NUTRIGNN: Graph Neural Networks for Nutrient Imputation in LLM-Enriched Food Knowledge Graphs

Jasmine Lesner and Sathvika Anand
University of California, Santa Barbara

***Abstract*—Food composition databases are essential for nutrition research and policy, yet they remain incomplete, with only 31% of food-nutrient pairs directly measured in the USDA database. We present a novel approach to predict missing nutrient values by enriching food-nutrient knowledge graphs with information derived from Large Language Models (LLMs). Our method combines USDA data with OpenAI embeddings and GPT-4o-generated nutrient groupings to create a comprehensive graph with over 645,000 edges connecting 8,170 nodes. We train Graph Neural Networks (GNNs) on this enriched structure to predict missing nutrient values. Our best model achieves 67.55% accuracy (predictions within $\pm$30% of true values), substantially outperforming the baseline food group median imputation method (30.29%). Through ablation studies, we demonstrate that most LLM-derived features improve prediction performance, though some—like GPT-4o-generated nutrient groups—unexpectedly reduced accuracy. Our approach demonstrates how domain knowledge encoded in LLMs can enhance structured prediction tasks, particularly when dealing with incomplete data. This work contributes to making food composition databases more complete and useful for nutrition science and public health applications.**

***Index Terms*—Nutrient Prediction, Knowledge Graphs, Large Language Models, LLMs, Graph Neural Networks, GNNs, Embedding Vectors, Data Imputation, Food Composition Databases,**

I. INTRODUCTION

Food composition databases like the USDA's Food Nutrition Data are vital tools for research, diet planning, and policy. They aim to document the nutrient content of foods to serve scientists, nutritionists, and consumers [1]. Yet, these databases remain incomplete. In our analysis of USDA data, only 31% of the possible 1.17 million food-nutrient pairs have been directly measured. About 24% are estimated, and the rest are missing (Figure 1).

This data gap poses a major challenge. Direct nutrient measurement is costly and slow, requiring lab equipment and protocols. As a result, many foods lack full nutrient profiles, especially rare foods or nutrients that are harder to measure. These gaps limit how useful these databases can be for work that needs complete nutrition data.

Existing methods to predict missing values often use regression or clustering [2]. These rely on patterns in the structured data. While sometimes helpful, they struggle to capture deeper relationships—like those found in literature, recipes, or cultural food knowledge.

Our method addresses this by enriching a food-nutrient knowledge graph with knowledge from Large Language Models (LLMs). LLMs encode broad information, including nutritional science and food-related knowledge. By using LLMs and a graph-based imputation approach, we aim to improve predictions of missing nutrient values.

II. CONTRIBUTIONS

Our work makes the following contributions:

- We introduce a new method for enriching food-nutrient knowledge graphs using LLM-derived information.
- We build a graph (Figure 2) that combines USDA data, GPT-4o nutrient groupings, and OpenAI embeddings for food and nutrient names. We also use clustering to form new food groups.
- We show that Graph Neural Networks can use this enriched graph to predict missing nutrient values more accurately.
- We run an ablation study to measure how each part of our approach contributes to prediction performance.

III. BACKGROUND

A. Graph Neural Networks

Graph Neural Networks (GNNs) are a class of deep learning models designed to process graph-structured data and are effective for modeling relational information data. Common architectures include Graph Convolutional Networks [3] that aggregate information from neighboring nodes, Graph Attention Networks [4] that introduce attention mechanisms during aggregation, and GraphSAGE [5], which samples neighborhoods for inductive learning on large graphs. GNNs have been used in chemistry [6], biology [7], and recommendation systems [8].

B. Food Nutrition Databases

Nutrition databases developed by governments and researchers aim to track the nutrient content of foods [9]. The USDA database is one of the most detailed, with data on about 8,000 foods and 150 nutrients [1]. Other efforts include the UK's McCance and Widdowson's database [10] and the EuroFIR project [11]. Still, these databases all face the same challenge: missing data, due to the high cost of full lab testing [12].

C. Graphs for Nutrient Prediction

Možina et al. [13] built a knowledge graph where foods and nutrients are nodes and relationships are edges. Using the ComplEx model [14], they showed that graphs can reveal food similarities that simpler models miss. Their method turned nutrient prediction into a classification task by grouping nutrient values into bins (e.g., "between x and y"). Their best model achieved an MRR of 0.81 and Hits@10 of 0.94.

Our approach differs in key ways. We treat nutrient prediction as a regression task instead of classification. Rather than predicting a category, we predict exact values. This matters because:

1) Regression keeps fine-grained details of nutrient values.
2) It allows more precise predictions across wide value ranges.
3) It better handles the skewed and multi-modal data common in food composition.

Možina et al. included food groups in their graph. We go further, adding LLM-derived groupings for both foods and nutrients. Also, their study used 351 foods and 25 nutrients, while ours covers 7,800 foods and 150 nutrients.

IV. Method

A. Dataset

We use the USDA Food Nutrition Database, which includes ∼7,800 foods and 150 nutrients [1]. Its schema (Figure 4) includes tables for foods, nutrients, food groups, and nutrient values.

As shown in Figure 3, measurements of nutrients:

- Span five orders of magnitude
- Have distributions that are tail heavy
- Have distributions are highly skewed

Because of these traits adaptive scaling is required before using them for machine learning.

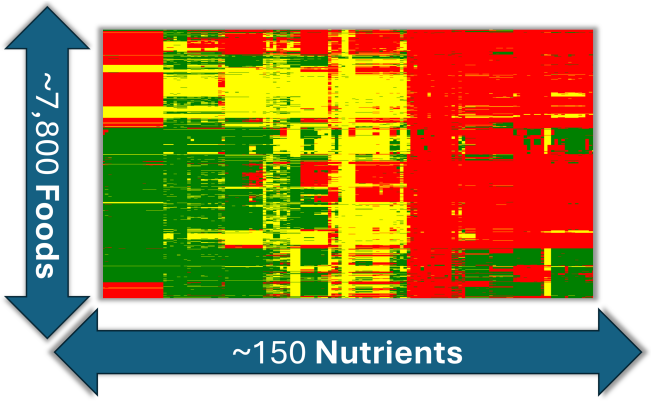

Fig. 1. Our analysis of USDA-tracked foods showing which nutrient values are measured (green), estimated (yellow), or missing (red). About 31% are measured, 45% are missing, and 24% are estimated. The clustered heatmap sorts similar foods (rows) and nutrients (columns) together to show missing data patterns.

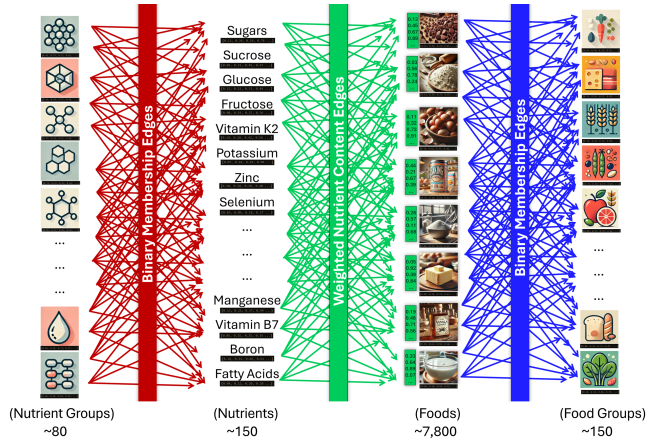

Fig. 2. Overview of our LLM-enriched knowledge graph for nutrient prediction. It combines USDA food composition data, OpenAI embeddings, and GPT-4o nutrient groups, creating a graph with over 645,000 edges.

B. Adaptive Scaling

To handle wide, skewed, and multi-modal nutrient distributions, we apply adaptive scaling. We first analyze each nutrient's distribution and then choose one of the transformation paths shown in Figure 7. This helps represent values in a way that suits neural network training, while keeping relative differences intact.

C. Graph Enrichment

We enrich the graph using four steps (Figure 5):

1. Vector Embedding: We use OpenAI's "text-embedding-3-small" model [15] to create vector embeddings of USDA food, nutrient, and group names. These serve as node features, capturing semantic meaning.

2. Nutrient Groups: Since the USDA does not define nutrient groups, we asked GPT-4o to create them and assign USDA nutrients to these groups. We first prompted the model to generate approximately 80 distinct nutrient groups, and then asked the model to categorize each of the 150 nutrients into these nutrient groups. For example, GPT-4o grouped calcium, iron, and zinc under 'Minerals'. This added ∼80 new nutrient groups.

3. Embedding Clusters: t-SNE plots (Figure 9) revealed clear subgroups within the USDA-provided food groups. We used K-means clustering with angular distance to define more granular groupings for both food and nutrient embeddings. Silhouette scores determined optimal cluster counts.

4. New Groups: We added the previously mentioned embedding clusters as food/nutrient group nodes, with edges connecting each food or nutrient to its respective group. This integration uses knowledge that large language models (LLMs) have learned from their extensive text training corpora. While embeddings are trained to predict words, and LLMs may generate plausible but false answers, we were curious if their general knowledge could still improve prediction.

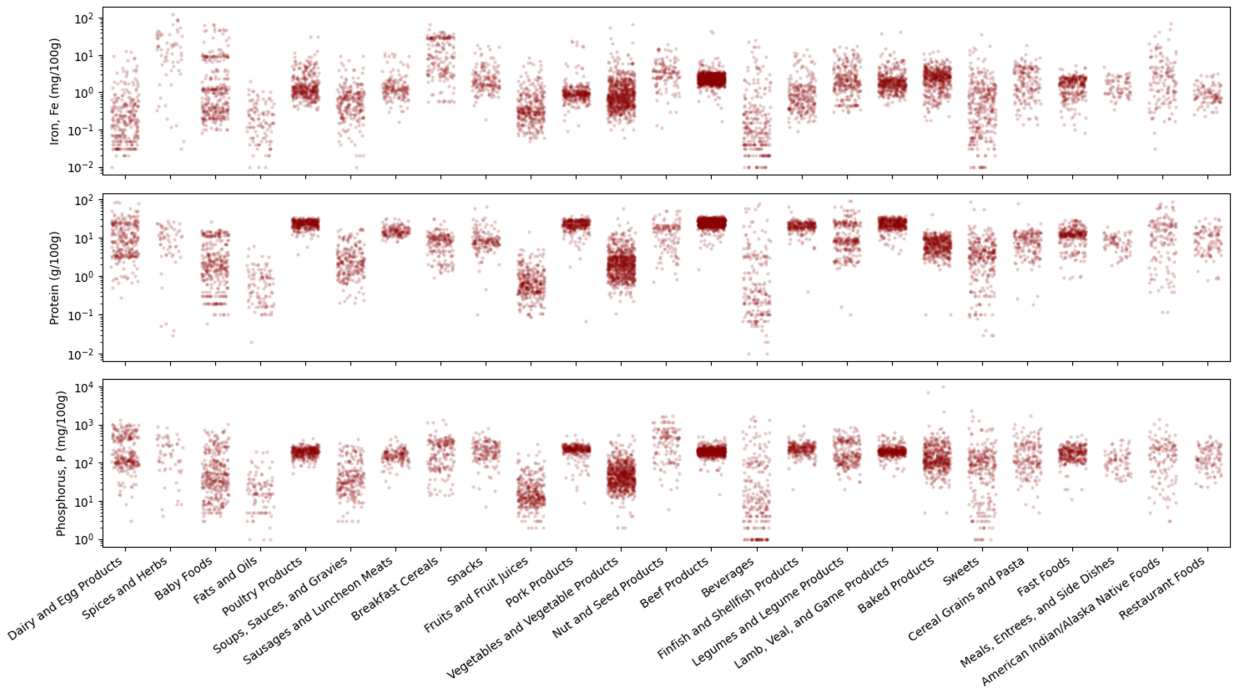

Fig. 3. Strip plots showing Iron, Protein, and Phosphorus levels in 7,800 foods, grouped by food type. USDA measures nutrients per 100g of edible food. The values span five orders of magnitude, are skewed, and often multi-modal.

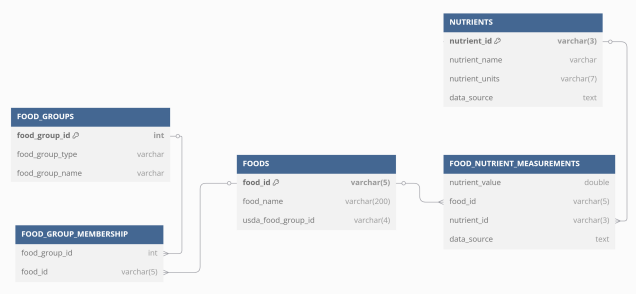

Fig. 4. USDA schema: The nutrient_value column in the food_nutrient_measurements table indicates how much of a nutrient (nutrient_id) is in a food (food_id).

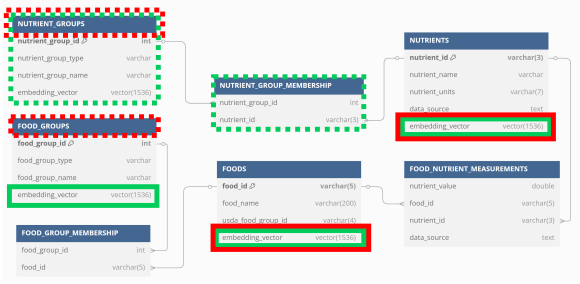

Fig. 5. To enrich the knowledge graph: (1) We added OpenAI embeddings as node features. (2) GPT-4o suggested nutrient groupings. (3) We clustered embedding vectors. (4) These clusters became additional food and nutrient groups.

D. Graph Composition

Our graph (Figure 6) includes: **Food nodes**: ∼7,800 foods tracked by USDA **Nutrient nodes**: ∼150 nutrients tracked by USDA. **Food group nodes**: ∼150 total, ∼25 from USDA, the rest from clustering. **Nutrient group nodes**: ∼150 total, ∼70 from GPT-4o, the rest from clustering. **Food-nutrient edges** show how much of a nutrient a food contains per 100g. **Food-food group edges** link foods to their groups. **Nutrient-nutrient group edges** link nutrients to their groups.

USDA does not supply nutrient groups and assigns foods to a single food group. We added nutrient groups (from GPT4-o), and unlike USDA's database schema, our graph allows foods and nutrients to be members of multiple groups.

E. GNN Architecture

Our best-performing GNN uses a 4-layer design with hidden dimension $h = 500$ and dropout rate $p = 0.1$. The model combines food and nutrient embeddings with bidirectional message passing across six relation types, including food-to-food group and food-to-nutrient edges. Node features are first transformed using layer normalization, followed by GraphSAGE convolutional layers with residual connections. Figure 8 shows a two-layer version of this architecture, which performed worse than the four-layer version but shares the same structure.

We use HuberLoss ($\delta = 1.0$) to reduce the effect of outliers in nutrient values, and optimize with AdamW ($\eta = 10^{-4}$, weight decay=10^{-4}). The learning rate decreases on plateau (factor=0.5, patience=20). Training runs for up to 3000 epochs with early stopping (patience=200), based on regularization loss. This setup supports strong nutrient prediction across diverse food types.

We also tested a GAT-based heterogeneous GNN with multi-head attention (heads=4). Our best performing version had

3 layers and used attention-based message passing over the same six relation types. While it applied attention weights to highlight key node relationships, it consistently underperformed compared to our GraphSAGE model. This is likely due to the fact that the GAT architecture does not support edge attributes. The GAT used the same layer normalization and residual connections, but struggled to model the complex structure of the nutrition graph.

We further explored a Transformer-based heterogeneous GNN using TransformerConv layers over the same six relation types. Although transformer layers can model long-range dependencies, the best Transformer-based model we found still performed poorly on our task.

F. Prediction Evaluation

We split our data 80/20 for training and validation and report performance only from validation. Metrics used: **Loss**: Mean Squared Error. **R^2**: Coefficient of determination. **Accuracy ±30%**: Fraction of predictions within ±30% of the true value. This range reflects natural variation in food composition caused by factors like soil, climate, and genetics. Nutrient values often vary by 20–30% even in lab settings, so this margin is a practical benchmark. No single threshold fits all nutrients or foods.

For our evaluation we did not rely on USDA's nutrient estimates (yellow in figure 1) but only considered actual nutrient measurements (green in figure 1).

V. Results

The common imputation method, which assigns missing values based on the median of each food group, achieves an accuracy (±30%) of 0.3029 on our dataset. By contrast, the first version of our GNN model (Section IV-E) with all enrichments included reaches an accuracy (±30%) of 0.6654. When graph enrichment is removed, accuracy drops to 0.6080. This 6-point drop shows that predictions improve when using an LLM-enriched knowledge graph.

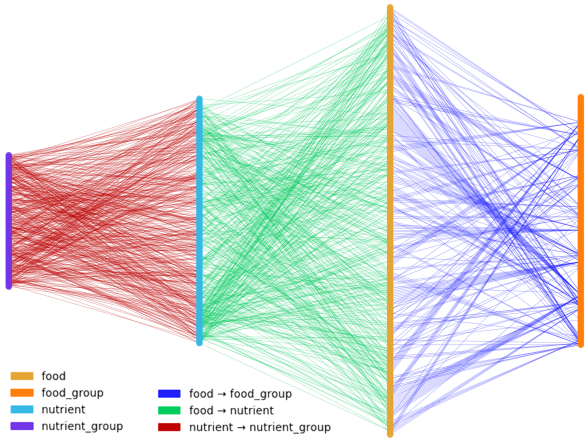

Fig. 6. Structure of our knowledge graph. Nodes: foods (yellow), nutrients (blue), food groups (orange), and nutrient groups (purple). Graph has 8,170 nodes and 645,209 edges.

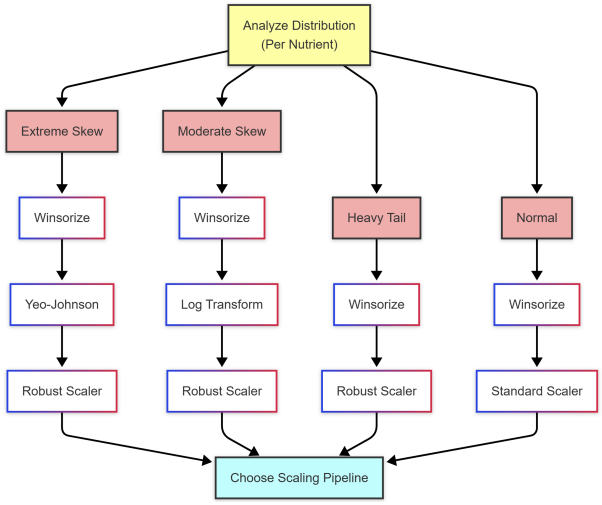

Fig. 7. Adaptive scaling: Each nutrient's distribution is analyzed, then a suitable transformation is applied.

In our first ablation study (Table I), we found that disabling nutrient embedding vectors actually improved performance to 0.6712. This suggests that including these embeddings was hurting accuracy.

We removed nutrient groups and repeated the ablation study (Table II). This led to our highest accuracy: 0.6755.

These nutrient groups—and their assignments—were generated entirely by GPT-4o. While the output initially appeared reasonable, our results suggest that GPT-4o did not produce nutrient groupings that improved prediction accuracy.

VI. Discussion

While our results show that GNNs can benefit from LLM-enhanced food nutrient graphs, several areas remain for future work:

- We tuned hyperparameters using a single RTX4090 GPU over two days. Further tuning may lead to better performance.

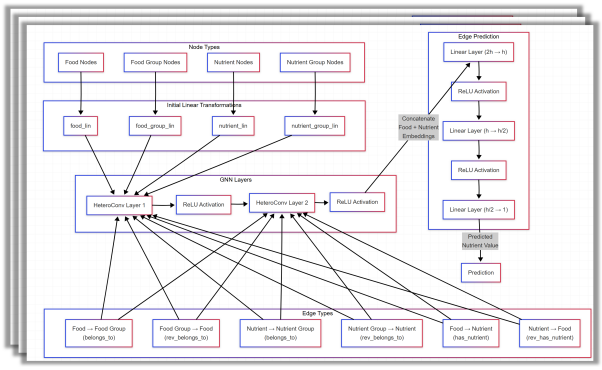

Fig. 8. GNN architecture for nutrient prediction. Various node types go through initial transformations, then heterogeneous graph convolutions. Our best performing models use four GNN layers.This diagram shows only two GNN layers for clarity.

TABLE I
SINGLE FEATURE ABLATION STUDY

Model Configuration	MSE (Loss)	R^2	Accuracy (±30%)
Imputation with Food Group Means	-	-	0.1918
Imputation with Food Group Medians	-	-	0.3029
All Enrichment Enabled	0.0501	0.8754	0.6654
All Enrichment Disabled	0.0796	0.7995	0.6080
Food Groups (FGs) Disabled	0.0507	0.8736	0.6669
Nutrient Groups (NGs) Disabled	0.0502	0.8748	0.6641
Food Embedding Vectors (FEV) Disabled	0.0705	0.8228	0.6293
Nutrient Embedding Vectors (NEV) Disabled	0.0519	0.8702	0.6569
Food Group Embedding Vectors (FGEV) Disabled	0.0496	0.8761	0.6665
Nutrient Group Embedding Vectors (NGEV) Disabled	**0.0481**	**0.8811**	**0.6718**

TABLE II
DUAL FEATURE ABLATION: NUTRIENT GROUP EMBEDDING VECTORS (NGEV) ALWAYS DISABLED

Model Configuration	MSE (Loss)	R^2	Accuracy (±30%)
Imputation with Food Group Means	-	-	0.1918
Imputation with Food Group Medians	-	-	0.3029
NGEV Disabled	0.0509	0.8733	0.6629
NGEV and FGs Disabled	0.0491	0.8783	0.6699
NGEV and NGs Disabled	**0.0484**	**0.8796**	**0.6755**
NGEV and FEV Disabled	0.0738	0.8148	0.6143
NGEV and NEV Disabled	0.0541	0.8656	0.6476
NGEV and FGEV Disabled	0.0491	0.8780	0.6667

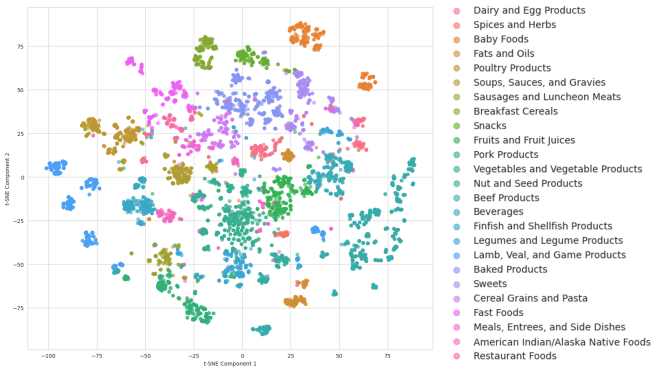

Fig. 9. t-SNE plot of food embedding vectors, colored by cluster. Clear separations show that the embeddings capture meaningful groupings.

- We did not explicitly model food processing methods (e.g., raw vs. cooked) or regional variations. Some of these are implicitly captured in USDA labels used for embedding vector computation.
- Prediction accuracy varies across nutrients, but we did not study which ones perform well or poorly. Future work should include this analysis and consider a confidence model to indicate prediction reliability.

VII. CONCLUSION

This paper presents a new approach to food nutrient prediction using GNNs applied to an LLM-enriched knowledge graph. By incorporating semantic knowledge from LLMs, we build a graph that helps predict missing nutrient values.

Our best GNN achieves 0.6755 accuracy (±30%), far outperforming the food group median baseline of 0.3029 on the same USDA dataset of 7,800 foods and 150 nutrients. While LLM-derived features generally improve performance, ablation studies are essential to identify which ones help. For example, we found that removing GPT-4o-generated nutrient groups led to better predictions. All other enrichments contributed to a 6-point gain.

Our results show the potential of using LLMs to enhance structured prediction, especially when data is incomplete. This method could apply to other domains where domain-specific knowledge from LLMs can improve predictions.

By improving estimates of missing nutrient values, our work helps make food composition databases more complete and useful. These databases are key to nutrition research, public health, and food policy.

AUTHOR CONTRIBUTIONS

J. L. used LLMs to enhance the USDA food nutrition dataset and implemented adaptive scaling. Both S. A. and J. L. evaluated GNNs for nutrient value prediction. S. A. explored different GNN architectures, and J. L. conducted ablation studies on the best-performing one. Both authors contributed to writing the manuscript.

REFERENCES

[1] D. B. Haytowitz, J. K. Ahuja, X. Wu, M. Somanchi, M. Nickle, Q. A. Nguyen, *et al.*, "USDA National Nutrient Database for Standard Reference, Legacy Release," 2019.

[2] S. F. Schakel, I. M. Buzzard, and S. E. Gebhardt, "Procedures for estimating nutrient values for food composition databases," *Journal of Food Composition and Analysis*, vol. 10, pp. 102–114, 1997. ARTICLE NO. FC970527.
[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.
[4] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations (ICLR)*, 2018.
[5] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.
[6] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*, pp. 1263–1272, PMLR, 2017.
[7] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.
[8] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 974–983, 2018.
[9] H. Greenfield and D. A. Southgate, "Food composition data: production, management, and use (food & agriculture org.)," 2003.
[10] R. A. McCance and E. M. Widdowson, *McCance and Widdowson's the Composition of Foods*. Royal Society of Chemistry, 2014.
[11] EuroFIR AISBL, "European food information resource network," 2018. Information resource for food composition data in Europe.
[12] J. K. Ahuja, A. J. Moshfegh, J. M. Holden, and E. Harris, "USDA food and nutrient databases provide the infrastructure for food and nutrition research, policy, and practice," *The Journal of nutrition*, vol. 143, no. 2, pp. 241S–249S, 2013.
[13] M. Možina, S. Žitnik, B. K. Seljak, and T. Eftimov, *Enhancing Food Composition Databases: Predicting Missing Values via Knowledge Graph Embeddings*. PhD thesis, Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2023.
[14] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International conference on machine learning*, pp. 2071–2080, PMLR, 2016.
[15] OpenAI, "OpenAI's text-embedding-3-small Embedding Model." https://platform.openai.com/docs/guides/embeddings, 2024. Semantic numerical representation of food_name and food_group.