What do we know about linguistic journals?

Applying NLP methods to a dataset of abstracts from linguistic journals (work in progress)

George Moroz, Asya Alekseeva, Timofei Dedov, Artyom Orekhov, Kirill Sidorov, Angelina Stepanova

2 May 2023

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclussions

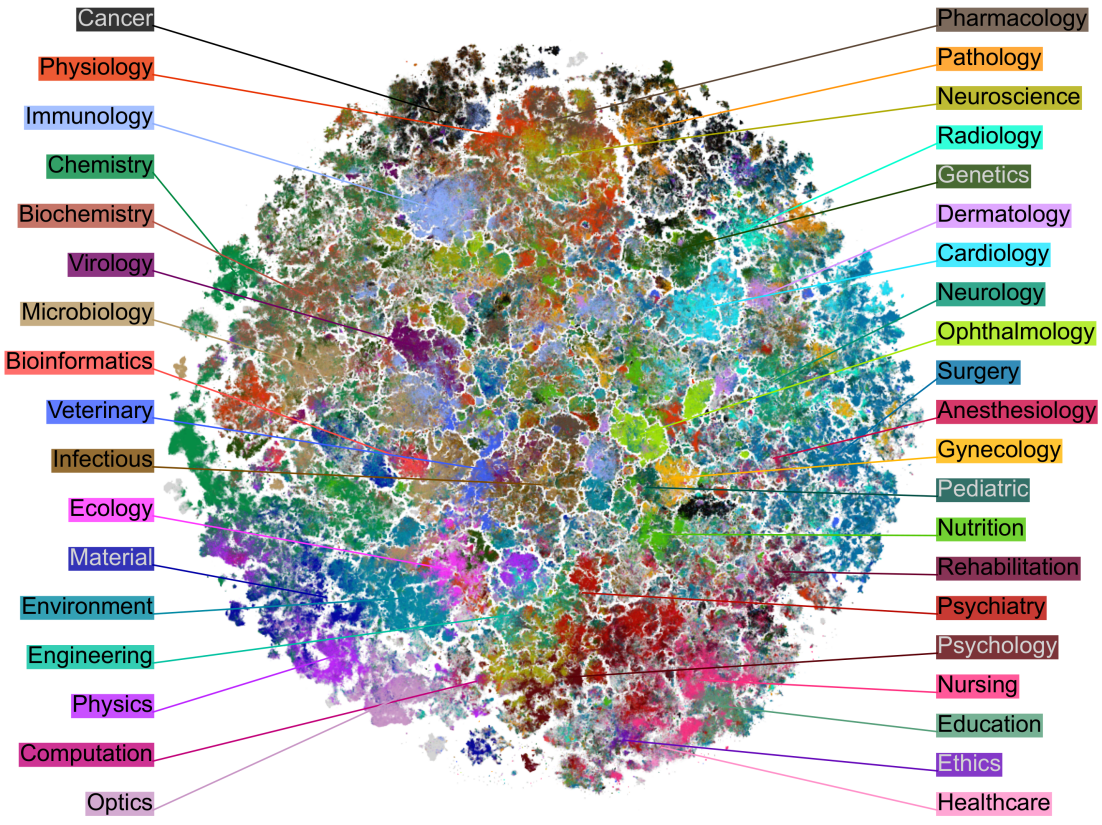
(Gonzalez-Marquez et al. 2023) The landscape of biomedical research

> *The number of publications in biomedicine and life sciences has rapidly grown over the last decades, with over 1.5 million papers now published every year. This makes it difficult to keep track of new scientific works and to have an overview of the evolution of the field as a whole. Here we present a 2D atlas of the entire corpus of biomedical literature, and argue that it provides a unique and useful overview of the life sciences research. <...>*

https://static.nomic.ai/pubmed.html (interactive version)

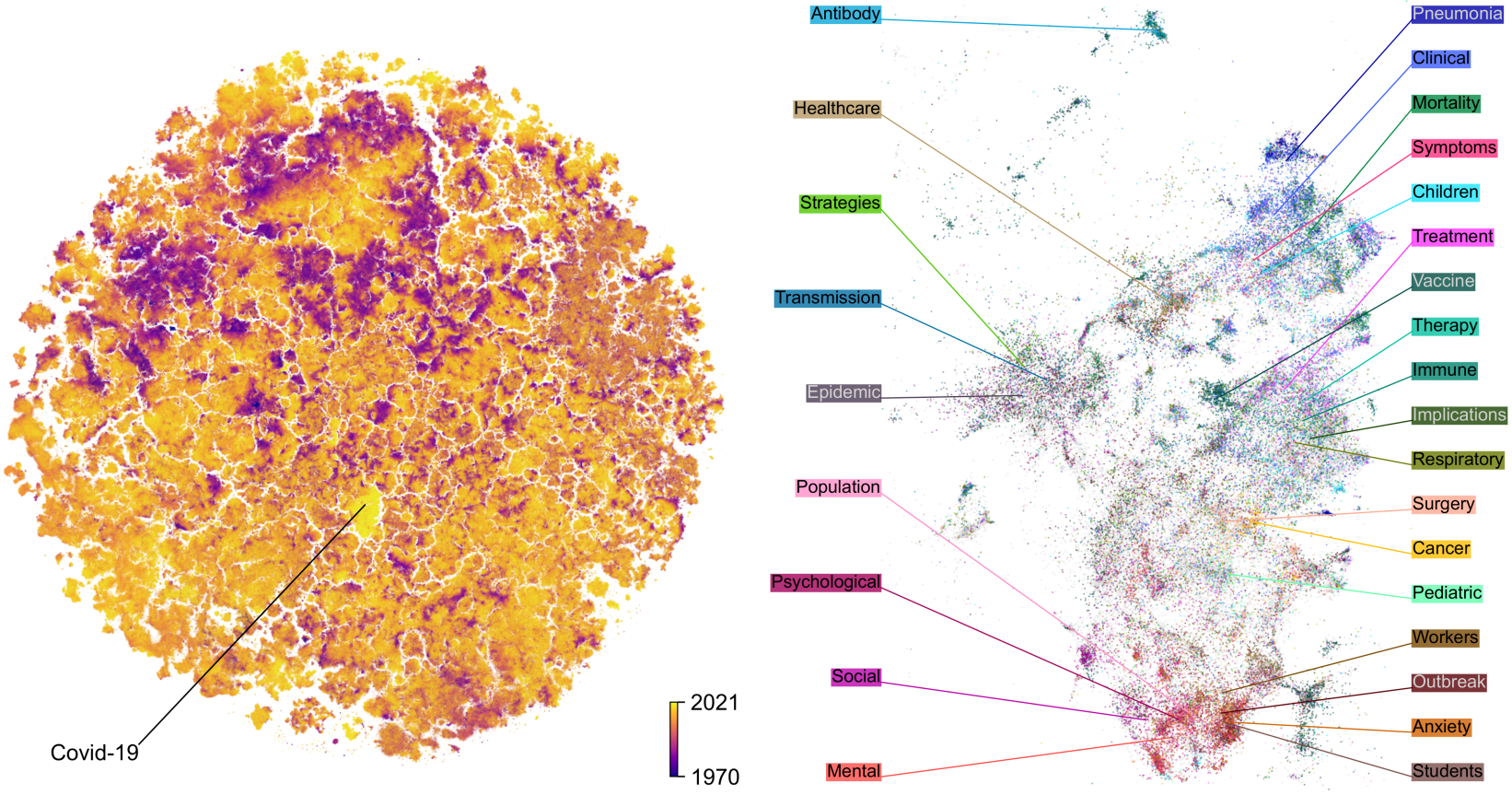
This is a preprint and has not been certified by peer review!

(Gonzalez-Marquez et al. 2023)



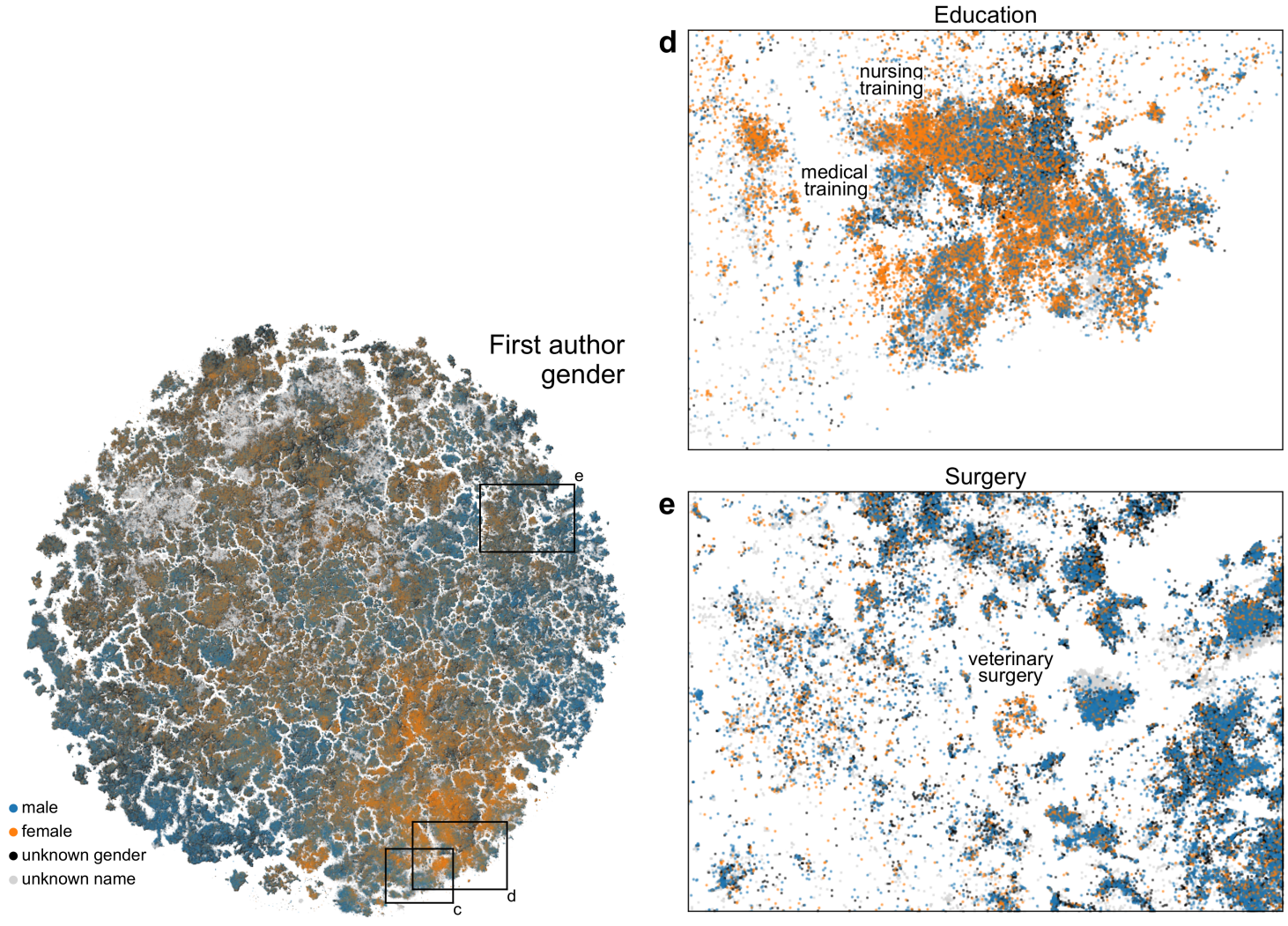
2D embedding of the PubMed dataset. Paper abstracts (n = 21 M) were transformed into 768-dimensional vectors with PubMedBERT (Gu et al. 2021) and then embedded in 2D with t-SNE (Van der Maaten and Hinton 2008). Coloured using labels based on journal titles. Unlabeled papers are shown in gray and are displayed in the background.

(Gonzalez-Marquez et al. 2023)



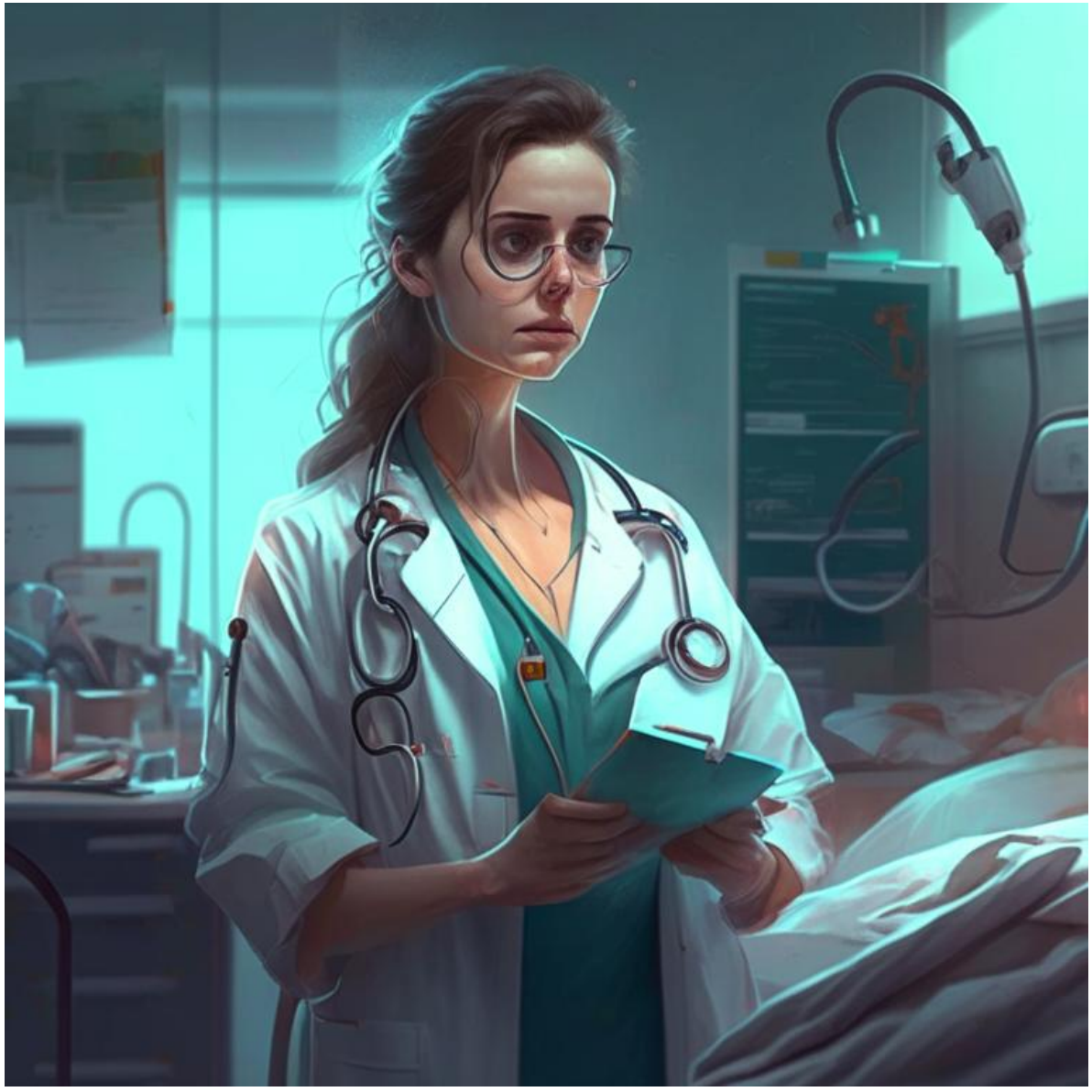
Covid-19 region of the map. Colours are assigned using labels based on paper titles. Unlabeled Covid papers are shown in the background in gray. This region in the embedding also contained some non-Covid papers (~15%) about other respiratory epidemics; they are not shown.

(Gonzalez-Marquez et al. 2023)



Papers coloured by the inferred gender of their first authors.(d–e) Regions of the map showing within-label heterogeneity in the distribution of first authors' gender.

Reaction of proffesional doctors and pharmacologist



AI generated picture

How wonderful would it be to have something like this for linguistics?

- zoning of different subfields of linguistics
- interaction of different subfields of linguistics
  - can we see a boundary between morphology and syntax?
  - can we see cancer-like behavior of the Computer linguistics among all subfields of modern linguistics?
  - can we visualize the emergence of Computer linguistics?
  - can we show how strong/vague are boundaries between philology, Digital Humanity and linguistics?
  - can we visualize the history of linguistics?
  - can we see the scientific track/preferences of different researchers? E. g. if the person is a phonetician will this person also be into the computational linguistics?
  - ...

Plan

- extract abstracts from publications in linguistic journals
- use them in order to create some space of linguistic publications (one dot – one publication)
- see how linguistic journals are distributed within this space

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclussions

Our team

- conceptualization: George Moroz, Boris Orekhov
- team leadership: George Moroz, Asya Alekseeva, Kirill Sidorov
- data curation, data analysis: George Moroz
- data gathering and annotation:
  - DH masters: Kirill Sidorov and Artyom Orekhov
  - bachelor students: Asya Alekseeva, Timofei Dedov and Angelina Stepanova
  - group of bachelor students who choose this project as a summer practice (2-week internship)

Outline of the talk

Introduction

Our team

Data collection

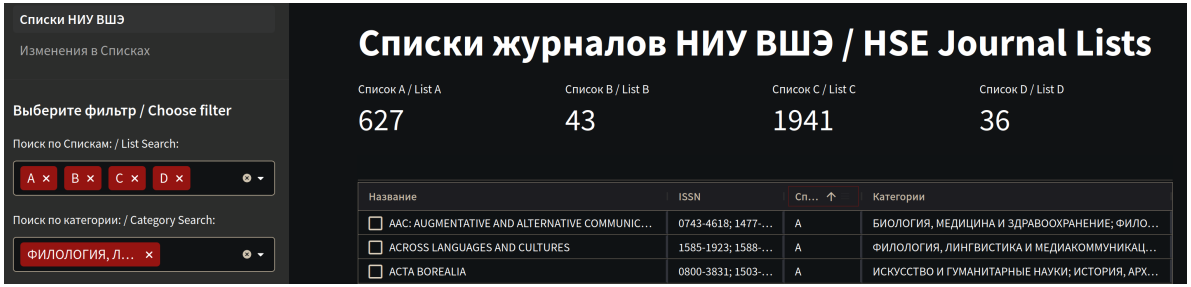
Data analysis

Conclussions

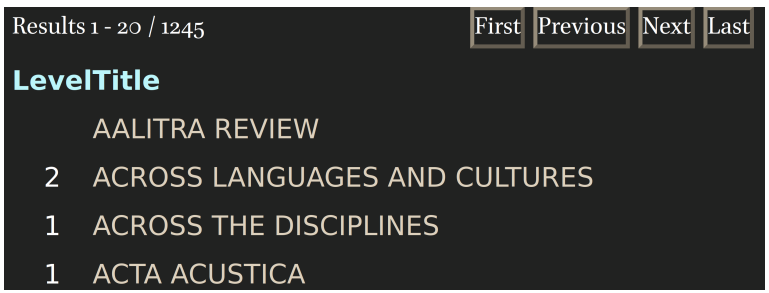
Journal lists

We have different journal lists

- Tag `филология, лингвистика, медиакоммуникации` from list of journals HSE uses for Academic Merit Bonus.[1]



- Tag `6162 Languages` from journal rankings from Finish Publication Forum



[1]As far as I know, we are not allowed to disclosure those lists outside the HSE. However, right now they are available without any checks for HSE affiliation.

Journal lists

After we gathered journal lists we annotated them according to personal beliefs on what is linguistics, and what is not (tags: `linguistics` (232); `interdisciplinary` (203); `language_learning` (26) and some others).

There is also a side-project on annotation of journals for literature science (tags: `literary_studies` (168); `interdisciplinary` (253); `philology` (12) and some others).

After manual annotation 232 (16%) journals out of 1421 have tag `linguistics`. 1334 journals have not yet been annotated.

	Helsenki level				
HSE level	a	b	c	d	NA
a	50	51	23	1	1
b	1	3	12	NA	3
c	NA	7	42	1	37

Abstracts extraction

- We planed to scrap all papers' metadata from the journals' web-pages

Abstracts extraction

- We planed to scrap all papers' metadata from the journals' web-pages
- Then we discovered the `crossref` database and the `rcrossref` package (Chamberlain et al. 2022) for R

Abstracts extraction

- We planed to scrap all papers' metadata from the journals' web-pages
- Then we discovered the `crossref` database and the `rcrossref` package (Chamberlain et al. 2022) for R
- Then we discovered the `openalex` database and the `openalexR` package (Aria and Le 2023) for R

Abstracts extraction: scrapping journal web-pages

Since journals can be grouped by the publishers it is possible to write a web-scrapper for the group of journals from one publisher. However there are some problems:

- blocking the IP address
- a lot of old publication are digitized in the form of a pdf of the first page of the paper
- change of the name and ISSN[2] of the journal
- change of the publisher of the journal
- there are still some mistakes on the journal websites

[2]ISSN (International Standard Serial Number) — eight-digit serial number used to uniquely identify a serial publication.

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclussions

Outline of the talk

Introduction

Our team

Data collection

Data analysis

Conclussions

Thank you for your attention!

References

Massimo Aria and Trang Le. *openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API*, 2023. URL https://CRAN.R-project.org/package=openalexR. R package version 1.0.2.9.

Scott Chamberlain, Hao Zhu, Najko Jahn, Carl Boettiger, and Karthik Ram. *rcrossref: Client for Various 'CrossRef' 'APIs'*, 2022. URL https://CRAN.R-project.org/package=rcrossref. R package version 1.2.0.

Rita Gonzalez-Marquez, Luca Schmidt, Benjamin M Schmidt, Philipp Berens, and Dmitry Kobak. The landscape of biomedical research. *bioRxiv*, 2023. doi: https://doi.org/10.1101/2023.04.10.536208.

References

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.