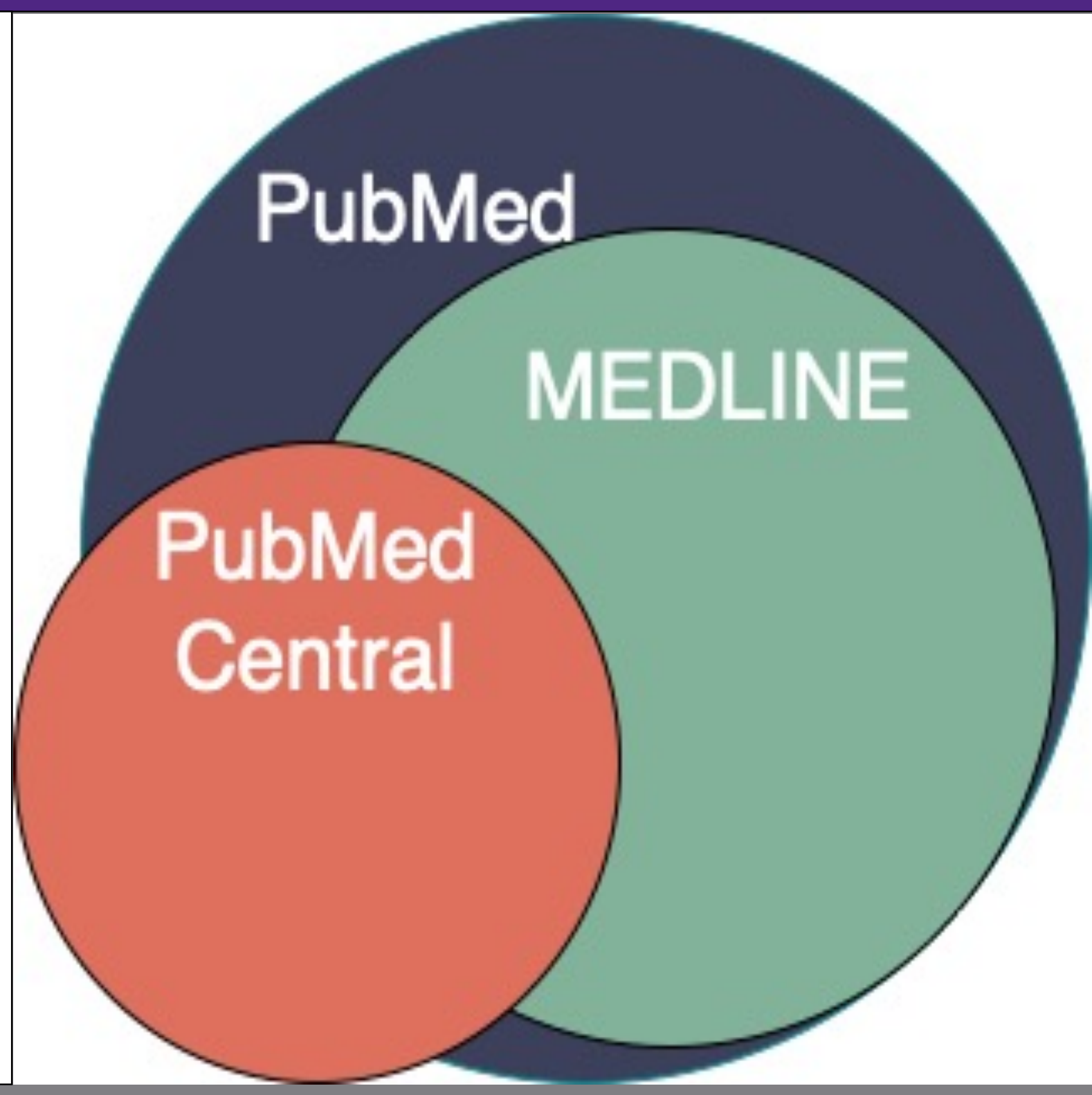
MeSHup: A Corpus for Full Text Biomedical Document Indexing

Xindi Wang[1,3], Robert E. Mercer[1,3], and Frank Rudzicz[2,3,4]

[1]Department of Computer Science, University of Western Ontario, London, Ontario, Canada
[2]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[3]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[4]Unity Health Toronto, Toronto, Ontario, Canada

BACKGROUND

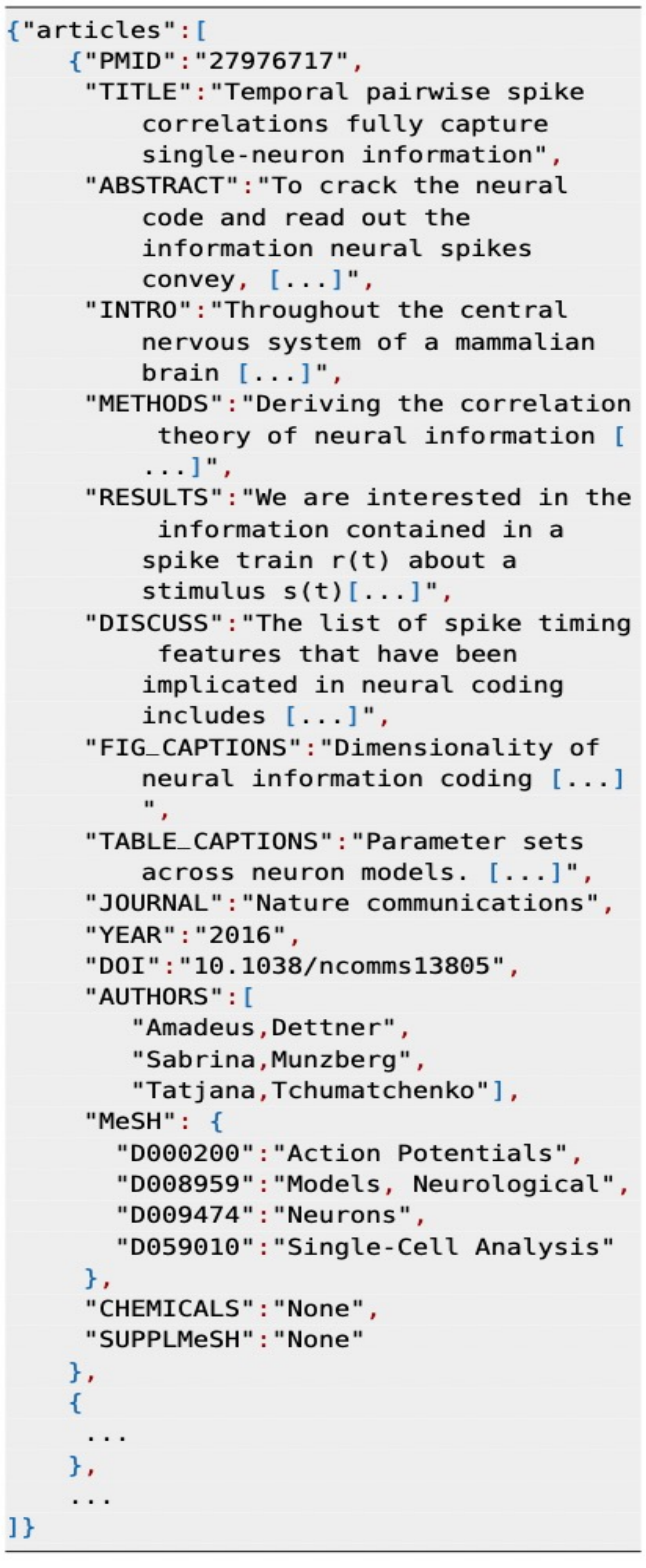
- MEDLINE
  - Core database, contains more than 28 million references to a specific set of journals in biomedical science.
- PubMed
  - A free access search engine for abstracting and indexing biomedical citations.
  - Comprises more than 33 million citations for biomedical literature from MEDLINE (as of Apr. 2022).
  - Links to articles from publisher's websites and PubMed Central.
- PubMed Central (PMC)
  - Full-text archive for biomedical and life sciences journal articles.
  - 7.9 million articles are archived in PMC



MOTIVATIONS

- Existing corpora only provide the title and abstract, while human annotators review the full text articles.
- Previous work focused on text information but are less concerned with metadata.

DATA SAMPLE



```
{"articles":[
    {"PMID":"27976717",
     "TITLE":"Temporal pairwise spike
         correlations fully capture
         single-neuron information",
     "ABSTRACT":"To crack the neural
         code and read out the
         information neural spikes
         convey, [...]",
     "INTRO":"Throughout the central
         nervous system of a mammalian
         brain [...]",
     "METHODS":"Deriving the correlation
          theory of neural information [
         ...]",
     "RESULTS":"We are interested in the
          information contained in a
         spike train r(t) about a
         stimulus s(t)[...]",
     "DISCUSS":"The list of spike timing
          features that have been
         implicated in neural coding
         includes [...]",
     "FIG_CAPTIONS":"Dimensionality of
         neural information coding [...]
         ",
     "TABLE_CAPTIONS":"Parameter sets
         across neuron models. [...]",
     "JOURNAL":"Nature communications",
     "YEAR":"2016",
     "DOI":"10.1038/ncomms13805",
     "AUTHORS":[
        "Amadeus,Dettner",
        "Sabrina,Munzberg",
        "Tatjana,Tchumatchenko"],
     "MeSH": {
        "D000200":"Action Potentials",
        "D008959":"Models, Neurological",
        "D009474":"Neurons",
        "D059010":"Single-Cell Analysis"
     },
     "CHEMICALS":"None",
     "SUPPLMeSH":"None"
    },
    {
     ...
    },
    ...
]}
```

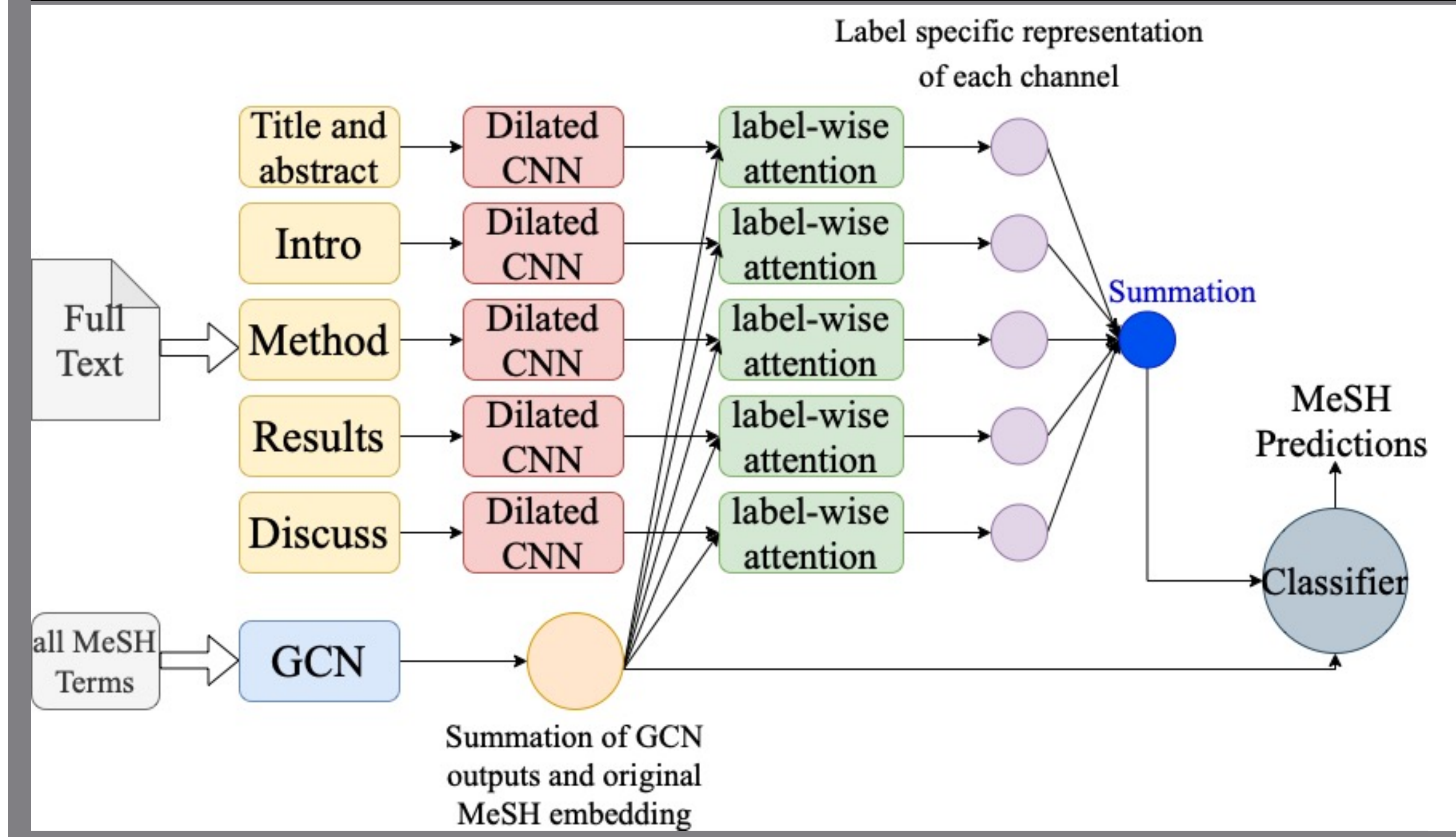
DATASET CONSTRUCTION

- Data resources
  - PubMed Central Open Access in BioC format (BioC-PMC)
  - MEDLINE / PubMed Annual Baseline Repository (MBR)
- Constrains
  - Articles indexed by human annotators only
  - English articles only
- Information extracted from BioC-PMC
  - Eight BioC sections are selected to construct the new corpus: title, abstract, introduction, methods, results, discussion, figure captions, and table captions.
- Information extracted from MBR
  - Metadata: PMID, authors, journal name, publication year, DOI, MeSH terms, supply MeSH, and chemical list.

MeSHup CORPUS

- Contains a set of 1, 342, 667 biomedical documents.
- Each article has full textual information and metadata associated with it.

BASELINE MODEL AND RESULTS



Bipartition evaluation		*Methods*	
		Titles and Abstracts	*Full Texts*
Example based	*EBF*	0.183	**0.259**
	EBP	0.503	**0.588**
	EBR	0.112	**0.166**
Micro-averaged	*MiF*	0.177	**0.259**
	MiP	0.473	**0.604**
	MiR	0.110	**0.164**
Macro-averaged	*MaF*	0.362	**0.367**
	MaP	0.798	**0.810**
	MaR	0.234	**0.237**

Table 3: Comparison using only titles and abstracts and full texts across bipartition evaluation. Bold: best scores in each row.

CONTRIBUTIONS

- We release a large-scale annotated MeSH indexing corpus, MeSHup.
- We train an end-to-end multichannel model that incorporates different sections of the full text article to show that full texts are more informative in the MeSH indexing tasks compared to the titles and abstracts only

LREC 2022
Marseille