Data Augmentation

Data augmentation is a major constituent of contrastive learning methods. In self-supervised visual representation learning tasks, contrasting congruent and incongruent views of images helps encoders to learn meaningful representations [23], [26], [27]. However, unlike the regular grid-like image data, the data augmentation on irregular graph data is not trivial, due to vertices and edges of graph not involving visually semantic contents as in the image [9]. In HsCTRD, we design two strategies for graph augmentation, edge perturbation for graph topology and node feature disturbance for node features.

For edge perturbation, we randomly add or drop a portion of edges in the original single-view attributed graph, while for node feature disturbance, we randomly switch a fraction of dimensions in node features. Formally, we first sample two random matrix $\boldsymbol{E} \in \{0,1\}^{N\times N}$ and $\boldsymbol{F} \in \{0,1\}^{N\times d}$ where each element is sampled from a Bernoulli distribution $\widetilde{\boldsymbol{E}}_{ij} \sim Bernoulli\,(p_e)$, $\widetilde{\boldsymbol{F}}_{gh} \sim Bernoulli\,(p_f)$ to determine whether to perturb the adjacency matrix at position (i, j) and switch the feature matrix at position (g, h). Here p_e and p_f is the corruption rate for graph topology and features, respectively. The perturbed adjacency matrix $\widetilde{\mathbf{A}}$ and generated node features matrix $\widetilde{\mathbf{X}}$ can be computed as:

$$\widetilde{\mathbf{A}} = \mathbf{A} \oplus \mathbf{E} \tag{2}$$

$$\widetilde{\mathbf{X}} = \mathbf{X} \oplus \mathbf{F} \tag{3}$$

where $\oplus$ is the XOR (exclusive OR) operation.

By leveraging these two strategies, we can get an augmented graph with different connectivity and features for each single-view attributed graph. Then we employ the GNN encoder f to generate the embedding of the augmented graphs. i.e. $\widetilde{H}^{(r)} = f(\widetilde{X}, \widetilde{A}^{(r)}) = \{\widetilde{h_1}^{(r)}, \widetilde{h_2}^{(r)}, ..., \widetilde{h_n}^{(r)}\}$.

Based on the embeddings of the original single-view attributed graph and the augmented graphs, we propose a hybrid-scales contrastive approach to learn the encoder f such that the final representation can achieve desired performance on downstream tasks such as herb classification and herb similarity search.

D. Hybrid-scales Contrastive Learning

Graph contrastive learning methods have achieved great success, because they can use the rich information in the data to guide the representation learning process [9], [21], [28]. However, these contrastive learning methods are based on a single scale, so as to overlook other scales' information. To address this issue, we propose a hybrid-scales (i.e., combing

cross-scale and same-scale) contrastive method to take better advantage of information on different scales.

Cross-scale Contrastive Learning(CCL). In HsCTRD, CCL learns representations by contrasting two graph elements in a different scale (i.e., node-graph). The objective of CCL is to maximize the MI between local (node-level) representations and the global (graph-level) representaion. To this end, for each single-view attributed graph, we first leverage a readout function $\mathcal{R}: \mathbb{R}^{n\times d'} \rightarrow \mathbb{R}^{d'}$ to summarize the node representations into a graph-level representation s.

$$s^{(r)} = R\left(f(X, A^{(r)})\right) = R\left(H^{(r)}\right) = \sigma\left(\frac{1}{n}\sum_{i=1}^{n} h_i^{(r)}\right), \tag{4}$$

where σ is the nonlinear activation function.

Based on the Jensen-Shannon divergence [29], we adopt a noise-contrastive type objective with a standard binary cross-entropy (BCE) loss between the samples from the joint (positive examples) and the product of marginals (negative examples) to maximize MI between local node representation ($h_i^{(r)}$) and global graph representation ($s^{(r)}$). The objective of CCL can be shown as follows:

$$\begin{aligned}\mathcal{L}_{ccl}^{(r)} = &\frac{1}{N+M}\left(\sum_{i=1}^{N}\mathbb{E}_{(X,A^{(r)})}\left[\log\mathcal{D}\left(h_i^{(r)}, s^{(r)}\right)\right] + \right.\\ &\left.\sum_{j=1}^{M}\mathbb{E}_{(\tilde{X},\tilde{A}^{(r)})}\left[\log\left(1-\mathcal{D}\left(\tilde{h}_j^{(r)}, s^{(r)}\right)\right)\right]\right),\end{aligned} \tag{5}$$

where N and M are the number of positive and negative samples, and $\mathcal{D}(\cdot,\cdot): \mathbb{R}^d \times \mathbb{R}^d \longmapsto \mathbb{R}$ is a discriminator that take in a node and a graph representation, and caculates the agreement between them. Here we simply implement the discriminator $\mathcal{D}$ as the dot product between two representations followed by a sigmoid function, i.e, $\mathcal{D}\left(h_i^{(r)}, s^{(r)}\right) = \text{sigmoid}\left(h_i^{(r)T}\cdot s^{(r)}\right)$.

Same-scale Contrastive Learning (SCL). Inspired by recent developments in contrastive visual learning [23], [30], we also employ SCL to further exploit local information in graphs. Different from the CCL, SCL can learn representations by contrasting two graph elements on a same scale, i.e., node-node contrasting.

Having generated the corresponding augmented graph for each single-view attributed graph, we consider the representations of the same node in two graphs as the positive pairs, and any different nodes as the negative pairs. Specifically, for any node u in original graph $\mathcal{G}^{(r)}$, we treat its representation $h_u^{(r)}$ and the corresponding representation $\widetilde{h_u}^{(r)}$ generated in the augumented graph $\widetilde{\mathcal{G}^{(r)}}$ as the positive pairs $(i.e., \{(h_u^{(r)}, \widetilde{h_u}^{(r)})|u \in \mathcal{V}\})$, and the representations of any different nodes as the negative pairs $(i.e., \{(h_u^{(r)}, \widetilde{h_v}^{(r)})|u, v \in \mathcal{V}, u \neq v\})$. Here, we follow SimCLR [23] and utilize the normalized temperature-scaled cross entropy loss (NT-Xent) [31], [32] to maximize the agreement of positive pairs and minimize it of negative pairs:

$$\begin{aligned}\mathcal{L}_{scl}^{(r)} = -\frac{1}{2|\mathcal{V}|}\sum_{u\in\mathcal{V}}\left[\log\frac{\exp\left(sim\left(h_u^{(r)}, \widetilde{h_u}^{(r)}\right)/\tau\right)}{\sum_{v\in\mathcal{V}}\exp\left(s\left(h_u^{(r)}, \widetilde{h_v}^{(r)}\right)/\tau\right)} + \right.\\ \left.\log\frac{\exp\left(sim\left(\widetilde{h_u}^{(r)}, h_u^{(r)}\right)/\tau\right)}{\sum_{v\in\mathcal{V}}\exp\left(sim\left(\widetilde{h_u}^{(r)}, h_v^{(r)}\right)/\tau\right)}\right],\end{aligned} \tag{6}$$

where sim(·,·) denotes cosine similarity between two learned node representations and τ is a temperature parameter. Please note that the SCL loss function contains two components due to the symmetry of the original and augmented graphs, which is the main difference with NT-Xent.

E. Semantic Fusion Component

By conducting hybrid-scales contrastive learning, we get single-view specific node representation $H^{(r)}$, that captures the local and global information in graph $\mathcal{G}^{(r)}(\forall r \in \mathcal{R})$. However, due to each $H^{(r)}$ trained independently for each single-view $r \in \mathcal{R}$, these representaions only contain the semantic-specific information regarding each view. This prompts us to achieve a self-supervised solution to fuse the representations learned from different views, so as to learn more semantically comprehensive representations of herb nodes.

Considering that different single-view attributed graphs have various semantic information, we use the attention mechanism to fuse them together to get the final embedding H. First, we calculate the weight of each single-view as follows:

$$\begin{aligned}w_r &= \frac{1}{|V|}\sum_{v\in V} f^{\top}\cdot\tanh\left(Wh_v^{(r)} + b\right),\\ \psi_r &= \frac{\exp(w_r)}{\sum_{i\in\mathcal{R}}\exp(w_i)},\end{aligned} \tag{7}$$

where $W \in \mathbb{R}^{d'\times d'}$ and $b \in \mathbb{R}^{d'\times 1}$ are learnable parameters, and f is the fusing attention vector. ψ_r is considered as the importance of the r-th single-view attributed graph to the final representation H. With the learned weights as coefficients, we can fuse these single-view specific node representaions to obtain the final representation H as follows:

$$H = \sum_{r\in\mathcal{R}}\psi_r \cdot H^{(r)} \tag{8}$$

After obtaining the representation $\widetilde{H}$ of the augmented multi-view network in this way, we need to maximize the disagreement between H and $\widetilde{H}$, which are formulated as follows:

$$\mathcal{L}_{sf} = -\|H - \tilde{H}\|_2, \tag{9}$$

where $\|\cdot\|_2$ is the Frobenius Norm of a matrix which calculates the euclidean distance of these two matrices.

Finally, we jointly optimize the sum of all the CCL loss (cf. Equation(5)), the SCL loss (cf. Equation(6)), and the semantic

fusion loss (cf. Equation(9)) to obtain the final objective L as follows:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} (\mathcal{L}_{ccl}^{(r)} + \mathcal{L}_{scl}^{(r)}) + \lambda_1 \mathcal{L}_{sf} + \lambda_2 \|\Theta\|^2, \qquad (10)$$

where λ_1 denotes the importance of the semantic fusion, λ_2 is a coefficient for l2 regularization on Θ.

V. Experiments

We now present our experiments to validate the effectiveness of the proposed framework. We begin by stating that all our source code and datasets are publicly available at https://github.com/Yonggie/HsCTRD for reproducibility.

A. Datasets

- **TCMRel** [16] is constructed based on the corpus of TCM literature, which contains entities such as herbs, symptoms, diseases, and the relationships among them.
- **ChP** [15] is built from the Pharmacopoeia of the People's Republic of China 2015 Edition, which is an unstructured corpus and contains various TCM information. We extract formula, herb, symptom and function to build TAHIN.

B. Baseline Methods

To evaluate the performance of the proposed HsCTRD, we compare it with two categories of baselines: semi-supervised methods {GCN [25], GAT [33], HAN [34]}, and unsupervised methods {metapath2vec [35], DGI [9], GRACE [24], and DMGI [10]}. Note that since the GCN, GAT, DGI, and GRACE are designed for homogeneous graphs and metapath2vec depends on a single meta-path, we test these baselines on all the single-view attributed networks and report the best performance.

C. Experimental settings

We implement HsCTRD by PyTorch and conduct experiments in the server with eight NVIDIA GeForce RTX 2080 GPUs. We set the node final embedding dimension $d' = 128$. For graph data augmentation, we set corruption rates, p_e and p_f, to be 0.5. In data preprocessing, we utilize the Bag Of Word (BOW) algorithm to generate the initial node feature with 2048 dimensions. About the training part, we apply Adam optimizer with a learning rate of 0.005, and employ an early stopping scheme with a patience of 50 to control epoch. The coefficients of objective L, λ_1 and λ_2, are set to be 0.1, 0.07, respectively.

D. Experiments Results

1) Node Classification: Since most of the herbal efficacy comes from ancient TCM literature, it has led to inconsistent and vague efficacy records [36]. For assisting herbal efficacy studying, we conduct node classification (according to herbal efficacy) experiments to evaluate HsCTRD. Specifically, after learning the node embeddings, we implement a logistic regression model on the learned embeddings in the training set, and then evaluate the test set. To achieve steady results, we repeat this process for 10 times and report the averaged Macro-F1, Micro-F1 in Table I.

TABLE I
Experiment results for the node classification task.

method	TCMRel		ChP	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
metapath2vec	0.043	0.1644	0.0524	0.187
GCN	0.1052	0.2759	0.3483	0.4755
GAT	0.1525	0.411	0.3172	0.3821
HAN	0.1987	0.3562	0.3261	0.48
DGI	0.1904	0.3972	0.3776	0.5082
GRACE	0.2376	0.4861	0.3346	0.5328
DMGI	0.1598	0.5278	0.2874	0.5082
HsCTRD	**0.276**	**0.5833**	**0.4507**	**0.6311**

As shown in Table I, we observe that HsCTRD outperforms all the other baselines on two datasets, even compared with several semi-supervised ones. For Semi-supervised methods, the method designed for heterogeneous graphs, HAN, obtain better results than ones for homogeneous graphs, e.g., GCN and GAT. We also observe the same phenomenon in unsupervised methods, where DMGI and our proposed HsCTRD outperform DGI. This observation reflects that our proposed TAHIN can well model the rich semantics and multiple relations among entities in TCM formulae, and the diversity of the TAHIN should be carefully handled. In addition, HsCTRD is competitive with HAN, which is designed for heterogeneous graphs and utilize label guidance. This demonstrates that by maximizing MI to learn the representation, HsCTRD can explore more global structural information, while HAN bias toward local neighborhoods. For unsupervised methods, HsCTRD outperforms single-scale contrastive learning methods (i.e., DGI, GRACE and DMGI). Therefore, conducting contrastive learning on hybrid scales is effective. This well indicates the considerable potential of hybrid-scales contrastive learning.

2) Node Clustering and Similarity Search: For the intention of herbal relations discovery, we also conduct the clustering task to evaluate the node embeddings learned by our proposed HsCTRD. We feed the learned embeddings to K-Means algorithm. Because the clustering performance is greatly affected by the centroids, we repeat the process for 10 times. Finally, we adopt Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) as the evaluation metrics to evaluate the clustering results. To assess the ability of HsCTRD to find similar herbs, a top-k similarity search experiment was performed. The task of similarity search is defined as follows. We first calculate the cosine similarity scores of the learned embedding between all pairs of herbal nodes. For each node, we then rank the nodes based on the similarity score. Finally, we compute the percentage of nodes belonging to the same class among the top 5 nodes (Sim@5). The results were reported in Table II.

As shown in Table II, the proposed HsCTRD performs the best on ChP. While HsCTRD performs the second best on TCMRel, its performance is still very competitive with that of the best baseline HAN. Like the classification task, the methods designed for multiplex network achieve better

TABLE II
EXPERIMENT RESULTS FOR THE NODE CLUSTERING AND SIMILARITY SEARCH TASK.

Method	TCMRel			ChP		
	NMI	ARI	Sim@5	NMI	ARI	Sim@5
metapath2vec	0.1679	0.0063	0.0074	0.1127	0.02	0.083
GCN	0.3271	0.0996	0.2297	0.4045	0.1608	0.4335
GAT	0.3201	0.0898	0.2524	0.2957	0.092	0.3351
HAN	**0.3418**	0.1069	0.1194	0.4169	0.1817	0.4344
DGI	0.0793	0.0144	0.279	0.2521	0.0631	0.3843
GRACE	0.3212	0.0734	0.2387	0.3338	0.0824	0.3742
DMGI	0.26	0.07	0.2815	0.3505	0.152	0.4454
HsCTRD	0.3278	**0.1255**	**0.2906**	**0.4183**	**0.2223**	**0.4856**

performance. With the assistance of local and global contrastive learning, HsCTRD performs significantly better than DGI, GRACE, and DMGI, which only conduct single-scale contrastive learning. We observe that consistent with the classification task, metapath2vec exhibits the weakest performance. The most likely causes of this are that metapath2vec cannot deal with multiply semantic information simultaneously and ignores the node feature. Furthermore, please note that the performance of all methods on clustering and similarity search tasks is not dazzling. This is because we choose herbal efficacy as its label. However, the compatible herbs, which often have different efficacy, tend to be closer in the latent space. This will be elaborated in more detail in the case study section. Based on the above observation, we can find that the proposed HsCTRD can learn effective representations by considering hybrid-scale information of TAHIN and achieve significant improvements.

E. Ablation Study

We conduct ablation studies to explore the importance of different components of our HsCTRD. Specifically, we explore the model's capabilities from the following three questions. ($Q1$) Is hybrid-scales contrastive learning superior to single-scale comparative learning? ($Q2$) Does replacing the encoder with GAT improve performance? ($Q3$) Does the *semantic fusion component* really have a positive impact on performance? We report the results obtained from the ChP dataset on all three tasks in Table III. Here, $\text{HsCTRD}_{\text{ccl}}$ and $\text{HsCTRD}_{\text{scl}}$ denotes our proposed model only conducts the cross-scale or same-scale contrastive learning, respectively. $\text{HsCTRD}_{\text{gat}}$ switches to adopting GAT as the encoder f to learn embeddings. $\text{HsCTRD}_{\text{avg}}$ fuse the final node representation H by computing the average of the node representations learned from each single-view attributed network. i.e., $\mathbf{H} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}\mathbf{H}^{(r)}$. For $Q1$, comparing HsCTRD with $\text{HsCTRD}_{\text{ccl}}$ and $\text{HsCTRD}_{\text{scl}}$, we can see hybrid-scales contrastive learning, that benefits from capturing both local and global information, clearly contribute the superior performance over two single-scale comparative learning. For $Q2$, the results of $\text{HsCTRD}_{\text{gat}}$ and HsCTRD suggest that GCN encoder does help to improve HsCTRD by a large margin, especially in the node classification task. For $Q3$, results of $\text{HsCTRD}_{\text{avg}}$ are slightly better on the NMI and ARI metrics. These results are meaningful and reveal that since the $\text{HsCTRD}_{\text{avg}}$ averages the embedding of three single-view networks, this leads to a greater influence of HFH and HSH and makes the final node representations suitable for the clustering task. However, HsCTRD adopts an attention mechanism for fusion, which takes more into account the influence of HH and captures the semantic information of herbal compatibility. Thus the compatible herbs will be closer in the latent space, yet they often have various efficacy labels.

TABLE III
RESULTS OF THE ABLATION STUDY ON CHP DATASET

	Macro-F1	Micro-F1	NMI	ARI	sim@5
$\text{HsCTRD}_{\text{ccl}}$	0.3769	0.5738	0.4066	0.2454	0.4788
$\text{HsCTRD}_{\text{scl}}$	0.4139	0.6066	0.4008	0.2382	0.4804
$\text{HsCTRD}_{\text{gat}}$	0.2339	0.4918	0.3425	0.1914	0.4242
$\text{HsCTRD}_{\text{avg}}$	0.342	0.5738	**0.4299**	**0.2573**	0.342
HsCTRD	**0.4507**	**0.6311**	0.4183	0.2223	**0.4856**

F. Case Study for herbal compatibility

In order to verify the actual performance of HsCTRD in the scenario of discovering TCM formula regularities, We select two target herbs to perform a task of herbal similarity search under the guidance of herbalists. As shown in the Fig. 4, the results suggest that our proposed HsCTRD achieves superior performance compared with GRACE and DMGI. The homogeneous graph model GRACE only relies on a single-view network constructed by one meta path. So, its sim@5 mainly include compatible herbs (green font), and it does not handle the semantic information of TCM well. For the heterogeneous graph model DMGI, sim@5 achieves a relatively good performance. Nevertheless, it has a poor performance on sim@-5, which is the same as GRACE. For HsCTRD, the herbs in sim@5 include not only ones with similar efficacy (blue font), but also many compatible herbs. This indicates that our model does capture the TCM semantic information contained in HH, HFH, and HSH. Also, it explains the reason for the non-dazzling clustering performance at the same time. In addition, we observe that the herbs in sim@-5 contain many incompatible herbs (red font), which are not allowed to be used with the target herbs in clinical practice. The above analysis shows that HsCTRD does discover regularities for herbal compatibility, which can assist TCM practitioners with prescription preparation and pharmaceutical companies for finding new formulae.

Target Herb		草乌 (Kusnezoff Monkshood Root)	甘草(Licorice)
GRACE	sim@5	独活、雨羊花、川乌、朱砂根、油松节 (Pubescent Angelica, Rhododendron Mole, Radix Aconiti, Ardisia Crenata, Pine Nodular Branch)	干漆、全蝎、王不留行、蕲蛇、骨碎补 (Toxicodendron vernicifluum, Scorpio, Vaccaria segetalis, Polygonum aviculare, drynaria rhizome)
	sim@-5	青蒿、西红花、全蝎、天仙藤、薏苡仁 (Artemisia Annua, Crocus, Scorpio, Dutchmanspipe Vine, Coix Seed)	穿心莲、益母草、紫花地丁、漏芦、白蔹 (Andrographis Paniculate, Motherwort, Viola Yedoensis, Rhapontic um Uniflorum, Radix Ampelopsis)
DMGI	sim@5	川乌、制草乌、制川乌、独活、油松节 (Radix Aconiti, Processed Radix Aconiti Kusnezoffii, Processed Radix Aconiti, Rhododendron Mole, Pine Nodular Branch)	刀豆、丁公藤、丁香、五味子、巴豆 (Radix Ampelopsis, Erycibe Obtusifolia, Clove, Schisandra Chinensis, Croton Fruit)
	sim@-5	野马追、柿蒂、刀豆、炙黄芪、贝母 (Lindleyanum, Calyx Kaki, Croton Fruit, Prepared Astragalua Root, Fritillary)	鱼腥草、地锦草、黄藤、茴麻子、飞扬草 (Houttuynia, Humifuse Euphorbia, Caulis Fibraureae, Euphorbia hirta)
HsCTRD	sim@5	川乌、制草乌、独活、油松节、肉桂 (Radix Aconiti, Processed Radix Aconiti Kusnezoffii, Pubescent Angelica Root, Pine Nodular Branch, Cassia)	益智、厚朴、丁香、刀豆、大枣 (Alpiniae Oxyphyliae, Officinal Magnolia Bark, Officinal Magnolia Bark, Jack Bean, Jujube)
	sim@-5	瓜蒌皮、贝母、半夏、益智、莲子心 (Pericarpium Trichosanthis, Fritillary, Pinellia, Alpiniae Oxyphyliae, Lotus Plumule)	鱼腥草、甘遂、茴麻子、大戟、赤小豆 (Houttuynia, Euphorbia Kansui, Chingma Abutilon Seed, Euphorbia, Rice Bean)

Fig. 4. An example of herbal compatibility search.

VI. CONCLUSION

In this paper, we propose an effective model HsCTRD to discover regularities in TCM formula compatibility by

MI maximization. For modeling rich semantics and complex relation in TCM formulae and literature, we construct a novel TAHIN. We divide TAHIN into multiple single-view networks according to the TCM semantics, and on these networks, hybrid-scales contrastive learning is performed. We further present a semantic fusion component that can efficiently fuse the node representations learned in each network. In experiments, HsCTRD achieves state-of-the-art results on two real-world datasets in the node classification, clustering and similarity search tasks. Ablation study also demonstrates the rationality of the design of the HsCTRD's components. In order to verify the practical application value of the model, we perform a case study that further demonstrates the effectiveness of our proposed model HsCTRD in the task of TCM prescription regularities discovery.

ACKNOWLEDGMENT

This paper is funded by NSFC Grant No. 61672161, and Dongguan Innovative Research Team Program Grant No. 2018607201008).

REFERENCES

[1] L. Yao, Y. Zhang, B. Wei, W. Zhang *et al.*, "A topic modeling approach for traditional chinese medicine prescriptions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1007–1021, 2018.

[2] J. Zhu, Y. Liu, S. Yang, S. Zhai *et al.*, "A supervised learning framework for prediction of incompatible herb pair in traditional chinese medicine," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1799–1802.

[3] Q. Hu, T. Yu, J. Li, Q. Yu, L. Zhu, and Y. Gu, "End-to-end syndrome differentiation of yin deficiency and yang deficiency in traditional chinese medicine," *Computer methods and programs in biomedicine*, vol. 174, pp. 9–15, 2019.

[4] Y. Jin, W. Zhang, X. He *et al.*, "Syndrome-aware herb recommendation with multi-graph convolution network," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2020, pp. 145–156.

[5] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 91–100.

[6] Y. Fan, Y. Ye, Q. Peng, J. Zhang, Y. Zhang, X. Xiao, C. Shi, Q. Xiong, F. Shao, and L. Zhao, "Metagraph aggregated heterogeneous graph neural network for illicit traded product identification in underground market," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 132–141.

[7] Q. Zhong, Y. Liu, X. Ao, B. Hu, J. Feng, J. Tang, and Q. He, "Financial defaulter detection on online credit payment via multi-view attributed heterogeneous information network," in *Proceedings of The Web Conference 2020*, 2020, pp. 785–795.

[8] Y. Zhang, Y. Fan *et al.*, "Key player identification in underground forums over attributed heterogeneous information network embedding framework," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 549–558.

[9] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax." *ICLR (Poster)*, vol. 2, no. 3, p. 4, 2019.

[10] C. Park, D. Kim, J. Han, and H. Yu, "Unsupervised attributed multiplex network embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5371–5378.

[11] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.

[12] W. Li, Z. Yang, and X. Sun, "Exploration on generating traditional chinese medicine prescription from symptoms with an end-to-end method," *arXiv preprint arXiv:1801.09030*, 2018.

[13] Z. Wang, J. Poon, and S. Poon, "Tcm translator: A sequence generation approach for prescribing herbal medicines," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 2474–2480.

[14] C. Li, D. Liu, K. Yang, X. Huang, and J. Lv, "Herb-know: Knowledge enhanced prescription generation for traditional chinese medicine," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 1560–1567.

[15] X. Chen, C. Ruan, Y. Zhang, and H. Chen, "Heterogeneous information network based clustering for categorizations of traditional chinese medicine formula," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 839–846.

[16] H. Wan, M.-F. Moens, W. Luyten, X. Zhou, Q. Mei, L. Liu, and J. Tang, "Extracting relations from traditional chinese medicine literature via heterogeneous entity networks," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 356–365, 2016.

[17] C. Ruan, J. Ma, Y. Wang, Y. Zhang, Y. Yang, and S. Kraus, "Discovering regularities from traditional chinese medicine prescriptions via bipartite embedding model." in *IJCAI*, 2019, pp. 3346–3352.

[18] F. Yang, F. Xue, Y. Zhang, and G. Karypis, "Kernelized multitask learning method for personalized signaling adverse drug reactions," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[19] F. Yang, Q. Zhang, X. Ji, Y. Zhang, W. Li, S. Peng, and F. Xue, "Machine learning applications in drug repurposing," *Interdisciplinary Sciences: Computational Life Sciences*, 2021.

[20] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[21] F.-Y. Sun, J. Hoffmann, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," *arXiv preprint arXiv:1908.01000*, 2019.

[22] Z. Peng, W. Huang, M. Luo, Q. Zheng, Y. Rong *et al.*, "Graph representation learning via graphical mutual information maximization," in *Proceedings of The Web Conference 2020*, 2020, pp. 259–270.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[24] Y. Zhu, Y. Xu, F. Yu, Q. Liu *et al.*, "Deep graph contrastive representation learning," *arXiv preprint arXiv:2006.04131*, 2020.

[25] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[26] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[27] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *arXiv preprint arXiv:1906.00910*, 2019.

[28] S. Wan, S. Pan, J. Yang, and C. Gong, "Contrastive and generative graph convolutional networks for graph-based semi-supervised learning," *arXiv preprint arXiv:2009.07111*, 2020.

[29] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 271–279.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[31] Z. Wu, Y. Xiong *et al.*, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[32] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[33] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[34] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *The World Wide Web Conference*, 2019, pp. 2022–2032.

[35] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 135–144.

[36] X. Chu, B. Sun, Q. Huang, S. Peng, Y. Zhou, and Y. Zhang, "Quantitative knowledge presentation models of traditional chinese medicine (tcm): A review," *Artificial intelligence in medicine*, vol. 103, p. 101810, 2020.