D3.2

Adaptation of MelanomaMine and LiMTox to the analysis of paediatric cancers and application to biomedical publications on paediatric cancers

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D3.2 / 1.0
Work package contributing to the deliverable	WP3
Due date	May, 2021 – M29
Actual submission date	31st May, 2021

Responsible organisation	Barcelona Supercomputing Center (BSC)
Editor	Davide Cirillo
Dissemination level	PU
Revision	1.0

Abstract	We report on the implementation of a text mining workflow for the extraction of relevant biomedical information from publicly available free text and its network representation.
Keywords	Text mining, named entity recognition, word embeddings, network biology

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Davide Cirillo (BSC)

Contributors (ordered according to beneficiary numbers)

Matteo Manica (IBM)

Salvador Capella-Gutierrez (BSC)

Alfonso Valencia (BSC)

Martin Krallinger (BSC)

Javier Omar Corvi (BSC)

José María Fernández (BSC)

Alejandro Canosa (BSC)

Elena De La Calle (BSC)

Joao Pita Costa (XLAB)

Jolanda Modic (XLAB)

Disclaimer

The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The content of this document reflects only the author`s view – the European Commission is not responsible for any use that may be made of the information it contains. The users use the information at their sole risk and liability.

Executive Summary

Most results in biomedical research are published as part of unstructured natural language texts. Traditionally, researchers access these sources through bibliographic databases, such as MEDLINE of the National Library of Medicine, collecting abstracts and publications, and their corresponding search engines, such as PubMed. The fast growing number of publications (MEDLINE contains more than 26 million[1] references to journal articles in biomedicine) makes the task of information retrieval an extremely tedious and complex endeavour, which requires the use of Natural Language Processing (NLP) and, in particular, text mining approaches.

This document reports on the implementation of a text mining workflow for extracting biomedical information from large volumes of paediatric cancer-related abstracts in PubMed. The workflow builds on the general framework of two text mining tools that have been previously developed at BSC: LimTox and MelanomaMine. We adapted core common components of LimTox and MelanomaMine to process PubMed abstracts and generate paediatric cancer-related corpora as well as key bio-entities that can be used to create relational graphs or networks. Additionally, some functionalities of the iPC text mining workflow, namely the generation of word embeddings, can be used by complementary NLP approaches, such as INtERAcT, developed by IBM, to infer molecular associations.

The document reports on the implementation of the iPC text mining workflow and three use cases. The relevance of text mining in paediatric oncology and the motivation of the objectives of D3.2 are exposed in the Chapter 1 (Introduction). MelanomaMine and LimTox tools and their use in projects related to and works supported by iPC are presented in Chapter 2 **and Chapter 3**. Chapter 4 is devoted to the description of the specific components of the iPC text mining workflow and the comparison with MelanomaMine and LimTox. Finally, **Chapter 5** shows applications of the iPC text mining workflow to paediatric oncology and future work, and **Chapter 6** offers a summary and the conclusions of the document.

All the files used in the analyses presented in sections 5.2 and 5.2 as well as the draft version of the research paper presented in section 2.2 are available at the iPC Nextcloud repository: https://data.ipc-project.bsc.es/s/tgYEkTC7bECWkor.

[1] https://www.nlm.nih.gov/bsd/medline_lang_distr.html

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Text mining in cancer research

Clinical research and practice in oncology and related fields generate large volumes of texts detailing patient descriptors, such as symptoms, diagnosis and treatment outcomes, as well as experimental findings and therapeutic strategies. Knowledge extraction from textual sources of biomedical information, including clinical documentation (e.g., notes from electronic health records and pathology reports) as well as scientific literature (e.g., scholarly articles published in scientific journals [1]), has a great potential in biomedicine, especially in the area of cancer research [2]. Nevertheless, even with several successful applications, such as adverse drug event surveillance [3] and literature-based discovery [4], text mined information is recognized as an underused source of knowledge for improved cancer care [5] mainly due to the difficulties associated with the domain-specific text processing.

Indeed, despite being rich in information, texts in free forms are not easy to process automatically. Moreover, terminology and language expressions in medicine and molecular biology make the analysis of biomedical unstructured data a notoriously difficult domain for Natural Language Processing (NLP) and text mining. In these areas, relevant sources comprise heterogeneous document types from various highly specialized subdomains with a highly ambiguous language, characterized by acronyms, abbreviations and ever-changing technical terms. Substantial progress has been made in the extraction of bio-entity mentions (genes, proteins, drugs, diseases, and others) and their relations, as evaluated in the bi-annual challenge BioCreative that BSC organizes (http://www.biocrative.org).

The content of a text needs to be made accessible for statistical analysis and modeling by rendering a more structured representation of it through information extraction (IE). IE consists in automatically extracting structured information from unstructured sources. NLP methods and, in particular, text mining techniques are fundamental tools to achieve precise IE and make the unstructured content of texts accessible for further analysis and learning tasks. iPC partners BSC, IBM and XLAB recently offered an in-depth discussion on the importance of text mining for knowledge retrieval in cancer research, with particular emphasis on paediatric oncology, in a blog posted on the iPC web page titled “Mining biomedical text to find insight that can save lives” [2].

In this document, we report on the development of a unified text mining workflow for IE in paediatric oncology. The workflow is based on the backbone of existing text mining tools that we previously designed, called MelanomaMine and LimTox, dedicated to distinct branches of cancer research, namely melanoma research and liver toxicology, respectively. The iPC text mining workflow facilitates the generation of network representation of the text mined information for further utilization in other iPC work packages, specifically WP4 (“Network-based meta-analysis of multi-omics and text-mined data”).

1.2 MelanomaMine and LimTox: a blueprint for text mining in cancer research

MelanomaMine and LimTox are two text mining applications that have been developed at BSC in collaboration with the National Centre for Oncological Research (CNIO) in Madrid, Spain.

MelanomaMine is a text mining application and database dedicated to the processing of melanoma related biomedical literature and knowledge resources. MelanomaMine uses IE and machine learning approaches to score and classify textual data and applies Named Entity Recognition (NER)

[2] https://ipc-project.eu/mining-biomedical-text-to-find-insight-that-can-save-lives/

methods to detect bio-entities of relevance to understand the molecular basis of melanoma, focusing on genes, proteins, mutations and chemicals/drugs.

LimTox [6] is a system to extract associations between compounds and liver toxicological endpoints, facilitating the establishment of toxicity thresholds of several substances. LimTox is intended to serve basic researchers and medical personnel as a domain-specific search and retrieval system to evaluate the relevance of a particular bio-entity to hepatotoxicity.

Beside specific functionalities required for their distinct domain of application, both methods share common processes. They retrieve associations among genes, proteins, mutations, chemicals/drugs through NER and use machine learning approaches, namely Support Vector Machines (SVMs), to classify documents. Additionally, both tools provide interfaces that enable heterogeneous search types, including general free text search and entity specific search options. The information is displayed in friendly user platforms that allow both specialists and nonspecialists visualizing and exploring the results.

MelanomaMine and LiMTox have been originally developed for melanoma and hepatotoxicity, but the text mining foundations of their internal design make those tools, or derived adaptations, particularly amenable for cancer research and, specifically, paediatric oncology.

1.3 Adaptation motivation

The efficient retrieval and processing of biomedical information from textual sources represents an important step to better characterize the molecular entities at play in paediatric cancers. The need to overcome the limited domain-specific applications of dedicated text mining tools represents the main motivation for the adaptation of MelanomaMine and LimTox to IE for paediatric oncology. Additionally, the interoperability between the outcomes of the iPC text mining workflow and state-of-the-art NLP tools to infer specific molecular associations, such as INtERAcT [7], developed by IBM, is highly desirable. Moreover, the possibility of generating network representations of the text mined information allows leveraging this knowledge in the computational solutions implemented in further iPC work packages, specifically WP4 (“Network-based meta-analysis of multi-omics and text-mined data”).

A project focused on the implementation of an initial text mining approach for pediatric oncology was awarded the “José Castillejo” mobility grant funded by the Spanish Ministry of Education in 2019-2020, which allowed iPC partners BSC and IBM to actively collaborate on this matter. The project led to the development of a prototype text mining tool for the content analytics of scientific abstract of biomedical publications referring to distinct childhood cancers. A short summary of this collaborative experience is described in a blog post that BSC has published on the iPC web page titled “Automatic extraction of biomedical knowledge from scientific publications” [3].

The iPC text mining workflow, which stemmed from this initial prototype, enables the automatic extraction of mentions of bio-entities (genes, proteins, drugs, diseases and others) and their associations, allowing to unify and interpret biomedical information of text sources, to further analyze it with INtERAcT, and to link it with the multi-omics datasets available within the iPC consortium (Figure 1).

[3] https://ipc-project.eu/automatic-extraction-of-biomedical-knowledge-from-scientific-publications/

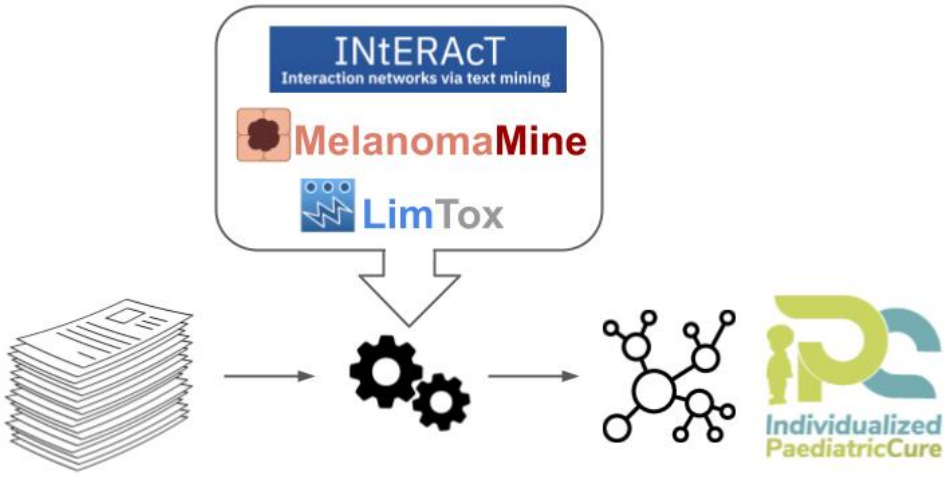

Figure 1: The use of text mining in the iPC project. The adaptation of core components of MelanomaMine and LimTox as well as the interoperability with INtERAcT allow generating network representations of information mined from paediatric cancer-specific texts that can be leveraged by the iPC consortium.

Chapter 2 MelanomaMine

2.1 MelanomaMine description

MelanomaMine is a text mining application dedicated to the processing of melanoma related biomedical literature and knowledge resources. The tool has been created in collaboration with the National Centre for Oncological Research (CNIO) in Madrid, Spain. A web service of MelanomaMine is visible at http://melanomamine.bioinfo.cnio.es/. The code of the MelanomaMine application is available at the public GitHub repository https://github.com/cirillodavide/melanomamine.

MelanomaMine uses IE and machine learning to score and classify textual data based on melanoma relevance detected by a text classifier, namely a Support Vector Machines (SVMs) with a linear kernel. The SVM model was trained on the content of a curated set of 4580 melanoma-related PubMed abstracts and 4580 randomly selected abstracts. As a testing set, abstracts from the Melanoma Molecular Map Project (MMMP, www.mmmp.org), were used, achieving high performances (0.95 Area Under the ROC curve in a five-fold cross-validation).

Along with document classification capabilities, MelanomaMine integrates NER tools for bio-entities based on open-source software[4], namely tmChem for chemical names, GNormPlus for gene/protein names, tmVar for sequence variants at protein and gene levels, and DNorm for disease names. A description of the MelanomaMine pipeline is presented in Figure 2.

MelanomaMine is intended as a domain-specific search and retrieval system to be used for both basic researchers as well as for medical personnel who want to know how relevant a particular gene or drug or mutation is to melanoma. More information about MelanomaMine implementation, the SVM training and performance evaluation can be found in the Supplemental Information file of the draft version of a research paper (see section 2.2) that is available at the iPC Nextcloud repository: https://data.ipc-project.bsc.es/s/tgYEkTC7bECWkor.

[4] https://www.ncbi.nlm.nih.gov/research/bionlp/tools/

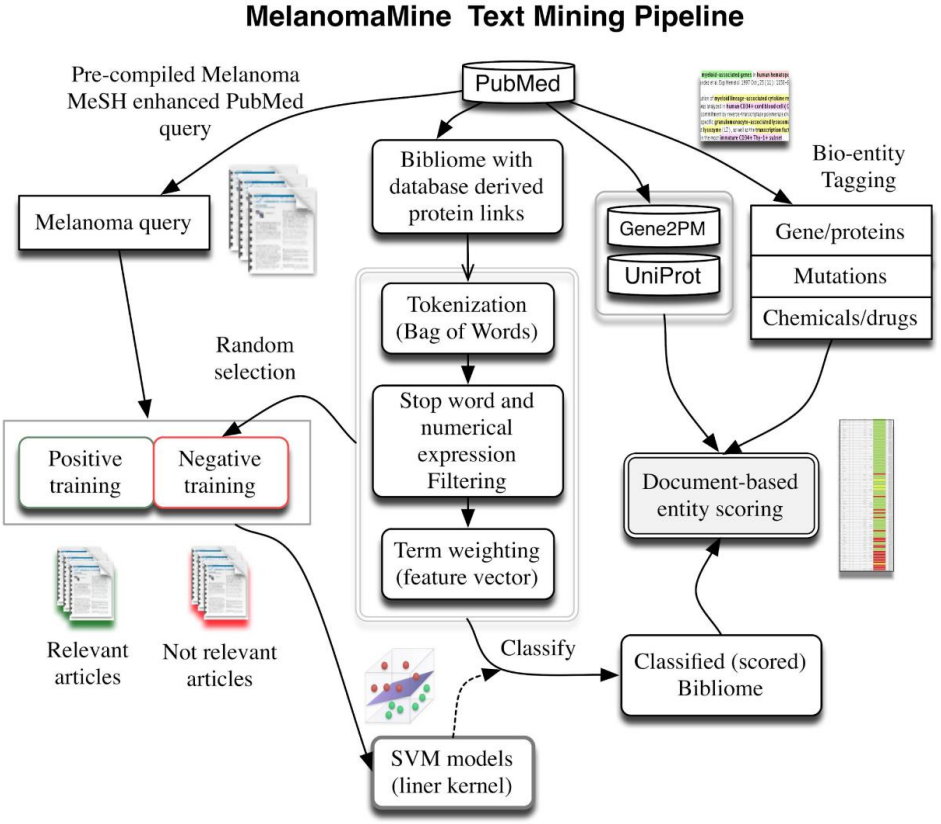

Figure 2: Flow chart of the MelanomaMine system pipeline. The various tasks that are part of the MelanomaMine processing pipeline are shown, from the initial document preprocessing to the detection of chemical entities to PubMed abstract scoring.

2.2 Research outcomes using MelanomaMine

BSC has employed MelanomaMine in a recent work on melanoma risk assessment supported by iPC. A draft version of the research paper, which is an invited contribution to be peer-reviewed for a special issue of the journal Cancers (ISSN 2072-6694), is available at the iPC Nextcloud repository: https://data.ipc-project.bsc.es/s/tgYEkTC7bECWkor

In this work, we use MelanomaMine to identify and analyze melanoma-specific scientific literature and discover relevant information that can accelerate biomedical discovery. Melanoma is the most deadly skin cancer, for which early detection and Precision Medicine interventions are crucial to survival. Several loci for susceptibility to various forms of Cutaneous malignant melanoma (CMM) have been mapped to distinct chromosomes, and somatic mutations have been identified in distinct genes. Despite recent efforts to identify other high-risk melanoma susceptibility genes, around 77% of melanoma-prone families present unknown susceptibility factors involved in the disease. In this view, it is of utmost importance to identify genes involved in unsuspected biological processes with potential implications in early detection and personalized treatment.

We applied MelanomaMine to score more than 16 milions PubMed abstracts and created a melanoma-specific knowledge corpus by selecting the top-ranking documents. This corpus of melanoma-specific abstracts was subsequently processed with IBM Watson Explorer (WEX), a general-purpose cognitive computing software that allows performing NLP analytics on huge volumes of unstructured data. In particular, we analyzed the content of the melanoma-specific knowledge corpus with WEX using healthcare-specific dictionaries obtained from the IBM Text Analytics Catalog (TAC). In this way, we generated a so-called Content Analytics graph, that is a weighted network of relevant keywords. We analyzed the Content Analytics graph to extract

actionable insights for melanoma healthcare and showed the potential of coupling text mining and cognitive computing for cancer research (Figure 3).

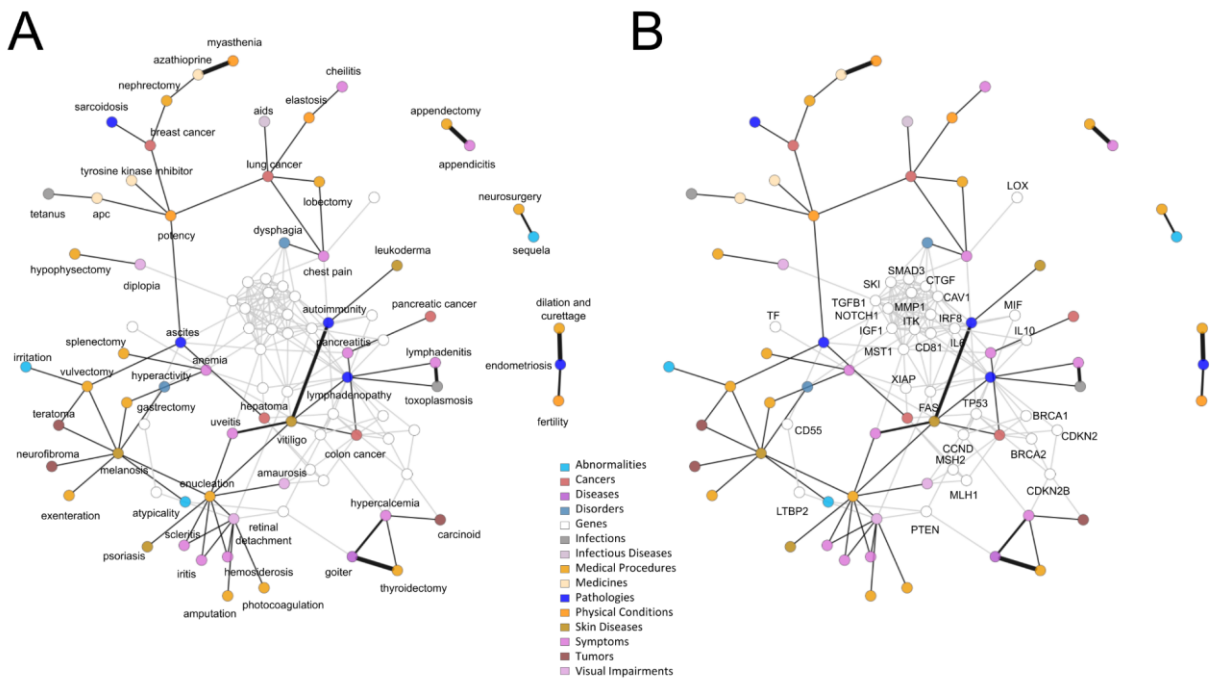

Figure 3: Content Analytics graph generated by WEX using TAC healthcare dictionaries and the melanoma-specific corpus distilled by MelanomaMine. Edges width is proportional to the WEX score. (A) TAC keywords associations; (B) genes from protein-protein interactions and gene-phenotype associations.

MelanomaMine proves to be an indispensable tool that permits both to pinpoint distinct melanoma-related bio-entities and to use them to provide WEX with melanoma-specific information. Our results shed light on distinct aspects of melanoma physiopathology. In particular, we corroborate that relevant gene alterations in melanoma are involved in cellular signaling and genome stability, and reveal new biological processes involved in pathology, indicating promising and uncharted avenues for biomedical research in skin malignancies.

Considering the little amount of existing text mining solutions dedicated to melanoma-related problems, MelanomaMine represents a comprehensive and user-friendly tool for domain-specific text mining. Indeed, text mining applications dedicated to melanoma research comprise custom applications, common procedures adapted to specific tasks, and functionalities embedded in general-purpose platforms.

Our results show how computational frameworks, designed to homogenize and contextualize biomedical information prior to downstream analyses, can be effectively used to allow specialized systems, such as WEX or INtERacT (see section 5.3), delivering valuable observations for biomedical discovery. This general paradigm has been strongly promoted by IBM developers since the inception of the Watson technology [8]. As a matter of fact, all individual IBM Watson commercial applications (e.g. the conversational agent Watson Assistant for Hospitality, designed for hotel guests, or Watson Personality Insights, designed for predicting consumption preferences) are implemented for domain-specific tasks, integrating additional machine learning components (e.g. tone analyzer or sentiment analysis) to the NLP hallmark features of the original Watson technology.

Chapter 3 LimTox

3.1 LimTox description

LimTox [6] (http://limtox.bioinfo.cnio.es/) is a web-based online biomedical search tool. It uses several sources, including adverse liver events, scientific abstracts, full text articles, and medical agency assessment reports. LimTox implements a biomedical text mining pipeline, which performs distinct tasks, such as NER and IE, using an array of methodological approaches, namely machine learning, rule- and pattern-based techniques, and term lookup strategies. LimTox search is specialized to distinct queries, including chemical compounds or drugs, genes, and biochemical liver markers. Although it is specific for liver toxicity, it can be expanded to other organs to provide insights regarding nephrotoxicity, cardiotoxicity, thyrotoxicity, and phospholipidosis. LimTox has been developed in the context of the eTOX project (http://www.etoxproject.eu/) aimed at creating models to support toxicity prediction by sharing historical toxicological preclinical data within pharmaceutical industries.

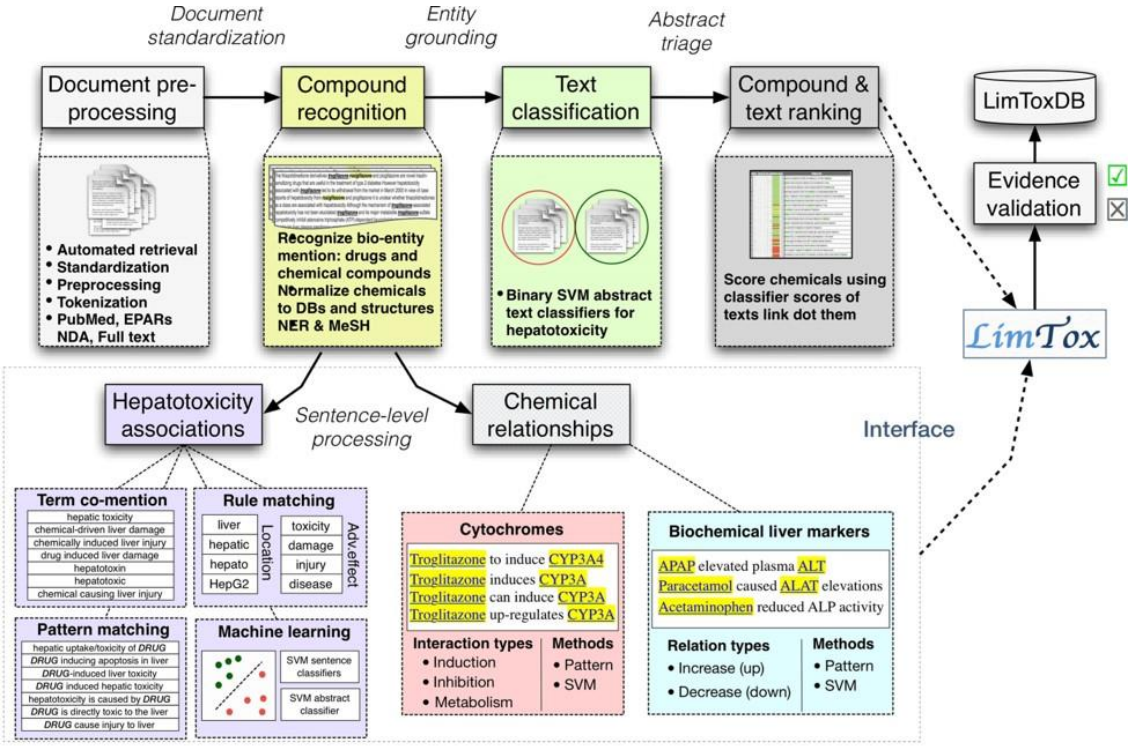

Figure 4: Flow chart of the LimTox system pipeline. The various tasks that are part of the LimTox processing pipeline are shown, from the initial document preprocessing to the detection of chemical entities to the hepatotoxicity text scoring approaches and relation extraction tasks.

3.2 Research outcomes using LimTox

LimTox has been employed inside a text mining pipeline that is currently under development in the context of the eTRANSAFE project [9] (https://etransafe.eu/). As a continuation of the eTOX project, eTRANSAFE aims to create a toxicological knowledge hub for preclinical and clinical data that leverages tools for data sharing, mining, analysis and predictive modelling to account for security and safety assessment in drug development. The text mining efforts of eTRANSAFE are devoted to the implementation of tools to support the identification of relevant information in the The European Federation of Pharmaceutical Industries and Associations (EFPIA) toxicology reports on drug tests

from preclinical studies. In particular, the information of interest consists of treatment-related findings (abnormal observations) that comply with a study report domain template and can be used to develop an annotated preclinical toxicological corpus.

The eTRANSAFE text mining pipeline, called PreTox, is a modular system composed of several components, including a standard pre-processing (tokenization, sentence splitter, part-of-speech tagging, etc.), a sentence classifier (detection of relevant toxicological sentences), and terminology and relation extraction (NER, dictionary annotations, additional taggers, etc.). The pipeline has been implemented following a modular organization using software containers for the different components and orchestrated using Nextflow as workflow manager. The retrieved information is exported to online resources that will be available to be queried and visualized by the end-users. LimTox dwells inside the terminology extraction component of the PreTox pipeline as an annotator of hepatotoxicity concepts (Figure 5).

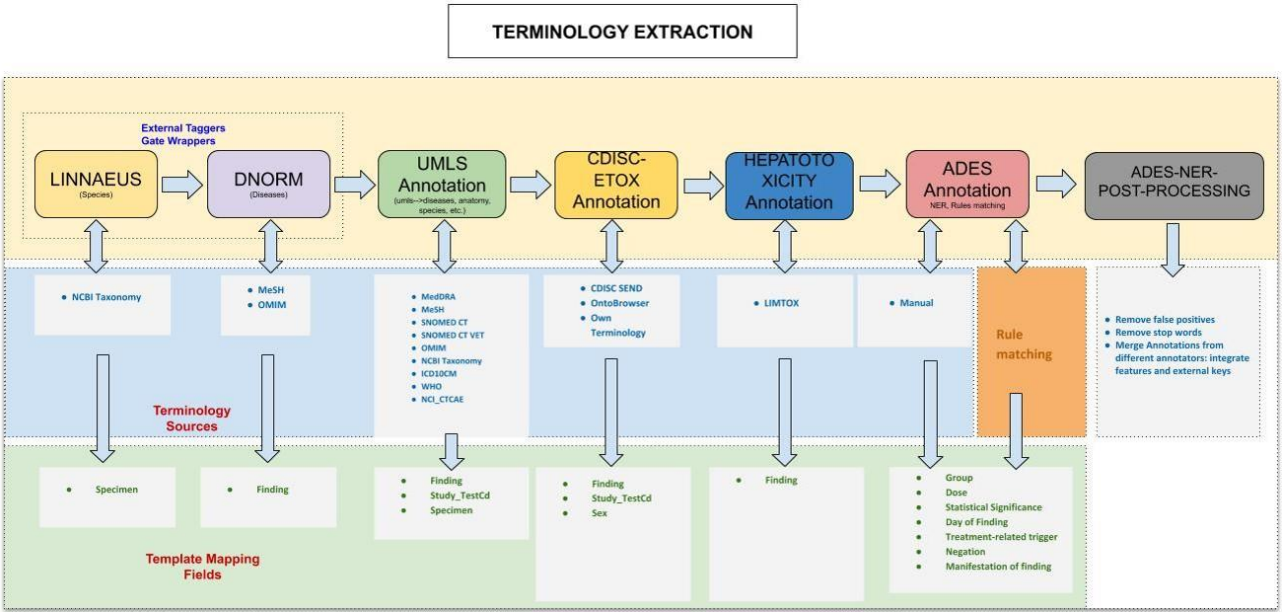

Figure 5: Terminology extraction component of the eTRANSAFE text mining pipeline. LimTox is embedded into the hepatotoxicity annotation module.

Chapter 4 A unified workflow for text mining in paediatric oncology

4.1 The iPC text mining workflow

The iPC text mining workflow allows generating a network representation of the information extracted from PubMed abstracts (Figure 6). The application, which is mostly written in Python, runs in a Docker container and it is publicly available as a GitHub repository at https://github.com/cirillodavide/ipc_textmining. The repository provides documentation to execute an illustrative example.

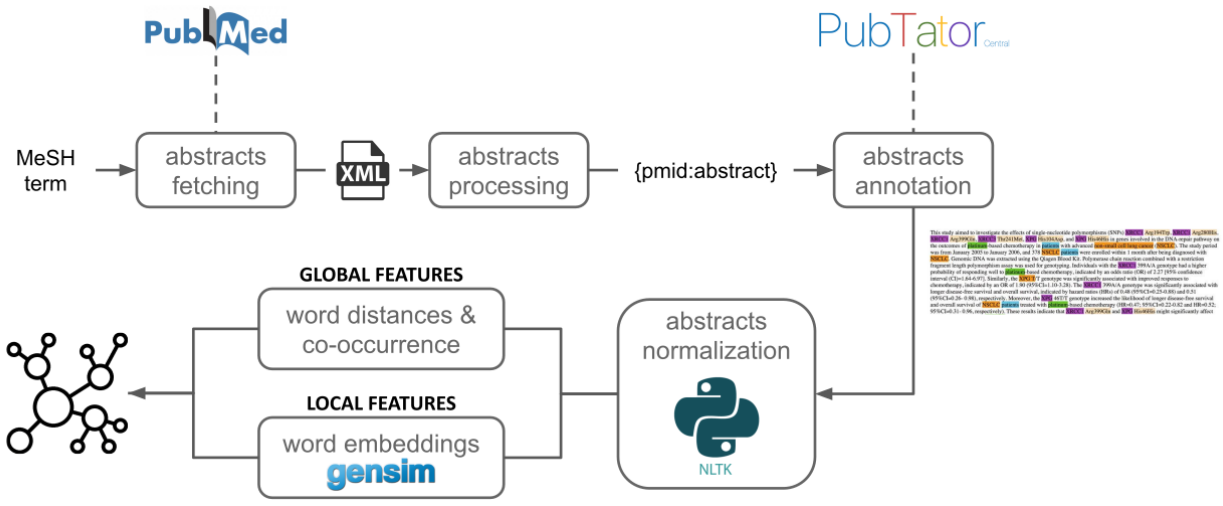

Figure 6: Diagram describing the components of the iPC text mining workflow.

The iPC text mining workflow is based on two main pillars, namely the extraction of bio-entities from selected PubMed abstracts and their network representation based on local and global associations inside the text. In particular, the workflow consists of five sequential steps:

1. **Abstracts fetching**. The abstracts to analyze are selected based on a user-defined query that uses the Medical Subject Headings (MeSH) associated with the paediatric cancer, or other content, of interest. Every article in the PubMed database is associated with a set of MeSH terms, a controlled vocabulary thesaurus describing its content. MeSH terms can be descriptors (main headings) and qualifiers (subheadings). Descriptors indicate the subject or topic, the genre or format, and the geographic locations, while qualifiers convey a particular aspect of a subject. All existing MeSH terms can be interrogated at https://www.ncbi.nlm.nih.gov/mesh/. Queries can be of different types:

    a. A query can consist of a single MeSH term. For example, the query “hepatoblastoma [mesh]” will return all PubMed abstracts indexed with the corresponding MeSH term D018197.

b. A query can include specific subheadings. For example, the query “hepatoblastoma/genetics [mesh]” will return all the PubMed abstracts that specifically address the genetic aspects of this pediatric liver cancer.
c. A query can be composed of several MeSH terms. For example, the query "carcinoma, hepatocellular/genetics [mesh] AND hepatoblastoma/genetics [mesh]" will return all PubMed abstracts that address the genetic aspects of adult and childhood liver cancer.

2. **Abstracts processing and annotation**. The fetched abstracts are downloaded as a single file in XML format. This file is processed in order to return a simple dictionary where the text of each abstract is associated with the corresponding PubMed identifier (PMID). The text of each selected abstract is annotated using PubTator Central (PTC) [10] that is an NCBI service providing automatic annotations of biomedical concepts. PTC is interrogated programmatically through a REST API. The annotations include mentions of disease names, gene/protein names, drugs/chemical, species, mutations at protein and/or DNA level, SNPs, and cell lines.

   To protect them from the next step of text normalization, the PTC annotations are converted into unique identifiers of 10 letters padded with “x”. For example, the sentence “The AXIN1 protein functions in the canonical pathway” is annotated by PTC as “The Gene:8312 protein functions in the canonical pathway” and the annotation is then masked as “The xorikmlyrmxx protein functions in the canonical pathway”. A vocabulary of PTC annotations and internal identifiers is provided.

3. **Abstracts normalization**. This step normalizes the annotated text of each abstract so that the content of all documents are harmonized and comparable. Normalization consists of a series of basic NLP tasks, namely

   a. *Tokenization*, that is breaking the text into tokens such as words, punctuation marks, numerical digits, etc.
   b. *Lemmatization*, that is converting the words in the second or third forms to their first form variants.
   c. *Stemming*, that is reducing the words to its root form.

   For example, the aforementioned annotated sentence “The xorikmlyrmxx protein functions in the canonical pathway” is normalized as “xorikmlyrmxx protein function canon pathway”. Normalization tasks are performed by using the popular NLP Python library called NLTK.

4. **Global and local text features**. The local features of a word are encapsulated in a vector embedding that reflects its local context (i.e., the surrounding words). Additionally, the global features of a word are captured by its distance to and co-occurrence with all other words in the body of the abstract.

   a. *Local text features*. Vector embeddings of each word of the normalized abstracts are created with the open-source Python library GenSim [11], distinguished for efficiency and scalability. To learn the word embeddings, GenSim employs the word2vec technique [12], which is based on a neural network model that can be trained using two possible arquitectures, called continuous bag-of-words (CBOW) and skip-gram (**Figure 7**). The skip-gram training is more apt for infrequent words [13]. The iPC text mining workflow is configured to use skip-gram training, with vector dimensionality of 150 and maximum distance of 3 between the current and predicted words. These parameters can be changed by the user.

   b. *Global text features*. To account for the spatial relationships between pairs of words in the body of the selected abstracts, we compute metrics of words’ distance and co-

occurrence. In particular, we calculate the average number of other words that separate each possible pair of words as well as the number of selected abstracts where such a pair appears.

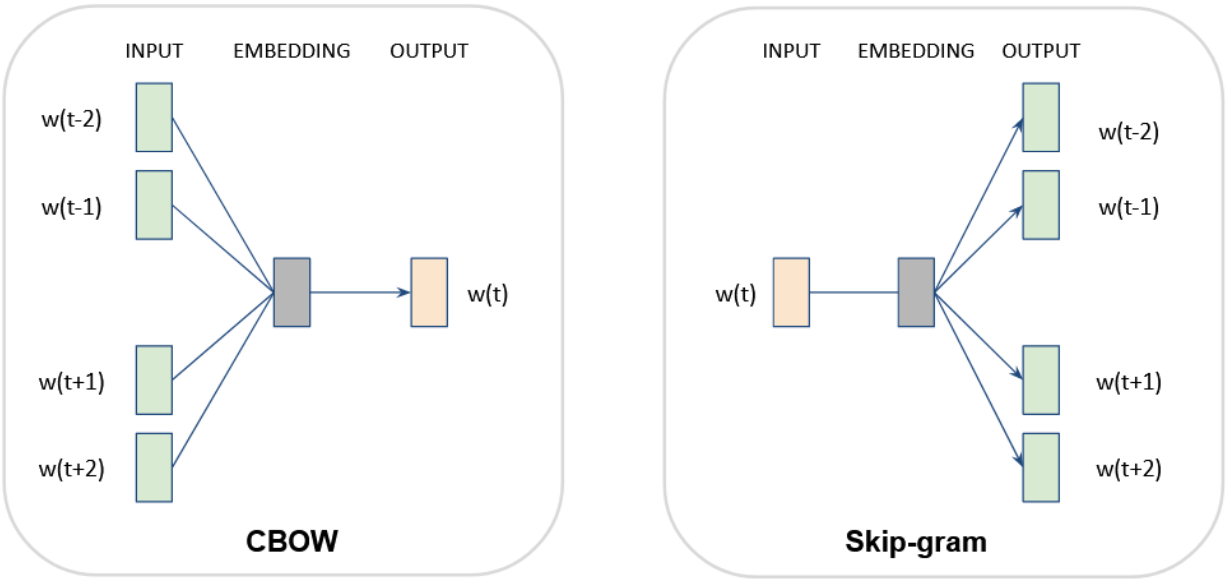

Figure 7: Two possible training architectures of the word2vec model to learn vector embeddings of words. Continuous bag-of-words (CBOW) generates the embeddings (grey box) while learning to predict the current word (orange box) from a surrounding window of context words (green boxes); skip-gram generates the embeddings while learning to predict the surrounding window of context words from the current word.

5. **Network representation**. Local and global text features are used to create a network representation of associations among words. In the case of local text features, the cosine similarity ($cos\theta$) between the vector embeddings of a pair of tokens ($\overrightarrow{a}$ and $\overrightarrow{b}$) is computed as follows:

$$cos\theta = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{||\overrightarrow{a}|| \cdot ||\overrightarrow{b}||}$$

In the case of global text features, the two empirical p-values are computed for distance and co-occurrence (P_k) and combined into one test statistic (χ^2) using the Fisher's method in order to identify pairs of tokens that are significantly close and co-occurrent ($\chi^2 < 0.01$):

$$-2\sum_k logP_k \sim \chi^2_{2k}$$

4.2 Comparison with MelanomaMine and LimTox

The main difference between the iPC text mining workflow and MelanomaMine and LimTox is the scope of their area of application. While the first is an IE modular workflow that can be used to retrieve information on paediatric cancers or other concepts of interest and generate networks, the others are domain-specific software based on pre-trained models dedicated to one specific task (i.e., document classification). Despite the differences in scope, distinct components of the iPC text mining workflow are inspired by those that are in common between MelanomaMine and LimTox and

improved. In particular, three main improvements of the iPC text mining workflow with respect to the domain-specific tools MelanomaMine and LimTox can de identified (Table 1):

1. The document selection is not based on predictions but on curated information. Specifically, the system allows the user to retrieve the PubMed abstracts of interest by using the desired combinations of MeSH descriptors and qualifiers. The main advantages of this approach are an increased flexibility in the content search and the independence from static pre-trained models of distinct biomedical concepts (e.g., melanoma and hepatotoxicity).

2. The bio-entity tagging in PubMed abstracts relies on the continuously updated service PTC instead of pre-defined taggers. PTC is an automated concept annotation in abstracts and full-text biomedical articles available for immediate download. PTC concept identification systems and a new disambiguation module based on deep learning ensure better performances in NER tasks as demonstrated in a BioCreative benchmarking [10].

3. While MelanomaMine and LimTox do not contain a network inference component, the iPC text mining allows generating such relational representations of the information that is extracted. This functionality is based on word embeddings and enables the interoperability between the iPC text mining workflow and other tools that used in the iPC project, such as INtERAcT. Moreover, the networks that are generated can be directly used as additional layers of information on paediatric tumors to integrate in the network-based approaches implemented in other work packages of the project, namely WP4.

Additionally, the iPC text mining workflow has been entirely implemented to be executed inside a Docker container. Packing, deploying, and running the application using Docker containers represent a lightweight and portable solution for any text mining workflow. This also accelerates the cycle of development, test, and production for newer versions of this system. Moreover, dockerization simplifies and automates the deployment process for possible future web applications.

Finally, it is important to notice that the iPC text mining workflow could be easily equipped with the possibility of training an *ad hoc* machine learning model on-the-fly, such as SVMs, to refine the document selection once a set of abstracts on a concept of interest have been retrieved. Anyways, such a training set of abstracts should be large enough to avoid overfitting, which is generally not the case considering the small amount of publications about specific paediatric tumors and rare diseases in general. For this reason, the most appropriate strategy for document selection for iPC is to leverage the wealth of curated MeSH terms and the possibility of combining them.

Tasks and components	MelanomaMine & LimTox	iPC text mining workflow	Comment
Document classification	SVM models requiring domain knowledge feature engineering.	MeSH thesaurus	MeSH thesaurus contains approximately 27,000 entries and is updated annually to reflect changes in the medical terminology. It is human-curated and based on predefined terms with specific definitions and synonyms. Thus, MeSH is effective for searching for meaning, rather than only for keywords or patterns of words ranked based on a static disease-specific model.
Bio-entity tagging	A selection of concept taggers: tmChem (chemical	PubTator Central (PTC)	PTC provides automatic annotations of biomedical concepts in PubMed abstracts and full-text articles. Bio-entity tagging

Tasks and components	MelanomaMine & LimTox	iPC text mining workflow	Comment
	names), GNormPlus (gene/protein names), tmVar (sequence variants), DNorm (disease names)		uses a series of state-of-the-art concept taggers with improved performance benchmarked using BioCreative corpora. PTC service is continuously synchronized with PubMed. The resource can be interrogated programmatically via a REST API (as in the case of the iPC text mining workflow), or downloaded in bulk via FTP.
Network inference	N.A.	Word embeddings and co-occurrence statistics	Word embeddings are efficient and dense representations in which words with similar semantic contexts are rendered with similar encodings. The similarity among word embeddings can be used to infer network representation of the information extracted from text.

Table 1: Main improvements introduced in the iPC text mining workflow in common tasks and components with MelanomaMine and LimTox ("Document classification" and "Bio-entity tagging") as well as new ones ("Network inference").

Chapter 5 Applications of the iPC text mining workflow to paediatric oncology and future work

5.1 Multilayer community trajectories of text-mined medulloblastoma genes

We employed the iPC text mining workflow in a recently published study on medulloblastoma [14]. Medulloblastoma is a malignant and fast-growing primary central nervous system tumor, which originates from embryonic cells of the brain or spinal cord with no known causes and a preferential manifestation in children. Despite being rare, it is the most common cancerous pediatric brain tumor. Four molecular disease subtypes of pediatric medulloblastomas with distinct clinicopathological features have been identified: WNT, SHH, Group 3, and Group 4 [15]. Seven genes exhibit recurrent genetic alterations in the four subgroups (*SHH* in SHH group, *CTNNB1* in WNT group, *MYC* and *MYCN* in Groups 3-4, *ERBB4*, *SRC* and *CDK6* in Group 4 [16,15,17–23].

In our study, we implement an analytical procedure that uses a complex network representation, called multilayer network, presenting relational associations among genes from large-scale repositories of biomedical information, including protein interactions, drug targets, genetic variants, cellular pathways, and metabolic reactions. The procedure is based on a new descriptor that we introduced, called *multilayer community trajectory*, which is defined as the sequence of multilayer network communities visited by a node at increasing values of modularity resolution. The definitions of multilter network community and modularity resolution are given in the following.

Although there is not a consensus definition for community in networks, it is widely accepted to consider a community as the group of nodes that are more densely connected with each other than the rest of the network [24]. A measure of such property is called modularity (Q), which is a quality function of a partition $\mathbf{c}$ of the network X that can be maximized in order to identify communities. A convenient algorithm to detect communities by maximizing the modularity is the Louvain algorithm [25]. In the case of a multilayer network, the sum of the modularities of each layer g is maximized, as in following equation:

$$max \sum_{g} Q_{\gamma}(X^{(g)}, \mathbf{c})$$

Modularity is parametrized to the modularity resolution (γ): the higher the resolution, the smaller the size of the detected communities. An adaptation of the Louvain algorithm for multilayer networks has been implemented in the software MolTi [26]. The multilayer community trajectory of each node of the multilayer network can be, thus, derived by progressively increasing the modularity resolution. Our results illustrate how multilayer community trajectories capture the complexity of multi-omics information despite the small sample size, making this analytical tool particularly suitable for the study of pediatric cancers and rare diseases in general.

To monitor the behaviour of multilayer communities containing medulloblastoma genes upon changes of the modularity resolution, we sought to take into account gene mentions in abstracts of scientific publications about medulloblastoma. By using core functionalities of the iPC text mining workflow (i.e., abstracts fetching, processing and annotation), we retrieved a total of 1941 multi-species genes, consisting of 1475 human genes (76%), 389 murine genes (20%), and 77 genes of other species (4%). We identified the multilayer communities to which the human genes (1387 out of 1475, represented in the multilayer network) belong in a range of modularity resolution of interest (see [14] for details).

As shown in Figure 8, there are plain differences in the trajectories of the communities that are visited by each gene. Interestingly, the trajectories of seven genes, whose recurrent genetic alterations are well-known hallmark features of the four molecular disease subgroups, branch off from well

separated communities, with the exception of *SRC* and *CTNNB1*, which encode proteins that are physical interactors (IntAct interaction accession: EBI-15951997).

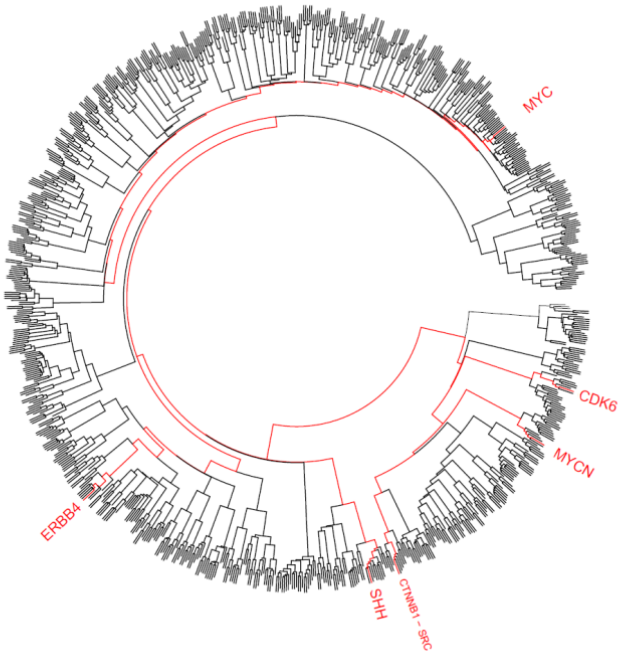

Figure 8: Dendrogram of multilayer community trajectories. The dendrogram represents the Hamming distance among the trajectories of the multilayer communities visited by each gene associated to medulloblastoma by text mining in a range of modularity resolution. Trajectories of the seven genes that are known to characterize the four medulloblastoma subgroups are highlighted in red (SHH in SHH group, CTNNB1 in WNT group, MYC and MYCN in Groups 3-4, ERBB4, SRC and CDK6 in Group 4).

The landscape of these multilayer community trajectories can be further explored to investigate the so-called operations on dynamic communities [27], such as birth (a new community appears), death (a community vanishes), and resurgence (a community disappears and appears again later on). Along the explored range of modularity resolution, the 2186 unique multilayer communities of the text-mined medulloblastoma genes experience a total of 2517 death events and 673 resurgence events (Figure 9), indicating a high level of instability (all communities disappear at least once) but also a high level of commutability (some communities reappear several times with the same exact composition). These observations led us to realize that each gene is characterized by its own journey throughout the communities found at different levels of resolution. For this reason, we further tested the hypothesis that tracing such trajectories for a set of disease-related genes could be exploited for patient clustering purposes, which we explored in the publication using proteogenomic data from two independent medulloblastoma cohorts.

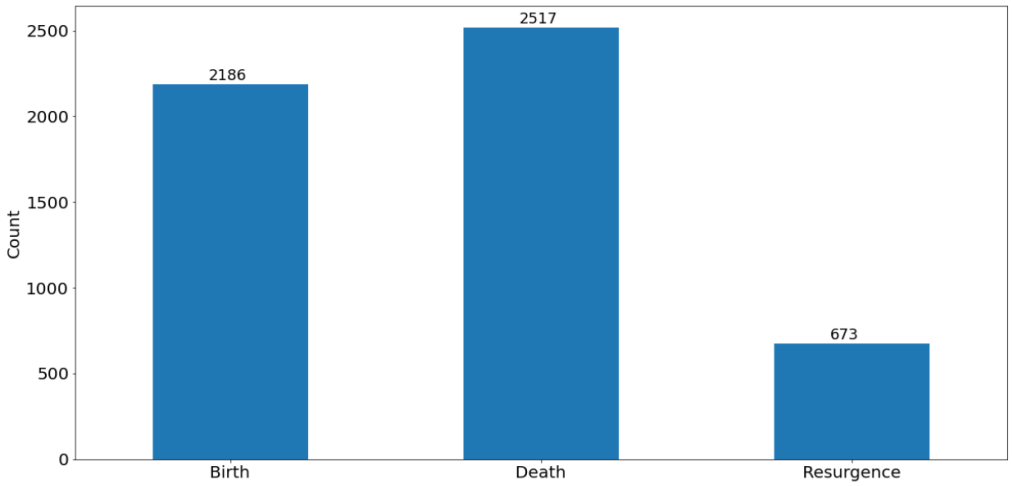

Figure 9: Operations on dynamic communities. Count of dynamic events (birth, death, and resurgence) in the multilayer communities that contain text-mined medulloblastoma genes.

5.2 Network representation of text mined bio-entities of iPC paediatric tumors

The iPC text mining workflow has been used to generate networks of bio-entities for each one of the main five paediatric tumors that are being studied by the iPC consortium, namely Ewing's sarcoma, hepatoblastoma, medulloblastoma, neuroblastoma, and acute lymphoblastic leukemia. PubMed abstracts indexed with such cancers were downloaded in February 2020. The MeSH terms used and the number of retrieved abstracts are reported in Table 2. All output files generated by the iPC text mining workflow for the five paediatric tumors are publicly available at the iPC Nextcloud repository: https://data.ipc-project.bsc.es/s/tgYEkTC7bECWkor

Paediatric tumor	MeSH term	Number of abstracts
Neuroblastoma	D009447	23,122
Acute Lymphoblastic Leukemia	D054198	22,898
Medulloblastoma	D008527	5,148
Ewing's sarcoma	D012512	5,024
Hepatoblastoma	D018197	1,557

Table 2: MeSH terms and number of retrieved abstracts of the main five paediatric tumors studied in the iPC project (abstracts downloaded in February 2020).

As an illustrative example, we report the counts of bio-entities that have been extracted from the abstracts indexed as 'hepatoblastoma' (Figure 10). It is important to notice that, although the abstracts of scientific publications are typically short texts (between 150 and 250 words) and the hepatoblastoma set of abstracts is the smallest in our collection (1,557 abstracts), the number of

extracted bio-entities of relevance for further investigations (i.e., genes, chemicals, diseases) is qualitatively high.

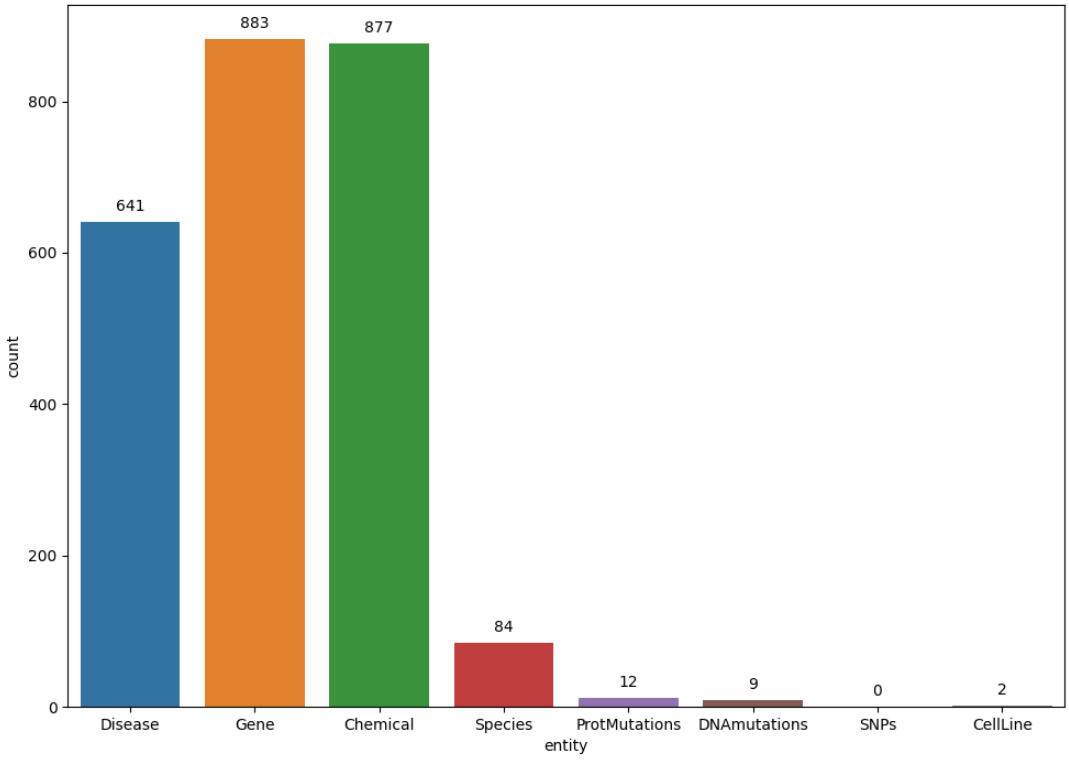

Figure 10: Counts of bio-entities extracted from abstracts indexed as 'hepatoblastoma' (MeSH ID: D018197) (see Table 2).

The network representation of hepatoblastoma content based on local and global text features reveals a high connectivity among different types of bio-entities (Figure 11). Although an in-depth analysis of this network is beyond the scope of this document, a closer look at its structural properties, in particular its community structure, indicates an expected affinity among the types of retrieved bio-entities. In particular, it can be observed how the composition of the communities is rather homogeneous and reflects associations among drugs (e.g., cisplatin, fluorouracil, vincristine), diseases (e.g., biliary atresia, rhabdomyosarcoma, nephroblastoma), and genes (e.g., CDKN2A, TP53, CTNNB1).

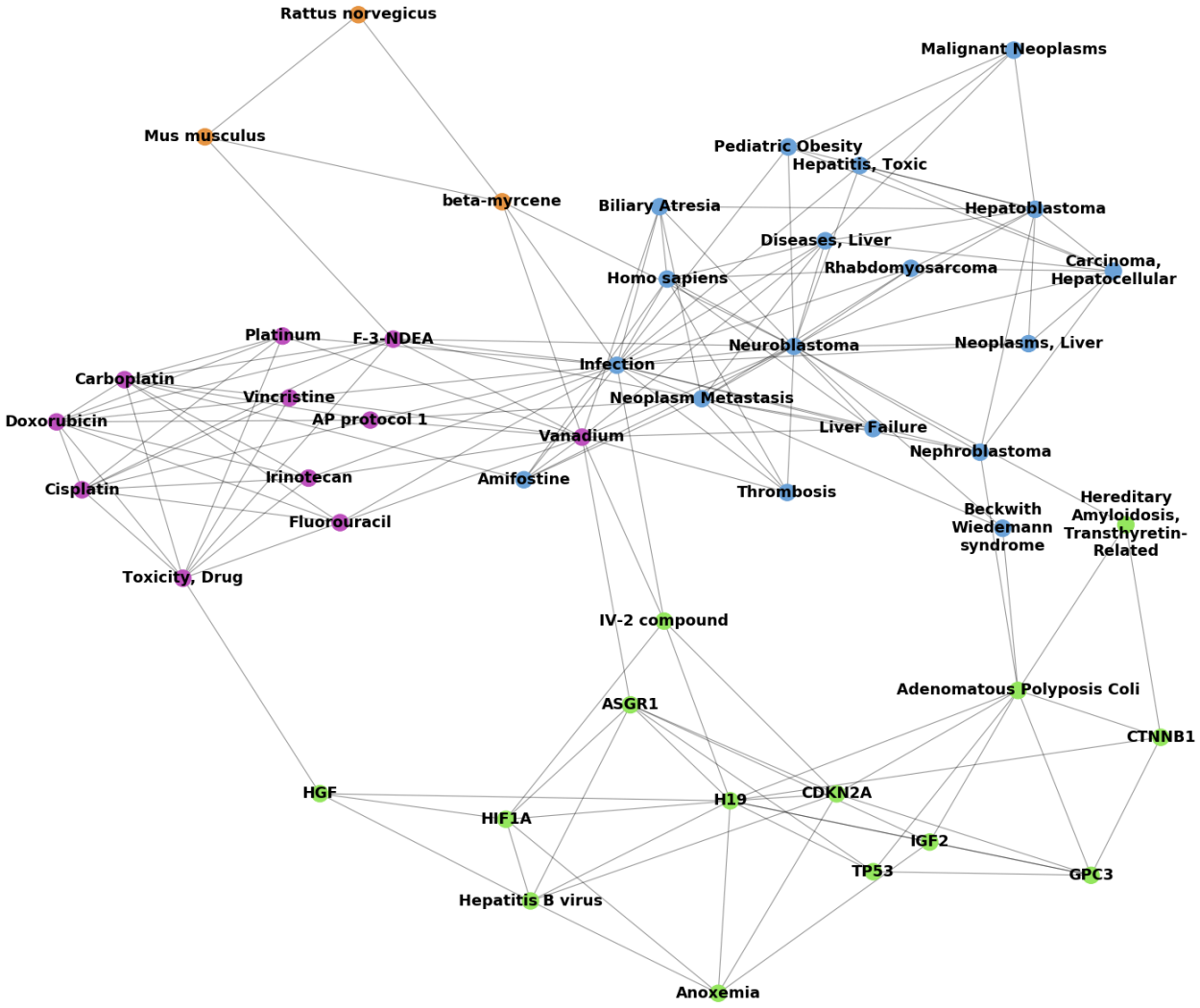

Figure 11: Network representation of the bio-entity associations extracted from PubMed abstracts indexed as 'hepatoblastoma'. For ease of visualization, associations (edges) are shown only among bio-entities (nodes) that are significantly close and co-occurrent ($\chi^2 < 0.01$) and in the 80th percentile of cosine similarity between the corresponding word embeddings (i.e., with most similar semantic context). Nodes are colored based on membership to the four network communities detected using the Clauset-Newman-Moore greedy modularity maximization algorithm implemented in the Python library Networkx.

5.3 INtERAcT using the word embeddings of iPC paediatric tumors

The word embeddings generated for the five paediatric tumors under study within the iPC project (see section 5.2) have been used to infer a protein-protein interaction network using INtERAcT [7].

INtERAcT was developed in the H2020 project PrECISE (ref. 668858), where it was used to extract protein-protein interactions. INtERAcT infers interactions using a novel metric that exploits word embeddings. It is important to notice that generating word embeddings does not require text labeling for training or domain-specific knowledge, and hence it can be easily applied to different scientific domains in a completely unsupervised way. INtERAcT defines its metric on a discretized space. A clustering of the word embedding space is performed, with the goal to define semantic word groups. The score for the interactions between two entities of interest is then defined using the Jensen-Shannon divergence between their neighbors' distributions over clusters' assignment.

The lists of protein-protein interaction inferred with INtERAcT using the word embedding generated with the iPC text mining workflow (see section 5.2) are available at the public GitHub repository https://github.com/drugilsberg/interact/tree/master/examples/ipc.

5.4 Future work

The European Genome-phenome Archive (EGA) is one the largest European platforms for sharing and reusing genetic and phenotypic data resulting from biomedical research projects. Many of the studies in EGA are fundamental to support research in paediatric oncology and the objectives of the iPC project. Although metadata querying is the most common way to retrieve EGA datasets, the inherent fuzziness of categorical descriptors, as well as the presence of synonyms and similar concepts, make this practice extremely inefficient.

In the next future, we aim to leverage the components of the iPC text mining workflow to increase the efficiency of the EGA metadata query system. In particular, the generation of word embeddings capturing the content of the available metadata can be used to model and expand the queries. Graph-based machine learning approaches applied on the network representations of the extracted information can provide ranked lists of relevant concepts. Such concepts can be mapped into the word embeddings in order to retrieve similar ones and provide supplementary terms to enrich the query. The expanded query can be finally used to search among all the available datasets and the results ranked according to relevance.

Chapter 6 Summary and Conclusion

The document reports on the development of a unified text mining workflow for information extraction in paediatric oncology. The iPC text mining workflow is based on distinct components of existing text mining tools that we previously designed, called MelanomaMine and LimTox, dedicated to specific areas of applications of cancer research, namely melanoma research and liver toxicology, respectively. Both MelanomaMine and LimTox have been employed in additional works related to or supported by the iPC project.

The iPC text mining workflow enables the automatic extraction of mentions of bio-entities (genes, proteins, drugs, diseases and others) and their associations, allowing to unify and interpret biomedical information of text sources and to further analyze it with complementary NLP approaches (e.g., INtERAcT). Moreover, it facilitates the generation of network representation of the text mined information for further utilization in other iPC work packages, specifically WP4 (“Network-based meta-analysis of multi-omics and text-mined data”), in order to integrate it with clinical and molecular information of multi-omics datasets available within the iPC consortium.

Three use cases where the iPC text mining workflow have been used are reported, namely the generation of multilayer community trajectories of text mined medulloblastoma genes, and the creation of network representations of text mined information as well as the inference of protein-protein interactions with INtERAcT based on the word embeddings of the main five paediatric cancers studied by the iPC consortium. In the future, components of the iPC text mining workflow will be used to improve the EGA metadata query system.

List of Abbreviations

Abbreviation	Translation
NLP	Natural Language Processing
WEX	Watson Explorer
TAC	Text Analytics Catalog
CBOW	Continuous Bag-Of-Words
MeSH	Medical Subject Headings
IE	Information extraction
EHR	Electronic Health Record
NER	Named Entity Recognition
EGA	European Genome-phenome Archive
PTC	PubTator Central
CMM	Cutaneous Malignant Melanoma
SVM	Support Vector Machine

Bibliography

[1] Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet 2012;13:829–39. https://doi.org/10.1038/nrg3337.

[2] Yim W, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol 2016;2:797. https://doi.org/10.1001/jamaoncol.2016.0213.

[3] Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. Journal of Biomedical Informatics 2015;57:333–49. https://doi.org/10.1016/j.jbi.2015.08.013.

[4] Pyysalo S, Baker S, Ali I, Haselwimmer S, Shah T, Young A, et al. LION LBD: a literature-based discovery system for cancer biology. Bioinformatics 2019;35:1553–61. https://doi.org/10.1093/bioinformatics/bty845.

[5] Dalianis H. Clinical Text Mining. Cham: Springer International Publishing; 2018. https://doi.org/10.1007/978-3-319-78503-5.

[6] Cañada A, Capella-Gutierrez S, Rabal O, Oyarzabal J, Valencia A, Krallinger M. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. Nucleic Acids Research 2017;45:W484–9. https://doi.org/10.1093/nar/gkx462.

[7] Manica M, Mathis R, Cadow J, Rodríguez Martínez M. Context-specific interaction networks from vector representation of words. Nature Machine Intelligence 2019;1:181–90. https://doi.org/10.1038/s42256-019-0036-1.

[8] Ferrucci DA. Introduction to “This is Watson.” IBM J Res & Dev 2012;56:1:1-1:15. https://doi.org/10.1147/JRD.2012.2184356.

[9] Pognan F, Steger-Hartmann T, Díaz C, Blomberg N, Bringezu F, Briggs K, et al. The eTRANSAFE Project on Translational Safety Assessment through Integrative Knowledge Management: Achievements and Perspectives. Pharmaceuticals 2021;14:237. https://doi.org/10.3390/ph14030237.

[10] Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Research 2019;47:W587–93. https://doi.org/10.1093/nar/gkz389.

[11] Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks, 2010, p. 45–50.

[12] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. ArXiv:13013781 [Cs] 2013.

[13] Google Code Archive - Long-term storage for Google Code Project Hosting. n.d. https://code.google.com/archive/p/word2vec/ (accessed May 1, 2021).

[14] Núñez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. IScience 2021;24. https://doi.org/10.1016/j.isci.2021.102365.

[15] Taylor MD, Northcott PA, Korshunov A, Remke M, Cho Y-J, Clifford SC, et al. Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol 2012;123:465–72. https://doi.org/10.1007/s00401-011-0922-z.

[16] Clifford SC, Lusher ME, Lindsey JC, Langdon JA, Gilbertson RJ, Straughton D, et al. Wnt/Wingless pathway activation and chromosome 6 loss characterize a distinct molecular sub-group of medulloblastomas associated with a favorable prognosis. Cell Cycle 2006;5:2666–70. https://doi.org/10.4161/cc.5.22.3446.

[17] Kool M, Korshunov A, Remke M, Jones DTW, Schlanstein M, Northcott PA, et al. Molecular

subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. Acta Neuropathol 2012;123:473–84. https://doi.org/10.1007/s00401-012-0958-8.

[18] Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, et al. Novel mutations target distinct subgroups of medulloblastoma. Nature 2012;488:43–8. https://doi.org/10.1038/nature11213.

[19] Northcott PA, Lee C, Zichner T, Stütz AM, Erkek S, Kawauchi D, et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 2014;511:428–34. https://doi.org/10.1038/nature13379.

[20] Kool M, Jones DTW, Jäger N, Northcott PA, Pugh TJ, Hovestadt V, et al. Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell 2014;25:393–405. https://doi.org/10.1016/j.ccr.2014.02.004.

[21] Ramaswamy V, Remke M, Bouffet E, Bailey S, Clifford SC, Doz F, et al. Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. Acta Neuropathol 2016;131:821–31. https://doi.org/10.1007/s00401-016-1569-6.

[22] Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, et al. The whole-genome landscape of medulloblastoma subtypes. Nature 2017;547:311–7. https://doi.org/10.1038/nature22973.

[23] Forget A, Martignetti L, Puget S, Calzone L, Brabetz S, Picard D, et al. Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling. Cancer Cell 2018;34:379-395.e7. https://doi.org/10.1016/j.ccell.2018.08.002.

[24] Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E 2006;74:016110. https://doi.org/10.1103/PhysRevE.74.016110.

[25] Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech 2008;2008:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

[26] Didier G, Valdeolivas A, Baudot A. Identifying communities from multiplex biological networks by randomized optimization of modularity. F1000Res 2018;7:1042. https://doi.org/10.12688/f1000research.15486.2.

[27] Cazabet R, Amblard F. Dynamic Community Detection. In: Alhajj R, Rokne J, editors. Encyclopedia of Social Network Analysis and Mining, New York, NY: Springer; 2014, p. 404–14. https://doi.org/10.1007/978-1-4614-6170-8_383.