Applied Network Science

RESEARCH Open Access

NETME: on-the-fly knowledge network construction from biomedical literature

Alessandro Muscolino[1†], Antonio Di Maria[2†], Rosaria Valentina Rapicavoli[1†], Salvatore Alaimo[2], Lorenzo Bellomo[4], Fabrizio Billeci[3], Stefano Borzì[3], Paolo Ferragina[4], Alfredo Ferro[2] and Alfredo Pulvirenti[2*]

*Correspondence: alfredo.pulvirenti@unict.it
†Alessandro Muscolino, Antonio Di Maria and Rosaria Valentina Rapicavoli have contributed equally to this work.
[2] Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy
Full list of author information is available at the end of the article

Abstract

Background: The rapidly increasing biological literature is a key resource to automatically extract and gain knowledge concerning biological elements and their relations. Knowledge Networks are helpful tools in the context of biological knowledge discovery and modeling.

Results: We introduce a novel system called *NETME*, which, starting from a set of full-texts obtained from *PubMed*, through an easy-to-use web interface, interactively extracts biological elements from ontological databases and then synthesizes a network inferring relations among such elements. The results clearly show that our tool is capable of inferring comprehensive and reliable biological networks.

Keywords: Network analysis, Knowledge graph, Text mining

Introduction

The increasing amount of scientific literature is posing new challenges for scientists. Identifying the most relevant articles dealing with a topic is not straightforward, leading to the high chance of missing essential references and relevant literature. In particular, in research areas like biology or bio-medicine, thanks to fast-track publication journals, the number of published papers increases significantly fast.

On the other hand, network analysis has become a critical enabling technology to understand mechanisms of life, living organisms, and in general, uncover the underlying fundamental biological processes. Examples of applications include: (i) analyzing disease networks for identifying disease-causing genes and pathways (Barabási et al. 2010); (ii) discovering the functional interdependence among molecular mechanisms through network inference and construction Szklarczyk et al. 2016; (iii) releasing Network-based inference models with application on drug re-purposing (Himmelstein et al. 2017).

In the last few years, thanks to the availability of sizeable open-access article repositories such as *PubMed* Central (Beck 2010), arxiv (https://arxiv.org) bioarxiv (https://www.biorxiv.org/) as well as ontology databases which hold entities and their relations (Lambrix et al. 2007), the research community has focused on text mining tools and machine learning algorithms to digest these corpora and extract valuable semantic knowledge from them. Text mining (Cohen 2005), and Natural Language

Springer Open

© The Author(s), 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Processing (Krallinger et al. 2005) tools employ information extraction methods to translate unstructured textual knowledge in a form that can be easily analyzed and used to build a functional network (i.e. a network in which the relations between two entities are not necessarily physical but can be indirect), or knowledge graphs (Szklarczyk et al. 2016; Dörpinghaus et al. 2019; Nicholson and Greene 2020). This technology allows us to infer putative relations among molecules, such as understanding how proteins interact with each other or determining which gene mutations are involved in a disease. In the context of biology and biomedicine, the Biological Expression Language (BEL) (Slater 2014), or Resource Description Framework (RDF) (McBride 2004) have been widely applied to convert a text in semantic triplets having the following form: <subject, predicate, object>. The subject and object represent biological elements, whereas the predicate represents a logical or physical relationship between them (Szklarczyk et al. 2016; Himmelstein and Baranzini 2015).

However, the implementation of biological text mining tools requires highly specialized skills in Natural Language Processing and Information Retrieval. Therefore, several ecosystems and tools have been implemented and made available to the bioscience community. Relevant tools include PubAnnotation (Kim et al. 2019), a public resource for sharing annotated biomedical texts based on the "Agile text mining" concept; PubTator (PTC) (Wei et al. 2019), a web service for viewing and retrieving bio-concept annotations (for genes/proteins, genetic variants, diseases, chemicals, species, and cell lines) in full-text biomedical articles. This latter tool annotates all *PubMed* abstracts and more than three million full texts. The annotations are downloadable in multiple formats (XML, JSON, and tab-delimited) through the online interface, a RESTful web service, and bulk FTP. Another interesting tool is SemRep (Rindflesch and Fiszman 2003), which extracts relationships from biomedical sentences in PubMed articles by mapping textual content to an ontology that represents its meaning. To establish the binding relation, SemRep relies on internal rules (called "indicator rules"), which map syntactic elements, such as verbs, prepositions, and nominalization, to predicates in the Semantic Network. We also mention Hetionet (Himmelstein et al. 2017), a heterogeneous network of biomedical knowledge that unifies data from a collection of several available databases and millions of publications. Also, the edges are extracted from omics-scale resources and consolidated through multiple studies or resources. Finally, in Yuan et al. (2019) authors propose a minimally supervised approach for knowledge-graph construction based on 24,687 unstructured biomedical abstracts. Authors included entity recognition, unsupervised entity and relation embedding, latent relation generation via clustering, relation refinement, and relation assignment to assign cluster-level labels. The proposed framework can extract 16,192 structured facts with high precision.

Starting from our previous work (Muscolino et al. 2021), we introduce *NETME* a novel web-based app (available at https://netme.click/ website, and https://github.com/alemuscolino/netme.git github repository), which is capable of extracting knowledge from a collection of full-text documents. The tool orchestrates two different technologies:

- A customized version of the entity-linker *TAGME* (Ferragina and Scaiella 2010) (called *OntoTAGME*) for extracting network nodes (i.e., genes, drugs, diseases) from a collection of full-text articles.
- A software module, developed on top of SpaCy (Honnibal et al. 2020) and NLTK (Loper and Bird 2002) libraries, that derives relations (edges) between pair of nodes. Edges are weighted according to their frequency within the collection of full-texts used to create the on-fly knowledge graph.

These inferred networks are handy in biomedicine, where it is essential to understand the difference between various components and mechanisms, such as genes and diseases, and their relations, such as up-regulation and binding. Therefore, the tool helps scientists fast identify reliable relations among the biological entities under investigation, based on their occurrences and mentions in *PubMed* 's articles.

The novelties with respect our previous work (Muscolino et al. 2021) include:

- The sentence's grammatical structure is extracted by Spacy linguistic annotations. Such a structure includes the word types (parts of speech) and how the words are related to each other. In the previous NETME release, the nltk bottom-up and top-down approach were employed for building the syntactic tree of each document sentence. Furthermore, the Spacy's Matcher has been used to identify verbs' passive forms. With this approach the system is now capable of properly establishing the correct edge direction.
- In Muscolino et al. (2021), the proposed system was able to build a network composed of only genes, diseases, and drugs. Now, thanks to the extension we made on *OntoTAGME*, our new system is able to build networks composed of much more biological entities such as: genes, variants, diseases, drugs, compounds, molecular function, biological proves, pathways, enzymes, etc.
- Finally, we designed and implemented a new module to handle the disambiguation among gene symbols and the acronyms of diseases or other biological elements. In fact, in many documents, the authors assign acronyms for very long biological elements that are usually equal to genes symbols.

To the authors' knowledge, *NETME* is the first tool that allows to interactively synthesize biological knowledge-graphs on-the-fly starting from a *PubMed* query.

The paper is organized as follows. Section "The *NETME* model" introduces *NETME* system together with its components. Section "The annotation tool" provides the technical details of the back-end and the front-end of *NETME*. Section "Experimental analysis" reports two different case studies that allow evaluating *NETME* 's prediction qualitatively. The first one is focused on: (i) recovering known gene interactions; (ii) avoid false-negative ones. For this purpose, we selected a subset of gene-gene interactions in KEGG/REACTOME (Kanehisa and Goto 2000; Kanehisa 2019, 2000; Fabregat et al. 2017) by making use of STRING API. More precisely, such interactions were obtained by selecting 100 random gene-gene interactions (manually curated in KEGG or REACTOME database) for each of the following STRING text-mining score intervals: 500-600,600-700, 700-800, 800-900, >= 900. Next, we selected the first 100

pairs of non-interacting genes from the Negatome 2.0 database (Blohm et al. 2013; Smialowski et al. 2009) in order to understand if *NETME* can avoid false-negative interactions. The experiment yielded accuracy values from 58% when the STRING text-minig score is in [500, 600] interval, to 84% when the value of such a score is higher than 900. Whereas, the second case study is focused on building a "CD147-genes" interaction network through selected papers containing valuable information about CD147 gene. We compared the network returned by *NETME* against a manually-curated network derived from these selected papers. The experiment yielded 98% sensitivity and 100% specificity. Therefore, both experiments clearly showed the high reliability of *NETME* inferred networks. Moreover, we have also assessed the *NETME* performance for inferring "CD147-diseases" interactions by selecting 100 random interactions from DisGenNET, and the same "abstracts" used by DisGenNET for inferring these interactions. *NETME* detected 63 True Positive values out of 100, revealing a sensitivity of 63% Sect. "Conclusion" ends the paper and sketches future research directions.

The *NETME* model

A Knowledge Graph (also known as a semantic network) is a systematic way to connect information and data to knowledge. It represents a collection of interlinked descriptions of entities, real-world objects, and events, or abstract concepts, obtained from knowledge-bases such as ontologies $(O_1, O_2, \cdots, O_k)$. Basically, a semantic network is defined as a graph $G = (V, E)$ where entities are in V, and relationships in E. Each relation represents a connection between entities of one (intra-relationship) or more (inter-relationship) ontologies (Nettleton 2014). Therefore, there might exist a relation $e = (v_1, v_2) \in E$ where $v_1 \in O_i$ and $v_2 \in O_j$ with $i \neq j$.

An ontology is a formal description of knowledge as a set of domain-based concepts in relationships among them. As a result, the ontology does not only introduce a shareable and reusable knowledge representation, but it can also provide new knowledge about the considered domain (Xiaoke and Lin 2012).

NETME builds a biomedical knowledge graph starting from a set of n documents obtained through a query to the *PubMed* database. Papers can be sorted by relevance (default) or publication date. Users can also provide a list of PMCID/PMID or a set of PDF documents. The inferred network contains biological elements (i.e., genes, diseases, drugs, enzymes) as nodes and edges as possible relationships.

In Fig. 1 we outline the architecture of *NETME*. The user provides the query terms to perform the search on *PubMed*, and she may directly provide PDFs or PMCIDs/PMIDs of other pertinent documents. Then *NETME* begins to create the network as follows:

1 First, *OntoTAGME* converts the full-text of the input documents into a list of entities (nodes) using literature databases and ontologies (such as GeneOntology Consortium 2004, Drug-Bank Wishart et al. 2017, DisGeNET Piñero et al. 2019, and Obofoundry Smith et al. 2007) as corpus. These entities will be the knowledge graph nodes. Note that, Obofoundry contains a several ontologies, but only the following have been currently used in our model: GO, DO, PW, BTO, PRO, AEO, PATO, CL and CLO.

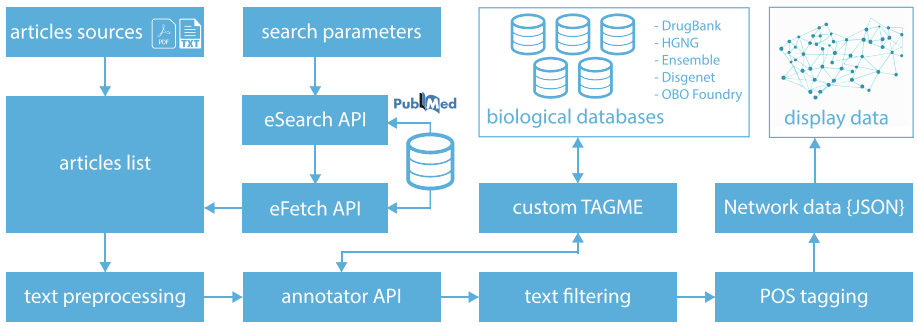


Fig. 1 *NETME* pipeline architecture

2 Next, an NLP model based on Python SpaCy (Honnibal et al. 2020), and NLTK (Loper and Bird 2002) libraries, is executed to infer the relations among nodes entity-nodes belonging to the same sentence (S_i) or to the adjacent ones (S_i, S_{i+1}) of the same document. Such relationships indicate disease treatment, genes regulations, molecular functions, gene-gene interactions, gene-disease interactions, gene-drug interactions, drug-disease interactions, disease-disease interactions and drug-drug interactions.

The final network will contain both directed and undirected edges according to the predictions made by the model. At the end of the process, the network will be rendered through Cytoscape JS. The following two subsections provide the details of these two phases.

OntoTAGME: Ontology oN Top Of *TAGME*

TAGME Ferragina and Scaiella (2010) is a state-of-the-art entity linker for annotating Wikipedia pages mentioned in an input text. The tool searches for sequences of words (spots) that can be linked to pertinent Wikipedia pages (entities) that explain those words in that context. The use of Wikipedia as corpus allows to enrich texts with explanatory links in order to provide a structured knowledge for any unstructured fragment of the text. These links are then used for drawing a network of relationships among the extracted spots.

To mitigate ambiguity and polysemy, *TAGME* computes a ρ value $\in [0, 1]$ for each Spot-Entity (Node) association, and keeps only those ones having the ρ value higher than an established user threshold. This value estimates the "goodness" of the annotation compared to other possible associations in the input text. A suitable use of ρ ensures the highest accordance among the extracted spots.

Due to the topics-generality of the Wikipedia corpus used by *TAGME*, several non-biological spots could be extracted during the annotation procedure. To overcome this limitation, we developed a customized version of *TAGME*, called *OntoTAGME*, which makes use of several ontology and literature databases, such as: GeneOntology (GO) (Consortium 2004), DiseaseOntology (DO) (Schriml et al. 2018), PathwayOntology (PW) (Petri et al. 2014), BRENDA tissueenzyme source (BTO) (Gremse et al. 2010), ProteinOntology(PRO) (Natale et al. 2016), Anatomical Entity Ontology (AEO)

(Bard 2012), Phenotype And Trait Ontology (PATO) (http://obofoundry.org/ontology/pato.html), Cell Ontology (CL) (Diehl et al. 2016), Cell Line Ontology (CLO) (Sarntivijai et al. 2014), DrugBank (Wishart et al. 2017), Disgenet (Piñero et al. 2019), HGNC (Gray et al. 2016), ENSEMBL (Birney 2004), CIViC (Griffith et al. 2017), and PharmGKB (Whirl-Carrillo et al. 2012). The usage of topic-specific ontology databases ensures reduced disambiguation errors and therefore yields highly reliable knowledge graphs inference.

The integration consisted of releasing a new intermediate python layer (Python Parser in Fig. 2), and a customized two-steps procedure (Wikipedia Adapter module in Fig. 2) for converting ontology databases in a *wikipedia-like* structure. The Python layer transforms a generic ontology or database in a list of CSV files: pages.csv, pageslink.csv and category.csv. The pages.csv stores the name of each biological element, and all possible synonyms. The pageslink.csv contains all the relationships among the nodes of the ontology. Finally, the category.csv has the type of each element extracted from the ontology or database entry (i.e Genes, Diseases, Drugs).

Next, a two-steps procedure is triggered to convert each row of the page.csv file into an XML file containing a unique ID generated by our system, the name (title), type (category) and the description (page's body) of the considered biological element. Since an element j could have several linked pages "LPs" (i.e. DOID:0002116 is a DOID:10124), or redirected pages "RPs" due to synonyms (CD147 is a synonym of BSG), the process generates a tuple $\langle uniqueID_j, uniqueID_k \rangle$ for each element k belonging to LPs, and a tuple $\langle uniqueID_j, uniqueID_i \rangle$ for each element i belonging to RPs. These tuples are then stored in the SQL files "wiki-latest-pagelinks" and "wiki-latest-redirect", respectively.

Finally, the SQL and XML files are used to generate the complete *OntoTAGME* network. It contains 331 thousand of main nodes, 700 thousand of synonyms, and 4 million of relationships.

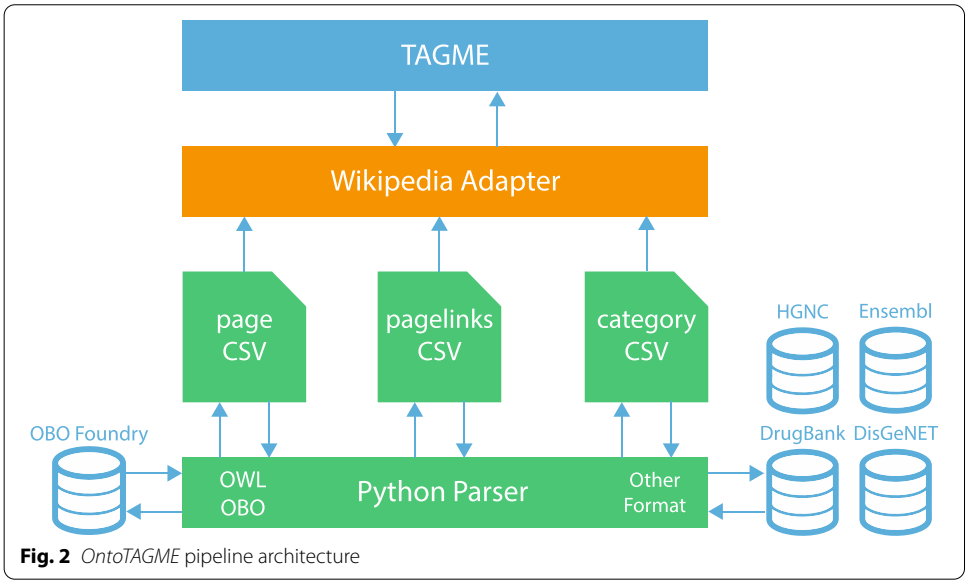


Fig. 2 *OntoTAGME* pipeline architecture

Ontology databases

In order to build the *OntoTAGME* annotation networks we used the following nine ontology and six bio-databases.

DrugBank Wishart et al. (2017) contains data about drugs name, drugs synonyms, drug-drug interaction, and other comprehensive drug-target information. The database release used in our project is the v5.1 which contains 13, 367 drugs entries, including 2, 611 approved small molecule drugs, 1, 300 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6, 315 experimental drugs. Additionally, 5, 155 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries.

HGNG (HUGO Gene Nomenclature Committee) Gray et al. (2016) assigns unique and informative gene symbols and names to human genes. Standardized HGNC approved nomenclature is used in publications and biomedical databases to remove ambiguity and facilitate communication between researchers worldwide. The last database release contains more than 40, 000 approved gene symbols of which over 19, 000 are for protein-coding genes. The HGNC also names a set of small and long non-coding RNA genes and pseudo-genes (659 since 2017). The genes are grouped on the basis of several shared characteristics such as homology, associated phenotype and encoded protein function.

Ensembl Birney (2004) contains genome annotation (i.e genes, variation, regulation and comparative genomics) across the vertebrate sub-phylum and key model organisms. This tool is also able to compute multiple alignments, predicts regulatory function and collects disease data. The last complete version of the Ensembl database has been downloaded through their FTP service, and then integrated in *OntoTAGME* thanks to Python Parser layer. All data in Ensembl are used in combination with those coming from HGNC to detect Genes name and symbols within a text.

Disgenet Piñero et al. (2019) contains collections of genes and variants associated with human diseases. It integrates data from scientific literature, GWAS catalogues, expert curated repositories and animal models. Additionally, several original metrics are provided to assist the prioritization of genotype–phenotype relationships. DisGeNET releases two types of databases, Gene-Disease Associations and Variant-Gene Associations.

CIViC Griffith et al. (2017) is an expert-crowd-sourced knowledge-base for Clinical Interpretation of Variants in Cancer describing the therapeutic, prognostic, diagnostic and predisposing relevance of inherited and somatic variants of all types. CIViC is committed to open-source code, open-access content, public application programming interfaces (APIs) and provenance of supporting evidence to allow for the transparent creation of current and accurate variant interpretations for use in cancer precision medicine.

PharmGKB Whirl-Carrillo et al. (2012) is an interactive tool for researchers investigating how genetic variation affects drug response. It displays genotype, molecular, and clinical knowledge integrated into pathway representations and Very Important Pharmacogene (VIP) summaries with links to additional external resources. A user may search and browse the knowledge-base by genes, variants, drugs, diseases, and pathways through the website: http://www.pharmgkb.org).

OBO Foundry Smith et al. (2007) is the Open Biological and Biomedical Ontology (OBO) Foundry. It provides well-formed and scientifically accurate ontology thanks to the collaboration of ontology developers. They contribute to develop an evolving set of principles and common syntax based on ontology models that ensure the proper functioning of the system. In *NETME*, we use the following list of ontology:

- Gene Ontology (GO) Consortium (2004) project provides a uniform way to describe the functions of gene products from organisms across all kingdoms of life and thereby enable analysis of genomic data. it contains more than 44 thousand GO terms, 8 millions of annotations, 1.5 millions of gene products and nearly 5 thousand species.
- Human Disease Ontology (DO) Schriml et al. (2018) is a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease.
- Pathway ontology (PW) Petri et al. (2014) is a controlled vocabulary for pathways that provides standard terms for the annotation of gene products.
- PRotein Ontology (PRO) Natale et al. (2016) defines taxon-specific and taxon-neutral protein-related entities in three major areas: proteins related by evolution; proteins produced from a given gene; and protein-containing complexes.
- BRENDA tissue / enzyme source (BTO) Gremse et al. (2010) is a structured controlled vocabulary for the source of an enzyme comprising tissues, cell lines, cell types and cell cultures.
- Anatomical Entity Ontology (AEO) Bard (2012) is an ontology of anatomical structures that expands CARO, the Common Anatomy Reference Ontology, to about 160 classes using the is_a relationship; it thus provides a detailed type classification for tissues. The AEO is useful in increasing the amount of knowledge in anatomy ontology, facilitating annotation and enabling interoperability across anatomy ontology.
- Phenotype And Trait Ontology (PATO) (http://obofoundry.org/ontology/pato.html) is used in conjunction with other ontologies such as GO or anatomical ontology to refer to phenotypes. Examples of qualities are red, ectopic, high temperature, fused, small, edematous and arrested.
- Cell Ontology (CL) Diehl et al. (2016) is designed as a structured controlled vocabulary for cell types. This ontology covers cell types from prokaryotes to mammals. However, it excludes plant cell types. One of the main uses of the CL is to describe samples used in transcriptomic and functional genomics studies, such as FANTOM5, ENCODE and LINCS.
- Cell Line Ontology (CLO) Sarntivijai et al. (2014) is a community-driven ontology that is developed to standardize and integrate cell line information and support computer-assisted reasoning.

The data relating to the number of nodes and relationships extracted from each mentioned ontology have been listed in Table 1

Table 1 Number of nodes and edges per ontology

Ontology name	Nodes number	Edges number
go	43917	142086
doid	10862	29938
pr	326811	846366
pw	2619	6210
cl	10809	34410
clo	44712	91966
aeo	248	523
bto	6515	9378
pato	4610	13027

Network edge inference

Once the network nodes have been extracted the system will annotate their position and their main characteristics within the text. We capture the significant elements in each sentence, by making use of the parts of speech (POS tags). Then through a syntactic analysis we verify the coherence of the extracted elements. Indeed, sentences have an internal organization that can be represented using a tree. Solving a syntax analysis problem for a sentence consists of looking for predefined syntactic forms which, like a tree, branch out from the single words. The main syntactic form is the sentence (S) which contains noun phrases (NP) or verb phrases (VP) that are formed by further elementary syntactic forms such as nouns (N), verbs (V), determiners (DET), etc (see Table 3). All these information will be used by the textual analysis phase to infer relations between them.

A transition-based dependency parser is then used to first check the syntactic coherence and then build the syntactic tree. The dependency parser component inside the spaCy library jointly learns sentence segmentation and labelled dependency parsing. The parser uses a variant of the non-monotonic arc-eager transition-system (Honnibal and Johnson 2015), with the addition of a break transition to perform the sentence segmentation. Nivre's (2005) pseudo-projective dependency transformation is also used to allow the parser to predict non-projective parses. The parser is trained through an imitation learning objective. It follows the actions predicted by the current weights and, at each state, it determines which actions are compatible with the optimal parse that could be reached from the current state. The weights are updated in a way that the scores assigned to the set of optimal actions is increased, while scores assigned to other actions are decreased. Note that more than one action may be optimal for a given state.

Once *OntoTAGME* have extracted the set of nodes $n_1, \ldots, n_z$ from a list of N full-text documents $[p_1, p_2, \ldots, p_N]$, the edge inference module of *NETME* (developed on top of the Python library NLTK Loper and Bird 2002 and spaCy (Honnibal et al. 2020)) starts to establish any verbal relationships between those pairs of nodes. When two or more nodes are detected within a sentence or adjacent sentences, the syntactic analyzer extracts the parts of speech and syntactic dependencies within the sentence. For each sentence we then get a set of labelled tokens $lt_1, lt_2 \ldots, lt_{k_i}$. Each token is a tuple of the following form $\{token, POS, dependency\_label\}$, where POS and Dependency label are valued with the data present in Table 3.

Irrelevant POS are filtered out (stop-words, URLs, etc.), we keep only the useful verb forms and the nodes which correspond to the noun parts. A final pruning phase is also executed in which we use: (i) POS tag labels and dependency labels to check if the syntactic link between the verb form and the annotations is correct and consistent, as described in the Fig. 3; (ii) a dictionary of biological verb forms to check if they are pertinent. The surviving nodes and verb forms will allow to generate network edges.

In our final network, each edge $e = (a, b)$ is weighted with three parameters: the term frequency and inverse document frequency (tf.idf), the medium relatedness (*mrho*) and the biological degree (bio). More specifically, tf.idf is a measure of how much information the edge provides, namely if it is common or rare across all input documents. In formula, we compute $\mathrm{tf.idf}(e, p, P) = \mathrm{tf}(e, p) * \mathrm{idf}(e, P)$.

Where, term frequency $\mathrm{tf}(e, p)$ is the frequency of edge e, is defined as $\mathrm{tf}(e, p) = f_{e,p} / \sum_{e' \in p} f_{e',p}$, with $f_{e,p}$ representing the number of times that edge e occurs in paper p. The inverse document frequency $\mathrm{idf}(e, P)$ is a measure of how much information the edge e provides. It is defined as $\mathrm{idf}(e, P) = logN/|\{p \in P : e \in p\}|$, where N is the number of documents analyzed by the query such that $N = |P|$, and $|\{p \in P : e \in p\}|$ is the number of documents where the edge e appears. The parameter *mrho* measures the relatedness of the labels starting from the ρ value assigned by *OntoTAGME* to the two annotations involved, i.e. $mrho(e) = \frac{\rho_a * \rho_b}{2}$. The *bio*-parameter is the cosine similarity (having a value ranging from 0 to 1) between the inferred relationship and a set of biological verb forms (see Table 2). Figure 4 provides an example of such an annotation.

The annotation tool

NETME is provided with a front-end developed in PHP and Javascript, in which the network rendering is performed through the CytoscapeJS library (Franz et al. 2015). Its back-end, which integrates *OntoTAGME*, is written in Java and communicates with both Python NLTK (Loper and Bird 2002) and SpaCy (Honnibal et al. 2020) libraries for the NLP module. *PubMed* search is performed with the Entrez Programming Utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/), a set of server-side programs providing a stable interface to the Entrez database and to the query system at the National Center for Biotechnology Information (NCBI).

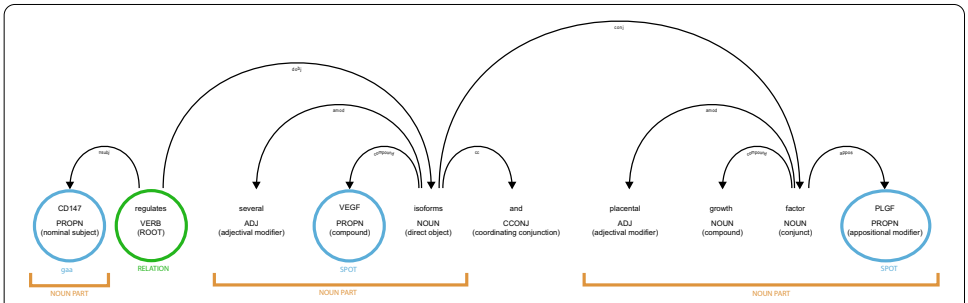


Fig. 3 *NETME* example of POS extraction and coherence checking for the sentence *[...] CD147 regulates several VEGF isoforms and placental growth factor (PLGF), and it has unique effects on trophoblastic function. [...]*. Through *OntoTAGME* we detect the spots [*"BSG", "VEGFA", "PGF"*]. After the syntactic analysis, three noun parts are identified (the phrase spots, highlighted via orange segments): two of them ("VEGF" and "PLGF") have a joint relationship with the first ("CD147"). The verbal part is the root between the two pairs of nouns ("CD147"—"VEGF"), (CD147—"PLGF")

Table 2 List of biological verb forms

Verb forms		
Activate	Downregulate	Reduce
Affect	Enhance	Regulate
Associates	Express	Release
Block	Find	Reveal
Cause	Inactivate	Stimulate
Contain	Increase	Trigger
Control	Induce	Ubiquitination
Decrease	Interacts	Upregulates
Detect	Overexpress	
Display	Produce	

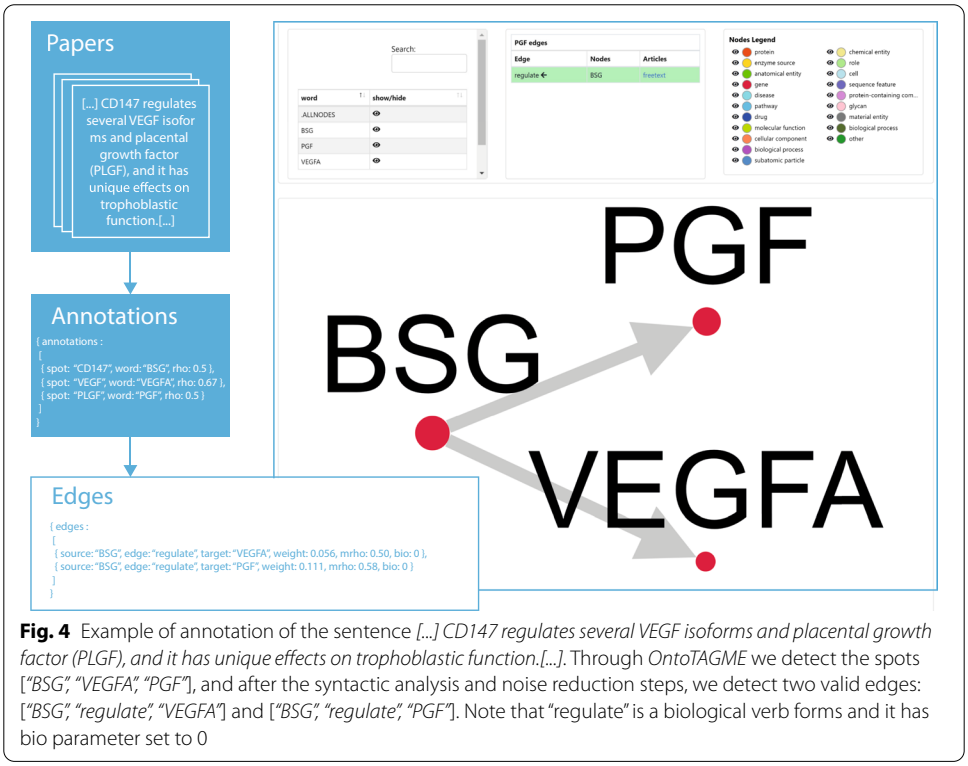


Fig. 4 Example of annotation of the sentence *[...] CD147 regulates several VEGF isoforms and placental growth factor (PLGF), and it has unique effects on trophoblastic function.[...]*. Through *OntoTAGME* we detect the spots ["*BSG*", "*VEGFA*", "*PGF*"], and after the syntactic analysis and noise reduction steps, we detect two valid edges: ["*BSG*", "*regulate*", "*VEGFA*"] and ["*BSG*", "*regulate*", "*PGF*"]. Note that "regulate" is a biological verb forms and it has bio parameter set to 0

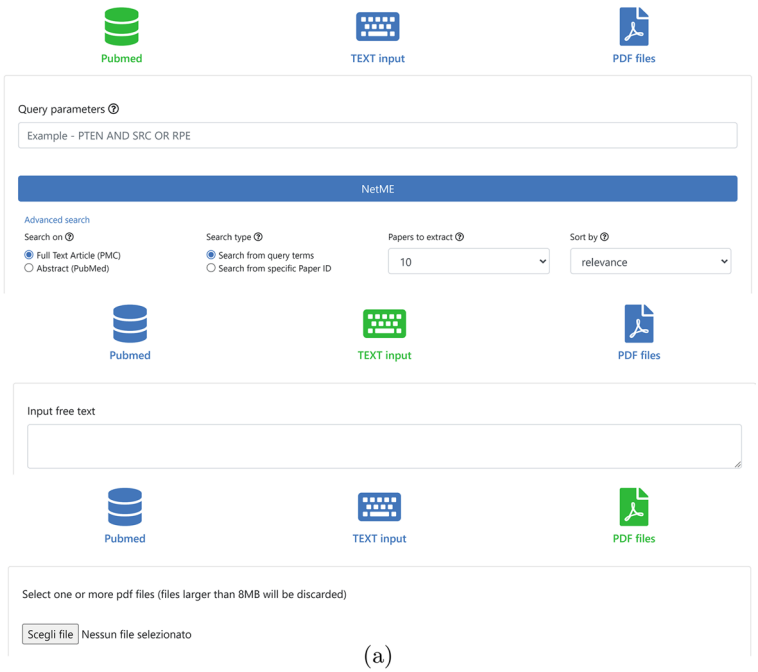
NETME is equipped with an easy-to-use web interface providing three major functions (see Fig. 5): (i) Pubmed query-based network annotation; (ii) user-provided free-text network annotation; (iii) user-provided PDF documents network annotation.

In the *query-based network annotation*, the user provides a list of keywords, which are employed to run a query on *PubMed*, or a list of article ids. The top resulting papers are retrieved and then the network inference procedure is run. Several parameters can be set by the user (or left with default values) such as: the number of top article to retrieve from *PubMed*, and the criteria used to sort papers (relevance or date).

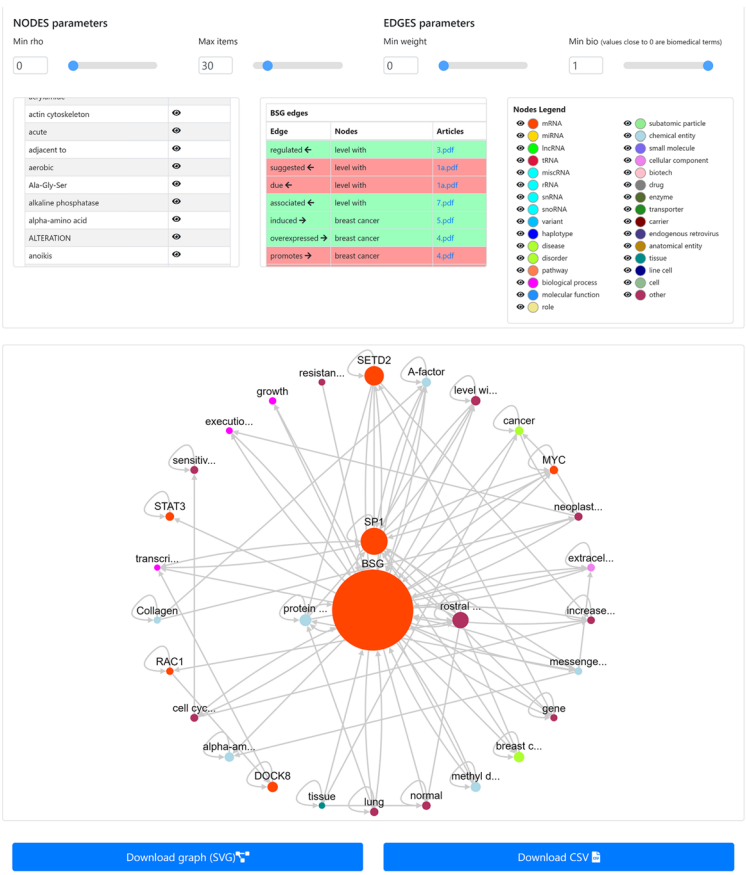
In the *user-provided free-text network annotation*, users provide a free text which is then input to the network inference procedure.

Table 3 List of POS tag and syntactic dependency labels

POS tag		Dependency label	
Symbol	**Meaning**	**Symbol**	**Meaning**
ADD	email	acl	clausal modifier of noun (adjectival clause)
AFX	affix	acomp	adjectival complement
CC	conjunction, coordinating	advcl	adverbial clause modifier
CD	cardinal number	advmod	adverbial modifier
DT	determiner	agent	agent
EX	existential there	amod	adjectival modifier
FW	foreign word	appos	appositional modifier
HYPH	punctuation mark, hyphen	attr	attribute
IN	conjunction, subordinating or preposition	aux	auxiliary
JJ	adjective	auxpass	auxiliary (passive)
JJR	adjective, comparative	case	case marking
JJS	adjective, superlative	cc	coordinating conjunction
LS	list item marker	ccomp	clausal complement
MD	verb, modal auxiliary	compound	compound
NFP	superfluous punctuation	conj	conjunct
NN	noun, singular or mass	csubj	clausal subject
NNP	noun, proper singular	csubjpass	clausal subject (passive)
NNPS	noun, proper plural	dative	dative
NNS	noun, plural	dep	unclassified dependent
PDT	predeterminer	det	determiner
POS	possessive ending	dobj	direct object
PRP	pronoun, personal	expl	expletive
PRP$	pronoun, possessive	intj	interjection
RB	adverb	mark	marker
RBR	adverb, comparative	meta	meta modifier
RBS	adverb, superlative	neg	negation modifier
RP	adverb, particle	nmod	modifier of nominal
SYM	symbol	npadvmod	noun phrase as adverbial modifier
TO	infinitival "to"	nsubj	nominal subject
UH	interjection	nsubjpass	nominal subject (passive)
VB	verb, base form	nummod	numeric modifier
VBD	verb, past tense	oprd	object predicate
VBG	verb, gerund or present participle	parataxis	parataxis
VBN	verb, past participle	pcomp	complement of preposition
VBP	verb, non-3rd person singular present	pobj	object of preposition
VBZ	verb, 3rd person singular present	poss	possession modifier
WDT	wh-determiner	preconj	pre-correlative conjunction
WP	wh-pronoun, personal	predet	None
WP$	wh-pronoun, possessive	prep	prepositional modifier
WRB	wh-adverb	prt	particle
		punct	punctuation
		quantmod	modifier of quantifier
		relcl	relative clause modifier
		xcomp	open clausal complement



(a)



(b)

Fig. 5 *NETME* web interface in (**a**), generated network in (**b**)

In the *user-provided PDF documents network annotation*, users give a set of PDF documents which are then input to the network inference procedure.

The result of the network inference procedure is a direct graph (network) which shows all inference details in three main tables containing: the list of extracted papers, the list of annotations, and the list of edges together with their weight.

The user can then click on a node of the network to view all incoming and outgoing connections, or she can click on an edge to display its type and the verbal relation between the nodes it connects.

Experimental analysis

To analyze the reliability of *NETME* knowledge graphs, we performed two case studies. The first one aims at providing a comprehensive analysis of *NETME* performance by checking its ability to predict known relations between genes drawn from Kyoto Encyclopedia of Genes and Genomes - KEGG (Kanehisa and Goto 2000; Kanehisa 2019, 2000) or REACTOME (Croft et al. 2010; Joshi-Tope 2004; Croft et al. 2013) pathways and, on the other hand, its ability to avoid inferring false connections between proteins by using the Negatome 2.0 database (Blohm et al. 2013; Smialowski et al. 2009). The second case study is more specific and focuses on building a network based on some selected publications that contain valuable information specific to the CD147 gene. Such a network is then compared against a manually-curated one derived from the same papers by a bio-expert. In both cases, the performance of *NETME* has been measured in terms of a precision/recall curve.

Case study 1

The first case study focuses on assessing *NETME* performance through its capability to recover known gene interactions. For this purpose, we selected a subset of gene-gene interactions from KEGG/REACTOME by making use of STRING API. More precisely, such interactions were obtained by selecting 100 random gene-gene interactions for each of the following STRING text-mining score intervals: 500–600, 600–700, 700–800, 800–900, $\geq$ 900 (listed in Additional files 1, 2, 3, 4, 5, respectively). These interactions form the true-positive set.

Next, we selected 100 random pairs of non-interacting genes from the Negatome 2.0 database as a true-negative set (listed in Table 5). For each interacting gene-pairs, we queried *NETME* with the papers used by STRING to infer the interactions. On the other hand, to annotate non-interacting genes, we queried *NETME* with the pair of genes of interest, selecting the top 20 papers from *PubMed*. Accuracy, sensitivity, specificity and PPV values, detected by *NETME*, are listed in Table 4 The results clearly show that NETME produces reliable results when the annotations are performed on top of relevant literature (STRING text-mining score higher than 700). On the other hand, when the STRING text-mining score is lower than 700, the NETME performances degrade in accordance with STRING predicted confidence as highlighted by their score . The reason behind such a behaviour is due: (i) not enough literature about these interactions; (ii) the interactions have been inferred by human curators as a combination of other interactions occurring in the text. Furthermore, when the text-mining score is small, STRING predictions could be wrong. In fact, as reported in Szklarczyk et al. (2016), a score of 500

would indicate that roughly every second term of an interaction might be erroneous (i.e., a false positive). Therefore, the computed value of accuracy, sensitivity, specificity and PPV could be incorrect.

Case study 2

Many tools (Alaimo et al. 2020) and computational models rely on existing network databases, such as KEGG (Kanehisa and Goto 2000; Kanehisa 2019, 2000) and Reactome (Croft et al. 2010; Joshi-Tope 2004; Croft et al. 2013). However, despite the enormous amount of available data, these databases are still incomplete and therefore have partial information (Menche et al. 2015). As an example, KEGG includes approximately one-third of the known genes. In this case study, we have chosen CD147, also known as Basigin (BSG) or EMMPRIN, as a starting point for the gene-gene interactions network construction. This gene represents an example of a biological element that should be supplemented to the KEGG network since it is not currently described in their pathways. Among the bibliography consulted to build the network manually, we have carefully selected 11 papers containing a significant amount of helpful information for our purpose. On the other hand, in this case study, we have also assessed the capabilities of *NETME* in inferring CD147-diseases relations. For this purpose we selected 100 random interactions from DisGenNET (Piñero et al. 2019), as well as the same abstracts used by DisGenNET for inferring such interactions (listed in Additional file 6).

CD147 is a transmembrane glycoprotein of the immunoglobulin superfamily, expressed in many tissues and cells, which is known to participate in several high biological and clinical relevance processes and is a crucial molecule in the pathogenesis of several human diseases (Xiong et al. 2014). Recently Wang et al. (2020) discovered an interaction between host cell receptor CD147 and SARS-CoV-2 spike protein, together with Angiotensin-Converting Enzyme 2 (ACE2), as an entry point for SARS-CoV-2.

In this direction, CD147 is an example of how a missing crucial gene within a biological network can compromise scientists' efforts to understand certain molecular phenomena. In literature, there are many valuable tools (Himmelstein et al. 2017; Himmelstein and Baranzini 2015) to integrate the missing information into bio-databases, such as KEGG. However, the most reliable approach in terms of accuracy and updated information remains the manual curation of such networks through careful and time-consuming literature analysis. On the other hand, a manually constructed network provides partial information due to the limited number of articles that a scientist could read. Our second case study affords this issue by providing a practical example of how *NETME* can create valuable networks by analyzing quickly and automatically larger sets of publications. The set of 11 selected papers, described in Fig. 7a, was analyzed by a bio-expert to derive a CD147-genes interactions network manually. This process resulted in 50 genes and 64 interactions, as shown in Fig. 7a. Next, by using the same set of papers, we run *NETME* with no upstream filter. The automatically generated network consisted of 86 genes and 139 relationships between them (see Fig. 7a, b). As the manually curated network consists of genes and proteins, only elements from these two categories were selected for the evaluation. This was performed by considering edges with the lowest "bio" score for each node pair. Qualitatively, this network includes most of the interconnections mentioned in the papers, thus providing a reliable and comprehensive overview of the molecular

function of Basigin. Quantitatively, *NETME* achieved an accuracy of 98.99%, a sensitivity of 100%, a specificity of 98.98%, and a positive predicted value of 46.32%.

Figure 6a–c depicts the precision/recall curve (AUC 0.997), the sensitivity/specificity curve and the True positive rate/False Positive Rate one. The construction of the curves considered all possible gene-pairs and their edges.

Finally, we queried *NETME* with the selected 100 random CD147-diseases interactions in DisGenNET, selecting the same PubMed abstract used by DisGenNET for inferring those interactions. *NETME* detected 63 True Positive values out of 100, revealing a sensitivity of 63%

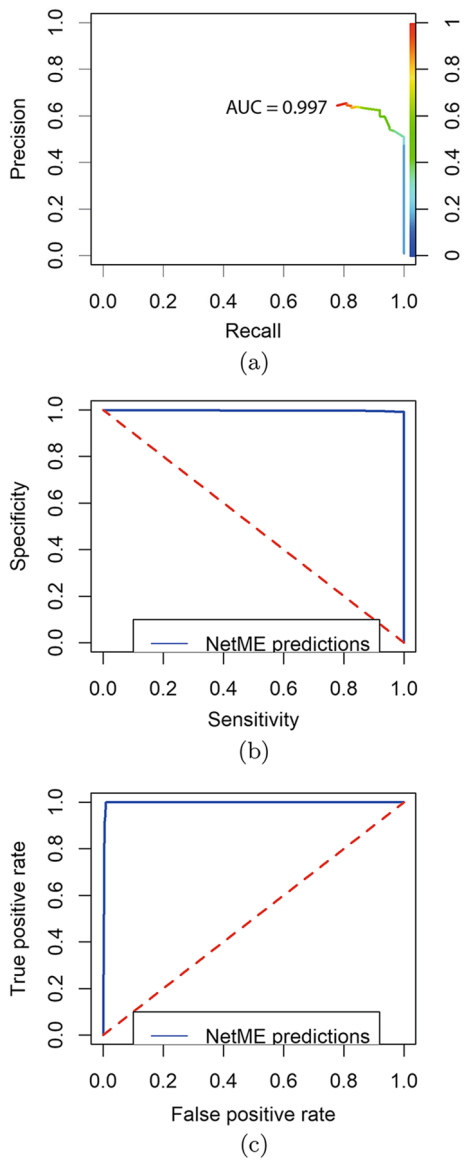


Fig. 6 Metrics of BSG-network performed by *NETME*. The plots show **a** Precision/Recall curve; **b** Sensitivity/Specificity; **c** True positive rate/False Positive Rate. The red dashed line in **b**, **c**, indicates the expected result if the used method was random that is any method which, given a pair of nodes, elects whether between them there is a link with a probability of 0.5

Table 4 Metrics on *NETME* 's ability to predict known interactions (from KEGG/Reactome) and non-interactions (from Negatome 2.0) between genes

Text-mining score interval	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)
500–600	58.5	31	86	68.8
600–700	66.5	47	86	77.05
700–800	72.5	59	86	80.8
800–900	73.5	61	86	81.3
$\geq$ 900	84	82	86	85.4

It is essential to stress that *NETME* allows us to extract a satisfactory and valid amount of information in a few minutes, compared to a manual search that may take days or weeks. We also believe that this case study is significant because, in the evaluation, we considered not only the presence of a link between two nodes but even more closely the type of edge, hence the adequacy and specificity of the annotated edge in its biological context.

Conclusions

In this paper, we have introduced *NETME* system to infer on-the-fly knowledge-graphs from a collection of either full-text papers obtained from *PubMed* or user-provided ones. It has been implemented upon a customized version of *TAGME*, called *OntoTAGME*, in connection to a syntactic analysis module developed on top of the Python NLTK and SpaCy libraries. Our results clearly show that *NETME* allows extracting reliable knowledge graphs in a few minutes or hours compared to a manual search that could take several days or weeks. The completeness of the extracted knowledge increases when the documents used by *NETME* comprehensively describe the desired topic under study. To evaluate *NETME*, we performed two case studies. The first one tested the ability of *NETME* in recovering relationships between genes. The experiment yielded accuracy ranging from 58%, when using low reliable relations (i.e. False Positives) from STRING,

(See figure on next page.)

Fig. 7 **a** Depicts the pathway constructed by hand from the selected papers (Jiang et al. 2014; Kong et al. 2014; Ke et al. 2012; Grass and Toole 2016; Xiong et al. 2014; Rucci et al. 2010; Ding et al. 2017; Ulrich and Pillat 2020; Wang et al. 2014; Kong et al. 2014; Kirk et al. 2000), with CD147(BSG) as the central node. **b** Shows the molecular mechanisms summarised in the knowledge network developed by *NETME* in accordance with the same papers used in **a** *NETME* shows that CD147 is a potent inducer of metalloproteinases (MMPs) such as MMP2, MMP14 and MMP9 as reported in Xiong et al. (2014); Rucci et al. (2010); Ding et al. (2017). Furthermore, the overexpression of CD147, which results in increased phosphorylation of PI3K(PIK3CA), Akt(AKT1), leads to the secretion of vascular endothelial growth factor (VEGFA) in several biological contexts such as KSHV infection Xiong et al. (2014); Rucci et al. (2010). In addition to its ability to induce MMPs, CD147 regulates spermatogenesis, lymphocyte reactivity and MCT system, in particular MCT1 and MCT4 (MCTS1 and SLC16A4) expression (Xiong et al. 2014; Kirk et al. 2000). Our results also show that CD147 can increase the expression of ATP-binding cassette transporter G2 (ABCG2) protein, regulating its function as a drug transporter, as mentioned by Xiong et al. for MCF-7 cells (Xiong et al. 2014). *NETME* identifies also BSG as an upstream activator of STAT3, highlighting its involvement in tumor development in agreement with the literature (Wang et al. 2014). As summarized by our knowledge network, CD147 is regulated by various inflammatory mediators, such as RANKL (TNFSF11), denoting its involvement in inflammatory processes (Grass and Toole 2016; Rucci et al. 2010). Among the potential activators of BSG, *NETME* also find the transcription factor c-Myc (MYC) (Kong et al. 2014)

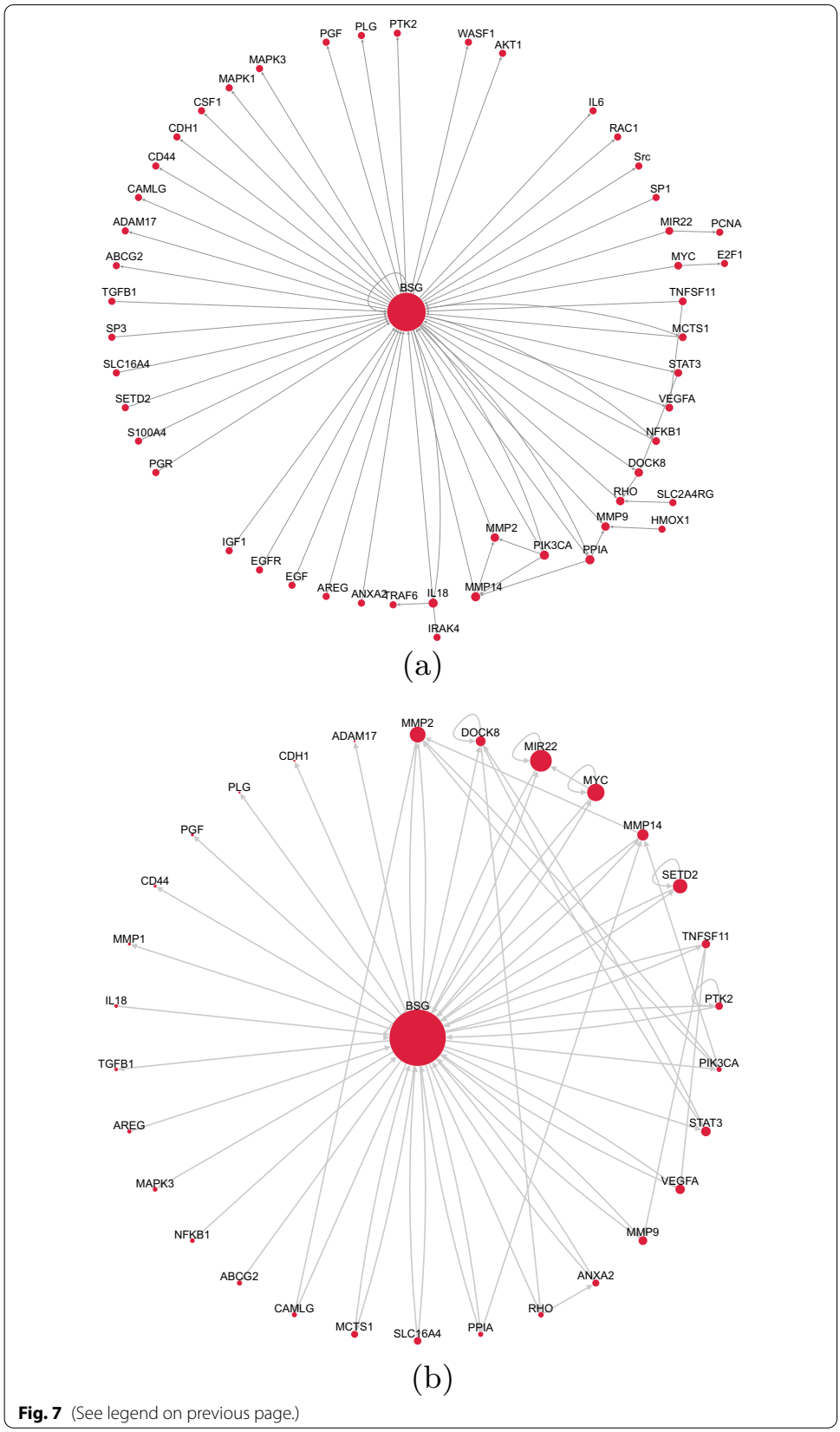


Fig. 7 (See legend on previous page.)

Table 5 List of the first 100 pairs of non-interacting genes from the Negatome 2.0 database.The column "SOURCE" indicates the starting gene, instead the column "TARGET" indicates the gene to which the action of the source gene is directed

Non-interacting genes from Negatome 2.0			
SOURCE	TARGET	SOURCE	TARGET
AKT1	TSC1	MAD2L2	MAD1L1
ARAF	BCL2L1	NCK1	EGFR
ARAF	BCL2	OSM	LIFR
BCL10	BIRC3	PARD3	LIMK1
BCL2L1	MAVS	PDGFC	FLT1
BMPR1A	TGFB1	PFN4	ACTB
BMPR1A	BMP5	PGF	KDR
BMPR1A	BMP6	PIAS3	STAT1
BMPR1B	TGFB1	PIK3CG	PIK3R2
BMPR1B	BMP5	PKN1	RPS6KA1
BMPR1B	BMP6	PKN1	RPS6KA3
BMPR2	BMP2	PKN1	MAP3K2
CCND1	MCM2	PKN2	RPS6KA1
CCR3	CCL3	PKN2	RPS6KA3
CCR3	CCL4L2	PKN2	MAP3K3
CD274	CD28	RB1	SMAD3
CD274	CTLA4	RBL2	SMAD3
CD274	ICOS	RIPK1	TNFRSF10A
CD3G	ZAP70	RIPK1	TNFRSF10B
CD74	NOTCH1	SFN	TSC1
CDKN1B	TSC1	SH3KBP1	TNFRSF14
CSF2	IL3RA	SMAD1	ANAPC10
CTNNB1	HSP90AA1	SMAD4	ANAPC10
CTNNB1	DDIT3	SOCS3	JAK2
CTNND1	IL2	STIM1	TRPC6
CTNND1	APC	TANK	RBCK1
CTNND1	CTNNA1	TBC1D7	TSC2
CTNND1	CTNND1	TFDP1	CDK2
CTNND1	CTNNB1	TFDP1	CCNA1
DKK1	WNT1	TICAM1	TLR4
DKK1	SOST	TJAP1	F11R
DVL1	TSC1	TJAP1	CLDN1
EIF3I	ACVR2A	TJAP1	TJP1
EIF3I	ACVR1	TNF	EGFR
EIF3I	TGFBR1	TRADD	TNFRSF10A
EP300	CD44	TRADD	TNFRSF10B
ERBB2	PIK3R2	TRAF6	IRF3
ETS1	CREBBP	TSC1	CDKN1B
FOXO1	TSC1	VAV1	SHC1
GRAP2	SOS1	VEGFB	KDR
GRAP2	CBL	VEGFB	FLT4
HDAC2	RELA	VEGFC	FLT1
HIPK2	MDM2	VIPR2	RAMP1
HSPA4	BAX	VIPR2	RAMP2
IGF2	IGF1R	VIPR2	RAMP3
IL15	IL2RA	VWF	F8
IL1A	EGFR	YWHAB	TSC1

Table 5 (continued)

Non-interacting genes from Negatome 2.0			
IL22	IL10RA	YWHAE	TSC1
IL4R	IL13	YWHAZ	TSC1
KDR	FLT1	NFKBIA	CREB3L2

to 84% when such STRING relations are very reliable. At the same time, the second case study tested the ability of *NETME* in integrating knowledge about genes starting from a selected set of papers. The experiment yielded 98% sensitivity and 100% specificity. Therefore, both experiments clearly showed the high reliability of *NETME* 's inferred networks.

Future work will include: (i) the construction of knowledge-graphs from all the open-access papers stored in PubMed Central; (ii) the integration of all Obofoundry ontology within *OntoTAGME*; (iii) the design of a more effective algorithm to select the pertinent papers on which *NETME* has to be applied (Ponza et al. 2019, 2020); and finally, add a methodology that allows to extract context-based relationships

Abbreviations

BEL: Biological Expression Language; RDF: Resource Description Frame-work; PTC: Pub-Tator; spots: Words; HGNC: HUGO Gene Nomenclature Committee; APIs: Application programming inter-faces; VIP: Very Important Pharmacogene; GO: Gene Ontology; DO: Human Disease Ontology; PW: Pathway ontology; PRO: PRotein Ontology; BTO: BRENDA tissue/enzyme source; AEO: Anatomical Entity Ontology; PATO: Phenotype And Trait Ontology; CL: Cell Ontology; CLO: Cell Line Ontology; POS: Parts of speech; S: Sentences; NP: Noun phrase; VP: Verb phrase; N: Nouns; V: Verbs; DET: Determiners; mrho: Medium relatedness; BSG: Basigin; ACE2: Angiotensin-Converting Enzyme 2; PLGF: Placental growth factor; MMPs: Metalloproteinases; VEGFA: Vascular endothelial growth factor; ABCG2: ATP-binding cassette transporter G2; MYC: Transcription factor c-Myc.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-021-00435-x.

Additional file 1. The json files storing all gene1-gene2 pairs used in the first case study having String scores ranging from 500 to 600. The main key of each record is the name ofthe two genes concatenated by "-". The lists of documents, are under the sub-keys "PMID" and "PMC".

Additional file 2. The json files storing all gene1-gene2 pairs used in the first case study having String scores ranging from 600 to 700. The main key of each record is the name ofthe two genes concatenated by "-". The lists of documents, are under the sub-keys "PMID" and "PMC".

Additional file 3. The json files storing all gene1-gene2 pairs used in the first case study having String scores ranging from 700 to 800. The main key of each record is the name ofthe two genes concatenated by "-". The lists of documents, are under the sub-keys "PMID" and "PMC".

Additional file 4. The json files storing all gene1-gene2 pairs used in the first case study having String scores ranging from 800 to 900. The main key of each record is the name ofthe two genes concatenated by "-". The lists of documents, are under the sub-keys "PMID" and "PMC".

Additional file 5. The json files storing all gene1-gene2 pairs used in the first case study having String scores greater than 900. The main key of each record is the name ofthe two genes concatenated by "-". The lists of documents, are under the sub-keys "PMID" and "PMC".

Additional file 6. The json files storing all BSG-Disease available in DisGenNET. The lists of documents, are under the sub-keys "PMCID".

Acknowledgements

Not applicable

Authors' contributions

AP, PF, SA, and AF conceived the work and coordinated the research. ADM and AM designed and developed the system. SB worked on the first version of ***OntoTAGME***. LB and FB realized the extension of ***OntoTAGME***. SA tested the system. VR

conducted the analysis of the performance of the system. ADM, AM, VR and AP wrote the paper. All authors read and approved the final manuscript.

Funding
AP, SA, AF, have been partially supported by the following research projects: MIUR PON BILIGeCT "Liquid Biopsies for Cancer Clinical Management" (CUP B96G18000590005); PO-FESR Sicilia 2014-2020 "DiOncoGen: Innovative diagnostics" (CUP G89J18000700007). AP, has been also partially supported by the following research project: "PROMOTE: Identificazione di nuovi biomarcatori per la diagnosi precoce di mesotelioma maligno pleurico in soggetti ex esposti a fibre asbestiformi", University of Catania - Piano di incentivi per la ricerca 2020-2022. PF and LB have been supported by the EU H2020 programmes "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (INFRAIA-01-2018-2019, Grant # 871042), and by "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (Grant # 820437).

Availability of data and materials
The datasets generated and analysed during the current study are available at the following URL https://netme.click/. Additional files for reproducibility purpose are provided as supplementary materials.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Physics and Astronomy, University of Catania, Catania, Italy. [2]Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy. [3]Department of Maths and Computer Science, University of Catania, Catania, Italy. [4]Department of Computer Science, University of Pisa, Pisa, Italy.

Received: 31 March 2021 Accepted: 21 September 2021
Published online: 06 January 2022

References

Alaimo S, Rapicavoli RV, Marceca GP, La Ferlita A, Serebrennikova OB, Tsichlis PN, Mishra B, Pulvirenti A, Ferro A (2020) Phensim: phenotype simulator. bioRxiv. https://doi.org/10.1101/2020.01.20.912279
Barabási A, Gulbahce N, Loscalzo J (2010) Network medicine: a network-based approach to human disease. Nat Rev Genet 12(1):56–68. https://doi.org/10.1038/nrg2918
Bard JBL (2012) The AEO, an ontology of anatomical entities for classifying animal tissues and organs. Front Genet. https://doi.org/10.3389/fgene.2012.00018
Beck J (2010) Report from the field: PubMed central, an XML-based archive of life sciences journal articles. In: Proceedings of the international symposium on XML for the Long Haul: issues in the long-term preservation of XML. Mulberry Technologies, Inc. https://doi.org/10.4242/balisagevol6.beck01
Birney E (2004) An overview of ensembl. Genome Res 14(5):925–928. https://doi.org/10.1101/gr.1860604
Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D (2013) Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Res 42(D1):396–400. https://doi.org/10.1093/nar/gkt1079
bioRxiv. https://www.biorxiv.org/
Cohen AM (2005) A survey of current work in biomedical text mining. Brief Bioinform 6(1):57–71. https://doi.org/10.1093/bib/6.1.57
Consortium GO (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32(90001):258–261. https://doi.org/10.1093/nar/gkh036
Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P (2013) The reactome pathway knowledgebase. Nucleic Acids Res 42(D1):472–477. https://doi.org/10.1093/nar/gkt1102
Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L (2010) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39(Database):691–697. https://doi.org/10.1093/nar/gkq1018
Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y, Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, Slyke CEV, Vasilevsky NA, Haendel MA, Blake JA, Mungall CJ (2016) The cell ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semant. https://doi.org/10.1186/s13326-016-0088-7
Ding P, Zhang X, Jin S, Duan B, Chu P, Zhang Y, Chen Z, Xia B, Song F (2017) Cd147 functions as the signaling receptor for extracellular divalent copper in hepatocellular carcinoma cells. Oncotarget 8(31):51151–51163. https://doi.org/10.18632/oncotarget.17712
Dörpinghaus J, Apke A, Lage-Rupprecht V, Stefan A (2019) Data Exploration and Validation on dense knowledge graphs for biomedical research. arXiv:1912.06194
Ex S (2018) Entrez programming utilities help. https://www.ncbi.nlm.nih.gov/books/NBK25501/

Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, D'Eustachio P, Stein L, Hermjakob H (2017) Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinform. https://doi.org/10.1186/s12859-017-1559-2

Ferragina P, Scaiella U (2010) TAGME. In: Proceedings of the 19th ACM international conference on information and knowledge management—CIKM '10. ACM Press. https://doi.org/10.1145/1871437.1871689

Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD (2015) Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv557

Ginsparg P. arXiv https://arxiv.org

Grass GD, Toole BP (2016) How, with whom and when: an overview of cd147-mediated regulatory networks influencing matrix metalloproteinase activity. Biosci Rep 36(1):25. https://doi.org/10.1042/bsr20150256

Gray KA, Seal RL, Tweedie S, Wright MW, Bruford EA (2016) A review of the new HGNC gene family resource. Hum Genom. https://doi.org/10.1186/s40246-016-0062-6

Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D (2010) The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Res 39(Database):507–513. https://doi.org/10.1093/nar/gkq968

Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK, Wagner AH, Skidmore ZL, Wollam A, Liu CJ, Jones MR, Bilski RL, Lesurf R, Feng Y, Shah NM, Bonakdar M, Trani L, Matlock M, Ramu A, Campbell KM, Spies GC, Graubert AP, Gangavarapu K, Eldred JM, Larson DE, Walker JR, Good BM, Wu C, Su AI, Dienstmann R, Margolin AA, Tamborero D, Lopez-Bigas N, Jones SJM, Bose R, Spencer DH, Wartman LD, Wilson RK, Mardis ER, Griffith OL (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 49(2):170–174. https://doi.org/10.1038/ng.3774

Himmelstein DS, Baranzini SE (2015) Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. PLoS Comput Biol 11(7):1004259. https://doi.org/10.1371/journal.pcbi.1004259

Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. https://doi.org/10.7554/elife.26726

Honnibal M, Montani I, Van Landeghem S, Boyd A (2020) spaCy: industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303

Honnibal M, Johnson M (2015) An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, Portugal, pp 1373–1378. https://doi.org/10.18653/v1/D15-1162

Jiang Z, Hu S, Hua D, Ni J, Xu L, Ge Y, Zhou Y, Cheng Z, Wu S (2014) β3gnt8 plays an important role in CD147 signal transduction as an upstream modulator of MMP production in tumor cells. Oncol Rep 32(3):1156–1162. https://doi.org/10.3892/or.2014.3280

Joshi-Tope G (2004) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33(Database issue):428–432. https://doi.org/10.1093/nar/gki072

Kanehisa M (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30. https://doi.org/10.1093/nar/28.1.27

Kanehisa M (2019) Toward understanding the origin and evolution of cellular organisms. Protein Sci 28(11):1947–1951. https://doi.org/10.1002/pro.3715

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30. https://doi.org/10.1093/nar/28.1.27

Ke X, Fei F, Chen Y, Xu L, Zhang Z, Huang Q, Zhang H, Yang H, Chen Z, Xing J (2012) Hypoxia upregulates cd147 through a combined effect of hif-1α and sp1 to promote glycolysis and tumor progression in epithelial solid tumors. Carcinogenesis 33(8):1598–1607. https://doi.org/10.1093/carcin/bgs196

Kim J, Wang Y, Fujiwara T, Okuda S, Callahan TJ, Cohen KB (2019) Open agile text mining for bioinformatics: the PubAnnotation ecosystem. Bioinformatics 35(21):4372–4380. https://doi.org/10.1093/bioinformatics/btz227

Kirk P, Wilson MC, Heddle C, Brown MH, Barclay AN, Halestrap AP (2000) CD147 is tightly associated with lactate transporters MCT1 and MCT4 and facilitates their cell surface expression. EMBO J 19(15):3896–3904. https://doi.org/10.1093/emboj/19.15.3896

Kong L-M, Liao C-G, Zhang Y, Xu J, Li Y, Huang W, Zhang Y, Bian H, Chen Z-N (2014) A regulatory loop involving mir-22, sp1, and c-myc modulates cd147 expression in breast cancer invasion and metastasis. Can Res 74(14):3764–3778. https://doi.org/10.1158/0008-5472.can-13-3555

Krallinger M, Erhardt RA, Valencia A (2005) Text-mining approaches in molecular biology and biomedicine. Drug Discov Today 10(6):439–445. https://doi.org/10.1016/s1359-6446(05)03376-3

Lambrix P, Tan H, Jakoniene V, Strömbäck L (2007) Biological ontologies. In: Semantic web. Springer, pp 85–99. https://doi.org/10.1007/978-0-387-48438-9_5

Loper E, Bird S (2002) NLTK. In: Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics. Association for Computational Linguistics. https://doi.org/10.3115/1118108.1118117

McBride B (2004) The resource description framework (RDF) and its vocabulary description language RDFS. In: Handbook on ontologies. Springer, pp 51–65. https://doi.org/10.1007/978-3-540-24750-0_3

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabasi A-L (2015) Uncovering disease-disease relationships through the incomplete interactome. Science 347(6224):1257601–1257601. https://doi.org/10.1126/science.1257601

Muscolino A, Di Maria A, Alaimo S, Borzì S, Ferragina P, Ferro A, Pulvirenti A (2021) NETME: on-the-fly knowledge network construction from biomedical literature. In: Complex networks & their applications IX. Springer, pp 386–397. https://doi.org/10.1007/978-3-030-65351-4_31

Natale DA, Arighi CN, Blake JA, Bona J, Chen C, Chen S, Christie KR, Cowart J, D'Eustachio P, Diehl AD, Drabkin HJ, Duncan WD, Huang H, Ren J, Ross K, Ruttenberg A, Shamovsky V, Smith B, Wang Q, Zhang J, El-Sayed A, Wu CH (2016) Protein

ontology (PRO): enhancing and scaling up the representation of protein entities. Nucleic Acids Res 45(D1):339–346. https://doi.org/10.1093/nar/gkw1075
Nettleton D (2014) Data representation. In: Commercial data mining. Elsevier, pp 49–66. https://doi.org/10.1016/b978-0-12-416602-8.00004-2
Nicholson DN, Greene CS (2020) Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J 18:1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017
Nivre J, Nilsson J (2005) Pseudo-projective dependency parsing. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05). Association for Computational Linguistics, Ann Arbor, Michigan, pp 99–106. https://doi.org/10.3115/1219840.1219853
Petri V, Jayaraman P, Tutaj M, Hayman G, Smith JR, De Pons J, JF Laulederkind S, Lowry TF, Nigam R, Wang S, Shimoyama M, Dwinell MR, Munzenmaier DH, Worthey EA, Jacob HJ (2014) The pathway ontology—updates and applications. J Biomed Semant 5(1):7. https://doi.org/10.1186/2041-1480-5-7
Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI (2019) The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz1021
Ponza M, Ferragina P, Chakrabarti S (2020) On computing entity relatedness in wikipedia, with applications. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2019.105051
Ponza M, Ferragina P, Piccinno F (2019) Swat: a system for detecting salient wikipedia entities in texts. Comput Intell 35(4):858–890. https://doi.org/10.1111/coin.12216
Rindflesch TC, Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 36(6):462–477. https://doi.org/10.1016/j.jbi.2003.11.003
Rucci N, Millimaggi D, Mari M, Del Fattore A, Bologna M, Teti A, Angelucci A, Dolo V (2010) Receptor activator of nfkb ligand enhances breast cancer-induced osteolytic lesions through upregulation of extracellular matrix metalloproteinase inducer cd147. Can Res 70(15):6150–6160. https://doi.org/10.1158/0008-5472.can-09-2758
Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, Schürer SC, Pang C, Malone J, Parkinson H, Liu Y, Takatsuki T, Saijo K, Masuya H, Nakamura Y, Brush MH, Haendel MA, Zheng J, Stoeckert CJ, Peters B, Mungall CJ, Carey TE, States DJ, Athey BD, He Y (2014) CLO: the cell line ontology. J Biomed Semant 5(1):37. https://doi.org/10.1186/2041-1480-5-37
Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates CP, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C (2018) Human disease ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 47(D1):955–962. https://doi.org/10.1093/nar/gky1032
Slater T (2014) Recent advances in modeling languages for pathway maps and computable biological networks. Drug Discov Today 19(2):193–198. https://doi.org/10.1016/j.drudis.2013.12.011
Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Rattei T, Frishman D, Ruepp A (2009) The negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res 38(suppl-1):540–544. https://doi.org/10.1093/nar/gkp1026
Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11):1251–1255. https://doi.org/10.1038/nbt1346
Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder SMS, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, Von Mering C (2016) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 45(D1):362–368. https://doi.org/10.1093/nar/gkw937
Ulrich H, Pillat MM (2020) CD147 as a target for COVID-19 treatment: suggested effects of azithromycin and stem cell engagement. Stem Cell Rev Rep 16(3):434–440. https://doi.org/10.1007/s12015-020-09976-7
Wang S-J, Cui H-Y, Liu Y-M, Zhao P, Zhang Y, Fu Z-G, Chen Z-N, Jiang J-L (2014) CD147 promotes src-dependent activation of rac1 signaling through STAT3/DOCK8 during the motility of hepatocellular carcinoma cells. Oncotarget 6(1):243–257. https://doi.org/10.18632/oncotarget.2801
Wang, K, Chen W, Zhang Z, Deng Y, Lian J-Q, Du P, Wei D, Zhang Y, Sun X-X, Gong L, Yang X, He L, Zhang L, Yang Z, Geng J-J, Chen R, Zhang H, Wang B, Zhu Y-M, Nan G, Jiang J-L, Li L, Wu J, Lin P, Huang W, Xie L, Zheng Z-H, Zhang K, Miao J-L, Cui H-Y, Huang M, Zhang J, Fu L, Yang X-M, Zhao Z, Sun S, Gu H, Wang Z, Wang C-F, Lu Y, Liu Y-Y, Wang Q-Y, Bian H, Zhu P, Chen Z-N, (2020) CD147-spike protein is a novel route for SARS-CoV-2 infection to host cells. Signal Trans Target Ther 5(1):5. https://doi.org/10.1038/s41392-020-00426-x
Wei C, Allot A, Leaman R, Lu Z (2019) PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res 47(W1):587–593. https://doi.org/10.1093/nar/gkz389
Wg OT (2018) Phenotype and trait ontology. http://obofoundry.org/ontology/pato.html
Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE (2012) Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther 92(4):414–417. https://doi.org/10.1038/clpt.2012.96
Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46(D1):1074–1082. https://doi.org/10.1093/nar/gkx1037
Xiaoke M, Lin G (2012) Biological network analysis: insights into structure and functions. Brief Funct Genomics 11(6):434–442. https://doi.org/10.1093/bfgp/els045
Xiong L, Edwards C, Zhou L (2014) The biological function and clinical utilization of CD147 in human diseases: a review of the current scientific literature. Int J Mol Sci 15(10):17411–17441. https://doi.org/10.3390/ijms151017411
Yuan J, Jin Z, Guo H, Jin H, Zhang X, Smith T, Luo J (2019) Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowl Inf Syst 62(1):317–336. https://doi.org/10.1007/s10115-019-01351-4

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

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com