

<https://doi.org/10.1038/s41538-025-00680-9>

A unified knowledge graph linking foodomics to chemical-disease networks and flavor profiles



Fangzhou Li^{1,2,3}, Jason Youn^{1,2,3}, Kaichi Xie^{1,3}, Trevor Chan^{1,2,3}, Pranav Gupta^{1,2,3}, Arielle Yoo^{2,3}, Michael Gunning^{1,2,3}, Keer Ni^{1,2,3} & Ilias Tagkopoulos^{1,2,3} ✉

Modern nutrition science still lacks a comprehensive, machine-readable map linking diet to molecular composition and biological effects. Here we present FoodAtlas, a large-scale knowledge graph that links 1430 foods to 3610 chemicals, 2181 diseases, and 958 flavor descriptors through 96,981 provenance-tracked edges. A transformer-based text-mining pipeline extracted 48,474 quantitative food–chemical associations from 125,723 literature sentences ($F_1 = 0.67$) and integrated them with 23,211 chemical–disease assertions from the Comparative Toxicogenomics Database, 15,222 chemical–bioactivity records from ChEMBL, 3645 flavor annotations from FlavorDB and PubChem, and 6429 taxonomic relationships. Graph embeddings revealed six dietary modules whose signature metabolites delineate distinct, multisystem disease-risk trajectories. Models built on FoodAtlas demonstrate practical utility: a bioactivity predictor achieved strong correlation with antioxidant assays ($R^2 = 0.52$; $\rho = 0.72$), and a substitution engine reduced simulated total disease risk by 11.9%.

The interplay between food, chemicals, and health represents a cornerstone of research in nutrition, toxicology, and sensory science. Understanding these complex relationships is critical for addressing pressing global challenges, such as improving public health through personalized nutrition, developing innovative food products, and assessing the safety of dietary components. However, existing resources often lack the granularity, comprehensiveness, or integration necessary to fully explore these domains¹. FoodAtlas² was developed to address these limitations by constructing a knowledge graph (KG) that combines data from diverse sources, enabling structured queries and advanced analytics.

Recent advances in the study of the interplay among foods, chemicals, and health have shifted the focus from isolated nutrient databases to integrated, data-rich platforms that harness multiple omics and artificial intelligence technologies^{3,4}. Early food composition resources, such as FoodDB⁵ and USDA FoodData Central⁶, provided detailed chemical and nutritional profiles, but often operated as standalone silos⁷. The emergence of foodomics, an approach that integrates food chemistry with genomics, proteomics, and metabolomics, has further enriched our understanding of bioactive compounds and their health impacts⁸. Concurrently, the application of knowledge graphs (KGs) in food science has enabled researchers to interlink heterogeneous datasets, thereby facilitating novel applications such as recipe development, diet–disease correlation discovery, and personalized nutrition recommendation^{9,2}.

Despite these promising developments, existing approaches exhibit notable limitations. Most traditional food composition tables (FCTs) and databases like USDA's FoodData Central or FoodDB focus on macronutrients and certain micronutrients, neglecting many bioactive phytochemicals that have potentially significant health effects^{7,10}. Further, ontological resources such as FoodOn¹¹ provide valuable standardization of food entities, but do not capture granular relationships (e.g., quantitative concentration data or exact part of relationships for foods). As a result, researchers often rely on labor-intensive data integration approaches that do not scale easily¹². Moreover, many existing KGs fail to incorporate advanced natural language processing (NLP) methods, which can unlock vast amounts of untapped data from the literature¹³. Traditional text-mining methods also face challenges in extracting numerical information (e.g., specific concentrations, complex units), reducing the ability to interpret or apply the data meaningfully.

Recent advances in transformer-based NLP models have enabled more accurate automated extraction of relationships from textual sources^{14,15}. Models like BioBERT¹⁶ have dramatically improved the automated extraction of food–chemical relationships from vast bodies of biomedical literature. For example, FoodChem¹⁷ employed these techniques to identify “contains” relations with high accuracy, automating the extraction of thousands of food–chemical interactions. Moreover, large language models can achieve high accuracy in extracting chemical food safety hazards from scientific abstracts, often exceeding 90% without extensive additional

¹Department of Computer Science, the University of California at Davis, Davis, CA, USA. ²Genome Center, the University of California at Davis, Davis, CA, USA.

³USDA/NSF AI Institute for Next Generation Food Systems (AIFS), Davis, CA, USA. ✉e-mail: itagkopoulos@ucdavis.edu

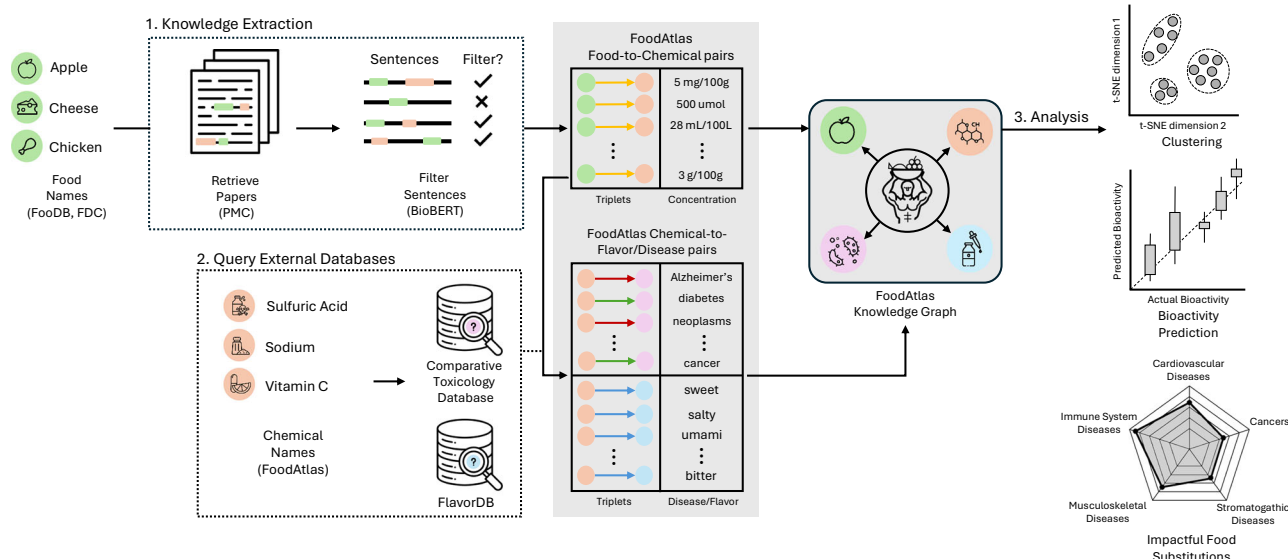


Fig. 1 | FoodAtlas pipeline for building and analyzing a multi-layer food-chemical knowledge graph. Food names from FooDB and USDA FoodData Central are used to retrieve full-text articles in PubMed Central; BioBERT filters comment sentences, which are converted into quantitative Food → Chemical triplets. Chemicals already cataloged in FoodAtlas are then linked to the Comparative Toxicogenomics Database to add Chemical → Disease edges and to FlavorDB for

Chemical → Flavor edges, yielding an integrated graph that unifies compositional, biomedical, and sensory information. This enriched FoodAtlas Knowledge Graph underpins downstream tasks—including t-SNE clustering of foods, supervised bioactivity prediction, and disease-focused food substitutions—enabling data-driven nutrition and precision-diet applications.

training¹⁸. These pipelines facilitate integration of data from external sources like the Comparative Toxicogenomics Database (CTD)¹⁹ and flavor repositories^{20,21}, enabling the KG to also encompass health and sensory attributes. However, a gap remains in terms of unifying these diverse data types, such as bioactivity and disease associations, flavor attributes, and chemical concentrations, into a single, consistent knowledge resource that can be readily accessed and queried.

Beyond these general limitations, several food-focused knowledge resources have emerged, but each covers only parts of the domain. FoodKG²² integrates recipes, ingredients, and nutrition, FlavorGraph²³ links ingredients with flavor compounds and recipe co-occurrence, and NutriChem²⁴ connects plant-based foods with phytochemicals and diseases via text-mining. Supplementary Table 1 summarizes how FoodAtlas compares with these resources in scale and scope. Notably, FoodAtlas is the first to unify food composition, chemical-bioactivity, and disease associations in a single graph, achieving broader entity coverage and relationship types than any prior resource.

In this work, we present a new framework to support evidence synthesis that we incorporate into FoodAtlas², which includes large language models to extract food knowledge from scientific literature, capable of extracting diverse metadata, such as concentration values. To support better query and synthesis capabilities, we have doubled the number of resources (9 sources total), incorporated 5 ontologies, and expanded relationships to sensory and disease attributes (Fig. 1). First, we detail the updated information extraction pipeline, which leverages state-of-the-art large language models. Then, we apply unsupervised machine learning to the FoodAtlas data, demonstrating that the learned representations form meaningful clusters related to food and health. Finally, we demonstrate potential real-world use cases of FoodAtlas by showcasing two relevant health applications. Specifically, a bioactivity predictor model and a simple diet recommendation supported by FoodAtlas data result in predictive results corroborated by published literature.

Results

FoodAtlas incorporates data from a variety of different sources

FoodAtlas markedly broadens both the breadth and resolution of the knowledge graph, interlinking 1430 foods with 3610 chemicals through

48,474 curated food-to-chemical edges (Fig. 2). A major feature is the integration of 23,211 chemical-to-disease assertions, 13,417 “treats” and 9794 “worsens” relationships, covering 2181 diseases extracted from the CTD. Each chemical node is normalized to its MeSH identifier, and every edge carries CTD provenance metadata, Direct Evidence plus PubMed citations, creating a transparent scaffold for toxicological and biomedical inquiry. FoodAtlas also incorporates the chemical-to-bioactivity linking all the chemicals contained in at least one food to 8 key bioactivities derived from ChEMBL, alongside 660 direct food-to-bioactivity antioxidant FRAP values. FoodAtlas also introduces a sensory layer: 3645 chemical-to-flavor links connecting 958 unique flavor descriptors compiled from FlavorDB and PubChem.

Large language model-based information extraction accurately extracts chemical composition

The test set consists of 356 sentences and was used to evaluate the performance of the OpenAI GPT-4, GPT-3.5, and fine-tuned GPT-3.5 models. We applied both a zero-shot prompt, containing only extraction instructions, and a one-shot prompt, containing one example along with the instructions to the pre-trained GPT-4 and GPT-3.5 models (Supplementary Note 1.1). Additionally, we fine-tuned the GPT-3.5 model using the zero-shot prompt, achieving a significantly higher F₁-score of 0.67 compared to 0.42 for GPT-4 with the one-shot prompt (Supplementary Figs. 2, 3).

FoodAtlas knowledge graph encoded meaningful food representation in chemical and health spaces

Projection of the food-by-chemical matrix into t-SNE²⁵ space of the raw composition matrix confirms that the metabolites driving disease divergence are the dominant axes of compositional variance (Fig. 3A). Following this, a density-based clustering²⁶ and a two-part hurdle test²⁷, chemical enrichment (q_{enrich}), and disease intensity ($q_{\text{intensity}}$), each with a false discovery rate²⁸ smaller than 0.10, resolved six significant disease composition clusters (Fig. 3B, Table 1, and Supplementary Note 1.4). Collectively, these six clusters reveal chemical composition patterns that correlate with divergent disease-risk trajectories, highlighting potential mechanisms linking dietary components to health outcomes. The *omega-3 marine oils* cluster contains oily fish and krill products, which are extremely enriched in

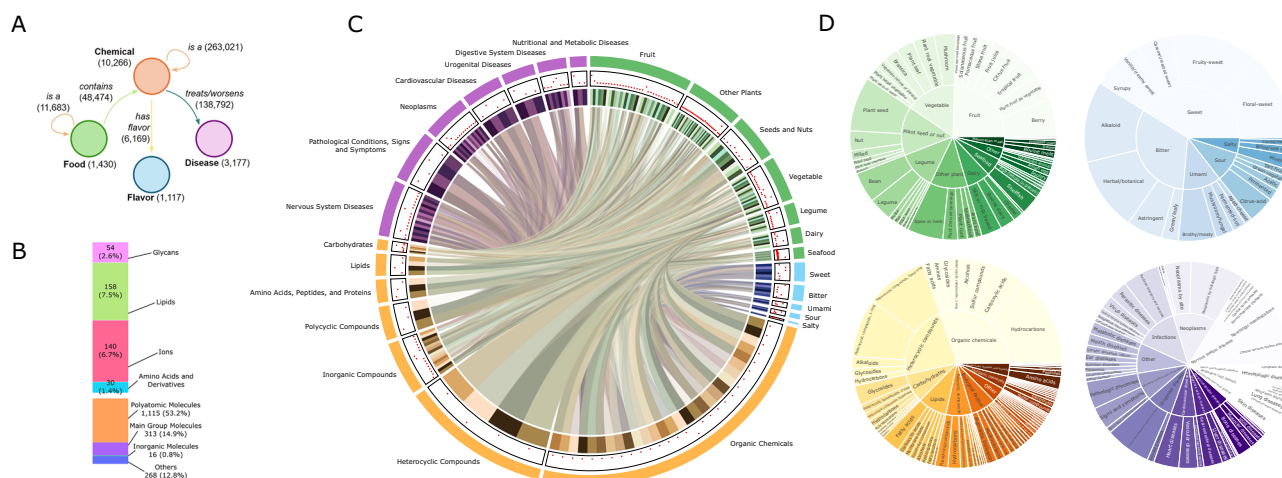


Fig. 2 | Composition of the FoodAtlas Knowledge Graph. **A** Entity-relation schema with node counts (in parentheses) and edge counts on the arrows: foods (1,430) connect to chemicals (10,266) through *contains* edges (48,474); chemicals link hierarchically via *is a* (263,021); chemicals map to diseases (3,177) through *treats/worsens* relations (138,792); chemicals map to flavors (1,117) through *has flavor* (6,169); foods are organized taxonomically by additional *is a* edges (11,683). **B** Chemical superclass breakdown shows that polyatomic molecules dominate (1,115; 53.2%), followed by main group molecules (313; 14.9%), lipids (158; 7.5%), ions (140; 6.7%), glycans (54; 2.6%), amino acid derivatives (30; 1.4%), inorganic

molecules (16; 0.8%) and other classes (268; 12.8%). **C** Integrated circos plot depicting the ecosystem of food-chemical-disease-flavor relationships across all entities in FoodAtlas. The outermost ring shows the category arcs. Inside this, a narrow white ring hosts a scatter-style evidence lane: red radial bars plot, for every sub-category, the volume of supporting metadata (bar height scales with evidence count). The innermost colored band comprises the sub-category arcs. **D** Sunburst plots showing hierarchical distributions of flavor descriptors and disease outcomes. Each sunburst represents the relative frequency of (outer ring) subcategories nested within (inner ring) major categories.

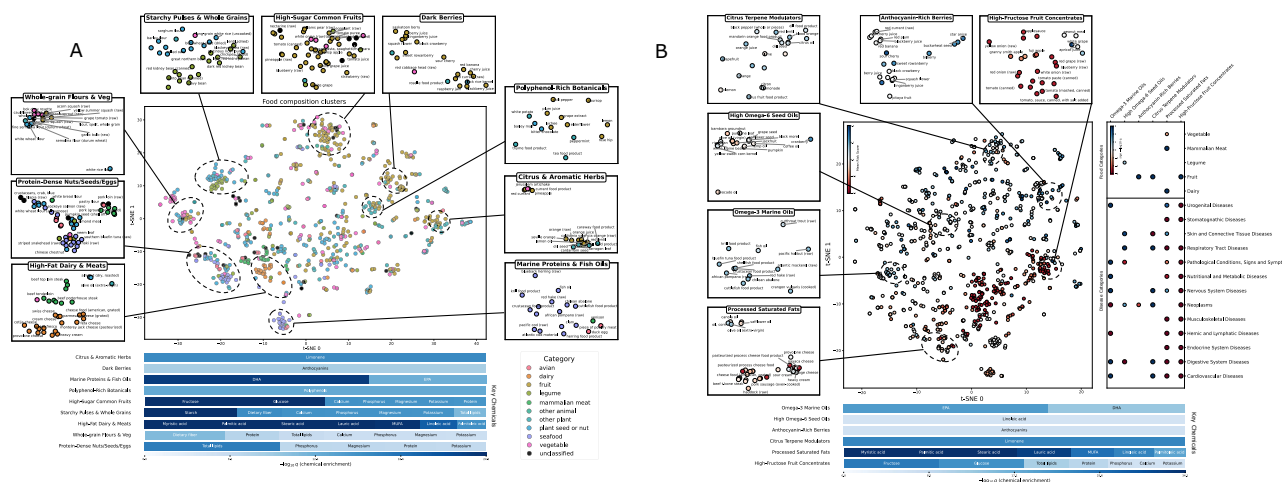


Fig. 3 | Concordant chemical and disease structure in FoodAtlas foods. **A** Composition space. A Node2Vec¹⁰⁵ embedding built only from food-to-chemical edges was projected with t-SNE (center plot). Density-based clustering (dashed ellipses) resolves nine chemically coherent groups whose insets (top and side panels) name representative members. The horizontal bar-stripe beneath the scatter ranks the chemicals that are significantly enriched within each group (signed- $\log_{10} q$ from the two-part hurdle test; darker blue = stronger enrichment). **B** Disease-aware space. Adding signed chemical-to-disease edges to the graph and repeating the embedding yields a second t-SNE in which node color encodes the composite, normalized

disease score (blue = protective, red = harmful). Six clusters that survive the dual chemical-plus-disease hurdle test are emphasized (dashed ellipses, labels). The dot-matrix to the right compares, for those clusters, (i) over-represented food categories (top sub-panel) and (ii) disease classes that differ significantly from the background (bottom; dot size $\propto -\log_{10} q$, color sign = protection or risk). The lower bar-stripe again lists the hallmark enriched chemicals, showing that the same metabolites—EPA/DHA, anthocyanins, limonene, palmitic acid, fructose/glucose, linoleic acid—drive separation in both purely compositional and health-augmented landscapes.

EPA ($q_{\text{enrich}} = 1.3 \times 10^{-20}$) and DHA ($q_{\text{enrich}} = 4.9 \times 10^{-14}$). This group shows the dataset's strongest protection, lowering cardiovascular risk ($q_{\text{intensity}} = 1.2 \times 10^{-7}$) and improving metabolic, digestive, and nervous-system scores (all $q_{\text{intensity}} < 10^{-2}$). The *high omega-6 seed oils* cluster, which includes sunflower, safflower, and rice-bran oils, shares significant linoleic acid enrichment ($q_{\text{enrich}} = 4.0 \times 10^{-7}$) but exhibits a modest yet significant rise in the composite "pathological-conditions" score ($q_{\text{intensity}} = 1.0 \times 10^{-3}$), echoing reports that excessive omega-6 intake can sustain low-grade inflammation²⁹. The *anthocyanin-rich berries*, such as blueberry, bilberry,

and black-currant, are highly enriched for total anthocyanins ($q_{\text{enrich}} = 2.3 \times 10^{-5}$) and deliver broad protection, most prominently against cardiovascular disease ($q_{\text{intensity}} = 1.0 \times 10^{-3}$) with auxiliary benefits in respiratory and metabolic domains ($q_{\text{intensity}} = 5.0 \times 10^{-2}$). The *citrus-terpene modulators* cluster, consisting of oranges, lemons, and limes, is dominated by limonene ($q_{\text{enrich}} = 9.2 \times 10^{-24}$), is linked to lower neoplasm scores ($q_{\text{intensity}} = 1.0 \times 10^{-6}$) yet higher skin/connective-tissue scores ($q_{\text{intensity}} = 1.0 \times 10^{-10}$), mirroring limonene's chemo-preventive activity alongside its well-known phototoxicity³⁰. The *fat-dense animal proteins*,

Table 1 | Chemistry and disease signatures of supported clusters

Cluster (ID)	Representative foods	Key chemical(s)	Leading disease effect	Supporting references
Omega-3 marine oils	Salmon, mackerel, krill oil	EPA (1.3×10^{-20}); DHA (4.9×10^{-14})	Cardiovascular score decreased (1.2×10^{-7})	69,70
High-LA seed oils	Sunflower, safflower oils	Linoleic acid (4.3×10^{-7})	Pathological-conditions score increased (3.5×10^{-9})	29,71
Anthocyanin berries	Blueberry, bilberry	Total anthocyanins (1.4×10^{-7})	Cardiovascular score decreased (3.1×10^{-9})	72,73
Citrus-terpene foods	Orange, lemon, lime	Limonene (9.3×10^{-24})	Neoplasm score decreased (6.6×10^{-6}); Skin/connective-tissue score increased (1.2×10^{-10})	30,74
Processed animal fats	Bacon, hard cheese	Palmitic acid (7.5×10^{-35}); Trans-FA ($\leq 2.0 \times 10^{-15}$)	Stomatognathic score increased (5.2×10^{-9})	75,76
Free-sugar fruits	Grape, apple juices	Fructose (3.5×10^{-25}); Glucose (3.6×10^{-22})	Nervous score increased (6.3×10^{-7}); Urogenital score increased (9.1×10^{-7})	77,78

Table showing the key chemicals and disease impact of each cluster, with supporting references.

including bacon, sausages, and hard cheeses, concentrate palmitic acid ($q_{\text{enrich}} = 4.0 \times 10^{-35}$) and industrial trans-fatty acids ($q_{\text{enrich}} = 2.3 \times 10^{-14}$). The cluster is associated with worsening stomatognathic disease ($q_{\text{intensity}} = 1.0 \times 10^{-3}$) and measurable endocrine and cardiovascular penalties ($q_{\text{intensity}} = 5.0 \times 10^{-2}$). Finally, the *high-fructose fruit concentrates* cluster, which involves grape and apple juices and high-Brix tomato products, shows extreme fructose ($q_{\text{enrich}} = 2.6 \times 10^{-22}$) and glucose ($q_{\text{enrich}} = 2.1 \times 10^{-19}$) enrichment, translating into increased nervous-system ($q_{\text{intensity}} = 1.0 \times 10^{-7}$) and urogenital ($q_{\text{intensity}} = 1.0 \times 10^{-7}$) disease scores and broader metabolic detriments ($q_{\text{intensity}} = 1.0 \times 10^{-2}$).

By integrating chemical composition, structural fingerprints, and potency data, we accurately predict food-level antioxidant capacity

The integration of FoodAtlas composition data with ChEMBL potency measurements culminated in a BFL set that maps 15,222 chemicals across 660 foods to quantitative bioactivity readouts via pChEMBL values. A preliminary attempt that relied on pChEMBL values alone was uninformative (Supplementary Note 1.5). In contrast, the ensuing BPM, which jointly leverages chemical concentrations, extended-connectivity fingerprints, and potency scores, captured 52% of the variance in literature FRAP measurements and achieved a Pearson correlation coefficient (PCC) of 0.72 ($p = 3.9 \times 10^{-14}$); three quarters of predictions deviated by $\pm 0.81 -\log_{10}$ (FRAP) units (Fig. 4A). Residual analysis confirmed stable model performance, showing a negligible correlation ($r = 0.039$) between predicted values and model residuals. Beyond reproducing reported values, BPM generalizes to previously uncharacterized items: Fig. 4B ranks the 50 common foods per their predicted antioxidant content, illustrating the model's utility for rapid triage of antioxidant potential in foods lacking experimental FRAP data.

Holistic food substitutions achieve significant disease-risk reduction and antioxidant gains

We modeled 14,580 disease-focused substitutions and 7798 antioxidant-focused substitutions (Fig. 5A and Supplementary Table 5). Across the full dataset, the mean disease prevention score rose by 11.9%. Improvements were equally weighted across all meals - breakfast (9.9%), lunch (10.0%), and dinner (11.1%) meals (Fig. 5B). All reductions remained significant after correction for multiple testing ($q < 10^{-3}$). Applying the same substitution framework to antioxidant bioactivity more than tripled predicted activity, yielding an average increase of 210.6%, with peak enhancements at lunch (257.9%), dinner (185.8%), and breakfast (176.4%). We first quantified the aggregate change in disease-risk score that resulted from category-matched food swaps (Fig. 5A). Radar plots of individual disease domains confirmed broad-spectrum improvements rather than benefits confined to a single system (Fig. 5B). To dissect the chemical basis of these effects, we extracted the ten most influential chemicals for each meal type and visualized both

their absolute contributions to risk and the direction of change induced by the swap (Fig. 5C, left versus right sub-panels). Protective shifts were consistently driven by polyphenolic flavonoids, including quercetin, kaempferol, and cyanidin, and long-chain omega-3 fatty acids (EPA, DHA), which exhibited uniformly negative differential scores (cool hues). Conversely, heterocyclic amines such as PhIP and MeIQx, advanced glycation end-products (CML, CEL), and saturated tri-acyl-glycerols (palmitic-, stearic-, and lauric-TAGs) remained dominant positive contributors (warm hues). The recurrence of these chemical signatures across breakfast, lunch, and dinner indicates that the underlying mechanisms are conserved and time-of-day-independent. Three-layer Sankey diagrams traced the pathways from food categories through chemical classes to disease domains, revealing how dietary risk propagates through the food system (Fig. 5D). By measuring how much the total chemical content of meals changed after substitutions and testing whether these changes were statistically meaningful, we identified the most robust pathways. Equalized node heights decoupled topological structure from magnitude, making “risk bottlenecks” visually salient. Processed-meat and refined-grain categories converged through heterocyclic amines and dicarbonyl AGEs onto neoplasm nodes, where food substitutions significantly increased cancer-related chemical content ($q = 1.8 \times 10^{-56}$ for breakfast; $q = 5.1 \times 10^{-77}$ for lunch; $q = 1.1 \times 10^{-68}$ for dinner), cardiovascular nodes, where substitutions increased heart disease risk ($q = 2.2 \times 10^{-5}$ for breakfast; $q = 8.5 \times 10^{-19}$ for lunch; $q = 7.0 \times 10^{-4}$ for dinner), and endocrine-metabolic nodes, where substitutions disrupted hormonal systems ($q = 8.7 \times 10^{-4}$ for breakfast; $q = 2.2 \times 10^{-35}$ for lunch; $q = 2.9 \times 10^{-30}$ for dinner). Conversely, berries, leafy greens, and olive products channeled flavonoids and monophenols towards protective endocrine-metabolic nodes ($q = 8.7 \times 10^{-4}$ for breakfast; $q = 2.2 \times 10^{-35}$ for lunch; $q = 2.9 \times 10^{-30}$ for dinner) and immune nodes ($q = 8.7 \times 10^{-4}$ for breakfast; $q = 5.1 \times 10^{-37}$ for lunch; $q = 1.3 \times 10^{-31}$) where substitutions significantly reduced disease risk (Tables 2 and 3).

Discussion

In this study, we introduce FoodAtlas, a substantially expanded and fully traceable knowledge graph that integrates compositional, toxicological, flavor, and bioactivity data for advanced dietary analysis. By linking 1430 foods to 3610 chemicals through 48,474 curated food-chemical edges, incorporating 23,211 chemical-disease assertions across 2181 diseases (13,417 ameliorative, 9794 exacerbating), 15,222 chemical-bioactivity assertions across eight bioactivities, 660 food-antioxidant bioactivity, and layering in 3645 chemical-flavor relationships covering 958 descriptors, FoodAtlas represents a major leap in both breadth and resolution over its predecessor. Our ontology-aware, transformer-driven ingestion pipeline achieved an F_1 of 0.67 on held-out extraction tasks, enabling reliable scaling of literature-derived composition facts. The Bioactivity Prediction Model further demonstrates generalizability, explaining 52% of variance in FRAP assays ($R^2 = 0.52$, $\text{PCC} = 0.72$) and ranking antioxidant potential for

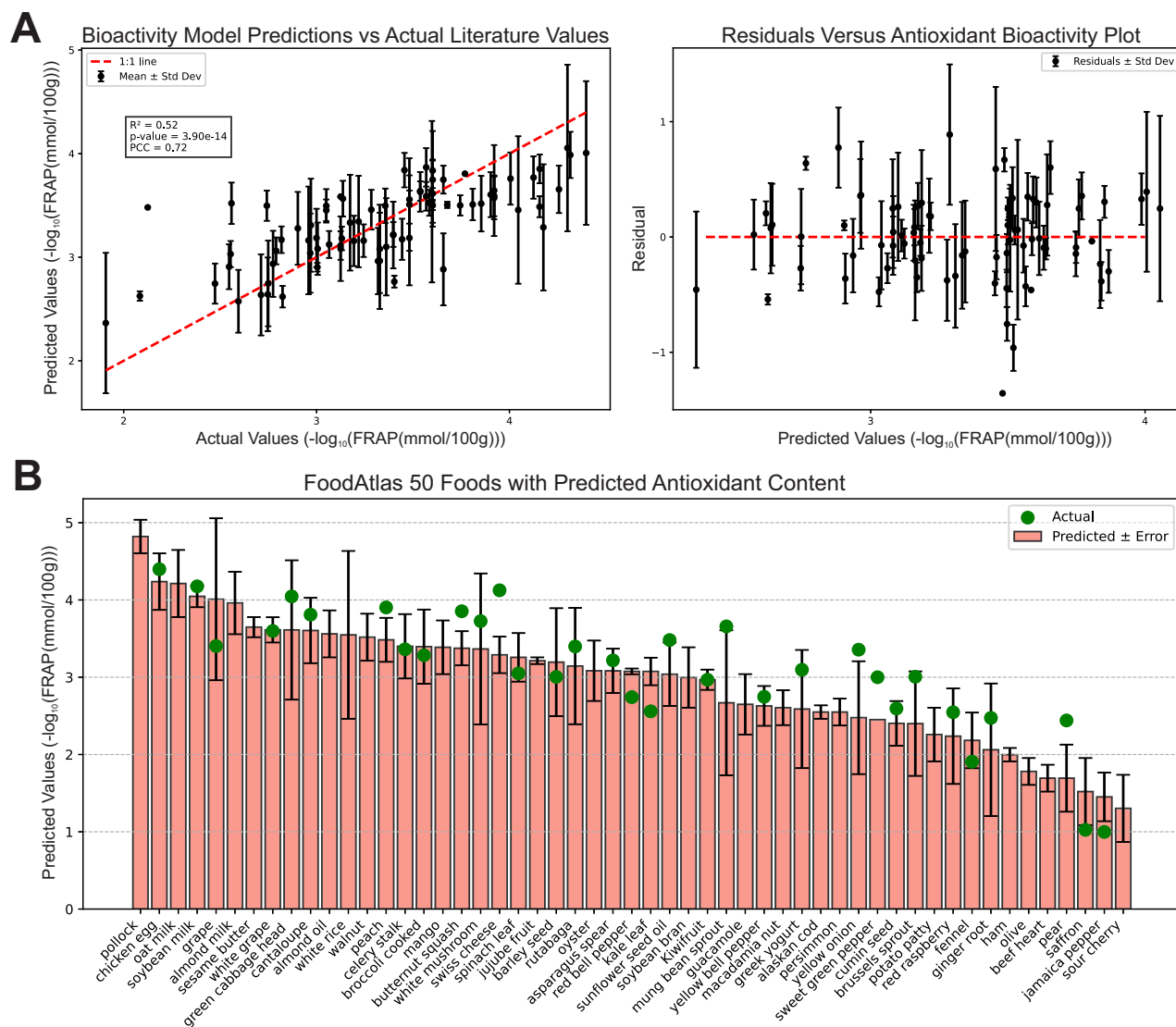


Fig. 4 | Predictive performance of the antioxidant Bioactivity Prediction Model (BPM). **A** Goodness of fit and diagnostics. The left panel contrasts BPM-predicted versus literature-reported antioxidant activity (negative log-transformed FRAP values) for foods with available measurements. The model attains an $R^2 = 0.52$ and a Pearson correlation coefficient (PCC) = 0.72. The right panel plots studentized residuals against observed antioxidant values, revealing homoscedastic error

dispersion around zero and no systematic bias across the response range.

B Extrapolative ranking of unmeasured foods. The bar chart lists the top-50 FoodAtlas foods by predicted antioxidant content (pink bars $\pm$ prediction error). Green circles indicate foods for which experimental FRAP data exist, showing close concordance with model estimates and underscoring the BPM's ability to prioritize candidate foods that currently lack empirical measurements.

previously uncharacterized foods. We used antioxidant activity as a benchmark endpoint because it is one of the most widely studied and quantitatively validated measures in food and nutrition research³¹, or for evaluating bioactive compounds' health benefits³². The BPM's reliance on chemical structure, potency, and concentration data illustrates extensibility to other bioactivities, such as those screened via AI for food-derived peptides with anti-obesity or anti-fatigue effects³³, while bioactive compounds' broader impacts on diseases like diabetes and cancer underscore the framework's biological relevance³⁴. Finally, our substitution framework, evaluating 14,580 disease-focused and 7798 antioxidant-focused one-hop swaps, yielded a mean 11.9% disease-risk reduction and 210.6% increase in predicted antioxidant activity across breakfast, lunch, and dinner. These results suggested FoodAtlas's dual role as both a mechanistic knowledge base and a decision-support engine. The enriched graph structure, complete with MeSH-normalized nodes and PubMed-cited provenance metadata, could empower in-depth toxicological and nutrigenomic investigations.

There are several rooms for improvement for our future work. First, our information-extraction F1 of 0.67 indicates errors in entity and relation recognition, particularly in complex or tabular contexts. In the future, we aim to utilize a more advanced large language model, such as GPT-5³⁵, incorporating techniques like multimodal reasoning³⁶ and structured output³⁷, as well as incorporating unstructured sources (e.g., tables) and hybrid architectures³⁸, to enhance the extraction performance. Second, CTD-derived disease associations are binary and do not capture dose-response dynamics or clinical outcomes. To improve, we consider incorporating quantitative toxicology (dose-response data, clinical endpoints) for predicting therapeutic success³⁹. Third, the BPM's reliance on existing pChEMBL and concentration data may introduce biases due to uneven assay coverage. In the future, we plan to expand coverage by developing an automated pipeline to extract bioactivity data from the broader scientific literature and implementing active learning approaches to prioritize underrepresented targets and compound classes⁴⁰. We also plan to expand our predictive bioactivity modeling to other endpoints, e.g., antidiabetic

Table 2 | Representative one-hop, within-category substitutions to maximize disease-risk reduction across meal types

Meal type	Meal foods	Category	Substitution (original → alt.)	Portion (g)	Key chemicals	Disease score change (%)	Supporting references
Breakfast	Whole-grain toast + <i>dill pickle</i>	Condiment	Dill pickle → olive	18	Quercetin +50.2%; oleic acid +23.4%	11.9	79,80
	Soy-latte + <i>soy milk splash</i>	Dairy substitute	Soy milk → almond milk	330	Genistein +92.6%; sucrose +6.6%	11.9	81,82
	Coffee + <i>oatmeal-raisin cookie</i>	Sweet bakery	Cookie → almond	60	Trans-resveratrol +88%; oleic acid +7%	11.8	83,84
Lunch	Spinach–quinoa bowl + <i>apple juice</i>	100% juice	Apple juice → grape juice	149.5	Sucrose +103.7%	12.2	85,86
	Grain wrap + <i>almond butter</i>	Plant-protein	Almond butter → hazelnut	48	Linoleic acid +83%; oleic acid +29%	11.8	87,88
	Garden salad + <i>dill pickle</i>	Condiment	Dill pickle → olive	18	Quercetin +50.2%; oleic acid +23.4%	12.1	79,80
Dinner	Spinach side-salad + <i>cranberry juice</i>	100% juice	Cranberry → pomegranate	248	Chlorogenic acid +51%; citric acid +25%	12.4	89,90
	Stir-fry sautéed in <i>sunflower oil</i>	Fat/Oil	Sunflower oil → coconut oil	8	Lauric acid ↑; total phenolics ↑	10.8	91,92
	Roast veg incl. <i>red cabbage</i>	Vegetable	Red cabbage → green cabbage	45	Kaempferol +101%	13.5	93,94

For each meal, the original food composition, the identified high-impact item, its replacement candidate, the primary food category, and the predicted percentage reduction in the normalized composite disease treatment score are shown, with supporting references.

Table 3 | Representative one-hop, within-category substitutions to maximize antioxidant bioactivity enhancement across meal types

Meal type	Meal foods	Category	Substitution (original → alt.)	Portion (g)	Key chemicals	Antioxidant score change (%)	Supporting references
Breakfast	Oats + <i>chia seed</i>	Plant-protein	Chia seed → pecan nut	10.5	Linoleic acid +85%; α-linolenic +19%	194.8	95,96
	Egg-melt + <i>Swiss cheese</i>	Dairy	Swiss cheese → reduced-fat cheese	21	Methionine +102%	182.2	97,98
	Muesli drizzle <i>flaxseed oil</i>	Fat/Oil	Flaxseed oil → soybean oil	4.5	Linoleic acid +398%; oleic acid +389%	244.4	99,100
Lunch	Kale-slaw w/ <i>green cabbage</i>	Vegetable	Green cabbage → red cabbage	22	Anthocyanins ↑; kaempferol +101%	200.6	101,102
	Med-wrap + <i>dill pickle</i>	Condiment	Dill pickle → black olive	18	Oleuropein +65%; oleic acid +24%	196.8	79,80
	Veg-burrito + <i>sweet green pepper</i>	Vegetable	Green pepper → beet (raw)	30	Betanin +352%; sucrose +413%	258.6	103,104
Dinner	Veg-lasagna + <i>cheese topping</i>	Dairy	Cheese → reduced-fat cheese	56.7	Methionine +102%	197.2	97,98
	Stir-fry finished w/ <i>soybean oil</i>	Fat/Oil	Soybean oil → pecan nut	10.5	Linoleic acid +85%; α-linolenic +19%	194	95,96
	Grilled entrée + <i>dill pickle</i>	Condiment	Dill pickle → black olive	18	Oleuropein +65%; oleic acid +24%	207.4	79,80

For each meal, the original food composition, the identified high-impact item, its replacement candidate, the primary food category, and the predicted percentage increase in the normalized antioxidant activity score are shown, with supporting references.

bioactivity, glycemic index values can serve as functional indicators of glucose regulation potential⁴¹, while anti-inflammatory activity can be approximated using the Dietary Inflammatory Index⁴². Fourth, due to the lack of a standardized flavor ontology, flavors are used as descriptors for chemicals rather than embedded as part of modeling. In the future, we plan to use an LLM-based human-in-the-loop tool to design a practical ontology for flavors^{43,44}, which will help with more versatile downstream tasks. Fifth, our one-hop substitution approach is simplified and does not reflect the multifaceted complexity of food-health interaction. Specifically, we do not account for multi-ingredient and chemical synergies, cultural dietary patterns, or palatability constraints, bioavailability, exposure, and metabolism^{45–50}, which could limit real-world adoption. For this work, we

focus on highlighting the potential of how the FoodAtlas Knowledge Graph could provide a data source even for simplistic downstream models to capture correlation corroborated with literature validation. As future work, we plan to continuously ingest newly published literature data into the FoodAtlas Knowledge Graph to support different use cases by scientists and collaborate with domain experts to advance the food-based health intervention research. Lastly, we continuously work to enhance the usability of our data and web interface, which we aim to extend ontologies, such as MONDO⁵¹ for diseases, for better searchability, different downloadable data formats, such as Parquet, JSON, RDF, and Neo4j, and periodic synchronization with external data sources, such as FoodOn¹¹, ChEBI⁵², and PubChem⁵³, to ensure interoperability.

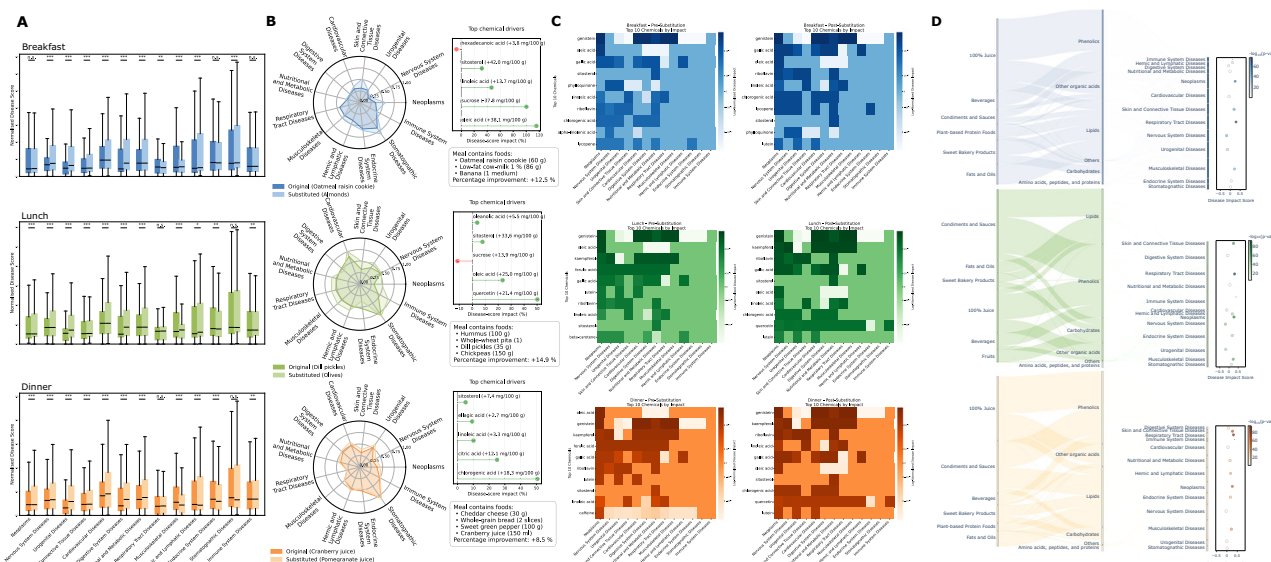


Fig. 5 | FoodAtlas-guided, category-matched ingredient swaps attenuate predicted disease risk and expose their chemical drivers across breakfast, lunch, and dinner. **A** Paired box-and-whisker plots of per-disease scores (22 MeSH classes) for breakfast (blue), lunch (green), and dinner (orange) meals before (lighter tint) and after (darker tint) replacing the single ingredient with the greatest weighted risk by a nutritionally superior food from the same category. Box centers show medians, hinges the inter-quartile range, whiskers $1.5 \times$ IQR, and dots outliers; horizontal bars denote Wilcoxon signed-rank significance ($p < 0.05$, $*p < 0.01$, $**p < 0.001$). All three meal types shift consistently toward lower risk after substitution. **B** Radar plots for one representative meal per time-of-day illustrate the multidimensional benefit of a single swap—nectarine $\rightarrow$ cranberry in a fruit-and-cottage-cheese breakfast (+21% aggregate improvement), hummus $\rightarrow$ olive in a pita-and-chickpea lunch (+18%), and oatmeal-raisin cookie $\rightarrow$ sweet-potato purée in a spinach-and-feta dinner (+4%). Solid polygons trace pre-swap scores, dashed polygons post-swap scores; adjoining dot-plots rank the five chemicals contributing most to each improvement. **C** For each meal type, paired heat-maps display the mean signed

contribution of the ten most influential chemicals (columns) to every disease class (rows): the left panel shows the change induced by the swap (baseline – optimized), the right panel the absolute post-swap impact. Meal-specific color maps (Blues, Greens, Oranges) reveal a conserved pattern in which polyphenol-rich flavonoids and omega-3 fatty acids align with risk reductions (cool hues), whereas heterocyclic amines, advanced glycation products, and saturated tri-acyl-glycerols drive risk elevations (warm hues). **D** Three-layer Sankey diagrams trace risk flow from food categories (left) through chemical classes (center) to disease classes (right). Node heights are equalized with a constant dummy flow to separate topology from magnitude, so dominant “risk bottlenecks” stand out visually. Companion lollipop charts plot the $-\log_{10}(p\text{-values})$ for enrichment of each disease class, confirming the statistical significance of the observed pathways. Together, the panels demonstrate that single, FoodAtlas-recommended swaps reliably lower predicted disease burden and that a small, mechanistically coherent set of chemicals mediates most of the benefit.

FoodAtlas provides a promising foundation for precision nutrition and food-innovation applications, uniting high-resolution compositional data, mechanistic disease links, predictive bioactivity models, and dietary guidance in a single resource.

Methods

Sentence retrieval

We initially gathered 1300 food names, including both common and scientific terms, from FooDB and FoodData Central (FDC)⁷. We then used these names to search PubMed⁵⁴ and PubMed Central (PMC)⁵⁵, which host 36 million and 9.8 million abstracts and full-text articles in biomedical and life sciences, respectively. The searches were conducted using the NCBI's Entrez Direct (EDirect)⁵⁶ to retrieve scientific literature that mentions these food names. We employed a search template “{food name} AND ((compound) OR (nutrient))” to ensure that the articles included not only the food name but also at least one of the keywords, ‘compound’ or ‘nutrient.’ This search template was empirically selected to enhance the accuracy of the results. Next, we tokenized the retrieved scientific literature into sentences using the NLTK library⁵⁷ and excluded sentences that were either too short or too long, following the approach used by LitSense in their study¹.

Sentence filtering

Since not all sentences in the retrieved articles above were related to food–chemical associations, we followed a two-step filtering procedure. First, we used a fuzzy matching⁵⁸ technique to keep only sentences containing one of the food terms described in the last section, resulting in about 10 million sentences mentioning food terms. Second, we fine-tuned a large

language model, BioBERT¹⁶, to classify sentences that contain valid food–chemical associations. This process resulted in 773,366 sentences that were highly likely ($p > 0.9$) to contain food–chemical associations. More details can be found in Supplementary Note 1.1.

Association extraction

For sentences that indicated food–chemical associations following the filtering step above, we passed those with a probability greater than 90% to the large language model-based information extraction models. We compared the performance of two different models: OpenAI GPT-4⁵⁹ (Model ID: gpt-4-0613) and GPT-3.5⁶⁰ with in-context learning or fine-tuning. For both models, we extracted food, food parts, chemicals, and chemical concentrations. For example, the sentence, “Chinese cabbage leaves contain Ca (1020 g kg⁻¹ FW), Fe (26 g kg⁻¹ FW), and total glucosinolates (10.926 micromol g⁻¹ DW),” would result in the following three tuple extraction: (Chinese cabbage, leaves, Ca, 1020 g kg⁻¹ FW), (Chinese cabbage, leaves, Fe, 26 g kg⁻¹ FW), and (Chinese cabbage, leaves, total glucosinolates, 10.926 micromol g⁻¹ DW). We provided more details in Supplementary Note 1.1.

Fine-tuning of GPT models using annotation rules

To improve the accuracy and specificity of extracted data, we fine-tuned the GPT-3.5 (Model ID: gpt-3.5-turbo-0125) model with a dataset curated based on an extensive set of annotation rules. We have two annotators label the same sentences independently and consolidate the disagreement. These rules were designed and validated through the analysis of over 1780 sentences containing food and chemical information. Specifically, we extracted the information in the format: {food}, {food part}, {chemical compound},

{*chemical concentration*}) by following key principles: First, extracting information exactly as presented in sentences to ensure fidelity. Second, annotating the most specific entities, such as individual compounds rather than broader chemical groups or food species, instead of generalized categories. Third, preserving contextual modifiers, such as adjectives and processing descriptors (e.g., “aqueous extract of *Mentha aquatica*”).

Fine-tuning focused on enhancing the model's ability to handle complex entity structures and relationships. For instance, sentences describing chemical concentrations within specific food parts (e.g., “celery leaves contain phenols”) were annotated to reflect both the food entity and its part. Additionally, cases involving synonyms, acronyms, and hierarchical relationships (e.g., “blueberry extract | *Vaccinium angustifolium* extract”) were fine-tuned to prioritize professional terminology and ensure completeness.

The fine-tuned models were trained to generate outputs in a structured format suitable for integration into the FAKG. Validation metrics, including precision, recall, and F_1 scores, demonstrated significant improvements over baseline models, particularly in extracting nuanced relationships and rare descriptors. Supplementary Note 1.1 shows the details for our fine-tuned model.

Knowledge graph construction

Our FoodAtlas knowledge graph (FAKG) was formed by a set of (*food*), contains, (*chemical*)-triplets. Each entity can be a food or a chemical, indexed by a unique FoodAtlas ID and other identifiers (e.g., FoodOn IDs for food and PubChem CIDs for chemical entities). Each *contains*-triplet is associated with metadata, providing additional information, including chemical concentration and references to the source where the information was extracted. To scalably expand FAKG, we developed a KG construction pipeline that harmonized and ingested the data extracted by the GPT models into the knowledge graph. First, the pipeline processed the CSV-formatted outputs from the LLM, discarding a few with incorrect CSV format. Second, chemical concentration information in the raw-string format was parsed into concentration value-unit pairs using regular expressions. The units convertible to each other were unified into a standardized unit, such as ‘g,’ ‘mg,’ ‘gram,’ and ‘kg,’ which were all converted into ‘100 g’ with corresponding changes in concentration value. Third, an entity linking system performed synonym resolution and string matching, mapping food and chemical names mentioned in sentences to ontology databases FoodOn¹¹ and ChEBI⁵², respectively. To enrich the interoperability and categorizability, identifiers from several other databases were retrieved (Supplementary Fig. 1 and Supplementary Note 1.2, 1.3).

Integration of disease data from the Comparative Toxicogenomics Database (CTD)

To enrich the FoodAtlas Knowledge Graph (FAKG) with disease-related data, we integrated chemical–disease associations from the Comparative Toxicogenomics Database (CTD)⁶¹. The CTD provides a robust dataset detailing relationships between chemicals and diseases and contains critical fields such as ChemicalName, ChemicalID (MeSH⁶² identifier), DiseaseName, DiseaseID (MeSH or OMIM⁶³ identifier), DirectEvidence (whether the chemical is a therapeutic or marker/mechanism for the disease, which we map as treats or worsens the disease, respectively), InferenceGeneSymbol, InferenceScore, and PubMedIDs (a | -delimited list of references).

We began by mapping the CTD chemicals to FoodAtlas using the ChemicalID field, which corresponds to MeSH identifiers. This alignment allowed us to cross-reference chemical entities in FoodAtlas with their associated diseases as recorded in CTD. Once the mappings were established, we extracted relationships of the type (chemical, associated with, disease) and populated the knowledge graph with these connections. The integration included metadata for these relationships, such as the original ChemicalIDs and DiseaseIDs found in CTD for the relationship, and whether the chemical exists in foods in FoodAtlas. Additionally, the linked PubMedIDs and their corresponding PMcIDs were incorporated as references

to enhance the verifiability and provenance of the data. For the diseases in these relationships, we extracted the DiseaseName, DiseaseID, AltDiseaseIDs (MeSH, OMIM, or DiseaseOntology identifiers), and the list of synonyms from CTD's disease information to create our FoodAtlas disease entities and their relationships to chemicals (Supplementary Note 1.3).

Integration of flavor data from FlavorDB and PubChem

Chemicals in FlavorDB are indexed using PubChem IDs, allowing for straightforward alignment with existing FoodAtlas entities. We scraped FlavorDB, iterating through PubChem IDs of FoodAtlas chemicals, and retrieving structured JSON data for each chemical. For chemicals missing from FlavorDB, we extended flavor data collection to PubChem's Hazardous Substances Data Bank (HSDB)²¹, extracting “Taste” and “Odor” descriptors. While PubChem provided a significant volume of additional data, its descriptors were often noisy and inconsistently formatted. Hence, fuzzy string matching techniques⁵⁸ (Supplementary Note 1.2) were employed to harmonize flavor descriptors, creating a unified set of standardized flavor terms. Each unique flavor descriptor was represented in the knowledge graph as an entity with the entity type flavor, and relationships between chemicals and their flavors were established using a newly defined *hasFlavor* relationship (Supplementary Note 1.3).

Integration of food and chemical bioactivity via Bioactivity Prediction Model

To integrate chemical bioactivity with food composition data in the FAKG, we employed a multi-step procedure for both data assembly and model development using machine learning. First, we mapped chemicals present in FA to corresponding records in the ChEMBL database⁶⁴, using InChIKey identifiers as the primary linkage. This allowed us to retrieve chemical pChEMBL values (i.e., $-\log_{10}(\text{nM})$) for the related assays that capture various types of bioactivity, including antioxidant, antimicrobial, anti-inflammatory, immunomodulatory, antiviral, neuroprotective, anticancer, and antidiabetic activities (Supplementary Note 1.5). After establishing the connections between chemicals in FA and their associated assays in ChEMBL, we linked each food to its constituent chemicals using concentration data available in FA. The resulting dataset enabled the formation of a Bioactivity Food Link (BFL), whereby each food inherited potential bioactivity from the chemicals it contains. This dataset was further enriched by including the Ferric Reducing Antioxidant Power (FRAP) values⁶⁵ for 159 food samples, providing ground truth for antioxidant capacity. The resulting Bioactivity Prediction Model (BPM) was designed to predict antioxidant capacity, quantified as $-\log_{10}(\text{FRAP})$, based on a machine learning architecture, Random Forest. The model relies on three primary input modalities. First, we generated Morgan fingerprints from SMILES strings using RDKit⁶⁶, encoding chemical structure as a binary vector. Second, we included pChEMBL values⁶⁷, which reflect comparable measures of concentrations to reach half-maximal potency/affinity/response (e.g., IC₅₀, EC₅₀) on a negative logarithmic scale. Third, we incorporated chemical concentration information from FA to capture the quantity of each chemical present in a given food. These three types of inputs were fed into machine learning models. Outputs from the model predicted $-\log_{10}(\text{FRAP})$ for each food. During model training, we employed an 80:20 train-test split, which we repeated across 25 bootstrap iterations to account for variability due to the limited dataset size. The Random Forest was tuned using a hyperparameter grid search. We tracked metrics such as mean absolute error (MAE), MSE, Pearson's correlation (PCC), and the coefficient of determination (R^2) between predicted and observed FRAP values. Residual analysis was used to gauge systematic errors and to verify the stability of predictions across foods with diverse chemical profiles (Supplementary Note 1.5).

One-hop food substitutions for disease- and antioxidant bioactivity-oriented diet improvement

We implemented a one-hop substitution framework to evaluate and enhance the healthfulness of real-world meals by jointly optimizing disease-

risk and antioxidant-bioactivity profiles. First, we ingested the USDA “What We Eat in America” (WWEIA) diet recall data⁶⁸, extracting each meal’s constituent foods and their portion sizes. For disease modeling, we mapped foods to associated conditions (e.g., cardiovascular, inflammatory) using curated chemical–disease relationships from the CTD. In parallel, we assigned antioxidant bioactivity scores derived from FRAP assays. To enable direct comparison and aggregation across disparate metrics, both disease-risk and bioactivity scores were min-max normalized between 0 and 1. We then computed a weighted aggregate score for each meal by summing individual food contributions in proportion to their serving size. Next, we identified “high-impact” foods, those whose removal yielded the greatest disease-risk reduction, and generated candidate alternatives drawn from the same primary food category (e.g., dairy, leafy greens). For each candidate swap, we recomputed the meal’s composite disease and antioxidant scores under the same portion-weighted scheme. Substitutions that produced the maximal net benefit, minimizing normalized disease risk while maximizing normalized antioxidant activity, were selected as optimal (Supplementary Note 1.6).

Data availability

All data is available at <https://github.com/IBPA/FoodAtlas-KGv2> and <https://foodatlas.ai>. Food Atlas may be using data that are restricted licenses for different uses, please consult individual license terms from each data source. *CTD compliance*. FoodAtlas integrates curated knowledge from the Comparative Toxicogenomics Database (CTD) by linking to CTD chemical/disease identifiers and PubMed references; we do not redistribute CTD-curated interaction tables in our public releases. Users can reproduce the CTD integration locally using our scripts, which download CTD data from the source and join via CTD IDs and PMIDs. Non-commercial use is free; commercial reuse requires a license from CTD’s licensing agent. Please consult CTD’s terms before downloading/using CTD content.

Code availability

All code and instructions on how to reproduce the results for knowledge graph construction and information extraction can be found at <https://github.com/IBPA/FoodAtlas-KGv2> and <https://github.com/IBPA/Lit2KG>, respectively.

Received: 17 July 2025; Accepted: 17 December 2025;

Published online: 20 January 2026

References

- Allot, A. et al. LitSense: making sense of biomedical literature at sentence level. *Nucleic Acids Res.* **47**, W594–W599 (2019).
- Youn, J., Li, F., Simmons, G., Kim, S. & Tagkopoulos, I. FoodAtlas: automated knowledge extraction of food and chemicals from literature. *Comput. Biol. Med.* **181**, 109072 (2024).
- Cifuentes, A. Food analysis and foodomics. *J. Chromatogr. A* **1216**, 7109 (2009).
- García-Cañas, V., Simó, C., Herrero, M., Ibáñez, E. & Cifuentes, A. Present and future challenges in food analysis: foodomics. *Anal. Chem.* **84**, 10150–10159 (2012).
- FoodDB. <https://fooddb.ca/>. [Accessed at 12/25/2025]
- McKillop, K., Harnly, J., Pehrsson, P., Fukagawa, N. & Finley, J. FoodData Central, USDA’s updated approach to food composition data systems. *Curr. Dev. Nutr.* **5**, 596–596 (2021).
- USDA FoodData Central. <https://fdc.nal.usda.gov/>. [Accessed at 12/25/2025]
- Capozzi, F. & Bordon, A. Foodomics: a new comprehensive approach to food and nutrition. *Genes Nutr.* **8**, 1–4 (2013).
- Min, W., Liu, C., Xu, L. & Jiang, S. Applications of knowledge graphs for food science and industry. *Patterns* **3**, 100484 (2022).
- Jahangir, M., Kim, H. K., Choi, Y. H. & Verpoorte, R. Health-affecting compounds in *Brassicaceae*. *Compr. Rev. Food Sci. Food Saf.* **8**, 31–43 (2009).
- Dooley, D. M. et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Sci. Food* **2**, 23 (2018).
- Eftimov, T., Ispirova, G., Potočnik, D., Ogrinc, N. & Koroušić Seljak, B. ISO-FOOD ontology: a formal representation of the knowledge within the domain of isotopes for food science. *Food Chem.* **277**, 382–390 (2019).
- Furukawa H. Deep Learning for End-to-End Automatic Target Recognition from Synthetic Aperture Radar Imagery. IEICE Technical Report; IEICE Tech. Rep. **117**, 35–40 (2018).
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. (2019).
- Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems*, 30. <https://doi.org/10.48550/arXiv.1706.03762> (2017).
- Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
- Cenik, G., Seljak, B. K. & Eftimov, T. FoodChem: A food-chemical relation extraction model. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1–8). IEEE. (2021).
- Özen, N., Mu, W., van Asselt, ED. & van den Bulk, LM. Extracting chemical food safety hazards from the scientific literature automatically using large language models. *Appl. Food Res.* **5**, 100679, <https://doi.org/10.1016/j.afres.2024.100679> (2025).
- Davis, A. P. et al. Comparative Toxicogenomics Database’s 20th anniversary: update 2025. *Nucleic Acids Res.* **53**, D1328–D1334 (2025).
- FlavorDB: a database of flavor molecules | Nucleic Acids Research | Oxford Academic. <https://academic.oup.com/nar/article/46/D1/D1210/4559748>.
- Fonger, G. C. Hazardous substances data bank (HSDB) as a source of environmental fate information on chemicals. *Toxicology* **103**, 137–145 (1995).
- Hausmann, S. et al. FoodKG: a semantics-driven knowledge graph for food recommendation. In *The Semantic Web – ISWC 2019* (eds Ghidini, C. et al.) Vol. 11779 146–162 (Springer International Publishing, Cham, 2019).
- 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.* **11**, 931 (2021).
- Ni, Y., Jensen, K., Kouskoumvekaki, I. & Panagiotou, G. NutriChem 2.0: exploring the effect of plant-based foods on human health and drug efficacy. *Database* **2017**, bax044 (2017).
- van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
- McInnes, L., Healy, J. & Astels, S. hdbSCAN: hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).
- Mullahy, J. Specification and testing of some modified count data models. *J. Econom.* **33**, 341–365 (1986).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
- Johnson, G. H. & Fritsche, K. Effect of dietary linoleic acid on markers of inflammation in healthy persons: a systematic review of randomized controlled trials. *J. Acad. Nutr. Diet.* **112**, 1041.e1–15 (2012).
- Crowell, P. L. Prevention and therapy of cancer by dietary monoterpenes. *J. Nutr.* **129**, 775S–778S (1999).

31. Benzie, I. F. F. & Choi, S.-W. Chapter One - Antioxidants in food: content, measurement, significance, action, cautions, caveats, and research needs. In *Advances in Food and Nutrition Research* (ed. Henry, J.) Vol. 71 1–53 (Academic Press, 2014).
32. Shahidi, F. & Ambigaipalan, P. Phenolics and polyphenolics in foods, beverages and spices: antioxidant activity and health effects – A review. *J. Funct. Foods* **18**, 820–897 (2015).
33. Chang, J., Wang, H., Su, W., He, X. & Tan, M. Artificial intelligence in food bioactive peptides screening: recent advances and future prospects. *Trends Food Sci. Technol.* **156**, 104845 (2025).
34. Alvarez-Leite, J. I. The role of bioactive compounds in human health and disease. *Nutrients* **17**, 1170 (2025).
35. GPT-5 System Card. <https://openai.com/index/gpt-5-system-card/> (2025).
36. Zhang, Z. et al. Multimodal chain-of-thought reasoning in language models. <https://openreview.net/forum?id=gDlsMWost9> (2023).
37. Liu, M. X. et al. 'We Need Structured Output': towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* 1–9 (Association for Computing Machinery, New York, NY, USA, 2024).
38. Lu, W. et al. Large language model for table processing: a survey. *Front. Comput. Sci.* **19**, 192350 (2025).
39. Li, F., Youn, J., Millsop, C. & Tagkopoulos, I. Predicting clinical trial success for Clostridium difficile infections based on preclinical data. *Front. Artif. Intell.* 710.3389/frai.2024.1487335 (2024).
40. Ren, P. et al. A survey of deep active learning. *ACM Comput. Surv.* **54**, 180:1–180:40 (2021).
41. Foster-Powell, K., Holt, S. H. & Brand-Miller, J. C. International table of glycemic index and glycemic load values: 2002. *Am. J. Clin. Nutr.* **76**, 5–56 (2002).
42. Shivappa, N., Steck, S. E., Hurley, T. G., Hussey, J. R. & Hébert, J. R. Designing and developing a literature-derived, population-based dietary inflammatory index. *Public Health Nutr.* **17**, 1689–1696 (2014).
43. Toro, S. et al. Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *J. Biomed. Semant.* **15**, 19 (2024).
44. Youn, J., Naravane, T. & Tagkopoulos, I. Using Word Embeddings to Learn a Better Food Ontology. *Front. Artif. Intell.* 310.3389/frai.2020.584784. (2020).
45. Jacobs, D. R. Jr & Tapsell, L. C. Food synergy: the key to a healthy diet. *Proc. Nutr. Soc.* **72**, 200–206 (2013).
46. Nemec, K. Cultural awareness of eating patterns in the health care setting. *Clin. Liver Dis.* **16**, 204–207 (2020).
47. Forde, C. G. & de Graaf, K. Influence of sensory properties in moderating eating behaviors and food intake. *Front. Nutr.* **9**, 841444 (2022).
48. Melse-Boonstra, A. Bioavailability of micronutrients from nutrient-dense whole foods: zooming in on dairy, vegetables, and fruits. *Front. Nutr.* **7**, 101 (2020).
49. Benford, D. et al. The principles and methods behind EFSA's guidance on uncertainty analysis in scientific assessment. *EFSA J* **16**, e05122 (2018).
50. Shannar, A. et al. Pharmacodynamics (PD), pharmacokinetics (PK) and PK-PD modeling of NRF2 activating dietary phytochemicals in cancer prevention and in health. *Curr. Pharmacol. Rep.* **11**, 6 (2024).
51. Vasilevsky, N. A. et al. Mondo: Unifying diseases for the world, by the world. Preprint at <https://doi.org/10.1101/2022.04.13.22273750> (2022).
52. Hastings, J. et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
53. Kim, S. et al. PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).
54. PubMed. *PubMed* <https://pubmed.ncbi.nlm.nih.gov/>. [Accessed at 12/25/2025].
55. PubMed Central (PMC). *PubMed Central (PMC)* <https://pmc.ncbi.nlm.nih.gov/>. [Accessed at 12/25/2025].
56. Kans, J. Entrez Direct: E-utilities on the Unix Command Line. In *Entrez Programming Utilities Help [Internet]* (National Center for Biotechnology Information (US), 2025).
57. Bird, S. & Loper, E. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics. (2004).
58. Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv.* **33**, 31–88 (2001).
59. OpenAI et al. GPT-4 technical report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2024).
60. Ye, J. et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. Preprint at <https://doi.org/10.48550/arXiv.2303.10420> (2023).
61. Davis, A. P., Wieggers, T. C., Rosenstein, M. C. & Mattingly, C. J. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database J. Biol. Databases Curation* **2012**, bar065 (2012).
62. Dhammi, I. K. & Kumar, S. Medical subject headings (MeSH) terms. *Indian J. Orthop.* **48**, 443–444 (2014).
63. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
64. Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
65. Carlsen, M. H. et al. The total antioxidant content of more than 3100 foods, beverages, spices, herbs and supplements used worldwide. *Nutr. J.* **9**, 3 (2010).
66. Laufkötter, O., Sturm, N., Bajorath, J., Chen, H. & Engkvist, O. Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. *J. Cheminformatics* **11**, 54 (2019).
67. Lenselink, E. B. et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminformatics* **9**, 45 (2017).
68. What We Eat In America (WWEIA) Database. Food Surveys Research Group <https://doi.org/10.15482/USDA.ADC/1178144> (2015).
69. Calder, P. C. Omega-3 fatty acids and inflammatory processes. *Nutrients* **2**, 355–374 (2010).
70. Mozaffarian, D. & Rimm, E. B. Fish intake, contaminants, and human health: evaluating the risks and the benefits. *JAMA* **296**, 1885–1899 (2006).
71. Ramsden, C. E. et al. Use of dietary linoleic acid for secondary prevention of coronary heart disease and death: evaluation of recovered data from the Sydney Diet Heart Study and updated meta-analysis. *BMJ* **346**, e8707 (2013).
72. Cassidy, A. et al. High anthocyanin intake is associated with a reduced risk of myocardial infarction in young and middle-aged women. *Circulation* **127**, 188–196 (2013).
73. Kalt, W. et al. Recent research on the health benefits of blueberries and their anthocyanins. *Adv. Nutr.* **11**, 224–236 (2020).
74. Hankinson, A., Lloyd, B. & Alweis, R. Lime-induced phytophotodermatitis. *J. Community Hosp. Intern. Med. Perspect.* **4**, 25090 (2014).
75. Sacks, F. M. et al. Dietary fats and cardiovascular disease: a presidential advisory from the American Heart Association. *Circulation* **136**, e1–e23 (2017).
76. de Souza, R. J. et al. Intake of saturated and trans unsaturated fatty acids and risk of all cause mortality, cardiovascular disease, and

- type 2 diabetes: systematic review and meta-analysis of observational studies. *BMJ* **351**, h3978 (2015).
77. Imamura, F. et al. Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction. *BMJ* **351**, h3576 (2015).
78. Yang, Q. et al. Added sugar intake and cardiovascular diseases mortality among US adults. *JAMA Intern. Med.* **174**, 516–524 (2014).
79. Schwingshackl, L. et al. Food groups and risk of all-cause mortality: a systematic review and meta-analysis of prospective studies. *Am. J. Clin. Nutr.* **105**, 1462–1473 (2017).
80. Rocha, J., Borges, N. & Pinho, O. Table olives and health: a review. *J. Nutr. Sci.* **9**, e57 (2020).
81. Li, S.-C. et al. Almond consumption improved glycemic control and lipid profiles in patients with type 2 diabetes mellitus. *Metabolism* **60**, 474–479 (2011).
82. Lee-Bravatti, M. A. et al. Almond consumption and risk factors for cardiovascular disease: a systematic review and meta-analysis of randomized controlled trials. *Adv. Nutr.* **10**, 1076–1088 (2019).
83. Martin, N., Germanò, R., Hartley, L., Adler, A. J. & Rees, K. Nut consumption for the primary prevention of cardiovascular disease. *Cochrane Database Syst. Rev.* **2015**, CD011583 (2015).
84. Jiang, R. Nut and peanut butter consumption and risk of type 2 diabetes in women. *JAMA* **288**, 2554 (2002).
85. Blumberg, J., Vita, J. & Chen, C. Concord grape juice polyphenols and cardiovascular risk factors: dose-response relationships. *Nutrients* **7**, 10032–10052 (2015).
86. Stein, J. H., Keevil, J. G., Wiebe, D. A., Aeschlimann, S. & Folts, J. D. Purple grape juice improves endothelial function and reduces the susceptibility of LDL cholesterol to oxidation in patients with coronary artery disease. *Circulation* **100**, 1050–1055 (1999).
87. Zhao, J., Wang, X., Lin, H. & Lin, Z. Hazelnut and its by-products: a comprehensive review of nutrition, phytochemical profile, extraction, bioactivities and applications. *Food Chem.* **413**, 135576 (2023).
88. Tey, S. L. et al. Effects of different forms of hazelnuts on blood lipids and α -tocopherol concentrations in mildly hypercholesterolemic individuals. *Eur. J. Clin. Nutr.* **65**, 117–124 (2011).
89. Jazinaki, M. S., Rashidmayvan, M. & Pahlavani, N. The effect of pomegranate juice supplementation on C-reactive protein levels: GRADE-assessed systematic review and dose-response updated meta-analysis of data from randomized controlled trials. *Phytother. Res.* **38**, 2818–2831 (2024).
90. Basu, A. & Penugonda, K. Pomegranate juice: a heart-healthy fruit juice. *Nutr. Rev.* **67**, 49–56 (2009).
91. Khaw, K.-T. et al. Randomised trial of coconut oil, olive oil or butter on blood lipids and other cardiovascular risk factors in healthy men and women. *BMJ Open* **8**, e020167 (2018).
92. Eyres, L., Eyres, M. F., Chisholm, A. & Brown, R. C. Coconut oil consumption and cardiovascular risk factors in humans. *Nutr. Rev.* **74**, 267–280 (2016).
93. Yang, D. K. Cabbage (*Brassica oleracea* var. *capitata*) protects against H_2O_2 -induced oxidative stress by preventing mitochondrial dysfunction in H9c2 cardiomyoblasts. *Evid. Based Complement. Alternat. Med.* **2018**, 2179021 (2018).
94. Jiang, Y. et al. Cruciferous vegetable intake is inversely correlated with circulating levels of proinflammatory markers in women. *J. Acad. Nutr. Diet.* **114**, 700–708.e2 (2014).
95. McKay, D., Eliasziw, M., Chen, C. & Blumberg, J. A pecan-rich diet improves cardiometabolic risk factors in overweight and obese adults: a randomized controlled trial. *Nutrients* **10**, 339 (2018).
96. Robbins, K. S., Gong, Y., Wells, M. L., Greenspan, P. & Pegg, R. B. Reprint of “Investigation of the antioxidant capacity and phenolic constituents of U.S. pecans. *J. Funct. Foods* **18**, 1002–1013 (2015).
97. Feeney, E. L., Lamichhane, P. & Sheehan, J. J. The cheese matrix: understanding the impact of cheese structure on aspects of cardiovascular health – a food science and a human nutrition perspective. *Int. J. Dairy Technol.* **74**, 656–670 (2021).
98. Rangel, A. H. D. N. et al. An overview of the occurrence of bioactive peptides in different types of cheeses. *Foods* **12**, 4261 (2023).
99. Lemke, S. L. et al. Dietary intake of stearidonic acid-enriched soybean oil increases the omega-3 index: randomized, double-blind clinical study of efficacy and safety. *Am. J. Clin. Nutr.* **92**, 766–775 (2010).
100. Baer, D. J., Henderson, T. & Gebauer, S. K. Consumption of high-oleic soybean oil improves lipid and lipoprotein profile in humans compared to a palm oil blend: a randomized controlled trial. *Lipids* **56**, 313–325 (2021).
101. Fang, S., Lin, F., Qu, D., Liang, X. & Wang, L. Characterization of purified red cabbage anthocyanins: improvement in HPLC separation and protective effect against H_2O_2 -induced oxidative stress in HepG2 cells. *Molecules* **24**, 124 (2018).
102. Wiczowski, W., Szawara-Nowak, D. & Topolska, J. Red cabbage anthocyanins: profile, isolation, identification, and antioxidant activity. *Food Res. Int.* **51**, 303–309 (2013).
103. Siervo, M. et al. Nitrate-rich beetroot juice reduces blood pressure in Tanzanian adults with elevated blood pressure: a double-blind randomized controlled feasibility trial. *J. Nutr.* **150**, 2460–2468 (2020).
104. Clifford, T., Howatson, G., West, D. & Stevenson, E. The potential benefits of red beetroot supplementation in health and disease. *Nutrients* **7**, 2801–2822 (2015).
105. Grover, A. & Leskovec, J. "node2vec: Scalable feature learning for networks." Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. (2016).

Acknowledgements

This work was supported by the USDA-NIFA AI Institute for Next Generation Food Systems (AIFS), USDA-NIFA award number 2020-67021-32855.

Author contributions

F.L. and J.Y. created the knowledge graph and built the sentence filtration and sentence extraction pipeline. K.X. annotated the information extraction dataset and fine-tuned GPT-3.5 on the data. T.C. performed analysis for results, generated figures, and wrote the paper. P.G. built the bioactivity prediction model. A.Y. integrated disease data into FA. M.G. integrated flavor data into FA. K.N. helped K.X. with annotation. I.T. conceived and supervised all aspects of the project, including project management, framework design, pipeline architecture, predictive model, data analysis, hypothesis generation and testing. All authors contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41538-025-00680-9>.

Correspondence and requests for materials should be addressed to Ilias Tagkopoulos.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026