Computers in Biology and Medicine 181 (2024) 109072

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FoodAtlas: Automated knowledge extraction of food and chemicals from literature

Jason Youn [a,b,c,1], Fangzhou Li [a,b,c,1], Gabriel Simmons [a,b,c], Shanghyeon Kim [b,c], Ilias Tagkopoulos [a,b,c,*]

[a] Department of Computer Science, University of California, Davis, Davis, CA, 95616, USA
[b] Genome Center, University of California, Davis, Davis, CA, 95616, USA
[c] USDA/NSF AI Institute for Next Generation Food Systems, Davis, CA, 95616, USA

ARTICLE INFO

Keywords:
Nutrition
Food chemical
Data mining
Knowledge graph
Deep learning
Large language model
Link prediction
Quality control

ABSTRACT

Automated generation of knowledge graphs that accurately capture published information can help with knowledge organization and access, which have the potential to accelerate discovery and innovation. Here, we present an integrated pipeline to construct a large-scale knowledge graph using large language models in an active learning setting. We apply our pipeline to the association of raw food, ingredients, and chemicals, a domain that lacks such knowledge resources. By using an iterative active learning approach of 4120 manually curated premise-hypothesis pairs as training data for ten consecutive cycles, the entailment model extracted 230,848 food-chemical composition relationships from 155,260 scientific papers, with 106,082 (46.0 %) of them never been reported in any published database. To augment the knowledge incorporated in the knowledge graph, we further incorporated information from 5 external databases and ontology sources. We then applied a link prediction model to identify putative food-chemical relationships that were not part of the constructed knowledge graph. Validation of the 443 hypotheses generated by the link prediction model resulted in 355 new food-chemical relationships, while results show that the model score correlates well ($R^2 = 0.70$) with the probability of a novel finding. This work demonstrates how automated learning from literature at scale can accelerate discovery and support practical applications through reproducible, evidence-based capture of latent interactions of diverse entities, such as food and chemicals.

1. Introduction

Mapping the chemical composition of food and ingredients is essential for unlocking their potential and informing decisions. From creating healthier and tastier food products[1,2] to enriching food with the right compounds[3,4] or building personalized diets[5–7], understanding what is in each ingredient and at what concentration is paramount. Food composition at the molecular level is usually found in food composition tables like the USDA's FoodData Central (FDC)[8] or the ANSES-CIQUAL database[9]. This enables several stakeholder groups, from researchers to policymakers, to assess the nutrition quality of various foods and their regulatory status and to use them in the respective industries[10]. However, despite the established importance of the food composition information, most of the food-chemical information that is present in the scientific literature is not captured in the structured databases[1]. For instance, the total size of food composition space is estimated at tens of thousands of chemicals[11], while FDC and ANSES-CIQUAL focus on only 500 compounds. To expand the coverage of chemicals in foods, several initiatives attempt to capture food composition from scientific literature, such as FooDB[12] (797 foods and 15,750 detected chemicals) and DietRx[13] (2222 foods and 6992 chemicals), which further aggregate data from several other databases like FDC[8], KNApSAcK[14], Dr. Duke's Phytochemical and Ethnobotanical Databases[15], Phenol-Explorer[16–18], and PhytoHub[19]. However, existing databases require laborious annotation effort from experts or lack consistent quality control as the majority of their food-chemical composition information is not linked to evidence that allows reproducible results. For example, less than 1 % of associations in FooDB, one of the most notable DBs in this space, have literature citations to support them (**Supplementary Information Section 1.1.1**).

* Corresponding author. Department of Computer Science, University of California, Davis, Davis, CA, 95616, USA.
E-mail address: itagkopoulos@ucdavis.edu (I. Tagkopoulos).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2024.109072
Received 27 May 2024; Received in revised form 16 July 2024; Accepted 22 August 2024
Available online 30 August 2024

Although manual extraction by the domain experts is often precise, it does not scale well with bibliographic literature sources such as PubMed [20], which contains 34 million citations and abstracts, and PubMed Central (PMC)[21], which includes 7.6 million full-text scientific literature articles. From PMC, we estimate we can extract at least 2 million unique food-chemical associations from the unstructured text data (**Supplementary Information Section 1.1.2**). The sheer amount of available scientific literature necessitates the need for an automated framework for constructing knowledge graphs (KGs), which is widely used thanks to their scalability and ability to reveal previously hidden patterns and relationships in the data, leading to better insights and more informed decision-making(X. [22]), especially in the life sciences domain(J. [23]; N. [24]). There has been prior work utilizing language models to construct domain-specific knowledge graphs from unstructured texts[25–30], with some combined with active learning (AL) to reduce human annotation[31–33]. Although some works have constructed food-relevant knowledge graphs, they are limited by the low ground truth precision of relation extraction[25] and the small number of chemical entities[29].

In this work, we present the Lit2KG framework (Fig. 1a) that extracts information from scientific literature using a large language model in an AL setting to construct a large-scale KG. The entailment model of the Lit2KG framework uses a premise from the scientific literature to extract and predict multiple hypotheses with high performance (F1 score of 83 %), with the predicted probabilities being highly correlated to the ground-truth annotations ($R^2 = 0.94$). We also tested four different AL strategies and found that selecting samples that maximize the likelihood leads to discovering new knowledge 38.2 % faster than the baseline. Applying graph-embedding link prediction models for graph completion followed by validation through literature search revealed 355 missed food-chemical composition associations that were further verified manually and 11 additional associations that were novel, 6 of which we have found strong evidence to support them. The resulting knowledge graph contains 285,077 triplets of three entity types (food, part, chemical) and four relation types (*contains*, *has part*, *is a*, *has child*) on three evidence quality levels (high, medium, low) with 4318 of them evaluated by human experts (Fig. 1b).

2. Material and methods

2.1. Premise-hypothesis pair generation

We collected a total of 1959 raw and non-processed food names that have a known National Center for Biotechnology Information (NCBI) Taxonomy ID[34] from multiple food databases (see Supplementary Fig. 1a). We then used the LitSense API[35], which is a search system for biomedical literature at the sentence level provided by the NCBI, to query for the search keyword "{*food name*} contains" (**Supplementary Data 1**; **Supplementary Information Section 1.2.1**). The LitSense API returns sentence-level text snippets from the PubMed abstracts and the PMC open-access full-text articles, as well as the named entity recognition (NER) service for species and chemical entities, along with their

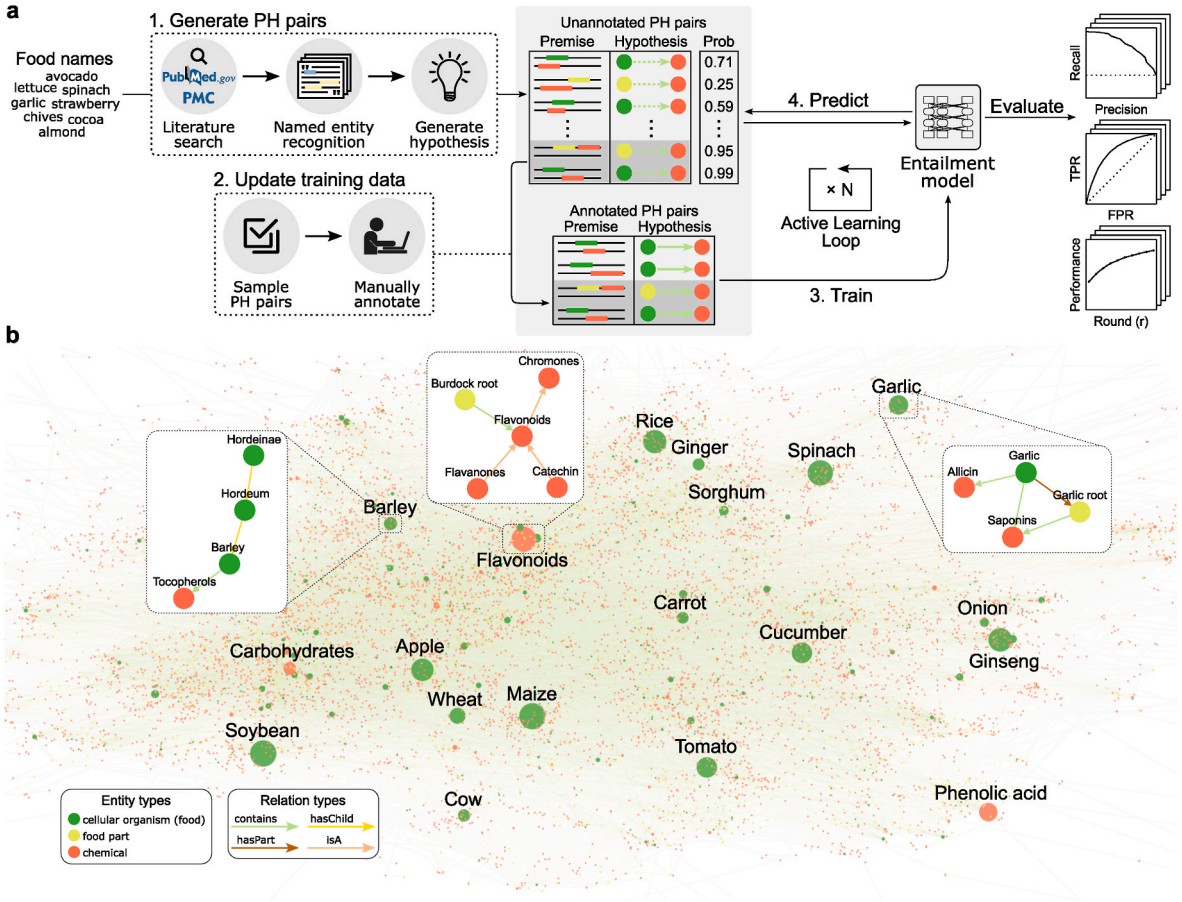

Fig. 1. Overview of the Lit2KG framework and the FoodAtlas Knowledge Graph. a, Scientific literature is queried using raw food names and retrieved sentences (premises) where the species and chemical entities are tagged (*e.g.*, … cocoa$_{[SPECIES]}$ is a good source of (−)-epicatechin$_{[CHEMICAL]}$ …). From these premises, hypothesis triplets are generated such as (*cocoa, contains, (−)-epicatechin*), which we refer to as premise-hypothesis (PH) pairs. The entailment model is then iteratively updated through active learning cycles, where a new batch of PH pairs is annotated in each cycle. Finally, both annotated and predicted positive PH pairs are used to populate the knowledge graph. **b**, Visualization of the FoodAtlas Knowledge Graph (FAKG), which contains 285,077 triplets of 3 entity types and 4 relation types. Each triplet in the FAKG is assigned one of three quality types and provides a reference to the publications that support it for reproducibility.

corresponding NCBI Taxonomy IDs and MeSH IDs, respectively. We further processed these text snippets by discarding non-food entities and tagging the part entities (*e.g.*, leaf and root) using our manually generated lookup table consisting of 70 food parts (**Supplementary Data 2**).

For each LitSense-returned sentence $s_i \in S$, which we refer to as a *premise* in our work, there exist three sets of named entities F_i, P_i, and C_i for food, parts, and chemicals, respectively, where P_i can be an empty set as not all sentences have parts in them. We then generated a set of hypotheses H_i for each premise s_i by taking the cartesian product of the entity sets F_i, P_i, and C_i as

$$H_i = \{template(f, p, c) \forall (f, p, c) \in F_i \times P_i \times C_i\} \cup \{template(f, c) \forall (f, c) \in F_i \times C_i\},$$

where $template(\cdot)$ is the hypothesis template that generates a triplet of type ({*food*} {*part*}, contains, {*chemical*}) or ({*food*}, contains, {*chemical*}), respectively. We refer to these pairs of premise and the extracted hypotheses as premise-hypotheses (PH) pairs in our work (see Supplementary Fig. 1b).

2.2. Premise-hypothesis pair annotation

We annotated the PH pairs to generate a dataset for training, validating, and testing the entailment model using the AL strategy (described in the following sections). During the annotation process, a given PH pair was assigned one of three possible classes *entails*, *does not entail*, and *skip*. More specifically, *entails* was assigned if the premise supported the underlying relationship used to construct the hypothesis, and *does not entail* was assigned if there was insufficient evidence in the premise to support the hypothesis. Note that the hypothesis from a PH pair marked as *does not entail* is not necessarily a negative, as another premise may support the hypothesis. Finally, *skip* was assigned if the premise the LitSense API returned was not formatted correctly or if the NER tagging by LitSense API was wrong (**Supplementary Information Section 1.2.2**). To ensure the annotation was of high quality, two experts annotated each PH pair independently, and only the PH pairs that had agreed annotation results by the two experts were kept. We randomly split the data into training, validation, and test sets with approximate ratios of 70 %, 15 %, and 15 %. To avoid data leakage, we ensured that the three datasets did not share the same premises or hypotheses during the splitting. In the end, we had a training set with 4120 PH pairs (1899 *entails*, 2221 *does not entail*), a validation set with 825 PH pairs (295 *entails*, 530 *does not entail*), and a test set with 840 PH pairs (312 *entails*, 528 *does not entail*) (**Supplementary Data 3**).

2.3. Entailment model

We trained the entailment model to predict whether the premise logically would entail the hypotheses. To this end, we used the BioBERT [36] over other language models[37–39] (**Supplementary Information Section 1.2.3**), as it was pre-trained on the same corpus as where the premises were extracted from(PubMed abstracts and PMC full-text articles) and have demonstrated improved performance on biomedical benchmarks[36]. We then fine-tuned the BioBERT entailment model by utilizing the binary classification schema, where the input sequence was formatted by concatenating the premise and hypothesis with the [SEP] token in between, and the model predicted if the given PH pair was *entails* or *does not entail*. We used the held-out validation set to optimize the hyperparameters, where the tunable hyperparameters were learning rate $= \{2 \times 10^{-5}, 5 \times 10^{-5}\}$, epochs $= \{3, 4\}$, and batch size $= \{16, 32\}$. The hyperparameter set with the best held-out validation precision was selected, and the performance of each round was reported using the held-out test set. Note that we trained a production entailment model using all the labeled data (*i.e.*, training, validation, and test sets) (**Supplementary Information Section 1.2.3**).

2.4. Active learning strategy

In this work, we tested four active learning (AL) strategies, *maximum likelihood*, *maximum entropy*, *stratified*, and *random*. We simulated the AL strategy by splitting the training pool with 4120 PH pairs into ten rounds $r = \{1, 2, \ldots, 10\}$, with 412 new PH pairs selected in each round and appended to the existing training data by the respective strategy. In other words, at round r, we trained the entailment model m_r using 412 x (r – 1) training PH pairs plus 412 new PH pairs selected from the remaining 412 x (10 – r + 1) PH pairs. We call this training and evaluation process a *run*, and we repeated 100 *runs* for each AL strategy to test the statistical significance. The *stratified* strategy first ranked the remaining PH pairs from high to low probability and split them into ten equally sized bins, randomly drawing the same number of samples from each bin. The *maximum likelihood* strategy chose the top 412 positive samples based on their probability score. The *maximum entropy* sampling strategy first computed the uncertainty for each PH pair as $\min(1-p, p)$, where p is the probability of the given PH pair predicted by the entail model. All PH pairs were then ranked using the uncertainty value from high to low, and the top 412 uncertain PH pairs were selected. Finally, the *random* sampling strategy chose 412 PH pairs randomly. Note that for the first round, all four AL strategies randomly selected the first round of PH pairs to train on, and for the last round, all four AL strategies were trained on a whole training pool of 4120 PH pairs regardless of the sampling strategy taken. More detailed information can be found in **Supplementary Information Section 1.2.3**, and a visual illustration of the sampling strategies is in Supplementary Fig. 2.

2.5. Knowledge graph generation

After LitSense and our entailment model performed knowledge extraction, we constructed our knowledge graph based on the schema shown in Fig. 2a. Our schema defined *contains* relationship between food and chemical entities and included taxonomical relationships derived from NCBI Taxonomy and MeSH tree ontologies. The FoodAtlas Knowledge Graph $FAKG = (E, R)$ encodes information using a bag of triplets (h, r, t), where $\{h, t\} \in E$ is the set of all entities (h for the head entity and t for the tail entity) and $r \in R$ is the set of all relation. Each triplet in the KG can have one or more sources and qualities. In this work, we define three qualities *high*, *medium*, and *low* for a triplet. The *high*-quality triplets have been validated by the FoodAtlas team and have PMID and/or PMCID. The *medium*-quality triplets are not validated by the FoodAtlas team but have PMID and/or PMCID. Taxonomy and ontology also are medium-quality triplets. The *low*-quality triplets are not validated by the FoodAtlas team and do not have PMID or PMCID. Please refer to **Supplementary Information Section 1.3** for the details of the FAKG design including the entity and relation types.

The first source of information was from the PH pair annotation process, where two relation types, *contains* and *has part*, exist. The triplets with the *contains* relation type were from the positive annotated PH pairs, whereas the triplets with the *has part* relation type were automatically extracted from the *contains* triplets. For example, a triplet (*coconut, has part, coconut seed*) was extracted from the triplet (*coconut seed, contains, lauric acid*). All triplets from this source were high-quality. The second source was the entailment model predictions, also with the *contains* and *has part* relation types. However, these were not annotated and thus were assigned a medium-quality. The third source was the enrichment through the NCBI Taxonomy and MeSH tree ontology. The NCBI Taxonomy, which contains medium-quality triplets with the *has child* relation type, encodes the hierarchical structure of the taxonomic lineage (*Cocos (genus), has child, Cocos nucifera (species)*). The MeSH tree, which contains medium-quality triplets with the *is a* relation type, encodes the ontological relationship of the chemical entities. We also included the triplets extracted from the external databases (Frida[40], FDC, and Phenol-Explorer) with either *medium*- or *low*-quality triplets with the *contains* relation type. Finally, we also included the link

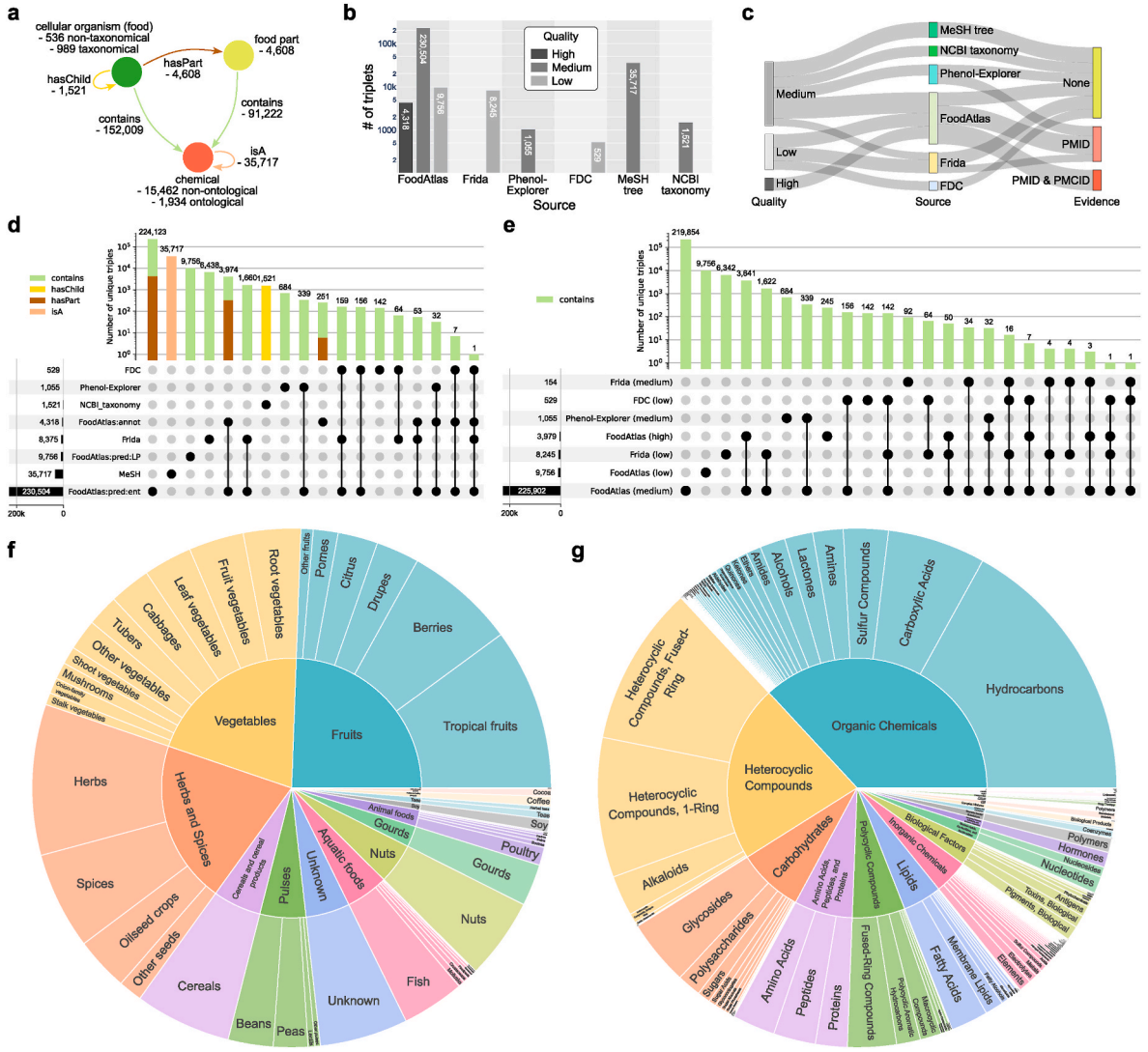

Fig. 2. Statistics of the FoodAtlas Knowledge Graph. a, Schema of the FAKG. The relation types *contains*, *hasPart*, *isA*, and *hasChild* encode the food-chemical composition relations, the food-food with part relations, the chemical ontological relations using the MeSH tree, and the taxonomical relations using the NCBI Taxonomy, respectively. **b**, Number of triplets per data source in the FAKG depending on the quality. **c**, Sankey graph showing the connections between quality, data source, and evidence. The thickness of the relations between the nodes represents the number of connections in the log scale. **d**, **e**, UpSet plot showing the number of unique triplets for all data sources for all relation types and all sources based on quality for only the contains triplets. Each row in the plot corresponds to a source, and the bar chart on the left shows the size of each source. Each column corresponds to an intersection, where the filled-in cells denote which source is part of an intersection. The bar chart for each column denotes the size of intersections. 'annot' stands for annotation, 'pred' stands for prediction, and 'LP' stands for link prediction. **f**, **g**, Classification of foods and chemicals in FoodAtlas.

prediction results (triplets with the *contains* relation type) as low-quality.

2.6. Link prediction

Link prediction is a widely studied field that refers to the task of predicting missing relationships or links between entities in a graph, (food, contains, chemical) triplet type in our case, and contributes to the enhancement and enrichment of knowledge graphs[41]. Using the Python library PyKEEN[42], we trained a set of benchmark link prediction models TransE[43], ER-MLP[44], DistMult[45], TransD[46], ComplEx [47], and RotatE(Z. [48]) on different versions of the FAKG (Fig. 5a and b), performed hyperparameter optimization on the held-out validation set using mean rank (MR), and reported the results on the held-out test set (**Supplementary Information Section 1.2.4**). The link prediction models were also calibrated using isotonic regression to provide an interpretable probability score. Link prediction models are commonly evaluated using rank-based metrics like mean rank (MR), mean reciprocal rank (MRR), hits@1, hits@3, and hits@10[49]. However, our end goal was to generate hypotheses that were either true or false, and therefore, we decided to also evaluate using standard binary classification metrics like confusion matrix, precision, and recall. To this end, we randomly sampled two negatives for each positive triplet in the validation and test set by corrupting the head and tail entity once, which resulted in a validation set with 1335 triplets (445 positives and 890 negatives) and a test set with 1341 triplets (447 positives and 894 negatives). Due to the nature of the graph-embedding models that cannot make predictions on test triplets with an entity that is never seen during the training, we report our binary classification metrics in a stricter *unfiltered* setting, where the test triplets that would be dropped in the *filtered* setting are kept and assigned a default majority label 0.

2.7. Link prediction literature validation

To validate the link prediction-generated food-chemical triplets, we searched the following four sources sequentially: PubChem taxonomy [50], Bing Chat, Google Scholar, and Google. Specifically, for a given food-chemical pair, we first checked if the Taxonomy section of PubChem entry for the chemical of interest lists the scientific name of the food and has a reference. If not, we then asked Bing Chat, a search engine based on a large language model, to find the reference (Supplementary Fig. 3). Next, we searched Google Scholar using a set of pre-defined search queries (**Supplementary Information Section 1.1.8.2**). If the initial Google Scholar search did not return the positive relationship within the first three pages (30 papers, 10 papers per page), we repeated the process with the synonyms of the entities. Finally, we searched the first 30 contents of Google using the same search method as Google Scholar. A complete procedure for the link prediction validation can be found in **Supplementary Information Section 1.1.8**.

3. Results

3.1. The FoodAtlas Knowledge Graph contains a wide spectrum of food-chemical composition information

We utilized the Lit2KG framework (Fig. 1a) to extract the food-chemical composition information from the PubMed abstracts and open-access articles using raw food ingredients as queries (see **Section 2.1**). From this search, we generated 3,596,755 premise-hypotheses (PH) pairs where the hypotheses are (food, contains, chemical) or (food part, contains, chemical) triplets. We then used BioBERT[36], a biomedical language representation model for triplet binary classification that we fine-tuned with 4318 manually curated positive triplets in an active learning setting. This resulted in 230,504 additional positive triplets, for a total of 234,822 unique positive triplets. In addition, we curated and added the food-chemical composition information based on quality criteria from three external databases (8375 triplets from Frida [40], 1055 triplets from Phenol-Explorer[18], and 529 triplets from FDC [8]), taxonomical information of the foods using the NCBI Taxonomy (1526 triplets), and ontological information of the chemicals using the MeSH tree (43,691 triplets) (Fig. 2d). Applying link prediction on the knowledge graph generated an additional 9756 triplets of food and chemical pairs, 355 of them manually validated as positives. The final FoodAtlas knowledge graph (FAKG, Fig. 1b) contains 536 food entities, 4608 food parts, 15,462 chemical entities, and 285,077 unique triplets about food-chemical composition with four different relation types and three different entity types (Fig. 2a–**g**).

In terms of triplet quality, FAKG has 4318 (1.5 %) high-quality (*i.e.*, validated by two experts), 264,455 (92.8 %) medium-quality (*i.e.*, with at least one reference, but not manually validated), and 16,304 low-quality (5.7 %) triplets (*i.e.*, no references, see **Section 2.5** and Fig. 2b). From those, 4,318, 226,437, and 9,756, respectively, have been uniquely captured by our Lit2KG pipeline and the link prediction analysis (**Supplementary Information Section 1.1.3** and Fig. 2b and c). The top five foods whose chemical composition is most well documented in the knowledge graph are soybean (Glycine max), maize (*Zea mays*), rice (*Oryza sativa*), cucumber (*Cucumis sativus*), followed by tomato (*Solanum lycopersicum*) (Supplementary Fig. 4).

3.2. FoodAtlas discovers complementary information to benchmark datasets

To test how good the coverage of the food-chemical composition triplets from the Lit2KG pipeline is, we compared them with FoodMine [51], a database that contains a manually curated chemical composition of two selected foods, cocoa (592 chemicals) and garlic (289 chemicals). Although there were initially 1289 cocoa and 1376 garlic chemicals in FAKG, we adopted the same method used by FoodMine to make chemicals in the two sources comparable and created an additional chemical identifier, specifically for matching FoodMine chemicals with those in FAKG (**Supplementary Information Section 1.1.4**). After this processing step, the FAKG has 379 cocoa and 406 garlic chemicals, whereas FoodMine has 301 and 176, respectively. Out of 575 chemicals for cocoa, 274 (47.7 %) chemicals were found in FAKG but not in FoodMine, 105 (18.3 %) chemicals were common between the two, and 196 (34.1 %) chemicals were not found in FoodAtlas (Fig. 3a). For garlic, FoodAtlas was able to capture 51.1 % (90 out of 176) of FoodMine chemicals, while 316 chemicals were unique to FAKG (Fig. 3b; see Supplementary Fig. 5 for a similar comparison with FooDB).

3.3. Maximum likelihood active learning strategy discovers knowledge 38 % faster than without

We fine-tuned the BioBERT-based entailment models based on four different AL strategies over ten rounds (see **Section 2.4**; Supplementary Fig. 6). Although all four AL strategies eventually discovered the same set of 1899 positives among the 4120 PH pairs in the training pool at the final round ($r = 10$), the maximum likelihood strategy identified the positives in training set by 38.2 % ± 27.3 % faster than the active learning baseline of choosing random pairs, followed by the maximum entropy (10.7 % ± 6.6 %) and stratified learning (9.3 % ± 5.3 %; Fig. 4a and b and **Supplementary Information Section 1.1.5**). This was because only the positive food-chemical relationships (i.e., *contains*) were added to the knowledge graph so that the maximum likelihood strategy was able to discover more positives via active sampling (see **Supplementary Information Section 1.2.3.3**). Concomitantly, we observed lower performance for the entailment models trained using the maximum likelihood strategy than the others on all metrics for rounds 2 through 4 (adjusted p-value $<3.6 \times 10^{-2}$). This was due to data imbalance, as the maximum likelihood strategy samples PH pairs that were highly probable, and thus its entailment models were trained on an unbalanced training set where on average, 74.9 % of the training data for rounds 2 through 4 was positive compared to 53.6 %, 52.2 %, and 46.1 % for maximum entropy, stratified, and random, respectively, (Supplementary Fig. 7).

For the final entailment models, AUCPR = 0.90 and AUROC = 0.94, where the baselines were 0.37 and 0.50, respectively (Fig. 4d and e and Supplementary Table 1). The model PH prediction probability was well-calibrated and highly correlated with the actual ground truth statistics after manual validation ($R^2 = 0.94$, Fig. 4c). For instance, 88.6 % out of all triplets with a probability ≥0.9 were positives, whereas only 3.9 % with a probability <0.1 were positives (Supplementary Fig. 8).

3.4. Sources of error and impact of large language model general knowledge

Not surprisingly, the entailment model predicted best on straightforward, simple sentence structure, while its performance deteriorated when domain expertise was needed or premises were hypotheses posed by the authors as shown in index 5–8 of Table 1 (see **Supplementary Information Section 1.1.6**). Variance across bootstrapped models was maximized with uncertainty: predictions with 40 %–60 % probability had a standard deviation of 0.31 vs. 0.04 for predictions with less than 10 % or more than 90 % probability (p-value $= 2.2 \times 10^{-177}$; see **Supplementary Data 3**). Furthermore, analyzing the entailment model prediction results based on which section of the literature the premise was taken from (*e.g.*, introduction, methods, etc.) revealed higher precision in certain sections. Unexpectedly, hypotheses stemming from the introduction and methods sections were associated with high precision (0.91 and 0.89, respectively) when compared to sections like abstract, title, and conclusion (0.77, 0.75, and 0.74, respectively; p-value: 9.7×10^{-76}) (see **Supplementary Information Section 1.1.7** and Supplementary Table 2).

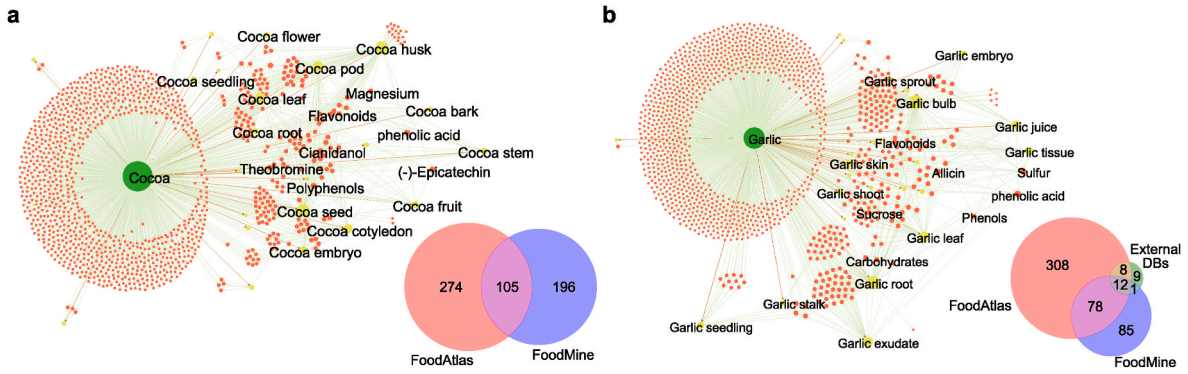

Fig. 3. Results of comparing cocoa and garlic to the benchmark dataset FoodMine. a, **b**, FoodAtlas subgraph of cocoa and garlic where whole food and food parts and their chemical composition are displayed. The label of the top 20 nodes with the largest degree is shown for each subgraph, and the size of the node is proportionate to its degree. The Venn diagram shows the overlap of FoodAtlas (entailment model annotation, entailment model prediction, and link prediction), external databases (Frida, Phenol-Explorer, and FDC), and FoodMine. Interestingly, none of the 3 external databases reported any chemical composition of cocoa.

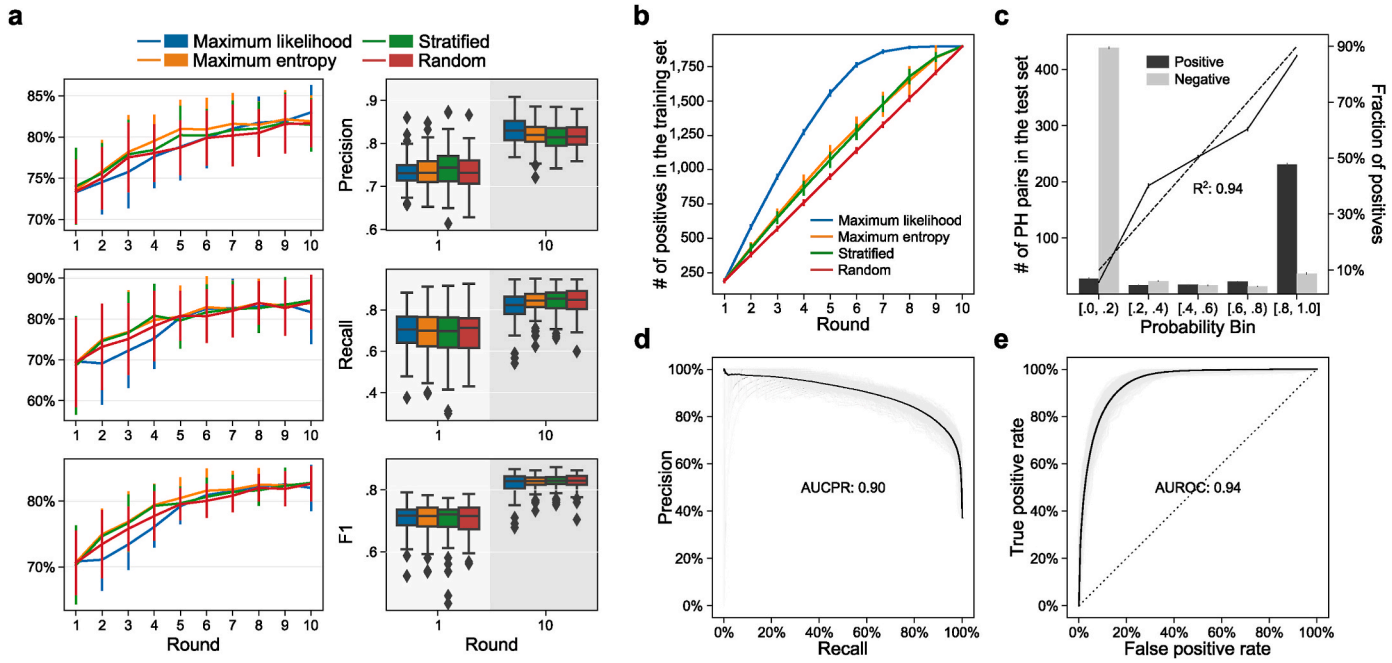

Fig. 4. Prediction performance of the entailment model. a, Precision, recall, and F1 score of the entailment models trained using the 4 different AL strategies for initial (r = 1) and final (r = 10) rounds (n = 100, 100 different random seeds). On the left, the line plot shows the mean value of each AL strategy, and the error lines denote the standard deviation of the 100 random seeds. On the right, the box represents the interquartile range, the middle line represents the median, the whisker line extends from minimum to maximum values, and the diamond represents outliers. **b**, Comparison of the new knowledge discovery rate compared between the 4 AL strategies. The plot shows how early on in the AL round the 1899 positive triplets within the simulated training pool of 4120 triplets are discovered. The error line shows the standard deviation of the 100 random seeds. **c**, Calibration plot showing a high correlation between the probability assigned by the entailment model and the ground-truth annotations on the test set ($R^2 = 0.94$). **d**, **e**, The precision-recall and receiver operating characteristic curves of the entailment model predictions compared to the ground-truth annotations in the test set at the final round (r = 10) averaged over all 400 runs with a different random seed (100 runs for each of the 4 AL strategies).

3.5. Link prediction, GPT model, and the impact of ontologies in performance

We trained a set of link prediction models for the *contains* relation between previously unknown food-chemical pairs (Fig. 5a). The best performance was from TransD trained on the $FA_{A,R}$ (TransD-$FA_{A,R}$) with an overall best performance (precision: 79.3 %, recall: 75.4 %, and F1: 77.2 %) (Fig. 5b). However, as these models cannot classify triplets with entities not seen during the training phase, we used the next best model, RotatE-$FA_{A,E,R,P80}$ that has this capacity (precision: 76.8 %, recall: 70.6 %, and F1: 73.5 %;**Supplementary Data 4**and Fig. 5c). Interestingly, the inclusion of ontological information (Enrichment in Fig. 5b), increases the F1 score by 22.2 % (63.2 % of FA_A vs. 77.2 % of $FA_{A,R}$; *p*-value $= 2.4 \times 10^{-5}$). Moreover, RotatE-$FA_{A,E,R,P80}$ is highly calibrated with $R^2 = 0.99$ (Fig. 5d) and has an AUCPR of 0.82 (baseline 0.33) and AUROC of 0.88 (baseline 0.5) (Fig. 5e and f). All link prediction models performed better than the generalized GPT-3.5 model (text-davinci-003), which was not fine-tuned using the KG (precision: 64.8 %, recall: 31.8 %, and F1: 42.7 %) (**Supplementary Information Section 1.2.4**).

3.6. Link prediction reveals previously unknown food-chemical relationships

The final FAKG contains 536 food entities (excluding food part

Table 1
Comparison of the entailment model predicted premise-hypotheses pairs and the ground-truth annotation. The probability column shows the mean and standard deviation of the probability scores assigned to the corresponding PH pair at the final round (r = 10) of active learning by the 400 entailment models (100 random seeds each for 4 active learning strategies). GT stands for ground truth class assigned by the consensus of two annotators based on the premise. Samples shown in this table are from the test set.

Index	Premise	Hypothesis	Section	GT	Prediction	Probability
1	Standardized extracts from the **leaves** of **Ginkgo biloba** contains 24 % ginkgo-flavone glycosides and 6 % terpenoids (**ginkgolides**, bilobalide)[84].	(Ginkgo biloba – leaves, contains, ginkgolides)	Intro	Entails	Entails	99.6 % ± 1.0 %
2	This Vaccinium myrtillus L extract is composed of **flavonoids**, and standardized to contain 36 % anthocyanins, with conformance to the USP 31 on 'Powdered **Bilberry** Extract' [85].	(Bilberry, contains, flavonoids)	Methods	Entails	Entails	51.3 % ± 33.7 %
3	RYNXC consisted of 9 traditional Chinese herbs, including **clove**, rhubarb, frankincense, myrrh, **borneol**, rhizoma corydalis, cowherb **seed**, Rosae rugosae, Garden balsam stem.(G. [86])	(Clove – seed, contains, borneol)	Intro	Does not entail	Does not entail	0.3 % ± 0.4 %
4	For this purpose, tablets were produced containing 16 mg of **ellagic acid** with 100 mg of pulp from the fruit of an evergreen tree called Cherimoya, soursop, **custard apple**, and other common names (Annona muricata)[87].	(Custard apple, contains, ellagic acid)	N/A	Does not entail	Does not entail	44.4 % ± 37.0 %
5	Previous investigations postulated that **polyunsaturated fatty acids** (PUFAs) are essential nutrients for the **common octopus**[88].	(Common octopus, contains, polyunsaturated fatty acids)	Intro	Does not entail	Entails	99.3 % ± 1.0 %
6	**Domoic acid** excretion in **dungeness crabs**, razor clams and mussels[89].	(Dungeness crabs, contains, Domoic acid)	Title	Does not entail	Entails	62.7 % ± 34.4 %
7	Antihyperlipidaemic and antihypercholesterolaemic effects of **Anethum graveolens** leaves after the removal of **furocoumarins**[90].	(Anethum graveolens, contains, furocoumarins)	Title	Entails	Does not entail	2.3 % ± 8.0 %
8	In the study by Keskiner et al. (2017), the patients in the test group received capsules containing 6.25 mg EPA and 19.19 mg **DHA** from **Atlantic salmon** (Vectomega tablet, Laboratoires Le Stum, Plage, France)[91].	(Atlantic salmon, contains, DHA)	Results	Entails	Does not entail	44.7 % ± 34.3 %

entities) and 15,462 chemical entities, which translates to 8,287,632 possible food-chemical pairs. Only 1.72 % (142,253 triplets) of these food-chemical pairs are connected via the *contains* relation, with the rest, 98.28 % (8,145,379 triplets), being unknown. We, therefore, used RotatE to assign probability scores to these unknown pairs (Fig. 6a), among which 9756 pairs (0.1 %) were assigned a positive prediction label (see **Section 2.6** and **Supplementary Data 5**). Validating 443 sampled hypotheses from these pairs through an extensive literature search (Fig. 6b and **Supplementary Information Section 1.1.8**) revealed 355 positive *contains* triplets between 203 foods and 153 chemicals (Fig. 6c), while 11 triplets remained yet unknown with no direct evidence (Supplementary Table 3 and **Supplementary Data 5**).

A closer look at the 355 triplets demonstrated the importance of link prediction for knowledge graph completion. Linolenic acid, which is an essential omega-3 fatty acid that must be obtained through the diet and helps reduce inflammation[52], lower blood pressure[53], and improve cholesterol levels[53], were validated to be found in 14 different foods (Fig. 6c). The link prediction also discovered evident relationships such as the iodide ion, which is an essential trace element for vertebrates, and manganese(2+), which is a cofactor for many enzymes involved in metabolism[54], including those that are important for bone development[55] and antioxidant defense[56], each with relationship to 10 different foods (Fig. 6c). When it comes to foods, we identified five foods, *Lota lota* (NCBI:txid69944), *Brassica oleracea* var. *italica* (NCBI:txid36774), *Lupinus albus* (NCBI:txid3870), *Panax ginseng* (NCBI:txid4054), *Musa x paradisiaca* (NCBI:txid89151), that have largest number of positively validated positives to 5 chemicals each (Fig. 6c).

3.7. AI-driven discovery of six food-chemical relationships

We performed additional analysis for the 11 potential novel food-chemical candidates not reported in the literature (Supplementary Table 3) and found strong evidence that supports the relationships for 6 of them. Fig. 6d shows 3 of these potentially novel food-chemical relationships, whereas the rest can be found in Supplementary Fig. 12 and **Supplementary Information Section 1.1.8**. For (Atlantic cod, beta-carotene), metabolic pathway analysis identified homologous enzymes directly associated with the synthesis or metabolism of the chemical in the food (**Supplementary Information Section 1.1.8**). Specifically, the enzyme beta-carotene-15,15′-dioxygenase, which metabolizes beta-carotene in human[57], had 58.5 % sequence similarity with beta, beta-carotene 15,15′-dioxygenase-like in the Atlantic cod. Similarly, for (dudaim melon, matairesinol), we found an enzyme secoisolariciresinol dehydrogenase for biosynthesis of matairesinol in genetically close species *Cucumis melo* and varietas *Cucumis melo* var. *makuwa*[58], as we did not have the *Cucumis melo* var. *dudaim* genome to run a direct search. For the (bearded tooth, lumisterol) pair, we found the existence of ergosterol in bearded tooth[59] that converts to lumisterol under UV irradiation[60].

4. Discussion

In this work, we created an automated framework to extract information from literature and create domain-specific knowledgebase graphs. Applying to food and chemical relationships created the first AI-driven resource in the field, summarizing findings through 285,077 triplets, with 106,082 (2091 high-, 94,095 medium-, and 9896 low-quality) of those associations (46.0 %) never been reported before in published databases (**Supplementary Information Section 1.1.9**). While 98.2 % of triplets from the Lit2KG pipeline were labeled as either medium or low quality (Fig. 2b), our results indicate high performance for both the entailment model (medium-quality triplets; precision of 0.82) and the link prediction model (low-quality triplets; precision of 0.77). Additionally, both models exhibit strong calibration (R^2 of 0.94 and 0.99, respectively); that is, the model's predicted probabilities accurately reflect the likelihood of outcomes, providing reliability, interpretability, and better decision support. Surprisingly, in many cases, there are no indexed references associated with the reported entries and unique standardized IDs for the foods and compounds, which made reproducibility and provenance very difficult (**Supplementary Information Section 1.3.3** and Supplementary Table 6). FoodAtlas, by design, addresses this challenge by associating one or more references to each association.

Similarly, we are surprised that most of the associations that we have mined from the literature are not part of the existing databases, which argues that there is a plethora of information to be identified, validated, and integrated into tools like FoodAtlas. This, in turn, will be a boon for data-driven tools and pipelines for various applications, compound and source identification, product formulations, and other R&D operations that currently are serendipitous, error-prone, and time-consuming. Concomitantly, the food-chemical composition knowledge coverage of what is currently in various databases varies (22 % of Frida, 35 % of

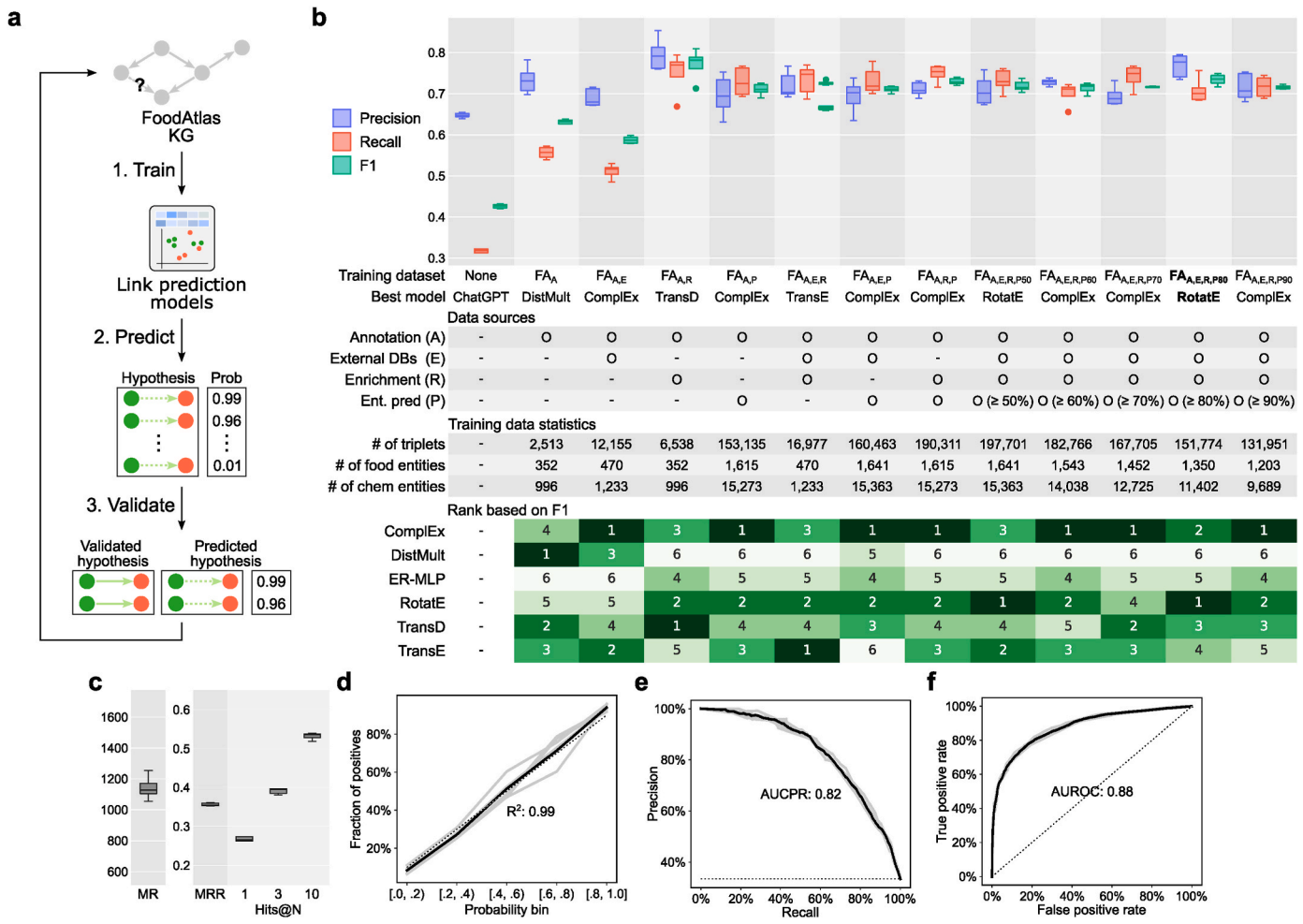

Fig. 5. Link prediction model performance. a, We use the FAKG to train a link prediction model whose objective is to generate hypotheses of type (food, contains, chemical) that is previously unknown in the graph. **b**, Ablation study result showing the performance of 6 different link prediction models trained using 12 different versions of the FAKG, where different data sources were added or removed to understand their importance. While the training data is different for each version of the dataset, the validation and test set remain the same for fair comparison (positive to negative ratio is 1–2; baseline precision: 0.33, recall: 1.0, F1: 0.46). The best model for each dataset is selected based on the F1 score. The box represents the interquartile range, the middle line represents the median, the whisker line extends from minimum to maximum values, and the diamond represents outliers. **c**, Standard rank-based metrics of the best model (RotatE) trained on the best training dataset ($FA_{A,E,R,P80}$). Lower is better for mean rank (MR), while higher is better for mean reciprocal rank (MRR), hits@1, hits@3, and hits@10. **d**, Calibration plot showing a high correlation between the probability assigned by the link prediction model and the ground-truth annotations on the test set (n = 5, 5 different random seeds). **e**, **f**, Precision-recall and receiver operating characteristic curves of the best link prediction model.

Phenol-Explorer, 61 % of FDC, and 49 % of FoodMine). There are two main reasons behind it. First, limitations to the NLP LitSense algorithms used by FoodAtlas may limit synonyms and exhaustive tagging of the various entities, co-occurrence of entities in windows that are further away in the text body, and information that is in tables, figures, or supplementary files[61]. Second, the lack of references that are indexed and unique IDs for either foods or chemicals may introduce false positives. Further experimental validation of findings, such as the 11 novel associations with indirect evidence proposed by our link prediction pipeline, will help in accelerating the discovery and achieving completeness of the domain knowledge.

Large language models worked well in the entailment model but not for link prediction. We tested state-of-the-art language models like KG-BERT[62] and KGLM[63] that have better MR metrics compared to the graph-embedding models and are generalizable to unseen entities or relations[64]. For example, we obtained an MR of 191 on the validation set by fine-tuning the KG-BERT architecture with the BioBERT as a pre-trained backbone instead of the BERT, which is a significant improvement over the RotatE MR of 1139. However, those models were not used as other metrics were significantly worse than simpler algorithms like RotatE (MRR: 0.12, hits@1: 0.08, hits@3: 0.11, and hits@10: 0.18), and training/inference time was much longer, making it infeasible to perform proper hyperparameter optimization over our large-scale FAKG. In addition, while the GPT-3.5 performance was impressive even without refinement on domain-specific data, it was not on par with the FoodAtlas pipeline, and the lack of source reference IDs defeats the purpose of one of the main pillars behind FoodAtlas: providing high-quality, trustworthy information with evidence provenance.

We identified a conflicting food-chemical relationship from the link prediction generated hypotheses. In some cases, this supports FoodAtlas's potential to challenge established knowledge and emphasizes the necessity of experimental checking of the solid, established data. For instance, the established absence of beta-carotene synthesis in Atlantic Cod (FDC food 171955) contrasts with a high probability score (0.84 ± 0.09; **Supplementary Data 5**) of the hypothesis (Atlantic Cod, contains, beta-carotene). We sought to reconcile this through literature validation, noting that while unused genes often degrade over time due to natural selection[65], the Atlantic Cod retains the beta,beta-carotene 15,15′-dioxygenase-like enzyme gene. Nevertheless, our further

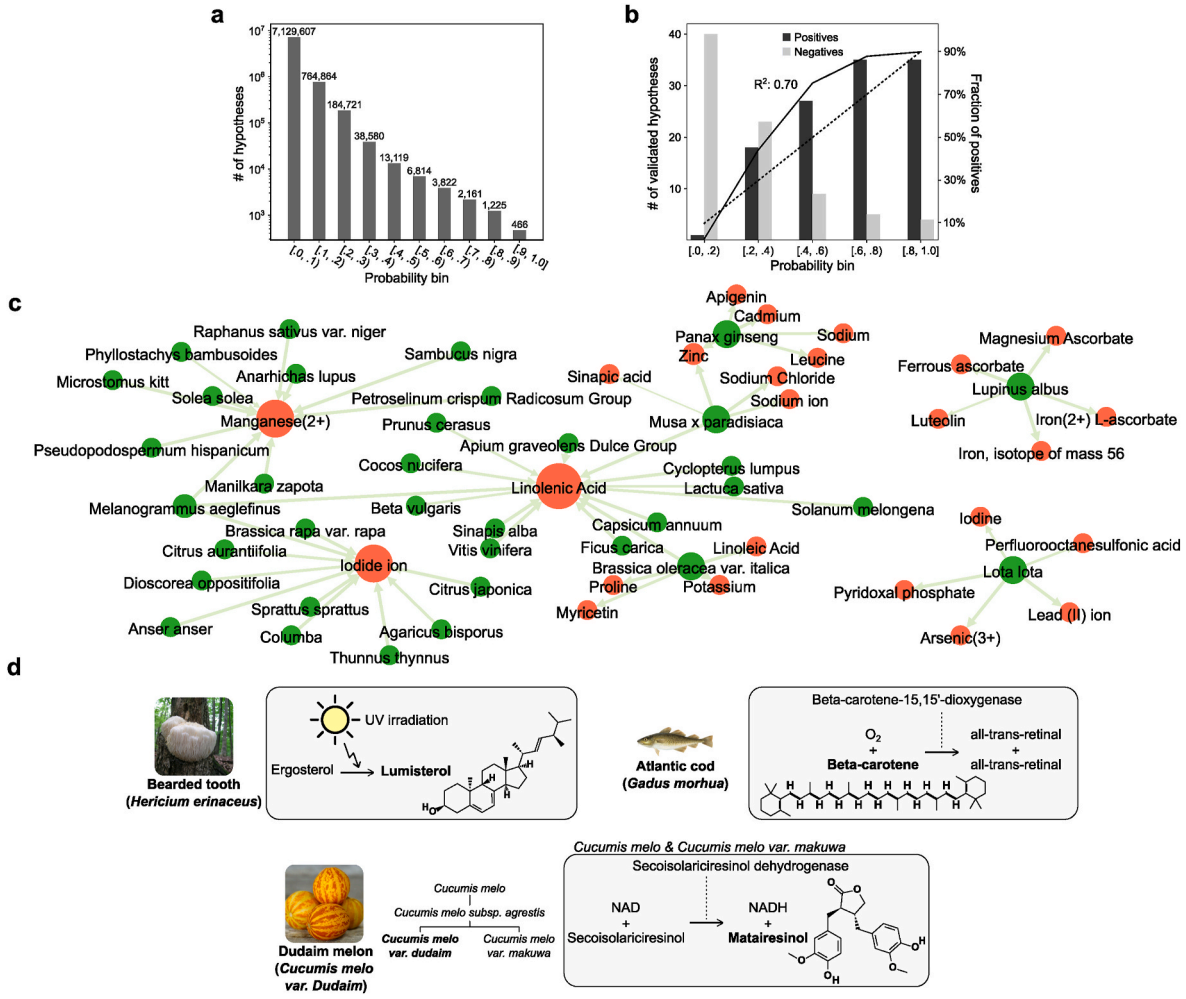

Fig. 6. Validation of link prediction generated hypotheses. a, Distribution of the 8,145,379 hypotheses in 10 equally spaced bins. **b**, Calibration plot of the link prediction model based on randomly selected hypotheses (40 per bin) validated through manual literature search. **c**, Visualization of positively validated link prediction hypotheses, where the 1-hop subgraph of the top 3 chemical and 5 food entities are shown. The edge width is proportionate to its probability score, and the size of the node is proportionate to its degree. **d**, Indirect evidence for the 3 food-chemical relationships not found in the manual literature search and suggested by the link prediction pipeline, where the food and chemical of interest are marked in bold.

investigation considered the Atlantic Cod's diet, particularly during its larval stage, which predominantly consists of crustaceans[66] rich in beta-carotene[67]. Thus, there exists a plausible dietary source for beta-carotene incorporation. Additionally, we discovered literature referencing the detection of beta-carotene in commercially processed cod liver oil, albeit the exact species of cod (Gadus morhua-Atlantic Cod or others like Gadus macrocephalus-Pacific Cod) was not specified[68].

The next version of FoodAtlas will address current limitations in data type, structure, and information source. First, we will work towards extending information extraction to the chemical concentration value in their source food (*e.g.*, cocoa contains 564 mg/serving of epicatechin) [69]. Second, not all information sources are equal, and we plan to introduce a quality score using the source trustworthiness[70]. Third, it was required that the food and chemical entities in the KG have a unique NCBI ID and PubChem ID, respectively, to ensure compatibility with existing data. Although food ontologies like FoodOn[71] exist, we need to create and adopt a unified vocabulary of all foods and food parts that can be further extended to include processes for processed foods. Fourth, expanding FoodAtlas to capture health conditions through a UMLS ontology[72] and dosage effects can link foods, ingredients, and their health effects in a way that can be useful for the discovery of new food-related bioactive compounds and sources[73], food formulation and substitutions[74], personalized diet recommendations(Y. [75]), among others. Compound information sources can also be expanded so that we include in more detail classes of molecules, such as terpenes, polyphenols, and peptides, that are of high interest[76]. Fifth, the identification and augmentation of data that the entailment model has difficulty handling, such as domain-specific hypotheses and complex sentence structures, could lead to improved performance[77,78]. Sixth, to allow our KG to continuously expand and improve along with new incoming publications, we will incorporate a never-ending learning scheme similar to NELL[79] to allow existing knowledge to infer new information and extend ontology. Recent work[80] also shows large language models, such as GPT-4[81], can be used for ontology expansion. Finally, we will investigate further the use of pre-trained and fine-tuned large language models with active learning strategies, including those based on the network analysis indicators like centrality and modularity[82,83], as the field has in the past few months produced striking results when it comes to efficiency, robustness, and scalability. We believe that the application of cutting-edge AI tools domains where computational science penetration has been traditionally limited, has the potential to revolutionize and pave the way for a paradigm-shift in those industries, with far reaching implications for our society and planet.

Data availability

All code and data are available at https://github.com/IBPA/FoodAtlas and http://foodatlas.ai.

CRediT authorship contribution statement

Jason Youn: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Fangzhou Li:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation. **Gabriel Simmons:** Methodology, Formal analysis. **Shanghyeon Kim:** Validation, Investigation. **Ilias Tagkopoulos:** Writing – review & editing, Supervision, Resources, Methodology, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank the members of the Tagkopoulos lab and Danielle Lemay from the U.S. Department of Agriculture Agricultural Research Service (USDA ARS) for helpful discussions and comments. We would also like to thank Alexis Allot from National Center for Biotechnology Information (NCBI) for running LitSense queries internally, Kyle McKillop and Kai Blumberg from the USDA ARS for providing FoodData Central (FDC) data, Anders Poulsen from the Technical University of Denmark (DTU) for providing the Frida data, Navneet Rai and Adil Muhammad from the Tagkopoulos lab for PH pair annotation, and Arielle Yoo for link prediction validation and analysis. This work was supported by the USDA-NIFA AI Institute for Next Generation Food Systems (AIFS), USDA-NIFA award number 2020-67021-32855.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109072.

References

[1] A.-L. Barabási, G. Menichetti, J. Loscalzo, The unmapped chemical complexity of our diet, Nature Food 1 (1) (2020) 33–37.
[2] I. Elmadfa, A.L. Meyer, Importance of food composition data to nutrition and public health, Eur. J. Clin. Nutr. 64 (3) (2010), https://doi.org/10.1038/ejcn.2010.202. Article 3.
[3] M. Diana, J. Quílez, M. Rafecas, Gamma-aminobutyric acid as a bioactive compound in foods: a review, J. Funct.Foods 10 (2014) 407–420, https://doi.org/10.1016/j.jff.2014.07.004.
[4] P. Reboredo-Rodríguez, M. Figueiredo-González, C. González-Barreiro, J. Simal-Gándara, M.D. Salvador, B. Cancho-Grande, G. Fregapane, State of the art on functional virgin olive oils enriched with bioactive compounds and their properties, Int. J. Mol. Sci. 18 (3) (2017), https://doi.org/10.3390/ijms18030668. Article 3.
[5] A. Eetemadi, N. Rai, B.M.P. Pereira, M. Kim, H. Schmitz, I. Tagkopoulos, The computational diet: a review of computational methods across diet, microbiome, and health, Front. Microbiol. 11 (2020). https://www.frontiersin.org/articles/10.3389/fmicb.2020.00393.
[6] A. Eetemadi, I. Tagkopoulos, Methane and fatty acid metabolism pathways are predictive of Low-FODMAP diet efficacy for patients with irritable bowel syndrome, Clinical Nutrition (Edinburgh, Scotland) 40 (6) (2021) 4414–4421, https://doi.org/10.1016/j.clnu.2020.12.041.
[7] J. Gan, J.B. Siegel, J.B. German, Molecular annotation of food – towards personalized diet and precision health, Trends Food Sci. Technol. 91 (2019) 675–680, https://doi.org/10.1016/j.tifs.2019.07.016.
[8] K. McKillop, J. Harnly, P. Pehrsson, N. Fukagawa, J. Finley, FoodData central, USDA's updated approach to food composition data systems, Curr. Dev. Nutr. 5 (Supplement_2) (2021) 596, https://doi.org/10.1093/cdn/nzab044_027.
[9] Ciqual. (n.d.). Retrieved July 15, 2024, from https://ciqual.anses.fr/.
[10] M. Kapsokefalou, M. Roe, A. Turrini, H.S. Costa, E. Martinez-Victoria, L. Marletta, R. Berry, P. Finglas, Food composition at present: new challenges, Nutrients 11 (8) (2019), https://doi.org/10.3390/nu11081714. Article 8.
[11] A. Scalbert, L. Brennan, C. Manach, C. Andres-Lacueva, L.O. Dragsted, J. Draper, S. M. Rappaport, J.J.J. van der Hooft, D.S. Wishart, The food metabolome: a window over dietary exposure, Am. J. Clin. Nutr. 99 (6) (2014) 1286–1308, https://doi.org/10.3945/ajcn.113.076133.
[12] Wishart, D. (n.d.). FooDB Version 1.0. Retrieved February 6, 2023, from https://foodb.ca/.
[13] N.K. Rakhi, R. Tuwani, J. Mukherjee, G. Bagler, Data-driven analysis of biomedical literature suggests broad-spectrum benefits of culinary herbs and spices, PLoS One 13 (5) (2018) e0198030, https://doi.org/10.1371/journal.pone.0198030.
[14] F.M. Afendi, T. Okada, M. Yamazaki, A. Hirai-Morita, Y. Nakamura, K. Nakamura, S. Ikeda, H. Takahashi, Md Altaf-Ul-Amin, L.K. Darusman, K. Saito, S. Kanaya, KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research, Plant Cell Physiol. 53 (2) (2012) e1, https://doi.org/10.1093/pcp/pcr165.
[15] Dr. Duke's Phytochemical and Ethnobotanical Databases. Retrieved July 15, 2024, from https://phytochem.nal.usda.gov/.
[16] V. Neveu, J. Perez-Jiménez, F. Vos, V. Crespy, L. du Chaffaut, L. Mennen, C. Knox, R. Eisner, J. Cruz, D. Wishart, A. Scalbert, Phenol-Explorer: an online comprehensive database on polyphenol contents in foods, Database 2010 (2010) bap024, https://doi.org/10.1093/database/bap024.
[17] J.A. Rothwell, M. Urpi-Sarda, M. Boto-Ordoñez, C. Knox, R. Llorach, R. Eisner, J. Cruz, V. Neveu, D. Wishart, C. Manach, C. Andres-Lacueva, A. Scalbert, Phenol-Explorer 2.0: a major update of the Phenol-Explorer database integrating data on polyphenol metabolism and pharmacokinetics in humans and experimental animals, Database 2012 (2012) bas031, https://doi.org/10.1093/database/bas031.
[18] J.A. Rothwell, J. Perez-Jimenez, V. Neveu, A. Medina-Remón, N. M'Hiri, P. García-Lobato, C. Manach, C. Knox, R. Eisner, D.S. Wishart, A. Scalbert, Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content, Database 2013 (2013) bat070, https://doi.org/10.1093/database/bat070.
[19] A. B. da Silva, F. Giacomoni, B. Pavot, Y. Fillatre, J. Rothwell, B.B. Sualdea, C. Veyrat, R. Garcia-Villalba, C. Gladine, R. Kopec, P. Hollman, R. Landberg, C. Morand, C. N. dos Santos, L. Nyström, E. Pujos-Guillot, M. Bronze, F. Tomas-Barberan, M. Urpi-Sarda, C. Manach, PhytoHub V1.4: a new release for the online database dedicated to food phytochemicals and their human metabolites. https://hal.archives-ouvertes.fr/hal-01607427, 2016.
[20] J. White, PubMed 2.0, Med. Ref. Serv. Q. 39 (4) (2020) 382–387, https://doi.org/10.1080/02763869.2020.1826228.
[21] R.J. Roberts, PubMed central: the GenBank of the published literature, Proc. Natl. Acad. Sci. USA 98 (2) (2001) 381–382, https://doi.org/10.1073/pnas.98.2.381.
[22] X. Chen, S. Jia, Y. Xiang, A review: knowledge reasoning over knowledge graph, Expert Syst. Appl. 141 (2020) 112948, https://doi.org/10.1016/j.eswa.2019.112948.
[23] J. Chen, H. Dong, J. Hastings, E. Jiménez-Ruiz, V. López, P. Monnin, C. Pesquita, P. Škoda, V. Tamma, Knowledge graphs for the life sciences: recent developments, challenges and opportunities, *DROPS-IDN/v2/Document/10.4230/TGDK.1.1.5*, https://doi.org/10.4230/TGDK.1.1.5, 2023.
[24] N. Zhang, Z. Bi, X. Liang, S. Cheng, H. Hong, S. Deng, J. Lian, Q. Zhang, H. Chen, OntoProtein: protein pretraining with gene ontology embedding, arXiv: 2201.11147 (2022), https://doi.org/10.48550/arXiv.2201.11147 arXiv.
[25] G. Cenikj, L. Strojnik, R. Angelski, N. Ogrinc, B. Koroušić Seljak, T. Eftimov, From language models to large-scale food and biomedical knowledge graphs, Sci. Rep. 13 (1) (2023), https://doi.org/10.1038/s41598-023-34981-4. Article 1.
[26] L.D. Dang, U.T.P. Phan, N.T.H. Nguyen, GENA: a knowledge graph for nutrition and mental health, J. Biomed. Inf. 145 (2023) 104460, https://doi.org/10.1016/j.jbi.2023.104460.
[27] A.D. Diaz Gonzalez, K.S. Hughes, S. Yue, S.T. Hayes, Applying BioBERT to extract germline gene-disease associations for building a knowledge graph from the biomedical literature. 2023 the 7th International Conference on Information System and Data Mining (ICISDM), 2023, pp. 37–42, https://doi.org/10.1145/3603765.3603771.
[28] A. Harnoune, M. Rhanoui, M. Mikram, S. Yousfi, Z. Elkaimbillah, B. El Asri, BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis, Computer Methods and Programs in Biomedicine Update 1 (2021) 100042, https://doi.org/10.1016/j.cmpbup.2021.100042.
[29] S. Haussmann, O. Seneviratne, Y. Chen, Y. Ne'eman, J. Codella, C.-H. Chen, D. L. McGuinness, M.J. Zaki, FoodKG: a semantics-driven knowledge graph for food recommendation, in: C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, F. Gandon (Eds.), The Semantic Web – ISWC 2019, vol. 11779, Springer International Publishing, 2019, pp. 146–162, https://doi.org/10.1007/978-3-030-30796-7_10.
[30] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J.F. Rousseau, X. Li, W. Xu, V. I. Torvik, Y. Bu, C. Chen, I.A. Ebeid, D. Li, Y. Ding, Building a PubMed knowledge graph, Sci. Data 7 (1) (2020), https://doi.org/10.1038/s41597-020-0543-2. Article 1.
[31] Z. Ahmad, A. Ekbal, S. Sengupta, A. Mitra, R. Rammani, P. Bhattacharyya, Active learning based relation classification for knowledge graph construction from conversation data, in: H. Yang, K. Pasupa, A.C.-S. Leung, J.T. Kwok, J.H. Chan, I. King (Eds.), Neural Information Processing, Springer International Publishing, 2020, pp. 617–625, https://doi.org/10.1007/978-3-030-63820-7_70.
[32] P. Ren, W. Hou, M. Sheng, X. Li, C. Li, Y. Zhang, MKGB: a medical knowledge graph construction framework based on data lake and active learning, in: S. Siuly,

H. Wang, L. Chen, Y. Guo, C. Xing (Eds.), Health Information Science, vol. 13079, Springer International Publishing, 2021, pp. 245–253, https://doi.org/10.1007/978-3-030-90885-0_22.
[33] L. Sun, W. Hu, K. Xu, Y. Chen, Q. Sun, J. Wang, ASRC:A knowledge graph relation construction model based on active learning and semantic recognition. 2022 IEEE International Conference on Big Data (Big Data), 2022, pp. 6025–6029, https://doi.org/10.1109/BigData55660.2022.10020502.
[34] C.L. Schoch, S. Ciufo, M. Domrachev, C.L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J.P. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, NCBI Taxonomy: a comprehensive update on curation, resources and tools, Database: The Journal of Biological Databases and Curation 2020 (2020) baaa062, https://doi.org/10.1093/database/baaa062.
[35] A. Allot, Q. Chen, S. Kim, R. Vera Alvarez, D.C. Comeau, W.J. Wilbur, Z. Lu, LitSense: making sense of biomedical literature at sentence level, Nucleic Acids Res. 47 (W1) (2019) W594–W599, https://doi.org/10.1093/nar/gkz289.
[36] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019), https://doi.org/10.1093/bioinformatics/btz682 btz682.
[37] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, D. Amodei, Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[38] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805) (2019), https://doi.org/10.48550/arXiv.1810.04805 arXiv.
[39] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, *RoBERTa: a robustly optimized BERT pretraining approach* (arXiv:1907.11692), arXiv (2019), https://doi.org/10.48550/arXiv.1907.11692.
[40] National Food Institute, Technical University of Denmark. (n.d.). Food data (frida.fooddata.dk), version 4.2, 2022. Retrieved May 19, 2023, from https://frida.fooddata.dk/.
[41] A. Rossi, D. Barbosa, D. Firmani, A. Matinata, P. Merialdo, Knowledge graph embedding for link prediction: a comparative analysis, ACM Trans. Knowl. Discov. Data 15 (2) (2021) 14:1–14:49, https://doi.org/10.1145/3424672.
[42] M. Ali, M. Berrendorf, C.T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, J. Lehmann, PyKEEN 1.0: a Python library for training and evaluating knowledge graph embeddings, J. Mach. Learn. Res. 22 (1) (2021) 3723–3782, 82,:3728.
[43] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, Adv. Neural Inf. Process. Syst. 26 (2013), in: https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.
[44] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang, Knowledge vault: a web-scale approach to probabilistic knowledge fusion. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 601–610, https://doi.org/10.1145/2623330.2623623.
[45] B. Yang, W. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, arXiv:1412.6575) (2015), https://doi.org/10.48550/arXiv.1412.6575 arXiv.
[46] G. Ji, S. He, L. Xu, K. Liu, J. Zhao, Knowledge graph embedding via dynamic mapping matrix. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 687–696, https://doi.org/10.3115/v1/P15-1067.
[47] T. Trouillon, J. Welbl, S. Riedel, E. Gaussier, G. Bouchard, Complex embeddings for simple link prediction. Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 2071–2080, in: https://proceedings.mlr.press/v48/trouillon16.html.
[48] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, RotatE: knowledge graph embedding by relational rotation in complex space, arXiv:1902.10197 (2019), https://doi.org/10.48550/arXiv.1902.10197 arXiv.
[49] J. Youn, I. Tagkopoulos, KGLM: integrating knowledge graph structure in language models for link prediction, arXiv:2211.02744) (2022), https://doi.org/10.48550/arXiv.2211.02744 arXiv.
[50] S. Kim, T. Cheng, S. He, P.A. Thiessen, Q. Li, A. Gindulyte, E.E. Bolton, PubChem protein, gene, pathway, and taxonomy data collections: bridging biology and chemistry through target-centric views of PubChem data, J. Mol. Biol. 434 (11) (2022) 167514, https://doi.org/10.1016/j.jmb.2022.167514.
[51] F. Hooton, G. Menichetti, A.-L. Barabási, Exploring food contents in scientific literature with FoodMine, Sci. Rep. 10 (1) (2020), https://doi.org/10.1038/s41598-020-73105-0. Article 1.
[52] R. Reifen, A. Karlinsky, A.H. Stark, Z. Berkovich, A. Nyska, α-Linolenic acid (ALA) is an anti-inflammatory agent in inflammatory bowel disease, J. Nutr. Biochem. 26 (12) (2015) 1632–1640, https://doi.org/10.1016/j.jnutbio.2015.08.006.
[53] P. Singer, W. Jaeger, I. Berger, H. Barleben, M. Wirth, E. Richter-Heinrich, S. Voigt, W. Gödicke, Effects of dietary oleic, linoleic and alpha-linolenic acids on blood pressure, serum lipids, lipoproteins and the formation of eicosanoid precursors in patients with mild essential hypertension, J. Hum. Hypertens. 4 (3) (1990) 227–233.
[54] G.D. Lawrence, D.T. Sawyer, The chemistry of biological manganese, Coord. Chem. Rev. 27 (2) (1978) 173–193, https://doi.org/10.1016/S0010-8545(00)80358-6.
[55] V.L. Schramm, Manganese in Metabolism and Enzyme Function, Elsevier, 2012.
[56] J.D. Aguirre, V.C. Culotta, Battles with iron: manganese in oxidative stress protection, J. Biol. Chem. 287 (17) (2012) 13541–13548, https://doi.org/10.1074/jbc.R111.312181.
[57] A. Nagao, M. Maeda, B.P. Lim, H. Kobayashi, J. Terao, Inhibition of β-carotene-15,15′-dioxygenase activity by dietary flavonoids, J. Nutr. Biochem. 11 (6) (2000) 348–355, https://doi.org/10.1016/S0955-2863(00)00090-5.
[58] J. Garcia-Mas, A. Benjak, W. Sanseverino, M. Bourgeois, G. Mir, V.M. González, E. Hénaff, F. Câmara, L. Cozzuto, E. Lowy, T. Alioto, S. Capella-Gutiérrez, J. Blanca, J. Cañizares, P. Ziarsolo, D. Gonzalez-Ibeas, L. Rodríguez-Moreno, M. Droege, L. Du, P. Puigdomènech, The genome of melon (*Cucumis melo* L.), Proc. Natl. Acad. Sci. USA 109 (29) (2012) 11872–11877, https://doi.org/10.1073/pnas.1205415109.
[59] P. Joradon, V. Rungsardthong, U. Ruktanonchai, K. Suttisintong, T. Iempridee, B. Thumthanaruk, S. Vatanyoopaisarn, N. Sumonsiri, D. Uttapap, Ergosterol content and antioxidant activity of lion's mane mushroom (hericium erinaceus) and its induction to vitamin D2 by UVC-irradiation. Proceedings of the 8th International Conference on Agricultural and Biological Sciences, 2022, pp. 19–28, https://doi.org/10.5220/0011594600003430.
[60] Y. Sun, F.K. Nzekoue, S. Vittori, G. Sagratini, G. Caprioli, Conversion of ergosterol into vitamin D2 and other photoisomers in Agaricus bisporus mushrooms under UV-C irradiation, Food Biosci. 50 (2022) 102143, https://doi.org/10.1016/j.fbio.2022.102143.
[61] J. Herzig, P.K. Nowak, T. Müller, F. Piccinno, J.M. Eisenschlos, TAPAS: weakly supervised table parsing via pre-training. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4320–4333, https://doi.org/10.18653/v1/2020.acl-main.398.
[62] L. Yao, C. Mao, Y. Luo, *KG-BERT: BERT for knowledge graph completion* (arXiv: 1909.03193), arXiv (2019), https://doi.org/10.48550/arXiv.1909.03193.
[63] J. Youn, I. Tagkopoulos, KGLM: integrating knowledge graph structure in language models for link prediction, in: A. Palmer, J. Camacho-collados (Eds.), Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Association for Computational Linguistics, 2023, pp. 217–224, https://doi.org/10.18653/v1/2023.starsem-1.20.
[64] H. Zha, Z. Chen, X. Yan, Inductive relation prediction by BERT, Proc. AAAI Conf. Artif. Intell. 36 (5) (2022), https://doi.org/10.1609/aaai.v36i5.20537. Article 5.
[65] R. Albalat, C. Cañestro, Evolution by gene loss, Nat. Rev. Genet. 17 (7) (2016), https://doi.org/10.1038/nrg.2016.39. Article 7.
[66] K. Hamre, Nutrition in cod (Gadus morhua) larvae and juveniles, ICES (Int. Counc. Explor. Sea) J. Mar. Sci. 63 (2) (2006) 267–274, https://doi.org/10.1016/j.icesjms.2005.11.011.
[67] T. Maoka, Carotenoids in marine animals, Mar. Drugs 9 (2) (2011), https://doi.org/10.3390/md9020278. Article 2.
[68] S. Luterotti, M. Franko, D. Bicanic, Ultrasensitive determination of β-carotene in fish oil-based supplementary drugs by HPLC-TLS, J. Pharmaceut. Biomed. Anal. 21 (5) (1999) 901–909, https://doi.org/10.1016/S0731-7085(99)00185-5.
[69] A. Crozier, I.B. Jaganath, M.N. Clifford, Dietary phenolics: chemistry, bioavailability and effects on health, Nat. Prod. Rep. 26 (8) (2009) 1001–1043, https://doi.org/10.1039/B802662A.
[70] H. Kyngäs, M. Kääriäinen, S. Elo, The trustworthiness of content analysis, in: H. Kyngäs, K. Mikkonen, M. Kääriäinen (Eds.), The Application of Content Analysis in Nursing Science Research, Springer International Publishing, 2020, pp. 41–48, https://doi.org/10.1007/978-3-030-30199-6_5.
[71] D.M. Dooley, E.J. Griffiths, G.S. Gosal, P.L. Buttigieg, R. Hoehndorf, M.C. Lange, L. M. Schriml, F.S.L. Brinkman, W.W.L. Hsiao, FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, Npj Science of Food 2 (1) (2018), https://doi.org/10.1038/s41538-018-0032-6. Article 1.
[72] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270, https://doi.org/10.1093/nar/gkh061.
[73] W. Min, C. Liu, L. Xu, S. Jiang, Applications of knowledge graphs for food science and industry, Patterns 3 (5) (2022) 100484, https://doi.org/10.1016/j.patter.2022.100484.
[74] A. Ławrynowicz, A. Wróblewska, W.T. Adrian, B. Kulczyński, A. Gramza-Michałowska, Food recipe ingredient substitution ontology design pattern, Sensors 22 (3) (2022), https://doi.org/10.3390/s22031095. Article 3.
[75] Y. Chen, A. Subburathinam, C.-H. Chen, M.J. Zaki, Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 544–552, https://doi.org/10.1145/3437963.3441816.
[76] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, ChEBI: a database and ontology for chemical entities of biological interest, Nucleic Acids Res. 36 (suppl_1) (2008) D344–D350, https://doi.org/10.1093/nar/gkm791.
[77] A.M. Issifu, M.C. Ganiz, A simple data augmentation method to improve the performance of named entity recognition models in medical domain. 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 763–768, https://doi.org/10.1109/UBMK52708.2021.9558986.
[78] T. Kang, A. Perotte, Y. Tang, C. Ta, C. Weng, UMLS-based data augmentation for natural language processing of clinical research literature, J. Am. Med. Inf. Assoc.: JAMIA 28 (4) (2021) 812–823, https://doi.org/10.1093/jamia/ocaa309.
[79] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, J. Welling, Never-

ending learning, Commun. ACM 61 (5) (2018) 103–115, https://doi.org/10.1145/3191513.
[80] S. Toro, A.V. Anagnostopoulos, S. Bello, K. Blumberg, R. Cameron, L. Carmody, A. D. Diehl, D. Dooley, W. Duncan, P. Fey, P. Gaudet, N.L. Harris, M. Joachimiak, L. Kiani, T. Lubiana, M.C. Munoz-Torres, S. O'Neil, D. Osumi-Sutherland, A. Puig, C.J. Mungall, Dynamic retrieval augmented generation of ontologies using artificial intelligence (DRAGON-AI), arXiv:2312.10904) (2024), https://doi.org/10.48550/arXiv.2312.10904 arXiv.
[81] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, B. Zoph, OpenAI, GPT-4 technical report, arXiv:2303.08774 (2024), https://doi.org/10.48550/arXiv.2303.08774 arXiv.
[82] F.A. Rodrigues, Network centrality: an introduction, in: E.E.N. Macau (Ed.), A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems, Springer International Publishing, 2019, pp. 177–196, https://doi.org/10.1007/978-3-319-78512-7_10.
[83] G.P. Wagner, M. Pavlicev, J.M. Cheverud, The road to modularity, Nat. Rev. Genet. 8 (12) (2007), https://doi.org/10.1038/nrg2267. Article 12.
[84] O.M.E. Abdel-Salam, N.A. Salem, M. El-Sayed El-Shamarka, N. Al-Said Ahmed, J. Seid Hussein, Z.A. El-Khyat, Cannabis-induced impairment of learning and memory: effect of different nootropic drugs, EXCLI Journal 12 (2013) 193–214.
[85] R.D. Steigerwalt, G. Belcaro, P. Morazzoni, E. Bombardelli, C. Burki, F. Schönlau, Mirtogenol potentiates latanoprost in lowering intraocular pressure and improves ocular blood flow in asymptomatic subjects, Clin. Ophthalmol. 4 (2010) 471–476, https://doi.org/10.2147/opth.s9899.
[86] G. Zhang, X. Jiang, Y. Liu, X. Hao, Y. Wang, X. Yan, N. Yuan, Y. Ma, M. Ma, Therapeutic efficiency of an external Chinese herbal formula of mammary precancerous lesions by BATMAN-TCM online bioinformatics analysis tool and experimental validation, Evid. base Compl. Alternative Med. : eCAM 2019 (2019) 2795010, https://doi.org/10.1155/2019/2795010.
[87] C. Bernier, C. Goetz, E. Jubinville, J. Jean, The new face of berries: a review of their antiviral proprieties, Foods 11 (1) (2021) 102, https://doi.org/10.3390/foods11010102.
[88] Ó. Monroig, R. de Llanos, I. Varó, F. Hontoria, D.R. Tocher, S. Puig, J.C. Navarro, Biosynthesis of polyunsaturated fatty acids in Octopus vulgaris: molecular cloning and functional characterisation of a stearoyl-CoA desaturase and an elongation of very long-chain fatty acid 4 protein, Mar. Drugs 15 (3) (2017) 82, https://doi.org/10.3390/md15030082.
[89] I.R. Schultz, A. Skillman, D. Woodruff, Domoic acid excretion in dungeness crabs, razor clams and mussels, Mar. Environ. Res. 66 (1) (2008) 21–23, https://doi.org/10.1016/j.marenvres.2008.02.012.
[90] R. Yazdanparast, M. Alavi, Antihyperlipidaemic and antihypercholesterolaemic effects of Anethum graveolens leaves after the removal of furocoumarins, Cytobios 105 (410) (2001) 185–191.
[91] A.B. Kruse, C.D. Kowalski, S. Leuthold, K. Vach, P. Ratka-Krüger, J.P. Woelber, What is the impact of the adjunctive use of omega-3 fatty acids in the treatment of periodontitis? A systematic review and meta-analysis, Lipids Health Dis. 19 (2020) 100, https://doi.org/10.1186/s12944-020-01267-x.