# Inferring context specific **PPI** networks using **unsupervised deep learning** for text mining

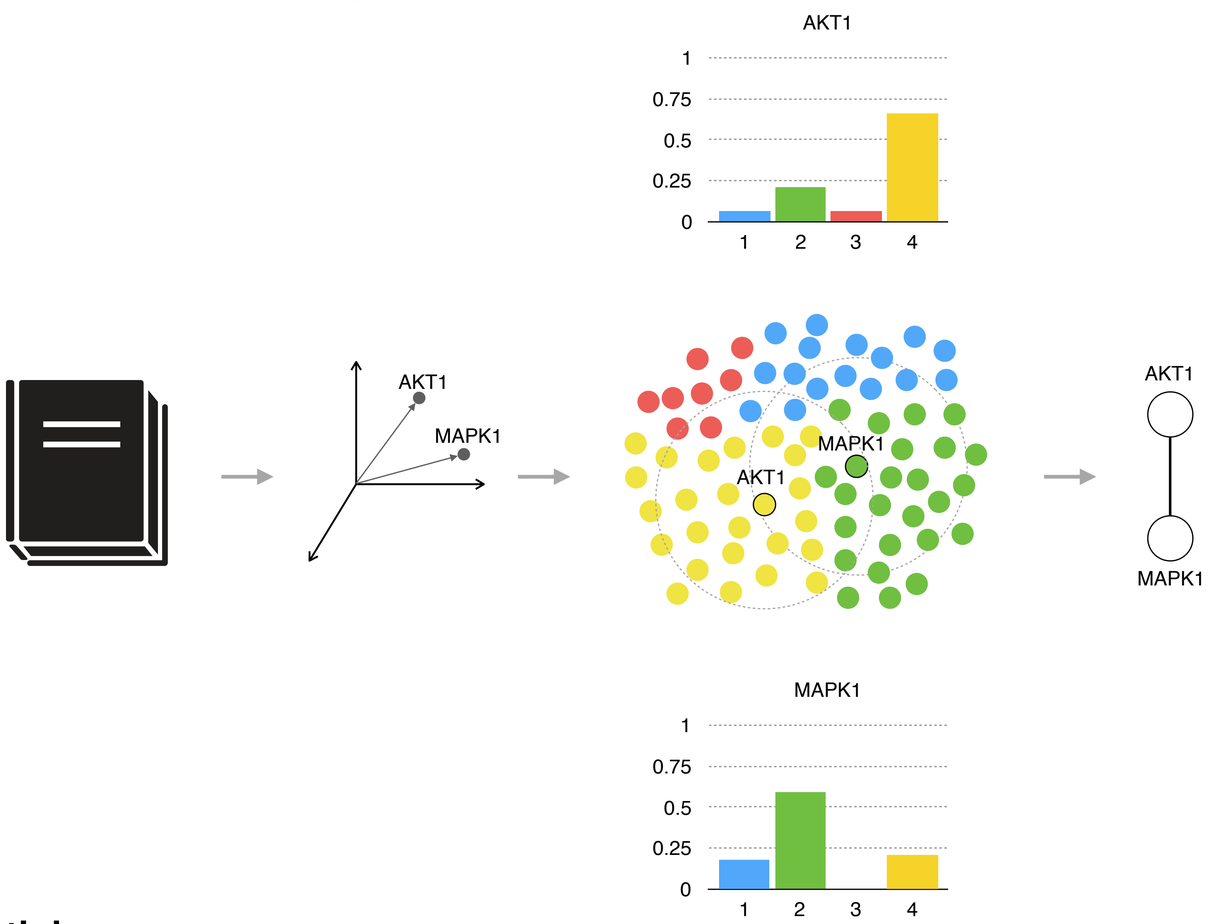
**What**
INtERAcT is a method to extract interaction networks from unstructured text.

- Fully **unsupervised**
  - **no** need for heavy **manual curation**
  - **no** need for a **language model**
  - **no annotations**
- Context-specific networks

**How**
Words in a vocabulary are mapped into vectors in a continuous, high dimensional space. In this representation, words that share a similar context in the corpus are located in close proximity in the word embedding vector space.

INtERAcT (Interaction Network InfErence from VectoR RepresentATion of Words) exploits **word vectors** trained on a **context-specific** corpus and defines a metric to score confidence of interactions.

Our proposed metric is based on a **clustering** of the embedded space and the distributional properties of word neighbors. This scoring is derived from the **Jensen-Shannon divergence** between the neighbor distributions of the words representing proteins.




**tl;dr**
1. Build a word embedding from context specific publications
2. INtERAcT will cluster embedded words and infer a network using neighbor distributions over clusters similarity
3. Profit

**Why**
As the number of scientific publications grows exponentially, search engines such as PubMed [`https://www.ncbi.nlm.nih.gov/pubmed`] provide an unprecedented amount of information in the form of **unstructured text**. Manual curation becomes infeasible.
Many approaches to **extract protein-protein interaction** (PPI) information from require feature engineering and expert-domain knowledge, hence preventing full automation.

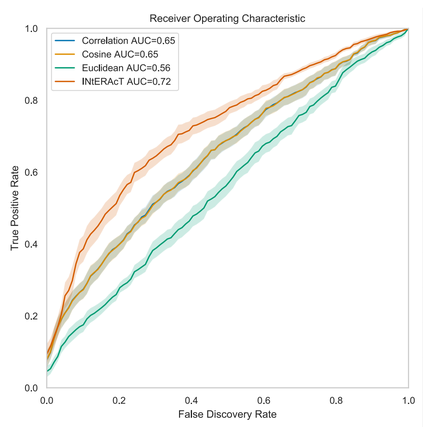
**Results**
We built context specific word embeddings for ten cancer types, e.g. from publications returned when searching PMC [`https://www.ncbi.nlm.nih.gov/pmc`] for “non small cell lung cancer”.

As reference we used (human) PPIs from the STRING database [`https://string-db.org`]. STRING combines multiple types of evidence, including text-mining and experimental evidence.

INtERAcT outperforms the Euclidean, cosine and correlation distances which are alternative, commonly used metrics for word vector similarity scoring.

For eight of the ten cancer types INtERAcT shows superior rank correlation compared to the text mining PPI predictions from STRING (excluding the text-mining evidence from the ground truth reference).

We tested INtERAcT using only abstract vs. full text and on a wide range of parameters for the embedding and scoring. Remarkably, INtERAcT performance is largely insensitive to the choice of embedding parameters.




| Pathway | Correlation with INtERAcT | Correlation with STRING text mining |
|---|---|---|
| KEGG Acute Myeloid | **0.326331** | 0.301880 |
| KEGG Bladder Cancer | **0.511700** | 0.336506 |
| KEGG Chronic Myeloid | 0.316876 | **0.386503** |
| KEGG Colorectal Cancer | **0.519800** | 0.344804 |
| KEGG Glioma | **0.430018** | 0.279987 |
| KEGG Non Small Cell Lung | **0.496619** | 0.333499 |
| KEGG Pancreatic Cancer | **0.552383** | 0.318142 |
| KEGG Prostate Cancer | **0.512520** | 0.285628 |
| KEGG Renal Cell Carcinoma | **0.555246** | 0.428592 |
| KEGG Small Cell Lung Cancer | 0.320010 | **0.412858** |

Context-specific interaction networks from vector representation of words
Matteo Manica, Roland Mathis, Joris Cadow and María Rodríguez Martínez
Nature Machine Intelligence 1(4), 181–190, 201

The project leading to this application received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 668858

INtERAcT

is open source and deployed on IBM Cloud as a web service.
Visit `ibm.biz/interact` to find all things related.

NOPrECISE

IBM **Research**