



# From metabolomics dark matter to bioactivity: a computational framework for prioritizing food-derived bioactive compounds



Pranav Gupta<sup>1,2,3</sup>, Michael Gunning<sup>1,2,3</sup>, Selena Ahmed<sup>4</sup> & Ilias Tagkopoulos<sup>1,2,3</sup> ✉

Despite the growing scale of untargeted food metabolomics, scalable approaches for translating raw LC-MS profiles into functionally prioritized, experimentally testable candidates remain limited. Most detected features remain structurally unresolved, and the few that are annotated cannot be reliably connected to whole-food biological activity through additive models alone. Here we present a multi-layer prioritization framework that addresses this bottleneck across 322 commonly consumed foods in the United States, comprising 30,343 small molecules. A retention-index model ( $R^2 = 0.95$ ,  $PCC = 0.97$ ) expanded putative structural coverage from 7.4% to over 78% of detected features, assigning confidence-tiered structural proxies through formula-constrained matching against 11.8 million PubChem compounds. A two-stage large language model workflow harmonized 39,158 assay protocols spanning 119,633 chemical-assay measurements from ChEMBL into standardized bioactivity classes, enabling compound-level evidence to be propagated across the food composition space. The framework was applied to antioxidant activity as a primary, well-characterized endpoint while additional bioactivity domains (anti-inflammatory, anticancer, antidiabetic, antiviral, and antibacterial effects) were incorporated as compound-level evidence layers to support metabolite prioritization. Integrating concentration-aware activity mapping with a machine learning model trained on whole-food antioxidant measurements yielded strong predictive performance ( $R^2 = 0.73$ ,  $PCC = 0.85$ ) and prioritized metabolites most strongly associated with model-predicted food-level antioxidant capacity via SHAP attribution. Structure-based analysis using Boltz-2 against the antioxidant sensor provided a final mechanistic triage layer, distinguishing high-priority candidates from lower-ranked compounds. Applied across 322 foods, this funnel reduces tens of thousands of untargeted LC-MS features to a confidence-ranked shortlist for targeted wet-lab validation and provides a generalizable strategy for translating food metabolomics into biologically interpretable and actionable hypotheses. This framework establishes a scalable foundation for transforming food composition data into actionable biological insight, advancing the integration of food, health, and data science. More broadly, these results highlight limitations of traditional nutrient-centric frameworks and support a shift toward modeling molecular patterns and compositional signatures to better capture the functional properties of foods.

<sup>1</sup>Department of Computer Science, University of California at Davis, Davis, CA, USA. <sup>2</sup>Genome Center, University of California at Davis, Davis, CA, USA. <sup>3</sup>USDA/NSF AI Institute for Next Generation Food Systems (AIFS), Davis, CA, USA. <sup>4</sup>The Periodic Table of Food Initiative (PTFI), American Heart Association, Dallas, TX, USA. ✉e-mail: [itagkopoulos@ucdavis.edu](mailto:itagkopoulos@ucdavis.edu)

The health and nutritional value of food extends far beyond macro- and micronutrients to a vast “dark matter” of thousands of metabolites that influence human health through diverse biological pathways<sup>1–3</sup>. Many of these metabolites have been directly linked to human health effects and frequently display a range of bioactivities measurable by in vitro and in vivo bioassays<sup>4</sup>. However, characterizing the bioactivity of these compounds across the global food supply remains a significant challenge due to the sparsity of comprehensive and standardized metabolomics data and the fragmentation of experimental results across studies. Traditionally, the health-relevant chemistry of foods has been characterized either by targeted quantification of a small, pre-selected panel of known nutrients or marker compounds, or by applying single bulk antioxidant assays (e.g., FRAP, DPPH, ABTS, or ORAC) to whole-food extracts. While valuable, such approaches probe only a limited fraction of food chemical space, rarely scale across the diversity of the global food supply, and largely overlook the thousands of uncharacterized small molecules that untargeted metabolomics now routinely detects. While the bioactive properties of hundreds of thousands of individual isolated compounds are well-documented in pharmacological databases such as ChEMBL<sup>5,6</sup>, comparable whole-food data remain rare because standardized measurement frameworks are limited. Furthermore, simply cataloging the presence of bioactive compounds within a food matrix is insufficient to infer the bioactivity of the food as a whole or within diets. Moreover, because biological effects are concentration-dependent, the mere presence of a compound is uninformative without an estimate of how much of it is present. Complex chemical interactions—including synergistic and antagonistic matrix effects—lead to highly non-linear functional outcomes that cannot be captured by simple additive models<sup>7,8</sup>.

Recent advances in food metabolomics, particularly *The Periodic Table of Food Initiative* (PTFI)<sup>9,10</sup> standardized untargeted metabolomics based on liquid chromatography-mass spectrometry (LC-MS), offer an unprecedented, high-resolution view of food composition<sup>2,11</sup>. Yet, the sheer volume and dimensionality of these data—typically comprising thousands of

unique metabolic features across hundreds of food items—necessitate sophisticated computational tools for meaningful analysis. Inter-food analysis is further complicated because many structurally complex metabolites are unique to only a handful of foods. To overcome this, comprehensive knowledge graphs like *FoodAtlas*<sup>1</sup> can be utilized to contextualize these high-resolution compositional profiles, mapping known chemicals to specific dietary sources and biological networks. Despite these informatic advances, there remains a critical gap: there is currently no scalable computational framework capable of translating these complex, whole-food LC-MS profiles into standardized, predictive bioactivity rankings. Historically, literature-mining efforts in this domain have been largely restricted to isolated antioxidant assays, leaving the functional potential of the broader food metabolome largely uncharacterized.

In this study, we bridge the gap between compositional metabolomics, predictive machine learning, and structure-aware bioactivity prioritization to systematically profile the functional potential of the food supply. Specifically, we address a fundamental informatic bottleneck: how to reduce thousands of untargeted food LC-MS features into a tractable set of biologically prioritized candidates. To achieve this, we developed a multi-layered computational framework (Fig. 1). First, we utilized Retention Index (RI)-guided structural narrowing to assign high-, medium-, and low-confidence structural proxies to previously unidentified metabolomic features. As a central intermediate step, we then converted relative LC-MS intensities into estimated concentrations by anchoring detected features to compound concentrations curated in FoodAtlas, enabling concentration-aware rather than presence/absence-based downstream analysis. Second, we deployed a novel, hierarchical Large Language Model (LLM) workflow to harmonize heterogeneous in vitro assay data from ChEMBL into standardized bioactivity classes, culminating in a standardized evidence layer that links compound-level assay activity to food composition data for downstream food and metabolite prioritization. Third, leveraging this foundational dataset, we introduce a non-linear machine learning pipeline capable of predicting whole-food bioactivities directly from complex metabolomic

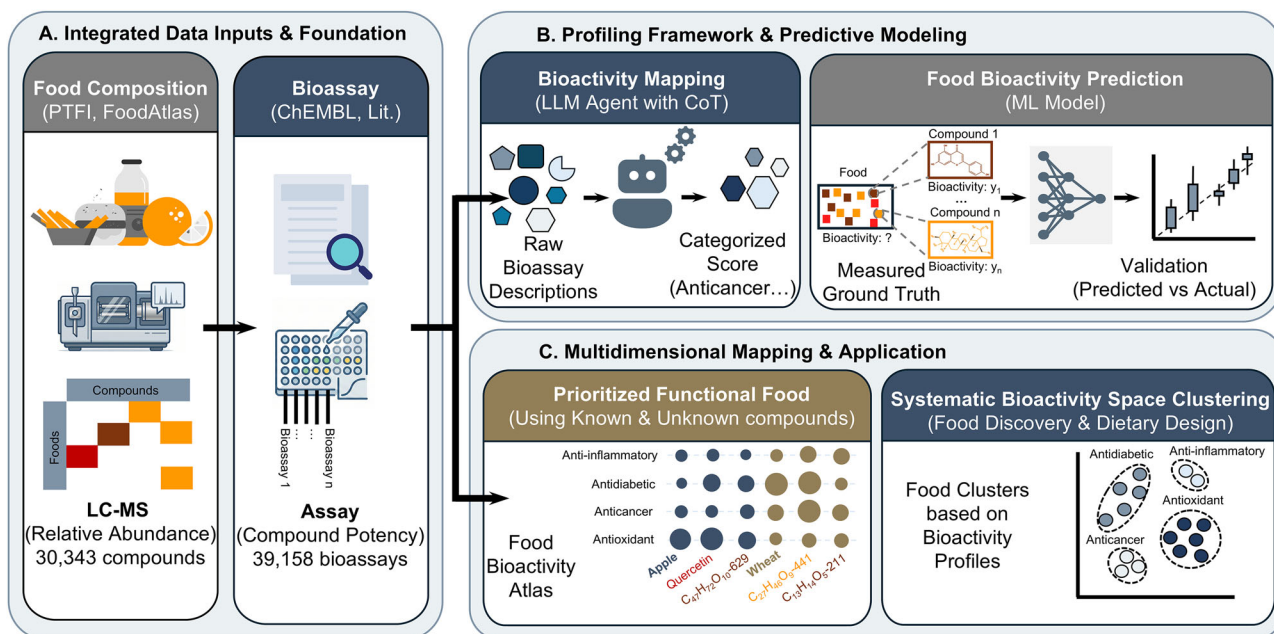


**Fig. 1 | Integrated computational framework for large-scale food bioactivity prediction and profiling.** **A Integrated Data Inputs & Foundation:** The pipeline merges high-resolution food composition data, comprising 30,343 small-molecules from databases such as PTFI via LC-MS and FoodAtlas via Literature, with extensive empirical bioassay records encompassing 39,158 assays. **B Profiling Framework & Predictive Modeling:** Raw assay descriptions were harmonized into standardized bioactivity classes using a two-stage large language model workflow, while retention-index-guided structural narrowing was used to assign confidence-tiered structural

proxies to previously unidentified LC-MS features. A whole-food bioactivity predictor with antioxidant-ranking component was then used to prioritize metabolites and foods for downstream follow-up. **C Multidimensional Mapping & Application:** The resulting framework generates a multidimensional Food Bioactivity Atlas that supports metabolomic clustering, compound-evidence-guided food prioritization, and shortlist generation for known (e.g., quercetin) and previously unresolved metabolites (e.g.,  $C_{47}H_{72}O_{10}$ -629), which can then be triaged by structure-based analysis for targeted validation.

profiles, successfully capturing synergistic and antagonistic matrix-level interaction effects. We integrated three model-guided ranking layers to prioritize metabolites and foods for downstream follow-up. We then use Boltz-2 analysis against the KEAP1 Kelch domain<sup>12,13</sup> as a mechanistic case study to further refine shortlisted candidates. We limited whole-food supervised prediction to antioxidant capacity measured by the Ferric Reducing Antioxidant Power (FRAP) assay, while other bioactivity categories are used as compound-level evidence layers for prioritization rather than food-level prediction targets. By applying this framework, we convert food metabolomic dark matter into confidence-ranked, testable molecules for wet-lab validation and functional food discovery. Importantly, unlike pharmacological systems that focus on isolated compounds, foods represent complex, multi-component mixtures where exposure, matrix interactions, and co-occurrence fundamentally shape biological outcomes, requiring a distinct analytical framework. The ability to perform this analysis is critically enabled by standardized metabolomics protocols, which ensure cross-sample comparability and allow models to generalize across foods and datasets. Together, this work establishes a scalable approach for translating high-dimensional food metabolomics into functional insight, enabling systematic prioritization of bioactive compounds and advancing our ability to link food composition to biological outcomes.

## Results

### Metabolite-based food clusters reveal unexpected similarities between disparate foods

The food metabolome was highly sparse and food-specific: of the 27,713 detected features, 44% were unique to a single food, and 85% were present in five foods or fewer, highlighting the extent of molecular heterogeneity across the dataset. HDBSCAN identified 4 clusters as well as 52 foods that were identified as outliers, or not belonging to any cluster (Fig. 2 and Table 1). These clusters aligned remarkably well with traditional food groups including fruits and vegetables; meat, dairy, and high-protein foods; whole grains, seeds, and nuts; and processed or compound foods (Chi-squared = 4.86<sup>-37</sup>), yet the chemical “fingerprints” of certain foods revealed surprising deviations from their culinary labels. Cluster 1 was significantly overrepresented by unclassified foods (no matching FoodOn ID), which in this dataset were primarily compound or processed foods rather than raw ingredients. Consistent with this interpretation, Cluster 1 was enriched for food-derived small molecules associated with processing or external material inputs. One notable example was N-Phenyl-1-naphthylamine, an antioxidant widely used in rubber and industrial materials<sup>14</sup> that can migrate into foods from processing equipment or packaging<sup>15</sup>, and it was one of the key features found in Cluster 1 (a food-contact migrant rather than an endogenous metabolite). Its presence helps distinguish this cluster from minimally processed foods and suggests that, in addition to intrinsic food chemistry, post-harvest handling and processing environments can measurably shape food metabolomic profiles. Cluster 2 was significantly overrepresented by fruits and vegetables and underrepresented by unclassified foods. Cluster 3 was overrepresented by dairy food products, and Cluster 4 largely consisted of foods in the “plant seed or nut” category Supplementary Fig. 2). Key metabolites for each cluster as well as a full food map can be found in Fig. 2.

Although many food items clustered predictably, there were some notable exceptions, such as white rice and wild rice. While wild rice was clustered together with other seeds and grains, white rice was clustered with more processed foods. Indeed, when examining the metabolomic profile of these two rice varieties, only 33 unique metabolites were identified in white rice, whereas 411 were identified in wild rice. Of these, only 11 were shared between the two foods, and of those 11, wild rice generally had 2–6x higher relative abundance. This pattern is consistent with the retention of bran/husk-associated secondary metabolites in less-polished wild rice<sup>16</sup> and the depletion of these compounds during white-rice polishing<sup>17</sup>, although species differences between *Oryza sativa* and *Zizania* also contribute. Another interesting case was xanthan gum, a food additive derived from sugars fermented by the bacteria *Xanthomonas campestris*<sup>18</sup>. Despite the

differences in food origin, xanthan gum was clustered with nuts and seeds. When compared to walnuts, 33 shared metabolites were identified, all at very similar relative abundances. Based on compound-level assay evidence from the harmonized ChEMBL-derived bioactivity layer, these shared metabolites are also associated with similar documented bioactivities (Fig. 2B).

### Retention-index prediction substantially increases putative structural coverage of the metabolome

Across 27,713 detected features, only 2038 (7.4%) possessed confirmed structural identities, leaving 25,675 unidentified features without database-resolved structures. To bridge this structural gap, we trained a Gradient Boosting regression model with the full 2432-dimensional descriptor set directly on PTFI-measured retention indices, achieving robust predictive accuracy on a held-out test set ( $R^2 = 0.95$ , MAE = 16.8 RI units, Pearson  $r = 0.97$ ; Supplementary Fig. 5). This native-platform model outperformed transfer learning from external reversed-phase databases ( $R^2 = 0.89$ ), consistent with the benefit of eliminating chromatographic domain shift when sufficient platform-specific training data are available.

Applying the model to PubChem enabled formula-constrained nearest-neighbor matching in RI space for all 25,675 previously unidentified features. This pipeline yielded 15,593 High-confidence assignments, 4046 Medium-confidence assignments, and 6,036 Low-confidence putative annotations (Supplementary Data 1). These RI-guided assignments provide scalable, structure-resolved identifiers suitable for downstream aggregation (e.g., enabling reliable linkage to ChEMBL bioactivity data and improving the accuracy of food-level bioactivity predictions), while remaining explicitly probabilistic and requiring orthogonal confirmation (e.g., MS/MS or standards) for definitive identification. In practical terms, this step reduces a largely unannotated LC-MS feature space into confidence-tiered candidates that can be ranked for biological follow-up.

Quantitative estimation anchored to FoodAtlas values further allowed conversion of relative LC-MS intensities into biologically meaningful molar concentrations for downstream modeling. Because concentration harmonization is a critical source of uncertainty, we explicitly characterized the conversion-factor (CF) step (Supplementary Fig. 6).

The resulting conversion factors span roughly seven orders of magnitude, which motivates modeling them on a log scale. Under leave-one-compound-out cross-validation, the Tanimoto-weighted k-nearest-neighbor imputation achieved an RMSE of 1.26 on the log10 scale, corresponding to an approximately 18-fold mean error factor. Critically, this error is not uniform across chemical space: it is concentrated among compounds that are structurally distant from the polyphenol-enriched anchor set, whereas for compounds possessing a close structural analog among the anchors (Tanimoto similarity  $\geq 0.5$ ) the error falls to roughly an 11-fold factor (RMSE = 1.06). A Naringenin multi-food metabolite example (Supplementary Table 1) in which grapefruit and lime cell-level CFs differ by 1.81-fold, illustrates the compound-level median CF assumption for a stable anchor. Thus, the CF layer supports approximate concentration-aware ranking but should not be interpreted as validated absolute quantification.

To assess whether CF uncertainty propagated into the downstream bioactivity layer, Monte Carlo perturbation of log10CF values showed moderate stability for food-level FRAP rankings (median Spearman  $\rho = 0.90$ , across 100 iterations) and stronger stability for SHAP-ranked metabolite candidates (median Spearman  $\rho = 0.94$ ). These results support the use of the BPM as a prioritization layer while reinforcing that food ranking is more sensitive to CF uncertainty than candidate nomination (Supplementary Fig. 6).

### Full metabolomics resolution improves machine learning prediction of whole-food antioxidant activity

We constructed a Bioactivity Prediction Model (BPM) (Supplementary Fig. 7) by integrating the complete PTFI (27,713) and FoodAtlas (3,610) small-molecule profiles, yielding 30,343 small-molecule and elemental

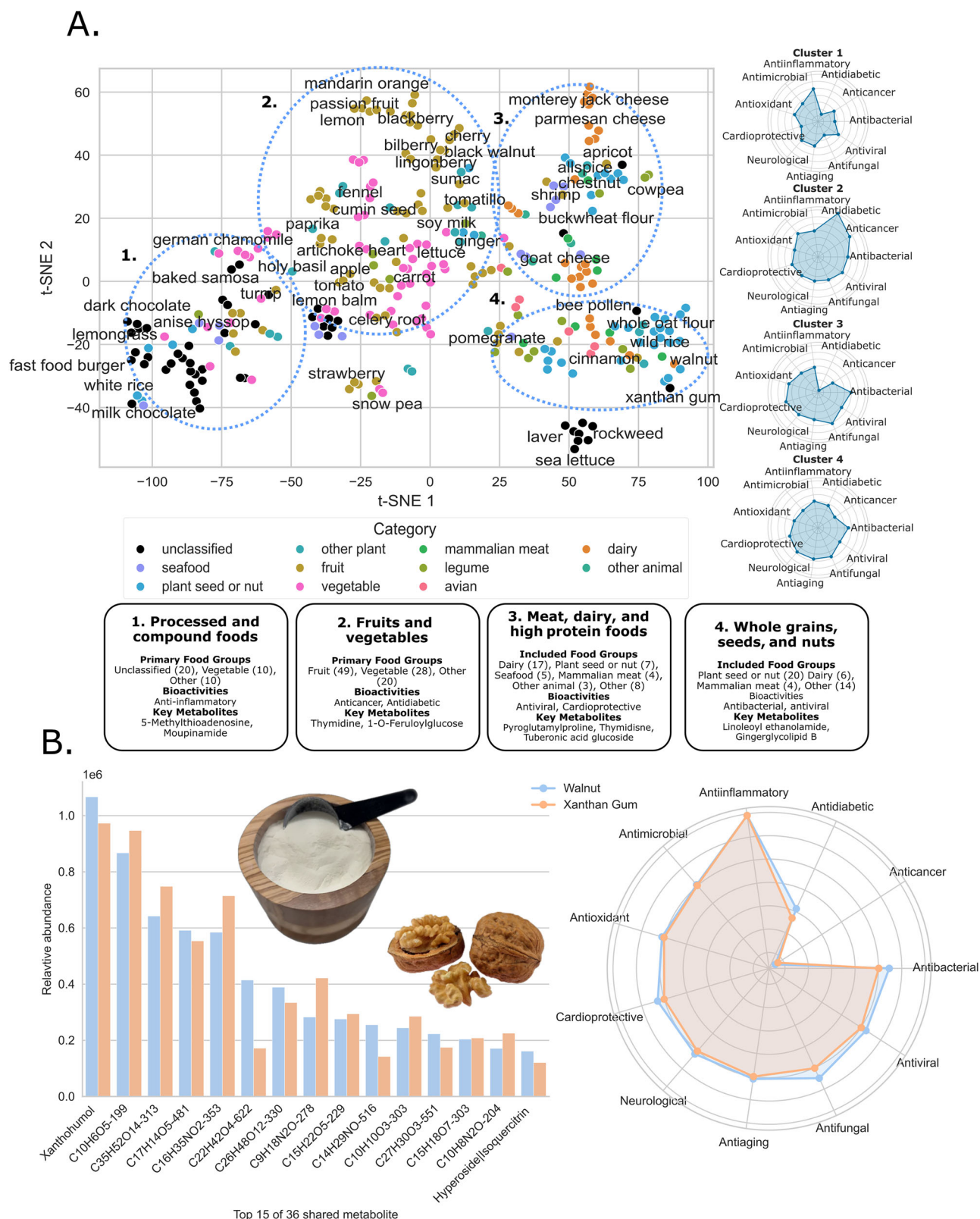


**Fig. 2 | Food metabolomics composition map.** **A** A t-SNE projection mapping the metabolomic composition of various foods, color-coded by food category. Clusters are strongly associated with food categories (Chi-squared = 4.86<sup>-37</sup>). Corresponding radar charts (right) display the aggregate bioactivity profiles for each cluster, highlighting group-level functional trends, while descriptive panels detail the defining

primary food groups, predicted bioactivities, and key driving metabolites. **B** A targeted comparative analysis between a whole food (Walnut) and a common food additive (Xanthan Gum). The bar graph (left) quantifies the relative abundance of their top 15 shared metabolites, and the accompanying radar chart (right) shows overall bioactivity profiles.

Table 1 | Compositional and metabolite-evidence signatures of food clusters

| Cluster (ID)                        | Representative foods         | Key Metabolites                                                       | Leading Bioactivity effect  | References |
|-------------------------------------|------------------------------|-----------------------------------------------------------------------|-----------------------------|------------|
| Processed and compound foods        | Burger, Baked Samosa, Turnip | 5-Methylthioadenosine, Moupinamide                                    | Anti-inflammatory           | 25–27      |
| Fruits and vegetables               | Lemon, Cherry, Soy, Ginger   | 1-O-Feruloyl glucose, (S)-Absciscic acid, Thymidine                   | Anticancer, Antidiabetic    | 28–30      |
| Meat, dairy, and high protein foods | Shrimp, Goat cheese          | Palmitoylethanolamide, uberonic acid glucoside, Stearoyl ethanolamide | Antiviral, Cardioprotective | 31–34      |
| Whole grains, seeds, nuts           | Walnut, Xanthan gum          | Gingerglycolipid B, Xanthohumol, L-Tryptophan, Linoleoyl ethanolamide | Antibacterial, Antiviral    | 22–24      |

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

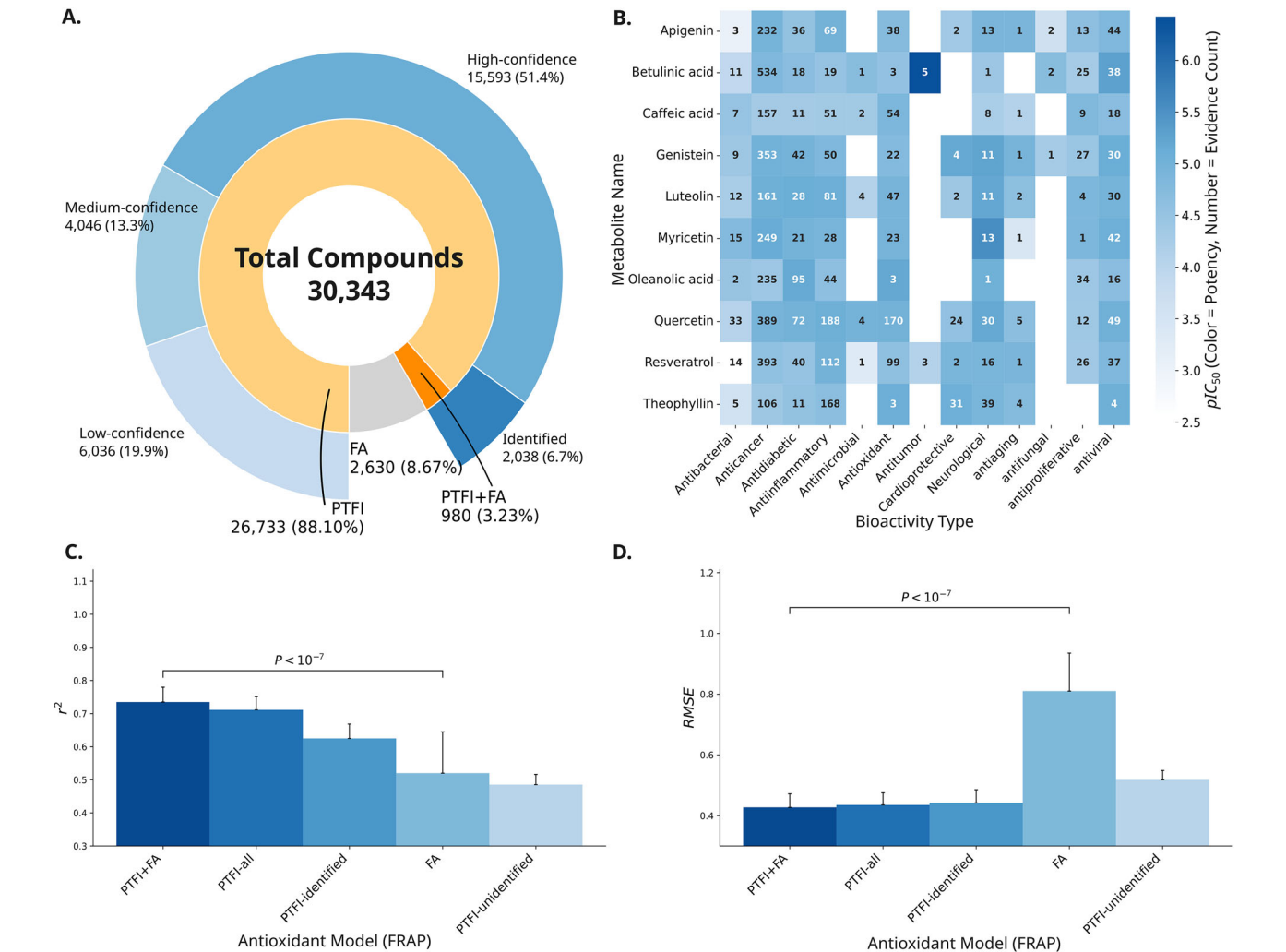


**Fig. 3 | Compound coverage, metabolite-evidence landscape, and predictive model performance.** **A** Composition of the small-molecule dataset used in the workflow, showing contributions from PTFI (27,713), FoodAtlas (3610), and their overlap (980), together with annotation-confidence strata for identified and RI-guided structurally narrowed features. **B** Aggregate compound-level bioactivity evidence for representative well-characterized metabolites, summarizing potency (color) and evidence count (number) across curated assay categories. **C, D** Predictive performance ( $R^2$ ) and Root Mean Square Error (RMSE) of the Random Forest FRAP antioxidant model evaluated across different feature subsets. The combined integration of datasets (PTFI + FA) achieves the highest correlation, significantly outperforming isolated or unidentified subsets ( $P < 10^{-7}$ ).

features across 322 foods, with whole-food FRAP measurements as the supervised antioxidant endpoint. Using only the 3610 small molecules with experimentally measured absolute concentrations from FoodAtlas produced suboptimal performance ( $R^2 = 0.52$ ). In contrast, using all small molecules (FoodAtlas and PTFI, the full 30,343 high-resolution profile, including the additional small-molecule whose concentrations were imputed via structural similarity) gave the best results (Fig. 3). This higher-resolution approach provides a broader compositional representation of whole-food antioxidant prediction by capturing subtle contributions from trace, untargeted, and structurally imputed compounds that would otherwise be missed.

Among the eleven machine learning algorithms evaluated, the Random Forest regressor demonstrated superior performance (Supplementary Fig. 8), likely due to its ability to model the complex, non-linear interactions

—both synergistic and antagonistic—among the myriad compounds present in the food matrix. Under repeated random subsampling validation (100 bootstrap iterations, 80/20 train/test split, 5-fold CV), the model explained 73% of the variance in experimental FRAP values ( $R^2 = 0.73$ ,  $\text{RMSE} = 0.43$ ) with a Pearson correlation coefficient of 0.85 ( $p = 2.7 \times 10^{-15}$ ) between predicted and observed antioxidant capacity. Residual analysis showed no systematic bias ( $r = 0.02$ ) (Supplementary Fig. 9). A preliminary model relying solely on the sum of all food small-molecule activity loads, without machine learning, was uninformative (Supplementary Fig. 10). Beyond reproducing known FRAP measurements, the BPM generalizes well to foods lacking experimental data. Figure 4 presents a food bioactivity wheel highlighting the top foods across multiple categories, including the highest-ranked foods for predicted antioxidant capacity, demonstrating the framework's utility for rapid in silico screening of antioxidant potential across the food landscape.

### Chemical bioactivities are log-normally distributed

The distribution of bioactivity scores per chemical is generally best fit by a log-normal distribution. Of 13 bioactivities tested, 11 were best fit by a log-normal distribution, albeit with the outlying 2 only containing 31 and 10 samples, respectively. Candidate distributions included log-normal, gamma, Weibull, truncated normal, and uniform. The Kolmogorov-Smirnov (K-S) test was used to determine the best fit. Of the bioactivities containing 300 or more samples, K-S distance was less than 0.1 in all cases, and of those containing 100 or more samples, K-S distance was 0.15 or less. Furthermore, the standard deviations and means of each bioactivity occupied tight ranges (Supplementary Fig. 3).

### Structure-based triage of food-derived candidates against the KEAP1 sensor

For the subset of prioritized metabolites with structure-level representations, we applied Boltz-2 against the KEAP1 Kelch domain as a secondary mechanistic filter following RI-guided structural narrowing and model-based ranking to generate putative target engagement hypotheses (Fig. 5). The model successfully differentiated between baseline controls. The positive control, Pterostilbene, achieved high binder probability (0.82) and structural confidence (0.87), with a favorable predicted affinity-like score ( $bIC_{50} = 1.4$ ). The negative control, Glucose, was accurately classified as inactive, exhibiting a low binder probability (0.14), a  $bIC_{50}$  of 3.4, and a structural confidence of 0.86. When evaluating food-derived metabolites that lacked prior KEAP1 antioxidant-related assay records, Quercetin—selected by the predictive model as an important feature—demonstrated favorable binding metrics with an affinity probability of 0.66, a  $bIC_{50}$  of 1.5, and a confidence score of 0.87. Conversely, Oleic Acid, which the model identified as a non-important feature, yielded an inactive probability of 0.28, a  $bIC_{50}$  of 2.7, and a confidence score of 0.88.

Importantly, this filtering step was also useful for previously unresolved metabolites carried forward from the RI-guided narrowing stage. For example, feature PTF42551 (formula  $C_{20}H_{30}O_3$ , RI = 181)—narrowed by the RI-guided stage to candidate CID 163747858 (InChIKey LNKXPOSAFDLORF-UHFFFAOYSA-N)—produced a binder probability of 0.76,  $bIC_{50} = 1.62$ , and structural confidence of 0.86, despite lacking prior KEAP1- or NRF2-related assay records in our curated evidence set. In the context of the full workflow, this illustrates the practical role of the structure-based layer: not to prove mechanism, but to reduce a large set of compositionally important and structurally plausible candidates into a smaller set of molecules that are more compelling for targeted experimental follow-up.

### A multi-layer prioritization funnel reduces untargeted metabolite features to a shortlist for mechanistic testing

Taken together, the structural, assay-curation, and model-ranking layers define a practical prioritization funnel from untargeted metabolomics to testable hypotheses. Of the 27,713 PTFI detected features, only 2038 possessed confirmed structural identities, whereas 25,675 initially lacked database-resolved structures. RI-guided matching expanded this space,

enabling provisional linkage of a much larger fraction of the metabolome to compound-level chemical space. These structurally narrowed candidates were then contextualized using thousands of LLM-harmonized ChEMBL assays and ranked using the antioxidant scoring and model-attribution framework. The global and local impact of individual small-molecule and elemental features on model predictions is visualized through SHAP (SHapley Additive exPlanations) values<sup>19</sup> (Supplementary Fig. 11). Examining the resulting prioritization list directly, the highest-ranked features include mechanistically plausible antioxidants such as quercetin-related and coumaric-acid-derived features, consistent with their well-established redox activity. Other high-ranking features, however, are less intuitive: these include simple sugar and elemental/mineral-associated features such as rhamnose, xylose, phosphorus, calcium, manganese, sodium, and barium from the integrated FoodAtlas/PTFI composition matrix. We therefore interpret SHAP importance as predictive association within the food matrix, not as evidence that every top-ranked feature is a direct chemical antioxidant. Elemental and mineral-associated signals were retained in the composition-level model because they are part of the feature table, but only structurally plausible organic compounds with valid ligand representations were advanced to the KEAP1/Boltz-2 mechanistic triage layer. The final structure-based KEAP1 analysis serves as an additional refinement step for shortlisted organic candidates, illustrating how a large untargeted food feature space can be reduced to a tractable set of functional molecules for targeted wet-lab testing.

### Integrated evidence prioritizes anise hyssop as an illustrative follow-up candidate

Integrating food-level predicted antioxidant potential with compound-level evidence mining from ChEMBL highlights a small set of foods that are consistently high-ranking and mechanistically interpretable. Notably, *Agastache foeniculum* (anise hyssop)—a plant native to North America that has been studied for its antioxidative, anticancer, and anti-inflammatory properties<sup>20,21</sup>—emerged as the illustrative high-ranking example. Anise hyssop achieved the highest overall summed bioactivity scores, as well as the highest specific anti-inflammatory bioactivity score of all 322 foods measured (Supplementary Fig. 4). Among more common foods, tomato had the highest compound-evidence score for anticancer-associated assays, and carrots ranked highly for antidiabetic and anticancer-associated compound-evidence layers (Fig. 4). These rankings are not interpreted as clinical efficacy claims; rather, they identify foods and chemical constellations that are suitable for targeted follow-up. This combined composition→prediction→evidence framework effectively prioritizes complex foods for follow-up experimentation and supports data-driven functional ingredient discovery.

### Discussion

The primary contribution of this study is a scalable prioritization framework that converts untargeted food metabolomics into testable candidate molecules. The robustness of this framework relies on both the standardized metabolomics protocols that ensure analytical consistency across samples, and the computational AI methods employed in FoodAtlas for factual correctness, thereby enabling reliable cross-food comparisons and model generalizability. The results show that standardized PTFI profiles, RI-guided structural narrowing, assay harmonization, concentration-aware mapping, and model-guided ranking can be combined to move from sparse LC-MS features to interpretable hypotheses. Known examples such as anise hyssop and quercetin function as internal sanity checks, whereas previously uncharacterized features such as PTF42551 illustrate the framework's ability to surface plausible follow-up candidates that would otherwise remain buried within untargeted LC-MS data.

A striking feature of the dataset is its extraordinary molecular diversity across foods. Of the 27,713 detected features, 44% were unique to a single food, and 85% were present in five foods or fewer, underscoring the highly sparse and heterogeneous nature of the food metabolome. This indicates that much of food's molecular diversity is not shared across the food supply but instead resides in food-specific molecular signatures. As a result,



**Fig. 4 | Food bioactivity wheel.** The top foods for each bioactivity are displayed, with the category of each food colored. A simple visual guide for selecting a food with a desired bioactivity. Top 10 foods are displayed for each of 11 bioactivity categories (Antiviral, Antioxidant, Antiaging, Anticancer, Antifungal, Antidiabetic, Neuroprotective, Antimicrobial, Anti-inflammatory, Antibacterial, and Cardioprotective), providing a visual guide for selecting foods with a desired functional profile. The inner wedge for each bioactivity shows representative example foods, while the outer ring lists the top 10 ranked foods ordered clockwise from highest to lowest, with each food colored by its FoodOn category (legend). Antioxidant rankings were generated using the FRAP whole-food Bioactivity Prediction Model. All remaining bioactivity categories were ranked by the compound-evidence method. Foods appearing across multiple wedges (e.g., *anise hyssop*, *bilberry*, *lingonberry*, *paprika*) represent candidates with broad multi-bioactivity potential, whereas category-specific top entries (e.g., *tomato* for anticancer, *lemon balm* for antioxidant) indicate more targeted functional profiles.

approaches based only on a limited set of common compounds or aggregated nutrient profiles are inherently incomplete, although our own concentration-estimation layer still depends on a limited anchor set and should be interpreted as semi-quantitative. Retention index-guided modeling therefore plays an important role by transforming a largely opaque

feature space into a structured, rankable set of provisional candidates. Although these assignments remain probabilistic, they provide a practical bridge between unresolved LC-MS features and external bioactivity knowledge, enabling downstream prioritization and hypothesis generation at scale. The superior performance of the native-platform RI model relative

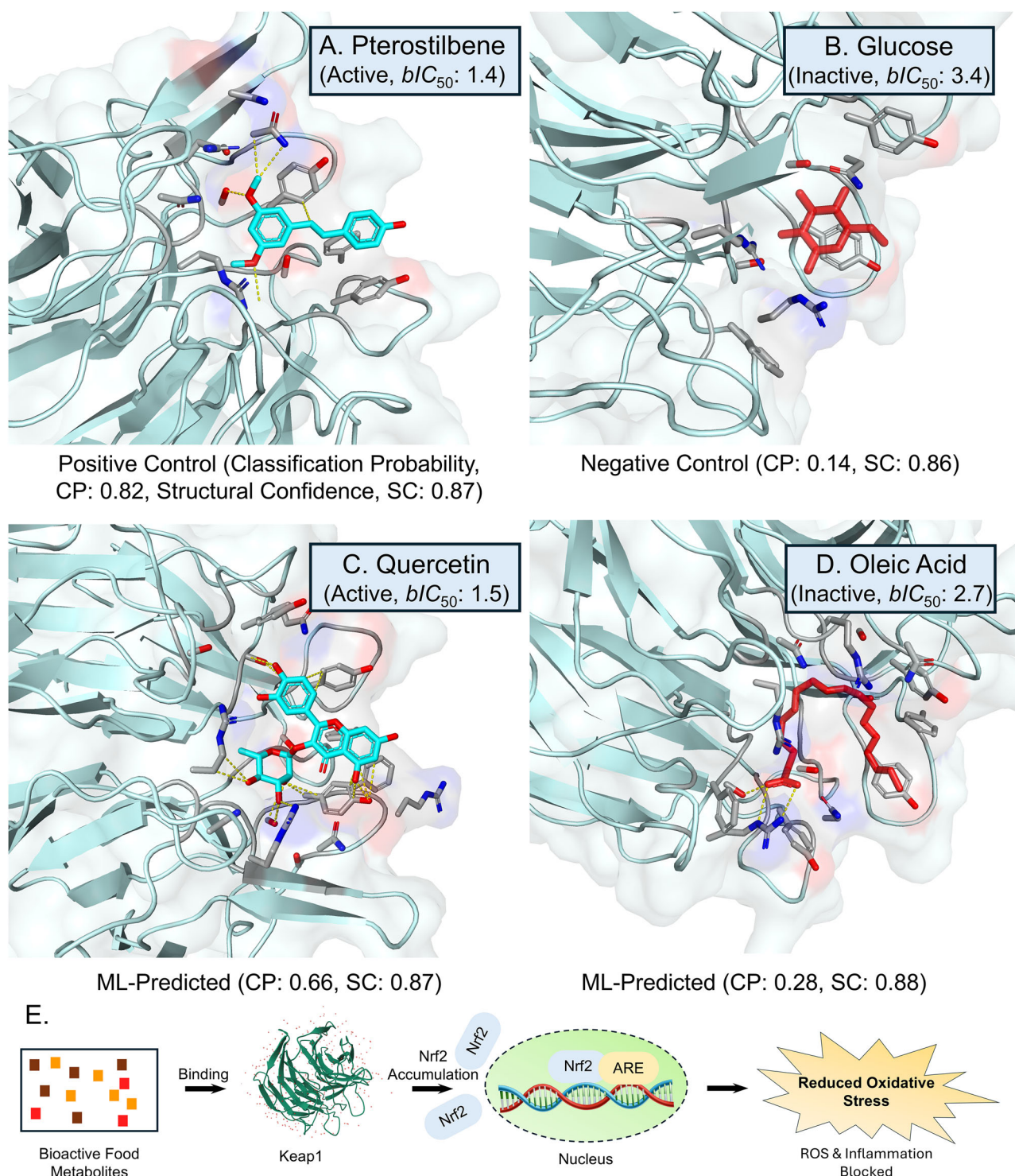


**Fig. 5 | Structural evaluation of model-prioritized metabolites against the KEAP1, an NRF2-pathway redox-regulatory protein.** A–D Boltz-2 predictions for representative controls and model-prioritized metabolites evaluated against the KEAP1 Kelch domain (rendered using PyMOL<sup>71</sup>). Pterostilbene and glucose were used as positive and negative reference compounds, respectively, while quercetin and oleic acid illustrate examples of metabolites ranked as more- and less-relevant by the composition-based prioritization workflow. Reported classification probabilities

and structural confidence values summarize the structure-based ranking output. **E** Schematic of the hypothesized antioxidant-related mechanism used for prioritization, in which food-derived metabolites may engage KEAP1 (PDB ID: 4IFJ; PDB DOI: 10.2210/pdb4ifj/pdb) and thereby support Nrf2-pathway activation. The subsequent accumulation and nuclear translocation of Nrf2 activate the Antioxidant Response Element (ARE), ultimately reducing cellular oxidative stress and blocking downstream inflammation.

to transfer learning further emphasizes the value of standardized, system-specific metabolomics data for large-scale inference.

Metabolite-based clustering both validates and challenges conventional food classification systems. Representative metabolites contributing

to the cluster-level compound-evidence signatures, together with supporting literature<sup>22–34</sup>, are summarized in Table 1. While broad groupings such as fruits and vegetables, dairy, and seeds or nuts were largely preserved, several notable deviations reveal that molecular composition, and by extension

functional potential, can diverge substantially from traditional culinary or taxonomic labels. For example, white rice (*Oryza sativa*) clustered with processed foods, reflecting a markedly reduced and altered metabolite profile relative to wild rice (*Zizania*), which retained substantially greater molecular diversity. While these grains share a common family (Poaceae), their divergence in both species and processing highlights how molecular composition can vary independently of broad crop classification. Conversely, xanthan gum, despite its microbial origin, clustered with nuts and seeds due to shared compositional and bioactivity profiles, illustrating that compositional convergence can arise across fundamentally different food types. In parallel, the detection of compounds associated with food-contact materials points to the measurable influence of processing environments and supply chains on food chemistry. Together, these findings challenge the assumption that food categories are reliable proxies for nutritional or functional equivalence and instead highlight molecular composition as a more precise and actionable determinant of food quality and biological activity.

A key finding is that whole-food bioactivity cannot be inferred from a limited subset of known compounds or simple additive models. Incorporating the full metabolomic profile—including trace, untargeted, and structurally inferred small molecules—substantially improved predictive performance, indicating that—within our dataset—these often-overlooked features collectively contribute<sup>8</sup> to predicted functional outcomes. This is consistent with prior reports that minor, trace, and matrix-associated constituents can meaningfully shape the measured bioactivity of whole foods<sup>3</sup>. The superior performance of nonlinear models further indicates that synergistic and antagonistic interactions within the food matrix are critical determinants of bioactivity. Together, these results support a shift away from reductionist, single-compound frameworks toward systems-level models that capture food as a complex chemical network.

Applying this integrated framework across the food landscape highlights its utility for discovery. The identification of anise hyssop (*Agastache foeniculum*) as a high-ranking candidate illustrates how combining composition, prediction, and evidence layers can surface foods with strong and multi-dimensional bioactivity profiles. Notably, such foods may be under-represented in conventional diets despite their functional potential, pointing to opportunities for expanding dietary diversity and identifying novel functional ingredients.

More broadly, understanding how foods affect our health is important not only for our current diet, but also for planning for a changing climate and agricultural landscape. Despite thousands of edible species, only 10–50 crops are responsible for about 95% of calories consumed worldwide<sup>35</sup>. Anise hyssop, as mentioned previously, is high in several bioactivities, serves as an excellent pollinator, and is native to North America<sup>36</sup>. Lemon balm, a hardy, fast-growing crop that is invasive in many places, topped the list of antioxidant foods<sup>37,38</sup>. Identifying functional foods like these is key to increasing biodiversity while prioritizing human health.

At the same time, profiling the bioactivities of common, well-known food items allows for more informed decisions for the everyday consumer and serves as the immediate entry point for public health. The tomato and carrot rankings reported in the Results are consistent with literature linking them with cancer protection<sup>39,40</sup>, tomatoes usually because of their lycopene content; however, tomatoes contain several potentially cancer-protective metabolites, including phlorizin, sphinganine, and more<sup>41,42</sup>. Likewise carrots were found to have high levels of digoxigenin, a chemical that has been linked with anticancer properties<sup>43</sup>. These observations reinforce the idea that the health effects of foods may arise from broader metabolite constellations rather than from a single canonical compound alone.

To characterize antioxidant-related bioactivity more comprehensively, we used a dual-modeling strategy that separates whole-food redox capacity from single-compound mechanistic prioritization. A machine learning model trained on whole-food FRAP measurements was used to predict ferric-reducing antioxidant capacity from food metabolomic profiles. Because FRAP is a single-electron-transfer assay, high FRAP values indicate greater reducing power under assay conditions and should not be

interpreted as direct in vivo antioxidant efficacy<sup>44</sup>. However, cellular redox homeostasis also depends heavily on endogenous defense systems. We therefore complemented the FRAP-based model with structure-based triage using Boltz-2 against KEAP1, the principal negative regulator of the Nrf2 antioxidant response pathway<sup>45</sup>. These two readouts answer different questions: FRAP provides a scalable bulk readout for complex food extracts, whereas KEAP1/Nrf2 engagement provides a compound-level mechanistic filter for shortlisted candidates. Thus, the KEAP1 analysis is not intended to validate whole-food antioxidant activity, but to prioritize structurally plausible compounds that may warrant targeted experimental follow-up. Integrating these mechanistic predictions into a unified knowledge graph linking foodomics and chemical-disease/gene networks<sup>1</sup> creates a practical route for rapid, scalable discovery. As more complete structural annotations and biological data become available, this framework can seamlessly scale to multi-endpoint models (e.g., anti-inflammatory or antidiabetic) and extend to whole-food extracts to better account for exposure, matrix effects, and compound co-occurrence.

Several limitations should be noted. First, the whole-food bioactivity model relies solely on FRAP and is therefore assay-specific as it captures only a single mechanism of antioxidant action. Combining complementary assays, such as DPPH, ABTS, or ORAC, would probe distinct mechanisms and improve prediction robustness. Second, RI-guided structural proxies remain probabilistic, especially for isomers and stereoisomers, and definitive compound identification will require orthogonal MS/MS matching and/or authentic standards. MS/MS was not used here because the dataset available for this analysis contained processed MS1 feature tables, RI, molecular formulas, and authentic-standard matches, but did not include redistributable MS/MS spectra for unidentified features under the applicable data-transfer and licensing terms. This is a dataset limitation rather than a methodological preference, as MS/MS would provide fragment-level, sub-structural, and compound-class evidence and would improve discrimination among candidate structures. Third, the concentration-estimation step is a major bottleneck. Conversion factors were inferred from FoodAtlas anchors rather than in-sample absolute quantification, then transferred to additional compounds by structural similarity. This introduces uncertainty from matrix-dependent ion suppression or enhancement, uneven chemical coverage of the anchor set, imperfect transferability of Tanimoto similarity, and annotation uncertainty for lower-confidence RI-based targets. While part of this error is expected given the extreme dynamic range of electrospray ionization efficiencies across metabolites, these values must be interpreted strictly as order-of-magnitude estimates rather than validated absolute quantifications. Although the downstream whole-food FRAP model is trained and evaluated against measured FRAP values and is therefore more robust than any individual imputed concentration, concentration-dependent compound-level interpretations should be treated cautiously. Finally, the framework does not yet account for digestion, bioavailability, metabolism, in vivo matrix effects, or inter-individual variability. The shortlisted foods and compounds should therefore be interpreted as preliminary, hypothesis-generating candidates<sup>46</sup> for targeted biochemical assays, cell-based testing, MS/MS or standard-based confirmation, and ultimately controlled intervention studies, rather than as definitive evidence of chemical identity, exposure, physiological activity, or clinical benefit.

These constraints also highlight a broader limitation of extending traditional nutrient-based frameworks, such as Recommended Dietary Allowances (RDAs), to the food metabolome. RDAs are designed around single nutrients with defined deficiency endpoints and largely linear dose-response relationships, whereas the metabolomic landscape revealed here is characterized by high dimensionality, extreme sparsity, and nonlinear interactions across thousands of co-occurring compounds. The observation that most metabolites are food-specific and that whole-food bioactivity emerges from complex, non-additive interactions indicates that reductionist, compound-by-compound targets are unlikely to capture the functional properties of foods. Instead, these results support a shift toward modeling molecular patterns, compositional signatures that reflect the

combined and context-dependent effects of multiple metabolites, as the basis for evaluating food quality and health relevance. Moving beyond RDAs will require the development of data-driven frameworks that link these molecular patterns to biological responses, initially through standardized in vitro and ex vivo assays, and ultimately through integration with human cohort data, controlled dietary interventions, and biomarker-based validation. In this context, the present framework provides a foundation for defining and testing such pattern-based models, enabling a transition from static nutrient thresholds to dynamic, biologically grounded representations of food function.

Looking forward, this framework can be substantially strengthened and expanded in several directions. Integrating tandem mass spectrometry fragmentation patterns with the current RI-guided annotation pipeline will markedly improve structural confidence—especially as MS/MS data become available—by providing fragmentation evidence that helps resolve isomeric ambiguity across thousands of dark-matter features. Once matched whole-food assay datasets become available for other therapeutic categories, the same non-linear modeling approach can be extended to other bioactivities. Systematic in vivo validation, bioavailability studies and controlled human intervention trials will be essential to bridge the gap between predicted and physiological effects. Incorporating gut microbiome interactions, multi-omics layers and individual response variability could enable a shift toward precision functional nutrition. In that sense, the framework provides not only a set of candidate foods and compounds but also catalyzes both basic research into food-health relationships and applied efforts to design next-generation functional foods that deliver measurable biological benefits. More fundamentally, this work reframes how food bioactivity can be studied at scale: rather than attempting to resolve every metabolite or relying on predefined nutrient frameworks, the approach leverages partial structural information, probabilistic annotation, and integrated evidence layers to generate actionable hypotheses from incomplete data. This represents a shift from a completeness-driven paradigm to an inference-driven paradigm, where biological insight can be extracted even in the presence of substantial uncertainty.

Taken together, this study demonstrates that translating food metabolomics into functional insight requires integration across structural inference, bioactivity evidence, and systems-level modeling. By linking these layers into a unified prioritization framework, this work provides a scalable route from high-dimensional food composition data to biologically interpretable and experimentally testable hypotheses, with potential applications in functional ingredient discovery, crop selection, dietary design, and precision nutrition.

## Methods

### Metabolomics data

The untargeted metabolomics dataset analyzed in this study was accessed through The Periodic Table of Food Initiative (PTFI) database (<https://foodperiodictable.org>) via a data-transfer agreement from the PTFI Secretariat at the American Heart Association. To make the upstream data-processing steps explicit, the PTFI/Odenkirk<sup>9</sup> workflow used lyophilized and homogenized composite food specimens, solid-phase extraction, reverse-phase LC, positive-mode electrospray HRMS, and precursor-only MS1 acquisition. The standardized method was evaluated across three laboratories using commercially available time-of-flight platforms: Waters Xevo G2-XS qTOF (Lab A), Agilent 6545 qTOF (Lab B), and Agilent 6560 IM-qTOF (Lab C). Raw centroided, mass-calibrated files were converted to mzML; feature detection was performed with XCMS/centWave after instrument-specific optimization of peak-width, ppm-error, and m/z-difference parameters. Retention times were converted to RI values with a second-order polynomial fit to the 33-compound internal retention time standard (IRTS) mixture using a Tukey M-estimator. This standardization reduced inter-laboratory retention variation from more than 20 s to less than 5 RI units in the method-development study and supported the cross-food comparability used here<sup>46–48</sup>.

For formula annotations, the upstream PTFI Metabolomics v1 pipeline matched experimental features to a consensus reference library assembled from chemical databases under defined elemental and adduct constraints. Redundant isotope and common-adduct signals were removed using IsoSpecPy. Molecular-formula confidence was evaluated using a composite score that incorporated mass difference, RI difference, isotopic-pattern similarity, adduct preference, and monoisotopic feature intensity, with a 15 ppm mass tolerance; library inclusion also required observation in two independent laboratories<sup>9</sup>. Thus, mass error and isotopic-pattern fit were used upstream for molecular-formula annotation quality control, whereas our downstream analysis began from the processed PTFI feature table.

Food samples in the PTFI database are accompanied by structured metadata on the source organism, production method, and food classification based on the FoodOn ontology. The PTFI release contains 500 metabolomics samples covering 322 unique food items; the remaining samples correspond to biological or technical replicates of the same food. To avoid inflation of per-food metrics (e.g., over-counting metabolites in foods represented by multiple samples, or biasing cluster membership toward foods with more replicates), all downstream analyses were conducted at the unique-food level ( $n = 322$ ), with replicate samples aggregated by median intensity per metabolite prior to analysis. Feature detection, retention-index calibration, and molecular formula assignment were performed upstream using the PTFI data-processing pipeline, and the resulting processed feature table served as the input for all downstream analyses in this study. Both the PTFI dataset and the analysis code used in this study are publicly accessible (see Code Availability).

### Retention-index guided structural assignment for unidentified metabolites

Following Metabolomics Standards Initiative guidelines<sup>49</sup>, the 2,038 features matched to authentic reference standards correspond to MSI Level 1 (confirmed identification). To expand structural annotation beyond the 2,038 identified metabolites with confirmed InChIKeys, we built a chromatography-specific retention prediction model<sup>50</sup> that learns the relationship between molecular structure and chromatographic retention directly from PTFI platform measurements. The RI-guided putative annotations correspond to MSI Level 3 (formula-level putative annotation with retention-based structural narrowing) and are represented by Formula-RI identifiers (e.g., C9H8O3-181), in which the molecular formula is assigned by the PTFI adduct-aware mass/RI annotation framework and the RI denotes normalized chromatographic retention on the PTFI platform. A supervised regression model was trained natively on the PTFI-identified compounds with confirmed SMILES and measured RI values. Molecular structures were encoded as a combined feature vector of 2432 descriptors: 217 two-dimensional physicochemical and topological descriptors<sup>51,52</sup>, (e.g., logP proxies, H-bond counts, polar surface area, ring metrics) computed via RDKit, 2048-bit Morgan circular fingerprints (ECFP4, radius 2), and 167 MACCS structural keys. Eight model families were systematically benchmarked across three feature set configurations on an 80:20 held-out test split, 5-fold cross-validation, and 25 bootstraps (Supplementary Fig. 5). Transfer learning from external reversed-phase databases was evaluated as a comparison strategy and is described in Supplementary Note 1.

For each PTFI metabolite feature lacking an InChIKey, we assembled a formula-constrained candidate set by retrieving 11,770,660 PubChem compounds with available structures whose elemental formula matched the feature's assigned molecular formula (Supplementary Note 1). For each candidate structure, we predicted RI in silico and ranked candidates by the absolute residual between observed and predicted retention ( $|\Delta\text{RI}|$ , expressed as RT-equivalent seconds under the PTFI calibration). The top-ranked candidate was reported as a putative annotation, and matches were stratified into confidence tiers based on retention agreement: High (1) if  $|\Delta\text{RT}| < 5$  s; Medium (2) if 5–25 s; and Low (3) otherwise. These confidence tiers formed the first screening layer used to reduce untargeted LC-MS features to a tractable candidate space for downstream bioactivity ranking.

Because retention agreement and molecular formula alone cannot uniquely resolve structural isomers (and cannot resolve stereochemistry), these annotations are treated as uncertainty-tagged structural proxies intended for scalable aggregation and hypothesis generation, with definitive identification requiring orthogonal confirmation (e.g., MS/MS matching or authentic standards).

### Dataset integration: unified food and metabolite concentration matrix

Downstream analyses operate on a unified  $322 \times 30,343$  food  $\times$  small-molecule matrix (includes endogenous metabolites as well as processing-derived, exogenous, and food-contact-associated small molecules) in which every populated cell is expressed as molar concentration (mol/L). The matrix combines two sources of distinct provenance: PTFI, providing a  $322 \times 27,713$  matrix of LC-MS relative abundances on a standardized platform, and FoodAtlas, a literature-mined knowledge graph providing per-compound, per-food absolute concentrations (mg/100 g). The two sources share 980 compounds by exact InChIKey match, which serve as empirical anchors for unit harmonization. Each cell is populated according to the compound's provenance class: (i) for the 980 PTFI  $\cap$  FoodAtlas compounds, FoodAtlas mg/100 g is converted to mol/L using the exact molecular weight from canonical SMILES (RDKit) and treated as ground truth; (ii) for the 26,733 PTFI-only compounds, PTFI relative abundance is converted to mg/100 g via a compound-specific conversion factor (CF), defined as the empirical FoodAtlas-to-PTFI ratio across the 980 anchors and imputed for non-anchor compounds by Tanimoto-weighted k-nearest-neighbor averaging<sup>53</sup> ( $k = 5$ , 217 RDKit physicochemical descriptors<sup>51,52</sup>; see Supplementary Note 2), then converted to mol/L as in (i); (iii) for the 2630 FoodAtlas-only compounds, literature concentrations are converted directly to mol/L, with food-compound pairs absent from the literature treated as zero.

The CF imputation treats CF as approximately compound-specific; in reality, CFs also vary across food matrices because ion suppression and enhancement are matrix dependent. We therefore report the food-to-food spread of CFs together with a chemical applicability-domain analysis, defined as the distribution of Tanimoto similarities between each imputed compound and its nearest usable anchor, in Supplementary Note 2. The empirical anchors used to fit CFs were compounds with FoodAtlas matches and usable InChIKeys; 665 were authentic-standard MSI Level 1 identifications, and the full 980-anchor set included high-confidence annotations as detailed in Supplementary Note 2.

### Metabolomics-based food clustering

322 unique foods derived from the PTFI database and matched to the FoodAtlas Knowledgebase were clustered based on their metabolomes. Because the metabolite matrix was highly sparse, feature selection was performed prior to clustering. The elbow method was used to retain the 305 metabolites detected in at least 12% of foods (Supplementary Fig. 1). To identify clusters in the dataset, t-SNE was employed for dimensionality reduction and then HDBSCAN was used to identify clusters in the lower dimensional space<sup>54,55</sup> (Fig. 2). To analyze the meaningfulness of the identified clusters, the amount of each metabolite in each cluster was compared via t-test to the amount of that metabolite in foods not in the cluster. Furthermore, the representation of food categories per cluster was also analyzed by Chi-squared test and by comparison of Pearson residuals per cluster and food category (Supplementary Fig. 2).

### Metabolite bioactivity scores

For the metabolites with InChIKeys, 119,633 chemical-assay data were extracted from ChEMBL (v36)<sup>6</sup>, yielding 39,158 unique assays. To map heterogeneous assay descriptions to broad bioactivity evidence classes (e.g., inflammation-relevant assays targeting COX-2, NF- $\kappa$ B, cytokines) (Supplementary Table 2), we implemented a hierarchical semantic indexing strategy inspired by LLM-agent frameworks<sup>56,57</sup> (Supplementary Note 3). This two-stage workflow isolated relevance detection from label assignment,

filtering assays based on evidence spans and assigning therapeutic directionality. The input to the agent was bioassay details including assay description, standard activity type, value with unit, bioassay ontology annotations (BOA), assay type, target type, and target organism metadata<sup>58</sup>. Only assays with explicit quantitative evidence were retained, and the agent was given the option to return an uncertain label when the available information was insufficient for confident classification. Cross-model (OpenAI ChatGPT-5.2-high and Google Gemini 3.1 Pro, accessed February 2026) agreement was assessed using Jaccard similarity ( $J = 0.92$ ), and a manual audit of 100 assays (randomly sampled and high-volume assays) was used to validate the final labels<sup>59</sup> (Supplementary Table 3). Importantly, these classes were used as semantic evidence labels rather than as claims that all assays within a class measured the same biochemical endpoint. For example, antioxidant-labeled assays could include radical-scavenging, reducing-power, oxidative-stress-protection, or pathway-related assays, but the underlying endpoint descriptions and quantitative records were retained at the assay-record level. Only assays with explicit quantitative evidence were retained, and uncertain or ambiguous labels were excluded. These curated assay labels were used as compound-level contextual evidence for prioritization rather than as direct food-level endpoints or supervised food-level prediction targets.

### From composition to bioactivity: concentration-potency mapping

Bioactivity data in ChEMBL is reported as  $IC_{50}$  values in molarity (mol/L), which are subsequently transformed into  $pIC_{50}$  units ( $-\log_{10}(IC_{50})$  or  $pChEMBL$ ). However, untargeted PTFI metabolomics and the FoodAtlas database typically report relative abundances and absolute concentrations in mass fractions ( $C_{mass}$ , mg/100 g) respectively. Calculating bioactivity directly from mass fractions using concentration-response curves<sup>60,61</sup> introduces a systematic error favoring low-molecular-weight compounds. To resolve this, all small-molecule concentrations were normalized to molarity using their exact molecular weights (MW, g/mol) via RDKit from SMILES strings (Supplementary Note 4).

For each individual metabolite we applied a bounded concentration-response<sup>61</sup> mapping using the Hill equation (sigmoid form) with Hill coefficient<sup>60</sup>  $h = 1$  to convert molar concentration and  $pIC_{50}$  potency into a fractional activity load bounded between 0 and 1 (Supplementary Note 4). For mixture baselines, we computed a conservative Concentration Addition / Hazard Index style estimate by aggregating activity loads across compounds (equivalent to summing the linear hazard quotients and taking the logarithm), yielding a “sum-of-parts” baseline that does not assume unbounded synergy<sup>62–64</sup>. This provides a biologically grounded, conservative estimate of a food's total potential bioactivity by capturing the additive cocktail effect of its components.

### Whole food bioactivity prediction model

To predict food-level antioxidant bioactivity from composition-derived metabolites, the model strictly treats different assays categorized under the same semantic label (e.g., FRAP, DPPH, ABTS, ORAC) as separate endpoints, rather than combining them into a single overarching variable. For this predictive task, the Ferric Reducing Antioxidant Power (FRAP) assay was selected as the primary, most readily available whole-food endpoint, using data collected from the literature<sup>65,66</sup>. FRAP quantifies ferric-reducing power and provides a practical proxy for direct chemical antioxidant capacity at the whole-food level. Food-level bioactivity is not strictly defined by chemical potency alone; rather, it is a complex function of potency, exposure, the food matrix, and the co-occurrence of multiple compounds.

A supervised machine learning approach was employed to predict whole-food bioactivity, utilizing empirical bioactivity scores as target labels. We implemented the primary Bioactivity Prediction Model as a Random Forest (RF) regressor. The RF algorithm was selected as the primary predictive engine because it empirically achieved the highest accuracy among the 11 models evaluated (Supplementary Fig. 8). It effectively captures the additive nature of complex chemical mixtures, models nonlinear effects, and

accounts for higher-order interactions among compositional features without requiring strong parametric assumptions. Model performance was evaluated using repeated random subsampling within the dataset. Across 100 bootstrap iterations, the dataset was iteratively split into 80% training and 20% testing subsets. Within the training sets, hyperparameters for all models were optimized using 5-fold cross-validation coupled with an exhaustive grid search (Supplementary Note 5). The antioxidant prediction model attribution was quantified using SHAP values<sup>19</sup>, and these feature-level rankings were subsequently used to nominate metabolites for structural follow-up and Boltz-2 analysis.

### Structure-based shortlist refinement using KEAP1 as a mechanistic case study

To transition from statistical prediction to mechanistic hypothesis, we prioritized metabolites with high model attribution (via SHAP values) that lacked prior bioactivity-related assay records in ChEMBL. To connect composition-based antioxidant prediction to plausible molecular mechanisms, we performed a targeted structure-based analysis against the KEAP1 Kelch domain (UniProt: Q14145), a sensor/repressor in the NRF2 pathway<sup>12,13,67</sup>. We applied Boltz-2<sup>68,69</sup> to predict protein-ligand complexes and associated affinity outputs for shortlisted candidates. Boltz-2 provides two relevant outputs: a binary affinity probability designed for binder-versus-decoy discrimination and an affinity-like value (affinity\_pred\_value), which we report as  $bIC_{50}$  and define as the model-predicted  $\log_{10}(IC_{50} [\mu M])$  score, where lower  $bIC_{50}$  values indicate stronger predicted binding. For reference, this score can be converted to an approximate  $pIC_{50}$  scale as  $pIC_{50} = 6 - bIC_{50}$  (Supplementary Note 6)<sup>70</sup>. In the present study, these outputs were used only as relative prioritization signals within a small candidate panel, not as calibrated estimates of biochemical potency or covalent reactivity. To anchor interpretation, we included Pterostilbene as an algorithmic positive control and Glucose as a negative control. We then compared model-prioritized candidates, such as Quercetin and RI-linked unknown-derived candidates, against lower-priority metabolites such as Oleic Acid. These calculations were used strictly for shortlist refinement and internal consistency. They do not establish electrophilic modification of KEAP1 sensor cysteines, direct KEAP1-NRF2 pathway activation, or downstream cellular antioxidant effects, all of which require orthogonal biochemical or cell-based validation.

### Data availability

The PTFI dataset analyzed in this study was accessed through the American Heart Association Precision Medicine Platform (<https://pmp.heart.org/ptfi-data-resource>), version v3\_2025\_12. PTFI data are distributed under the CC-BY-NC-4.0 license. External databases were retrieved from: ChEMBL v36 (<https://www.ebi.ac.uk/chembl/>); PubChem (FTP snapshot 2026-01-25; <https://ftp.ncbi.nlm.nih.gov/pubchem/>); FoodAtlas (<https://foodatlas.ai>); UniProt Q14145 for KEAP1. The paper and the Supplementary Information contain all necessary data to assess the conclusions.

### Code availability

All code and instructions on how to reproduce the results can be found at <https://github.com/IBPA/food-metabolomics-bioactivity>.

Received: 25 April 2026; Accepted: 1 August 2026;

Published online: 25 August 2026

### References

- Li, F. et al. A unified knowledge graph linking foodomics to chemical-disease networks and flavor profiles. *npj Sci. Food*, <https://doi.org/10.1038/s41538-025-00680-9> (2026).
- Zhang, J. et al. Emerging trends and applications of metabolomics in food science and nutrition. *Food Funct.* **14**, 9050–9082 (2023).
- Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nat. Food* **1**, 33–37 (2019).
- Bland, J. S. The dark matter of nutrition: dietary signals beyond traditional nutrients. *Integr. Med. Clin. J.* **18**, 12–15 (2019).
- Davies, M. et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **43**, W612–W620 (2015).
- Zdravil, 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).
- Caesar, L. K. & Cech, N. B. Synergy and antagonism in natural product extracts: when 1 + 1 does not equal 2. *Nat. Prod. Rep.* **36**, 869–888 (2019).
- Chen, X., Li, H., Zhang, B. & Deng, Z. The synergistic and antagonistic antioxidant interactions of dietary phytochemical combinations. *Crit. Rev. Food Sci. Nutr.* **62**, 5658–5677 (2022).
- Odenkirk, M. T. et al. A standardized nontargeted metabolomics method for cross-laboratory comparison of food profiles. *Food Chem.* **493**, 145934 (2025).
- Jarvis, A. et al. Periodic Table of Food Initiative for generating biomolecular knowledge of edible biodiversity. *Nat. Food* **5**, 189–193 (2024).
- Wu, W. et al. Emerging applications of metabolomics in food science and future trends. *Food Chem. X* **16**, 100500 (2022).
- Bhakkialakshmi, E. et al. Pterostilbene-mediated Nrf2 activation: mechanistic insights on Keap1:Nrf2 interface. *Bioorg. Med. Chem.* **24**, 3378–3386 (2016).
- Kansanen, E. et al. Electrophilic nitro-fatty acids activate NRF2 by a KEAP1 cysteine 151-independent mechanism. *J. Biol. Chem.* **286**, 14019–14027 (2011).
- Lu, L. et al. Migration of antioxidants from food-contact rubber materials to food simulants. *J. Food Eng.* **318**, 110904 (2022).
- Gupta, R. K. et al. Migration of chemical compounds from packaging materials into packaged foods: interaction, mechanism, assessment, and regulations. *Foods* **13**, 3125 (2024).
- Qiu, Y., Liu, Q. & Beta, T. Antioxidant properties of commercial wild rice and analysis of soluble and insoluble phenolic acids. *Food Chem.* **121**, 140–147 (2010).
- Kim, S.-H., Yang, Y.-J. & Chung, I.-M. The effect of degree of milling on the nutraceutical content in ecofriendly and conventional rice (*Oryza sativa* L.). *Foods* **9**, 1297 (2020).
- Food and Drug Administration. 21 CFR § 172.695 - Xanthan gum (FDA, 2024).
- Lundberg, S. M. & Lee, S.-I. A Unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
- Strilbytska, O. M. et al. Anise hyssop *Agastache foeniculum* increases lifespan, stress resistance, and metabolism by affecting free radical processes in *Drosophila*. *Front. Physiol.* **11**, 596729 (2020).
- Nechita, M.-A. et al. Agastache species: a comprehensive review on phytochemical composition and therapeutic properties. *Plants* **12**, 2937 (2023).
- Ishida, T. et al. Linoleoyl ethanolamide reduces lipopolysaccharide-induced inflammation in macrophages and ameliorates 2,4-dinitrofluorobenzene-induced contact dermatitis in mice. *Eur. J. Pharmacol.* **699**, 6–13 (2013).
- Astiti, M. A. et al. LC-QTOF-MS/MS based molecular networking approach for the isolation of  $\alpha$ -glucosidase inhibitors and virucidal agents from *Coccinia grandis* (L.) Voigt. *Foods* **10**, 3041 (2021).
- Zhang, Q., Ashraf, J. & Xu, B. Natural solutions for food safety: antibacterial activity of xanthohumol against food spoilage microorganisms and foodborne pathogens and its application prospect in food. *Food Biosci.* **63**, 105767 (2025).
- Hevia, H. et al. 5'-methylthioadenosine modulates the inflammatory response to endotoxin in mice and in rat hepatocytes. *Hepatology* **39**, 1088–1098 (2004).

26. Yeschi, K. et al. Revisiting inflammatory bowel disease: pathology, treatments, challenges and emerging therapeutics including drug leads from natural products. *J. Clin. Med.* **9**, 1273 (2020).
27. Jiang, Y., Yu, L. & Wang, M.-H. N-trans-feruloyltyramine inhibits LPS-induced NO and PGE2 production in RAW 264.7 macrophages: Involvement of AP-1 and MAP kinase signalling pathways. *Chem. Biol. Interact.* **235**, 56–62 (2015).
28. Magnone, M. et al. Chronic intake of micrograms of abscisic acid improves glycemia and lipidemia in a human study and in high-glucose fed mice. *Nutrients* **10**, 1495 (2018).
29. Zhang, J., Li, L., Kim, S.-H., Hagerman, A. E. & Lü, J. Anti-cancer, anti-diabetic and other pharmacologic and biological activities of penta-galloyl-glucose. *Pharm. Res.* **26**, 2066–2080 (2009).
30. El-Maslahi, H. et al. Functions and synthesis of abscisic acid (ABA) in humans—insights from computational approaches. *Int. J. Mol. Sci.* **26**, 11115 (2025).
31. Palladin Institute of Biochemistry, National Academy of Sciences of Ukraine, Department of Lipid Biochemistry, Kyiv et al. Historical overview of lipid biochemistry research: from initial hypotheses to understanding the biological role of N-acylethanolamines. *Ukr. Biochem. J.* **97**, 88–101 (2025).
32. Clayton, P., Hill, M., Bogoda, N., Subah, S. & Venkatesh, R. Palmitoylethanolamide: a natural compound for health management. *Int. J. Mol. Sci.* **22**, 5305 (2021).
33. Rao, A., Skinner, R. & Briskey, D. The efficacy of palmitoylethanolamide (Levagen+) on the incidence and symptoms of upper respiratory tract infection—a double blind, randomised, placebo-controlled trial. *Nutrients* **15**, 4453 (2023).
34. Quagliariello, V., Buccolo, S., Iovine, M., Paccone, A. & Maurea, N. P142 palmitoylethanolamide (PEA) reduces inflammation in human cardiomyocytes and vascular endothelial cells exposed to doxorubicin and ANTI-HER2 monoclonal antibody through PPAR- $\alpha$  and NLRP3-related pathways. *Eur. Heart J. Suppl.* **24**, 137 (2022).
35. Bundela, A. K., Abhilash, P. C. & Peñuelas, J. Securing wild edible plants for planetary healthy diet. *Anthr. Sci.* 1–3, <https://doi.org/10.1007/s44177-023-00054-4> (2023).
36. NC State Extension. Agastache foeniculum. *North Carolina Extension Gardener Plant Toolbox*. <https://plants.ces.ncsu.edu/plants/agastache-foeniculum/>.
37. Nowak, D., Kłębukowska, L. & Gośliński, M. Antioxidant properties and antibacterial activity of selected herbal teas. *Sci. Rep.* **15**, 41438 (2025).
38. Melissa officinalis (Balm, Balm Mint, Common Balm, Lemon Balm) | North Carolina Extension Gardener Plant Toolbox. <https://plants.ces.ncsu.edu/plants/melissa-officinalis/>.
39. Palozza, P., Simone, R. E., Catalano, A. & Mele, M. C. Tomato Lycopene and Lung Cancer Prevention: From Experimental to Human Studies. *Cancers* **3**, 2333–2357 (2011).
40. Deding, U., Baatrup, G., Kaalby, L. & Kobaek-Larsen, M. Carrot intake and risk of developing cancer: a prospective cohort study. *Nutrients* **15**, 678 (2023).
41. Ahn, E. H., Chang, C.-C. & Schroeder, J. J. Evaluation of sphinganine and sphingosine as human breast cancer chemotherapeutic and chemopreventive agents. *Exp. Biol. Med.* **231**, 1664–1672 (2006).
42. Dhyani, P. et al. Phloridzin as a nutraceutical for cancer prevention and therapy: a comprehensive review of its mechanisms, bioavailability challenges and future applications. *Food Sci. Nutr.* **13**, e70744 (2025).
43. Ma, M. et al. Digoxigenin activates autophagy in hepatocellular carcinoma cells by regulating the PI3K/AKT/mTOR pathway. *Cancer Cell Int* **24**, 405 (2024).
44. Shrivastava, A. K. et al. Evaluation of antioxidant and anti-inflammatory activities, and metabolite profiling of selected medicinal plants of Nepal. *J. Trop. Med.* **2023**, 1–15 (2023).
45. Itoh, K. et al. Keap1 represses nuclear activation of antioxidant responsive elements by Nrf2 through binding to the amino-terminal Neh2 domain. *Genes Dev.* **13**, 76–86 (1999).
46. Mendoza Cantu, A., Gauglitz, J. M. & Bittremieux, W. Tracing the origins of molecular signals in food through integrative metabolomics and chemical databases. *Npj Sci. Food*. <https://doi.org/10.1038/s41538-026-00802-x> (2026).
47. C., C. Metabolomics: V1 data processing pipeline. *Alliance Bioversity International - CIAT* <https://alliancebioversityciat.org/publications-data/metabolomics-v1-data-processing-pipeline> (2025).
48. Odenkirk, M. T. Nontargeted Metabolomics (Reverse phase) - Analysis of Foods Standard Operating Procedure (SOP). <https://hdl.handle.net/10568/179764> (2025).
49. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics Off. J. Metabolomic Soc.* **3**, 211–221 (2007).
50. Bonini, P., Kind, T., Tsugawa, H., Barupal, D. K. & Fiehn, O. Retip: retention time prediction for compound annotation in untargeted metabolomics. *Anal. Chem.* **92**, 7515–7522 (2020).
51. Breslin, W. & Pham, D. Machine learning and drug discovery for neglected tropical diseases. *BMC Bioinform.* **24**, 165 (2023).
52. Katubi, K. M. et al. Predicting the multiple parameters of organic acceptors through machine learning using RDKit descriptors: An easy and fast pipeline. *Int. J. Quantum Chem.* **123**, e27230 (2023).
53. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
54. Maaten, van der, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
55. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* (eds Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G.) 160–172 (Springer, 2013).
56. Deng, Y., Ericksen, S. S. & Gitter, A. Assay2Mol: large language model-based drug design using bioassay context. <https://doi.org/10.18653/v1/2025.emnlp-main.1390> (2025).
57. Liu, Y. et al. XMC-agent: dynamic navigation over scalable hierarchical index for incremental extreme multi-label classification. in *Findings of the Association for Computational Linguistics ACL 2024* 5659–5672 (Association for Computational Linguistics, 2024).
58. Smit, I. et al. Integrating artificial intelligence and manual curation to enhance bioassay annotations in ChEMBL. *J. Cheminform.* **18**, 24 (2026).
59. Vishwanath, K. et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nat. Med.* <https://doi.org/10.1038/s41591-026-04431-5> (2026).
60. Gutierrez, J.-J. G. et al. Multi-output prediction of dose–response curves enables drug repositioning and biomarker discovery. *Npj Precis. Oncol.* **8**, 209 (2024).
61. Randáková, A., Nelic, D. & Jakubík, J. A critical re-evaluation of the slope factor of the operational model of agonism: When to exponentiate operational efficacy. *Sci. Rep.* **13**, 17587 (2023).
62. Gutiérrez, S., Fernández, C., Escher, B. I. & Tarazona, J. V. A new hazard index of complex mixtures integrates bioconcentration and toxicity to refine the environmental risk assessment of effluents. *Environ. Int.* **34**, 773–781 (2008).
63. Hadrup, N. et al. Concentration addition, independent action and generalized concentration addition models for mixture effect prediction of sex hormone synthesis in vitro. *PLoS ONE* **8**, e70490 (2013).
64. Tanaka, Y. & Tada, M. Generalized concentration addition approach for predicting mixture toxicity. *Environ. Toxicol. Chem.* **36**, 265–275 (2016).

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. Rico, D., Cano, A. B. & Martín-Diana, A. B. Pulse-cereal blend extrusion for improving the antioxidant properties of a gluten-free flour. *Molecules* **26**, 5578 (2021).
67. Cuadrado, A. et al. Therapeutic targeting of the NRF2 and KEAP1 partnership in chronic diseases. *Nat. Rev. Drug Discov.* **18**, 295–317 (2019).
68. Passaro, S. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. <https://doi.org/10.1101/2025.06.14.659707> (2025).
69. Bret, G., Sindt, F. & Rognan, D. Assessing Boltz-2 performance for the binding classification of docking hits. *J. Chem. Inf. Model.* **66**, 1511–1521 (2026).
70. Yan, M. et al. Molecular mechanism of ziresovir targeting the fusion glycoprotein of respiratory syncytial virus. *PLoS Pathog.* **22**, e1013864 (2026).
71. Yuan, S., Chan, H. C. S. & Hu, Z. Using PYMOL as a platform for computational drug design. *WIREs Comput. Mol. Sci.* **7**, e1298 (2017).

## Acknowledgements

We would like to thank the members of the Tagkopoulos lab for helpful discussions and comments, and the Periodic Table of Food Initiative (PTFI) for the provision of the data on 322 foods. This work was supported by the USDA-NIFA AI Institute for Next Generation Food Systems (AIFS), USDA-NIFA award number 2020-67021-32855, and a grant from The Rockefeller Foundation.

## Author contributions

P.G. designed and performed analyses. S.A. provided the dataset and assisted in the discussion on the implications of the data findings and contributed to drafting the manuscript. P.G., M.G., and I.T. contributed to the evaluation, figures, and drafting of the manuscript. I.T. conceived, assisted with the critical analysis, and supervised all aspects of the project. All authors reviewed and approved the final 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/s41538-026-01066-1>.

**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

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.