Journal of Biomedical Informatics 143 (2023) 104398

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

A novel gluten knowledge base of potential biomedical and health-related interactions extracted from the literature: Using machine learning and graph analysis methodologies to reconstruct the bibliome

Martín Pérez-Pérez [a,b,*], Tânia Ferreira [c,d], Gilberto Igrejas [c,d,e], Florentino Fdez-Riverola [a,b]

[a] CINBIO, Universidade de Vigo, Department of Computer Science, ESEI – Escuela Superior de Ingeniería Informática, 32004 Ourense, Spain
[b] SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Spain
[c] Department of Genetics and Biotechnology, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[d] Functional Genomics and Proteomics Unit, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[e] LAQV-REQUIMTE, Faculty of Science and Technology, Nova University of Lisbon, Lisbon, Portugal

ARTICLE INFO

Keywords:
Literature curation
Knowledge representation
Ontology-base methods
Relation extraction
Social media
Gluten database

ABSTRACT

Background: In return for their nutritional properties and broad availability, cereal crops have been associated with different alimentary disorders and symptoms, with the majority of the responsibility being attributed to gluten. Therefore, the research of gluten-related literature data continues to be produced at ever-growing rates, driven in part by the recent exploratory studies that link gluten to non-traditional diseases and the popularity of gluten-free diets, making it increasingly difficult to access and analyse practical and structured information. In this sense, the accelerated discovery of novel advances in diagnosis and treatment, as well as exploratory studies, produce a favourable scenario for disinformation and misinformation.

Objectives: Aligned with, the European Union strategy "*Delivering on EU Food Safety and Nutrition in 2050*" which emphasizes the inextricable links between imbalanced diets, the increased exposure to unreliable sources of information and misleading information, and the increased dependency on reliable sources of information; this paper presents GlutKNOIS, a public and interactive literature-based database that reconstructs and represents the experimental biomedical knowledge extracted from the gluten-related literature. The developed platform includes different external database knowledge, bibliometrics statistics and social media discussion to propose a novel and enhanced way to search, visualise and analyse potential biomedical and health-related interactions in relation to the gluten domain.

Methods: For this purpose, the presented study applies a semi-supervised curation workflow that combines natural language processing techniques, machine learning algorithms, ontology-based normalization and integration approaches, named entity recognition methods, and graph knowledge reconstruction methodologies to process, classify, represent and analyse the experimental findings contained in the literature, which is also complemented by data from the social discussion.

Results and conclusions: In this sense, 5814 documents were manually annotated and 7424 were fully automatically processed to reconstruct the first online gluten-related knowledge database of evidenced health-related interactions that produce health or metabolic changes based on the literature. In addition, the automatic processing of the literature combined with the knowledge representation methodologies proposed has the potential to assist in the revision and analysis of years of gluten research. The reconstructed knowledge base is public and accessible at https://sing-group.org/glutknois/

1. Introduction

Cereal crops are a very important part of the human diet and were introduced about 10.000 years ago, during the transition from hunting to settled farming. Since then, the consumption of cereals has increased, and in 1941 the Nutrition Society established as its main objective the

* Corresponding author at: Department of Computer Science, ESEI – Escuela Superior de Ingeniería Informática, 32004 Ourense, Spain.
E-mail addresses: martiperez@uvigo.es (M. Pérez-Pérez), tania.rmf@hotmail.com (T. Ferreira), gigrejas@utad.pt (G. Igrejas), riverola@uvigo.es (F. Fdez-Riverola).

https://doi.org/10.1016/j.jbi.2023.104398
Received 17 October 2022; Received in revised form 12 May 2023; Accepted 15 May 2023
Available online 23 May 2023

improvement of wheat cultivation, to study how its nutrition and applications could help in health maintenance [1]. Nowadays, wheat production reaches 700 million tons per year [2].

In return for their nutritional properties and broad availability, cereal crops have been associated with different disorders and symptoms, with the majority of the responsibility being attributed to gluten, a high molecular weight protein found in wheat, barley and rye. Gluten-related disorders can be classified into three categories: (*i*) autoimmune, which includes celiac disease (CD), dermatitis herpetiformis (DH) and gluten ataxia (GA); (*ii*) allergic reactions, which includes wheat allergy; and (*iii*) immune-mediated in the form of non-coeliac gluten sensitivity (NCGS) [1]. Nevertheless, nowadays there are also an incremental number of exploratory and experimental studies that relate gluten involvement to various psychiatric illnesses such as autism [3] and schizophrenia [4]; and chronic diseases such as diabetes type 1 [5] or irritable bowel syndrome [6].

In this sense, gluten-related diseases are one of the most common and studied autoimmune disorders due to (*i*) their reported prevalence of 0.5–1% of the general population [7], (*ii*) their fastest-growing as one of the most common autoimmune disorders in the last years and, in addition, (*iii*) their high prevalence as one of the most common genetic diseases in the West [7–9]. To date, the only therapy available for gluten-related disorders is a lifelong gluten-free diet (GFD) and its adherence leads to the mitigation of the symptoms in most cases. However, nowadays, a great number of consumers are following a GFD as a self-prescribed lifestyle, although most of them have not been previously diagnosed with a related disease; which further increases the number of studies related to this protein [10,11]. In the last years, gluten-free diets have grown in popularity strongly promoted by: (*i*) the widespread media coverage, (*ii*) the strong economic interests and (*iii*) the considerable amount of disinformation and misinformation [12,13]. For example, nearly 50% of 910 athletes (including world-class and Olympic medallists) adhere to GFD because they perceive it as more healthy and provides energy benefits [14]. The issue here is that patients increasingly seek information from non-traditional and fringe groups, suggesting susceptibility to misinformation and disinformation. Besides, these unreliable information sources are not distinguishable by most patients.

In terms of health and nutrition, the internet and social media are increasingly being used by different interested individuals and patients to acquire new information, discover experimental disease treatments, and share experiences on personal health symptoms.

Social media technologies are inexpensive, accessible, and user-friendly and can attract a large proportion of health consumers becoming an excellent channel to disseminate and support information [15]. Nevertheless, several authors have already reported that exists a high risk of misinformation in social media, both video and text-based, which can mislead patients affecting their health safety [16–19]. Research studies could be difficult to retrieve and understand by the general population. Besides experimental and exploratory studies can be misunderstood and lead to the spread of misinformation. Food faddism, or an exaggerated belief in the effects of food or nutrition on health or disease fuels nutrition fraud. So, one of the most significant societal concerns remains to achieve food security for a healthy population. In this regard, the European Union strategy "*Delivering on EU Food Safety and Nutrition in 2050*" [20] emphasizes the inextricable links between: (i) the imbalanced diets due to over-reliance on (perceived) "healthy foods" or specific dietary regimes, (ii) the inadequate food safety and nutrition literacy, loss of food traditions and increased exposure to unreliable sources of information, (iii) the abundance of voluntary food material and increased opportunity for misleading information, and (iv) the increased consumer dependency on digital services for dietary choices, with healthy societies and the avoidance of future illnesses.

In terms of online information sources, the quality of the data is often unfiltered, becoming inaccurate and can mislead patients and lead to unhealthy choices [21,22]. Therefore, in a recent study, some researchers exposed how the respondents of their work could only identify about 52.4% of the food safety misinformation on average, and only 0.5% of the respondents had correctly identified all the misinformation [23]. In the same line, this other work analysed 98 celiac-related websites to evaluate their comprehensiveness, accuracy, transparency, and readability [24]. The authors determined that the knowledge provided by many websites was not sufficiently accurate, comprehensive, and transparent, or presented at an appropriate reading grade level, to be considered sufficiently trustworthy and reliable for patients, health care providers, CD support groups, and the general public. This susceptible situation leads patients and healthy people to make wrong decisions that can lead to nutritional risk behaviours associated with inadequate macronutrient intake and dietary imbalances for the general population who is adhered to a GFD without a medical prescription [25].

To make matters more complex, the information overload in the bibliome already exceeds the ability of researchers to digest it. The number of scientific articles published in the biomedical domain is growing at an unprecedented rate, further complicating access to structured knowledge within a particular domain of study [26].

So, keeping up to date about the different experimental developments presents an immense challenge to individuals and scientists, due to the tedious and time-consuming manual reading and analysis of a large amount of biomedical literature research published every day. Although, this triage process is unequivocally required by the general public to be informed and by scientists in order to: (*i*) advance in the identification of the most susceptible biomarkers, (*ii*) improve the efficacy of the current treatments, (*iii*) keep updated on novel discoveries and (*iv*) contrast and legitimize the novel hypothesis.

Consequently, manual-curated scientific knowledge or gold standards becomes undoubtedly a valuable resource. For this reason, there is an increased interest and effort to assist researchers in the curation, structuring and analysing of a vast amount of literature articles looking for relevant health-related semantic knowledge. Discover insights into genetic alterations and signalling pathways, disease comorbidities and interactions between the metabolism and specific genes or compounds are some relevant examples.

In this situation, literature-based knowledge reconstruction supported by semi-automated extraction tasks based on computational text-mining methodologies and knowledge graph visualisation techniques appears to have the potential to enhance human productivity and reduce the time-consuming nature of reviewing, curating and structuring the literature [27–30].

Under this scenario, this work presents a novel semantic knowledge base, as well as their reconstruction methodology, to process, structure and represent the discussed biomedical and health-related knowledge described in the gluten-related literature. In addition, the developed platform includes different external information such as (*i*) references to state-of-the-art biomedical databases, (*ii*) different bibliometric statistics and (*iii*) social media discussion to give a broad-range biomedical analysis of the gluten domain. It is expected that this new platform enhance the literature review productivity and assist scientists to (*i*) discover and analyse evidenced health-related interactions and patterns discussed in the literature, and (*ii*) identify the research opportunities or establish novel hypotheses in the same way that similar existing databases [31–34].

2. Background

In the last years, the synthesis and representation of the literature knowledge are showing the potential to systematically curate, organize, retrieve and interpret biomedical content in ways that are well-suited to human understanding. Besides, it helps to prevent public misinformation and disinformation and also assists the general public to interpret and analyse scientific knowledge. In this area, computed methods such as text mining, machine learning and graph-based representation

techniques becoming increasingly popular to assist in the processing and structuring of the contrasted and refuted evidence exposed in the bibliome.

In this sense, the work of Lamurias et al. [35] carried out a detailed review of how the development of different specialised methods and tools for the systematic processing and integration of large-scale scientific literature, biological databases and experimental data is a contemporary and well-recognised challenge of Bioinformatics. Their work highlights how computational methods could assist scientists to organise the highly available but deeply unstructured scientific knowledge as well as to carry out new literature-based hypotheses. In other work, Ammari et al. [36] also highlighted the relevance of biocurators and their syncretization and integration work to structure the available biomedical data into knowledge of practical use to save time and money in future research studies.

Consequently, many researchers are turning to the application of novel text mining and machine learning algorithms for the generation of structured information sources providing pipelines that support the automatic identification of relevant documents, biomedical concepts and biomedical relation interactions in a specific domain of the literature. For example, in order to structure the literature knowledge related to specific bacteria, Jorge et al. [37] and Pérez-Pérez et al. [38] apply a semi-curation pipeline that combines text mining methodologies, manual curation and graph analysis methods for the reconstruction of antimicrobial peptides-drugs combinations to create a comprehensive knowledge map and a novel database of potential anti-quorum sensing agents, concerning the bacterium *P. aeruginosa.* In terms of ontology-based text mining methods and knowledge enrichment to process and structure the bibliome, Hur et al. [39] proposed the application of ontology normalization capabilities and graph analysis methods to detect various gene interactions in the PubMed vaccine-related literature to extract scientific insights on E. coli vaccine research and development. In a similar line, Karaa et al. [40] propose the combination of natural language processing tools and ontology-based methods for the automatic extraction of gene-disease-food relations from MEDLINE with effective results. On the other hand, the work of Doğan et al. [41] evidenced how the application of a deep-learning-based prediction system and their combination with knowledge graph representations methodologies was an excellent proposal to infer, analyse, and established a novel database of biological pathways concerning genes, proteins and diseases using a large-scale biomedical data. In a similar way, Delmas et al. [42] proposed a web-based knowledge graph based on literature mining, to help to explore new hypotheses related to metabolomics pathways. Besides, they demonstrate how the integration of meta-information as the Medical Subject Headings (MeSH) descriptors associated with the articles could help to find relevant bibliometric statistics at a higher level. Related to the structuration and analysis of article *meta*-information to extract new knowledge, the works of Donthu et al. [43], Yuan et al. [44] and Guo et al. [45] highlight how the bibliometric analysis has gained huge popularity in recent years thanks to (*i*) their effectiveness to exploring and analysing large volumes of scientific data, (*ii*) their potential to discover emerging trends, collaboration patterns, and research constituents, and (*iii*) their potential to producing a high research impact. In this line, Yang et al. [46] explored the use of ontologies, such as Disease Ontology, to annotate and normalize disease mentions in the Arizona disease corpus obtaining a valuable performance achievement. Complementarily, Dandan Tao et al. [47] provided an overview of the data sources, computational methods, and applications of text data in the analysis of food science and nutrition. Through their review of different text mining techniques such as word-level analysis (e.g., frequency analysis), word association analysis (e.g., network analysis), and advanced techniques (e.g., text classification, text clustering, topic modelling, information retrieval or sentiment analysis), the authors discussed valuable insights on how different text data analysis methods can be used to help address critical issues to improve food production, food safety and human nutrition. In a complementary way, the work of Bakhtin et al. [48] evidenced the challenge posed by the overwhelming amount of available information, which surpasses the capacity of manual filtering or expert knowledge. To specifically address this issue, Bakhtin et al. proposed a text-mining methodology to analyze over 30 million documents related to science and technology in food production. Methods such as tokenization, lemmatization or term similarity analysis were successfully applied to extract and identify core science and technology fields and emerging topics in agriculture and food production.

In terms of combining social mining and literature knowledge to perform a biomedical large-scale analysis, the work of Jurca et al. [49] explored how the integration of text mining methods and graph analysis could generate valuable hypotheses about breast cancer biomarkers. Finally, this review by Edo-Osagie et al. [50] discussed how the combination of social media analysis with traditional research methods in a health context can present new practical and affordable solutions for implementing disease monitoring and surveillance systems in different health-related areas.

In this scenario, the present study presents a novel gluten knowledge database of potential biomedical and health-related interactions extracted from the literature using a semi-supervised curation workflow that combines (*i*) natural language processing (NLP) techniques, (*ii*) machine learning (ML) algorithms such as deep learning (DL) and random forest (RF), (*iii*) ontology-based normalization and integration techniques, (*iv*) named entity recognition (NER) methods, and (*v*) graph knowledge reconstruction techniques to process, classify, represent and analyse the unstructured knowledge contained in the literature. In addition, the developed platform includes different external knowledge such as links to state-of-the-art biomedical databases, bibliometrics statistics, and social media discussion to give a broad-range biomedical analysis of the gluten domain.

3. Materials and methods

This section resumes the different steps followed to curate, process, and structure the online GlutKNOIS (GLUTen KNOwledge InteractionS) database, as well as the metadata integration and biomedical concept normalisation methodologies applied to standardise and enrich the final knowledge base.

3.1. Knowledge base establishment

The NCBI (National Center for Biotechnology Information) Entrez Utilities Web services were used to access the PubMed library and retrieve the up-to-date gluten-related articles as well as their publication details, including the titles, abstracts, authors, funding agencies, affiliations, keywords and MESH terms; to be further processed [51]. The purpose was to establish a semi-automatic annotated corpus to identify and analyse the relevant biomedical topics and evidence of health-related interactions supported in the literature. Therefore, the final objective is the identification of domain mentions or categories such as anatomical terms (e.g., Pancreas), cell types (e.g., leukocytes), compounds (e.g., Glycan), variety of diets (e.g., keto), diseases (e.g., autism), food or food products (e.g., oats), genes (e.g., HLA-DQ2), organisms-viruses (e.g., Lactobacillus or COVID19), proteins (e.g., IgG) and symptoms (e.g., rash) for the reconstruction of a knowledge database of evidenced health-related interactions related to gluten. In this sense, the following relation categories were established: (*i*) related health issue, (*ii*) improve, (*iii*) aggravate, (*iv*) stimulation, (*v*) inhibition, (*vi*) activation, (*vii*) upregulation, (*viii*) increase symptoms, (*ix*) reduce symptoms, (*x*) weak relation, and (*xi*) no effect. Supplementary material 1 provides additional information about each annotated relation category established by an expert in the problem domain.

In addition, the NCBI Entrez Utilities Web services were also used to access the PubMed library and download a comprehensive dataset of 13,238 document abstracts related to the general query “gluten”,

including associated publication details. However, it should be noted that not all of the 13,238 documents were relevant to the domain of this work (i.e., gluten concerning diseases or changes in the human body), some of them were related to gluten concerning manufacturing processes or other related aspects. Consequently, a total of 5814 documents, ranked first by their PubMed relevance, were manually annotated by a domain expert with the goal of identifying studies that contain relevant health-related knowledge (i.e., articles that support meaningful biomedical interactions) and exclude from this sub-sample documents that, in basis on the criteria of the domain expert, did not provide relevant health-related information. Finally, due to the time-consuming nature of the manual reviewing, curating and structuring of the remaining dataset, the lasting 7424 "gluten" related articles were automatically processed by applying the expert-curated knowledge through the use of different machine learning and text mining techniques explained in the following sections.

3.2. Knowledge retrieval and normalization

In order to enrich the final knowledge reconstructed database and improve the user experience, different external and third-party information sources were integrated to present in relation to the obtained knowledge. In this way, the vocabulary and the *meta*-information related to the following selected ontologies were integrated to be able to link to other state-of-the-art related databases: the protein catalogue of Uniprot [52], Chemical Entities of Biological Interest (ChEBI) lexicon [53], Disease Ontology [54], FoodOn ontology [55], Symptom (SYMP) ontology [56], KEGG [57], PharmGKB [58], Medical Subject Headings (MeSH) [59], DrugBank lexicon [60], Foundational Model of Anatomy (FMA) ontology [61] and the National Cancer Institute Thesaurus (NCIt) [62]. Besides, the integration of these different ontologies and lexicons enabled the recognition of relevant domain terms, as well as the normalization of concepts with similar meanings.

In this way, the standardisation of the concepts to (*i*) enhance the named entity recognition steps and (*ii*) unificate the vocabulary of the final knowledge base, was carried out through three main pillars: (*i*) the normalization capabilities of the selected state-of-the-art NER taggers (presented in Section 3.3), (*ii*) the ontology-based normalization and reasoning capabilities like inferring related semantical terms using their common identifier or the curated synonyms list, and (*iii*) the application of lemmatization techniques to recognise concepts with the same lexical root.

This minimizes the final complexity of the knowledge base reconstruction (i.e., reduced the nomenclature diversity) maximizing the cohesion and standardization of the final vocabulary and establishing a lexicon of more than 500,000 term entries to support the entity recognition task.

As a result, synonyms, lexemes and variances of the same concept were identified and normalized reducing the semantic noise (e.g., diabetes type one, diabetes type I, type I diabetes, DBT 1, Diabetes type I, Diabetes mellitus I or Diabetes mellitus type one; were normalized to diabetes type one).

3.3. Document annotation and literature classification

To establish the first knowledge database related to gluten it was required to process the whole related literature and identify those documents that were relevant to the problem domain (i.e., focused on the study of proteins, compounds, and foods that produces health or metabolic changes) and filter not relevant documents (discussing manufacturing processes and the elasticity of gluten in food). However, carrying out this annotation and classification process only with manual methods can be a very tedious and time-consuming task, even more considering that the final objective of the present work was the processing and annotation of the up-to-date gluten-related literature published in PubMed.

Consequently, to save manual triage efforts, a semi-automatic annotation workflow and a machine learning document classification technique were established based on the previous works of the authors in the problem domain [63]. Therefore, a semi-automatic document curation task supported by different automatic domain recogniser methods was established to assist the curator and reduce the annotation effort and improve the final annotation quality. For this process, the next six state-of-the-art NER taggers were used to assist in the annotation task and save efforts in the semi-automatic annotation workflow: TMCHEM [64] to identify chemical, drug brand and trade names; LINNAEUS [65] to annotate species; DNORM [66] to recognize disease names; ABNER [67] to annotate genes and proteins; and OSCAR4 [68] to recognize chemical names, reaction names, enzymes, chemical prefixes, and adjectives. Besides, to complement the annotation process and annotate the domain categories of diets, foods, cell types, anatomical terms, and symptoms an in-house ontology-based NER was established. The developed NER entailed a dictionary lookup as well as a pattern and rule-based lookup to perform an inverted recognition strategy in which sentence words were used as patterns to be matched against the gluten-related lexicon established in the previous section.

In contrast to the previous work established by the authors, the present version of the semi-automatic annotation workflow has been improved with abbreviation resolution and hypernym normalization for enhanced entity recognition.

On the other hand, in terms of the document classification model and document triage assistance, the previous study established by the authors proposed a methodology to create an inferred model based on different state-of-the-art NER taggers and boosted by different domain ontologies to classify articles in a specific domain [63]. In this regard, based on the performance results of the different evaluated models, for this work, a random forest (RF) model was trained to filter the relevant documents in the problem domain and assist in the manual document annotation.

RF is an "ensemble learning" technique based on the aggregation of many decision trees to reduce the variance of applying a single decision tree. The idea is to train different trees against different samples to perform an average of the predictions of each decision tree. In this sense, each tree might have a high variance concerning a particular set of training data, but overall, the entire forest will have a lower variance but not at the cost of increasing the bias. The most common and relevant parameters for RF are the number of trees, the criterion on which features will be selected for splitting, the maximal depth of trees, the voting strategy and pruning strategies. In this regard, to find the best performance a "Hyperparameter Optimization" or "Hyperparameter Tuning" strategy was carried out to find the best model performance.

Therefore, after the application of the trained and optimised model, 6,684 documents were labelled as irrelevant whilst 6,554 were classified as relevant combining the manual dataset with the machine learning inference techniques.

3.4. Document relation extraction

In order to support the relation extraction of the gluten-related literature and identify the relation categories defined by the domain experts, different machine learning models were established following the previously proposed relation extraction methodology and the semantic vector space presented by the authors [69]. The presented approach incorporates a novel vector space that combines (*i*) high-level lexical and syntactic inference features as Wordnets and Health-related domain ontologies, (*ii*) unsupervised semantic resources as word embeddings, (*iii*) semantical and syntactic sentence knowledge, (*iv*) abbreviation resolution support, (*v*) several state-of-the-art Named-entity recognition methods, and, finally, (*vi*) different feature construction and optimization approaches; to support a relation extraction model be able to infer evidenced health-related interactions.

Therefore, for each established relation category, it was selected the

model with the greatest performance be able to provide a confidence metric. This decision was driven to provide confidence selection capabilities on the developed database and allow potential users to filter and visualise the automatically extracted health-related interactions that they feel confident in. In line with the mentioned work, three different machine learning algorithms obtained the best performance compared to the other alternatives. Gradient boosted trees (GBT), Fast large Margin (FLM) and Deep Learning (DNN), being the DNN model the most predominant selection for all established relation categories.

3.5. Knowledge reconstruction and integration

The reconstruction of knowledge graphs of large volumes of information provided by the processing of texts has proven to be a very powerful tool to obtain new knowledge and discover non-trivial domain patterns. In this regard, a knowledge graph database was reconstructed with the biological interactions identified in the literature to visualise and analyse all curated evidence. Besides, different layers of knowledge as the *meta*-information of the articles and references to external sources of information were integrated to enrich the final experience.

Therefore, the following equations define formally the rationale behind reconstructing the gluten-related literature in a knowledge graph.

A domain, D, is represented by a set of representative articles (Eq. (1)):

$$D = \{A_0, A_i, \cdots A_{N-1}\} \tag{1}$$

where A_i is the i-th article in corpus D and N represents the total number of articles in the problem domain D.

In the same line, an article, A, is represented by a set of meaningful concepts (Eq. (2)):

$$\mathrm{A} = \{c_0, c_i, \cdots c_{J-1}\} \tag{2}$$

where c_i is the i-th concept associated with document A and J represents the total number of identified concepts in the document.

A concept, c, for the studied domain, is represented by a set of terms (n-grams) of similar meaning (Eq. (3)):

$$c = \{t_0, t_j, \cdots t_{J'-1}\} \tag{3}$$

where t_i is the i-th term associated with concept c, and J' is the total number of terms associated with concept $c \in D$.

In terms of the knowledge graph, a vertex v_i is described by the set of the articles A that mentioned the concept c, and can be represented as a vector (Eq. (4)):

$$v_i = < A_{0c_i}, A_{ic_i} \cdots A_{N-1c_i} > \tag{4}$$

$A_{\mathrm{i-th}c_i}$ is 1 if the concept c_i is mentioned in the article A_i and 0 otherwise and N represents the total number of articles in the problem domain D. Therefore, the ontology-based inference approaches applied enabled the simplification and normalisation of the final graph enclosing in the same vertex different terms with the same meaning (e.g., Synonyms or acronyms such as IBD, inflammatory bowel disease or disease of bowel inflamed are identified as a unique vertex).

Finally, an edge or a relation e_{c_i,c_j} between two concepts, c_i and c_j, exist only if there is a semantic biomedical association or a meaningful relationship between them in the evaluated literature.

Therefore, in terms of knowledge representation: (i) a graph vertex denotes a unique normalised concept identified (i.e., unique annotated terms normalised using the different ontologies) whereas a graph edge denotes the existence of at least one experimental evidence identified in the literature; (*ii*) the vertex category is dependent on the biomedical annotated category (i.e., disease) whereas the vertex size is dependent on its degree (i.e., the number of interactions found in the literature concerning it); and (*iii*) the edge width indicates the number of documents that support the interaction, whereas edge category represents the category of identified evidence (e.g., aggravate).

Another valuable part of the reconstructed knowledge base was the bibliometric *meta*-information related to the processed articles. Knowledge such as the related funding agencies, funding countries, authors, author affiliations, MeSH terms, article keywords, and related references were also integrated to evaluate their relevance in connection with a specific vertex, annotated category or relation category.

In this sense, an article A could be represented from another point of view by the related metadata associated with the record (Eq (5)):

$$A = \{\mathrm{M}_0, \mathrm{M}_i, \cdots \mathrm{M}_{\mathrm{K}-1}\} \tag{5}$$

where M_i is the i-th related metadata of a specific nature (e.g., an author or funding agencies) and K represents the total number of metadata registers of a specific type. Accordingly, the knowledge graph could be reformulated to show evidence about any kind of related metadata, so v_i could be redefined to (Eq (6)):

$$v_i = \ll M_{0c_i} M_{ic_i} \cdots M_{T-1c_i} > \cdots < M_{0c_i} M_{ic_i} \cdots M_{T-1c_i} \gg \tag{6}$$

where $M_{\mathrm{i-th}c_i}$ is 1 if exist a concept c_i related to the metadata (i.e., is mentioned in the article A and $\mathrm{M}_i \in A$) and 0 otherwise; and T represents the total number of the metadata in the problem domain D.

In this way, the developed knowledge database integrates this meta-information, as well as the yet-mentioned third-party databases to enrich the user experience and provide practical statistics such as, for example, the most cited author in relation to a specific disease, the most related funding agency to a domain category as diets or genes, or the different articles concerning identified evidence in the literature. In summary, Fig. 1 represents the knowledge graph reconstruction and the concept normalization methodology.

The proposed reconstruction methodology enabled a holistic, multi-layered, and statistical analysis to acquire new knowledge, look into different levels of detail, and extract different knowledge subgraphs and perspective views. In this regard, different graph knowledge and statistical metrics were calculated on-demand and presented in the developed platform to assist to analyse the reconstructed knowledge. State-of-the-art graph metrics such as degree distribution, cluster coefficient, and betweenness centrality or closeness centrality were some of the standard graph metrics implemented to evaluate the queried sub-graphs in real-time [70].

On the other hand, the relation of how the different annotated categories were mentioned, without an implicit biomedical relation, can also provide a valuable information layer on how the knowledge was discussed in the literature. For example, recognising if diets were more associated with symptoms or diseases, or if genes were more related to compounds than proteins could provide domain lexical details. In consequence, a coefficient of association [71] between the identified annotations was implemented to be displayed in the platform, and is defined as (Eq (7)):

$$\phi_{c_i c_j} = \frac{D_{c_i \cap c_j} D_{c_{i'} \cap c_{j'}} - D_{c_i \cap c_{j'}} D_{c_{i'} \cap c_j}}{\sqrt{D_{c_i} D_{c_{i'}} D_{c_j} D_{c_{j'}}}} \tag{7}$$

where D_{c_i} represents the number of documents containing the annotated category c_i, D_{c_j} represents the number of documents containing the annotated category c_j,$D_{c_{i'}}$ stands for the number of documents not containing the category c_i, $D_{c_i \cap c_j}$ indicates the number of documents containing both categories c_i and c_j, $D_{c_{i'} \cap c_{j'}}$ indicates the number of documents not containing both categories c_i and c_j, and $D_{c_i \cap c_{j'}}$ represents the number of documents containing the category c_i but not category c_j. In this context, the ϕ coefficient ranges between -1 to $+1$ representing the extent to which articles tend to discuss one category but not the other, none of the categories or both semantic categories together.

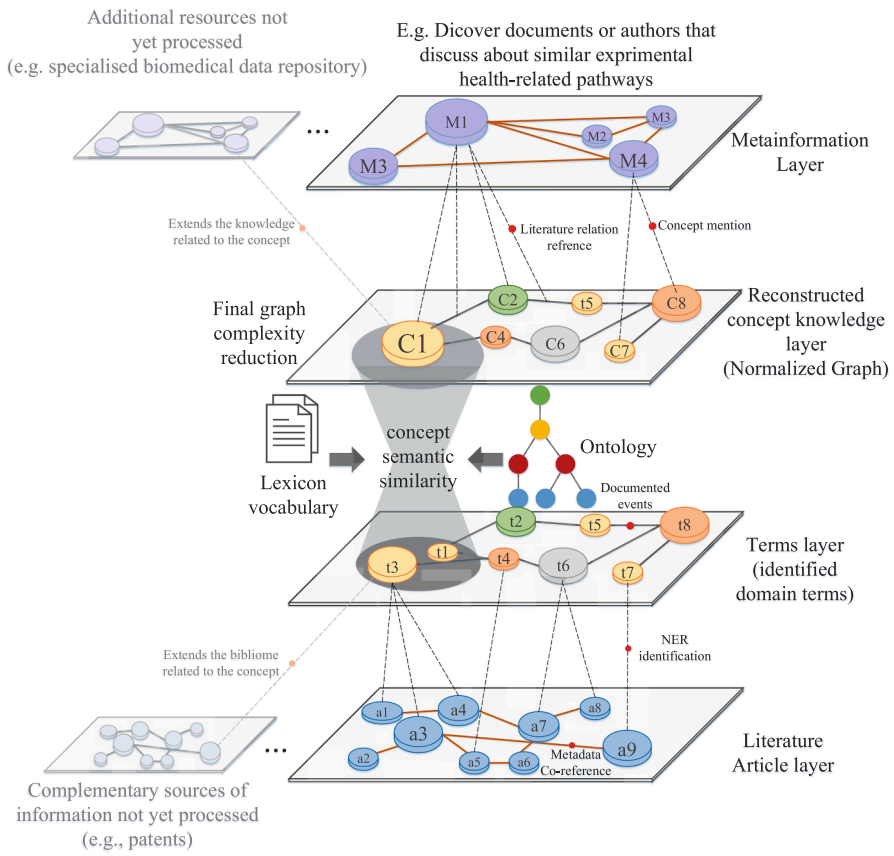

Fig. 1. Knowledge graph reconstruction methodology. Extracted literature knowledge enabled a holistic, multi-layered information level. Besides graph reconstruction and ontology-based methods allows the mathematical description of the graph at different levels and the normalization of the domain concepts reducing the final graph complexity and in consequence the final database complexity.

3.6. Knowledge visualisation

To make it easier for anyone to browse, visualise and analyse the reconstructed semantic knowledge, this work presents an online public database that represents the curated and integrated information in an interactive way (https://sing-group.org/glutknois/).

The proposed website was developed in the backend using the Spring boot MVC framework [72], the MySQL database [73] and the Jgraph library to allow Dijkstra searches and discover interesting paths between different biomedical annotated entities [74,75]. On the other hand, the most important web-based visualization technologies used were Jquery [76], Bootstrap [77], Angular.js [78], Amcharts [79], Brat document representation library [80] and the Cytoscape Web plugin [81]to make the website interactive.

In this sense, the developed platform presents (*i*) an interactive holistic, multi-layered graph representation and knowledge browsing perspective, (*ii*) on-demand state-of-the-art graph statistical analysis, (*ii*) an association coefficient of the annotated terms, (*iii*) the most referenced articles of the presented graph, (*iv*) the interactive visualisation of the annotated documents, and (*v*) different on-demand bibliometrics statistics concerning to the visualised reconstructed knowledge (e.g., most common fundings or affiliations). Fig. 2 depicts the main visualisation possibilities of the platform to represent the processed knowledge.

3.7. Social discussion enrichment

Although GlutKNOIS is primarily designed for researchers seeking a simple and visual way to look up possible evidence of associations between the different biomedical entities mentioned in gluten literature, the platform also includes a social media discussion related to the topic to obtain a more comprehensive comparison of both domains.

In this way, it is possible to revise at the same time how society behaves around the searched topics in the problem domain. This allows potential users to gain an overall perception and be informed about how people in general is dealing with specific biomedical topics related to gluten and how the general public perceives the latest experimental studies. Therefore, social networks have become an important health information resource for scientists and individuals to stay up and discuss research and scientific trends [82–84], as well as to discover different disinformation or misinformation areas [85].

In this sense, the presented platform tries to work as a knowledge hub, assisting as a starting point for reviewing both literary and social data and serving as a bridge between the two domains in order to assist potential users to connect and synthesize information from both sources. Consequently, the developed knowledge base can be a valuable reference tool for detecting discrepancies, misinformation and gaps between the literary and social domains and, in this way, aid in the development of more effective social media awareness-raising campaigns and, ultimately contribute to the promotion of healthier lifestyles.

In this sense, Fig. 3 shows some of the different social media statistics and retrieved methods implemented in the developed platform. Valuable information such as (*i*) the related tweets with the queried topic, (*ii*) the most shared hyperlinks (*iii*) the most relevant hashtags and (*iv*) the most relevant users in the community were included to be analysed in connection with the literature search and evaluate the social and bibliome information as a whole. For example, if the term “autism” were requested, a set of tweets and social statistics related to autism in the gluten-free community could be recovered in association with the articles that also contain this term.

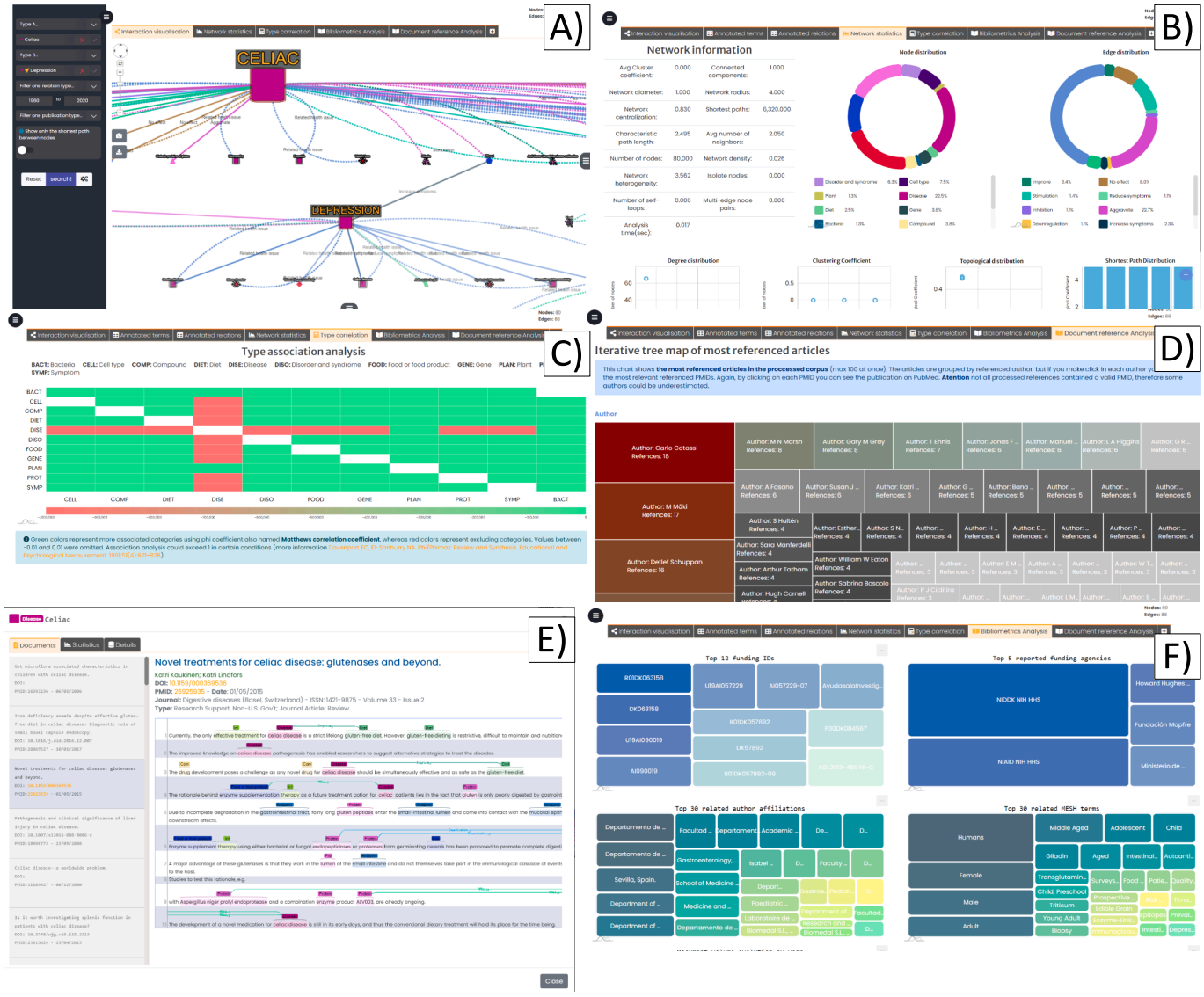

Fig. 2. (A) Knowledge graph visualization. (B) Visualization of the graph statistics. (C) Detailed information about the category correlation obtained in the analysed articles. (D) Bibliometric information about the most referenced articles in relation to the query. (E) example of annotated one document visualization. (F) Bibliometric analysis of the most related funding ids, agencies authors and MeSH terms in relation to the query.

4. Results and discussion

4.1. Database & performance statistics

As a result of applying the manual curation and the automatic text mining algorithms to the gluten literature, a semantic knowledge database of different; (*i*) classified documents, (*ii*) annotated terms and (*iii*) extracted relevant biomedical interactions; was generated. Accordingly, Table 1 resumes the statistics of the final database to clarify the amount of knowledge extracted by manual revision and automatically inferred.

In this way, the reconstructed knowledge database is the product of the processing of 5814 manually annotated documents and 7424 fully processed documents. Nevertheless, only the relevant labelled articles in the problem domain (i.e., health-related documents) were taken into account for the database reconstruction. Therefore, 6684 documents were labelled as irrelevant whilst 6 554 were classified as relevant using manual and machine learning inference techniques.

In terms of automatic literature processing, Table 2 introduces several state-of-the-art metrics showing the performance achieved after the application of the processing techniques comprising the proposed methodology.

It is important to emphasize that the accuracy of the entity recognition process is directly related to the most recent annotation round (i.e., 500 random documents used in the last step of the semi-automatic annotation) [63]. At this point, domain experts have already refined the annotation process, resulting in a more precise and targeted approach to entity recognition. Furthermore, the presented performance is highly dependent on the state-of-the-art NER taggers applied in Section 3.3.

4.2. Reconstructed biomedical knowledge graph analysis

This section evaluates the potential of the developed platform by analysing the final reconstructed graph and performing different analyses of the inferred knowledge.

To gain an in-depth understanding of the identified interactions (i.e., manual and inferred) in the gluten-related literature, Fig. 4 represents the final reconstructed graph and Fig. 5 depicts the most identified health-related interactions between each domain annotated category and the meaning of each graph colour present in the figures. The

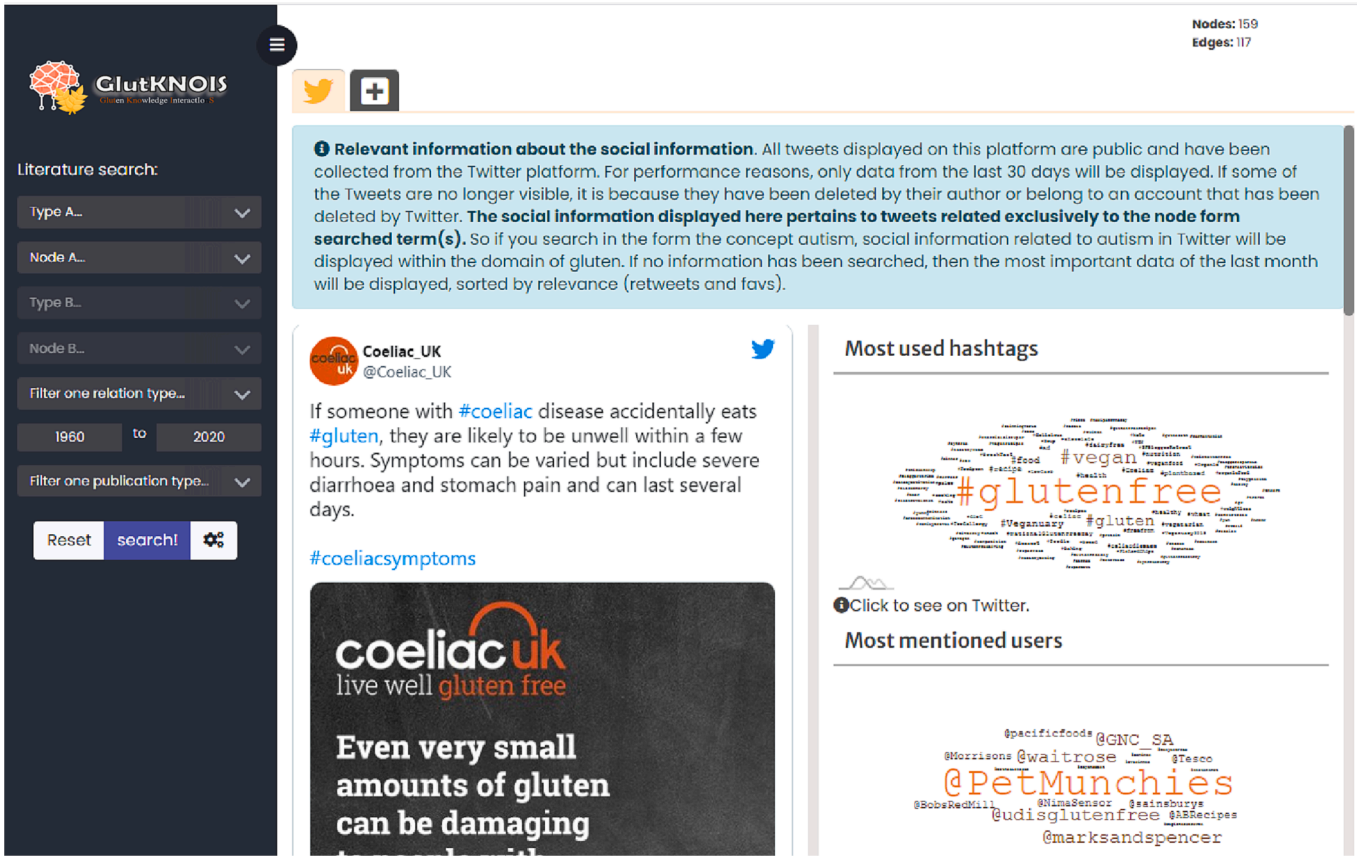

Fig. 3. Snapshot of the social media web analysis interface available in GlutKNOIS showing social-related information.

Table 1
Manual revision and automatically inferred knowledge processed.

Resource	#Manually revision	#Automatic inferred
Classified documents	5814	7424
Annotations	121,000	263,000
Interactions	9000	42,000*

* It includes interactions with a wide range of confidence values (≥0.3), to allow potential users to choose biomedical interactions with which confidence they feel most comfortable.

Table 2
F.score, precision and recall belonging to the different stages executed for the reconstruction of the GlutKNOIS database.

Processing stage	F.score	Precision	Recall
Named entity recognition	0.932	0.933	0.931
Document classification	0.846	0.798	0.901
Relation extraction	0.739*	0.757*	0.728*

* Mean ($\overline{x}$) obtained evaluating all models trained for each relation category.

complete knowledge graph (Fig. 4) contains 10,000 vertexes (i.e., unique annotated concepts) and 51,000 edges (i.e., health interactions identified). Edges were weighted based on the number of documents in the literature that support the experimental interaction, whereas the vertex size is based on their degree (i.e., the number of interactions related to the concept). The colour of the vertexes represents the annotated category whereas the edge colour represents the relation category.

In terms of graph description, the obtained knowledge graph (Fig. 4) had a diameter of 10, a radius of 5 and an average number of neighbours of 6. That means that each identified concept has a minimum path to almost 5 other concepts (i.e., if you peak a pair of vertexes that were spatially furthest from each other, it would take usually 5 hops or documents to reach the other) and a maximum of 10, and each concept usually has 6 directly connected concepts or existent interactions.

An important measure for the characterization of any complex graph is the heterogeneity, measured in terms of the diversity of connection reflected through its node degree. Heterogeneity reflects the tendency of a graph to contain hub nodes. In general, a hub is a vertex highly connected with many other nodes. Usually, biological graphs as the current graph tend to be very heterogeneous. While some "hub" nodes are highly connected, the majority of nodes tend to have very few connections. In light of this, the heterogeneity of the graph was 8.2 which means that there were a significant number of hubs [86].

In terms of graph cohesiveness and organization, the clustering coefficient is a density measure of local connections, or "cliquishness". Generally, highly organized graphs manifest higher clustering coefficient values ($0 \leq$ clustering coefficient ≤ 1), whereas random graphs manifest values near zero. In this regard, the clustering coefficient of the present graph was 0.4, which is a reasonably good value.

Regarding, the frequency of the identified mentions presented in Fig. 4 and Fig. 5, diseases (41%) and disorders (6%), proteins (18%) and symptoms (18%) were the categories with the biggest occurrence. On the other hand, regarding the frequency of the identified relation, related health issue (70%), stimulation (14%), aggravate (6%) and inhibition (2.7%) were the categories with the biggest occurrence in the literature. These results were consistent with the final research object of this paper (i.e., discover papers that discuss proteins, compounds, and foods that produce body changes in terms of disease, disorders, symptoms or specific organism reactions) and validate the satisfactory work of the document classifier semi-supervised workflow used. In conclusion, taking into account the volume and the diameter of the

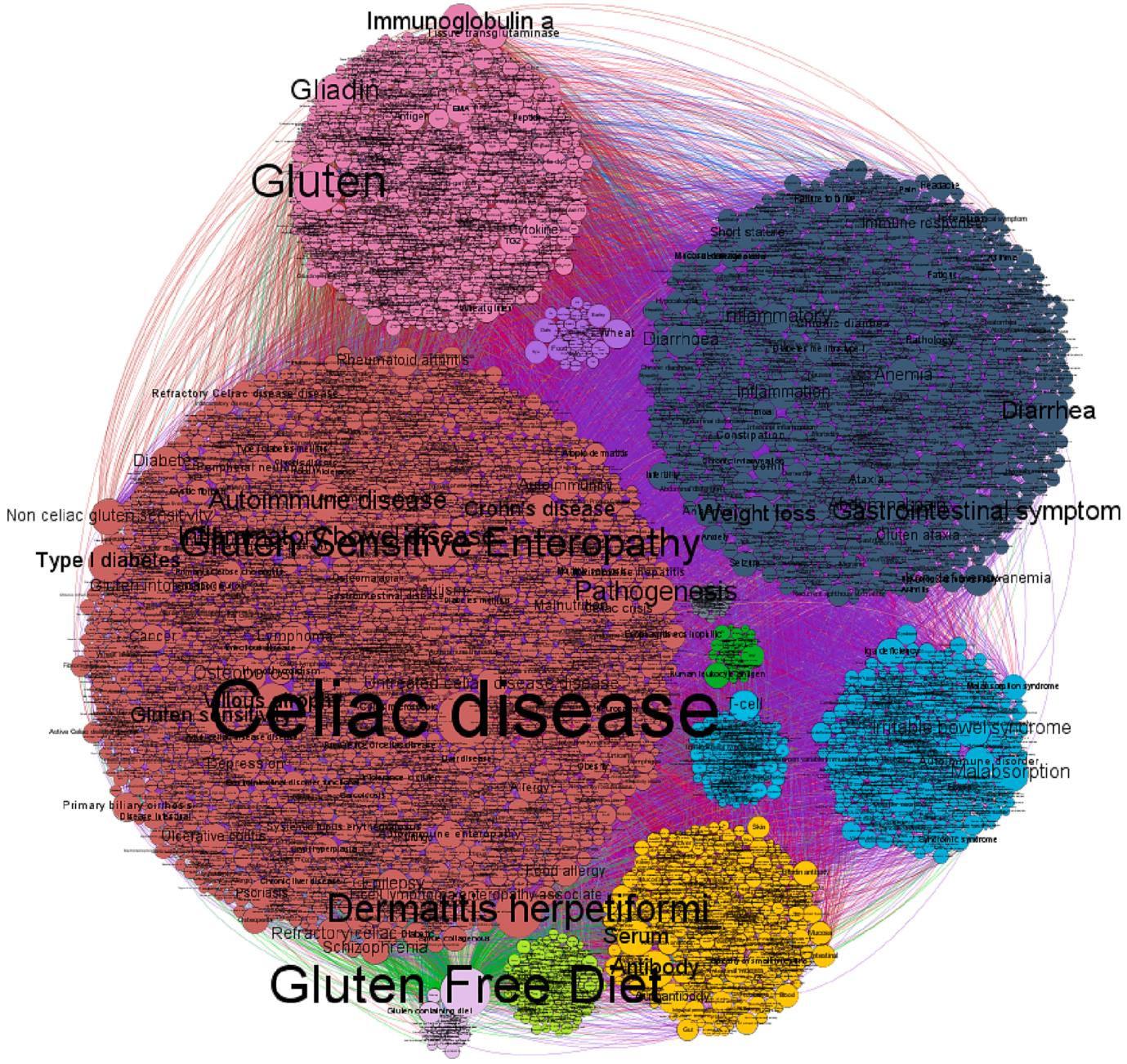

Fig. 4. Interaction graph reconstruction of the extracted interactions found in the gluten-related literature. The vertex colour denotes the biomedical domain category annotated (e.g., disease) whereas the vertex size is dependent on its degree (i.e., the number of annotated interactions). The edge colour represents the relation category annotated (e.g., downregulation) whereas the width of the edge indicates the number of documents that evidenced the interaction.

reconstructed knowledge graph, it is evidenced the suitability of the developed database to assist in the browsing and analysis of the identified health-related interactions.

4.2.1. Gluten health-issues experimental analysis

In order to examine the most often mentioned health-related interactions (i.e., disease, symptoms and disorders) concerning gluten protein and the GFD, Fig. 6 summarized the most identified patterns found in the literature. Ignoring the inherent interactions with bowel-related conditions, some different mental health-related conditions and symptoms were discussed concerning the gluten protein (e.g., Autism, Depression, Schizophrenia, Anxiety), different chronic conditions (e.g., Epilepsy, Diabetes, Autoimmune Thyroid) and other relevant health issues affecting to the human health (e.g., short stature, Osteoporosis, Vitiligo or Alopecia). Therefore, taking into account the large number of medical-related terms mentioned in the present subgraph, it is highlighted the relevance of the developed platform. The GlutKNOIS platform makes it easier to determine which health-related conditions are being discussed more and read the participant literature articles. Assisting in this way in the possible identification of misinformation or disinformation by the representation of well-researched associations between different health issues.

For example, by revising the reconstructed knowledge graph (Fig. 6), it is possible to see that there is an established debate in the literature on the association between gluten, diabetes and thyroid conditions. In this line and in a similar way to other work, the reconstructed knowledge base could assist in the revision of the literature to prevent and early diagnose comorbidities in the base of the scientific literature [87–91].

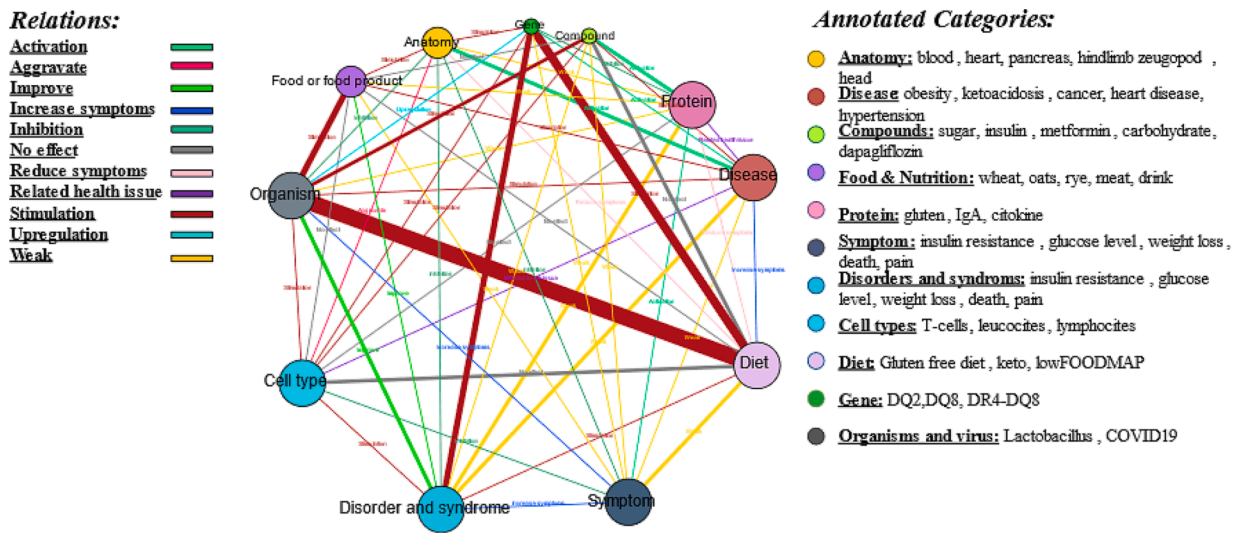

Fig. 5. Relation graph showing the top relation categories between each annotated category. The vertex colour denotes the biomedical domain category annotated (e.g., disease) whereas the vertex size is dependent on its degree (i.e., the number of annotated interactions), the edge colour represents the relation category annotated (e.g., aggravate) whereas the width of the edge indicates the number of documents that describe the interaction.

4.2.2. *Nutritional facts evidenced in the literature*

To illustrate how the reconstructed knowledge graph can simplify the review and analysis of the literature, Fig. 7 represents the main interactions between the most frequently mentioned diets, foods, compounds, cell types organisms, and viruses in the gluten literature.

In this sense, examining the present graph it is possible to easily visualise an extended literature discussion between a gluten-free diet with calcium and Vitamin D. This is logically associated with some of the health conditions mentioned in the previous section such as short stature and osteoporosis. On the other hand, in a similar way that other literature reviews expose, the reconstructed knowledge graph enables the opportunity to easily determine which micronutrient deficiencies or imbalances need to be taken into account in the patient's clinical and biomedical proves to prevent disease complications and risk factors based on the literature [92].

4.3. *Annotated biomedical concept analysis*

To gain an in-depth understanding of the scientific background and the database annotated knowledge, Fig. 8 represents the most relevant annotated concepts comparing the last 20 years of scientific advances to the previous years (i.e., before the 2000 s and after). Domain stop-words such as gluten, celiac disease or GFD were removed to help to discover more significant differences.

Concerning the different annotated categories (Fig. 8A), anatomical concepts and food-related concepts were the identified categories with the biggest differences in the different time windows analysed (i.e., taking into account the total number of documents published in each time window and the number of mentions). By comparing the differences between the two-time windows (Fig. 8B and C), it was possible to detect which research concepts were having a higher level of attention in the last years and which ones are being put on the back burner.

In this regard, Wheat, Flour and Starch were the discussed concepts with the biggest significant growth from the beginning of the millennium. Therefore, by examining their related interactions, disease, symptoms and anatomical parts were the annotated categories with the biggest association with them. This reveals an increased literature association of gluten-containing foods with various medical conditions. On the other hand, other concepts such as Gene, Genome or Immunoglobulins also had a significant association increment from the publication of the first draft of the human genome sequence in 2001 [93]. Considering the annotation differences in other specific concepts, Food, Ingredient, Weight loss, Supplements and different nutritional components (e.g., Fibre, Fat, Mineral, Antioxidant, Or Sugar) also had a rapidly alimentary adoption in the last years, and much more so with the widespread adoption of social networks from 2008 to 2010 [94]. These data reflect the importance and impact that social media networks have had in gluten research in recent years.

To conclude, the increasing mention of specific concepts such as Diabetes, Irritable bowel syndrome or Autoimmune disorder, as well as others not depicted such as Autism or Depression, and the increased mention of the general categories such as symptoms, disorders and diseases (Fig. 8A) reflect the growing number of articles associating and evaluating the effect of gluten or gluten-free diets and foods with different health conditions. Along the same line, the increased number of articles mentioning organisms (e.g., Lactobacillus, *Escherichia coli* and general microbes or bacteria) reflects the community research interest to find possible novel treatments taking into account the gut microflora.

4.4. *Analysis of social media discussion*

To illustrate the potential of the platform to contrast the social discussion against the literature, this section exemplifies how the combination of the bibliome with Twitter discussions could be a valuable tool for scientists and individuals to complement their knowledge. By leveraging the strengths of both sources, including the comprehensive and in-depth knowledge available in the literature, as well as the diverse and socially-near nature of Twitter, the present knowledge base could be a valuable starting point for gaining a more complete understanding of topics being discussed and thereby contributing to reveal discrepancies, gaps and/or misinformation around health-related topics and gluten.

For example, one of the most controversial topics related to gluten is the claim made by some individuals on social media that gluten is the primary cause of some neural-related disorders such as Alzheimer's disease and autism. They also suggest that the rising prevalence of these disorders in humans is directly connected with gluten. Some even claim that a gluten-free diet can cure autism (Fig. 9A). To provide a more comprehensive understanding of these facts, Fig. 9B presents a screenshot of the developed platform displaying tweet messages (if exist) related to the search query, along with the literature knowledge related to autism and Alzheimer diseases (Fig. 9C). In this way, the complementation of the knowledge from the literature with the social discussion offers a powerful perspective for evaluating gaps between the different health-related topics in discussion as well as, if the user

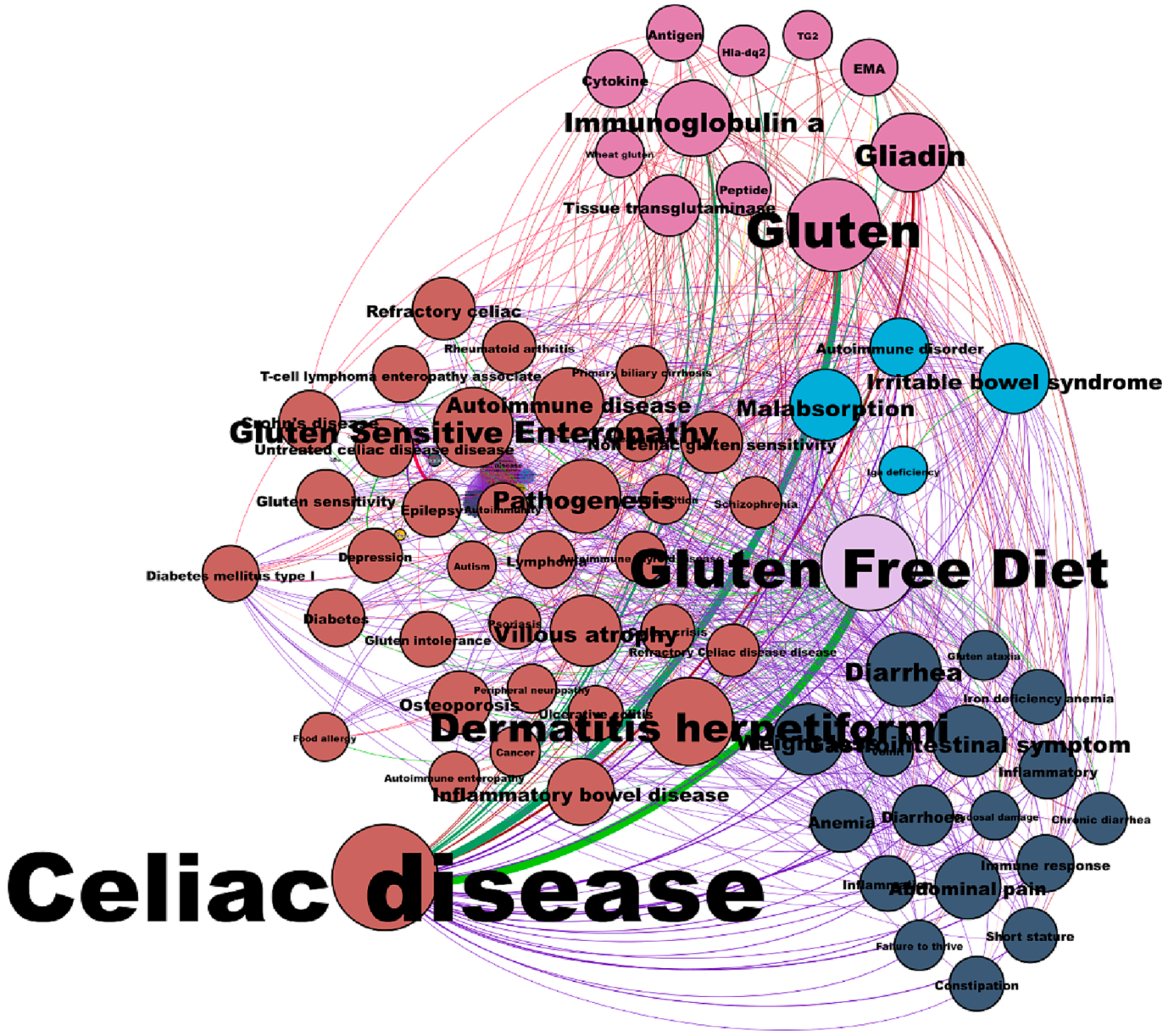

Fig. 6. Gluten-related health conditions associated with gluten and gluten-free diet. The vertex colour denotes the biomedical domain category annotated (e.g., disease) whereas the vertex size is dependent on its degree (i.e., the number of annotated interactions), the edge colour represents the relation category annotated (e.g., improve) whereas the width of the edge indicates the number of documents that describe the interaction.

considers it appropriate, rapidly debunk misinformation, carry out awareness decisions referencing available literature and, in the last instance, promote the experimental information.

5. Discussion

5.1. Limitations

Due to the use of the proposed semi-automatic curation workflow, not all experimental interactions exposed in the literature were recovered and reconstructed, in consequence, the lack of interactions between two different concepts does not imply that no exists proven evidence between them. The developed platform contains knowledge inferred through different manual and computational approaches, so it should be used as a supporting platform to be consulted and not as conclusive proof. On the other hand, due to the difficult integration of the different semantical sources (i.e., ontologies, lexicons and state-of-the-art NER taggers), it has not been possible to achieve a perfect standardisation of all concepts. In this way, there might be terms that belong to the same concept (e.g., anaemia and anaemia) or incorrect classified concepts that have not been manually revised because they did not appear in the manual document curation task.

In this line, the development of better and more robust domain ontologies by human experts enables better concept normalization and improved text mining methods and databases as present in this work. Therefore, the concept normalization was one of the most difficult challenges in this project due to the intrinsic difficulty of integrating all semantical resources. Taking into account the 500,000 term entries in the final lexicon, it is challenging to check all entries manually and identify possible errors and defects.

5.2. Conclusions

Understanding and structuring the empirically proven effects of

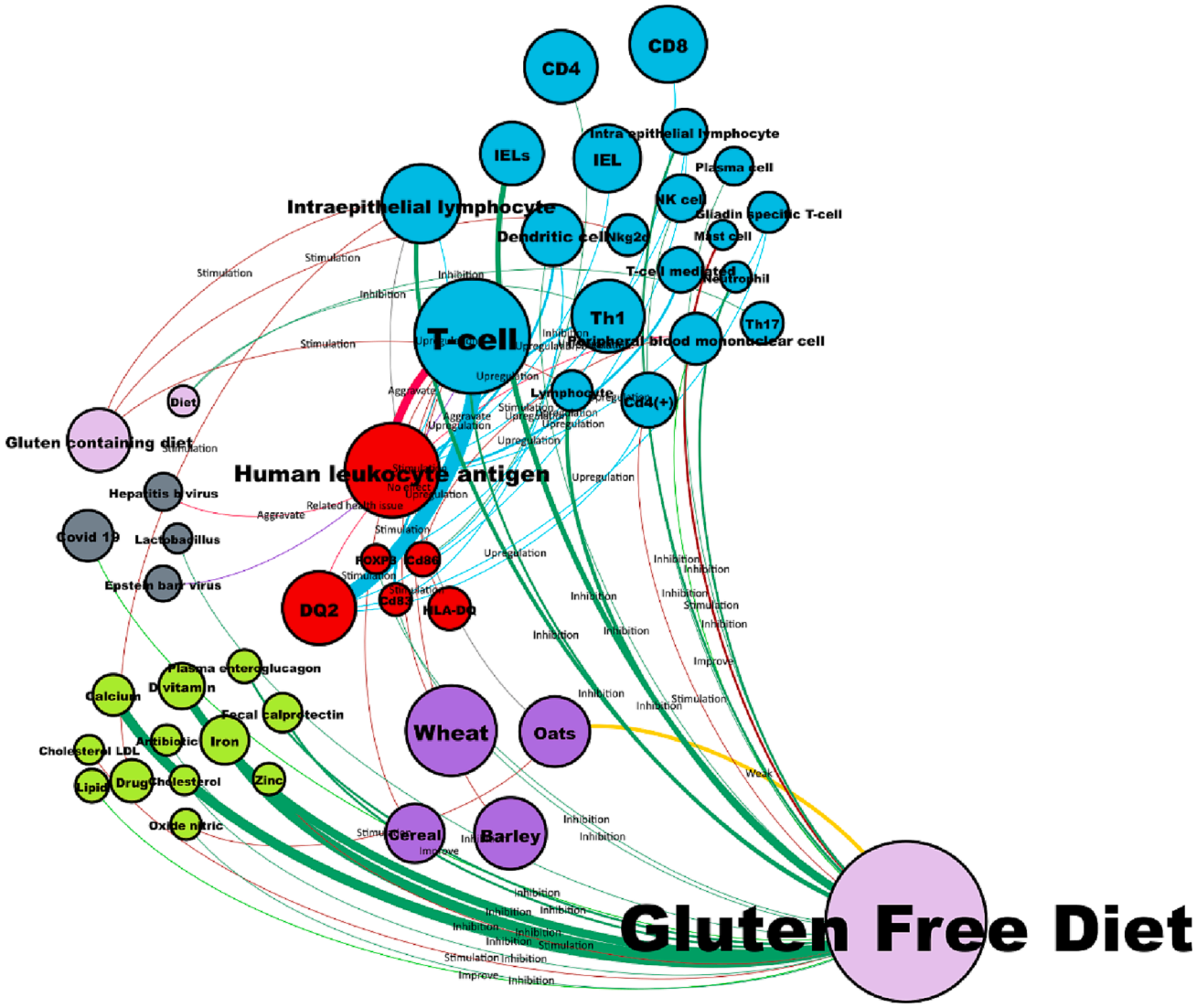

Fig. 7. Relation graph showing the most discovered interactions concerning diets, foods, compounds, cell types and organisms and viruses. The vertex colour denotes the biomedical domain category annotated (e.g., disease) whereas the vertex size is dependent on its degree (i.e., the number of annotated interactions), the edge colour represents the relation category annotated (e.g., inhibition) whereas the width of the edge indicates the number of documents that describe the interaction.

gluten and their intrinsic interactions with other autoimmune and chronic diseases is of utmost importance for researchers that intend to discover novel gluten-related therapies or prevent nutritional complications. Besides, in terms of alimentary disease and diets, gluten-related information is one of the dietary topics with the most volume of scientific publications, and with the most volume of misinformation and disinformation, producing a high economic and human cost.

In this line, this work presented a knowledge base reconstruction with a graph representation and integration methodology to visualise and analyse the up-to-date literature related to gluten protein and evaluate the evidenced experimental health interactions discussed in the literature. The practical relevance of the presented platform lies in the processing of 13,238 PubMed articles, using manual curation as well as fully text-mining workflows, for the reconstruction of the first gluten knowledge database based on the literature. For this purpose, a broad range of domain ontologies and state-of-the-art NER taggers were employed to identify and normalise domain concepts and semantical relations. Consequently, 9000 manual interactions were annotated and a total of 42,000 potential gluten-related interaction categories were automatically inferred using diverse machine-learning models.

In this way, this literature-based database structures and includes a broad range of practical knowledge related to gluten protein. The proposed graph representation technique, combined with the bibliometrics and social media integration, provides a novel enhanced way to search, visualise and analyse information on this topic. It is expected that the developed platform assists scientists to explore gluten-related evidence, discover patterns, explore the less researched scientific pathways or establish novel hypotheses in the same way that similar existing evidence-based databases in other domains allow. In addition, the processed knowledge database and the visualisation platform could help individuals and patients to easily and quickly access evidence-based knowledge and assist them to contrast and evaluate the disinformation and misinformation existing on social media.

Regarding the provided results, the reconstructed knowledge base has the potential to assist in the revision and analysis of years of gluten research, discovering relevant works and evidenced research interactions to support future studies and new therapeutic and preventive strategies. Finally, although automated document processing cannot completely replace human judgement, it can save substantial processing time and effort and enables the curation of a large number of articles on a specific topic in time. Besides the automatic processing of the literature combined with knowledge representation methodologies proposed

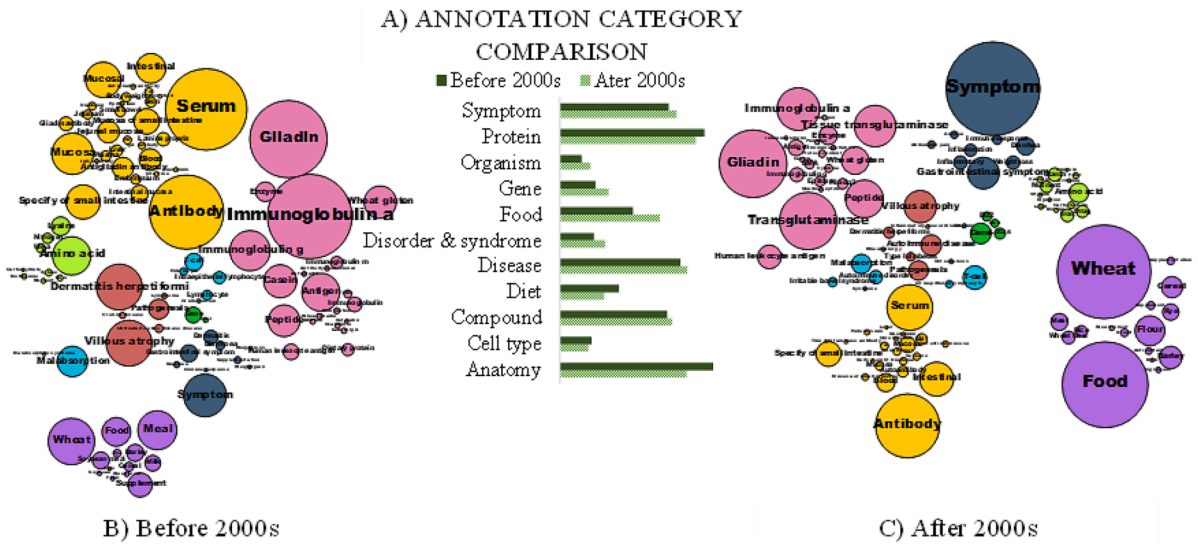

Fig. 8. Annotated biomedical concept comparison. Depicts the top annotated concepts by category comparing the years before 2000 and later. It was obtained from the reconstruction of the identified concepts found in the gluten-related literature. Rendering is based on the Circle Pack layout. The vertex colour denotes the semantic category whereas the vertex size represents the number of documents that contain the identified concept. (A) Top annotated concepts comparison per category. (B) Top annotated concepts before 2000. (C) Top annotated concepts after 2000.

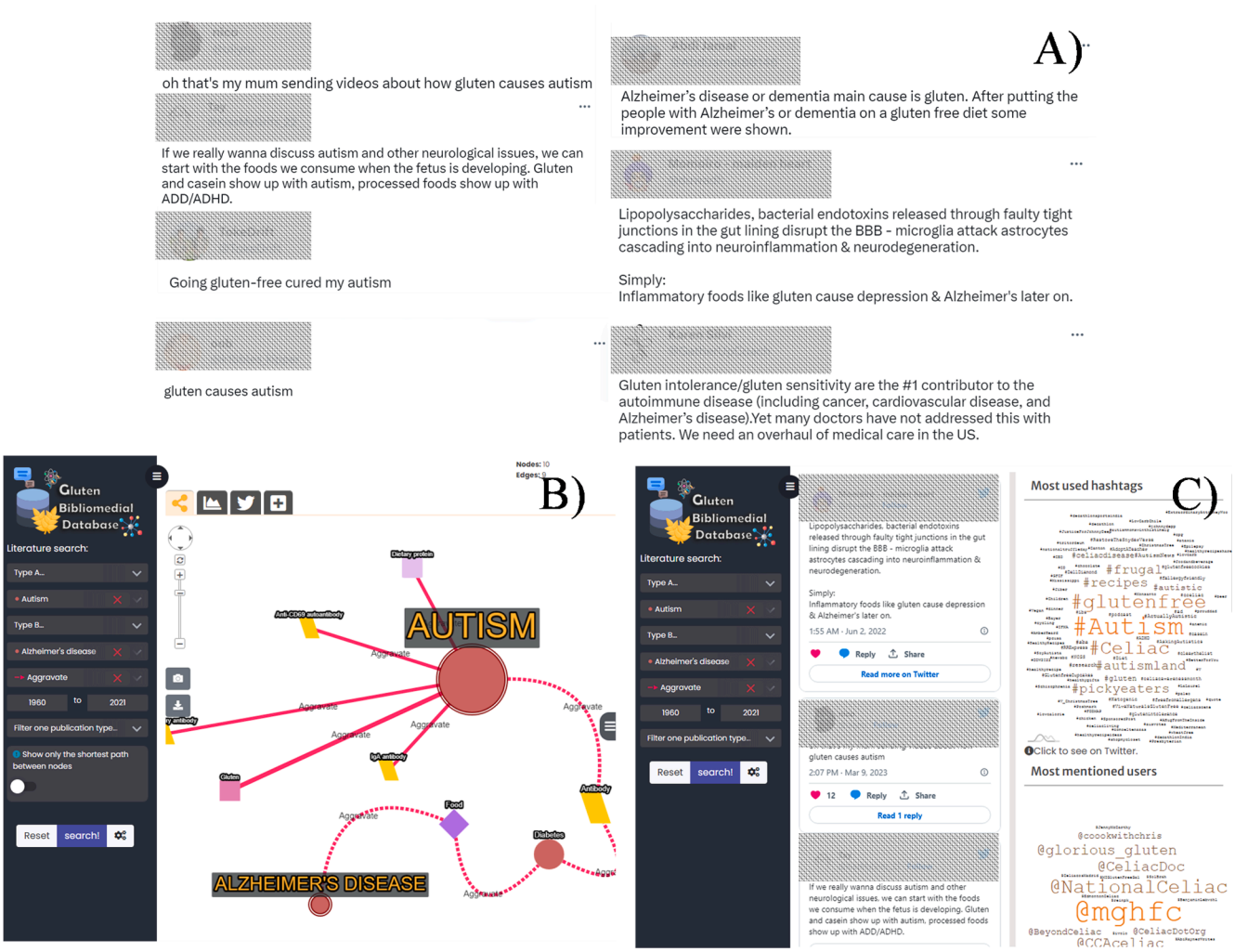

Fig. 9. Literature knowledge and social media discussion related to autism and Alzheimer diseases processed by GlutKNOIS. (A) Examples of messages that could include misinformation. (B) Knowledge network of biomedical relations found in the gluten literature related to autism and Alzheimer diseases. (C) Social media information related to the search query.

condensing the information, making it more manageable and visualisable to easily understand and infer new hypotheses.

To conclude, it is valuable to highlight that the literature-based knowledge of GlutKNOIS may extend beyond researchers to include the wider society. With its efficient visual data representation capabilities and user-friendly interface, the knowledge base enables agile exploration of extensive scientific research, empowering individuals with varying levels of scientific expertise to use it as a valuable reference point to discover experimental advancements related to their gluten-related field of interest. Therefore, the platform has the potential to facilitate greater engagement and participation in scientific advancements and promote a more informed society.

5.3. Future work

Future work will be centred on the integration of new metadata knowledge such as demographic information about the authors (e.g., infer their gender or their country), the improvement of the ontology normalisation, and the integration of journal metadata such as the impact factor or their quartile. In terms of knowledge representation, upcoming developments will be focused on representing the collaboration of the different agencies, authors or countries using also knowledge graphs. Taking into account the literature processed this database could be expanded with the integration of new literature knowledge related to other alimentary allergies or autoimmune affections that have been associated with gluten.

CRediT authorship contribution statement

Martín Pérez-Pérez: Investigation, Methodology, Writing – original draft. **Tânia Ferreira:** Investigation, Data curation, Visualization, Writing – original draft. **Gilberto Igrejas:** Conceptualization, Formal analysis, Resources, Writing – review & editing. **Florentino Fdez-Riverola:** Conceptualization, Project administration, Supervision, Writing – review & editing.

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.

Acknowledgements

SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from the University of Vigo for hosting its IT infrastructure. This work was supported by: the Associate Laboratory for Green Chemistry - LAQV financed by the Portuguese Foundation for Science and Technology (FCT/MCTES) [grant number UIDB/50006/2020]; the Conselleria de Cultura, Educación e Universidade (Xunta de Galicia) under the scope of the strategic funding of Competitive Reference Group [grant number ED431C 2022/03-GRC], the "Centro singular de investigación de Galicia" (accreditation 2019-2022) funded by the European Regional Development Fund (ERDF) [grant number ED431G2019/06]. The authors also acknowledge the postdoctoral fellowship of Martín Pérez-Pérez, funded by Xunta de Galicia [grant number ED481B-2019-032]. Funding for open access charge: Universidade de Vigo/CISUG.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104398.

References

[1] I. Aziz, F. Branchi, D.S. Sanders, The rise and fall of gluten!, in: Proc. Nutr. Soc., 2015, pp. 221–226. <https://doi.org/10.1017/S0029665115000038>.

[2] FAO, FAO Cereal Supply and Demand Brief | FAO | Food and Agriculture Organization of the United Nations, Fao, 2017.

[3] C.M. Pennesi, L.C. Klein, Effectiveness of the gluten-free, casein-free diet for children diagnosed with autism spectrum disorder: based on parental report, Nutr. Neurosci. 15 (2012) 85–91, https://doi.org/10.1179/1476830512Y.0000000003.

[4] A.E. Kalaydjian, W. Eaton, N. Cascella, A. Fasano, The gluten connection: the association between schizophrenia and celiac disease, Acta Psychiatr. Scand. 113 (2006) 82–90, https://doi.org/10.1111/j.1600-0447.2005.00687.x.

[5] G. Goodwin, Type 1 diabetes mellitus and celiac disease: distinct autoimmune disorders that share common pathogenic mechanisms, Horm. Res. Paediatr. 92 (5) (2020) 285–292. https://doi.org/10.1159/000503142.

[6] A. Testa, N. Imperatore, A. Rispo, M. Rea, R. Tortora, O.M. Nardone, L. Lucci, G. Accarino, N. Caporaso, F. Castiglione, Beyond irritable bowel syndrome: the efficacy of the low fodmap diet for improving symptoms in inflammatory bowel diseases and celiac disease, Dig. Dis. 36 (2018) 271–280, https://doi.org/10.1159/000489487.

[7] G. Caio, U. Volta, A. Sapone, D.A. Leffler, R. De Giorgio, C. Catassi, A. Fasano, Celiac disease: a comprehensive current review, BMC Med. 17 (2019) 1–20, https://doi.org/10.1186/s12916-019-1380-z.

[8] A. Rubio–Tapia, R.A. Kyle, E.L. Kaplan, D.R. Johnson, W. Page, F. Erdtmann, T. L. Brantner, W.R. Kim, T.K. Phelps, B.D. Lahr, A.R. Zinsmeister, L.J. Melton, J. A. Murray, Increased prevalence and mortality in undiagnosed celiac disease, Gastroenterology 137 (1) (2009) 88–93. https://doi.org/10.1053/J.GASTRO.2009.03.059.

[9] J.F. Ludvigsson, A. Rubio-Tapia, C.T. van Dyke, L.J. Melton, A.R. Zinsmeister, B. D. Lahr, J.A. Murray, J.A. Murray, Increasing incidence of celiac disease in a North American population, Am. J. Gastroenterol. 108 (2013) 818–824, https://doi.org/10.1038/ajg.2013.60.

[10] G.A. Gaesser, S.S. Angadi, Navigating the gluten-free boom, J. Am. Acad. Phys. Assist. 28 (8) (2015) 1–7. https://doi.org/10.1097/01.JAA.0000469434.67572.a4.

[11] C. Newberry, L. McKnight, M. Sarav, O. Pickett-Blakely, Going gluten free: the history and nutritional implications of today's most popular diet, Curr. Gastroenterol. Rep. 19 (2017) 1–8, https://doi.org/10.1007/s11894-017-0597-2.

[12] J. Masih, A. Sharma, R. Teodor, Study on consumer behaviour and economic advancements of gluten-free products, Niger. Orig. Res. Artic. Masih Sharma. AJEA (2016) 24737, https://doi.org/10.9734/AJEA/2016/24737.

[13] M. Pérez-Pérez, G. Igrejas, F. Fdez-Riverola, A. Lourenço, A framework to extract biomedical knowledge from gluten-related tweets: the case of dietary concerns in digital era, Artif. Intell. Med. 118 (2021), 102131, https://doi.org/10.1016/j.artmed.2021.102131.

[14] D.M. Lis, T. Stellingwerff, C.M. Shing, K.D.K. Ahuja, J.W. Fell, Exploring the popularity, experiences, and beliefs surrounding gluten-free diets in nonceliac athletes, Int. J. Sport Nutr. Exerc. Metab. 25 (2015) 37–45, https://doi.org/10.1123/ijsnem.2013-0247.

[15] M. Househ, E. Borycki, A. Kushniruk, Empowering patients through social media: the benefits and challenges, Health Inform. J. 20 (2014) 50–58, https://doi.org/10.1177/1460458213476969.

[16] J.A. Greene, N.K. Choudhry, E. Kilabuk, W.H. Shrank, Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook, J. Gen. Intern. Med. 26 (2011) 287–292, https://doi.org/10.1007/s11606-010-1526-3.

[17] K. Omurtag, P.T. Jimenez, V. Ratts, R. Odem, A.R. Cooper, The ART of social networking: how SART member clinics are connecting with patients online, Fertil Steril. 23 (2014) 88–94, https://doi.org/10.1016/j.fertnstert.2011.10.001.The.

[18] C.B. Thielst, Social media: ubiquitous community and patient engagement, Front. Health Serv Manage. 28 (2) (2011) 3–14. https://doi.org/10.1097/01974520-201110000-00002.

[19] C.A. Clerici, L. Veneroni, G. Bisogno, A. Trapuzzano, A. Ferrari, Videos on rhabdomyosarcoma on youtube: An example of the availability of information on pediatric tumors on the web, J. Pediatr. Hematol. Oncol. 34 (8) (2012) e329–e331.

[20] European Commission, Delivering on EU food safety and nutrition in 2050 – Future challenges and policy preparedness, 2016.

[21] K.J. Fortinsky, M.R. Fournier, E.I. Benchimol, Internet and electronic resources for inflammatory bowel disease: a primer for providers and patients, Inflamm. Bowel Dis. 18 (2012) 1156–1163, https://doi.org/10.1002/ibd.22834.

[22] K. Lee, A. Agrawal, A. Choudhary, Mining social media streams to improve public health allergy surveillance, in: Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2015, 2015: pp. 815–822. https://doi.org/10.1145/2808797.2808896.

[23] R. Wang, Y. He, J. Xu, H. Zhang, Fake news or bad news? Toward an emotion-driven cognitive dissonance model of misinformation diffusion, Asian J. Commun. 30 (2020) 317–342, https://doi.org/10.1080/01292986.2020.1811737.

[24] S.L. McNally, M.C. Donohue, K.P. Newton, S.P. Ogletree, K.K. Conner, S. E. Ingegneri, M.F. Kagnoff, Can consumers trust web-based information about celiac disease? Accuracy, comprehensiveness, transparency, and readability of information on the internet, Interact. J. Med. Res. 1 (2012) e1. https://doi.org/10.2196/ijmr.2010.

[25] G. Vici, L. Belli, M. Biondi, V. Polzonetti, Gluten free diet and nutrient deficiencies: a review, Clin. Nutr. 35 (2016) 1236–1241, https://doi.org/10.1016/j.clnu.2016.05.002.

[26] H.C. Lyson, G.M. Le, J. Zhang, N. Rivadeneira, C. Lyles, K. Radcliffe, R.J. Pasick, G. Sawaya, U. Sarkar, D. Centola, Social media as a tool to promote health awareness: results from an online cervical cancer prevention study, J. Cancer Educ. 34 (2019) 819–822, https://doi.org/10.1007/s13187-018-1379-8.

[27] P. Turina, P. Fariselli, E. Capriotti, ThermoScan: semi-automatic identification of protein stability data from PubMed, Front. Mol. Biosci. 8 (2021), https://doi.org/10.3389/fmolb.2021.620475.
[28] P.D. Karp, Can we replace curation with information extraction software?, Database. 2016. https://doi.org/10.1093/database/baw150.
[29] R. Rak, R.T. Batista-Navarro, A. Rowley, J. Carter, S. Ananiadou, Text-mining-assisted biocuration workflows in Argo, Database. 2014 (2014) 1–14, https://doi.org/10.1093/database/bau070.
[30] D.G. Jamieson, M. Gerner, F. Sarafraz, G. Nenadic, D.L. Robertson, Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database, Database. 2012 (2012), https://doi.org/10.1093/database/bas023.
[31] G. Zhu, A. Wu, X.J. Xu, P.P. Xiao, L. Lu, J. Liu, Y. Cao, L. Chen, J. Wu, X.M. Zhao, PPIM: a protein-protein interaction database for maize, Plant Physiol. 170 (2016) 618–626, https://doi.org/10.1104/pp.15.01821.
[32] R.K.R. Kalathur, J. Pedro Pinto, B. Sahoo, G. Chaurasia, M.E. Futschik, HDNetDB: a molecular interaction database for network-oriented investigations into Huntington's disease, Sci. Rep. 7 (2017) 1–12, https://doi.org/10.1038/s41598-017-05224-0.
[33] G. Babbi, P.L. Martelli, G. Profiti, S. Bovo, C. Savojardo, R. Casadio, eDGAR: a database of disease-gene associations with annotated relationships among genes, BMC Genom. 18 (2017) 25–34, https://doi.org/10.1186/s12864-017-3911-3.
[34] C.Y. Lin, J.Y. Lee, S.H. Huang, Y.C. Hsu, N.Y. Hsu, J.M. Yang, FooDisNET: a database of food-compound-protein-disease associations, in: Proc. - IEEE 20th Int. Conf. Bioinforma. Bioeng. BIBE 2020, Institute of Electrical and Electronics Engineers Inc., 2020: pp. 190–195. https://doi.org/10.1109/BIBE50027.2020.00039.
[35] A. Lamurias, F.M. Couto, Text mining for bioinformatics using biomedical literature, Encycl. Bioinforma. Comput. Biol. ABC Bioinforma. 1–3 (2019) 602–611, https://doi.org/10.1016/B978-0-12-809633-8.20409-3.
[36] M. Ammari, A. Chatr Aryamontri, H. Attrill, A. Bairoch, T. Berardini, J. Blake, Q. Chen, J. Collado, D. Dauga, J.T. Dudley, S. Engel, I. Erill, P. Fey, R. Gibson, H. Hermjakob, G. Holliday, D. Howe, C. Hunter, D. Landsman, R. Lovering, D. Manthravadi, A. Marchler-Bauer, B. Matthews, E.M. McDonagh, B. Meldal, G. MMicklem, D. Mietchen, C.J. Mungall, K. Pruitt, V. Sagar Rajamanickam, J.M. Reecy, A. Rey, K. Shameer, A. Luisa Toribio, M. Ann Tuli, P. Uetz, U. Wittig, V. Wood, T. Attwood, A. Bateman, T. Berardini, L. Bougueleret, P. Gaudet, J. Harrow, T. Imanishi, R. Kania, L. Richardson, M. Robinson-Rechavi, O. White, O. White, I. Xenarios, C. Yamasaki, C.N. Arighi, R. Balakrishnan, M.J. Cherry, M. Haendel, S.E. Lewis, P. McQuilton, M. Muñoz-Torres, C. O'Donovan, S. Orchard, S. Poux, A. Su, N. Vasilevsky, Z. Zhang, Biocuration: Distilling data into knowledge, PLoS Biol. 16 (2018) e2002846. https://doi.org/10.1371/journal.pbio.2002846.
[37] P. Jorge, M. Perez-Perez, G.P. Rodriguez, F. Fdez-Riverola, M.O. Pereira, A. Lourenco, Construction of antimicrobial peptide-drug combination networks from scientific literature based on a semi-automated curation workflow, Database. 2016 (2016) 14310–111093, https://doi.org/10.1093/database/baw143.
[38] M. Pérez-Pérez, P. Jorge, G. Pérez Rodríguez, M.O. Pereira, A. Lourenço, Quorum sensing inhibition in Pseudomonas aeruginosa biofilms: new insights through network mining, Biofouling. 33 (2017) 128–142, https://doi.org/10.1080/08927014.2016.1272104.
[39] J. Hur, A. Özgür, Y. He, Ontology-based literature mining of E. coli vaccine-associated gene interaction networks, J. Biomed. Semantics. 8 (2017), https://doi.org/10.1186/S13326-017-0122-4.
[40] W. Ben Abdessalem Karaa, M. Mannai, N. Dey, A.S. Ashour, I. Olariu, Gene-disease-food relation extraction from biomedical database, in: Adv. Intell. Syst. Comput., Springer, Cham, 2018: pp. 394–407. https://doi.org/10.1007/978-3-319-62521-8_34.
[41] T. Doğan, H. Atas, V. Joshi, A. Atakan, A. Rifaioglu, E. Nalbat, A. Nightingale, R. Saidi, V. Volynkin, H. Zellner, R. Cetin-Atalay, M. Martin, V. Atalay, CROssBAR: Comprehensive resource of biomedical relations with knowledge graph representations, Nucl. Acids Res. 49 (16) (2021) e96.
[42] M. Delmas, O. Filangi, N. Paulhe, F. Vinson, C. Duperier, W. Garrier, P.-E. Saunier, Y. Pitarch, F. Jourdan, F. Giacomoni, C. Frainay, FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases, Bioinformatics 37 (2021) 3896–3904. https://doi.org/10.1093/bioinformatics/btab627.
[43] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, W.M. Lim, How to conduct a bibliometric analysis: an overview and guidelines, J. Bus. Res. 133 (2021) 285–296, https://doi.org/10.1016/J.JBUSRES.2021.04.070.
[44] F. Yuan, J. Cai, B. Liu, X. Tang, Bibliometric analysis of 100 top-cited articles in gastric disease, Biomed Res. Int. 2020 (2020) 1–8. https://doi.org/10.1155/2020/2672373.
[45] Y. Guo, Z. Hao, S. Zhao, J. Gong, F. Yang, Artificial intelligence in health care: Bibliometric analysis, J. Med. Internet Res. 22 (2020), e18228. https://doi.org/10.2196/18228.
[46] L. Yang, Y. Zhou, Y. Zheng, Annotating the Literature with Disease Ontology, Chinese, J. Electron. 26 (2017) 1261–1268, https://doi.org/10.1049/CJE.2017.09.020.
[47] D. Tao, P. Yang, H. Feng, Utilization of text mining as a big data analysis tool for food science and nutrition, Compr. Rev. Food Sci. Food Saf. 19 (2020) 875–894, https://doi.org/10.1111/1541-4337.12540.
[48] P. Bakhtin, E. Khabirova, I. Kuzminov, T. Thurner, The future of food production–a text-mining approach, Technol. Anal. Strateg. Manag. 32 (2020) 516–528, https://doi.org/10.1080/09537325.2019.1674802.
[49] G. Jurca, O. Addam, A. Aksac, S. Gao, T. Özyer, D. Demetrick, R. Alhajj, Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends, BMC Res. Notes. 9 (2016) 1–35, https://doi.org/10.1186/s13104-016-2023-5.
[50] O. Edo-Osagie, B. De La Iglesia, I. Lake, O. Edeghere, A scoping review of the use of Twitter for public health research, Comput. Biol. Med. 122 (2020), 103770, https://doi.org/10.1016/j.compbiomed.2020.103770.
[51] T. Barrett, J. Beck, D.A. Benson, C. Bollin, E. Bolton, D. Bourexis, J.R. Brister, S. H. Bryant, K. Canese, K. Clark, M. Dicuccio, I. Dondoshansky, S. Federhen, M. Feolo, K. Funk, L.Y. Geer, V. Gorelenkov, M. Hoeppner, B. Holmes, M. Johnson, V. Khotomlianski, A. Kimchi, M. Kimelman, P. Kitts, W. Klimke, S. Krasnov, A. Kuznetsov, M.J. Landrum, D. Landsman, J.M. Lee, D.J. Lipman, Z. Lu, T. L. Madden, T. Madej, A. Marchler-Bauer, I. Karsch-Mizrachi, T. Murphy, R. Orris, J. Ostell, C. O'Sullivan, A. Panchenko, L. Phan, D. Preuss, K.D. Pruitt, W. Rubinstein, E.W. Sayers, V. Schneider, G.D. Schuler, S.T. Sherry, K. Sirotkin, K. Siyan, D. Slotta, A. Soboleva, V. Soussov, G. Starchenko, T.A. Tatusova, B. W. Trawick, D. Vakatov, Y. Wang, M. Ward, W.J. Wilbur, E. Yaschenko, K. Zbicz, Database resources of the National Center for Biotechnology Information, Nucl. Acids Res. 43 (2015) D6–D17, https://doi.org/10.1093/nar/gku1130.
[52] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L.G. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. Macdougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, B.E. Suzek, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, M.S. Yerramalla, J. Zhang, UniProt: a hub for protein information, Nucl. Acids Res. 43 (2015) D204–D212, https://doi.org/10.1093/nar/gku989.
[53] P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, C. Steinbeck, Chemical entities of biological interest: an update, Nucl. Acids Res. 38 (suppl_1) (2010) D249–D254. https://doi.org/10.1093/nar/gkp886.
[54] W.A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C.J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson, L.M. Schriml, D. Ontology, update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data, Nucl. Acids Res. 43 (2015) (2015) D1071–D1078, https://doi.org/10.1093/nar/gku1011.
[55] 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, Food on: a harmonized food ontology to increase global food traceability, quality control and data integration, Npj Sci. Food. 2 (2018) 1–10, https://doi.org/10.1038/s41538-018-0032-6.
[56] L.M. Schriml, Symptom Ontology, 2018. <http://www.obofoundry.org/ontology/symp.html%0Ahttps://bioportal.bioontology.org/ontologies/SYMP> (Accessed December 11, 2019).
[57] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucl. Acids Res. 45 (2017) D353–D361, https://doi.org/10.1093/nar/gkw1092.
[58] C.F. Thorn, T.E. Klein, R.B. Altman, PharmGKB: the pharmacogenomics knowledge base, Methods Mol. Biol. 1015 (2013) 311–320, https://doi.org/10.1007/978-1-62703-435-7_20.
[59] S.J. Nelson, W.D. Johnston, B.L. Humphreys, Relationships in Medical Subject Headings (MeSH), in: Springer, Dordrecht, 2001: pp. 171–184. https://doi.org/10.1007/978-94-015-9696-1_11.
[60] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: amajor update to the DrugBank database for 2018, Nucl. Acids Res. 46 (D1) (2018) D1074–D1082. https://doi.org/10.1093/nar/gkx1037.
[61] C. Rosse, J.L. V. Mejino, The Foundational Model of Anatomy Ontology, in: Anat. Ontol. Bioinforma., Springer London, 2008, pp. 59–117. https://doi.org/10.1007/978-1-84628-885-2_4.
[62] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, B. Parsia, The national cancer institute's thesaurus and ontology, SSRN Electron. J. (2018), https://doi.org/10.2139/ssrn.3199007.
[63] M. Pérez-Pérez, T. Ferreira, A. Lourenço, G. Igrejas, F. Fdez-Riverola, Boosting biomedical document classification through the use of domain entity recognizers and semantic ontologies for document representation: the case of gluten bibliome, Neurocomputing. 484 (2022) 223–237, https://doi.org/10.1016/j.neucom.2021.10.100.
[64] R. Leaman, C.H. Wei, Z. Lu, TmChem: a high performance approach for chemical named entity recognition and normalization, J. Cheminform. 7 (2015) S3, https://doi.org/10.1186/1758-2946-7-S1-S3.
[65] M. Gerner, G. Nenadic, C.M. Bergman, LINNAEUS: a species name identification system for biomedical literature, BMC Bioinformatics. 11 (2010) 85, https://doi.org/10.1186/1471-2105-11-85.

[66] R. Leaman, R.I. Doğan, Z. Lu, DNorm: Disease name normalization with pairwise learning to rank, Bioinformatics 29 (2013) 2909–2917. https://doi.org/10.1093/bioinformatics/btt474.
[67] B. Settles, ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text, Bioinformatics. 21 (2005) 3191–3192, https://doi.org/10.1093/bioinformatics/bti475.
[68] D.M. Jessop, S.E. Adams, E.L. Willighagen, L. Hawizy, P. Murray-Rust, OSCAR4: a flexible architecture for chemical textmining, J. Cheminform. 3 (2011) 41, https://doi.org/10.1186/1758-2946-3-41.
[69] M. Pérez-Pérez, T. Ferreira, G. Igrejas, F. Fdez-Riverola, A deep learning relation extraction approach to support a biomedical semi-automatic curation task: the case of the gluten bibliome, Expert Syst. Appl. 195 (2022), 116616, https://doi.org/10.1016/j.eswa.2022.116616.
[70] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, M. Albrecht, Computing topological parameters of biological networks, Bioinformatics. 24 (2008) 282–284. https://doi.org/10.1093/bioinformatics/btm554.
[71] E.C. Davenport, N.A. El-Sanhurry, Phi/Phimax: review and synthesis, Educ. Psychol. Meas. 51 (4) (1991) 821–828. https://doi.org/10.1177/0013164491051004003.
[72] A. Bawiskar, P. Sawant, V. Kankate, B.B. Meshram, Spring framework: a companion to JavaEE, IJCEM Int. J. Comput. Eng. Manag. (2012).
[73] Oracle, MySQL :: About MySQL, Oracle Corp. (2020).
[74] JGraph Ltd, JGraphT, JGraphT. (2016).
[75] E.W. Dijkstra, A note on two problems in connexion with graphs, Numer, Math 1 (1) (1959) 269–271. https://doi.org/10.1007/BF01386390.
[76] R. Steyer, jQuery, JQuery (2018), https://doi.org/10.3139/9783446456518.fm.
[77] L.P. Jhawar, D.D. Katyar, A review paper : bootstrap responsive framework, Int. J. Sci. Res. Dev. (2017).
[78] B. Sutar, D.R. Pratibha Adkar, Angular JS and its important component, Icon. Res. Eng. J. (2019).
[79] V. Veselá, Data better understanding by using of interactive visualization tools, Turkish Online J. Educ. Technol. (2015).
[80] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, BRAT: a web-based tool for NLP-assisted text annotation, in: Proc. Demonstr. 13th Conf. Eur. Chapter Assoc. Comput. Linguist. (2012) 102–107. https://dl.acm.org/citation.cfm?id=2380942 (accessed November 2, 2017).
[81] C.T. Lopes, M. Franz, F. Kazi, S.L. Donaldson, Q. Morris, G.D. Bader, Cytoscape Web: an interactive web-based network browser., Bioinformatics. 26 (2010) 2347–2348. https://doi.org/10.1093/bioinformatics/btq430.
[82] K. Collins, D. Shiffman, J. Rock, S. Goffredo, How are scientists using social media in the workplace? PLoS One 11 (10) (2016), e0162680. https://doi.org/10.1371/JOURNAL.PONE.0162680.
[83] Social media for scientists, Nat. Cell Biol. 2018 2012. 20 (2018) 1329–1329. https://doi.org/10.1038/s41556-018-0253-6.
[84] H.M. Bik, M.C. Goldstein, An introduction to social media for scientists, PLOS Biol 11 (2013), e1001535. https://doi.org/10.1371/JOURNAL.PBIO.1001535.
[85] V. Suarez-Lledo, J. Alvarez-Galvez, Prevalence of health misinformation on social media: systematic review, J. Med. Internet Res. 23(1) (2021) e17187. <https://doi.org/10.2196/17187>.
[86] J. Dong, S. Horvath, Understanding network concepts in modules, BMC Syst. Biol. 1 (2007), https://doi.org/10.1186/1752-0509-1-24.
[87] G. Serena, S. Camhi, C. Sturgeon, S. Yan, A. Fasano, The role of gluten in celiac disease and type 1 diabetes, Nutrients. 7 (2015) 7143–7162, https://doi.org/10.3390/NU7095329.
[88] O.N. Nadhem, G. Azeez, R.D. Smalligan, S. Urban, Review and practice guidelines for celiac disease in 2014, Postgrad. Med. 127 (2015) 259–265. <https://doi.org/10.1080/00325481.2015.1015926>.
[89] R. Minelli, F. Gaiani, S. Kayali, F. Di Mario, F. Fornaroli, G. Leandro, A. Nouvenne, F. Vincenzi, G.L. De'angelis, Thyroid and celiac disease in pediatric age: a literature review, Acta Biomed. 89 (2018) 11–16. <https://doi.org/10.23750/abm.v89i9-S.7872>.
[90] H.J. Freeman, Endocrine manifestations in celiac disease, World J. Gastroenterol. 22 (2016) 8472–8479, https://doi.org/10.3748/wjg.v22.i38.8472.
[91] J.Z. Zhang, D. Abudoureyimu, M. Wang, S.R. Yu, X.J. Kang, Association between celiac disease and vitiligo: a review of the literature, World J. Clin. Cases. 9 (2021) 10430–10437. https://doi.org/10.12998/wjcc.v9.i34.10430.
[92] F. Valitutti, C.M. Trovato, M. Montuori, S. Cucchiara, Pediatric celiac disease: follow-up in the spotlight, Adv. Nutr. 8 (2017) 356–361, https://doi.org/10.3945/an.116.013292.
[93] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R. K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A. T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T. L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P. J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E. J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R. S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, JoAnn Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S. R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G.R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F.A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M.J. Morgan, Initial sequencing and analysis of the human genome, Nature 409 (6822) (2001) 860–921. https://doi.org/10.1038/35057062.
[94] E.O. Ospina, The rise of social media – Our World in Data, Our World Data, 2019. <https://ourworldindata.org/rise-of-social-media> (Accessed May 30, 2022).