machine learning & knowledge extraction

MDPI

Article

Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes

Luís Rita [1], Joshua Southern [2], Ivan Laponogov [1], Kyle Higgins [1,3] and Kirill Veselkov [1,4,*]

[1] Division of Cancer, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London W12 0NN, UK; l.rita19@imperial.ac.uk (L.R.); i.laponogov@imperial.ac.uk (I.L.); kyle.higgins@childrens.harvard.edu (K.H.)
[2] Department of Computing, Faculty of Engineering, Imperial College London, London SW7 2AZ, UK; joshua.southern17@imperial.ac.uk
[3] Department of Neurobiology, Boston Children's Hospital and Harvard Medical School, Boston, MA 02115, USA
[4] Department of Environmental Health Sciences, Yale University, New Haven, CT 06520, USA
* Correspondence: kirill.veselkov04@imperial.ac.uk

Abstract: In the emerging field of computational gastronomy, aligning culinary practices with scientifically supported nutritional goals is increasingly important. This study explores how large language models (LLMs) can be applied to optimize ingredient substitutions in recipes, specifically to enhance the phytochemical content of meals. Phytochemicals are bioactive compounds found in plants, which, based on preclinical studies, may offer potential health benefits. We fine-tuned models, including OpenAI's GPT-3.5-Turbo, DaVinci-002, and Meta's TinyLlama-1.1B, using an ingredient substitution dataset. These models were used to predict substitutions that enhance the phytochemical content and to create a corresponding enriched recipe dataset. Our approach improved the top ingredient prediction accuracy on substitution tasks, from the baseline 34.53 $\pm$ 0.10% to 38.03 $\pm$ 0.28% on the original substitution dataset and from 40.24 $\pm$ 0.36% to 54.46 $\pm$ 0.29% on a refined version of the same dataset. These substitutions led to the creation of 1951 phytochemically enriched ingredient pairings and 1639 unique recipes. While this approach demonstrates potential in optimizing ingredient substitutions, caution must be taken when drawing conclusions about health benefits, as the claims are based on preclinical evidence. This research represents a step forward in using AI to promote healthier eating practices, providing potential pathways for integrating computational methods with nutritional science.

Keywords: ingredient substitution; nutritional optimization; large language models

Citation: Rita, L.; Southern, J.; Laponogov, I.; Higgins, K.; Veselkov, K. Optimizing Ingredient Substitution Using Large Language Models to Enhance Phytochemical Content in Recipes. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2738–2752. https://doi.org/10.3390/make6040131

Received: 29 September 2024
Revised: 16 November 2024
Accepted: 18 November 2024
Published: 26 November 2024

1. Introduction

In recent years, computational gastronomy has emerged as an interdisciplinary field that applies computational techniques such as data mining and machine learning to the study of food and cooking. The aim is to understand and model the complex interactions between ingredients, cooking methods, and the human perception of taste and nutrition. One of the primary goals of this field is to develop methods for ingredient substitution, aimed at improving the nutritional content, preserving the flavor integrity, and aligning meals with specific dietary needs. A key focus has been the integration of phytochemically enriched ingredients into diets, which has shown in silico potential to target biological networks of chronic diseases like cancer [1], Alzheimer's disease (AD) [2], and COVID-19 [3].

Phytochemicals, bioactive compounds found in plants, have gathered significant attention due to their antioxidant, anti-inflammatory, and anti-carcinogenic properties. Preclinical studies suggest that these compounds may play a role in disease prevention and treatment. For instance, brassinolide, a phytochemical present in tea, has shown potential to inhibit tumor growth and induce apoptosis in cancer cells [4]. In the context of AD,

quercetin, found in extra virgin olive oil, has been linked to improved brain health by exhibiting antioxidant and anti-inflammatory effects [5]. Moreover, genistein, a phytochemical in blackcurrant, has been investigated for its immune-supporting properties, including its potential to modulate inflammation and interfere with viral replication, making it relevant in the study of COVID-19 [6].

Initial attempts at ingredient substitution utilized statistical methods, such as Term Frequency–Inverse Document Frequency (TF-IDF), to identify potential substitutes based on occurrence patterns within large recipe datasets. TF-IDF is a numerical statistic that reflects how important a word (or ingredient) is to a document (or recipe) in a corpus. It does this by considering both the frequency of the word in the document and the rarity of the word across all documents. In the context of ingredient substitution, TF-IDF helps highlight ingredients that are significant within certain recipes but not ubiquitous across all recipes, thus identifying potential substitutes based on uniqueness and relevance [7–9]. Later, co-occurrence-based methods refined this approach by constructing ingredient networks that map relationships across recipes. In these networks, ingredients are represented as nodes, and the edges between them indicate how frequently they appear together. By analyzing these networks, researchers could suggest substitutes based on ingredients' mutual presence in culinary contexts, identifying clusters of ingredients that are often used interchangeably [10–14]. The introduction of language model-based methods marked a significant evolution, utilizing natural language processing techniques such as word2vec [15], BERT [16], and R-BERT [17] to capture semantic relationships between ingredients. Word2vec generates vector representations of words (or ingredients) based on their contexts in the text, allowing the model to identify ingredients with similar contexts or meanings. BERT (Bidirectional Encoder Representations from Transformers) goes further by understanding the bidirectional context of words, providing a deeper semantic understanding. R-BERT specializes in relation extraction, identifying and classifying relationships between entities—in this case, between ingredients. This approach proved effective in improving ingredient substitution tasks through learned embeddings that capture complex semantic relationships [18], although language models require substantial computational resources and may not always capture the full culinary context, such as flavor profiles or cooking techniques.

More recently, graph neural networks (GNNs) have been utilized to combine the relational information encoded in ingredient graphs with the specific context of given recipes, leading to a deeper understanding of ingredient interactions [18]. GNNs are designed to operate on graph structures, modeling the dependencies and relationships between nodes (ingredients) and edges (their relationships). Large-scale graphs, such as FlavorGraph, have been introduced to explore ingredient substitutions and food pairings [19]. FlavorGraph connects ingredients based on shared flavor compounds and culinary usage, providing a rich dataset for analyzing how ingredients relate on a molecular level. This graph-based approach allows for the identification of substitutes that are not only contextually appropriate but also compatible in terms of flavor and chemistry. However, success in this area relies heavily on the quality and curation of the underlying graph data; inaccuracies or omissions can significantly affect the model's performance. Building on this approach, GISMo was introduced—a GNN-based model that incorporates both recipe-specific contexts and ingredient relationships from FlavorGraph. By constructing a benchmark dataset, Recipe1MSubs, which includes ingredient substitution pairs extracted from user comments, GISMo significantly outperforms previous methods in ranking plausible ingredient substitutions. Specifically, it achieved a performance improvement of at least 14% in the top substitute ingredient prediction, as measured by the Hit@1 metric, over existing models [20]. In the context of ingredient substitution, Hit@1 evaluates whether the model's first (most confident) suggested substitute matches the actual substitute used in the recipe. This metric is important because it reflects the model's effectiveness in providing accurate substitution suggestions on the first attempt, which is essential for real-world application.

The latest stage in this evolving field is represented by LLMs, which promise to overcome the limitations of previous approaches by leveraging their capacity for understanding and generating human-like text [21,22]. The introduction of LLMs, such as GPT-3 developed by OpenAI [21], presents an approach to address the limitations of previous methods for ingredient substitution. Furthermore, while language model-based methods and GNNs represent significant advancements, they still face challenges in capturing the full culinary context and ensuring gastronomically sensible substitutions [20]. LLMs, trained on extensive and diverse culinary datasets, can potentially offer more contextually aware and accurate ingredient substitutions by leveraging their understanding of both the syntax and semantics of culinary texts [23]. This capacity for high-level language comprehension and manipulation allows for considering factors such as ingredient compatibility. Importantly, LLMs can be fine-tuned for specific tasks such as ingredient substitution [22].

Recognizing the limitations of statistical, co-occurrence, language, and GNN-based methods, our research proposes a unique approach by leveraging the capabilities of LLMs for ingredient substitution. LLMs, such as GPT-3.5 [22], DaVinci [21], and Meta's TinyLlama [24], have demonstrated state-of-the-art performance across a range of natural language processing tasks, from text generation to semantic understanding [16,25]. By fine-tuning these models on a dataset of recipes and ingredient substitutes, we aim to develop an algorithm that not only understands the interplay of flavors and nutritional aspects in cooking but also tailors suggestions to the preferences and requirements of each user. In this paper, we benchmarked our ingredient substitution algorithm against the current state of the art GISMo to demonstrate its superiority in generating contextually appropriate ingredient substitutions. We used the Hit@1 accuracy metric to benchmark our models' performance against state-of-the-art methods. After identifying phytochemically enriched substitutes, we generated a new set of recipes aimed at targeting biological networks associated with cancer, AD, and COVID-19 (Figure 1).

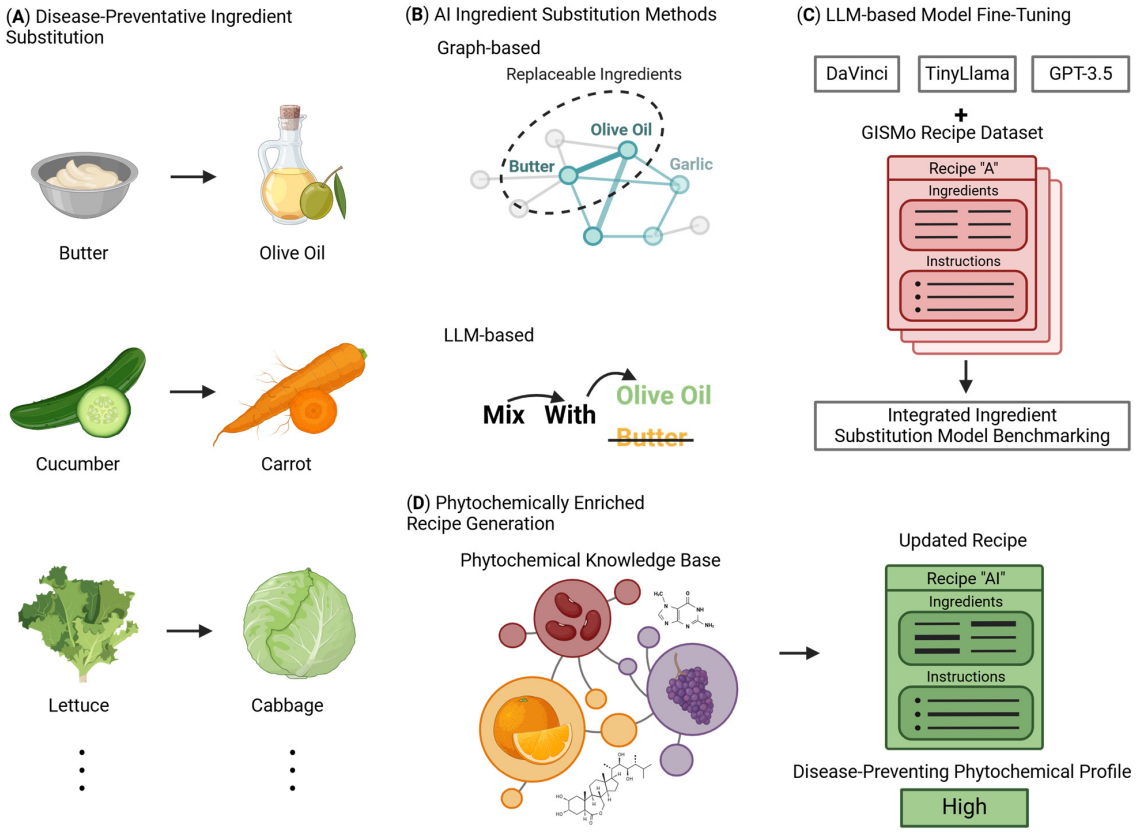

Figure 1. Ingredient substitution methods. (**A**) Disease-preventative ingredient substitution: it illustrates the process of substituting ingredients to enhance the phytochemical profile of recipes

focused on disease-specific prevention. Examples include substituting butter with olive oil, cucumber with carrot, and lettuce with cabbage. (**B**) Ingredient substitution methods: it compares graph-based approaches, which rely on ingredient co-occurrence and relational data, with LLM-based approaches, which utilize advanced language models for more context-sensitive substitutions. (**C**) LLM-based model fine-tuning: it details the fine-tuning of LLMs such as DaVinci, GPT-3.5, and TinyLlama for the ingredient substitution task using the Recipe1MSubs dataset. It includes the benchmarking process against the GISMo model to evaluate performance improvements. (**D**) Phytochemically enriched recipe generation: it details the generation of recipes with enhanced phytochemical profiles tailoring diseases' biological networks with the best performer model.

The research hypothesis of the article is that LLMs can achieve higher accuracy in ingredient substitution tasks compared to the current state-of-the-art GISMo model when evaluated on a standardized dataset. The main contributions of this paper are (1) enhanced accuracy in ingredient substitution, (2) a novel dataset filtration process, and (3) the generation of phytochemically enriched recipes. The rest of this paper is organized as follows. In Section 2, we detail the materials and methods employed, including the datasets used, the fine-tuning of LLMs, and the evaluation metrics for ingredient substitution accuracy. Section 3 presents the results of our experiments, comparing the performance of fine-tuned LLMs with the GISMo model and showing the generation of phytochemically enriched recipes. Section 4 provides a discussion of our findings, highlighting the improvements achieved, the implications for computational gastronomy, and the limitations of our approach. Finally, Section 5 concludes the paper by summarizing our contributions and suggesting directions for future research in integrating AI with nutritional science to promote healthier eating practices.

2. Materials and Methods

2.1. Recipe and Ingredient Substitution Datasets

Our research started with the study of the Recipe1MSubs dataset, provided by Meta, containing 70,520 pairs of ingredient substitutes with the respective recipes [20], which is a subset of the Recipe1M dataset [26]. The Recipe1MSubs dataset was separated into 49,044 data points designated for training, 10,729 for validation, and 10,747 for testing. Each recipe within this dataset is organized in a structured format, beginning with the recipe title, followed by the list of ingredients, with associated quantities, and finally, the cooking instructions. The original GISMo model was trained on this dataset using a methodology focused on ingredient context and co-occurrence as the benchmark for our study.

2.2. GISMo Benchmark

To establish a baseline for comparison with our new LLM-based models, we re-implemented and re-ran the GISMo as described in the original study. We set the learning rate to 5×10^{-5}, weight decay to 0.0001, and used an embedding dimension of 300 to represent ingredients in a continuous vector space. The model consists of two graph convolutional layers, each with 300 hidden units, and applies a dropout rate of 0.25 to reduce overfitting. Training was conducted over 400 epochs using regular negative sampling, where for each positive substitution pair, negative examples were generated by randomly selecting non-substitutable ingredients, and embeddings were initialized randomly. Average pooling was used for contextual embedding to aggregate information from neighboring nodes, enhancing the model's context sensitivity without altering the original dataset's composition. We re-ran GISMo not only to replicate the results of the original study but also to establish a standard benchmark against which we could evaluate the performance of our newer LLM-based models, ensuring that any improvements in ingredient substitution accuracy were attributable to the capabilities of the LLMs rather than differences in the experimental setup. Furthermore, we introduced enhancements to the GISMo model by incorporating each ingredient's food category as an additional node feature in the graph, providing higher-level semantic knowledge to potentially improve substitution sugges-

tions (as described in Section 2.3). Additionally, we applied a dataset filtration process (as described in Section 2.4) to the original Recipe1MSubs dataset to remove incorrect or unsuitable substitutions, training GISMo on this filtered dataset to assess whether cleaner training data could enhance the model's performance.

2.3. Incorporation in GISMo of a Food Category Feature

Using GPT-4-0613, the latest of OpenAI's language models, we categorized ingredients into predefined culinary groups. This process involved a Python script utilizing the pandas library for dataset manipulation and the openai library for API interactions. A function, categorize_ingredient, was used to query GPT-4-0613 with each ingredient, requesting its classification into one of 23 categories ranging from common food groups like Fruits and Vegetables to more specialized ones such as Confectioneries and Aquatic foods (Appendix A). By setting the temperature parameter to 0, the script prioritized reproducibility to ensure consistency in GPT-4-0613's responses. This approach processed a CSV file of ingredients, appending a category column with the GPT-4-0613-determined categories to the dataset. The augmented dataset, saved as a new CSV file, served as a tool for ingredient substitution models, enabling more contextually relevant substitutions.

2.4. Dataset Filtration Based on Substitution Validity

To enhance the ingredient substitution model's accuracy, we used GPT-3.5-Turbo with an asynchronous Python script to evaluate the validity of the proposed ingredient substitutions. This process involved sending detailed prompts to GPT-3.5-Turbo, asking if one ingredient could feasibly substitute another within a specific recipe, and classifying responses into Correct, Potential, or Incorrect to determine their suitability. By processing substitutions in multiple batches using the aiohttp library for asynchronous HTTP requests, we efficiently assessed the 70,520 substitutions, thereby accelerating the evaluation process. Substitutions categorized as Correct were considered suitable and retained, Potential indicated possible suitability requiring further consideration, and Incorrect were considered inappropriate, leading to their removal from the dataset. The final results, saved into a JSON file, formed a filtered dataset for retraining the model, ensuring it was based on accurate substitution data. Key settings included a prediction temperature of 0.5, a limit of 10 output tokens, and five runs to evaluate the prediction stability. In addition, there was a batch size of 100 substitutions with respective recipes to avoid reaching the maximum number of requests per second.

2.5. Fine-Tuning Language Models for Substitution Predictions

We used GPT-3.5-Turbo-1106, DaVinci-002, and TinyLlama-1.1B to predict viable ingredient substitutions, fine-tuning each with consistent specifications to ensure comparability. Key settings included a prediction temperature of 0.5, a limit of 10 output tokens, and five runs to evaluate the prediction stability, all conducted over a single epoch.

For the fine-tuning process for TinyLlama-1.1B models in our experimental configuration, we refined our model's fine-tuning process with selected hyperparameters encapsulated within the TrainingArguments setup. This configuration specified an output directory, a per-device train batch size of 8 (due to memory constraints) and applied gradient accumulation over 4 steps to efficiently balance computational demand and memory constraints. The model optimization was conducted using paged_adamw_32bit with a learning rate set at 5×10^{-4}, and a cosine learning rate scheduler was employed for optimal learning rate adjustments throughout the training phase. A save strategy based on epochs was utilized, coupled with logging and evaluation intervals set at 25 and 50 steps, respectively, aligning with an evaluation strategy that triggers at specified steps to closely monitor the model's performance. The training was streamlined to complete within 1 epoch to ensure quick adaptation while preventing overfitting, without setting a maximum step limit and avoiding mixed precision training to maintain computational accuracy. The SFTTrainer was used in the training process, directly interfacing with the training and validation datasets, and

was configured with peft_config for tailored pre-fine-tuning adjustments. This allowed us to set our specified hyperparameters and training configurations. Text preprocessing was managed using a specified dataset_text_field and tokenizer, with packing disabled and a maximum sequence length of 512 to standardize input data handling. This approach aimed at enhancing the model's learning efficiency, prioritizing a balance between optimizing the computational resources and achieving high-quality model training.

Building upon the filtration methodology outlined above, we randomly chose one of the filtered datasets and fine-tuned four final models considering only the Correct substitutions to further refine the accuracy of predictions. TinyLlama-1.1B, DaVinci-002, GPT-3.5-Turbo-1106, and GISMo models were fine-tuned incorporating these high-quality substitutions.

Training samples were provided in prompt completion format for DaVinci-002 and TinyLlama-1.1B and chat completion format for GPT-3.5-Turbo-1106. The number of epochs, training steps, and batch sizes chosen are detailed in Appendix B.

2.6. Evaluation of Ingredient Substitution Accuracy

To validate the accuracy of the ingredient substitution predictions generated by LLMs, we developed an algorithm to standardize and process ingredient names before comparing them to a ground truth dataset derived from the Recipe1M dataset. We began by extracting predictions from the model output, where each line contained an original ingredient, its corresponding ground truth substitute, and the predicted substitution. To ensure consistency across ingredient names, several preprocessing steps were applied, including converting all text to lowercase, removing numeric values, and applying predefined rules to replace or eliminate special characters. This normalization was intended to maintain uniformity in ingredient representation. After preprocessing, a clustering mechanism was used to group similar ingredients, accounting for variations in lexical forms such as singular and plural versions or different types of the same ingredient (e.g., basmati rice and long grain rice). Each ingredient was assigned a unique cluster identifier to ensure that similar ingredients were treated as equivalent during comparison.

Once the LLM-predicted ingredient names were uniformized and categorized, the core of the evaluation involved comparing the predicted substitutions against the ground truth using the Hit@1 metric. This metric assessed the model's precision by determining whether the first predicted substitution matched the ground truth or fell within the same ingredient cluster. For example, if the ground truth substitution was barley and the model predicted basmati rice, both ingredients would be considered correct if they belong to the same grain cluster. Hit@1 focuses on measuring the accuracy of the model's top recommendation, as this is the most critical in real-world applications where users often act on the first suggestion. By prioritizing precision in the initial substitution, Hit@1 provides a measure of the LLM's ability to generate viable and contextually appropriate ingredient substitutions.

2.7. Phytochemically Enriched Recipe Generation

Finally, we integrated phytochemically enriched ingredients based on their ability to target molecular networks responsible for disease development in cancer [1], AD [2], and COVID-19 [3]. By applying the best-performing model from our comparative analysis, we substituted all ingredients across our dataset with alternatives that elevated the content of the targeted phytochemicals. The recipes were then evaluated and ranked based on their cumulative phytochemical profile. Only salads were considered given the lower number of cooking processes involved in their preparation and, consequently, the higher chances of phytochemical preservation [27] (Figure 2).

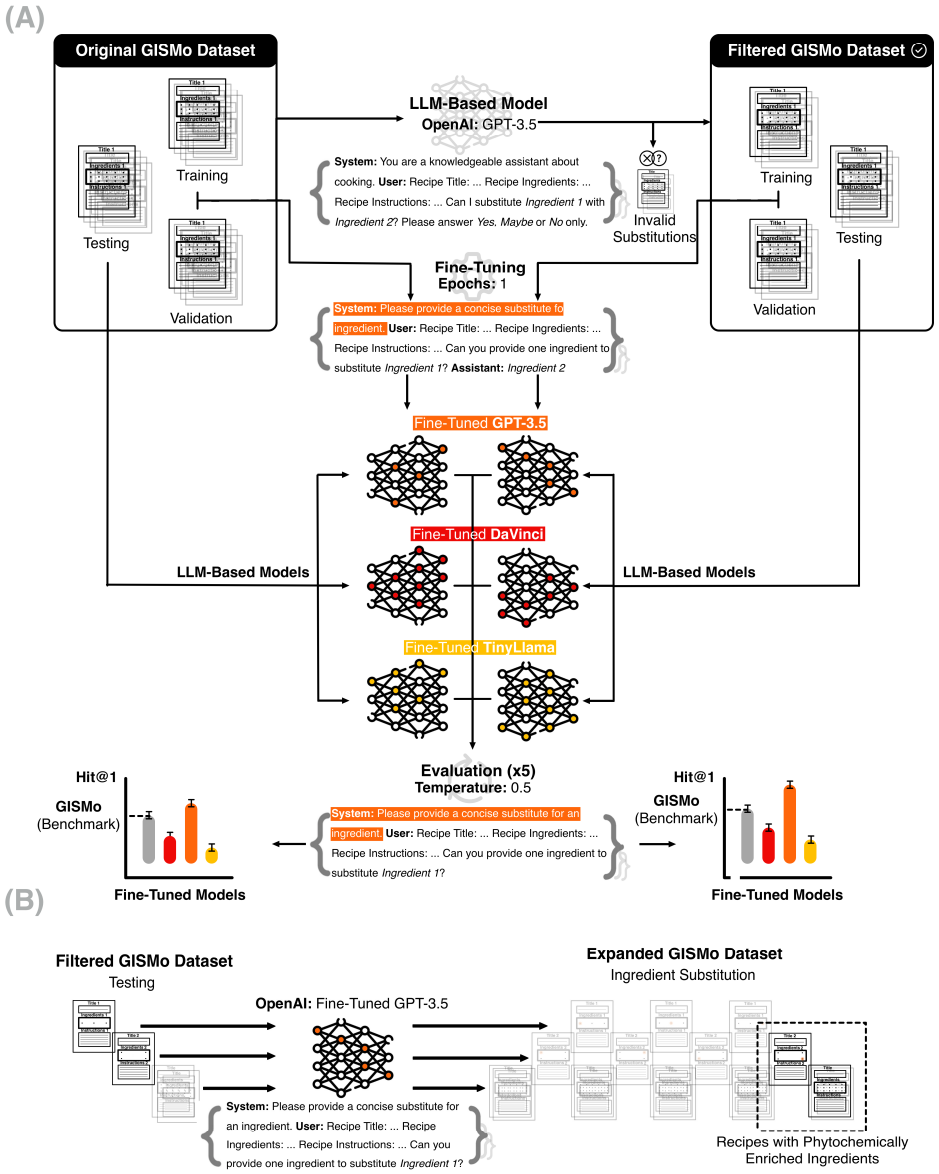

Figure 2. This figure is divided into two parts, (**A**,**B**), providing an overview of the methodology in our study. (**A**) The fine-tuning and evaluation process for three LLMs—OpenAI's GPT-3.5, DaVinci, and Meta's TinyLlama—benchmarked against the state-of-the-art GISMo model for ingredient substitution. The figure starts with the Original GISMo Dataset, which undergoes initial processing through GPT-3.5 to identify invalid substitutions, producing a Filtered GISMo Dataset used for training, validation, and testing. Each LLM was fine-tuned with a single epoch, using prompt engineering to structure responses for precise ingredient substitution. Prompts included recipe context such as title, ingredients, and instructions, asking the model to propose a suitable substitute ingredient. The fine-tuned models were then evaluated using the Hit@1 metric, where the top prediction was checked against ground truth substitutions, with a temperature setting of 0.5 to balance creativity and accuracy. The Hit@1 evaluation results are shown, with significant improvements in accuracy for the fine-tuned GPT-3.5 model over the GISMo benchmark, demonstrating the effectiveness of LLM-based methods in generating contextually appropriate substitutions. Fine-tuned models GPT-3.5, DaVinci and TinyLlama are represented with the colors orange, red and yellow, respectively. (**B**) The application of the best-performing model (fine-tuned GPT-3.5) for generating new recipes. Using the Filtered GISMo Dataset for testing, the model identifies suitable ingredient substitutions to create an Expanded GISMo Dataset featuring recipes with enhanced phytochemical content. This process involves using the LLM to suggest ingredient substitutions that increase the phytochemical profile,

targeting health benefits associated with bioactive compounds. The expanded dataset thus contains recipes with optimized nutritional content, focusing on the potential health benefits of phytochemically enriched ingredients. The workflow includes model selection, recipe generation, and selection based on phytochemical content, illustrating a pathway from dataset refinement to practical application in creating health-focused recipes.

3. Results

3.1. Dataset Filtering and Preparation

The original Recipe1MSubs dataset comprises 70,520 ingredient substitutions, partitioned into 49,044 for training, 10,729 for validation, and 10,747 for testing. To enhance the quality of the dataset, we applied a filtration process using GPT-3.5-Turbo-1106 across five separate runs. This process evaluated the validity of each substitution, classifying them as Correct, Potential, or Incorrect based on their suitability within specific recipes.

The filtration resulted in five filtered datasets with the following average sample sizes for training, validation, testing, and total: 31,819 $\pm$ 67, 7094 $\pm$ 25, 7085 $\pm$ 21, and 45,998 $\pm$ 85, respectively (see Appendix C for detailed statistics).

From these, we randomly selected one filtered dataset containing 44,615 ingredient substitutions, divided into 31,063 for training, 6831 for validation, and 6721 for testing. This refined dataset served as the basis for rerunning the GISMo model and fine-tuning the LLMs.

3.2. Performance of GISMo Model

We first evaluated the GISMo model by running it five times on both the original and filtered datasets. The performance was assessed using the Hit@1 accuracy metric, which measures the proportion of times the top predicted substitute matches the ground truth. On the original and filtered dataset, the Hit@1 accuracy was 34.53 $\pm$ 0.10% and 40.24 $\pm$ 0.36%, respectively.

These results indicate that filtering the dataset improved the GISMo model's performance by approximately 6%. Additionally, we experimented with incorporating food categories (Appendix A) as additional node features into the GISMo model. However, this modification resulted in a similar Hit@1 accuracy of 34.62% (Appendix D), suggesting no significant benefit from this approach.

3.3. Fine-Tuning Large Language Models

Subsequent experiments with fine-tuned LLMs on the original GISMo ingredient substitution dataset yielded Hit@1 values of 20.09 $\pm$ 0.31% for DaVinci-002, 20.93 $\pm$ 0.29% for TinyLlama-1.1B, and 38.03 $\pm$ 0.28% for GPT-3.5-Turbo-1106. Using the filtered dataset in conjunction with these fine-tuned models resulted in improved Hit@1 values of 29.43 $\pm$ 0.30%, 34.53 $\pm$ 0.32%, and 54.46 $\pm$ 0.29%, respectively (Appendix D). The models were evaluated five times each, and the average Hit@1 accuracies are presented in Table 1.

Table 1. Performance of fine-tuned models on the Recipe1MSubs dataset. Hit@1 accuracies are reported as mean $\pm$ standard deviation. The best performance for each dataset is highlighted in bold.

Recipe1MSubs Dataset	Fine-Tuned Model	Hit@1 (%)
Filtered	**GPT-3.5-Turbo-1106**	**54.46 $\pm$ 0.29**
	GISMo	40.24 $\pm$ 0.36
	TinyLlama-1.1B	34.53 $\pm$ 0.32
	DaVinci-002	29.43 $\pm$ 0.30
Unfiltered	**GPT-3.5-Turbo-1106**	**38.03 $\pm$ 0.28**
	GISMo	34.55 $\pm$ 0.11
	TinyLlama-1.1B	20.93 $\pm$ 0.29
	DaVinci-002	20.09 $\pm$ 0.31

3.4. Generation of Phytochemically Enriched Recipes

Leveraging the higher performance of the fine-tuned GPT-3.5-Turbo-1106 model on the filtered dataset, we generated phytochemically enriched ingredient substitutions. This aimed to enhance recipes with ingredients containing bioactive compounds associated with potential health benefits. We obtained 1951 phytochemically enriched substitutions across 1639 unique recipes generated, featuring at least one phytochemically rich ingredient. As examples, we highlight Watercress Salad (predicted in silico to have potential relevance for COVID-19 mitigation), Kale and Quinoa Salad (associated in silico with AD and COVID-19) and Thai-Style Beef Salad (optimized to target cancer, AD, and COVID-19). Detailed descriptions and analyses of these recipes are provided in Appendix E.

4. Discussion

Our study validated the research hypothesis that LLMs can achieve higher accuracy in ingredient substitution tasks compared to the current state-of-the-art GISMo model when evaluated on a standardized dataset. The fine-tuned GPT-3.5-Turbo-1106 model achieved a Hit@1 accuracy of 54.46% on the filtered Recipe1MSubs dataset, significantly outperforming GISMo's 40.24%. This substantial improvement demonstrates that LLMs have a higher capacity to understand and generate contextually appropriate ingredient substitutions. Building upon this validation, we discuss in more detail the aspects that contributed to the improved performance of the LLMs over GISMo. The following subsections discuss the incorporation of food category features, the impact of dataset filtration based on substitution validity, the fine-tuning process of the LLMs, the generation of phytochemically enriched recipes, and the ethical and economic considerations of our approach.

4.1. Incorporation in GISMo of a Food Category Feature

An initial strategy we explored was the enhancement of the GISMo model through the incorporation of an additional node feature—food categories for each ingredient, classified into one of the 23 categories utilized in FooDB, based on classifications retrieved via the GPT-4. However, contrary to our expectations, this modification did not yield any improvements in the model's performance. This outcome may be attributed to several factors. Firstly, including this additional categorical information might have led to overfitting the model to the training data, compromising its ability to generalize to unseen data in the test set (available in our repository). Additionally, another potential reason could be that part of the value of ingredient categorization might have been indirectly achieved by the model's consideration of ingredient co-occurrence in recipes alongside the presence of flavor molecules. These inherent features within the training data might already provide a basis for the model to make substitution predictions without the need for explicit categorical labels.

4.2. Dataset Filtration Based on Substitution Validity

With the goal of optimizing ingredient substitution, our study introduced an improvement by integrating the capabilities of GPT-3.5 with the GISMo model. While GISMo independently showcased a threefold enhancement in performance compared to prior methods [28], our approach to refine the GISMo model through the preliminary filtration of the original dataset via GPT-3.5's API further increased this improvement. This filtration process involved the exclusion of Potential and Incorrect substitutions from the dataset, thereby ensuring a higher quality of data for model training and application.

The filtration step encompassed five different datasets, and although one was randomly selected to rerun GISMo, the improved results are generalizable across all, due to their almost perfect ingredient substitute similarity across the training, validation, and testing datasets. To demonstrate the consistency of our filtration process, here are examples of substitutions that were consistently classified across the five runs: (A) correct substitutions: orange juice to pineapple juice, carrot to red pepper, black bean to chickpea, basil to dried oregano, onion to shallot; (B) potential substitutions: lemon to orange, apple to peach, apple to apricot, water to wine, blueberry to strawberry; (C) incorrect substitutions:

seedless watermelon to lime, fresh cilantro to ground coriander, horseradish to honey, carrot to seasoning salt, clove to garlic.

4.3. LLM Fine-Tuning for Ingredient Substitution

Using Recipe1MSubs dataset, our experiments explored the benefits of fine-tuning DaVinci, TinyLlama, and GPT-3.5. The first two models did not demonstrate any performance enhancements over the initial method. In contrast, the fine-tuned model leveraging the GPT-3.5 showed a 4% improvement in performance over the GISMo model. Building upon this Recipe1MSubs filtered dataset, we ventured to fine-tune the same three models. Again, the GPT-3.5 model was the only one that showed an increase in performance (20%) when compared with current state of the art.

The findings of this study underscore the importance of data quality and model compatibility in the development of ingredient substitution algorithms. The superior performance achieved through the combination of GPT-3.5's advanced language processing capabilities and the GISMo model's framework highlights the potential of leveraging state-of-the-art AI technologies to refine and enhance existing computational models.

4.4. Phytochemically Enriched Recipe Generation

We specifically selected examples of recipes with ingredients phytochemically enriched targeting COVID-19; COVID-19 and AD; and COVID-19, AD, and cancer molecular networks. Those were Watercress Salad, Kale and Quinoa Salad, and Thai-Style Beef Salad, respectively (Appendix E). We exclusively considered salads in this analysis due to their minimal food processing steps. This choice was made because fewer processing steps generally help preserve the phytochemicals with the health benefits discussed. Salads undergo minimal thermal processing, which helps maintain the integrity of essential nutrients and active compounds compared to more extensively cooked dishes [27].

4.5. Ethical and Economical Considerations

Our research advances computational gastronomy with significant economic and ethical implications, as highlighted in studies on LLMs in food science [29–31]. Economically, LLMs enable cost reduction and innovation through ingredient substitution and recipe optimization, promoting personalized nutrition services and creating new revenue streams for the food industry and healthcare sectors. Additionally, AI-driven personalized recommendation systems, including multimedia food logging and geolocation-based food maps, enhance customer satisfaction and loyalty [30]. Ethically, the deployment of LLMs raises concerns about data biases, misinformation, and privacy, necessitating careful data curation and transparency to ensure fairness and to prevent misleading consumers. Integrating QR code technologies into food labeling further promotes ethical practices by providing transparent detailed product information, thereby enhancing food safety and consumer trust [29]. Balancing these economic benefits with ethical considerations is essential to responsibly harness AI's potential in food science.

4.6. Limitations

One inherent limitation is the diversity of the training datasets used to fine-tune the LLMs. Although these datasets are extensive, they may not fully capture the vast diversity of global cuisines and dietary preferences, potentially impacting the model's ability to generalize across different culinary traditions and suggest culturally and regionally appropriate substitutions. Additionally, the methodology primarily focuses on textual data, which might not capture the full spectrum of culinary contexts, including taste profiles, textures, and the interplay of flavors. LLMs, while proficient in parsing and generating text, have limited capacity to understand and replicate the sensory experiences of cooking and eating.

Additionally, the fine-tuning process, especially when using a limited set of high-quality substitutions, poses a risk of overfitting, where models may become overly specialized to the training data and less capable of generalizing to unseen recipes or ingredients.

Furthermore, the reliance on the Hit@1 metric, while providing a clear measure of the model's ability to suggest the correct first substitution, does not capture the overall utility and flexibility of the model in providing a range of suitable alternatives.

Finally, the computational resources required for fine-tuning and deploying LLMs may also limit the accessibility of these advanced tools to researchers and practitioners with limited resources.

5. Conclusions

By integrating state-of-the-art LLMs such as OpenAI's GPT-3.5 and DaVinci and Meta's TinyLlama, we (1) enhanced the accuracy of ingredient substitution tasks, outperforming the current state-of-the-art GISMo model with an increase in the Hit@1 metric; (2) introduced a novel dataset filtration process using GPT-3.5-Turbo to eliminate less valid ingredient substitutions, leading to higher-quality training data and improved performance of the fine-tuned LLMs; and (3) utilized the best-performing fine-tuned LLM to generate phytochemically enriched ingredient pairings and create unique recipes targeting at least one of the following disease networks—cancer, AD, and COVID-19. As we continue to refine the models and expand our datasets, we anticipate that incorporating domain-specific knowledge, such as clinical and biochemical data, will be crucial for further enhancing the accuracy and relevance of ingredient substitutions. Future research should focus on rigorous validation of these substitutions through clinical trials and controlled dietary studies to assess their efficacy in improving health outcomes. These developments hold promise for revolutionizing personalized nutrition and optimizing dietary practices in a scientifically robust and clinically validated manner.

Author Contributions: Conceptualization, K.V. and I.L.; methodology, K.V., I.L., L.R. and J.S.; investigation, L.R. and J.S.; data curation, L.R.; writing—original draft preparation, L.R., J.S. and K.H.; writing—review and editing L.R., J.S. and K.H.; visualization, L.R. and K.H.; supervision, K.V. and I.L.; project administration, K.V. and I.L.; funding acquisition, K.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundação para a Ciência e a Tecnologia, grant number 2021.05460.BD; ERC-Consolidator, 724228 (LEMAN); ERC Proof of Concept, 899932 (Hyperfoods); UK Research and Innovation, 10058099; European Union, 101095359; and Vodafone Foundation, CORONA-AI/DRUGS DreamLab. We also acknowledge the large citizen science community that made the discovery of phytochemicals possible through the use of the DreamLab app (version 2.6.7).

Data Availability Statement: According to MDPI's commitment to open scientific communication, the fine-tuned models GISMo and TinyLlama, along with the scripts used for filtering the Recipe1MSubs dataset and fine-tuning the GISMo, DaVinci, GPT-3.5, and TinyLlama models, are hosted on the Bitbucket repository. Additionally, data on the categories each ingredient from Recipe1M belongs to, as well as the script used to retrieve these categories, are also made available. The repository can be accessed at the following URL: https://bitbucket.org/iAnalytica/ingredient-substitution (accessed on 12 November 2024). However, due to the use of proprietary models by OpenAI and limitations set by third-party data sources, some restrictions apply to the direct sharing of model weights and specific algorithmic implementations.

Acknowledgments: Figure 1 was created in BioRender (version 2.0.1): https://BioRender.com/f03m948 (accessed on 12 November 2024).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

All ingredients in the Recipe1MSubs dataset were classified into one of the following 23 food categories as identified in the FooDB database [32]. The categories are:

1. Herbs and Spices
2. Fats and Oils
3. Unclassified
4. Baby Foods
5. Snack Foods
6. Dishes
7. Baking Goods
8. Confectioneries
9. Eggs
10. Milk and Milk Products
11. Animal Foods
12. Aquatic Foods
13. Beverages
14. Cocoa and Cocoa Products
15. Soy
16. Coffee and Coffee Products
17. Gourds
18. Teas
19. Pulses
20. Cereals and Cereal Products
21. Nuts
22. Fruits
23. Vegetables

Appendix B

Training configurations for DaVinci-002, GPT-3.5-Turbo-1106, and TinyLlama-1.1B, as well as their filtered dataset variants, are detailed below. This table includes the number of epochs, training steps, and batch sizes used for each model.

Model	Epochs	Steps	Batch Size
DaVinci-002	1	1533	32
DaVinci-002 (Filtered)	1	1554	20
GPT-3.5-Turbo-1106	1	1533	32
GPT-3.5-Turbo-1106 (Filtered)	1	1554	20
TinyLlama-1.1B-1.1B *	1	1532	8
TinyLlama-1.1B-1.1B (filtered) *	1	970	8

(*) Models highlighted had their training parameters manually optimized to enhance performance.

Appendix C

Performance evaluation of the GISMo model, after being trained, validated, and tested on five datasets generated by filtering the original dataset through the GPT-3.5-Turbo-1106. The table below details the Hit@1 accuracy metric, and the number of recipes used in each phase—training, validation, and testing—across all runs.

Filtering Run	Hit@1 (%)	Training	Validation	Testing	Total
Run 1	40.28	31,733	7082	7080	45,895
Run 2	40.82	31,795	7097	7056	45,948
Run 3	40.21	31,797	7083	7096	45,976

Filtering Run	Hit@1 (%)	Training	Validation	Testing	Total
Run 4	39.98	31,908	7073	7113	46,094
Run 5	39.90	31,860	7136	7081	46,077
Final	40.24 ± 0.36	31,819 ± 67	7094 ± 25	7085 ± 21	45,998 ± 85

Appendix D

Comparative analysis of three models—DaVinci-002, GPT-3.5-Turbo-1106, and TinyLlama-1.1B—evaluating their performance on the Recipe1MSubs dataset using the Hit@1 accuracy metric across multiple runs. The analysis includes descriptions of the datasets used.

Dataset Descriptions:

- Original Dataset: 70,520 ingredient substitutions, with 49,044 for training, 10,729 for validation, and 10,747 for testing.
- Filtered Dataset: 44,615 ingredient substitutions, with 31,063 for training, 6831 for validation, and 6721 for testing.

The table below shows the Hit@1 accuracy for each model across five runs, along with the final average and standard deviation. The models are ordered in the table from best to worst performance based on their Hit@1 accuracy.

Recipe1MSubs Dataset	Fine-Tuned Model	Run 1	Run 2	Run 3	Run 4	Run 5	Final Hit@1 (%)
Filtered	**GPT-3.5-Turbo-1106**	**54.05**	**54.77**	**54.69**	**54.40**	**54.37**	**54.46 ± 0.29**
	GISMo	40.28	40.82	40.21	39.98	39.90	40.24 ± 0.36
	TinyLlama-1.1B	35.07	34.22	34.46	34.53	34.37	34.53 ± 0.32
	DaVinci-002	28.97	29.59	29.61	29.27	29.70	29.43 ± 0.30
Unfiltered	**GPT-3.5-Turbo-1106**	**38.08**	**38.25**	**37.96**	**38.28**	**37.59**	**38.03 ± 0.28**
	GISMo	34.55	34.42	34.68	34.54	34.45	34.53 ± 0.10
	TinyLlama-1.1B	20.44	20.78	20.35	20.18	20.16	20.38 ± 0.25
	DaVinci-002	20.39	19.73	19.77	20.29	20.26	20.09 ± 0.31

The best performer fine-tuned models with the original and filtered datasets are in bold.

Appendix E

Three examples of recipes showcasing substitutions that account for the dish's flavor profile but also its nutritional value using phytochemically enriched ingredients:

- Thai Style Beef Salad: This recipe includes substitutions such as replacing mung bean sprouts with cabbage to increase the recipe's content of glucosinolates, known for their cancer-preventive properties. The use of olive oil instead of sesame oil increases the content of healthy fats and antioxidants, supporting cognitive health and cardiovascular health.
- Super Corn Salad: substitutions in this recipe include using olive oil instead of vegetable oil to provide healthier monounsaturated fats and antioxidants. Carrots replace pepper to increase the beta-carotene content, beneficial for immune function, and dill is used instead of tarragon, providing a different set of phytonutrients beneficial for inflammation reduction.
- Pineapple-Cabbage Salad: In this recipe, radishes are substituted with carrots to increase the beta-carotene content, and peas are replaced with more carrots to further enhance the dish's vitamin A content, which is crucial for immune system function and vision.

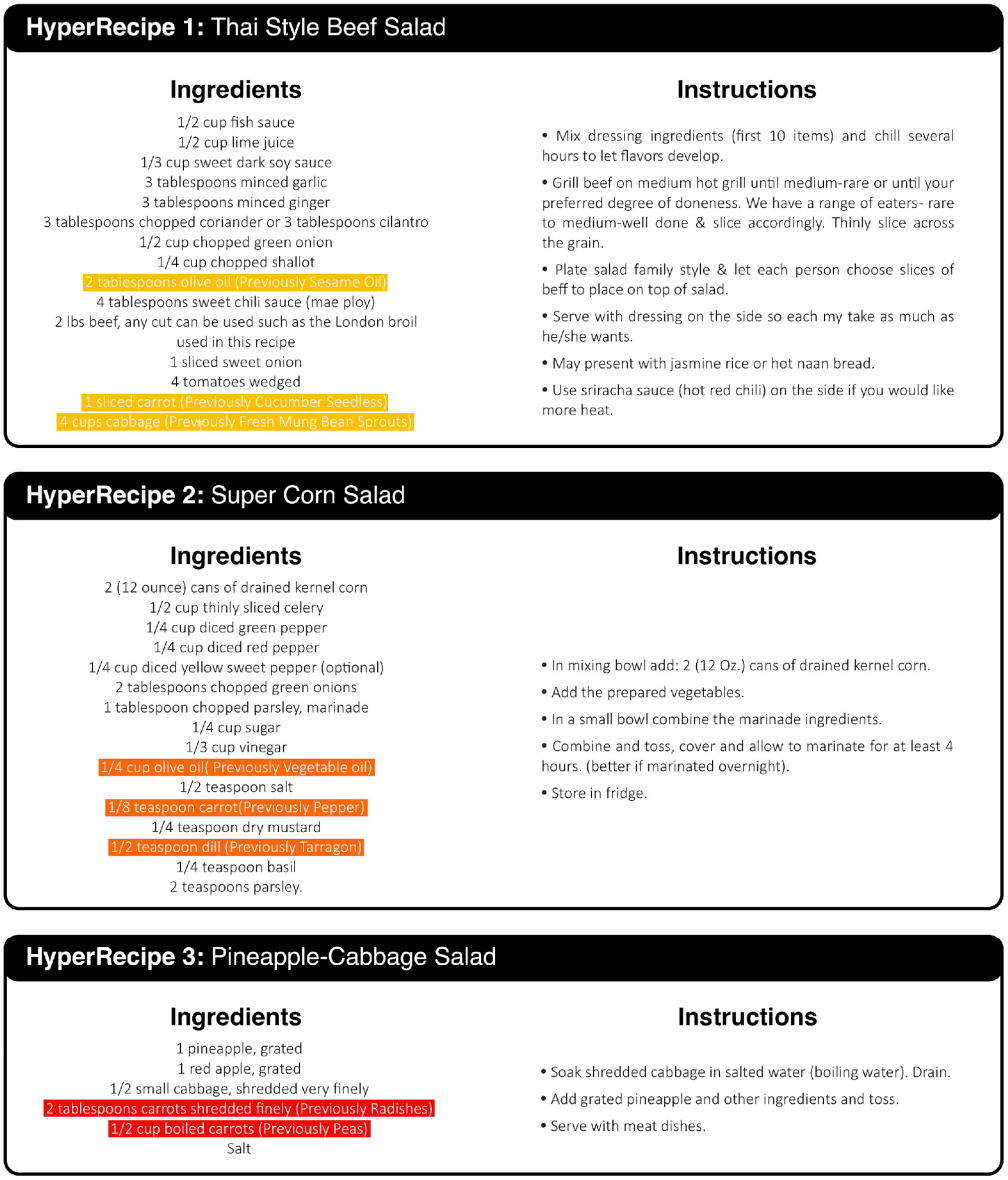

Figure A1. Phytochemically enriched recipes. Three examples showing the old and substituted ingredients considering the whole context of each recipe. Yellow, orange and red highlights emphasize the substituted ingredients in each of the three recipes.

References

1. Veselkov, K.; Gonzalez, G.; Aljifri, S.; Galea, D.; Mirnezami, R.; Youssef, J.; Bronstein, M.; Laponogov, I. HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. *Sci. Rep.* **2019**, *9*, 9237. [CrossRef] [PubMed]
2. Rita, L.; Neumann, N.R.; Laponogov, I.; Gonzalez, G.; Veselkov, D.; Pratico, D.; Aalizadeh, R.; Thomaidis, N.S.; Thompson, D.C.; Vasiliou, V.; et al. Alzheimer's disease: Using gene/protein network machine learning for molecule discovery in olive oil. *Hum. Genom.* **2023**, *17*, 57. [CrossRef] [PubMed]
3. Laponogov, I.; Gonzalez, G.; Shepherd, M.; Qureshi, A.; Veselkov, D.; Charkoftaki, G.; Li, K.; Zhao, L.; Su, L. Network machine learning maps phytochemically rich "Hyperfoods" to fight COVID-19. *Hum. Genom.* **2021**, *15*, 1. [CrossRef] [PubMed]
4. Ma, F.; An, Z.; Yue, Q.; Zhao, C.; Zhang, S.; Sun, X.; Li, K.; Zhao, L.; Su, L. Effects of brassinosteroids on cancer cells: A review. *J. Biochem. Mol. Toxicol.* **2022**, *36*, e23026. [CrossRef] [PubMed]
5. Elreedy, H.A.; Elfiky, A.M.; Mahmoud, A.A.; Ibrahim, K.S.; Ghazy, M.A. Neuroprotective effect of quercetin through targeting key genes involved in aluminum chloride induced Alzheimer's disease in rats. *Egypt. J. Basic Appl. Sci.* **2023**, *10*, 174–184. [CrossRef]
6. Liu, J.; Zhang, L.; Gao, J.; Zhang, B.; Liu, X.; Yang, N.; Liu, X.; Liu, X.; Cheng, Y. Discovery of genistein derivatives as potential SARS-CoV-2 main protease inhibitors by virtual screening, molecular dynamics simulations and ADMET analysis. *Front. Pharmacol.* **2022**, *13*, 961154. [CrossRef]

7. Shirai, S.S.; Seneviratne, O.; Gordon, M.E.; Chen, C.-H.; McGuinness, D.L. Identifying Ingredient Substitutions Using a Knowledge Graph of Food. *Front. Artif. Intell.* **2021**, *3*, 621766. [CrossRef]
8. Yamanishi, R.; Shino, N.; Nishihara, Y.; Fukumoto, J.; Kaizaki, A. Alternative-ingredient Recommendation Based on Co-occurrence Relation on Recipe Database. *Procedia Comput. Sci.* **2015**, *60*, 986–993. [CrossRef]
9. Boscarino, C.; Nedović, V.; Koenderink, N.J.J.P.; Top, J.L. Automatic extraction of ingredient's substitutes. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, New York, NY, USA, 13–17 September 2014; pp. 559–564.
10. Ahn, Y.-Y.; Ahnert, S.E.; Bagrow, J.P.; Barabási, A.-L. Flavor network and the principles of food pairing. *Sci. Rep.* **2011**, *1*, 196. [CrossRef]
11. Achananuparp, P.; Weber, I. Extracting Food Substitutes From Food Diary via Distributional Similarity. *arXiv* **2016**, arXiv:1607.08807.
12. Kazama, M.; Sugimoto, M.; Hosokawa, C.; Matsushima, K.; Varshney, L.R.; Ishikawa, Y. A neural network system for transformation of regional cuisine style. *arXiv* **2017**, arXiv:1705.03487. [CrossRef]
13. Lawo, D.; Böhm, L.; Stevens, G. *Veganaizer: AI-Assisted Ingredient Substitution*; University of Siegen: Siegen, Germany, 2020.
14. Morales-Garzon, A.; Gomez-Romero, J.; Martin-Bautista, M.J. A Word Embedding-Based Method for Unsupervised Adaptation of Cooking Recipes. *IEEE Access* **2021**, *9*, 27389–27404. [CrossRef]
15. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
16. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
17. Wu, S.; He, Y. Enriching Pre-trained Language Model with Entity Information for Relation Classification. *arXiv* **2019**, arXiv:1905.08284.
18. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *arXiv* **2019**, arXiv:1901.00596. [CrossRef]
19. Park, D.; Kim, K.; Kim, S.; Spranger, M.; Kang, J. FlavorGraph: A large-scale food-chemical graph for generating food representations and recommending food pairings. *Sci. Rep.* **2021**, *11*, 931. [CrossRef]
20. Fatemi, B.; Duval, Q.; Girdhar, R.; Drozdzal, M.; Romero-Soriano, A. Learning to Substitute Ingredients in Recipes. *arXiv* **2023**, arXiv:2302.07960.
21. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
22. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, L. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
24. Zhang, P.; Zeng, G.; Wang, T.; Lu, W. TinyLlama: An Open-Source Small Language Model. *arXiv* **2024**, arXiv:2401.02385.
25. Lample, G.; Conneau, A. Cross-lingual Language Model Pretraining. *arXiv* **2019**, arXiv:1901.07291.
26. Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Ofli, F.; Weber, I.; Torralba, A. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3068–3076.
27. Nicoli, M.C.; Anese, M.; Parpinel, M. Influence of processing on the antioxidant properties of fruit and vegetables. *Trends Food Sci. Technol.* **1999**, *10*, 94–100. [CrossRef]
28. Hasan, I.; Hossain, A.; Rahman, H.; Sohel; Ahsan, A.; Soikot, S.H.; Islam, N.; Amin, M.R.; Jain, D.K. Galangin for COVID-19 and Mucormycosis co-infection: A potential therapeutic strategy of targeting critical host signal pathways triggered by SARS-CoV-2 and Mucormycosis. *Netw. Model. Anal. Health Inform. Bioinform.* **2023**, *12*, 26. [CrossRef]
29. Dospinescu, O. The Use of Information Technology Toward the Ethics of Food Safety. *Ecoforum. J.* **2018**, *7*, 70.
30. Rostami, A. *An Integrated Framework for Contextual Personalized LLM-Based Food Recommendation*; University of California: Irvine, CA, USA, 2024.
31. Ma, P.; Tsai, S.; He, Y.; Jia, X.; Zhen, D.; Yu, N.; Wang, Q.; Ahuja, J.K.C.; Wei, C.-I. Large language models in food science: Innovations, applications, and future. *Trends Food Sci. Technol.* **2024**, *148*, 104488. [CrossRef]
32. Harrington, R.A.; Adhikari, V.; Rayner, M.; Scarborough, P. Nutrient composition databases in the age of big data: FoodDB, a comprehensive, real-time database infrastructure. *BMJ Open* **2019**, *9*, e026652. [CrossRef]