# Building a large gene expression-cancer knowledge base with limited human annotations

Stefano Marchesin [1,*], Laura Menotti [1], Fabio Giachelle [1], Gianmaria Silvello [1] and Omar Alonso [2,†]

[1]Department of Information Engineering, University of Padova, Via G. Gradenigo 6b, Padova 35131, Italy
[2]Applied Science, Amazon, 3075 Olcott St., Santa Clara, California 95054, USA

*Corresponding author: Email: stefano.marchesin@unipd.it
†Work does not relate to the author's position at Amazon.

## Abstract

**Cancer prevention is one of the most pressing challenges that public health needs to face. In this regard, data-driven research is central to assist medical solutions targeting cancer. To fully harness the power of data-driven research, it is imperative to have well-organized machine-readable facts into a knowledge base (KB). Motivated by this urgent need, we introduce the Collaborative Oriented Relation Extraction (CORE) system for building KBs with limited manual annotations. CORE is based on the combination of distant supervision and active learning paradigms and offers a seamless, transparent, modular architecture equipped for large-scale processing. We focus on precision medicine and build the largest KB on 'fine-grained' gene expression–cancer associations—a key to complement and validate experimental data for cancer research. We show the robustness of CORE and discuss the usefulness of the provided KB.**

**Database URL:** https://zenodo.org/record/7577127

## Introduction

In 2020, there were 19.2 million cancer cases worldwide. The World Health Organization estimates a 32% overall increase by 2040.[1] With this growing global burden, cancer prevention is one of the century's most pressing public health challenges and data-driven research is crucial for assisting the development of medical solutions to address them. In the last few years, cancer research increasingly relied on microarray and next-generation sequencing technologies, which provide raw 'experimental data' about 'gene expression–cancer interactions' [1, 2]. These interactions hold vital information to guide diagnosis, assess prognosis or predict therapy response [3, 4]. To fully harness the potential of this data, they must be made readily accessible and organized in a comprehensive manner. This is where Knowledge Bases (KBs) come into play, providing machine-readable knowledge that connects cutting-edge technologies with data-driven research to support the fight against cancer [5].

Although experimental data are invaluable for cancer research, they require further investigation, processing and validation by experts to be used to populate a KB. Luckily, in most cases, the analysis and interpretation of experimental data are described in peer-reviewed publications,

making the scientific literature a critical source to complement and validate the data. Given the high economical and time costs to manually extract knowledge ( e.g. scientific facts) from the domain-specific literature [6–8], in recent years, Machine Learning (ML) methods and automated techniques for Knowledge Base Construction (KBC) have flourished [9–11]. The main bottleneck for KBC systems is the requisite of large and expensive labeled training data to perform Named Entity Recognition and Disambiguation () and Relation Extraction (RE).

Distant supervision [12, 13] and active learning [14, 15] are the main paradigms adopted to address this limitation [16, 17]. In this work, we use both paradigms in conjunction to build a modular, pluggable, transparent and scalable KBC system focusing on the discovery of 'gene expression–cancer' associations.

We note that there are a handful of knowledge resources containing data about gene expression–cancer associations [18–23] and most of them only contain experimental data [18–21]. A small minority—e.g. BioXpress [22] and OncoMX [23]—also integrates knowledge extracted from the biomedical literature. To do so, they rely on Disease-Expression Relation Extraction from Text (DEXTER) [24], a state-of-the-art method based on 'pattern matching' techniques. CoMAGC [25] and OncoSearch [26] are other literature-based resources [27, 28], modeling also the

[1]https://gco.iarc.fr/tomorrow/en/dataviz/bubbles?sexes=0&mode=population

particularly valuable fine-grained aspects of gene expression–cancer associations. Still, CoMAGC only consists of 821 sentences on prostate, breast and ovarian cancers, while OncoSearch is not maintained.

More general and large-scale resources on gene–disease associations—i.e. DisGeNET [27] and the Literature-derived Human Gene-Disease Network (LHGDN) [28]—store coarse-grained information expressing the existence of an association between gene expression and cancer, which is often insufficient to model such complex, faceted relationships effectively.

Hence, there is a need for KBC systems that can scale to large text corpora and stay up to date while generating fine-grained information about gene expression–cancer associations. We present the Collaborative Oriented Relation Extraction (CORE) system, a KBC system based on the combination of automated ML-based methods and domain experts' feedback. CORE features a seamless, transparent, modular architecture, where the different components can be easily plugged-in. It also employs active learning to bootstrap the KB to be produced and exploits the fine-grained aspects involved in gene expression–cancer associations to perform iterative tests that measure the reliability of the data to be stored in the KB. Finally, it returns small, selected samples to domain experts for annotation. The high-quality data generated by this process are then used as reinforcement to retrain the ML models from scratch. Active learning makes CORE suited to iterative KB versioning. Therefore, with the data annotated by domain experts, retrained ML models are deployed to build subsequent versions of the KB.

The experimental evaluation shows the robustness of the proposed approach by highlighting how CORE scales to large text corpora with little human annotations. Additionally, when compared to the state-of-the-art in a KB reconstruction task, the performance of CORE confirms the system effectiveness in this domain. We have made the KB derived by CORE available as Open Data [29], provided with a SPARQL Protocol and RDF Query Language (SPARQL) endpoint for querying: http://w3id.org/corekb/sparql. The KB can also be accessed via COREKB [30], an intuitive and easy-to-use search platform: https://gda.dei.unipd.it. The source code is available at https://github.com/GDAMining/core.

The rest of the article is organized as follows: Section (Background) presents the required background; Section (The CORE system) outlines the CORE system; Sections (NERD)–(Active learning) describe the components of CORE; Section (Implementation and experiments) presents the system settings and the experiments; Section (KB exploration) performs some exploratory queries and analyses on the generated KB; Section (Search platform) showcases the search platform and Section (Conclusions and future work) draws final remarks.

## Background
### Precision medicine and gene expression
Generally, precision medicine is about tailoring of medical treatment to the characteristics of an individual patient and moves beyond the traditional approach of stratifying patients into treatment groups based on phenotypic markers [31]. Among the different fields of medicine, precision medicine has the greatest impact in cancer research. In this context,

precision medicine involves the use of an individual patient's genomics information to guide diagnosis, prognosis, treatment and prevention of cancer for that patient. In other words, precision medicine is a multifaceted approach that involves several critical aspects such as pharmacogenomics, pharmacodynamics and the impact of genetic variations on an individual's response to cancer treatments like chemotherapeutics [32]. It also considers various factors, including age, ethnicity, sex and lifestyle habits, as prognostic indicators, as well as diagnostic aspects related to disease identification and its severity. Among the factors that identify the different risk outcomes in patients, alterations in gene expression patterns play a central role [3]. In fact, genes contain the information required to create proteins and dictate cellular functions, but it is the 'gene expression' that determines the cellular phenotype—and therefore the disease development. Moreover, abnormalities in the expression of microRNAs—small RNAs that post-transcriptionally regulate the expression of their target genes—have recently been associated with cancer [33, 34]. Thus, identifying genes and microRNAs whose expression levels interact with cancer status is imperative to advance cancer research [3].

### Gene expression–cancer resources
Most knowledge resources about 'gene expression–cancer associations' consist of experimental data obtained through microarray and next-generation sequencing technologies [18–21]. A relevant example is GENT2 [18], which provides a search platform for gene expression patterns across different normal and tumor tissues, compiled from public gene expression datasets. Another platform is the Metabolic gEne RApid Visualizer [19], which provides access to gene expression datasets and compares gene expressions across human tissues and cell types. The International Cancer Genome Consortium [20] is a collaborative effort to characterize genomic abnormalities in 50 different cancer types, which provides a data portal containing data from 24 cancer projects, including The Cancer Genome Atlas [21]. As explained earlier, these resources are valuable but provide raw data that need to be further processed and validated to be effectively used in precision medicine.

Beyond experimental data, BioXpress [22] and OncoMX [23] also integrate knowledge extracted—either manually or automatically—from the biomedical literature [22, 23]. BioXpress is a KB storing genes that are differentially expressed in adjacent normal and tumor tissues from the same patients. On the other hand, OncoMX is a KB for exploring cancer biomarkers, which encompasses more than 1000 unique biomarker entries mapped to 20 576 genes that have either mutations or differential expressions in cancer. In particular, both resources consider data automatically extracted by DEXTER [24], a text-mining expert system that identifies gene and microRNA expressions in disease samples from sentences selected from the relevant literature. To the best of our knowledge, DEXTER is the most advanced text-mining system for gene expression–cancer associations. However, the use of pattern matching techniques—based on manually defined regular expressions—to extract relationships limits DEXTER flexibility and hampers its applicability to a broad range of new, unseen sentences. In this regard, we show that DEXTER fails to extract knowledge from sentences whose syntactic structure differs from its predefined patterns.

Thus, it appears clear that more adaptive RE methods are required to build KBC systems capable of scaling to large-scale text corpora, made of heterogeneous documents written in different styles.

Regarding literature-based resources, CoMAGC (25) is a corpus developed to train RE methods targeting gene expression–cancer associations. In CoMAGC, each sentence is annotated with different concepts that together express (i) how a gene changes, (ii) how a cancer changes and (iii) the interaction between gene and cancer changes. Together, the different concepts can be used to infer the prospective roles of genes in cancer and to classify genes into classes according to the inferred roles. Relying on the CoMAGC annotation schema, OncoSearch (26) retrieves sentences mentioning gene expression changes in cancers from Medline abstracts. Specifically, OncoSearch queries seek (i) whether a gene is up- or downregulated, (ii) whether a given cancer progresses or regresses based on the given gene expression change and (iii) the expected role of the gene in the cancer. While relevant, CoMAGC is small-scale and OncoSearch is not maintained, making them limited resources. On the other hand, general, large-scale and widely used resources about gene–disease associations, such as DisGeNET (27) and LHGDN (28), store information that is not specific enough to be effectively used to model gene expression–cancer relationships.

## Knowledge base construction

KBs have gained a great deal of attention recently as a key component for supporting search and recommendations at the Web scale. From seminal works such as DBPedia (35), Freebase (36) and YAGO (37) to community-driven projects such as WikiData (38), KBs have become a central asset in several applications. To build them, KBC spans different areas of data management and artificial intelligence. Data integration (39, 40), data cleaning (41), NERD (42–44), RE (45–47) and active learning (14, 15, 48) are critical to ensure the accurate and scalable construction of KBs. In this regard, several systems that include the latest advances in these areas have been built (9–11, 49).

KBs are also increasingly used by large companies and organizations as a means for organizing and understanding their data (50–53). Industry-scale KBs are typically derived from numerous sources and contain a wealth of information that is used in downstream applications (54). One of the few published reports of an end-to-end industrial system to build, maintain and use such KBs is the work by Deshpande et al. (50). We follow a similar focus on data quality and clear methodology for bootstrapping a KB. Besides, we agree that an imperfect KB is still useful for real-world applications and that maintenance is also important as facts evolve over time.

## The CORE system
### Preliminaries

Let us consider a directed graph $G = (V, E)$, where $E \subseteq \{(v_1, v_2) \mid (v_1, v_2) \in V \times V\}$ is the set of edges connecting ordered pairs of vertices. Given an edge $e = (v_1, v_2) \in E$, we call $v_1$ the source vertex and $v_2$ the target vertex. In our context, the nodes of $G$ are entities and the edges are the relationships between them.

**Table 1.** Description of the aspects involved in gene expression-cancer associations. For each aspect, we report its domain values and the corresponding descriptions.

| Aspect | Value | Description |
| --- | --- | --- |
| CGE | up | The expression of a gene is increased. |
| | down | The expression of a gene is decreased. |
| | notinf | The change of gene expression is unknown. |
| Change of cancer status (CCS) | progression | The cell or tissue acquires cancerous properties as gene expression level changes. |
| | regression | The cell or tissue loses cancerous properties as gene expression level changes. |
| | notinf | The change of cancerous properties of cell or tissue is unknown. |
| Gene–cancer interaction (GCI) | causality | There is a cause-effect relationship between CGE and CCS. |
| | correlation | There is a correlation between CGE and CCS. |
| | notinf | The interaction between CGE and CCS is unknown. |

**Definition 1.** (Aspect) We call aspect an attribute of a relationship between a pair of entities. An aspect $A_i$ has a name and a domain $D = \{a_{i1}, \dots, a_{in}\}$, where $a_{ij} \in A_i$ is the $j^{th}$ *aspect value* of $A_i$. $Dom(A_i) = D$ returns the domain of $A_i$ (when it is clear from the context, the aspect value $a_{ij} \in A_i$ is referred to as $a_j$).

**Example 1.** Let us consider gene–cancer associations. There are three aspects describing a possible relationship ($e$) between gene ($v_1$) and cancer ($v_2$): the change of gene expression (CGE), the change of cancer status (CCS) and the gene–cancer interaction (GCI). Following Definition 1, CGE, CCS and GCI are the names of the aspects with the following domains: $Dom(\text{CGE}) = \{\texttt{up}, \texttt{down}, \texttt{notinf}\}$, $Dom(\text{CCS}) = \{\texttt{progression}, \texttt{regression}, \texttt{notinf}\}$ and $Dom(\text{GCI}) = \{\texttt{causality}, \texttt{correlation}, \texttt{notinf}\}$.

Details about the aspect domains are given in Table 1.

**Definition 2.** (multi-aspect relationship) Given a graph $G(V, E)$ and a set of aspects $\mathcal{A} = \{A_i\}_{i=1}^n$, a tuple of aspect values $(a_{1j}, \dots, a_{nj})$ associated with $e = (v_1, v_2) \in E$ defines a multi-aspect relationship between $v_1$ and $v_2$.

**Definition 3.** (signature function) Given a set of aspects $\mathcal{A} = \{A_i\}_{i=1}^n$ and an alphabet $\Sigma$, we define $s : \prod_{i=1}^n A_i \to S \subseteq \Sigma^*; s(a_{1j}, \dots, a_{nj}) \mapsto \texttt{type}$ as the signature function that maps a multi-aspect relationship to a $\texttt{type}$ in $S$, called the signature set.

**Table 2.** Inference rules for gene classes. For each combination of CGE, CCS and GCI, we report the expected gene class. Gene classes refer to the role that a given gene plays in a specific disease. Following (25, 26), a biomarker represents a gene that exhibits altered expression levels in cancer, but which is not (yet) identified as an oncogene or a tumor suppressor gene. In Rule 5, CGE and CCS can assume any value between {up,down} and {progression,regression}.

| Rule number | CGE | CCS | GCI | Gene Class |
|---|---|---|---|---|
| 1 | up | progression | causality | oncogene |
| 2 | up | regression | causality | tumor suppressor gene |
| 3 | down | regression | causality | oncogene |
| 4 | down | progression | causality | tumor suppressor gene |
| 5 | up\|down | progression\|regression | observation | biomarker |

The signature function defines a set of mapping rules depending on the domain of interest. We use the signature function to map multi-aspect gene expression–cancer relationships to gene prospective roles in cancer ( e.g. oncogene or biomarker). Table 2 provides the inference rules used to derive the expected gene roles. Gene roles allow us to distinguish the genes that are responsible for oncogenesis from those that are not; this is essential information for effective cancer research and therapy design (55).

> **Definition 4.** (tagging function) Given an edge $e \in E$ and the signature set $S$, we define $\sigma : E \to S; \sigma(e) \mapsto$ type as the function tagging an edge with a signature type.

The tagging function associates a signature type with an edge of the graph. In this work, we use the tagging function to label edges with gene prospective roles. The graph represents gene expression–cancer associations as gene prospective roles in cancer.

## Overview

The goal of the CORE system is to harvest facts from text corpora to populate KBs. We model a KB as a directed graph $G$ made up of entities connected by typed relationships. Facts (or statements) are $(v_1, e, v_2)$ triples, where $v_1, v_2 \in V$, $e = (v_1, v_2) \in E$ and $\sigma(e) \in S$.

To obtain facts, CORE collects the scientific literature from different sources, identifies sentences containing pairs of entities relevant to the considered task and extracts aspects from them. Depending on the combination of extracted aspect values, a sentence expresses a specific signature type. Note that, for a given pair of entities, different sentences can express various signature types, as we show in the next example.

> **Example 2.** See these two sentences from the biomedical literature:
>
> (i) ***Colorectal cancer** (CRC) growth and progression is frequently driven by RAS pathway activation through upstream growth factor receptor activation or through mutational activation of KRAS or **BRAF**.*
> (ii) *Somatic mutations of the BRAF gene, causing constitutive activation of **BRAF**, have been found in various types of human cancers such as malignant melanoma, and **CRC**.*

In both sentences, the following entities are extracted $v_1$ = BRAF and $v_2$ = CRC. Considering the aspects introduced in Example 1, for sentence A, we find CGE = up, CCS = progression and GCI = causality, leading to the signature type s((up, progression, causality)) = oncogene. On the other hand, the aspect values of sentence B are CGE = up, CCS = progression and GCI = correlation, leading to the signature type s((up, progression, correlation)) = biomarker.

From Example 2, we see that different sentences may lead to different signature types. In the scientific discourse, it is not surprising that there are different viewpoints and that various studies can lead to different conclusions—even in contradiction with each other. Hence, we need to consider this potential uncertainty when facts are extracted from the literature. The CORE system models this inherent uncertainty by assigning the likelihood of being true to each aspect value. This probability is based on the evidence we can extract from the literature. Given a set of sentences concerning the same two entities, the more an aspect value is consistent in the set, the higher the probability for that value to be true.

> **Definition 5.** (aspect–probability set) Given an aspect $A_i = \{a_j\}_{j=1}^m$ such that each aspect value $a_j$ carries a likelihood $\Pr(a_j)$, we call $AP_i = \{(a_j, \Pr(a_j))\}_{j=1}^m$ the aspect–probability set of $A_i$.

> **Definition 6.** (multi-aspect function) Let $G = (V, E)$ be a directed graph and $\mathcal{AP} = \{AP_i\}_{i=1}^n$ a set of aspect–probability sets. We define $\phi : E \to \prod_{i=1}^n AP_i; \phi(e) \mapsto (\{(a_{1j}, \Pr(a_{1j}))\}_{j=1}^{|A_1|}, ..., \{(a_{nj}, \Pr(a_{nj}))\}_{j=1}^{|A_n|})$ as the multi-aspect function that, given an edge, returns the $n$-tuple of aspect–probability sets.

Thus, for each pair of target entities, CORE computes the probabilities for all the aspect values and combines them into tuples of aspect–probability sets—i.e. a probability distribution over multi-aspect relationships. Sentences serve as supporting or contradicting evidence that strengthens or weakens the likelihood of a fact.

## Architecture

Figure 1 gives an overview of the CORE architecture, depicting modules and processes, and Figure 2 zooms into it providing further details. The system acquires text from the literature and processes and normalizes it to obtain sentences, from
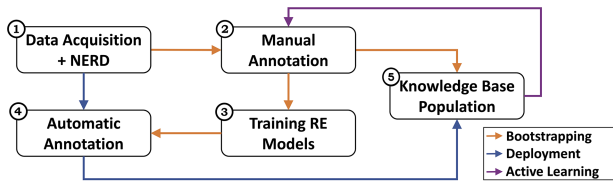
**Figure 1.** Overview of the CORE architecture. The system consists of five main modules and three processes. The modules represent the data acquisition and NERD components (1), the manual annotation activities (2), the training of the RE models (3), the subsequent automatic annotation (4), and the KB population (5). The processes reflect the different workflows: bootstrapping (orange) sets up the KBC process via expert involvement; deployment (blue) scales it through automated RE methods; and active learning (purple) allows refining the process through subsequent iterations.

which a NERD component detects and annotates the entity pairs (Module 1). These entity-annotated sentences undergo two different processes: bootstrapping and deployment. In the bootstrapping workflow, experts manually annotate multi-aspect relationships between the entities (Module 2), producing a set of 'relation-annotated sentences' used to train RE methods (Module 3) and to populate the KB (Module 5).

In the deployment workflow, the automatic annotations expressing multi-aspect relationships between entities are provided by the RE methods (Module 4) previously trained in the bootstrapping phase. Then, in the last module (Module 5), relation-annotated sentences are grouped by entity pairs and used to generate facts to further populate the KB. Module 5 is composed of (i) a knowledge enrichment component computing the probabilities for all the aspect values and combining them into tuples of aspect–probability sets and (ii) a reliability testing component that uses these probabilities to perform multiple tests to tag the facts as 'reliable' or 'unreliable'. Only the facts tagged as 'reliable' are used to populate the KB. When the deployment workflow is complete, 'unreliable' facts are ranked by ascending reliability score and the top-*k* automatically annotated sentences associated with them are reannotated by experts. This process triggers an active learning workflow that reinforces the RE methods.

## Versioning

The active learning workflow makes CORE suitable to iterative KB versioning. We define a KB version as the graph $G_j = (V_j, E_j)$ obtained after the $j^{th}$ iteration of the bootstrap and deployment workflows. Once the $j^{th}$ version of the KB has been deployed, the active learning workflow starts by generating the batch of unreliable sentences for bootstrapping the $j^{th} + 1$ version of the KB. The unreliable sentences are manually annotated and used to increase the size of the datasets to retrain the RE methods from scratch, which then generate a new set of automatic annotations to be included in the $j^{th} + 1$ KB version. Hence, once the bootstrap and deployment workflows are completed, the $j^{th} + 1$ version of the KB is rebuilt from scratch and comprises all the available annotations.

## NERD

CORE recognizes gene and cancer entities from text and links them to relevant and authoritative KBs. In our
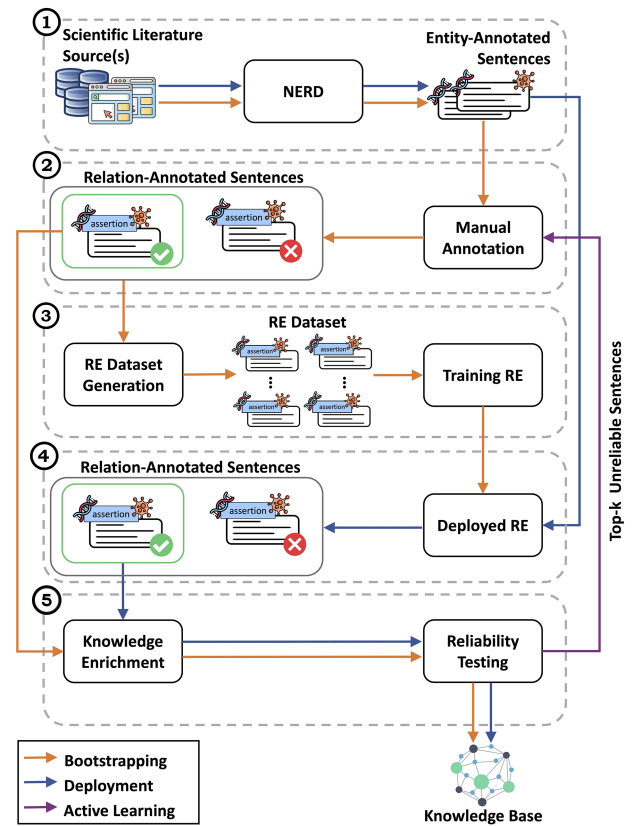


**Figure 2.** Detailed view of the CORE architecture. In module (1), CORE acquires text from biomedical literature and then performs NERD to generate entity-annotated sentences. These sentences are then manually annotated by experts in module (2) to produce relation-annotated sentences, which are used to generate the datasets for training RE methods in module (3). Once trained, in module (4), the RE methods are deployed over entity-annotated sentences to automatically generate relation-annotated sentences. Finally, in module (5), relation-annotated sentences undergo a knowledge enrichment component, which generates facts, and a reliability testing component, which tags facts as 'reliable' or 'unreliable'. Facts tagged as *'reliable'* are used to populate the KB, whereas *'unreliable'* facts are returned to experts for re-annotation.

setting, gene entities are linked to the National Center for Biotechnology Information (NCBI) Gene database (56), whereas cancer entities are linked to the Unified Medical Language System (UMLS) (57). The choice of UMLS as the reference KB for cancer aims to maximize the interoperability of the CORE system with different existing biomedical resources, such as DisGeNET, BioXpress and OncoMX.

As NERD component, CORE integrates the PubTator system (58–60). Given biomedical text, PubTator provides automated annotations from state-of-the-art text mining systems for genes/proteins, genetic variants, diseases, chemicals, species and cell lines. In particular, PubTator normalizes annotated genes to NCBI Gene identifiers and annotated diseases to MeSH (61) identifiers. However, the CORE system requires UMLS identifiers for diseases. Therefore, a mapping process normalizes MeSH identifiers to UMLS Concept Unique Identifiers (CUIs). Then, to restrict to cancer, CORE only keeps UMLS CUIs that belong to the neoplastic process' semantic type. Once gene and cancer entities have been extracted and linked to the reference KBs, CORE splits biomedical text into

sentences and keeps only those sentences containing gene–cancer pairs. When a sentence contains multiple gene–cancer pairs, CORE returns separate entity-annotated sentences for each pair.

## Manual annotation

CORE involves experts to manually annotate some sentences with multi-aspect relationships about gene expression–cancer associations. For annotation, CORE adopts a common and shared schema in the biomedical domain (e.g. CoMAGC (25) and OncoSearch (26)), where experts are required to annotate sentences with three different aspects: CGE, CCS and GCI. CGE represents the change of the gene expression level, CCS represents the change of the cancer status and GCI indicates the interaction occurring between CGE and CCS aspects. Each aspect can assume different values: $Dom$(CGE) = {up, down, notinf}; $Dom$(CCS) = {progression, regression, notinf}; $Dom$(GCI) = {causality, correlation, notinf}.

Even though a huge amount of sentences contains gene–cancer pairs in the biomedical literature, only a small fraction actually describes gene expression–cancer associations. It is therefore essential to have a tool capable of limiting the amount of noise introduced in the annotation process and maximizing sentence utility. In this regard, CORE requires the annotation of an additional aspect, the gene–cancer context (GCC). GCC indicates the coarse-grained association between gene and cancer and serves as a filter that helps differentiating between gene–cancer associations related to changes in the gene expression levels and those encompassing other types of gene–cancer relationships. To this end, GCC has the following domain: $Dom$(GCC) = {expression, other}. The expression value indicates that an altered gene expression is associated with cancer, whereas the other value represents any other gene–cancer association(note that other can be broken down into different and finer values, thus leaving room for the integration of different types of gene–cancer associations in the CORE system.)—including the absence of association. Thus, the GCC aspect assesses the sentence utility in context, because it is a filter limiting manual and automatic analysis of sentences containing gene–cancer pairs not inherent to gene expression–cancer associations.

Based on the annotation schema presented earlier, domain experts perform multi-aspect manual annotations between gene–cancer pairs and return relation-annotated sentences (Module 2). Depending on the considered workflow, the sentences to be annotated come from different modules at different stages. At the beginning of bootstrapping, (entity-annotated) sentences come from Module 1 as the output of the NERD component. After active learning, (unreliable) sentences come from Module 5 due to the reliability testing. In both cases, any errors associated with the NERD component are corrected too. For training RE methods (Module 3), the CORE system employs the complete set of relation-annotated sentences, whereas to populate the KB (Module 5), CORE keeps only the sentences with GCC = expression. The sentences annotated with CGE = notinf are excluded to limit noise injection, because CGE is the main aspect gene expression–cancer associations driver (25).

## Relation extraction

CORE's RE methods are trained to automatically annotate multi-aspect relationships on sentences. Once trained, the RE methods are applied on new, unseen sentences to generate knowledge and thus update the KB.

For each aspect to be annotated, a different RE method is trained (Module 3) and deployed (Module 4). Together, different aspect-based annotations compose the multi-aspect relationship. Although simple, this approach reflects the transparent and modular architecture of the CORE system, where different components can be easily plugged in and plugged out since every RE method can be retrained—or changed—without affecting others.

The RE methods serve two purposes: classify sentence utility and extract gene expression–cancer aspects. Every RE method presents the same architecture but addresses a different aspect. As the underlying ML model, all RE methods adopt SciBERT (62), a pretrained language model based on BERT (63). SciBERT addresses the lack of high-quality, large-scale labeled scientific data by pretraining on scientific papers from Semantic Scholar (64). On top of it, a linear layer takes SciBERT pooled output. Predictions are scores in $[0, 1]$ for target values. The higher the score for an aspect value, the more the RE method believes the sentence expresses that particular value. We formally define prediction scores as follows.

> **Definition 7.** (score function) Let $A_i$ be an aspect, $a_j$ one of its values and $T$ a set of sentences. We define score : $A_i \times T \to \mathbb{R}_{[0,1]}$; score$(a_j, t) \mapsto r$ as the score function that given a sentence $t$ returns how close the aspect value $a_j$ is to the truth.

Every RE method instantiates the score function tailored for its aspect extraction task. A RE method returns a specific sentence-aspect score always in the range of $[0, 1]$ of real numbers.

> **Remark 1.** Given a sentence $t$ and an aspect $A_i = \{a_j\}_{j=1}^m$, then $\sum_{a_j \in A_i}$ score$(a_j, t) = 1$. Prediction scores for an aspect $A_i$ given a sentence $t$ are a probability distribution over the aspect values $a_j$.

Thus, given an entity-annotated sentence (Module 1), CORE first masks gene and cancer entities with special tokens to avoid bias and then applies RE methods to extract CGE, CCS, GCI and GCC aspects. GCC extraction serves to assess sentence utility, while CGE, CCS and GCI extraction to compose multi-aspect relationships. For each aspect, the CORE system keeps the value associated with the highest score. For manual annotation, the scores are set to 1 if the aspect value is present and 0 otherwise.

Afterward, relation-annotated sentences where the extracted GCC value is expression are kept, whereas those with GCC = other are discarded. As in manual annotation, sentences with CGE = notinf are also filtered out. The retained set of automatic, relation-annotated sentences is used to populate the KB (Module 5).

## Knowledge enrichment

The relation-annotated sentences obtained from manual annotation (Module 2) and RE deployment (Module 4) pass through a knowledge enrichment component, which groups annotated sentences by gene–cancer pairs and generates facts. However, for a given gene–cancer pair, different sentences can have different multi-aspect annotations. This situation occurs because, in the literature, different studies and viewpoints can lead to different conclusions. To address this intrinsic uncertainty, the CORE system assigns to each aspect a likelihood to be true.

> **Definition 8.** (aspect value likelihood) Let $(v_1, v_2)$ be a pair of entities, $T_{(v_1, v_2)}$ the set of sentences annotated with both $v_1$ and $v_2$ and $a_j \in A_i$ the target aspect value. Then, the aspect value likelihood is
>
> $$\Pr(A_i = a_j \mid (v_1, v_2)) = \frac{\sum_{t \in T_{(v_1,v_2)}} \text{score}(a_j, t) \cdot \mathbb{1}(a_j, t)}{\sum_{t \in T_{(v_1,v_2)}} \max_{a_k \in A_i}(\text{score}(a_k, t))}, \quad (1)$$
>
> where $\mathbb{1}(\cdot, \cdot)$ represents the indicator function, defined as
>
> $$\mathbb{1}(a_j, t) = \begin{cases} \frac{1}{|\arg\max_{a_k \in A_i}(\text{score}(a_k, t))|}, & a_j \in \arg\max_{a_k \in A_i}(\text{score}(a_k, t)) \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

By modeling aspect value likelihoods this way, the CORE system takes the beliefs of RE methods into account. The more the RE methods are confident about an aspect value $a_j$ over the others, the more its likelihood increases. Vice versa, if RE methods have a larger degree of uncertainty across the different aspect values, then the likelihood for $a_j$ also decreases accordingly.

> **Remark 2.** Given a pair of entities $(v_1, v_2)$, its set of annotated sentences $T_{(v_1, v_2)}$ and an aspect $A_i$, then $\sum_{a_j \in A_i} \Pr(A_i = a_j \mid (v_1, v_2)) = 1$.

> **Example 3.** Let us consider a gene–disease pair $(v_1, v_2)$, its set of annotated sentences $T_{(v_1, v_2)} = \{t_1, t_2, t_3, t_4\}$ and the CGE, CCS and GCI aspects. For each sentence, the candidate aspect value–score pairs are as follows:

$t_1$ : CGE (up, 0.7), CCS(progression, 0.6), GCI(notinf, 0.9)
$t_2$ : CGE (down, 0.8), CCS (regression, 0.9), GCI (causality, 0.6),
$t_3$ : CGE (notinf, 0.8), CCS (progression, 0.9), GCI(notinf, 0.9),
$t_4$ : CGE (up, 1.0), CCS (regression, 1.0), GCI (observation, 1.0).

First, sentence $t_3$ is filtered out since CGE = notinf. Hence, the sentences used for computation are $t_1$, $t_2$ and $t_4$. Then, following Definition 8, CGE value likelihoods are computed as $\Pr(\text{up}) = (0.7 + 0.0 + 1.0)/(0.7 + 0.8 + 1.0) = 0.68$ and $\Pr(\text{down}) = (0.0 + 0.8 + 0.0)/(0.7 + 0.8 + 1.0) = 0.32$, leading to the aspect–probability set $AP_{\text{CGE}} = \{(\text{up}, 0.68), (\text{down}, 0.32)\}$. Similarly, CCS and GCI lead to aspect–probability sets $AP_{\text{CCS}} = \{(\text{progression}, 0.24), (\text{regression}, 0.76),$ (notinf, 0.00)$\}$ and $AP_{\text{GCI}} = \{(\text{observation}, 0.40), (\text{causality}, 0.24), (\text{notinf}, 0.36)\}$. Thus, given the fact $(v_1, e, v_2)$ obtained from the gene–disease pair $(v_1, v_2)$, we have that $\phi(e) = (AP_{\text{CGE}}, AP_{\text{CCS}}, AP_{\text{GCI}})$ consists of the aspect–probability sets defined earlier.

For each fact, CORE combines CGE, CCS and GCI aspects into the tuple of aspect–probability sets, which represents a probability distribution over multi-aspect relationships and performs reliability tests to decide if the fact is reliable enough to populate the KB.

## Reliability testing

The facts generated through the knowledge enrichment component undergo a set of reliability tests, which are used by CORE to identify those facts that are reliable enough to populate the KB. These reliability tests are based on aspect–probability sets and follow the inference rules defined in (25, 26) and reported in Table 2 to map multi-aspect relationships to signature types. Indeed, multi-aspect relationships can be used to infer the prospective roles of genes in cancer and to classify genes into three mutually exclusive classes according to the inferred role: oncogene, tumor suppressor gene and biomarker (as in (25, 26), a gene classified as biomarker represents a gene that exhibits altered expression levels in cancer, which, however, is not (yet) identified as oncogene or tumor suppressor gene). For instance, an oncogene can be inferred from (up, progression, causality) or (down, regression, causality) multi-aspect relationships (Rules 1 and 3 of Table 2). These mutually exclusive classes represent the signature set $S$ and are associated with edges of the KB through the tagging function $\sigma(\cdot)$.

Thus, based on aspect–probability sets and inference rules, CORE performs a two-stage reliability test that first verifies that facts have sufficient evidence and then assesses the degree of contradicting evidence. The two stages are divided into sufficiency and consistency checks.

Given a fact $(v_1, e, v_2)$, a sufficiency check monitors whether the likelihood of not-informative aspect values is large enough to undermine the reliability of the fact. CORE applies the sufficiency check to CCS = notinf and GCI = notinf aspect values. Hence, a fact fails the sufficiency check and therefore is deemed 'unreliable' if $\Pr(\text{CCS} = \text{notinf}) > \alpha \vee \Pr(\text{GCI} = \text{notinf}) > \alpha$, where $\alpha$ is a fixed system threshold.

The facts that pass the sufficiency check are further inspected for consistency. Given a tuple of aspect–probability sets, associated with a fact $(v_1, e, v_2)$ through $\phi(e)$, the consistency check verifies that mutually exclusive signature types are not similarly probable.

> **Definition 9.** (signature type likelihood) Let $(v_1, e, v_2)$ be a fact and $S$ the set of mutually exclusive signature types. Then, the signature type likelihood is defined as
>
> $$\Pr(\sigma(e) = \text{type}) = \sum_{\substack{\{(a_{1j},\ldots,a_{nj})\} \text{ s.t.} \\ s((a_{1j},\ldots,a_{nj}))=\text{type}}} \prod_{i=1}^{n} \Pr(a_{ij}), \quad (3)$$

where $\Pr(a_{ij})$ is the aspect value likelihood and $\sigma(\cdot)$ and $s(\cdot)$ are the tagging and signature functions, respectively.

Since gene expression–cancer aspects can be treated as independent events (25, 26), the signature type likelihood can be computed for the gene classes. For instance, according to Rules 1 and 3 from Table 2, the likelihood of the oncogene class is

$$\Pr(\texttt{oncogene}) = \Pr(\texttt{up}) \cdot \Pr(\texttt{progression}) \cdot$$
$$\Pr(\texttt{causality}) + \Pr(\texttt{down}) \cdot$$
$$\Pr(\texttt{regression}) \cdot \Pr(\texttt{causality}).$$

Given that gene classes are mutually exclusive, the consistency check verifies whether the class likelihoods are too close to each other. Indeed, similar likelihoods imply that a fact is supported by contradictory evidence, thus showing some inconsistency. Vice versa, a large difference between likelihoods suggests a strong tendency towards a specific gene class, and therefore a more consistent support for the fact.

Hence, for a target fact $(v_1, e, v_2)$, CORE takes gene classes type-1 and type-2 with largest likelihoods and verifies that the condition $(\Pr(\texttt{type-1}) - \Pr(\texttt{type-2})) > \beta$ is satisfied, where $\beta$ is a fixed system threshold. A fact that fails the condition is therefore considered 'unreliable'. In other words, when no gene class has a likelihood large enough to overcome the others by a margin of $\beta$, CORE tags the fact as 'unreliable'. Note that the consistency check admits that only one gene class satisfies the condition.

**Example 4.** Let us consider two facts $f_1 = (v_1, e_1, v_2)$ and $f_2 = (v_3, e_2, v_4)$. The not-informative likelihoods associated with each fact are as follows:

$f_1 : \Pr(\texttt{CCS} = \texttt{notinf}) = 0.1, \Pr(\texttt{GCI} = \texttt{notinf}) = 0.3,$
$f_2 : \Pr(\texttt{CCS} = \texttt{notinf}) = 0.6, \Pr(\texttt{GCI} = \texttt{notinf}) = 0.5.$

The signature type likelihoods associated with each fact, and sorted in decreasing order of probability, are as follows:

$f_1 : \Pr(\texttt{oncogene}) = 0.7, \Pr(\texttt{tsg}) = 0.2, \Pr(\texttt{biomarker}) = 0.1,$
$f_2 : \Pr(\texttt{oncogene}) = 0.5, \Pr(\texttt{tsg}) = 0.4, \Pr(\texttt{biomarker}) = 0.1.$

Then, let us set the sufficiency threshold $\alpha$ to 0.7 and the consistency threshold $\beta$ to 0.4. In this scenario, both $f_1$ and $f_2$ pass the sufficiency check, as $\Pr(\texttt{CCS} = \texttt{notinf})$ and $\Pr(\texttt{GCI} = \texttt{notinf})$ are lower than $\alpha$ for both facts. On the other hand, only $f_1$ passes the consistency check, since none of the signature type likelihoods of $f_2$ are large enough to overcome the others by a margin of $\beta$. In this regard, for $f_1$, we have $\Pr(\texttt{oncogene}) - \Pr(\texttt{tsg}) > \beta$, which makes oncogene the candidate gene class for the fact. Conversely, for $f_2$, we have $\Pr(\texttt{oncogene}) - \Pr(\texttt{tsg}) < \beta$, which provides no candidate gene class for the fact. Therefore, $f_1$ is tagged as 'reliable' and $f_2$ as 'unreliable'.

The facts that pass both sufficiency and consistency checks are tagged as 'reliable' and used to populate the KB. Prior to population, the edges of 'reliable' facts are labeled through the tagging function $\sigma(\cdot)$ with the gene class having the highest likelihood. Note that we do not claim that gene classes are definitive. Rather, gene classes—and supporting sentences—should be treated as complementary, textual evidence that strengthens the hypotheses on the expected roles of genes in cancer obtained through experimental data.

## Active learning

The facts deemed as 'unreliable' by the reliability testing component (Module 5) are taken over by the active learning process, which ranks them by ascending reliability score and returns the top-$k$ automatically annotated sentences to domain experts for annotation.

**Definition 10.** (reliability score) Let $(v_1, e, v_2)$ be a fact, $\{A_i\}_{i=1}^l$ a subset of the aspects associated with $e$ and $S$ the set of signature types. Then, by taking a specific value $a_{ij}$ for each aspect $A_i$ of the subset, the reliability score is computed as

$$\text{rel}(e) = -\frac{\sum_{i=1}^l \Pr(a_{ij})}{l} \cdot \text{H}(S), \qquad (4)$$

where $\text{H}(S)$ is the entropy of the signature set $S$, computed as

$$\text{H}(S) = -\sum_{\texttt{type} \in S} \Pr(\sigma(e) = \texttt{type}) \cdot \log \Pr(\sigma(e) = \texttt{type}). \quad (5)$$

In this work, we compute the reliability score by considering the subset of CCS and GCI aspects and by taking their not-informative values $\{\texttt{CCS} = \texttt{notinf}, \texttt{GCI} = \texttt{notinf}\}$. Once computed, CORE uses the reliability score to perform uncertainty sampling (65). In other words, CORE ranks 'unreliable' facts by ascending order of reliability score. Then, the top-$k$ automatically annotated sentences associated with these facts are returned to domain experts for manual annotation (Module 2).

## Implementation and experiments

We use CORE to build a KB for gene expression–cancer associations. To this end, we conducted comprehensive experiments to quantify the extracted knowledge and evaluate the RE methods used to build the KB. In addition, we performed a KB reconstruction task against the state-of-the-art showing CORE effectiveness.

### Knowledge base creation

**Data processing.** We use different resources to build the KB, which increase with each subsequent iteration of the KB construction process. Table 3 reports statistics for the resources used to build each KB version. In the first iteration (KB0), we only consider manually annotated data coming from CoMAGC, OncoSearch and BioXpress. We revised these annotations to make them compliant with the annotation schema presented in Section (Manual annotation).

Then, in the second iteration (KB1), we introduce DisGeNET data, on which the CORE system deploys the RE methods. DisGeNET collects data on different 'coarse-grained' gene-disease associations from several

**Table 3.** Raw statistics for the KB versions. Rows represent the number of raw instances considered to build the KB.

| | | KB0 | KB1 | KB2 | KB3 |
|---|---|---|---|---|---|
| Manual | CoMAGC (revised) | 821 | 821 | 821 | 821 |
| | OncoSearch (revised) | 157 | 157 | 157 | 157 |
| | BioXpress (revised) | 74 | 74 | 74 | 74 |
| | DisGeNET (batch 1) | – | – | 250 | 250 |
| | DisGeNET (batch 2) | – | – | – | 249 |
| Automatic | DisGeNET (batch 1) | – | 184,859 | 184,609 | 184,609 |
| | DisGeNET (batch 2) | – | – | 184,858 | 184,609 |
| | PubMed (citing papers) | – | – | – | 2,841,096 |
| Total | | 1,052 | 185,911 | 370,769 | 3,211,865 |

**Table 4.** Partition, absolute and conditional statistics for KB.

| | | KB0 | KB1 | KB2 | KB3 |
|---|---|---|---|---|---|
| Partition | Manual | 655 | 585 | 605 | 592 |
| | Automatic | – | 96 531 | 95 282 | 435 283 |
| Absolute | Sentence | 655 | 97 116 | 95 887 | 435 875 |
| | Article | 411 | 69 462 | 65 236 | 161 449 |
| | Gene | 329 | 9,483 | 9981 | 21 005 |
| | Cancer | 98 | 1479 | 1554 | 1665 |
| | Fact | 512 | 71 554 | 89 999 | 153 016 |
| Conditional | Sentence/article | 1.59 | 1.40 | 1.47 | 2.70 |
| | Sentence/fact | 1.28 | 1.67 | 1.56 | 3.10 |
| | Article/fact | 1.09 | 1.67 | 1.56 | 2.10 |

**Table 5.** Reduction statistics for unreliable facts. For each KB version (rows), we report the number of unreliable facts present in that version that are also found in subsequent versions (columns).

| | | KB0 | KB1 | KB2 | KB3 |
|---|---|---|---|---|---|
| Insufficient | KB0 | 10 | 5 | 5 | 5 |
| | KB1 | – | 9055 | 2308 | 1135 |
| | KB2 | – | – | 4515 | 2452 |
| Inconsistent | KB0 | 22 | 18 | 15 | 17 |
| | KB1 | – | 6135 | 3837 | 3704 |
| | KB2 | – | – | 11 380 | 7786 |

**Table 6.** Signature type statistics for each KB version.

| Signature type | KB0 | KB1 | KB2 | KB3 |
|---|---|---|---|---|
| Biomarker | 390 | 59,147 | 69,409 | 105,089 |
| Oncogene | 87 | 8,833 | 13,501 | 35,520 |
| Tumor Suppressor Gene | 35 | 3,574 | 7,089 | 12,407 |

resources and covers most human diseases. Regarding gene expression–cancer associations, DisGeNET contains automatically extracted data that have been identified from the literature using text-mining techniques (28, 66, 67). For each gene–disease association, DisGeNET provides the publication(s) supporting the association, a representative sentence from each publication, the original source, as well as information on the gene and disease involved in the association. Thus, sentences within DisGeNET can be used as a high-quality starting point from which multi-aspect relationships can be extracted.

After the construction of KB1, the active learning process ranks 'unreliable' facts and returns the top-$k$ sentences for manual annotation. This new set of manually annotated sentences—together with a second batch from DisGeNET—are added to previously used data and employed to build KB2. In the last iteration (KB3), we collect from PubMed the articles citing those stored within KB2. Then, the CORE system relies on the NERD component to extract gene and cancer entities from titles and abstract sentences and deploys RE methods on them. Finally, PubMed and top-$k$ 'unreliable' sentences from KB2 are integrated into KB3 construction.

**Manual annotation.** The manual annotation process has been carried out by a clinical expert. The annotator has been given the target sentence to annotate/validate, together with the corresponding PubMed article from which it has been extracted—from either the title or the abstract.

**System parameters.** The parameters required by CORE are the sufficiency and consistency thresholds $\alpha$ and $\beta$ and the number $k$ of sentences to be returned for manual annotation during active learning. Sufficiency and consistency thresholds regulate the degree of reliability of the facts in the KB. A low sufficiency combined with a high consistency threshold leads to fewer facts but with a high level of reliability. Empirically, we set $\alpha = 0.7$ and $\beta = 0.4$. We set $k = 250$, meaning that 250 sentences are reannotated after each iteration. Note that system parameters can be adjusted as the KB size increases.

**KB statistics.** From the statistics reported in Table 4, we draw some considerations. First, we can see that the ratio between the sentences stored in the KB and the input ones decreases at each iteration. From the first iteration, CORE uses 62% of the input sentences to build KB0, and we move to 52% to build KB1, 26% for KB2 and only 14% for KB3. Such a decrease reflects the use of reliability tests and active learning, which makes the system more selective and accurate. In particular, active learning leads the system to refine the RE methods at each iteration, thus reducing false positives as well as 'unreliable' facts as shown in Table 5, which presents the reduction statistics of 'unreliable' facts. We see that the number of facts deemed as 'unreliable' in one iteration decreases in the next ones, confirming the effectiveness of active learning.

Second, the large number of different genes and cancers in KB3 highlights the scalability of the approach. In this regard, KB3 contains 21 005 genes, which cover 70% of the 30 000 estimated genes in the human genome.[2] On the other hand, through the integration of DisGeNET data, KBs 1–3 contain most of the (known) cancer types involved in gene expression–cancer associations. Together, this large number of genes and cancer types leads to more than 150 000 'reliable' facts. Table 6 presents the distribution of these facts according to the corresponding signature type.

Finally, KB3 represents one of the largest literature-derived KBs with fine-grained facts about gene expression–cancer associations. Compared to KB3, BioXpress and OncoMX—both relying on DEXTER text-mined results—contain less

---

[2]https://www.genome.gov/human-genome-project/

**Table 7.** Statistics of the aspect extraction datasets. We provide the percentage increase from one version to the next

| Aspect | Value | DS0 | DS1 | DS2 |
|---|---|---|---|---|
| CGE | up | 524 | 604 (+15%) | 679 (+12%) |
| | down | 219 | 263 (+20%) | 311 (+18%) |
| | notinf | 309 | 430 (+39%) | 550 (+28%) |
| CCS | progression | 605 | 719 (+19%) | 829 (+15%) |
| | regression | 134 | 147 (+10%) | 162 (+10%) |
| | notinf | 313 | 431 (+38%) | 549 (+27%) |
| GCI | causality | 189 | 227 (+20%) | 266 (+17%) |
| | observation | 548 | 634 (+16%) | 719 (+13%) |
| | notinf | 315 | 436 (+38%) | 555 (+27%) |
| Total | | 1,052 | 1,297 (+23%) | 1,540 (+19%) |

**Table 8.** Statistics of the sentence utility classifier dataset.

| Class | Training | Validation | Test |
|---|---|---|---|
| expression | 18,555 | 6,185 | 1,540 |
| other | 18,876 | 6,292 | 825 |

**Table 9.** Aspect extraction performances

| Dataset | Aspect | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| DS0 | CGE | 0.8812 | 0.8870 | 0.8812 | 0.8792 |
| | CCS | 0.8593 | 0.8650 | 0.8593 | 0.8600 |
| | GCI | 0.8194 | 0.8305 | 0.8194 | 0.8212 |
| DS1 | CGE | 0.8543 | 0.8574 | 0.8543 | 0.8526 |
| | CCS | 0.8404 | 0.8436 | 0.8404 | 0.8400 |
| | GCI | 0.8150 | 0.8269 | 0.8150 | 0.8142 |
| DS2 | CGE | 0.8760 | 0.8813 | 0.8760 | 0.8746 |
| | CCS | 0.8481 | 0.8515 | 0.8481 | 0.8478 |
| | GCI | 0.8266 | 0.8314 | 0.8266 | 0.8259 |

literature-derived data. Specifically, BioXpress integrates DEXTER gene expression–cancer associations for 2024 genes in lung cancer, 115 glycosyltransferases in 62 cancers and 826 microRNAs in 171 cancers (24). On the other hand, OncoMX integrates 22 904 gene expression–cancer associations between 5524 genes/microRNAs and 272 cancer types, extracted by DEXTER from 36 196 sentences in 25 860 PubMed articles. Although larger, OncoMX is still an order of magnitude smaller than KB3. Besides, both BioXpress and OncoMX only report CGE values between cancer and normal samples, thus providing less comprehensive information than CORE to model gene expression–cancer associations. A different situation occurs with OncoSearch, which contains 451 798 sentences expressing 7555 genes and 1717 cancer types, leading to 2295 oncogenes, 1549 tumor suppressor genes and 6779 biomarkers. Compared to OncoSearch, KB3 contains less sentences and cancer types. However, OncoSearch does not perform reliability tests and therefore ingests any annotated sentence. If we also consider the facts deemed as 'unreliable' by CORE when building KB3, then the number of sentences and cancer types becomes 1 037 845 and 1767, respectively. Thus, KB3 integrates a smaller number of sentences and cancer types to seek for a higher quality.

## Relation extraction evaluation

**Datasets.** We evaluate the effectiveness of the CGE, CCS and GCI extraction methods using three incremental sets of manually annotated data. Table 7 reports the statistics of these aspect extraction datasets. The first dataset (DS0) derives from the seed batch of manually annotated data used to build KB0. The second (DS1) and third (DS2) ones integrate additional data coming from the subsequent sets of 250 sentences returned by the active learning process. DS0 contains 1052 annotated sentences, which increased by 23% in DS1 and a further 19% in DS2.

Regarding the GCC extraction method, which serves as a sentence utility binary classifier, we use DisGeNET to create a large-scale semi-automatically annotated dataset. Similar to (68), we employ automatically extracted data from DisGeNET to build training and validation sets while relying on manually curated data for the test set. Table 8 reports the statistics for the sentence utility classifier dataset. For training and validation, DisGeNET sentences conveying a gene expression–cancer association were labeled as expression and those conveying any other type of association as other. For test, DS2 sentences were used as expression candidates and manually curated sentences from DisGeNET as other.

We create a unique dataset for the sentence utility classifier as the method is only applied to PubMed sentences during KB3 construction. PubMed is very general, and most of the sentences are not about gene expression–cancer associations, so the sentence utility classifier is critical for the CORE extraction process. Conversely, the sentence utility classifier is not needed on DisGeNET sentences because they are of high quality, and a filtering process has already taken place before their integration within it.

**Set-up.** For training, we set the batch size to 16 and the learning rate to 2e-5 with linear warm-up followed by linear decay (63), as suggested in (62). The CGE, CCS and GCI extraction methods perform multi-class classification and are trained using a standard cross entropy loss function. The sentence utility classifier performs binary classification and employs a binary cross-entropy loss.

We perform 10-fold cross-validation to evaluate CGE, CCS and GCI methods. For each iteration, we train the RE methods for 10 epochs, choose the best epoch on a validation set consisting of 25% of the training folds and report the corresponding results for the test fold. Instead, given the large size of the GCC extraction dataset, we train the sentence utility classifier for five epochs, pick the best epoch on the validation set and report the results on the test set.

**Results.** Table 9 reports the average performances of the CGE, CCS and GCI extraction methods on the different dataset versions.

We can see that all the three methods perform well on the task—above 0.80 for each measure—with peak performances on CGE and slightly lower performances on GCI. These results underline the differences between aspects, where CGE is most explicit in sentences—and therefore easier to extract—whereas GCI is less evident—and therefore more difficult to predict. CCS extraction is in between.

This experiment shows the effectiveness of the RE methods and their stability as they do not regress as the dataset size increases. In this regard, we recall that RE methods are retrained from scratch at each iteration and not fine-tuned with new data from the active learning process. Thus, such consistent performances across dataset versions highlight the robustness and reliability of the RE methods.

Table 10. CORE system performance on the BioXpress reconstruction task. We also report DEXTER performance on DS2.

| Dataset | Method | Accuracy | Precision | Recall | F1 |
|---------|--------|----------|-----------|--------|-----|
| BioXpress | CORE0 | 0.9544 | 0.9601 | 0.9544 | 0.9572 |
| | CORE1 | 0.9703 | 0.9831 | 0.9703 | 0.9766 |
| | CORE2 | 0.9706 | 0.9827 | 0.9706 | 0.9766 |
| DS2 | DEXTER | 0.3256 | 0.6034 | 0.3256 | 0.2882 |

Regarding GCC extraction, the sentence utility classifier achieves an accuracy of 0.8825 and a precision, a recall and an F1 value of 0.8824, 0.8825 and 0.8803, respectively. The results highlight the viability of training coarse-grained RE methods using automatically annotated data from DisGeNET (68) and show the effectiveness of the trained method on a manual test set. Thus, the sentence utility classifier is reliable enough to be used as filter on new and heterogeneous sentences gathered from PubMed.

### Knowledge base reconstruction

**Setup.** We further evaluate the effectiveness of the CORE system on a KB reconstruction task, in which we hold out a portion of an existing KB with associated sentences and we assess CORE ability to recover it. To this end, we hold out from BioXpress the set of 9636 sentences annotated by DEXTER (SoTA for gene expression–cancer annotations), and we evaluate the CORE system on them. Note that such sentences are not part of those used to train the CORE RE methods. Given that BioXpress only reports CGE values between cancer and normal samples, we restrict our evaluation to CGE extraction. As a further experiment, we also apply DEXTER to DS2 to evaluate its ability to generalize to heterogeneous sentences, whose syntactic structure can differ from its predefined patterns.

**Results.** Table 10 reports the CORE system performance on the BioXpress reconstruction task after each (re-)training of the RE methods, as well as DEXTER performance on DS2. We can see that each CORE version consistently achieves performances above 0.95 for each measure. In particular, CORE1 improves over CORE0 by 2% and reaches a performance plateau, where CORE2 also stabilizes with an accuracy of 0.9706 and a precision, a recall and an F1 value equal to 0.9827, 0.9706 and 0.9766, respectively. The results show the effectiveness of the CORE system in recovering BioXpress using a limited amount of manual annotations to train the RE methods. On the other hand, the poor performance of DEXTER on DS2 highlights a lack of flexibility that hampers its applicability to heterogeneous sentences. To further support this intuition, we observe that for DEXTER, recall presents the worst performance (0.3256) if compared to precision (0.6034). This underlines DEXTER's expert system nature based on pattern-matching, which, although precise, fails to generalize beyond its set of predefined patterns.

### Knowledge base exploration

We perform some exploratory queries to analyse the contents of the largest KB produced by CORE, that is KB3. The SPARQL queries used to explore KB3 can be found in Appendix A.

### Genes most involved in cancer diseases

Figure 3 illustrates the top ten oncogenes, biomarkers and tumor suppressor genes associated with cancer. Among the oncogenes, AKT1 emerges as the predominant gene implicated in cancer diseases within KB3. AKT1 exhibits widespread expression in various tissues (69, 70). Other known oncogenes are MAPK1 and MAPK3, frequently involved in oncogenesis, tumor progression and drug resistance (71) and STAT3 (72). Regarding biomarkers, there are several known proto-oncogenes such as ERBB2 (73), EGFR (74) and BCL2 (75). Proto-oncogenes fit our definition of biomarkers, i.e. genes that show altered expression levels in cancer but do not (yet) have enough evidence to be identified as oncogenes or tumor suppressor genes. A different situation occurs with TP53, which presents an interesting scenario as it is a biomarker and a tumor suppressor gene for many diseases. Over the years, the scientific understanding of TP53 has evolved, initially classifying it as an oncogene (76), then recognizing it as a tumor suppressor (77) and, more recently, under certain conditions, acknowledging its re-emergence as an oncogene (78). Thus, thanks to its probabilistic, fact-centric and evidence-based approach, the CORE system can capture such a dynamic scenario—which is proper for scientific discourse.

### Most discussed genes, cancer diseases and facts

Figure 4 presents the genes, diseases and facts that have garnered the most attention in the scientific literature. Naturally, the most discussed genes align with the ones most involved in cancer diseases. The most discussed topics concerning cancer predominantly revolve around breast, colorectal, prostate and lung cancer types. This outcome is fitting, as these cancer types are the four most common cancer types worldwide.[3] As a consequence, the most discussed facts pertain to gene expression–cancer associations involving the aforementioned genes and diseases.

### Longest-discussed fact in the literature

Figure 5 showcases the temporal progression of publications concerning the fact most extensively discussed in KB3 (i.e. ERBB2, BIOMARKER, mammary neoplasms). ERBB2 is a known proto-oncogene that plays an important role in human malignancies and is amplified or overexpressed in 30% of human breast cancers (73). Therefore, the relevance of ERBB2 in breast cancer well motivates its prominence within the scientific discourse.

### Search platform

The KB generated by CORE can also be accessed via COREKB (30), an intuitive and easy-to-use search platform for searching scientific facts over gene expression–cancer associations. COREKB allows users to search for gene–cancer associations and entities using free-text or structured search queries. The interface provides several features, including autocomplete facilities, entity cards summarizing the major gene–cancer relationships and entity landing pages, and users can easily switch between free-text and structured search interfaces. The system also offers a simple toggle button to
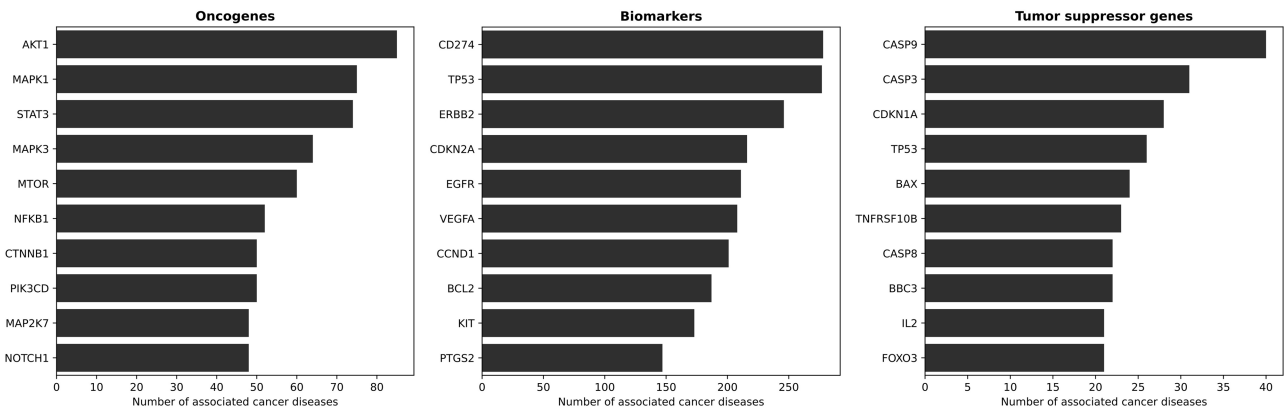
---

[3]https://www.wcrf.org/cancer-trends/worldwide-cancer-data/

**Figure 3.** The ten most involved genes (and their roles) in cancer diseases. From left to right, the figures present the ten most involved oncogenes, biomarkers and tumor suppressor genes, respectively. AKT1 is the most prominent oncogene, with wide expression in various tissues. Other known oncogenes include MAPK1, MAPK3 and STAT3. Proto-oncogenes such as ERBB2, EGFR and BCL2 show altered expression levels in cancer, but lack sufficient evidence to be identified as oncogenes, thus fitting our definition of biomarkers. TP53 represents an interesting case, as it functions as a biomarker and a tumor suppressor gene for several diseases, with its classification evolving over time.
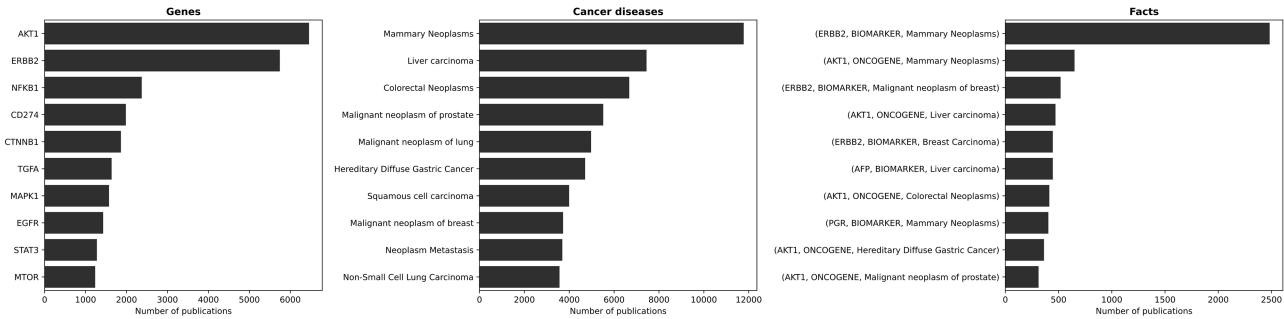


**Figure 4.** The ten most discussed genes, cancer diseases, and facts within the literature. The most discussed genes are those most involved in cancer diseases, with a focus on breast, colorectal, prostate, and lung cancer—i.e., the most common cancer types worldwide. Consequently, the most discussed facts refer to gene expression-cancer associations involving these specific genes and diseases.
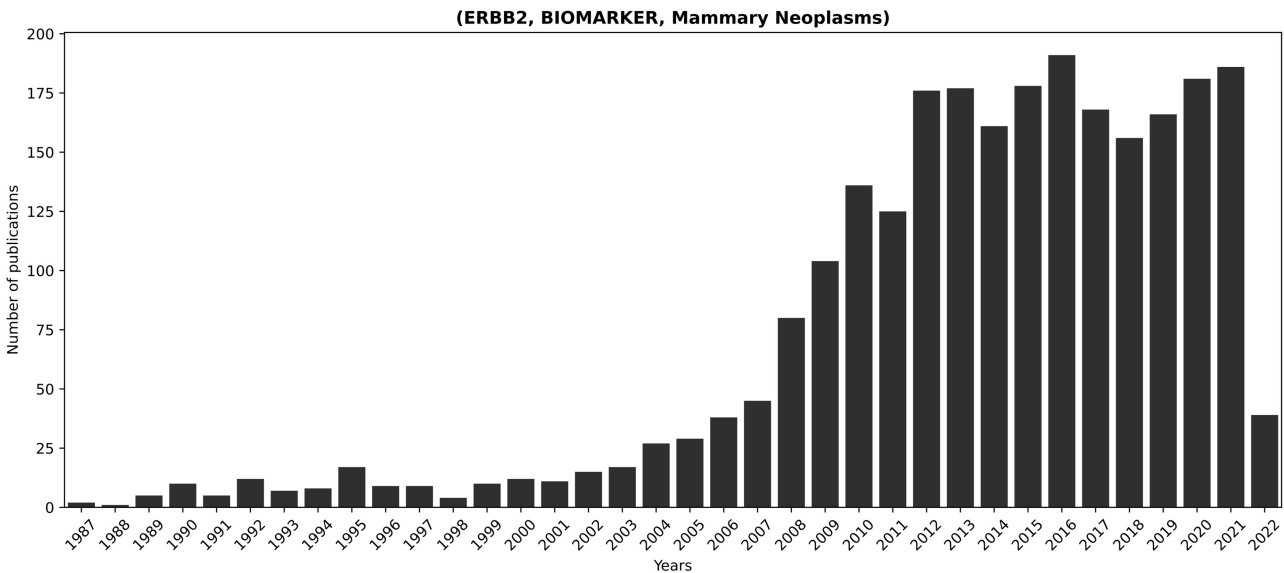


**Figure 5.** Temporal progression of publications concerning the longest-discussed fact in literature: (ERBB2, BIOMARKER, Mammary Neoplasms). ERBB2 is a known proto-oncogene, amplified or overexpressed in around 30% of human breast cancers (73). Its relevance in breast cancer justifies the prominent presence of the corresponding fact in the scientific discourse.
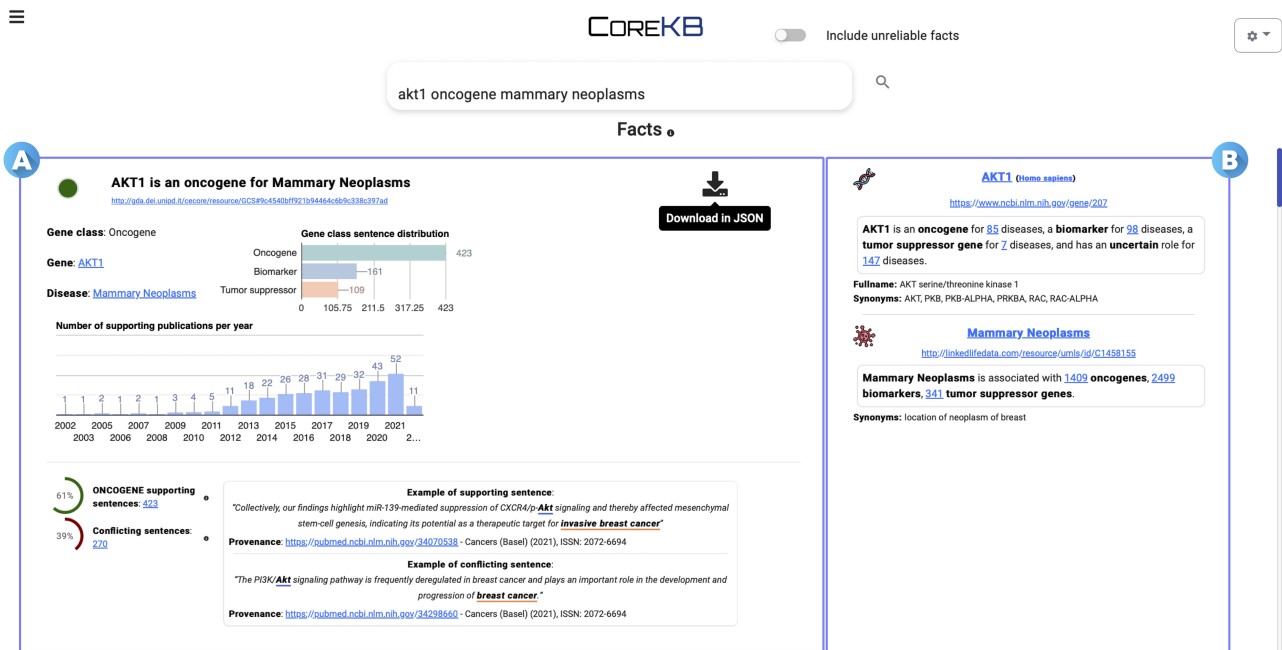
**Figure 6.** COREKB Search Engine Result Page first result for the query 'AKT1 oncogene mammary neoplasms'. The retrieved facts are organized as cards providing several information concerning (A) the gene, cancer and their relationship and (B) specific information concerning the entities—i.e. gene and the related cancer expression—involved in the association. In addition, card (A) includes infometrics and bibliometrics information to provide further insights. The contents of the cards are available for download in JSON format through the dedicated download button.

include/exclude unreliable facts from the search results. The search results are presented as a list of cards showing the information concerning the scientific facts matching the user-provided query. Card information can also be downloaded in JSON format via the dedicated download button. Figure 6 shows the first result of the Search Engine Result Page for the query 'AKT1 oncogene mammary neoplasms'.

## Architecture

COREKB's architecture consists of multiple components synergically cooperating to facilitate the search and retrieval of scientific facts—i.e. gene expression–cancer associations supported by the scientific literature. The architecture includes a web-based front–end interface built with React.js and a back-end for the business logic, Representational State Transfer Application Programming Interfaces and services built with the Python web framework Django. The system relies on a PostgreSQL database coupled with a Virtuoso Resource Description Framework triple store to memorize the KB. Moreover, Redis is exploited as an efficient in-memory data store and access broker. A search and retrieval component implemented in Python performs NERD on the user-provided queries to identify entity mentions and, in turn, perform a structured search on the database. To this aim, a Redis in-memory dictionary of entities is exploited for fast entity identification.

Specifically, when a user query is received, the system assigns a score to each entity based either on an exact match (if it occurs) or on the number of matching terms in the case of a partial match. The score is normalized based on the entity's length to avoid favoring longer entities at the expense of shorter ones. Then, the retrieved facts are ordered according to their scientific evidence support. In the case of multiple

recognized entities, the system promotes gene–cancer pairs with the most matching associations.

## Interface

The interface reports the search results by organizing them into cards; it provides information such as gene class, symbol, cancer label, supporting and conflicting sentences, associated publications, gene class distribution and bibliometrics. The fact claim is emphasized using boldface and a colored circle, indicating the informativeness and reliability of the fact—i.e. green, red and gray colors, respectively, for reliable, unreliable and non-informative facts. Moreover, the interface includes links and references to related entries in external platforms like NCBI and Linked Life Data.[4]

For each gene or cancer entity, there is a dedicated landing page that displays comprehensive information. The landing page consists of two major cards. The first card presents detailed information about the entity, for instance, in the case of a gene, it shows its symbol, full name, type, synonyms, designations, last modified date, summary and gene class distribution for different cancer diseases. Long textual information can be expanded or collapsed on click for space-saving purposes. Instead, a second card shows the sentences involving the entity of interest, presented in a tabular form with filtering and sorting features. Users can resize columns, hover over sentences for getting information via tooltips or click on sentences for a separate pop-up view.

## Conclusions and future work

In this work, we presented CORE, a KBC system based on the combination of automated ML-based methods and domain

[4]http://linkedlifedata.com

experts. CORE presents a seamless, transparent and modular architecture that can be easily modified and where different components can be replaced without affecting the others. Among its main features, the reliability tests and the active learning process make the system suited to iterative KB versioning. That is, CORE performs iterative tests that measure the reliability of the extracted data and return small, selected samples to domain experts for annotation. The high-quality data generated through active learning are then used to reinforce CORE subsequent versions. We used CORE to build one of the largest literature-derived KBs containing fine-grained facts about gene expression–cancer associations. To show the robustness of the approach, we conducted extensive experiments that highlighted the ability of CORE to scale to large collections of heterogeneous data with limited human annotations. The KB generated by CORE can be accessed via a SPARQL endpoint (http://w3id.org/corekb/sparql) or through the COREKB search platform (https://gda.dei.unipd.it).

The CORE system is an ongoing effort carried out in partnership with medical centers. The expertise and insights of clinicians have been instrumental in developing a robust KBC system. Future work aims to improve the system by integrating advanced large language models (LLMs) as input sources. Robust validation mechanisms and collaboration with experts will be crucial to identify and ingesting reliable content generated by LLMs.

## Acknowledgements

## Data availability

The KB derived by CORE is available at https://zenodo.org/record/7577127. The SPARQL endpoint for querying the KB can be found at http://w3id.org/corekb/sparql. The KB can also be accessed via COREKB at https://gda.dei.unipd.it. The source code is available at https://github.com/GDAMining/core.

## References

1. Manzoni,C., Kia,D.A., Vandrovcova,J. *et al.* (2016) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinformatics*, **19**, 286–302.
2. Borry,P., Bentzen,H.B., Budin-Ljøsne,I. *et al.* (2018) The challenges of the expanded availability of genomic information: an agenda-setting paper. *J. Community Genet.*, **9**, 103–116.
3. Neary,B., Zhou,J. and Qiu,P. (2021) Identifying gene expression patterns associated with drug-specific survival in cancer patients. *Sci. Rep.*, **11**, 1–12.
4. Dugger,S., Platt,A. and Goldstein,D. (2018) Drug development in the era of precision medicine. *Nat. Rev. Drug. Discov.*, **17**, 183–196.
5. Li,X. and Warner,J.L. (2020) A review of precision oncology knowledgebases for determining the clinical actionability of genetic variants. *Front. Cell Dev. Biol.*, **8**, 1–48.
6. Liu,F., Chen,J., Jagannatha,A., *et al.* (2016) Learning for biomedical information extraction: methodological review of recent advances. *CoRR*, **1606.07993**, 1–21. abs/1606.07993.
7. Krallinger,M., Rabal,O., Akhondi,S.A. *et al.* (2017) Overview of the BioCreative VI chemical-protein interaction track. In: *Proc. of the sixth BioCreative challenge evaluation workshop*, *Bethesda, Maryland, USA*, *October 18–20, 2017*.
8. Miranda,A., Mehryary,F., Luoma,J. *et al.* (2021) Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. In: *Proc. of the seventh BioCreative challenge evaluation workshop*, *November 8–10, 2021*, Online.
9. Weikum,G., Dong,X.L., Razniewski,S., *et al.* (2021) Machine knowledge: creation and curation of comprehensive knowledge bases. *Found. Trends Databases*, **10**, 108–490.
10. Wright,D., Gentile,A.L., Faux,N., *et al.* (2022) BioAct: biomedical knowledge base construction using active learning. *bioRxiv*.
11. Ernst,P., Siu,A. and Weikum,G. (2018) HighLife: Higher-arity fact harvesting. In: *Proc. of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*. ACM, Lyon, France. pp. 1013–1022, 23–27 April 2018.
12. Mintz,M., Bills,S., Snow,R., *et al.* (2009) Distant supervision for relation extraction without labeled data. In: *Proc. of the 47th Annual Meeting of the Association for Computational Linguistics (ACL 2009) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. ACL, Singapore, pp. 1003–1011, 2–7 August 2009.
13. Surdeanu,M., Tibshirani,J., Nallapati,R., *et al.* (2012) Multi-instance multi-label learning for relation extraction. In: *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012*. ACL, Jeju Island, Korea. pp. 455–465, 12–14 July 2012.
14. Settles,B. (1995) Active learning literature survey. *Science*, **10**, 237–304.
15. Olsson,F. (2009) A literature survey of active machine learning in the context of natural language processing. *SICS Technical Report*, Swedish Institute of Computer Science, **1**, 1–59.
16. Angeli,G., Tibshirani,J., Wu,J., *et al.* (2014) Combining distant and partial supervision for relation extraction. In: *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*. ACL, Doha, Qatar. pp. 1556–1567, 25–29 October 2014.
17. Sterckx,L., Demeester,T., Deleu,J., *et al.* (2014) Using active learning and semantic clustering for noise reduction in distant supervision. In: *Proc. of the 4th Workshop on Automated Base Construction at NIPS 2014 (AKBC-2014)*, *Montreal, Canada*, *December 13, 2014*, pp. 1–6.
18. Park,S.J., Yoon,B.H., Kim,S.K., *et al.* (2019) GENT2: an updated gene expression database for normal and tumor tissues. *BMC Medical Genom.*, **12**, 1–8.
19. Shaul,Y.D., Yuan,B., Thiru,P. *et al.* (2016) MERAV: a tool for comparing gene expression across human tissues and cell types. *Nucleic Acids Res.*, **44**, 560–566.
20. Zhang,J., Baran,J., Cros,A. *et al.* (2011) International cancer genome consortium data portal - a one-stop shop for cancer genomics data. *Database J. Biol. Databases Curation*, **2011**, bar026.
21. Weinstein,J.N., Collisson,E.A., Mills,G.B. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.*, **45**, 1113–1120.
22. Dingerdissen,H., Torcivia-Rodriguez,J., Hu,Y. *et al.* (2018) BioMuta and BioXpress: mutation and expression knowledge-bases for cancer biomarker discovery. *Nucleic Acids Res.*, **46**, D1128–D1136.
23. Dingerdissen,H.M., Bastian,F., Vijay-Shanker,K. *et al.* (2020) OncoMX: a knowledgebase for exploring cancer biomarkers in the

context of related cancer and healthy data. *JCO Clin. Cancer Inform.*, **4**, 210–220.

24. Gupta,S., Dingerdissen,H., Ross,K.E. *et al*. (2018) DEXTER: disease-expression relation extraction from text. *Database J. Biol. Databases Curation*, **2018**, bay045.

25. Lee,H.J., Shim,S.H., Song,M.R. *et al*. (2013) CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations. *BMC Bioinform.*, **14**, 323.

26. Lee,H.J., Cuong Dang,T., Lee,H., *et al*. (2014) OncoSearch: cancer gene search engine with literature evidence. *Nucleic Acids Res.*, **42**, 416–421.

27. Piñero González,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J. *et al*. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.

28. Bundschus,M., Bauer-Mehren,A., Tresp,V. *et al*. (2010) Digging for knowledge with information extraction: a case study on human gene-disease associations. In: *Proc. of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010*. ACM, Toronto, Ontario, Canada. pp. 1845–1848, 26–30 October 2010.

29. Marchesin,S., Menotti,L., Silvello,G. *et al*. (2023) CORE: gene expression-cancer knowledge base. January. https://zenodo.org/record/7577127.

30. Giachelle,F., Marchesin,S., Silvello,G., *et al*. (2023) Searching for reliable facts over a medical knowledge base. In: *Proc. of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023*. ACM, Taipei, Taiwan. 23–27 July 2023.

31. Ciardiello,F., Arnold,D., Casali,P.G. *et al*. (2014) Delivering precision medicine in oncology today and in future—the promise and challenges of personalised cancer medicine: a position paper by the European Society for Medical Oncology (ESMO). *Ann. Oncol.*, **25**, 1673–1678.

32. Nalejska,E., Mączyńska,E. and Lewandowska,M.A. (2014) Prognostic and predictive biomarkers: tools in personalized oncology. *Mol Diagn Ther.*, **18**, 273–284.

33. Fabian,M.R., Sonenberg,N. and Filipowicz,W. (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.*, **79**, 351–379.

34. Blenkiron,C. and Miska,E.A. (2007) miRNAs in cancer: approaches, aetiology, diagnostics and therapy. *Hum. Mol. Genet.*, **16**, R106–R113.

35. Lehmann,J., Isele,R., Jakob,M. *et al*. (2015) Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, **6**, 167–195.

36. Bollacker,K.D., Evans,C., Paritosh,P.K. *et al*. (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*. ACM, Vancouver, BC, Canada. pp. 1247–1250, 10–12 June 2008.

37. Suchanek,F.M., Kasneci,G. and Weikum,G. (2007) Yago: a core of semantic knowledge. In: *Proc. of the 16th International Conference on World Wide Web, WWW 2007*. ACM, Banff, Alberta, Canada. pp. 697–706, 8–12 May 2007.

38. Vrandecic,D. and Krötzsch,M. (2014) Wikidata: a free collaborative knowledgebase. *Commun. ACM*, **57**, 78–85.

39. Doan,A., Halevy,A.Y. and Ives,Z.G. (2012) *Principles of Data Integration*. Morgan Kaufmann: Burlington, Massachusetts.

40. Lenzerini,M. (2002) Data integration: a theoretical perspective. In: *Proc. of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM, *Madison, Wisconsin, USA*. pp. 233–246, 3–5 June.

41. Ilyas,I.F. and Chu,X. (2019) *Data Cleaning*. ACM: New York, USA.

42. Li,J., Sun,A., Han,J., *et al*. (2022) A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, **34**, 50–70.

43. Sevgili,O., Shelmanov,A., Arkhipov,M.Y. *et al*. (2022) Neural entity linking: a survey of models based on deep learning. *Semantic Web*, **13**, 527–570.

44. Shen,W., Li,Y., Liu,Y. *et al*. (2023) Entity linking meets deep learning: techniques and solutions. *IEEE Trans. Knowl. Data Eng.*, **35**, 2556–2578.

45. Kumar,S. (2017) A survey of deep learning methods for relation extraction. *CoRR*, **1705.03645**, 1–8. abs/1705.03645.

46. Smirnova,A. and Cudré-Mauroux,P. (2019) Relation extraction using distant supervision: a survey. *ACM Comput. Surv.*, **51**, 106:1–106:35.

47. Shi,P. and Lin,J. (2019) Simple BERT models for relation extraction and semantic role labeling. *CoRR*, **1904.05255**, 1–6. abs/1904.05255.

48. Ren,P., Xiao,Y., Chang,X. *et al*. (2022) A survey of deep active learning. *ACM Comput. Surv.*, **54**, 180:1–180:40.

49. Hogan,A., Blomqvist,E., Cochez,M. *et al*. (2021) Knowledge graphs. *ACM Comput. Surv.*, **54**, 71:1–71:37.

50. Deshpande,O., Lamba,D.S., Tourn,M. *et al*. (2013) Building, maintaining, and using knowledge bases: a report from the trenches. In: *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013*. ACM, New York, NY, USA. pp. 1209–1220, 22–27 June 2013.

51. Dong,X., Gabrilovich,E., Heitz,G. *et al*. (2014) Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*. ACM, New York, NY, USA. pp. 601–610. 24–27 August 2014.

52. Grubb,T., Andersen,B. and Alonso,O. (2022) On reliability scores for knowledge graphs. In: *Proc. of the Web Conference 2022 (WWW '22 Companion)*. ACM, Lyon, France. pp. 1–5, 25–29 April 2022.

53. Ilyas,I.F., Rekatsinas,T., Konda,V. *et al*. (2022) Saga: a platform for continuous construction and serving of knowledge at scale. In: *SIGMOD '22: International Conference on Management of Data*. ACM, Philadelphia, PA, USA. pp. 2259–2272, 12–17 June 2022.

54. Fridman Noy,N., Gao,Y., Jain,A. *et al*. (2019) Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM*, **62**, 36–43.

55. Haber,D. and Settleman,J. (2007) Cancer: drivers and passengers. *Nature*, **446**, 145–146.

56. Maglott,D.R., Ostell,J., Pruitt,K.D., *et al*. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, 52–57.

57. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, 267–270.

58. Wei,C.H., Kao,H.Y. and Lu,Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.*, **41**, 518–522.

59. Wei,C.H., Leaman,R. and Lu,Z. (2016) Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinform.*, **32**, 1907–1910.

60. Wei,C.H., Allot,A., Leaman,R., *et al*. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.*, **47**, W587–W593.

61. Lipscomb,C.E. (2000) Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, **88**, 265.

62. Beltagy,I., Lo,K. and Cohan,A. (2019) SciBERT: a pretrained language model for scientific text. In: *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*. ACL, Hong Kong, China. pp. 3613–3618, 3–7 November 2019.

63. Devlin,J., Chang,M.W., Lee,K., *et al*. (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*. ACL, Minneapolis, MN,

USA, Vol. 1 (Long and Short Papers), pp. 4171–4186, 2–7 June 2019.

64. Ammar,W., Groeneveld,D., Bhagavatula,C. *et al.* (2018) Construction of the Literature Graph in Semantic Scholar. In: *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018.* ACL, New Orleans, Louisiana, USA, Vol. 3 (Industry Papers), pp. 84–91, 1–6 June 2018.

65. Lewis,D.D. and Catlett,J. (1994) Heterogeneous uncertainty sampling for supervised learning. In: *Proc. of the Eleventh International Conference on Machine Learning (ICML).* Morgan Kaufmann, New Brunswick, NJ, USA. pp. 148–156, 10–13 July 1994.

66. Bravo,A., Piñero González,J., Queralt-Rosinach,N. *et al.* (2015) Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinform.*, **16**, 55:1–55:17.

67. Bundschus,M., Dejori,M., Stetter,M. *et al.* (2008) Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinform.*, **9**, 207.

68. Marchesin,S. and Silvello,G. (2022) TBGA: a large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinform.*, **23**, 111.

69. Testa,J.R. and Bellacosa,A. (2001) AKT plays a central role in tumorigenesis. *Proc. of the National Academy of Sciences*, **98**, 10983–10985.

70. Cohen,M.M. (2013) The AKT genes and their roles in various disorders. *American Journal of Medical Genetics Part A*, **161**, 2931–2937.

71. Braicu,C., Buse,M., Busuioc,C. *et al.* (2019) A comprehensive review on MAPK: a promising therapeutic target in Cancer. *Cancers*, **11**, 2019.

72. Bromberg,J.F., Wrzeszczynska,M.H., Devgan,G. *et al.* (1999) Stat3 as an Oncogene. *Cell*, **98**, 295–303.

73. Slamon,D.J., Godolphin,W., Jones,L.A. *et al.* (1989) Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*, **244**, 707–712.

74. Velu,T.J., Beguinot,L., Vass,W.C. *et al.* (1987) Epidermal-growth-factor-dependent transformation by a human EGF receptor proto-oncogene. *Science*, **238**, 1408–1410.

75. Kroemer,G. (1997) The proto-oncogene Bcl-2 and its role in regulating apoptosis. *Nat. Med.*, **3**, 614–620.

76. Eliyahu,D., Raz,A., Gruss,P. *et al.* (1984) Participation of p53 cellular tumour antigen in transformation of normal embryonic cells. *Nature*, **312**, 646–649.

77. Baker,S.J., Fearon,E.R., Nigro,J.M. *et al.* (1989) Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science*, **244**, 217–221.

78. Soussi,T. and Wiman,K.G. (2015) TP53: an oncogene in disguise. *Cell Death & Differentiation*, **22**, 1239–1249.

# Appendix SPARQL queries

Queries A1–A3 can be used to find the most involved genes (and their roles) in cancer diseases. Queries A4–A6 can be used to identify the most discussed genes, cancer diseases and facts. Query A7 can be used to find out the fact most extensively discussed in the literature. The code to cut and paste the queries in the SPARQL endpoint without syntax issues is available here: https://github.com/GDAMining/core/blob/main/sparql/example-queries.txt.

---

**Querie A1.** Find the ten most involved oncogenes in cancer diseases.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
SELECT ?gene (COUNT(?fact) AS ?numFacts)
WHERE
{
    ?fact core: expressedBy ?gene;
          core: hasType ``ONCOGENE''^^xsd:
              string.
}
GROUP BY ?gene
ORDER BY DESC(?numFacts)
LIMIT 10
```

---

**Querie A2.** Find the ten most involved biomarkers in cancer diseases.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
SELECT ?gene (COUNT(?fact) AS ?numFacts)
WHERE
{
    ?fact core: expressedBy ?gene;
          core: hasType ``BIOMARKER''^^ xsd:
              string.
}
GROUP BY ?gene
ORDER BY DESC(?numFacts)
LIMIT 10
```

**Querie A3.**  Find the ten most involved tumor suppressor genes in cancer diseases.

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
SELECT ?gene (COUNT(?fact) AS ?numFacts)
WHERE
{
    ?fact core: expressedBy ?gene;
          core: hasType ``TSG''^^ xsd:string.
}
GROUP BY ?gene
ORDER BY DESC(?numFacts)
LIMIT 10
```

**Querie A4.**  Find the ten most discussed genes in the literature.

```
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
SELECT ?gene (COUNT(DISTINCT ?article) AS
       ?numArticles)
WHERE
{
    ?fact core: expressedBy ?gene;
          core: supportedBy ?evidence.

    ?evidence core: extractedFrom ?article.
}
GROUP BY ?gene
ORDER BY DESC(?numArticles)
LIMIT 10
```

**Querie A5.**  Find the ten most discussed cancer diseases in the literature.

```
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
PREFIX umls: <http://linkedlifedata.com/
             resource/umls/id/>

SELECT ?disease (COUNT(DISTINCT ?article) AS
       ?numArticles)
WHERE
{
    ?fact core: involves ?disease;
          core: supportedBy ?evidence.

    ?evidence core: extractedFrom ?article.
    FILTER (?disease NOT IN (umls:C0086692,
            umls:C0027651, umls:C0006826,
            umls:C1306459))
}
GROUP BY ?disease
ORDER BY DESC(?numArticles)
LIMIT 10
```

**Querie A6.**  Find the ten most discussed facts in the literature.

```
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
PREFIX umls: <http://linkedlifedata.com/
             resource/umls/id/>
SELECT ?disease (COUNT(DISTINCT ?article) AS
       ?numArticles)
WHERE
{
    ?fact core: involves ?disease;
          core: supportedBy ?evidence.

    ?evidence core: extractedFrom ?article.
    FILTER (?disease NOT IN (umls:C0086692,
            umls:C0027651, umls:C0006826,
            umls:C1306459))
}
GROUP BY ?disease
ORDER BY DESC(?numArticles)
LIMIT 10
```

**Querie A7.**  Find the longest-discussed fact in the literature.

```
PREFIX core: <http://gda.dei.unipd.it/cecore/
             ontology/>
SELECT ?gene ?relation ?disease ?pubDate
       (COUNT(DISTINCT ?article) AS
       ?numArticles)
WHERE
{
   ?fact core: expressedBy ?gene;
         core: hasType ?relation;
         core: involves ?disease;
         core: supportedBy ?evidence.

   ?evidence core: extractedFrom ?article.
   ?article core: publicationYear ?pubDate.

   {
     SELECT ?fact (COUNT(DISTINCT ?pubDate) AS
            ?numYears)
     WHERE
     {
     ?fact core: supportedBy ?evidence.

     ?evidence core: extractedFrom ?article.
     ?article core: publicationYear ?pubDate.
     }
     GROUP BY ?fact
     ORDER BY DESC(?numYears)
     LIMIT 1
   }
}
GROUP BY ?gene ?relation ?disease ?pubDate
ORDER BY ASC(?pubDate)
```