

Beyond Similarity: Heterogeneous Graph Learning for Multi-Objective Food Substitution in Charitable Food Agencies

Naimur Rahman Chowdhury

The work was done prior to joining Amazon.
Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC, USA
nchowdh2@ncsu.edu

Limon Bin Hossain

Department of Industrial and Production Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
mdlimonbinhossain@gmail.com

Abstract—Charitable food agencies play an important role in alleviating food insecurity by distributing donated food to people in need. However, they rely on ad hoc in-kind donations and often face shortages of specific foods, so they offer substitutes. A good food substitution requires matching household preferences, nutritional needs, and item similarity. Agencies have limited direct records of consumption behavior due to resource constraints, making it challenging to make an appropriate substitution decision that meets multiple criteria. In this study, we propose a heterogeneous graph neural network (HeteroGNN), a source-grounded recommendation framework for food substitution in charitable food agencies. We first build a unified relational graph from large-scale public data sources, combining household behavior on food consumption and food nutrient information in the United States (US) context. We treat the substitution recommendation as a multi-objective ranking problem with three targets, including behavior affinity, health suitability, and substitution similarity. We train and validate the proposed framework under standard graph relationship and adverse cold-start settings by removing relational edges from the graph. Our results show that the proposed framework leverages relational information beyond node features in predicting consumption behavior. Additionally, the proposed framework remains robust with sparsity when the model receives incomplete information about behavior and nutrient features. Finally, we show the weak correlation among different objectives, thereby justifying the multi-objective framing as a replacement for an aggregated decision. The proposed framework can help downstream charitable agency decision-makers make context-specific substitution recommendations with limited information available.

Index Terms—Food Substitution, Heterogeneous Graph Neural Networks, Multi-Objective Recommendation, Public Nutrition

I. INTRODUCTION

In the United States (US), one in six people is served by food banks and their associated partner agencies, such as food pantries and soup kitchens, which play an important role in reducing food insecurity [1]. However, these charitable food agencies are heavily dependent on external in-kind donations, which limit their ability to cater to food-insecure communities [2], [3]. Hence, these organizations often have to provide substitutes for requested or preferred items to food-insecure

households from their limited catalog [4], [5]. However, being a good food substitute requires several components, including familiarity within the household, nutritional suitability for household members, and similarity to the requested item [6], [7]. A substitute for a requested food item is less attractive if the household does not recognize it, has a dietary restriction that conflicts with it, or receives a nutritionally poor option. This substitution decision for a charitable agency volunteer is non-trivial, since households served are heterogeneous in demographic characteristics and often have little or no directly observed preference history to the agencies with limited data management capacity [8]. Public data on household food consumption provide dietary patterns and familiarity [9], [10], but no current public dataset provides structured records of household context, requested foods, and offered substitutes, so a recommendation system may rely on indirect behavioral signals.

The most relevant studies on charitable food agencies have focused on forecasting donation quantity and timing [3], [11], [12], but modeling the substitution decision itself still remains unexplored. In food substitution recommendations for requested items, to match preferences across multiple dimensions (affinity, nutritional values, and similarity), the system needs to rely on behavioral evidence, keeping familiarity, health, and similarity rather than a general recommender that is trained on direct user-item interaction [13], [14] to optimize a single score [15], [16]. This process generalizes to foods and household groups that it has not directly observed.

To this end, although the final substitution score can be analytically obtained when all relevant information is available, the challenge arises with sparse, highly interdependent inputs that are scattered across separate data sources. Moreover, an analytical form struggles to produce a recommendation score when the input is incomplete. This can be addressed by learning latent representations across the relational structure, inferring recommendations from weakly observed relationships, and then generalizing them to unseen foods or household groups.

In this study, we propose a heterogeneous graph neural network (HeteroGNN) to complement heuristic alternatives requiring enumeration in every combination of nutrition, consumption behavior, and food similarity that becomes harder to maintain as the number of foods, household groups, and relations grow. We leverage extensive public data sources [9], [10], [17] to extract household consumption behavior, nutritional information, and similarity. Finally, we propose a recommendation framework that produces a ranked list of substitutes with three objectives, i.e., behavior affinity, health suitability, and substitution similarity. Keeping these three scores distinct lets a downstream evaluator inspect the trade-offs behind a recommendation. The objectives of this study are as follows.

- To develop a heterogeneous graph framework that integrates a sparse public dataset into a unified relational structure for food substitute recommendation.
- To jointly learn behavior affinity, health suitability, and substitution similarity as complementary objectives, preserving downstream decision makers’ ability to evaluate trade-offs in recommendations.
- To evaluate the robustness of the learning framework in reliable recommendation performance under incomplete and sparse information, where traditional systems struggle.

II. METHODS

We formulate a decision problem where $\mathcal{U}$ denotes the set of household-group contexts, and $\mathcal{F}$ the set of foods in the agency catalog. Here, $\mathcal{R} \subseteq \mathcal{F}$ is the set of possible requested foods by a household. For a request from context $u \in \mathcal{U}$ for food $r \in \mathcal{R}$, the system receives a candidate set $\mathcal{C}(r) \subseteq \mathcal{F}$ of substitutes. We aim to rank the candidates $f \in \mathcal{C}(r)$ along three objectives shown in Eq. 1.

$$S(u, r, f) = (S_b(u, f), S_h(u, f), S_s(r, f)), \quad (1)$$

where S_b is behavior affinity, S_h is health suitability, and S_s is substitution similarity. Our learning model estimates $\hat{S}_b$, $\hat{S}_h$, and $\hat{S}_s$ separately.

A. Dataset and Graph Construction

We construct the relational structure from four public data sources. Each source contributes a distinct type of information to the pipeline. First, we use the USDA FoodData Central Foundation Foods [17] that provides nutrient profiles for basic food items. Second, we use the Food and Nutrient Database for Dietary Studies (FNDDS) [18], which provides standardized food codes. Together, these two datasets form the basic food catalog.

The National Health and Nutrition Examination Survey (NHANES) [9] data provide demographic information, dietary records, and health indicators. Additionally, the FoodAPS [10] dataset provides household and individual food acquisition records. These two sources are combined to obtain information on population and behavior. Moreover, we incorporate nutrition guidance from Healthy Eating Research [19] to rank foods

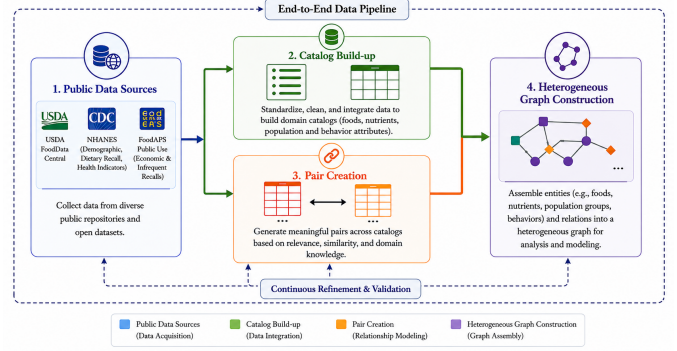


Fig. 1: Complete data construction with food catalog, population and behavior relations for heterogeneous graph.

TABLE I: Scale of the data in the graph construction.

Quantity	Count
People	26,250
Household/respondent contexts	16,759
Demographic household groups	42
Foods with nutrient evidence	9,903
Food–nutrient records	102,441
Consumption lines	188,148
Acquisition items	254,587
Group–food behavior edges	59,552

specifically for the charitable food supplies. The construction process follows two streams as shown in Figure 1. One stream builds the food catalog and substitution pairs. The other builds a relational database of persons, households, and observed food behavior. We merge them finally into a single heterogeneous graph.

For the food catalog, we use a list of generic items relevant to charitable food agencies [20] and link each item to the USDA FoodData Central Foundation Foods record. Any missing target nutrients are filled from mapped FNDDS records. We then construct directional substitution pairs using culinary-use overlap, food category, form, and nutrient closeness.

The largest construction involves the relational database linking person and household records from NHANES and FoodAPS. We combine the records to build household-context groups using income band, household size, and SNAP participation. We aggregate reported NHANES consumption and FoodAPS acquisition records at the household-group level for behavioral evidence.

This produces a heterogeneous graph with four node types: household groups, foods, nutrients, and categories. The main relations families are household-group–food behavior, food–nutrient content, food–category membership, and food–food substitution similarity. Table I summarizes the scale of the constructed relational structure.

B. Objective Targets

In this section, we define the three targets in the objective in substitute ranking described in Eq. 1.

1) *Behavior Affinity Target*: For each observed household-group-food pair (u, f) , the database records the number of occasions and the total grams acquired. The behavior target is constructed from observed occasions defined in Eq. 2.

$$y_b(u, f) = \frac{\log(1 + \text{occasions}_{u,f})}{\max_{f'} \log(1 + \text{occasions}_{u,f'})}. \quad (2)$$

Eq. 2 follows implicit feedback [21], in which the strength of an observed interaction is modeled as a monotonic function of its frequency rather than assumed to scale linearly with raw counts. For the binary behavior head, observed group-food pairs are treated as positive behavior evidence, and unobserved pairs are sampled as negative examples, as defined in Bayesian Personalized Ranking (BPR) [22]. A similar approach is used in graph-based recommenders, including Neural Graph Collaborative Filtering [23] and LightGCN [24].

2) *Health Suitability Target*: The relational database constructed from NHANES and FoodAPS records contains six profile-specific health scores, i.e., general, diabetes, hypertension, obesity, child nutrition, and senior nutrition, for members in a household group. Each food-profile score is a nutrient-density score based on beneficial and limiting nutrients, following practical guidance for charitable food environments [19]. Beneficial nutrients are mapped as in Eq. 3, whereas Eq. 4 maps limiting nutrients.

$$c^+(f, n) = \min\left(\frac{\max(a_{f,n}, 0)}{\rho_n}, 1\right), \quad (3)$$

$$c^-(f, n) = 1 - \min\left(\frac{\max(a_{f,n}, 0)}{\rho_n}, 1\right), \quad (4)$$

where $a_{f,n}$ is the per-100 g nutrient amount and ρ_n is the profile reference amount.

The profile-specific score is defined in Eq. 5 that follows the Nutrient-Rich Foods (NRF) index, which scores foods on a per-reference-amount basis [25]–[27].

$$H(f, p) = \frac{\sum_{n \in \mathcal{N}_f} w_{p,n} c(f, n)}{\sum_{n \in \mathcal{N}_f} w_{p,n}}. \quad (5)$$

For household group u , the profile mixture $\pi_{u,p}$ is estimated from observed health indicators in the group. The final group-food health target is defined in Eq. 6.

$$y_h(u, f) = \sum_{p \in \mathcal{P}} \pi_{u,p} H(f, p). \quad (6)$$

3) *Substitution Similarity Target*: We build directional candidate pairs from the food catalog for substitution. Each pair combines culinary-use, category, form, and nutrient-profile components. Only pairs whose request and substitute foods map to nodes in the FNDDS graph. The similarity target is denoted by $y_s(r, f) \in [0, 1]$.

C. HeteroGNN

In this section, we describe the main components of the HeteroGNN architecture.

1) *Relational Message Passing*: Let x_v^τ denote the input feature vector for node v of type τ . A type-specific encoder maps each input into a hidden representation,

$$h_v^{(0)} = \sigma(W_\tau x_v^\tau + b_\tau). \quad (7)$$

For each message-passing layer ℓ , relation-specific transformations aggregate information from neighboring nodes,

$$\tilde{h}_v^{(\ell+1)} = W_{\tau(v)}^{\text{self}} h_v^{(\ell)} + \sum_{\rho \in \mathcal{R}_G} \sum_{u \in \mathcal{N}_\rho(v)} \omega_{u,v}^\rho W_\rho h_u^{(\ell)}, \quad (8)$$

where ρ indexes graph relations and $\omega_{u,v}^\rho$ is the edge weight. The output is normalized and passed through a nonlinearity,

$$h_v^{(\ell+1)} = \text{ReLU}\left(\text{LayerNorm}(\tilde{h}_v^{(\ell+1)})\right). \quad (9)$$

Multiple message-passing layers are used in an inductive graph representation [13], [14].

2) *Heads and Training Objective*: After HeteroGNN encoding, three prediction heads operate on node-pair embeddings,

$$\hat{S}_b(u, f) = g_b([h_u, h_f]), \quad (10)$$

$$\hat{S}_h(u, f) = g_h([h_u, h_f]), \quad (11)$$

$$\hat{S}_s(r, f) = g_s([h_r, h_f]). \quad (12)$$

Each head is a small feed-forward network with a sigmoid output. The model is trained with a joint loss,

$$\mathcal{L} = \lambda_b \mathcal{L}_b + \lambda_h \mathcal{L}_h + \lambda_s \mathcal{L}_s, \quad (13)$$

where the current implementation sets all three weights to one. Behavior uses binary cross entropy,

$$\mathcal{L}_b = -\frac{1}{|\mathcal{D}_b|} \sum_{(u,f,y) \in \mathcal{D}_b} \left[y \log \hat{S}_b(u, f) + (1 - y) \log (1 - \hat{S}_b(u, f)) \right], \quad (14)$$

and health and similarity use mean-squared error,

$$\mathcal{L}_h = \frac{1}{|\mathcal{D}_h|} \sum_{(u,f,y) \in \mathcal{D}_h} (\hat{S}_h(u, f) - y)^2, \quad (15)$$

$$\mathcal{L}_s = \frac{1}{|\mathcal{D}_s|} \sum_{(r,f,y) \in \mathcal{D}_s} (\hat{S}_s(r, f) - y)^2. \quad (16)$$

D. Weighted Multi-Objective Score

The model produces three predicted scores per candidate. A weighted score is used for the composite scoring as in Eq. 17.

$$R(u, r, f) = \alpha_b \hat{S}_b(u, f) + \alpha_h \hat{S}_h(u, f) + \alpha_s \hat{S}_s(r, f), \quad (17)$$

α_b , α_h , and α_s are the weights for each objective, providing downstream decision-makers with the flexibility to evaluate trade-offs among recommendations. For this study, we use a uniform weight across the three targets. Table II demonstrates the three objectives and their interpretation.

TABLE II: Interpretation of the multi-objective recommendation.

Objective	Supervision	Interpretation
Behavior affinity	Observed group–food relation vs. sampled unobserved relation	The probability that the food is familiar or behaviorally plausible for the household group.
Health suitability	Profile-weighted food-health score	Nutrient-density suitability for the health-profile mix associated with the group.
Substitution similarity	Directional mapped catalog-pair score	The degree to which a candidate can play the same role as the requested item.
Composite ranking	Weighted combination over the three predictions	Inspectable decision support rather than a single uninterpretable score.

III. EXPERIMENTAL SETUP

The objective of the experiments is to evaluate the performance of HeteroGNN on the three substitution targets, thus the final composite score, and how that performance degrades when foods or population groups are unseen during behavior supervision. Table III shows all four node-level features used for training. All continuous features are standardized before training.

We use three validation settings. The standard setting uses a random 70/10/20 partition of observed behavior edges, with the final 20 percent held out for evaluation. To evaluate HeteroGNN’s performance when relational supervision is unavailable, we perform two cold-start settings. In the cold-start food setting, 1,774 food nodes are withheld from the behavior relation, and all food similarity edges connected to these held-out foods are removed from the training graph. Similarly, in the cold-start household-group setting, 8 of the 42 household groups are withheld from the behavior supervision. In both cold-start settings, node attributes remain available to see whether the model generalizes from feature information despite the absence of a behavioral relationship.

We construct the training pipeline to reflect the three training settings. In the standard split, we randomly partition the observed behavior edges, whereas in the cold-start-food split, label edges that connect held-out foods are removed from the training graph. In the cold-start-household-group split, behavior edges touching held-out groups are removed. At inference time, the trained model receives a household group, a requested food, and a set of feasible candidates, and predicts behavior, health, and similarity scores for each candidate.

Finally, to compare the model’s performance, we select baselines ranging from simple heuristics to advanced graph models to establish performance bounds and isolate our contributions. We begin with the simplest approach, i.e., Bayesian Personalized Ranking (BPR), that optimizes for observed items to rank above unobserved ones. However, it struggles in understanding auxiliary relations (e.g., recipes, store locations) [28] and is highly sensitive to data sparsity. To capture more complex patterns, we use a Multi-Layer Perceptron (MLP) that learns a flexible, nonlinear interaction between user and item feature vectors. It captures more complex patterns, and performance generally improves with model depth. However, it does not inherently capture relational structure. Recognizing the importance of relational structure, LightGCN [24] is a simplified Graph Neural Network (GNN) that removes

nonlinear activations and feature transformations, focusing on neighborhood aggregation and layer combination.

For our final training architecture, we leverage a tree-structured Parzen Estimator sampler and an early-pruner to find the best hyperparameter [29]. The best trial selects a hidden dimension of 64, 2 message-passing layers, ReLU activation, sum aggregation, and a dropout of 0.218. Table IV summarizes the main training settings.

IV. RESULTS AND DISCUSSION

We organize the results around the questions the experimental design was built to answer. We first report the evaluation accuracy on the standard setting and then discuss cold-start generalization. We highlight, in particular, the demographic segments that drive the cold-start household-group gap. Finally, we discuss the multi-objective trade-offs produced by the three heads.

A. Evaluation and Comparisons

The performance of the proposed HeteroGNN model was evaluated against baseline models MLP, BPR, and LightGCN across three evaluation settings (shown in Table ??). All evaluations across HeteroGNN and baseline models are performed under the identical split and supervision. Additionally, baseline models are trained on the same information and features, with the same supervision.

In the **standard** setting, HeteroGNN achieves the highest difference in performance on behavior affinity, with an ROC-AUC of 0.9480 and a PR-AUC of 0.9027. These results marginally surpass LightGCN (0.937, 0.884) and outperform MLP (0.901, 0.781). This demonstrates that the good relational structure is the primary mechanism driving behavior affinity prediction. While MLP leverages household and food attributes and captures nonlinear interactions, it lacks the capacity to propagate information through food–nutrient, food–category, and food–food relations. BPR benefits from optimizing the implicit ordering of behavior but lacks auxiliary relational evidence. LightGCN performs more closely to HeteroGNN due to its graph information propagation, but its homogeneous filtering architecture does not explicitly differentiate among semantic relation types. HeteroGNN surpasses these baselines by maintaining relation-specific message passing and heterogeneous representation learning. Furthermore, HeteroGNN achieves a Precision@5 of 0.995 and a Recall@5 of 0.026, reflecting that a household group’s positive set may

TABLE III: Node feature construction for the heterogeneous graph.

Node type	Feature name	Description	Encoding type
Household group	Member count	Number of respondent contexts in the group	Discrete
Household group	Mean household size	Average household size within the group	Discrete
Household group	Mean income-to-poverty ratio	Average income-to-poverty ratio within the group	Continuous
Household group	SNAP share	Share of group members participating in SNAP	Continuous proportion
Household group	Poverty band	Discretized income-to-poverty category	One-hot binary
Household group	Household-size band	Discretized household-size category	One-hot binary
Household group	SNAP band	SNAP or non-SNAP group label	One-hot binary
Household group	Health-profile mixture	Estimated mixture over nutrition profiles described in Section II-B2	Continuous proportions
Food	Log nutrient values	Log-transformed nutrient density values for the nutrients	Continuous
Food	Nutrient indicators	Indicator for whether each nutrient value is missing	Binary
Food	Source release	Food source release (e.g., FND DS cycle)	One-hot binary
Food	Category code	FND DS food category code	Continuous/numeric
Nutrient	Nutrient direction	Whether higher values are beneficial or limiting	Signed numeric
Nutrient	Reference amount	Nutrient-specific reference value used for score normalization	Continuous, log-transformed
Nutrient	Average profile weight	Mean importance of the nutrient across the six health profiles	Continuous
Category	Category code	Numeric food-category identifier	numeric
Category	Food count	Number of foods observed in the category	Continuous

TABLE IV: Model hyperparameters and training Settings.

Hyperparameter	Value
Node embedding dimension	64
Message-passing layers	2
Activation function	ReLU
Aggregation function	Sum
Dropout	0.218
Optimizer	AdamW
Learning rate	0.01
Weight decay	0.0001
Negative samples per positive behavior edge	2
Maximum health examples per training split	60,000

contain significantly more than five behaviorally plausible foods. Moreover, a top-5 recommendation list can capture only a small portion of the positives in a large set, even when all recommended items are correct. These results demonstrate that HeteroGNN is highly effective at ensuring precision at the top of the recommendation list.

In contrast with the behavior affinity, the health suitability prediction results show a different pattern. To evaluate performance, we report MAE to measure the accuracy of predicted health scores and the Spearman correlation to assess how well the predicted ranking of substitutions aligns with the ground-truth. Here, MLP achieves better performance with an MAE of 0.0128 and a Spearman correlation of 0.982, outperforming HeteroGNN (0.0182, 0.962), while BPR and LightGCN perform worse than HeteroGNN. These results indicate that when explicit nutrient attributes are present, a feature-based MLP can predict health suitability with high accuracy and outperform strong learning mechanisms that capture relational structure. From this result, we infer that health suitability is primarily a feature-based prediction task, as nutrient composition and demographic health profiles are

sufficiently represented in node features and provide little benefit to relational propagation, instead introducing noise.

For the substitution similarity objective, HeteroGNN and MLP exhibit comparable performance, with MAE values of 0.0222 and 0.0344, respectively, whereas BPR and LightGCN perform worse. The limited supervision set for similarity likely accounts for the lack of advantage observed with graph-based methods, as there is insufficient relational signal to propagate, allowing the simpler MLP to remain competitive. BPR’s near-random health predictions, indicated by a Spearman ρ of 0.101, highlight that optimizing for pairwise preference ranking does not provide a useful inductive bias for nutrient-density regression.

Finally, the weighted multi-objective score for HeteroGNN is 0.9628, compared to 0.9164 for MLP, 0.9211 for BPR, and 0.9267 for LightGCN. MLP’s weaker performance on behavior tasks, despite using the same node features, confirms that the graph structure encodes relational information not captured by node features alone. Finally, this finding demonstrates that a single ranking mechanism cannot effectively address all three objectives. LightGCN’s strong performance in behavior prediction but poor results in health suitability prediction also confirm that health suitability prediction is not a collaborative filtering problem, but it requires nutrient composition and demographic features rather than relational propagation through behavior edges.

B. Cold-Start Generalization

The cold-start evaluation settings are intended to reveal how graph networks generalize to unseen entities when the entire node loses behavioral supervision. Notably, this differs from performance under incomplete data, which demonstrates robustness (we discuss this in Section IV-C).

TABLE V: Performance comparison under the **standard** evaluation setting. Best results in **bold**.

Metric		HeteroGNN	MLP	BPR	LightGCN
<i>Behavior affinity</i>	ROC-AUC	0.9480	0.9009	0.9144	0.9369
	PR-AUC	0.9027	0.7809	0.8467	0.8838
	Precision@5	0.9952	0.8000	0.9667	0.9667
	Recall@5	0.0262	0.0186	0.0244	0.0240
	NDCG@5	0.9965	0.8039	0.9724	0.9699
<i>Health suitability</i>	MAE	0.0182	0.0128	0.0756	0.0765
	RMSE	0.0266	0.0177	0.1052	0.1053
	Spearman ρ	0.9617	0.9823	0.1013	0.1182
<i>Substitution similarity</i>	MAE	0.0222	0.0344	0.0820	0.0689
	RMSE	0.0561	0.0844	0.1720	0.1533
	Spearman ρ	0.6973	0.6900	0.6515	0.6699
<i>Weighted Multi-Objective Score</i>		0.9628	0.9164	0.9211	0.9267

Under **cold-start** setting, HeteroGNN’s behavior affinity ROC-AUC decreases from 0.9480 to 0.8381 for unseen foods and further to 0.6070 for unseen household groups (shown in Table VI). This indicates that HeteroGNN’s primary source of predictive power remains the observed group–food behavioral structure, suggesting that the model struggles when behavioral edges are absent. Under the cold-start group setting, the behavior ROC-AUC falls to 0.6070, and the PR-AUC to 0.4367. However, in terms of health suitability, the MAE remains 0.0245, with a Spearman correlation of 0.9641; for substitution similarity, the MAE is 0.0333, with a Spearman correlation of 0.6932. This suggests that the model does not fail under group cold start, but struggles to recover unseen household behavioral preferences.

To elaborate on the performance trend, the divergence is architecturally expected in HeteroGNN, since health suitability is primarily a function of food nutrient composition and the household’s health profile, whereas substitution similarity is primarily a function of the relationship between the requested and candidate foods. Both can therefore be inferred from relatively stable feature information. Behavior affinity, however, depends strong relational structure. When an entire group is unseen during behavior supervision, there is no direct relational evidence from which to infer its latent food preferences.

On the other hand, the MLP’s behavior affinity under cold start is stable and outperforms HeteroGNN in cold-start settings. This suggests that MLP is relying heavily on static node attributes, which remain available in both cold-start experiments, whereas HeteroGNN exploits relational behavior information that disappears when the relevant nodes are held out. Consequently, the MLP becomes unusually competitive when relational supervision is deliberately removed.

We observe a similar trend in health suitability, with MLP outperforming HeteroGNN. This indicates that health suitability can be predicted from node features and structural similarity to known foods, without requiring direct behavioral edges. Similarly, similarity MAE rises substantially from 0.0222 to 0.1260 in cold-start food for HeteroGNN, a wider degradation than any other objective, because similarity edges connected to held-out foods are removed, eliminating the primary relational

signal for this objective. In the cold-start household group, however, similarity MAE rises to only 0.0333, indicating that similarity prediction depends on food-food relations rather than household-specific information.

1) *Fairness Across Demographic Segments:* The aggregate cold-start household group ROC-AUC of 0.607 hides considerable variation across household segments. Table VII breaks the same held-out predictions down by poverty band, SNAP participation, and household size. We observe that the generalization is best for households below the poverty line, ROC-AUC 0.781, and worst for small households, ROC-AUC 0.324, which is worse than a random ranker. Large households generalize moderately well (0.640), and single-person households sit in between (0.659). On the contrary, the SNAP status produces a smaller spread than the poverty band or household size. The results particularly reveal that the model generalizes poorly to small households. For implication, practitioners should treat cold-start behavior predictions for small household groups with particular caution.

C. Sparsity Analysis

We use sparsity analysis to address data shortages at the facility level. Charitable food banks often serve groups for whom only partial information is available. In humanitarian logistics, this ensures equity by distributing aid in proportion to actual needs, even when those needs are not well-documented in existing datasets [30].

We perform different sparsity analyses. First, we randomly remove interaction edges (as opposed to completely removing all edges in a cold-start setting), which is defined as behavior sparsity. Second, we remove nutritional feature columns, a phenomenon known as nutrient sparsity. This tests the loss of relational and attribute information separately. Removing behavior relations affects behavior affinity prediction and impacts health scoring, whereas removing nutrient features impacts health scoring but can also affect behavioral affinity. Therefore, we report the ROC AUC for behavioral affinity and the MAE for health suitability in both analyses.

The sparsity experiments (Table VIII) provide an important explanation for why HeteroGNN remains useful despite this cold-start limitation. The HeteroGNN outperforms MLP at all

TABLE VI: Performance comparison under the cold-start evaluation settings. CS-F = cold-start food; CS-G = cold-start household group.

Metric		HeteroGNN		MLP		BPR		LightGCN	
		CS-F	CS-G	CS-F	CS-G	CS-F	CS-G	CS-F	CS-G
<i>Behavior affinity</i>	ROC-AUC	0.8381	0.6070	0.8973	0.8971	0.4959	0.3883	0.4887	0.5323
	PR-AUC	0.6932	0.4367	0.7739	0.7618	0.3347	0.2749	0.3511	0.3910
	Precision@5	0.5143	0.2500	0.7857	0.8750	0.3333	0.1500	0.2238	0.3750
	Recall@5	0.0108	0.0008	0.0179	0.0024	0.0057	0.0006	0.0039	0.0009
	NDCG@5	0.4656	0.2500	0.8012	0.8453	0.3390	0.1517	0.2405	0.3750
<i>Health suitability</i>	MAE	0.0492	0.0245	0.0140	0.0143	0.0772	0.0762	0.0771	0.0825
	RMSE	0.0616	0.0329	0.0200	0.0196	0.1071	0.1076	0.1071	0.1100
	Spearman ρ	0.8609	0.9641	0.9807	0.9795	0.0140	0.0941	0.0151	0.1015
<i>Substitution similarity</i>	MAE	0.1260	0.0333	0.0672	0.0491	0.1234	0.0818	0.1416	0.0751
	RMSE	0.2977	0.1130	0.1883	0.1521	0.2732	0.1971	0.3217	0.1511
	Spearman ρ	0.6848	0.6932	0.6786	0.6493	0.6401	0.6216	0.6384	0.6691
<i>Weighted Multi-Objective Score</i>		0.7894	0.7666	0.9146	0.9082	0.7054	0.6592	0.6907	0.7351

TABLE VII: Behavior Affinity Performance by Household Demographic Segment (Cold-Start Group Split)

Attribute	Level	ROC-AUC
Poverty band	Above $2\times$ poverty	0.470
	Below poverty	0.781
	$1-2\times$ poverty	0.667
SNAP band	Non-SNAP	0.590
	SNAP	0.619
Household size	Large	0.640
	Single	0.659
	Small	0.324

removal levels, with ROC AUC dropping only 0.7% (0.9444 to 0.9378) from 0% to 50% removal, while MLP remains steady at 0.889. The slight increase at 10% removal shows GNN’s robustness to moderate data loss, indicating removed edges were not critical. MLP’s flat AUC demonstrates insensitivity to behavioral sparsity, since it does not benefit from relational input. On the other hand, MLP has lower health MAE than GNN at all behavior removal levels (0%: 0.0285 vs. 0.0262; 50%: 0.0292 vs. 0.0265).

Under nutrient sparsity, GNN is more robust, leading to a higher ROC AUC, while MLP remains stable but at a lower level. For health suitability MAE, MLP starts better but drops sharply at 50% removal (GNN: 0.0379, MLP: 0.0481). GNN’s robustness shows it leverages relational structure when features are lost. Across eight perturbations on sparsity, HeteroGNN’s AUC drops marginally and remains robust under severe sparsity. MLP matches or surpass GNN performance on for health suitability when nutrient features are present, but under information loss, GNN’s relational modeling performs better in MAE.

D. Multi-Objective Trade-offs

Table IX reports the Pearson and Spearman correlation among the three predicted objectives. Behavior affinity and health suitability are weakly negatively correlated, -0.067 .

This is consistent with the common practice that familiar, frequently consumed foods are not always the most nutrient-rich. Behavior affinity and similarity are weakly positively correlated (0.090), whereas health suitability and substitution similarity are weakly positively correlated (0.070). None of these correlations is strong enough to claim that one objective serves as a reliable proxy for another, which provides quantitative justification for keeping the three separate targets as proposed in our framework. However, a practical consideration is that the weighted score in Eq. (17) is sensitive to how the three weights α_b , α_h , and α_s are set. Since the objectives are only weakly related, shifting weights can change which candidate ranks first. Hence, a downstream decision maker reviewing a candidate should set condition-driven weights for ranking. For example, for a person with health conditions, the weight on α_h should be higher than the other two factors to dominate the health suitability component of the composite score. The Spearman correlations also reflect that the objectives remain sufficiently independent to justify multi-objective optimization rather than a single target.

V. RELATED WORK

A. Food Assistance Operations and the Substitution Decision Gap

Food assistance research focuses on minimizing the gap between what a food bank has on hand and what a household needs [5], which includes allocating scarce donated inventory across facilities [31], forecasting donation volume and timing [11]. These studies focus on distributing donated food upon its arrival. However, once a household requests an item that is unavailable at distribution time, choosing a substitute is a ranking problem over a candidate set conditioned on that household. To our best knowledge, none of the existing studies focus on this aspect, leaving a gap between planning how much food arrives and deciding what to hand over when a specific request cannot be met.

TABLE VIII: Robustness of HeteroGNN and MLP in incomplete information.

Removed	Behavior Sparsity				Nutrient Sparsity			
	Behavior ROC AUC $\uparrow$		Health MAE $\downarrow$		Behavior ROC AUC $\uparrow$		Health MAE $\downarrow$	
	GNN	MLP	GNN	MLP	GNN	MLP	GNN	MLP
0%	0.9444	0.8890	0.0285	0.0262	0.9444	0.8890	0.0285	0.0262
10%	0.9474	0.8889	0.0283	0.0263	0.9433	0.8895	0.0329	0.0254
30%	0.9442	0.8889	0.0290	0.0265	0.9388	0.8890	0.0327	0.0285
50%	0.9378	0.8889	0.0292	0.0265	0.9469	0.8876	0.0379	0.0481

TABLE IX: Correlations among the three objectives.

	Pearson			Spearman		
	Behavior	Health	Similarity	Behavior	Health	Similarity
Behavior affinity	1.000	−0.067	0.090	1.000	0.521	0.311
Health suitability	−0.067	1.000	0.070	0.521	1.000	0.421
Substitution similarity	0.090	0.070	1.000	0.311	0.421	1.000

B. Public Nutrition and Acquisition

A set of public data collection describes what people eat and what households acquire, without addressing substitution preferences directly. NHANES records dietary recall and health indicators at The person level [9], [32], whereas FoodAPS records household-level acquisition [10], and FNDDS links the food codes in both surveys to nutrient composition [17], [18]. Existing work treats these sources as population-level descriptive statistics rather than as relational evidence linking people, households, foods, and nutrients. This study uses the same public instruments but keeps observed behavior and accepted substitution explicit throughout, and links persons, household groups, foods, nutrients, and categories into a single evidence graph rather than a set of separate survey tables.

C. Multi-Objective Recommendation

Classic recommender systems reduce a recommendation to one relevance score, typically learned from collaborative filtering over dense user-item interaction histories [15]. Hybrid and multi-objective approaches show that collapsing distinct evidence into one number hides tradeoffs a reviewer prefers to see [16]. For food bank substitution, ranking by familiarity alone reinforces existing nutritional gaps in frequently acquired staples, whereas ranking by health score alone can suggest foods a household cannot prepare or recognize, and ranking by similarity alone can ignore a real health constraint. As the correlation analysis in Table IX shows, these three signals are only weakly associated with one another across the collected evidence, which provides an empirical basis for treating them as separate targets rather than one blended score.

D. Graph Neural Networks for Heterogeneous Recommendation

GNNs offer a natural representation for recommendation problems involving users, items, and their relations. Graph convolutional networks establish neighborhood aggregation for propagating information across a graph [13], and GraphSAGE

made this aggregation inductive, letting a trained model embed nodes unseen during training as long as they carry features [14]. Web-scale systems built on graph convolution confirm that this representation improves recommendation quality whenever relation structure carries real signal [23], [33]. This body of work, however, is built around a single node type and single relation type: users, items, and their interactions.

In sum, food-bank operations research plans supply, but require modeling the substitution decision. Additionally, public nutrition and acquisition data describe behavior, but a strong relational structure across them is absent. Finally, graph neural models support heterogeneous, inductive representation learning in principle, which is leveraged in this study to close the previously mentioned gaps.

VI. CONCLUSION AND FUTURE WORK

This study formalizes food substitution as a multi-objective ranking problem with three separate objectives: behavior affinity, health suitability, and substitution similarity. These dimensions cannot be collapsed into a single score without losing decision-relevant information. To address this, we develop heteroGNN that integrates four disparate public data sources into a unified relational structure. Our results show that relational information adds predictive value beyond node features alone, delivering a practical, source-grounded decision-support framework for food bank volunteers who currently make substitution decisions without systematic guidance. Our results show that health and similarity predictions remain robust under cold-start conditions, whereas behavior prediction collapses, indicating that a single architecture cannot serve all objectives well. However, the sparsity analyses show that graph relations can compensate for incomplete behavioral or nutritional information, supporting the motivation for heterogeneous graph learning.

Several limitations follow directly from the findings of this study. First, the substitution similarity head is trained on only 56 mapped directional labels after the 124 catalog-level pairs

are mapped to the graph's food population. This small label count is the main reason the substitution similarity MAE degrades the most under cold-start-food. Second, the cold-start-household group result is estimated from only 8 held-out groups out of 42, so the aggregate ROC-AUC of 0.607 is itself a noisy estimate, and the segment breakdown in Section IV-B1 shows that this aggregate conceals a range from 0.324 to 0.781. Third, we use the single architecture under standard settings with optimal hyperparameters and then reuse across all validation splits, which require a setting-specific tuning.

Future work should account for feasible constraints, such as food preparation requirements and cultural acceptability beyond what consumption data capture. Additionally, incorporation of richer household context beyond the current demographic grouping, including cultural food preferences, preparation constraints, and household member age distributions, can make the graph more informed. Further studies consider adaptive objective weighting, allowing agencies to focus on nutritional suitability, familiarity, or functional similarity according to the specific household and operational context.

REFERENCES

- [1] A. Park, J. Chen, and A. Bustamante, "Adverse consequences of food insecurity among us adults beyond health outcomes," *American journal of preventive medicine*, vol. 66, no. 1, pp. 146–153, 2024.
- [2] N. R. Chowdhury, R. I. Siddiquee, and J. S. Ivy, "Influence of norms in alliance characteristics of humanitarian food agencies: Capability, compatibility and satisfaction," in *2025 Winter Simulation Conference (WSC)*. IEEE, 2025, pp. 652–663.
- [3] E. Sharma, L. Davis, J. Ivy, and M. Chi, "Data to donations: Towards in-kind food donation prediction across two coasts," in *2021 IEEE Global Humanitarian Technology Conference (GHTC)*. IEEE, 2021, pp. 281–288.
- [4] I. Sengul Orgut and E. J. Lodree, "Equitable distribution of perishable items in a food bank supply chain," *Production and Operations Management*, vol. 32, no. 10, pp. 3002–3021, 2023.
- [5] C. Bazerghi, F. H. McKay, and M. Dunn, "The role of food banks in addressing food insecurity: A systematic review," *Journal of Community Health*, vol. 41, no. 4, pp. 732–740, 2016.
- [6] E. Short, J. Sharma, D. I. Thompson, D. Taren, R. Gonzalez, and M. Hingle, "Food assistance use among food bank clients affected by type 2 diabetes," *Journal of nutrition education and behavior*, vol. 54, no. 4, pp. 288–298, 2022.
- [7] A. Simmet, J. Depa, P. Tinnemann, and N. Stroebele-Benschop, "The nutritional quality of food provided from food pantries: a systematic review of existing literature," *Journal of the Academy of Nutrition and Dietetics*, vol. 117, no. 4, pp. 577–588, 2017.
- [8] T. Hasnain, I. Sengul Orgut, and J. S. Ivy, "Elicitation of preference among multiple criteria in food distribution by food banks," *Production and Operations Management*, vol. 30, no. 12, pp. 4475–4500, 2021.
- [9] Centers for Disease Control and Prevention, "NHANES august 2021–august 2023: Questionnaires, datasets, and related documentation," National Center for Health Statistics, 2024, accessed August 18, 2026. [Online]. Available: <https://www.cdc.gov/nchs/nhanes/ContinuousNhanes/default.aspx?Cycle=2021-2023>
- [10] U.S. Department of Agriculture, Economic Research Service, "FoodAPS: National household food acquisition and purchase survey," 2026, data product page updated January 12, 2026; accessed August 18, 2026. [Online]. Available: <https://www.ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey>
- [11] L. B. Davis, S. X. Jiang, S. D. Morgan, I. A. Nuamah, and J. R. Terry, "Analysis and prediction of food donation behavior for a domestic hunger relief organization," *International Journal of Production Economics*, vol. 182, pp. 26–37, 2016.
- [12] E. Sharma, L. Davis, J. Ivy, and M. Chi, "FoodRL: A reinforcement learning ensembling framework for in-kind food donation forecasting," *arXiv preprint arXiv:2511.04865*, 2025.
- [13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.
- [14] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [15] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Introduction and challenges," in *Recommender Systems Handbook*, 2nd ed. Springer, 2015.
- [16] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, pp. 331–370, 2002.
- [17] U.S. Department of Agriculture, Agricultural Research Service, "FoodData Central: Downloadable data," 2026, accessed August 18, 2026. [Online]. Available: <https://fdc.nal.usda.gov/download-datasets/>
- [18] —, "Usda food and nutrient database for dietary studies (FNDDS) 2021–2023," 2024, accessed August 18, 2026. [Online]. Available: <https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fndds-download-databases/>
- [19] Healthy Eating Research, "Healthy eating research nutrition guidelines for the charitable food system," Healthy Eating Research, Tech. Rep., 2020, expert Panel Report; accessed August 18, 2026. [Online]. Available: <https://healthyeatingresearch.org/research/healthy-eating-research-nutrition-guidelines-for-the-charitable-food-system/>
- [20] R. Levi, M. Schwartz, E. Campbell, K. Martin, and H. Seligman, "Nutrition standards for the charitable food system: challenges and opportunities," *BMC Public Health*, vol. 22, no. 1, p. 495, 2022.
- [21] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM)*, 2008, pp. 263–272.
- [22] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009, pp. 452–461.
- [23] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 165–174.
- [24] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 639–648.
- [25] A. Drewnowski, "Defining nutrient density: Development and validation of the nutrient rich foods index," *Journal of the American College of Nutrition*, vol. 28, no. 4, pp. 421S–426S, 2009.
- [26] V. L. Fulgoni III, D. R. Keast, and A. Drewnowski, "Development and validation of the nutrient-rich foods index: A tool to measure nutritional quality of foods," *Journal of Nutrition*, vol. 139, no. 8, pp. 1549–1554, 2009.
- [27] A. Drewnowski, "The nutrient rich foods index helps to identify healthy, affordable foods," *American Journal of Clinical Nutrition*, vol. 91, no. 4, pp. 1095S–1101S, 2010.
- [28] A. Krohn-Grimberghe, L. Drumond, C. Freudenthaler, and L. Schmidt-Thieme, "Multi-relational matrix factorization using bayesian personalized ranking for social network data," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 173–182.
- [29] S. Watanabe, "Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance," in *arXiv:2304.11127*, 2023.
- [30] A. F. Rivera, N. R. Smith, and A. Ruiz, "A systematic literature review of food banks' supply chain operations with a focus on optimization models," *Journal of Humanitarian Logistics and Supply Chain Management*, vol. 13, no. 1, pp. 10–25, 2023.
- [31] I. S. Orgut, L. G. Brock, L. B. Davis, J. S. Ivy, S. Jiang, S. D. Morgan, R. Uzsoy, C. Hale, and E. Middleton, "Achieving equity, effectiveness, and efficiency in food bank operations: Strategies for feeding america with implications for global hunger relief," in *Advances in Managing Humanitarian Operations*. Springer, 2016, pp. 229–256.
- [32] Centers for Disease Control and Prevention, "August 2021–august 2023 dietary data," National Center for Health Statistics, 2024, accessed

August 18, 2026. [Online]. Available: <https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Dietary&Cycle=2021-2023>

- [33] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2018, pp. 974–983.