A Progressively-Passing-then-Disentangling Approach to Recipe Recommendation

Chunlai Dong, Haochao Ying*, *Member, IEEE*, Renjun Hu, Yuyang Xu, Jintai Chen, Fuzhen Zhuang, Jian Wu, *Member, IEEE*

***Abstract*—The increasing popularity of online food blogs and food ordering services has made personalized recipe recommendation a vital aspect of our emotional well-being. However, existing solutions, mainly based on graph neural networks, still face significant challenges, such as (a) focusing on exploiting the user-recipe interactions while neglecting other crucial pairwise and high-order relationships, and (b) failing to explicitly distinguish the distinct factors, *e.g.,* hedonic and healthy, that influence recipe selection. To address these issues, we propose a progressively-passing-then-disentangling approach named P2D. Our approach utilizes a three-stage progressive message-passing mechanism for better representation learning. Specifically, we incorporate the extra pairwise relationships between recipes and nutrients, ingredients, and visual contents to create fine-grained and multimodal recipe representations. We next refine these representations via message passing between high-order recipe relationships to learn people's shared food preferences. Based on them, we could derive comprehensive user representations, which are subsequently transformed into disentangled forms that correspond to various decision factors through contrastive and mutual information regularization. Experimental results demonstrate both the superiority and the rationality of our method: (a) P2D outperforms the state-of-the-art recipe recommendation methods by a large margin under various metrics, (b) ablation studies confirm the positive impact of each of its components, and (c) our visualization analysis empirically supports the advantage of explicitly disentangling decision factors.**

***Index Terms*—Recipe recommendation, hypergraph neural network, disentangled representation learning.**

I. Introduction

With the abundance of information available on the Internet, it is easy to feel overwhelmed by the sheer volume of content. Recommender systems could substantially alleviate such information overload and among the many types of systems developed, recipe recommenders [1] are receiving growing interest owing to the following. First, food plays an essential role in our health and happiness [2], making recipe recommendations even more valuable. Additionally, with the rise of online food ordering and sharing services, making dietary choices has become a habitual part of our daily lives. Thirdly, many people still struggle with the challenge of eating healthily while also enjoying delicious food. To build a successful recipe recommendation model, it is crucial to comprehensively encode user and recipe information as representations that could facilitate a deeper understanding of recipes from various aspects and allow for learning of people's diverse food preferences.

Historical interactions have been used in many existing recommender systems to learn user preferences and make item recommendations [3], including recipes [4]. Early studies [5] have utilized collaborative filtering algorithms to learn people's food preferences by directly embedding recipes with their IDs. Motivated by the achievements of deep learning, researchers have turned to neural networks to aggregate multimodal information and learn better representations for recommendation [6]. More recently, graph neural networks (GNNs) have emerged as powerful tools for handling structural data. Note that users, recipes, and other related entities naturally form certain relationships, which could be consumed by GNNs to learn content and collaborative signals for recipe recommendation [7]. Finally, a recent study named SCHGN [8] proposes a heterogeneous GNN to further model the relationships between ingredients and the influence of food calories on users' comprehensive food preferences.

Although the models mentioned above have proven to be effective, they still face several key challenges. **(C1)** Most of the current methods focus solely on modeling the user-recipe interaction relationships, neglecting other pairwise and high-order relationships. It has already been verified that modeling recipes at the more fine-grained ingredient level could have a positive impact on recommendation performance [8]. Moreover, we argue in this paper that directly learning user-recipe interactions is prone to prioritize personalized food preferences while being ineffective to capture those shared patterns. Alternatively, the latter could be revealed by learning the high-order relationships between recipes. To sum up, this lack of attention to other potential relationships makes representations of users and recipes less expressive, leading to sub-optimal recommendation performance. **(C2)** Existing

Chunlai Dong, Yuyang Xu, and Jintai Chen are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310012, China. They are also with the State Key Laboratory of Transvascular Implantation Devices of the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China , and the Institute of Wenzhou, Zhejiang University, Hangzhou 325036, China. (E-mail: dongcl@zju.edu.cn, xuyuyang@zju.edu.cn, jtchen721@gmail.com).

Haochao Ying is with the State Key Laboratory of Transvascular Implantation Devices of the Second Affiliated Hospital and School of Public Health, Zhejiang University School of Medicine, Hangzhou 310009, China. He is also with the Institute of Wenzhou, Zhejiang University, Hangzhou 325036, China. (E-mail: haochaoying@zju.edu.cn)

Renjun Hu is an independent author. (E-mail: renjun0hu@gmail.com).

Fuzhen Zhuang is with the Institute of Artificial Intelligence, Beihang University, Beijing, 100191, China. He is also with Zhongguancun Laboratory, Beijing, 100191, China. (zhuangfuzhen@buaa.edu.cn).

Jian Wu is with the State Key Laboratory of Transvascular Implantation Devices of the Second Affiliated Hospital and School of Public Health, Zhejiang University School of Medicine, Hangzhou 310009, China. (E-mail: wujian2000@zju.edu.cn).

Manuscript received April 19, 2021; revised August 16, 2021.

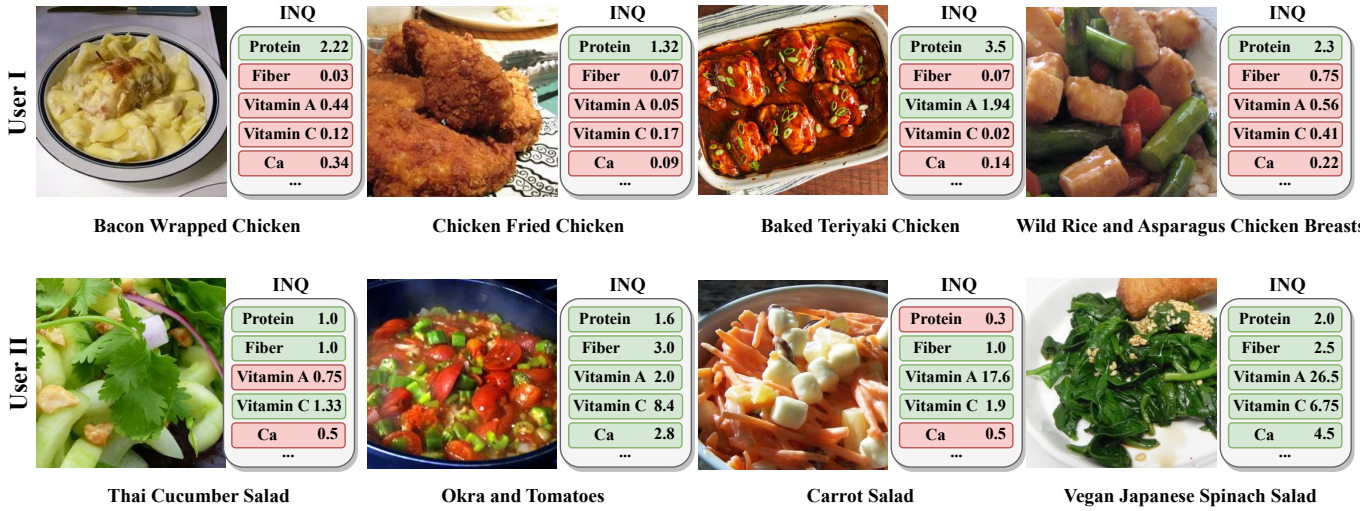


Fig. 1. Illustration of distinct factors, *i.e.,* healthy and hedonic, influencing people's choices for recipes. Based on the nutrition theory, each recipe is classified as healthy with respect to a nutrient (marked in green) if the corresponding INQ ≥ 1 and less healthy (marked in pink) otherwise.

approaches to recipe recommendation fail to explicitly and effectively account for the distinct key influencing factors for food choices. Psychological research [9], [10] has shown that users make trade-offs between two decision-making systems when choosing food, *i.e.,* healthy factors (rationality) and hedonic factors (sensibility). Take the two users and their recent recipe choices illustrated in Fig. 1 as an example. The healthiness of each recipe is assessed according to the nutrition theory such that it is regarded as healthy (marked in green) with respect to a nutrient if the corresponding index of nutritional quality (INQ) [11] is greater than or equal to 1 and less healthy (marked in pink) otherwise. As such, it is easy to observe from the figure that User I is more sensibility-driven while User II cares more about dietary healthiness. However, current recipe recommendations remain entangled and an attempt to decouple these unique decision factors is missing in the literature yet.

Present work. To tackle the aforementioned challenges, we creatively propose a progressively-passing-then-disentangling (namely P2D) approach for recipe recommendation. Our approach introduces two primary novelties. Firstly, we employ a three-stage progressive message-passing mechanism to effectively integrate a range of pairwise and high-order relationships that have an impact on recipe choices. Secondly, we decouple the blended user representations into disentangled forms, corresponding explicitly to the distinct influencing factors of food choices.

More specifically, the stage-one message passing considers the pairwise relationships between recipes and recipe attributes (says ingredients, nutrients, and visual contents), enabling multimodal and fine-grained recipe modeling. This step recovers the intrinsic correlations between related recipes, which are ignored by ID-based embedding. Additionally, the stage-two message passing further refines the recipe representations with hyper-GNN learning on the high-order recipe relationships. A hypergraph is built on recipes where each hyperedge connects the recipes consumed by a user. These hyperedges are user-agnostic in the sense that users only provide structural information while their identities are neglected. That is, each user creates an anonymous context within which the involved recipes interact to learn shared user preferences. An advantage of shared patterns over personalized ones is that they are more robust to user activeness imbalance and we expect our approach to performing better for less active users. Finally, comprehensive user representations are obtained through the stage-three message passing from recipes to users.

To better distinguish the decision factors, our approach also develops a disentangled representation learning module that transforms comprehensive user representations into disentangled forms. We utilize both contrastive learning and mutual information minimization to regularize the disentangling process. With either comprehensive or disentangled user representations, we adopt a hierarchical attention network [6] to estimate the affinity score, and multiple scores are finally combined to yield final recommendation results. In summary, the main contributions of this work are as follows:

- We propose a novel P2D approach for recipe recommendation. Regarding the P part, our approach seamlessly incorporates diverse pairwise and high-order relationships through a progressive message-passing mechanism to obtain fine-grained recipe encoding.
- We explicitly decouple user representations into disentangled forms to reflect the distinct healthy and hedonic factors that influence recipe-choosing decisions in the D part of P2D. We are among the first to introduce related psychological research to improve recipe recommendation performance.
- We evaluate the effectiveness and rationality of our approach on the Allrecipes dataset [6]. We find that P2D could outperform state-of-the-art methods by a large margin under various metrics. Our ablation studies also verify the positive impact of each component and the visualization analysis empirically supports the explicit disentangling of decision factors.

II. Related Work

In this section, we review related work on recipe recommendations, graph neural network-based recommendation, and

disentangled learning in recommender systems.

A. Recipe Recommendation

Recipe recommendation has become a critical domain for both individuals and society. Unlike other types of recommendation systems, recipe recommendation is more complex due to the multifaceted nature of food. When people choose a recipe, they are influenced by various factors such as ingredients, pictures, and cooking steps. This complexity makes it challenging to learn user preferences from historical interactions alone [12]. Technically speaking, recipe recommendation can be classified into three types: *collaborative filtering*, *content-based approaches*, and *hybrid approaches*.

Collaborative filtering-based methods frequently use classic Singular Value Decomposition (SVD) [13] and Matrix Factorization (MF) [14] to project interacted entities into latent embeddings for recommendation. For instance, Ge et al. [15] fused rating information and user tags with an MF approach for recipe recommendation and outperformed standard matrix factorization baselines. Trattner et al. [16] experimented with different collaborative filtering models and showed the superior performance of Latent Dirichlet Allocation (LDA) and weighted matrix factorization. Content-based approaches, on the other hand, focus on the information of food contents/attributes) on recommendations. For example, some studies [17], [18] considered the ingredients of a recipe as features and used them to make predictions. Hybrid approaches use collaborative signals and content information cooperatively for recommendation, enhancing the feature embeddings with more sources of information. The work in [6] proposed a model that captures user preferences using user-recipe interaction, recipe ingredients, and recipe image information. Other studies [7], [8] attempted to build a heterogeneous graph to explore relationships among users, recipes, and ingredients, employing message passing to obtain more comprehensive representations for recipe recommendation. However, these methods failed to incorporate richer multimodal recipe attribute information into the dataset and model implicit relationships between recipes. For instance, there might be commonalities among recipes interacted with by the same user.

In this work, we propose a novel model that attends to a range of pairwise and high-order relationships in the raw data and considers the psychological characteristics of users when making food choice decisions.

B. GNN-based Recommendation

Graph Neural Networks (GNNs) have gained significant attention in recent years for their ability to extract structured information and have been used in various fields [19], [20]. Some studies have built models using graph convolutional networks in the spectral domain, such as PinSage [21] and NGCF [22]. Alternatively, GNNs can be understood as to encode each user and item node into an embedding and update these representations by iteratively aggregating useful information from their neighbor nodes via message passing. Recent studies have leveraged GCNs to model more complex relationships besides user-item interactions, such as hierarchical structures based on user-outfit interactions and outfit-item mappings to aggregate item information into an outfit representation [23]. Our work constructs a heterogeneous graph that fuses different recipe attribute data and combines collaborative signals with recipe content information to improve representation learning.

Hyper-GNNs also offer expressive power in modeling high-order relationships [24], which has been utilized in some recent work to improve recommendation performance. For instance, MHCN [25] modeled high-order user relationships with a multi-channel hypergraph convolutional network, while DHCF [26] proposed a dual channel hypergraph collaborative filtering framework to model high-order correlations among subjects. Following these methods, we introduce a hypergraph neural architecture that models high-order recipe relationships to learn people's shared food preferences.

C. Disentangled Representation Learning

Recent years have seen the development of disentangled representation learning for recommender systems. For example, Ma et al. [27] learned disentangled representations of user behavior using Variational Auto-Ecoders (VAE), while Wang et al. [28] disentangled factors from user and item representations without knowing the meaning of those factors. Motivated by psychological research, our P2D approach incorporates rich contexts for explicit user intent factor decoupling and introduces contrastive learning and mutual information minimization methods to ensure the independence of learned factor-dependent user representations.

III. Preliminaries

In this section, we first introduce the task background and formalize our problem. Afterward, we present the embedding methods for heterogeneous entities involved in the task. The notations used in this paper are listed in Table I.

A. Background and Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ with cardinality $N_U = |\mathcal{U}|$ and $N_I = |\mathcal{I}|$ be the sets of users and recipes, respectively. The interaction matrix $\mathrm{Y} \in \mathbb{R}^{N_U \times N_I}$ records the interactions relationships between users and recipes such that $\mathrm{Y}_{ui} = 1$ indicates that user u has interacted with recipe i and $\mathrm{Y}_{ui} = 0$ otherwise. Each recipe i is associated with some attributes and we denote these attributes with a triple $\mathcal{F}_i = (v_i, g_i, m_i)$, where v_i, g_i, and m_i denote the image, ingredient, and nutrient attributes of recipe i, respectively. Formally, all three types of attributes are represented as vectors, with v_i being the output of an image encoder, $g_i \in \{0,1\}^{N_G}$ being an N_G-dimensional multi-hot binary encoding vector such that the k-th entry $g_i^k = 1$ if and only if recipe i contains ingredient k, and $m_i \in \mathbb{R}^{N_M}$ being an N_M-dimensional non-negative vector encoding the normalized nutrient quantities in recipe i. That is, we consider N_G ingredients and N_M nutrients for fine-grained recipe modeling. The task of recipe recommendation is then to predict a user's preferences of interaction on a recipe based on the interaction records and the recipe attribute features:

TABLE I
SUMMARY OF MATHEMATICAL NOTATIONS

Notation	Description
$\mathcal{U}$, $\mathcal{I}$	Sets of users and recipes
u, N_U	A user and the total number of users
i, N_I	A recipe and the total number of recipes
Y	The interaction matrix between users and recipes
N_G, N_M	The total number of ingredients and nutrients
$\mathcal{F}_i = (v_i, g_i, m_i)$	Visual, ingredient, and nutrient attributes of recipe i
$s_{(u,i)}$	Recommendation score of user u and recipe i
e_u, e_i, e_g, e_m	Embedding vector of different entities
e_{v_i}, e_{g_i}, e_{m_i}	Attribute representations of recipe i
$\mathcal{G} = (\mathcal{I}, \mathcal{E})$	Hypergraph of recipes
$\mathcal{N}(\cdot)$	A set of associated entities

- **Input:** Set of users $\mathcal{U}$ and recipes $\mathcal{I}$, user-recipe interaction matrix Y, recipes image vectors $[v_1, \dots v_{N_I}]$, recipes ingredient vectors $[g_1, \dots g_{N_I}]$, and recipe nutrient vectors $[m_1, \dots m_{N_I}]$.
- **Output:** A predictive function $s_{(u,i)} = f(u, i, v_i, g_i, m_i, \mathrm{Y})$ whose output is the estimated score that user u would interact with the recipe i.

B. Heterogeneous Entity Embedding

Before diving deep into our proposed approach, we explain the embedding method that helps to unify the above heterogeneous entities in the same framework. Following the common paradigm for recommender systems, we represent each user u and recipe i with the embedding vectors $e_u \in \mathbb{R}^d$ and $e_i \in \mathbb{R}^d$, respectively. Here d refers to the embedding dimension. Formally, let $h_u \in \mathbb{R}^{N_U}$ and $h_i \in \mathbb{R}^{N_I}$ be the corresponding one-hot encoding of the user and recipe IDs. The user and recipe embedding vectors are derived by:

$$e_u = \mathbf{E}_U^\top h_u, e_i = \mathbf{E}_I^\top h_i, \tag{1}$$

where matrices $\mathbf{E}_U \in \mathbb{R}^{N_U \times d}$ and $\mathbf{E}_I \in \mathbb{R}^{N_I \times d}$ denote the embedding matrices for all users and recipes. These embedding matrices are usually initialized randomly and learned simultaneously with the predictive function f. Similarly, we can represent each (says the k-th) ingredient and nutrient with the ID-based embedding vectors e_g^k and e_m^k. The ingredient representation of recipe i is then obtained as the average of the associated ingredient embeddings:

$$e_{g_i} = \text{mean}\{e_g^k \mid g_i^k = 1\}. \tag{2}$$

And the nutrient representation of recipe i is derived as the weighted sum of all nutrient embeddings. Note that m_i^k is the normalized quantity of the k-th nutrient contained in recipe i, which is a fixed value, but not a model parameter:

$$e_{m_i} = \sum_k m_i^k e_m^k. \tag{3}$$

In addition, food images also largely influence people's choices for food and contain a wealth of information [29]. Taking inspiration from recent successes in computer vision, we utilize a pre-trained CLIP [30] model, which is a vision model trained with natural language as supervision signals, to extract image features from the original image. We adopt the output v_i of CLIP's image encoder as the image attribute of recipe i, which is a 512-dimensional vector. For compatibility with other entities, we further use a linear layer to project v_i into a d-dimensional vector, which is expressed as:

$$e_{v_i} = W v_i + b. \tag{4}$$

Here, $W \in \mathbb{R}^{d \times 512}$ and $b \in \mathbb{R}^d$ are trainable parameters[1] and $e_{v_i} \in \mathbb{R}^d$ is the image embedding. To summarize, we create four types of entity embeddings, *i.e.,* e_u, e_i, e_g, and e_m, and three types of attribute representations for recipes, *i.e.,* e_{v_i}, e_{g_i}, and e_{m_i}. These vectors are of the same dimensionality and could be directly manipulated by our model introduced in the next section.

IV. METHODOLOGY

We now present our proposed progressively-passing-then-disentangling (P2D) approach for recipe recommendation. The framework overview is illustrated in Fig 2, which consists of three key modules:

- **Progressive Representation Learning** utilizes a three-stage message-passing mechanism to effectively integrate a range of pairwise and high-order relationships that have an impact on recipe choices. More specifically, this enables us to integrate the massive multimodal and fine-grained information into recipe representations and learn shared food preferences through the high-order relationships between recipes.
- **Disentangled Representation Learning** further decouples the comprehensive and blended user representations into disentangled forms such that each of them corresponds explicitly to a distinct influencing factor for food choices. This module is inspired by psychological studies and we adopt contrastive learning and mutual information minimization to regularize the disentangling process.
- **Personalized Recipe Recommendation** is responsible for producing the final recommendation result for each user-recipe pair. It first estimates multiple affinity scores with a hierarchical attention network and each score is calculated based on a user representation (either blended or disentangled) and other entity embeddings. These scores are then fed into an MLP to derive the final result.

We next introduce these modules in detail.

A. Progressive Representation Learning

The progressive representation learning module implements a hierarchical three-stage message passing process, as illustrated in Fig. 2(a). Specifically, for a recipe $i \in \mathcal{I}$, it first integrates the various multimodal information into recipe representations to obtain an attribute-aware recipe embedding e_i' (stage-one). It further refines the recipe embedding e_i' with a hyper-GNN (stage-two). Finally, it obtains a comprehensive user representation $\widetilde{e_u}$ for each user $u \in \mathcal{U}$ in the stage-three message passing from recipes to users.

[1]When it is clear from the context, we will omit the description of trainable parameters in the rest of the paper. Note that due to various types of entities and message-passing stages, we do not strictly distinguish trainable parameters with subscriptions. In other words, distinct parameters in different equations might appear under the same notations.

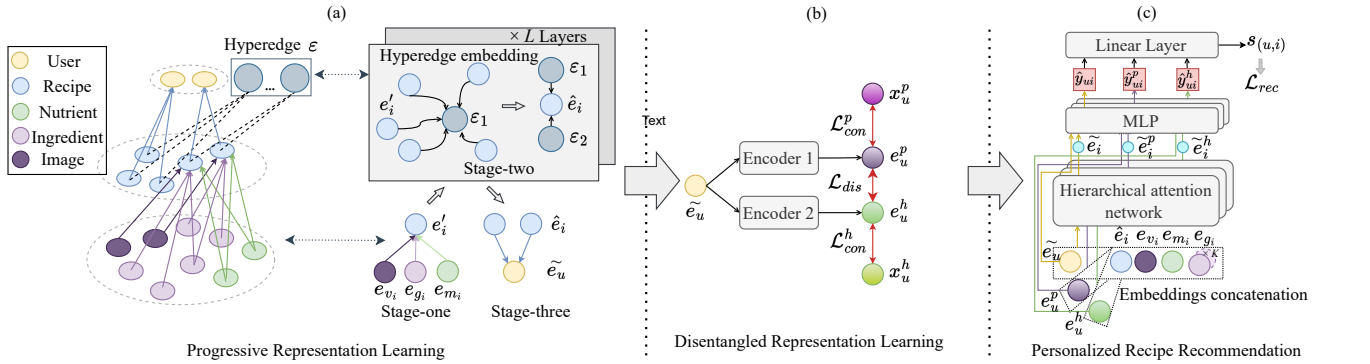


Fig. 2. Framework overview of our progressively-passing-then-disentangling (P2D) approach. It consists of three key modules: progressive representation learning for better exploiting the relationships crucial for the task, disentangled representation learning for explicitly modeling the distinct decision factors (*i.e.,* healthy and hedonic), and personalized recipe recommendation for comprehensive recommendation generation.

Stage-one. Recall that each recipe i is associated with a triple $\mathcal{F}_i = (v_i, g_i, m_i)$ of visual, ingredient, and nutrient attributes, and we derive e_{v_i}, e_{g_i}, and e_{m_i} as the corresponding attribute representations of recipe i. We explore these intrinsic pairwise relationships between recipes and recipe attributes to obtain attribute-aware recipe embeddings. This creates a connection between related recipes that have the same ingredients/nutrients or look similar, which is not available in ID-based embeddings. This stage is implemented by a simple message passing operator [31] as follows:

$$e_i{}' = W_1 e_i + W_2 e_{v_i} + W_3 e_{g_i} + W_4 e_{m_i}, \tag{5}$$

where $W_* \in \mathbb{R}^{d\times d}$, $e_i \in \mathbb{R}^d$ is the raw ID projection embedding of recipe i, and its counterpart $e_i{}' \in \mathbb{R}^d$ is enhanced with multimodal and fine-grained attribute information.

Stage-two. Motivated by the strength of hypergraph learning [32], we also adopt a hyperGNN to exploit the implicit high-order relationships between recipes. It helps to refine the recipe representation, denoted as $\hat{e}_i$ of recipe i, to encode people's shared food preferences.

More specifically, we first build a hypergraph $\mathcal{G} = (\mathcal{I}, \mathcal{E})$ on recipes where the node set is the recipe set and the hyperedge $\varepsilon_u \in \mathcal{E}$ connects all recipes consumed by user u in the training set, as illustrated in Fig. 2(a). We set the weights of the hyperedges to be equal for simplicity. Note that these hyperedges are user-agnostic in that users only provide structural information and their identities are ignored when building the hypergraph. Then, we adopt a convolutional operator [33] to achieve our hypergraph message passing as below:

$$C_{\varepsilon_u} = W_1 \text{mean}\{e_i' \mid i \in \mathcal{N}_I(u)\}, \tag{6}$$

$$\hat{e}_i = W_2 e_i' + W_3 \text{mean}\{C_{\varepsilon_u} \mid \varepsilon_u \in \mathcal{N}_E(i)\}. \tag{7}$$

Here, $W_* \in \mathbb{R}^{d\times d}$, and (a) C_{ε_u} is the embedding of hyperedge ε_u, (b) e_i' is the recipe embedding obtained from stage-one message passing, (c) $\mathcal{N}_I(u)$ is the set of recipes consumed by user u, *i.e.,* connected by the hyperedge ε_u, (d) $\mathcal{N}_E(i)$ is the set of hyperedges containing recipe i. That is, we first aggregate information from recipes to hyperedges, which reflects certain user preference patterns, and such piece of information on hyperedges is subsequently passed back to recipes for refinement. Note that the above two-round updates could be repeated in multiple L layers with each layer using the latest recipe representations for message passing, which enables recipes to receive information from multi-hop neighbors.

Finally, it is worth noting that, compared with traditional bipartite GNNs, our message passing in stage-two remains user-agnostic, *i.e.,* no user information is incorporated in Eq. (6). This brings a significant advantage to our P2D approach in that it is allowed to learn common user preferences for recipes in a reserved stage in our model pipeline.

Stage-three. The message passing of this stage exploits the user-recipe collaborative relationships. As illustrated in Fig. 2(a), we use another simple message passing operator similar to the stage-one to generate the comprehensive user representation $\widetilde{e_u}$ of user u, which can be formulated as:

$$\widetilde{e_u} = W_1 e_u + W_2 \text{mean}\{\hat{e}_i \mid i \in \mathcal{N}_I(u)\}, \tag{8}$$

where $W_* \in \mathbb{R}^{d\times d}$, $\hat{e}_i$ is the representation of recipe i from stage-two, $\mathcal{N}_I(u)$ is the set of recipes consumed by user u, and e_u/$\widetilde{e_u}$ are the raw/enhanced embedding of u.

Through our progressive representation learning module, we can obtain fine-grained recipe representations and comprehensive user representations. In particular, $\hat{e}_i$ aggregates multi-modal recipe attribute features, and $\widetilde{e_u}$ contains the recipe preference of user u, which also aggregates the user's preference for recipe attributes (ingredients, nutrients, and visual contents) with recipes as the bridge.

B. Disentangled Representation Learning

We next devise a disentangled representation learning module to transform the comprehensive user representation $\widetilde{e_u}$ into disentangled forms that correspond to the healthy factor e_u^h and hedonic factor e_u^p of user u in recipe choices. As illustrated in Fig. 2(b), we utilize two encoders (*i.e.,* MLP) to generate the disentangled representations, in which the disentangling process is aligned by two types of regularization losses according to prior psychological research [9], [10].

To separate and extract the healthy preference, we turn to the nutrient consumption in users' historical recipe choices. Specifically, we aggregate the nutrient preference of user u as the average of the nutrient attribute representations of her/his chosen recipes:

$$x_u^h = \text{mean}\{e_{m_i} \mid i \in \mathcal{N}_I(u)\}. \tag{9}$$

We then maximize the similarity between the generated e_u^h and x_u^h under the contrastive learning paradigm. That is, e_u^h should be more similar with x_u^h than $x_{u'}^h$ of other users. Formally, for user u, we minimize the contrastive objective $\mathcal{L}_{con}^h$ based on InfoNCE loss [34]:

$$\mathcal{L}_{con}^h = -\log \frac{\exp({e_u^h}^\top x_u^h/\tau)}{\sum_{u' \in \mathcal{B}_u} \exp({e_u^h}^\top x_{u'}^h/\tau)}, \tag{10}$$

where $\mathcal{B}_u$ is the set of random users in the same batch as u during optimization and τ is a temperature parameter. Similarly, we could separate and extract hedonic factor e_u^p with another contrastive objective $\mathcal{L}_{con}^p$, which utilizes the average of image attribute representations, *i.e.,* $x_u^p = \text{mean}\{e_{v_i} \mid i \in N_I(u)\}$, for regularization. It is noteworthy that our work drew this inspiration from the research outlined in [35], highlighting the substantial impact of visual cues on consumer emotions, hedonic perceptions, and food purchase intentions. This aligns seamlessly with our approach of employing images to reflect users' hedonic preferences.

Besides, to disentangle the healthy factor e_u^h and hedonic factor e_u^p, we further minimize the Mutual Information (MI) between them. However, estimating and minimizing MI in high-dimensional spaces of e_u^h and e_u^p is a great challenge. Then following [36], we perform the MI minimization algorithm using the vCLUB-based MI upper bound estimator which can be formulated as follows:

$$\begin{aligned} \mathrm{I}_{\mathrm{vCLUB}}(e_u^h; e_u^p) = \mathbb{E}_{p(e_u^h, e_u^p)}[\log q_\theta(e_u^p|e_u^h)] \\ - \mathbb{E}_{p(e_u^h)}\mathbb{E}_{p(e_u^p)}[\log q_\theta(e_u^p|e_u^h)]. \end{aligned} \tag{11}$$

Note that the variational distribution $q_\theta(e_u^p|e_u^h)$ with parameter θ is to approximate the conditional distribution $p(e_u^p|e_u^h)$ due to the conditional relation between variables is unavailable.

Specifically, we use two steps to implement the MI minimization algorithm like [36]. Firstly, we train the parameters θ to get a good variational approximation $q_\theta(e_u^p|e_u^h)$, in order to estimate the MI upper bound. Secondly, we freeze θ and minimize the upper bound $\mathcal{L}_{dis}$ by training other parameters in our model with loss function as follows:

$$\mathcal{L}_{dis} = \mathrm{I}_{\mathrm{vCLUB}}(e_u^h; e_u^p). \tag{12}$$

C. Personalized Recipe Recommendation

Based on the representations computed in the previous two modules, we now introduce the details of personalized recipe recommendation, as illustrated in Fig. 2(c). Specifically, for a user-recipe pair (u, i), we estimate three affinity scores $(\hat{y}_{ui}, \hat{y}_{ui}^h, \hat{y}_{ui}^p)$, each of which corresponds to a blended or disentangled user representation $(\widetilde{e_u}, e_u^h, e_u^p)$. These scores are then concatenated and fed into a linear layer to yield the final recommendation result. Note that affinity scores are estimated in the same manner except for using different user embeddings, and we only elaborate the computation of $\hat{y}_{ui}$ based on $\widetilde{e_u}$ in the following.

For accurate personalized recipe recommendation, the model should be aware of the personalized and diverse preferences of users on various aspects of recipes. For example, two users may have different interests in ingredients even though they have consumed the same recipe. At the attribute granularity, different users are also likely to express different interests in the visual, ingredient, and nutrient attributes of the same recipe. Inspired by these, we exploit a hierarchical attention network [6] to obtain user-specific recipe representations, which could capture users' various preference patterns.

The hierarchical attention network consists of two layers. In the first attention layer, we capture users' preferences for different ingredients, *i.e.,* aggregating the set of ingredient representations e_g^k contained in recipe i (that is, $g_i^k = 1$) into a single representation $\hat{e}_{g_i} \in \mathbb{R}^d$ as follows :

$$\hat{e}_{g_i} = \sum_{k:\ g_i^k = 1} \alpha(\widetilde{e_u}, e_g^k) e_g^k, \tag{13}$$

where $\alpha(\widetilde{e_u}, e_g^k)$ is the user-specific attention weight for the k-th ingredient e_g^k. Specifically, the attention weights are calculated as follows with the projection $W_* \in \mathbb{R}^{d \times d}$:

$$\begin{aligned} a(\widetilde{e_u}, e_g^k) &= v_1^\top \tanh(W_1 \widetilde{e_u} + W_2 e_g^k + b_1), \\ \alpha(\widetilde{e_u}, e_g^k) &= \frac{\exp(a(\widetilde{e_u}, e_g^k))}{\sum_{k':\ g_i^{k'}=1} \exp(a(\widetilde{e_u}, e_g^{k'}))}. \end{aligned} \tag{14}$$

By far we have obtained the user embedding $\widetilde{e_u}$, recipe embedding $\hat{e}_i$, and three recipe attribute representations $\hat{e}_{g_i}$, e_{v_i} and e_{m_i}. In the second attention layer, we explore the user preferences at the attribute granularity by attentively aggregating the four recipe-related embeddings:

$$\widetilde{e_i} = \sum_{e \in \{\hat{e}_i, \hat{e}_{g_i}, e_{v_i}, e_{m_i}\}} \beta(\widetilde{e_u}, e) e. \tag{15}$$

Similarly, $\beta(\widetilde{e_u}, e)$ is the attention weight to reflect the different interests of users for attributes, *e.g.,* one may favor good nutrition while another may care more about recipe taste. With the projection $W_* \in \mathbb{R}^{d \times d}$, these attention weights can be calculated as:

$$\begin{aligned} b(\widetilde{e_u}, e) &= v_2^\top \tanh(W_3 \widetilde{e_u} + W_4 e + b_2), \\ \beta(\widetilde{e_u}, e) &= \frac{\exp(b(\widetilde{e_u}, e))}{\sum_{e' \in \{\hat{e}_i, \hat{e}_{g_i}, e_{v_i}, e_{m_i}\}} \exp(b(\widetilde{e_u}, e'))}. \end{aligned} \tag{16}$$

To estimate the affinity score with regard to the blended user representation $\widetilde{e_u}$, we concatenate it with the user-specific recipe representation $\widetilde{e_i}$ in Eq. (15) and their element-wise product $\widetilde{e_u} \odot \widetilde{e_i}$, and utilize an MLP to combine the user, recipe, and interaction information:

$$\hat{y}_{ui} = h^\top \mathrm{ReLU}(W \begin{bmatrix} \widetilde{e_u} \\ \widetilde{e_i} \\ \widetilde{e_u} \odot \widetilde{e_i} \end{bmatrix} + b_3). \tag{17}$$

Here, $h \in \mathbb{R}^d$, $W \in \mathbb{R}^{d \times 3d}$, and $b_3 \in \mathbb{R}^d$ are learnable parameters. As for e_u^h and e_u^p, we could use the same architecture to calculate the affinity scores $\hat{y}_{ui}^h$ and $\hat{y}_{ui}^p$, which emphasize the impacts of different decision factors in the recipe choosing process. Finally, we concatenate these scores into a 3-dimensional vector $[\hat{y}_{ui}, \hat{y}_{ui}^h, \hat{y}_{ui}^p]$ and feed it into a linear layer to yield the final recommendation score:

$$s_{(u,i)} = \mathrm{LinearLayer}([\hat{y}_{ui}, \hat{y}_{ui}^h, \hat{y}_{ui}^p]) \in \mathbb{R}. \tag{18}$$

TABLE II
DATASET STATISTICS

Description	Allrecipes
# of users	68,768
# of recipes	45,630
# of ingredients	33,147
# of nutrients	20
# of interactions	1,093,845

D. Model Training

In this study, we tackle the recipe recommendation task under the ranking paradigm. We employ a pairwise ranking objective to learn the model parameters. The underlying assumption is that for a given user u, the model should score higher for items i^+ with observed interactions than those i^- without interactions. We leverage the Bayesian Personalized Ranking (BPR) [37] optimization criterion as the main loss for recipe recommendation:

$$\mathcal{L}_{rec} = \sum_{(u,i^+,i^-)\in\mathcal{D}} -\ln \sigma(s_{(u,i^+)} - s_{(u,i^-)}), \qquad (19)$$

where $\mathcal{D}$ denotes the training set and $\sigma(\cdot)$ is the logistic sigmoid function.

Combining the recommendation loss with the aforementioned disentangling regularization objectives, we optimize our model by minimizing the following:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha \cdot (\mathcal{L}^h_{con} + \mathcal{L}^p_{con} + \mathcal{L}_{dis}) + \lambda \left\|\Theta\right\|_2, \qquad (20)$$

where α and λ are hyper-parameters to balance different losses and $\left\|\Theta\right\|_2$ is the L_2 regularization term for weight decay. The learning process of P2D with pseudocode is summarized in Algorithm 1.

V. EXPERIMENTS

In this section, we conduct experiments on a large real-world recipe dataset to evaluate the effectiveness of our proposed P2D approach. Through our extensive tests, we aim to answer the following research questions:

- **RQ1:** How does our proposed P2D perform for the task, compared with state-of-the-art methods?
- **RQ2:** How do the key components of our model impact the recommendation performance?
- **RQ3:** How do the key hyper-parameters influence P2D and is it easy to tune these parameters?
- **RQ4:** How does P2D benefit from the disentangled representation learning module?

A. Experimental Settings

Dataset. The experiments are conducted on the Allrecipes dataset collected by Gao et al. [6] from a real-world website Allrecipes.com. It is one of the largest food-oriented online social networks with 225 million page views per month. This dataset contains 1,093,845 user-recipe interactions, where recipes are associated with images and detailed information about their ingredients and nutrients.[2] Some statistical information about the dataset is summarized in Table II. Following previous work [6], we use the latest 30% of interaction records as the test set, and the remaining data are split into 60% and 10% for training and validation, respectively.

Evaluation Protocols. Due to the large number of user-recipe pairs in the dataset, evaluating the recommendation results for the entire pair set is prohibited. Therefore, following a

[2]During the preparation of the manuscript, this is the only dataset that is suitable to test our approach.

Algorithm 1 Algorithm Flow

Inputs:
Set of users $\mathcal{U}$ and recipes $\mathcal{I}$, user-recipe interaction matrix Y, recipes image vectors $[v_1, \dots v_{N_I}]$, recipes ingredient vectors $[g_1, \dots g_{N_I}]$, recipe nutrient vectors $[m_1, \dots m_{N_I}]$, embedding dimension d, and hyper-parameters $\{\alpha, \lambda\}$.

Outputs:
Trained parameters Θ.

Progressive representation learning module:
  Stage-one: Aggregate information of recipes into the embedding e'_i according to Equation (5).
  Stage-two: Construct hyperedges and apply convolutional function to refine the recipe representation $\hat{e}_i$ according to Equation (6,7).
  Stage-three: Generate user representation $\widetilde{e_u}$ of user u according to Equation (8).
Disentangled representation learning module:
  Generate the disentangled user representation e^h_u and e^p_u.
  Calculate the contrastive loss $\mathcal{L}^h_{con}$, $\mathcal{L}^p_{con}$ and minimize the mutual information via the loss $\mathcal{L}_{dis}$ according to Equation (10-12).
Personalized recipe recommendation module:
  Perform hierarchical attention to inject diverse preferences of users on various aspects of recipes into the user-specific recipe representation $\widetilde{e_i}$ according to Equation (13-16).
  Calculate the final estimated score $s_{(u,i)}$ according to Equation (17,18).
Main process:
Initialize all parameters in Θ.
for each epoch **do**
  for each user-recipe interaction (u, i) and its corresponding triplet $\mathcal{F}_i = (v_i, g_i, m_i)$ **do**
    Calculate embeddings $(e_u, e_i, e_{v_i}, e_{g_i}, e_{m_i})$
    Conduct **progressive representation learning module** to obtain $(\widetilde{e_u}, \hat{e}_i)$
    Perform the **disentangled representation learning module** to generate features e^h_u and e^p_u related to the psychological aspects of user u.
    Product the final **recipe recommendation** result for the user-recipe interaction, with the estimated score $s_{(u,i)}$.
    Calculate the overall loss $\mathcal{L}$ according to Equation (19,20).
  end for
  Update θ to get a good variational approximation $q_\theta(e^p_u|e^h_u)$ according to Equation (11).
  Freeze θ and update other parameters of the model via minimizing loss $\mathcal{L}$ according to Equation (20).
end for

recent recipe recommendation study [8], we randomly sample 500 negative recipes for each observed interaction pair. Notice that the negative samples are drawn from the recipes that the users have not interacted with ever. We adopt three commonly-used evaluation metrics: Area Under the Roc Curve (AUC), Normalized Discounted Cumulative Gain (NDCG@K), and Recall@K with $K \in \{10, 20, 50\}$. For all three metrics, higher values indicate better performance.

Baselines. We compare our P2D approach with a variety of baselines, including both classic recommendation and recent recipe-oriented methods, to validate its effectiveness.

- **BiasMF [14]** augments traditional matrix factorization models with user and item bias vectors to better capture the observed information.
- **FM [38]** combines the advantages of Support Vector Machines (SVM) with factorization models. It could model feature interactions under a high degree of sparsity, which is particularly useful for recommender systems.
- **NCF [3]** develops the first deep learning architecture for collaborative filtering. It replaces the inner product operation in traditional CF models with neural networks to enhance model expressiveness.
- **LightGCN [39]** proposes a simplified Graph Convolution Network (GCN) framework, which only contains the most essential neighborhood aggregation component of GCN, and retrains similar effectiveness.
- **DGCF [28]** learns to disentangle implicit influencing factors for recommendations. It explores fine-grained user intent by disentangling user interactions into multiple latent factor representations.
- **HAFR [6]** comprehensively captures the impacts of different factors (*e.g.,* user-recipe interactions, recipe ingredients, and recipe images) on users' food choices, using the hierarchical attention mechanism.
- **SCHGN [8]** uses a heterogeneous graph to combine user, recipe, ingredient, and calorie information for representation learning and explore complex relationships between ingredients in different recipes via self-supervised ingredient prediction.

Implementation details. We implement all tested approaches with PyTorch. Each model is optimized with the Adam optimizer [40] with a learning rate of $1e^{-3}$, epoches of 50, and batch size of $2,048$ on a single 11G RTX 2080 Ti GPU server. For a fair comparison, we fix the embedding dimension $d = 64$ for all approaches and the weight decay parameter $\lambda = 0.1$. Besides, we grid search the balancing weight α in $\{0.1, 0.5, 1, 5, 10, 20, 30, 50, 100, 200\}$.

We next present our findings.

B. Overall Performance Comparison (RQ1)

To evaluate the overall performance, we optimize all tested models with training and validation data, collect the recommendation results on the test set, and compute the metrics. The results of all approaches are summarized in Table III, from which we observe the following.

Among all baselines, BiasMF and NCF perform the worst. This is because both methods only exploit the user-recipe interaction relationships while ignoring other important features of recipes. This has empirically justified the necessity of incorporating recipe content features into modeling for the recipe recommendation task. Classical recommendation methods that solely model interaction relationships perform significantly worse compared to subsequent recommendation methods that explicitly introduce multimodal recipe attribute information for modeling.

When feeding the same set of input features, it could be found that methods based on Graph Neural Networks generally perform better than others. For instance, by learning from the user-recipe interactions, LightGCN is better than BiasMF and NCF under all tested metrics. We note that the repeated message-passing process along neighboring nodes enables capturing more complex relational patterns, which are crucial for personalized recommendations. In addition, SCHGN performs the best among all baselines, which also supports the advantage of fusing multi-modal features with GNNs such that the fine-grained and complex relationships could be encoded in representations to improve the performance of recipe recommendation.

Moreover, we find that our proposed P2D model consistently outperforms all baselines under the seven metrics. Specifically, P2D is approximately 6.0%, 38.6%, and 30.7% better than the state-of-the-art recipe recommendation method (*i.e.,* SCHGN) under AUC, NDCG@10 and Recall@10, respectively. We attribute the improvement to two reasons. First, our model implements a three-stage progressive message-passing process between involved entities, which effectively injects a variety of preferential patterns into user and recipe representations. Second, through disentangled representation learning, our model could explicitly ascribe people's recipe choices to different influencing decision factors, *e.g.,* rationality and sensibility. This is the first application of psychological research in recipe recommendations.

Finally, we point out that the improvement of our model is less significant in AUC than in NDCG@K and Recall@K, compared with existing solutions. This is due to that both NDCG and Recall focus on the top-K recommended recipes while AUC is evaluated on the entire ranking. This also explains that the relative improvement in terms of NDCG@K and Recall@K also decreases with the increment of K.

C. Time efficiency Comparison (RQ1)

We provide the time efficiency of our method and the baselines in Table V. It becomes apparent that although our model necessitates more training time compared to the baseline models, it is important to note that this training occurs offline. The cumulative training duration of 5.53 hours is generally deemed acceptable, especially when considering the vast dataset consisting of millions of interactive records. Regarding inference time, our model aligns with the same order of magnitude as the best-performing baseline model, SCHGN. Notably, our proposed P2D demonstrates substantial performance improvements across various metrics when compared to its counterparts. We attribute the increase in training time primarily to the introduced disentangled rep-

TABLE III
OVERALL PERFORMANCE OF RECIPE RECOMMENDATION. THE REPORTED NUMBERS OF HAFR AND SCHGN ARE BORROWED FROM THEIR ORIGINAL PAPERS [6], [8] AS WE COULD NOT REPRODUCE THEIR RESULTS.

Model	AUC	NDCG@10	NDCG@20	NDCG@50	Recall@10	Recall@20	Recall@50
Classic Recommendation Methods							
BiasMF	$0.511_{(\pm 0.001)}$	$0.036_{(\pm 0.002)}$	$0.048_{(\pm 0.003)}$	$0.073_{(\pm 0.003)}$	$0.053_{(\pm 0.004)}$	$0.091_{(\pm 0.007)}$	$0.190_{(\pm 0.010)}$
NCF	$0.521_{(\pm 0.005)}$	$0.037_{(\pm 0.001)}$	$0.050_{(\pm 0.002)}$	$0.076_{(\pm 0.002)}$	$0.057_{(\pm 0.001)}$	$0.101_{(\pm 0.003)}$	$0.202_{(\pm 0.004)}$
FM	$0.571_{(\pm 0.003)}$	$0.040_{(\pm 0.001)}$	$0.054_{(\pm 0.002)}$	$0.079_{(\pm 0.002)}$	$0.061_{(\pm 0.002)}$	$0.106_{(\pm 0.004)}$	$0.211_{(\pm 0.009)}$
DGCF	$0.581_{(\pm 0.005)}$	$0.041_{(\pm 0.003)}$	$0.055_{(\pm 0.004)}$	$0.083_{(\pm 0.004)}$	$0.062_{(\pm 0.004)}$	$0.109_{(\pm 0.005)}$	$0.213_{(\pm 0.008)}$
LightGCN	$0.592_{(\pm 0.001)}$	$0.043_{(\pm 0.001)}$	$0.058_{(\pm 0.002)}$	$0.088_{(\pm 0.002)}$	$0.063_{(\pm 0.002)}$	$0.110_{(\pm 0.002)}$	$0.224_{(\pm 0.004)}$
Explicitly Content-Oriented Recommendation Methods							
HAFR	0.644	0.046	0.060	0.090	0.067	0.116	0.225
SCHGN	0.721	0.057	0.077	0.117	0.088	0.157	0.313
P2D	$\mathbf{0.764}_{(\pm 0.004)}$	$\mathbf{0.079}_{(\pm 0.004)}$	$\mathbf{0.101}_{(\pm 0.004)}$	$\mathbf{0.141}_{(\pm 0.005)}$	$\mathbf{0.115}_{(\pm 0.006)}$	$\mathbf{0.187}_{(\pm 0.008)}$	$\mathbf{0.350}_{(\pm 0.011)}$
Improvement	+6.0%	+38.6%	+31.2%	+20.5%	+30.7%	+19.1%	+11.8 %

TABLE IV
ABLATION STUDY OF P2D

Model variants	AUC	NDCG@10	NDCG@50	Recall@10	Recall@50
Complete P2D	**0.764**	**0.079**	**0.141**	**0.115**	**0.350**
without hyper-GNN	0.665	0.057	0.106	0.083	0.270
without $\mathcal{L}_{con}$	0.718	0.072	0.131	0.104	0.328
without $\mathcal{L}_{dis}$	0.696	0.070	0.122	0.101	0.299
without disentangled learning	0.651	0.062	0.114	0.094	0.284

TABLE V
TIME EFFICIENCY COMPARISON

Model	Overall Training Time	Inference Response Time
BiasMF	0.33 hours	0.6 ms / query
NCF	0.49 hours	0.8 ms / query
FM	0.41 hours	0.8 ms / query
DGCF	1.33 hours	13 ms / query
LightGCN	1.28 hours	11 ms / query
HAFR	1.31 hours	10 ms / query
SCHGN	1.71 hours	10 ms / query
Our P2D	5.53 hours	28 ms / query

resentation learning module. Within this module, we implement the MI minimization algorithm using the vCLUB-based MI upper bound estimator, necessitating additional training steps. However, our below ablation study underscores the significant performance benefits introduced by this module. Consequently, we contend that the incurred extra computation cost is justifiable in light of the substantial performance gains achieved.

D. Ablation Study (RQ2)

We next investigate the effectiveness of the key components, *i.e.,* hyper-GNN and disentangled representation learning with two regularization objectives, in P2D. To do this, we remove one of these components from the complete P2D and obtain four model variants. Similarly, we train these model variants and test the performance on the test set. The results are given in Table IV and we conclude the following.

Removing the hyper-GNN component would lead to severe performance degradation for all metrics. Specifically, the decrease in AUC, NDCG@10 and Recall@10 compared to the complete P2D is 13.0%, 27.8%, and 27.8%, respectively. This result strongly demonstrates the positive impact of our hyper-GNN component on recommendations. Observe that users usually have a limited number of recipe consumption records in the data, compared with the total number of recipes, *e.g.,* 10 on average vs. 45,630 on Allrecipes (Table II). The constructed hyper-graph between recipes allows our model to learn shared recipe preferences with interactions from all users. After that, we could estimate more reasonable recommendations for each user given only limited interactions.

We utilize two types of regularization objectives in the disentangled representation learning module. The results verify that both types of objectives are essential to decouple the blended user representations into disentangled forms. Specifically, explicitly disentangled factors via contrastive learning can capture more accurate preferential patterns to benefit recommendations. Moreover, the independence of factors ensured by the minimization of MI is necessary for disentangled representation learning. Combining the above, the entire disentangled representation learning module is another crucial component in our P2D approach.

E. Hyper-parameter Sensitivity (RQ3)

Our ablation study has proven the importance of the regularization objectives in disentangled learning. Recall that in Eq. (20) we use a hyper-parameter α to balance the recommendation and the regularization losses. Therefore, in this set of tests, we further investigate the sensitivity of P2D with regard to a. The result is reported in Fig. 3.

When varying α, we find that all performance metrics first increase and then decrease with the increment of α. When α is small, *e.g.,* $\alpha \leq 1$, our model does not perform well since the disentangled learning module is not sufficiently trained. On the other hand, with a large α, the performance would rapidly decrease as the model at this stage focuses too much on regularization while neglecting the main recommendation task. Overall, the best performance is achieved with $\alpha = 30$, and the unimodal trend assures that we could find a good α with reasonable efforts for parameter tuning.

F. Visualization Analysis of Disentangled Learning (RQ4)

Our disentangled learning module is inspired by psychological research. In the last set of tests, we qualitatively study the impacts and rationale of the module through visualization

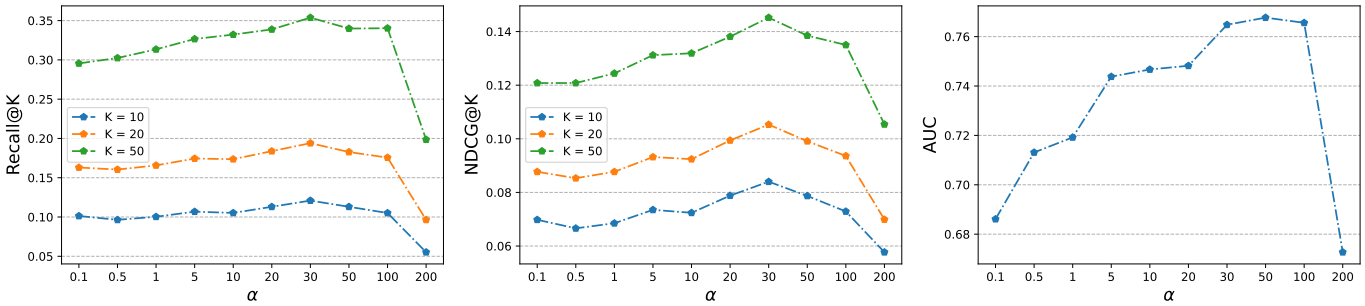
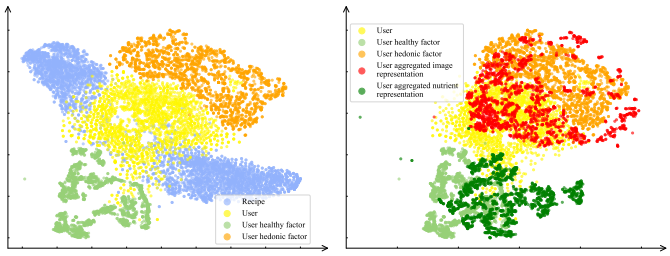


Fig. 3. Performance of P2D with different α in Eq. (20). We find that $\alpha = 30$ is a good choice for our model with all metrics.



(a) (b)

Fig. 4. Visualization of various representations learned by P2D.

analysis. Specifically, we plot in Fig. 4a the three sets of user representations and the recipe representations, *i.e.,* $\widetilde{e_u}$, e_u^h, e_u^p, and $\hat{e}_i$, learned by P2D with the t-SNE [41] method.

As can be seen from the Figure, the blended user representations and recipe representations are mixed together in the space, which is not conducive to personalized recommendation. In such a situation, the recommendations are likely to be made toward a small fraction of popular recipes. After exploiting disentangled representation learning, we obtain two sets of user representations that could be well separated in the space. The Figure 4b illustrates that throughout the decoupling process, user nodes (depicted in yellow) gradually shift towards preference anchor nodes (red and green). And the generated disentangled nodes, *i.e.*, user hedonic factor and healthy factor (orange and light green) maintain a distance from each other. This implies that the users' preferences for foods are possibly formed by multiple distinct factors. The decoupling process brings several advantages to recipe recommendations. First, the model could recommend more diverse recipes to users according to their different emphasis on the disentangled factors. Second, it is possible to also output the contributions of different factors along with the recommendations, adding some transparency to the complex underlying model.

VI. Conclusion

In this work, we tackled recipe recommendation from the perspective of (a) integrating a variety of crucial pairwise and high-order relationships and (b) explicitly distinguishing the distinct factors that influence recipe selections. Specifically, we proposed a P2D approach, which consists of a progressive message-passing module, a disentangled representation learning module, and a personalized recipe recommendation module. Extensive experiments on a large real-world dataset demonstrated the superiority of our P2D approach compared to the state-of-the-art baselines. Additional ablation and visualization analysis also illustrated the effectiveness of each of the key components in P2D. In the future, we plan to explore the following: (1) identifying more fine-grained user intent factors in disentangled representation learning for more accurate and interpretable recommendations and (2) incorporating self-supervised signals into GNNs to extract richer relational information and further enhance representation expressiveness.

References

[1] T. Theodoridis, V. Solachidis, K. Dimitropoulos, L. Gymnopoulos, and P. Daras, "A survey on ai nutrition recommender systems," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 540–546. [Online]. Available: https://doi.org/10.1145/3316782.3322760

[2] M. Chen, X. Jia, E. Gorbonos, C. T. Hong, X. Yu, and Y. Liu, "Eating healthier: Exploring nutrition information for healthier recipe recommendation," *Inf. Process. Manag.*, vol. 57, p. 102051, 2020.

[3] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[4] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2659–2671, 2020.

[5] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 426–434. [Online]. Available: https://doi.org/10.1145/1401890.1401944

[6] X. Gao, F. Feng, X. He, H. Huang, X. Guan, C. Feng, Z. Ming, and T.-S. Chua, "Hierarchical attention network for visually-aware food recommendation," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1647–1659, 2020.

[7] Y. Tian, C. Zhang, Z. Guo, C. Huang, R. Metoyer, and N. V. Chawla, "Reciperec: A heterogeneous graph learning model for recipe recommendation," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 7 2022, pp. 3466–3472, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/481

[8] Y. Song, X. Yang, and C. Xu, "Self-supervised calorie-aware heterogeneous graph networks for food recommendation," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 1s, feb 2023. [Online]. Available: https://doi.org/10.1145/3524618

[9] N. Sullivan, C. Hutcherson, A. Harris, and A. Rangel, "Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed," *Psychological Science*, vol. 26, no. 2, pp. 122–134, 2015, pMID: 25515527. [Online]. Available: https://doi.org/10.1177/0956797614559543

[10] D. Kahneman and S. Frederick, *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*. Cambridge University Press, 2002, pp. 49–81.

[11] A. W. Sorenson, B. W. Wyse, A. J. Wittwer, and R. G. Hansen, "An index of nutritional quality for a balanced diet," *Journal of the American Dietetic Association*, vol. 68, no. 3, pp. 236–242, 1976. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002822321063495

[12] C. Trattner and D. Elsweiler, "Food recommender systems: Important contributions, challenges and future research directions," *ArXiv*, vol. abs/1711.02760, 2017.

[13] M. Harvey, B. Ludwig, and D. Elsweiler, "You are what you eat: Learning user tastes for rating prediction," ser. SPIRE 2013. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 153–164. [Online]. Available: https://doi.org/10.1007/978-3-319-02432-5_19

[14] Y. Koren, R. M. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, 2009.

[15] M. Ge, M. Elahi, I. Fernaández-Tobías, F. Ricci, and D. Massimo, "Using tags and latent factors in a food recommender system," in *Proceedings of the 5th International Conference on Digital Health 2015*, ser. DH '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 105–112. [Online]. Available: https://doi.org/10.1145/2750511.2750528

[16] C. Trattner and D. Elsweiler, "Investigating the healthiness of internet-sourced recipes: Implications for meal planning and recommender systems," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, pp. 489–498. [Online]. Available: https://doi.org/10.1145/3038912.3052573

[17] J. Freyne and S. Berkovsky, "Intelligent food planning: Personalized recipe recommendation," in *Proceedings of the 15th International Conference on Intelligent User Interfaces*, ser. IUI '10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 321–324. [Online]. Available: https://doi.org/10.1145/1719970.1720021

[18] C. Teng, Y. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Web Science 2012, WebSci '12, Evanston, IL, USA - June 22 - 24, 2012*, N. S. Contractor, B. Uzzi, M. W. Macy, and W. Nejdl, Eds. ACM, 2012, pp. 298–307. [Online]. Available: https://doi.org/10.1145/2380718.2380757

[19] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=F72ximsx7C1

[20] Z. Li, Z. Cui, S. Wu, X. Zhang, and L. Wang, "Fi-gnn: Modeling feature interactions via graph neural networks for ctr prediction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 539–548. [Online]. Available: https://doi.org/10.1145/3357384.3357951

[21] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[22] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

[23] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, "Hierarchical fashion graph network for personalized outfit recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 159–168. [Online]. Available: https://doi.org/10.1145/3397271.3401080

[24] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3558–3565, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4235

[25] J. Yu, H. Yin, J. Li, Q. Wang, N. Q. V. Hung, and X. Zhang, "Self-supervised multi-channel hypergraph convolutional network for social recommendation," in *Proceedings of the Web Conference 2021*, ser. WWW '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 413–424. [Online]. Available: https://doi.org/10.1145/3442381.3449844

[26] S. Ji, Y. Feng, R. Ji, X. Zhao, W. Tang, and Y. Gao, "Dual channel hypergraph collaborative filtering," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 2020–2029. [Online]. Available: https://doi.org/10.1145/3394486.3403253

[27] J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu, "Learning disentangled representations for recommendation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 5712–5723. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html

[28] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua, "Disentangled graph collaborative filtering," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1001–1010. [Online]. Available: https://doi.org/10.1145/3397271.3401137

[29] M. Chokr and S. Elbassuoni, "Calories prediction from food images," in *AAAI Conference on Artificial Intelligence*, 2017.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[31] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017.

[32] J. Wang, K. Ding, L. Hong, H. Liu, and J. Caverlee, "Next-item recommendation with sequential hypergraphs," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1101–1110. [Online]. Available: https://doi.org/10.1145/3397271.3401133

[33] S. Bai, F. Zhang, and P. H. S. Torr, "Hypergraph convolution and hypergraph attention," *Pattern Recognit.*, vol. 110, p. 107637, 2019.

[34] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[35] P. Chonpracha, R. Ardoin, Y. Gao, P. Waimaleongora-ek, G. Tuuri, and W. Prinyawiwatkul, "Effects of intrinsic and extrinsic visual cues on consumer emotion and purchase intent: A case of ready-to-eat salad," *Foods*, vol. 9, no. 4, 2020. [Online]. Available: https://www.mdpi.com/2304-8158/9/4/396

[36] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1779–1788. [Online]. Available: https://proceedings.mlr.press/v119/cheng20b.html

[37] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, USA: AUAI Press, 2009, pp. 452–461.

[38] S. Rendle, "Factorization machines," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 995–1000.

[39] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[41] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

Chunlai Dong received his B.S. degree from Jilin University, in 2023. He is now studying for the Ph.D. degree in the College of Computer Science and Technology, Zhejiang University. His major research interest focuses on data mining.

Haochao Ying is currently an assistant professor in the School of Public Health, Zhejiang University. He received the Ph.D. degree in the College of Computer Science from Zhejiang University in 2019, and the B.S. degree in computer science and technology from Zhejiang University of Technology in 2014. His research interests include data mining for healthcare and personalized recommender system. He has authored some papers at prestigious international conferences and journals, such as TKDE, TMI, IJCAI, ICML, and CVPR.

Renjun Hu received the BE and Ph.D. degrees in computer science from Beihang University in 2014 and 2020, respectively. He has also spent time as a visiting student at Rutgers, the State University of New Jersey from 2017 to 2018, and as a research intern at Baidu Research's Business Intelligence Lab from 2018 to 2019. His research interests include graph and tabular data mining, as well as robust and explainable machine learning. He has contributed regularly to reputable conference proceedings and journals, including ICDE, AAAI, IJCAI, KDD, WSDM, IEEE TKDE, IEEE TMC, and ACM TKDD.

Yuyang Xu received his B.S. degree from Jilin University, in 2021. He is now studying for the Ph.D. degree in College of Computer Science and Technology, Zhejiang University. His major research interest is data mining in medical field.

Jintai Chen received the Ph.D. degree in the College of Computer Science and Technology, Zhejiang University, Hangzhou, China in 2023. His research interests include AI for healthcare, computer vision and deep tabular learning. He has published over 30 papers in top-tier journals and conference proceedings, including Nature Communications, ICML, ICLR, CVPR, AAAI, IJCAI, and MICCAI. He actively contributes to the academic community by serving as a program committee member and reviewer for prominent conferences such as CVPR, ICCV, NeurIPS, ICML, ICLR, AAAI, IJCAI, etc.

Fuzhen Zhuang received the Ph.D. degrees in the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He is currently a full Professor with the Institute of Artificial Intelligence, Beihang University. He has published more than 100 papers in some prestigious refereed journals and conference proceedings such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, ACM Transactions on Knowledge Discovery from Data, ACM Transactions on Intelligent Systems and Technology, Information Sciences, Neural Networks, SIGKDD, IJCAI, AAAI, TheWebConf, SIGIR, ICDE, CIKM, WSDM, SIAM, SDM, and ICDM. His research interests include transfer learning, machine learning, data mining, multitask learning, knowledge graph and recommendation systems. He is a Senior Member of CCF. He was a recipient of the Distinguished Dissertation Award of CAAI in 2013.

Jian Wu is a full Professor at Zhejiang University. He is currently the director of Real Doctor AI Research Centre of Zhejiang University. His research interests include Artificial Intelligence, Data Mining and their applications in healthcare and biomedicine. He received his Ph.D. degree in Computer Science and Technology from Zhejiang University. He has published more than 200 papers in some prestigious refereed journals and conference proceedings such as IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MEDICAL IMAGING, CVPR, IJCAI, AAAI, ICML, MICCAI. He is a Distinguished Member of CCF.