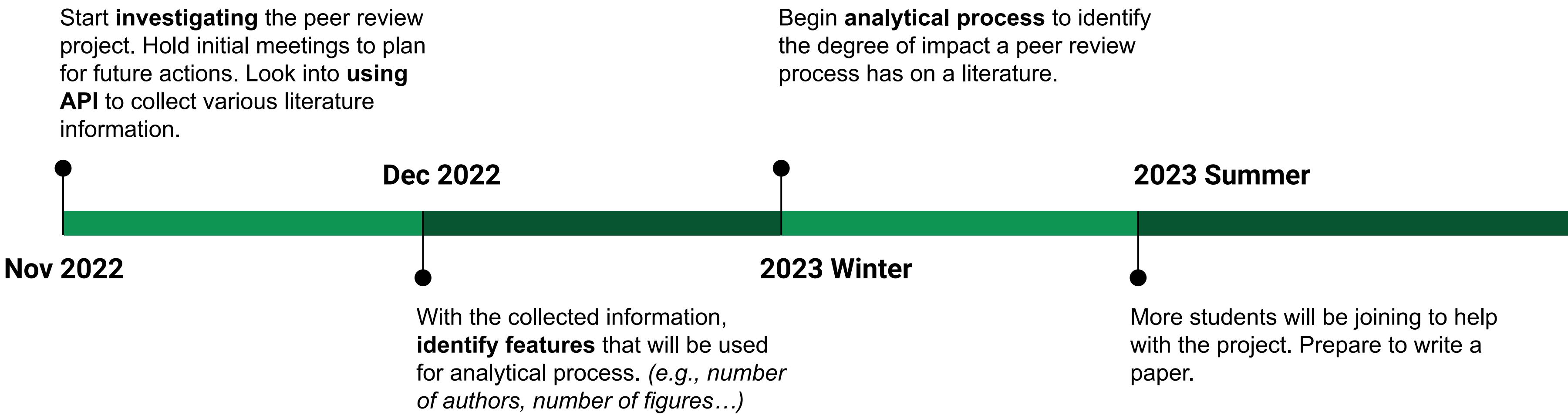
UNIVERSITY OF CALGARY

Peer Review Project 2022-2023

Jenny Lee, Albert Leung, Daniel Chang

🔗 GitHub Workflow

Progress



Nov 2022

Start **investigating** the peer review project. Hold initial meetings to plan for future actions. Look into **using API** to collect various literature information.

Dec 2022

With the collected information, **identify features** that will be used for analytical process. *(e.g., number of authors, number of figures…)*

2023 Winter

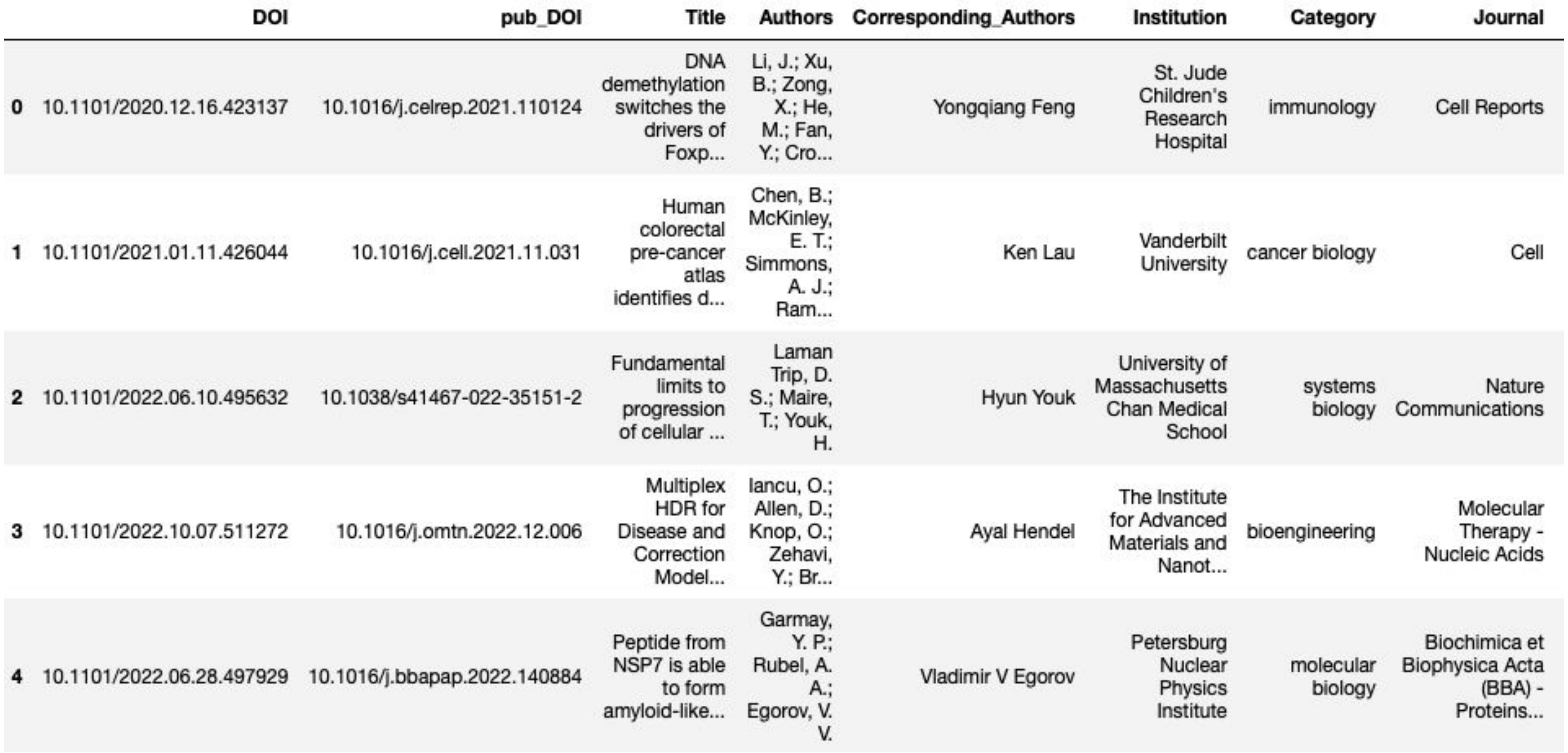
Begin **analytical process** to identify the degree of impact a peer review process has on a literature.

2023 Summer

More students will be joining to help with the project. Prepare to write a paper.

Biorxiv Query Duration: **2022/12/01 ~ 2022/12/31**

Step 1. Identify prepublished literatures with Biorxiv API



	DOI	pub_DOI	Title	Authors	Corresponding_Authors	Institution	Category	Journal
0	10.1101/2020.12.16.423137	10.1016/j.celrep.2021.110124	DNA demethylation switches the drivers of Foxp...	Li, J.; Xu, B.; Zong, X.; He, M.; Fan, Y.; Cro...	Yongqiang Feng	St. Jude Children's Research Hospital	immunology	Cell Reports
1	10.1101/2021.01.11.426044	10.1016/j.cell.2021.11.031	Human colorectal pre-cancer atlas identifies d...	Chen, B.; McKinley, E. T.; Simmons, A. J.; Ram...	Ken Lau	Vanderbilt University	cancer biology	Cell
2	10.1101/2022.06.10.495632	10.1038/s41467-022-35151-2	Fundamental limits to progression of cellular ...	Laman Trip, D. S.; Maire, T.; Youk, H.	Hyun Youk	University of Massachusetts Chan Medical School	systems biology	Nature Communications
3	10.1101/2022.10.07.511272	10.1016/j.omtn.2022.12.006	Multiplex HDR for Disease and Correction Model...	Iancu, O.; Allen, D.; Knop, O.; Zehavi, Y.; Br...	Ayal Hendel	The Institute for Advanced Materials and Nanot...	bioengineering	Molecular Therapy - Nucleic Acids
4	10.1101/2022.06.28.497929	10.1016/j.bbapap.2022.140884	Peptide from NSP7 is able to form amyloid-like...	Garmay, Y. P.; Rubel, A. A.; Egorov, V. V.	Vladimir V Egorov	Petersburg Nuclear Physics Institute	molecular biology	Biochimica et Biophysica Acta (BBA) - Proteins...

n: 175

Step 1. Identify prepublished literatures with Biorxiv API 🔗

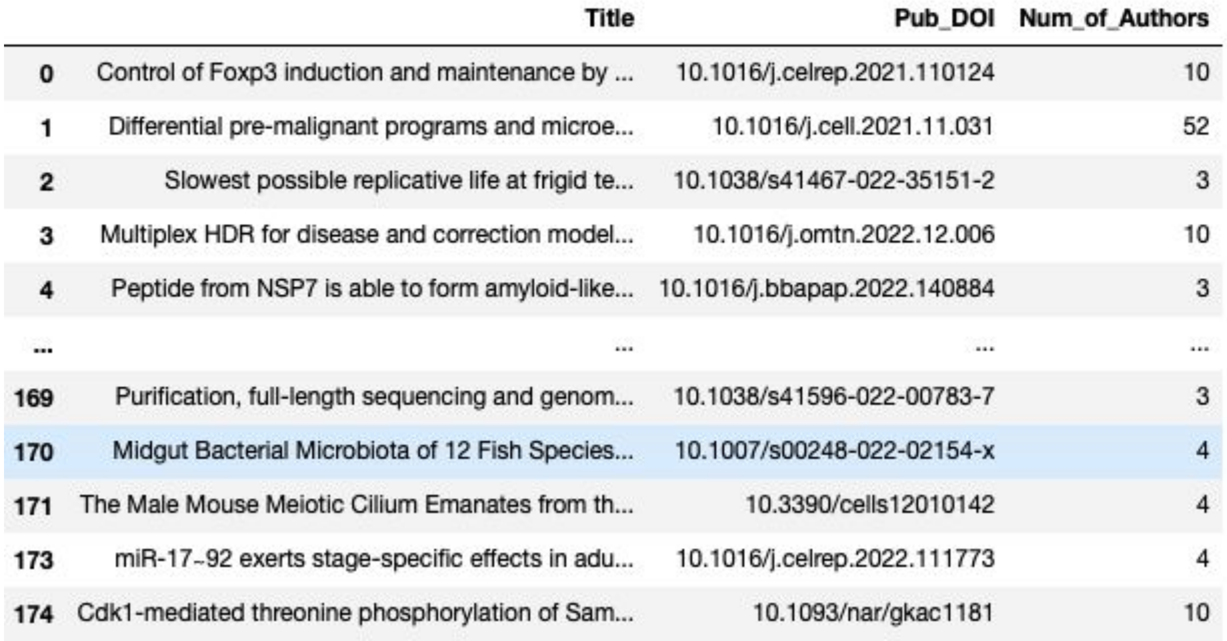
Retrievable Information

- DOI
- Published DOI
- Prepublished Title
- Authors
- Corresponding Authors
- Institution
- Date
- Version
- Type
- Category
- XML link

Retrieved Information

- DOI
- Published DOI
- Published Title
- Journal
- Institution
- Number of authors
- Category
- PDF Link

Step 2. Identify published literatures with PubMed & MetapubAPI



	Title	Pub_DOI	Num_of_Authors
0	Control of Foxp3 induction and maintenance by ...	10.1016/j.celrep.2021.110124	10
1	Differential pre-malignant programs and microe...	10.1016/j.cell.2021.11.031	52
2	Slowest possible replicative life at frigid te...	10.1038/s41467-022-35151-2	3
3	Multiplex HDR for disease and correction model...	10.1016/j.omtn.2022.12.006	10
4	Peptide from NSP7 is able to form amyloid-like...	10.1016/j.bbapap.2022.140884	3
...	...	...	...
169	Purification, full-length sequencing and genom...	10.1038/s41596-022-00783-7	3
170	Midgut Bacterial Microbiota of 12 Fish Species...	10.1007/s00248-022-02154-x	4
171	The Male Mouse Meiotic Cilium Emanates from th...	10.3390/cells12010142	4
173	miR-17~92 exerts stage-specific effects in adu...	10.1016/j.celrep.2022.111773	4
174	Cdk1-mediated threonine phosphorylation of Sam...	10.1093/nar/gkac1181	10

n: 170 (-5)

Step 2. Identify published literatures with PubMed & MetapubAPI 🔗

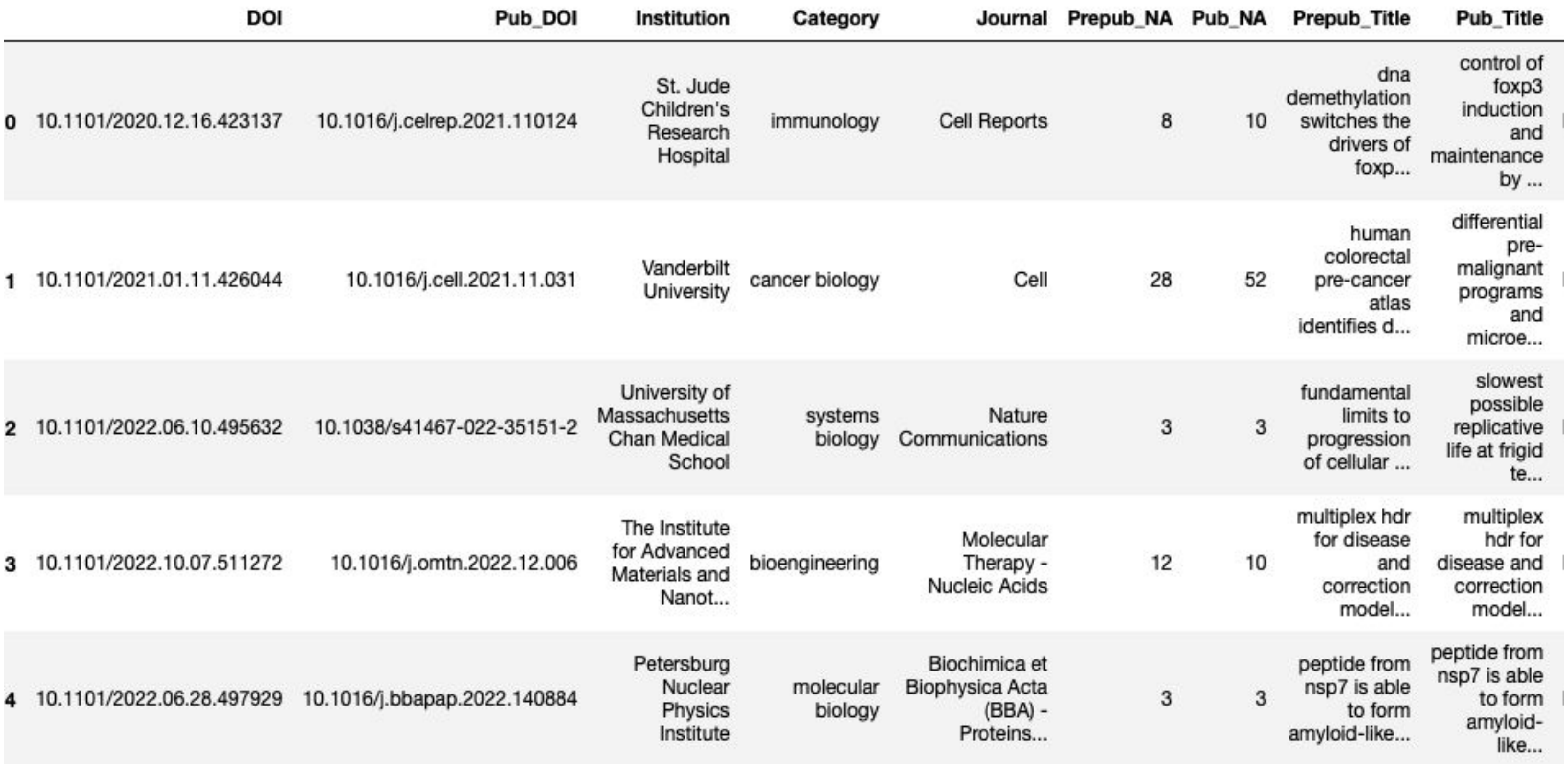
Retrieved Information from PubMed

- Published DOI
- Published Title
  - But is **unstable**, therefore not used.
- Number of authors
- Category

Retrieved Information from MetaPub

- Published DOI
- Published Title
- PDF Link

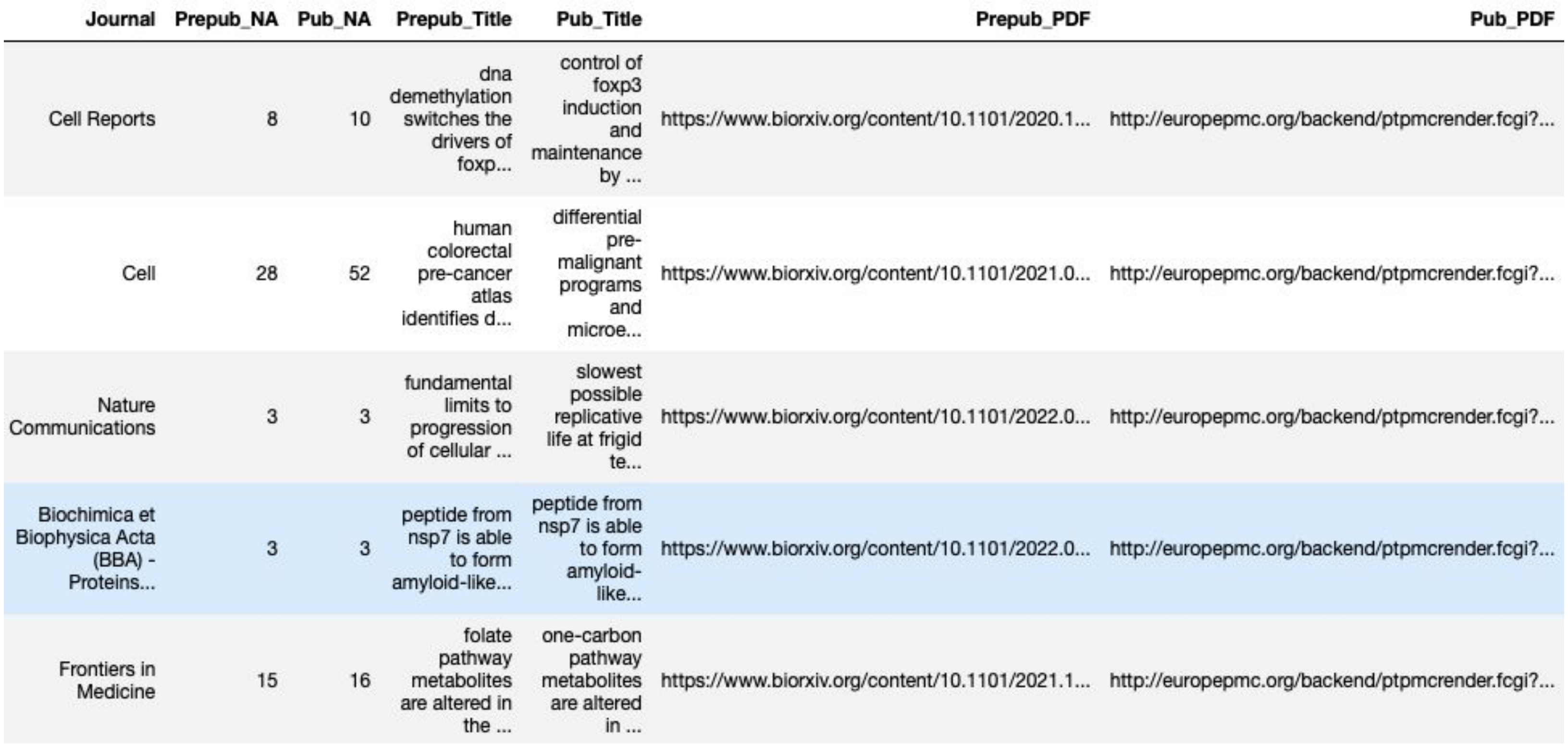
Step 3. Merge the outcomes from step 1 and 2



	DOI	Pub_DOI	Institution	Category	Journal	Prepub_NA	Pub_NA	Prepub_Title	Pub_Title
0	10.1101/2020.12.16.423137	10.1016/j.celrep.2021.110124	St. Jude Children's Research Hospital	immunology	Cell Reports	8	10	dna demethylation switches the drivers of foxp...	control of foxp3 induction and maintenance by ...
1	10.1101/2021.01.11.426044	10.1016/j.cell.2021.11.031	Vanderbilt University	cancer biology	Cell	28	52	human colorectal pre-cancer atlas identifies d...	differential pre-malignant programs and microe...
2	10.1101/2022.06.10.495632	10.1038/s41467-022-35151-2	University of Massachusetts Chan Medical School	systems biology	Nature Communications	3	3	fundamental limits to progression of cellular ...	slowest possible replicative life at frigid te...
3	10.1101/2022.10.07.511272	10.1016/j.omtn.2022.12.006	The Institute for Advanced Materials and Nanot...	bioengineering	Molecular Therapy - Nucleic Acids	12	10	multiplex hdr for disease and correction model...	multiplex hdr for disease and correction model...
4	10.1101/2022.06.28.497929	10.1016/j.bbapap.2022.140884	Petersburg Nuclear Physics Institute	molecular biology	Biochimica et Biophysica Acta (BBA) - Proteins...	3	3	peptide from nsp7 is able to form amyloid-like...	peptide from nsp7 is able to form amyloid-like...

n: 170 (-5)

Step 4. Retrieve PDF link for published article using Metapub API



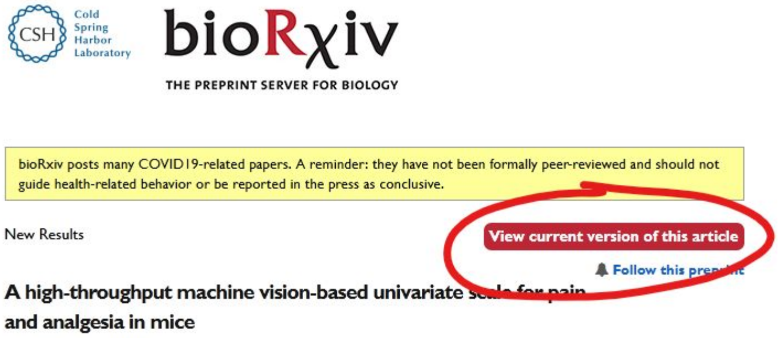
Journal	Prepub_NA	Pub_NA	Prepub_Title	Pub_Title	Prepub_PDF	Pub_PDF
Cell Reports	8	10	dna demethylation switches the drivers of foxp...	control of foxp3 induction and maintenance by ...	https://www.biorxiv.org/content/10.1101/2020.1...	http://europepmc.org/backend/ptpmcrender.fcgi?...
Cell	28	52	human colorectal pre-cancer atlas identifies d...	differential pre-malignant programs and microe...	https://www.biorxiv.org/content/10.1101/2021.0...	http://europepmc.org/backend/ptpmcrender.fcgi?...
Nature Communications	3	3	fundamental limits to progression of cellular ...	slowest possible replicative life at frigid te...	https://www.biorxiv.org/content/10.1101/2022.0...	http://europepmc.org/backend/ptpmcrender.fcgi?...
Biochimica et Biophysica Acta (BBA) - Proteins...	3	3	peptide from nsp7 is able to form amyloid-like...	peptide from nsp7 is able to form amyloid-like...	https://www.biorxiv.org/content/10.1101/2022.0...	http://europepmc.org/backend/ptpmcrender.fcgi?...
Frontiers in Medicine	15	16	folate pathway metabolites are altered in the ...	one-carbon pathway metabolites are altered in ...	https://www.biorxiv.org/content/10.1101/2021.1...	http://europepmc.org/backend/ptpmcrender.fcgi?...

n: 83 (-87)

Step 4: Retrieve PDF link for published article using web browsing

Suggested idea:

https://www.biorxiv.org/content/10.1101/2022.12.29.522204v1
https://www.biorxiv.org/content/10.1101/2022.12.29.522204v2



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

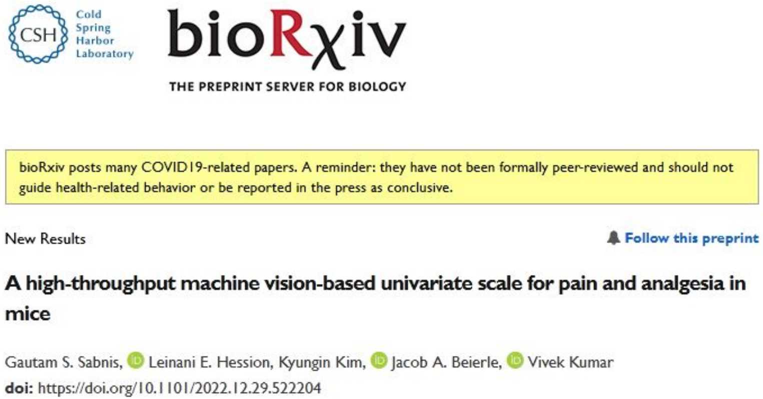
bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

View current version of this article

Follow this preprint

A high-throughput machine vision-based univariate scale for pain and analgesia in mice



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Follow this preprint

A high-throughput machine vision-based univariate scale for pain and analgesia in mice

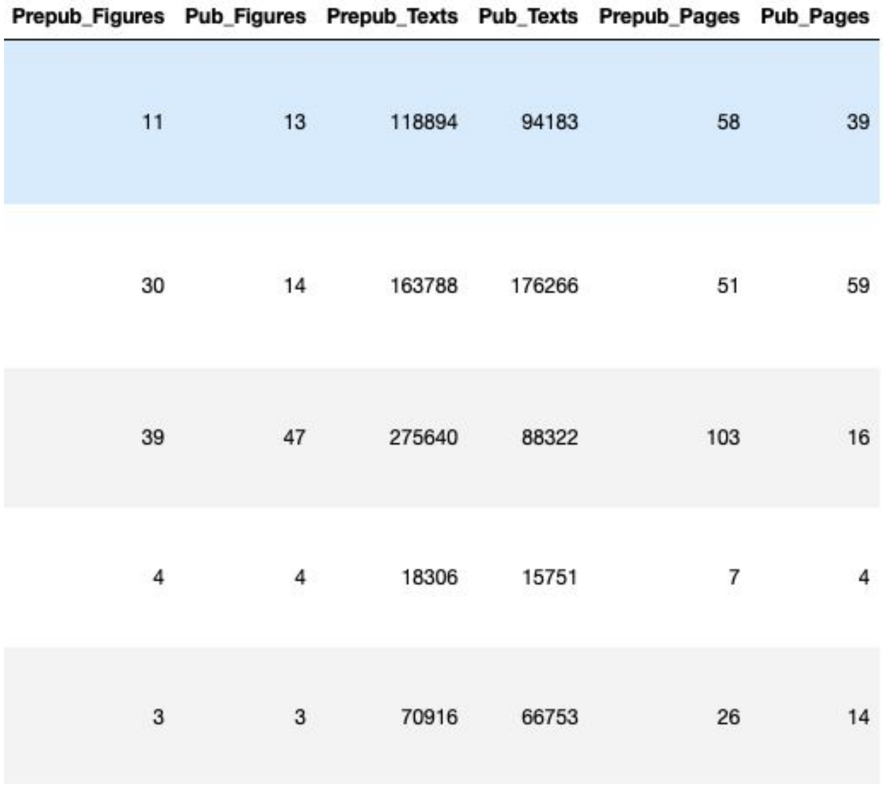
Gautam S. Sabnis, Leinani E. Hession, Kyungin Kim, Jacob A. Beierle, Vivek Kumar

doi: https://doi.org/10.1101/2022.12.29.522204

- No success in finding the link to the published article through web browsing
- Use API which is more reliable

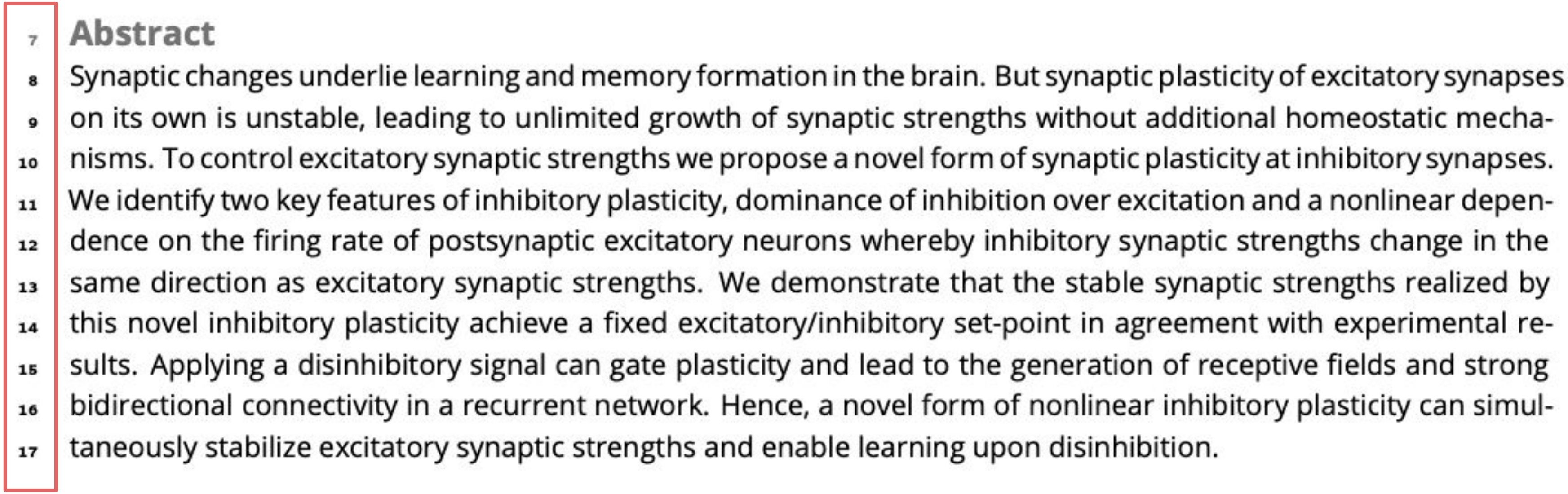
Step 5. Analyze PDF

- **Number of Figures & Texts & Pages:**
    - Using PDF parsing (fitz & PyPDF)
        - Reads all PDF into text strings



Prepub_Figures	Pub_Figures	Prepub_Texts	Pub_Texts	Prepub_Pages	Pub_Pages
11	13	118894	94183	58	39
30	14	163788	176266	51	59
39	47	275640	88322	103	16
4	4	18306	15751	7	4
3	3	70916	66753	26	14

Step 5. Concerns with PDF Parsing



7 **Abstract**
8 Synaptic changes underlie learning and memory formation in the brain. But synaptic plasticity of excitatory synapses
9 on its own is unstable, leading to unlimited growth of synaptic strengths without additional homeostatic mecha-
10 nisms. To control excitatory synaptic strengths we propose a novel form of synaptic plasticity at inhibitory synapses.
11 We identify two key features of inhibitory plasticity, dominance of inhibition over excitation and a nonlinear depen-
12 dence on the firing rate of postsynaptic excitatory neurons whereby inhibitory synaptic strengths change in the
13 same direction as excitatory synaptic strengths. We demonstrate that the stable synaptic strengths realized by
14 this novel inhibitory plasticity achieve a fixed excitatory/inhibitory set-point in agreement with experimental re-
15 sults. Applying a disinhibitory signal can gate plasticity and lead to the generation of receptive fields and strong
16 bidirectional connectivity in a recurrent network. Hence, a novel form of nonlinear inhibitory plasticity can simul-
17 taneously stabilize excitatory synaptic strengths and enable learning upon disinhibition.

Step 5. Available Features

- Change in **article title**
- Change in **number of authors**
- Change in **number of figures**
- Change in **total number of characters**
- Change in **number of pages**

Of **175** queried Biorvix articles, **73 articles** have been fully analyzed.

Comparison to the **Previous Version**

Old Version (Daniel's Codes)

- Used selenium.webdriver to extract article information from the U of C database.
- Did not filter out unpublished articles first.
- PDF files were downloaded for further analysis.
- PDF analysis was done in R.
  - Number of figures in the article were counted by pdf_extractImages built-in function in metagear package.
  - This built-in function is ***not accurate***.

Current Version (2022)

- Uses PyMed and Metapub API to retrieve article information.
- Filters out published articles at the beginning of the workflow.
- PDF files are viewed in URL format; no need to download the files.
- PDF analysis was done with fitz and PyPDF2 packages in Python.
  - Number of figures in the article are found with NLP (regex).
  - Also not 100% accurate, but more accurate than before.

Current Concerns

- Not all literatures are in **PubMed database**.
    - Among the first **175** literatures in the sample dataset, **5** literatures were missing from the database.
- Not all literatures have **open PDF access**.
    - Out of **170** literatures, **87** literatures did not have open PDF access.
- Takes **long time** to run through the process.
    - The entire process took about 3200 seconds (53.3 minutes) to run.

- **Possible solution**
    - Use **web scraping** (from Daniel's codes) to retrieve data from UBC / U of C engine.
        - However, this would require us to *download* all PDF files.
    - Use **multi-processing** to reduce run time.

Discussions

- Decide on what **features** to be used for further processes.
    - What are we going to compare between published and prepublished articles?
        - Number of authors, number of figures, words.
- Decide on **sample sizes**.
    - How many articles are we going to sample?
    - If sample size is small (< 200), we could consider downloading all PDF files for analysis.
- Decide on the **field of studies** to focus on.
    - Currently, we are only using prepublished information from the Biorxiv API. Are we planning to look at any other field of studies? (*E.g.*, sociology)
        - Countries, mathematics vs. biology (e.g. **STEM fields**).
- Decide on **methods of machine learning** for exploratory & descriptive analysis.

Future Actions

- **Albert:** Redesign some parts of the code to speed up the process. Verify codes to see if they work well without errors.

- **Daniel:** Help with analysing PDF files. We would potentially like to look at the number of characters by each section of an article (*E.g.*, Introduction, Methods…)

Future Actions: PDF Text Analysis

- PDF to text
    - OCR with Tesseract-OCR (https://github.com/tesseract-ocr/tesseract)
- Text Analysis: some suggested methods
    - Word frequencies
    - Sequence matching (dynamic programming)
    - Sentence similarity with SBERT

Thank you for listening!

🔗 GitHub Workflow