eLIFE

Open-Source Data Science Projects at eLife

Daniel Ecer
22th November 2019

eLife

Open-Source Data Science Projects at eLife

What is eLife?

- A non-profit backed by **research funders** to drive reform in research communication
- We invest heavily in open-source technology development and innovation **on behalf of the community**

LIBERO

eLIFE

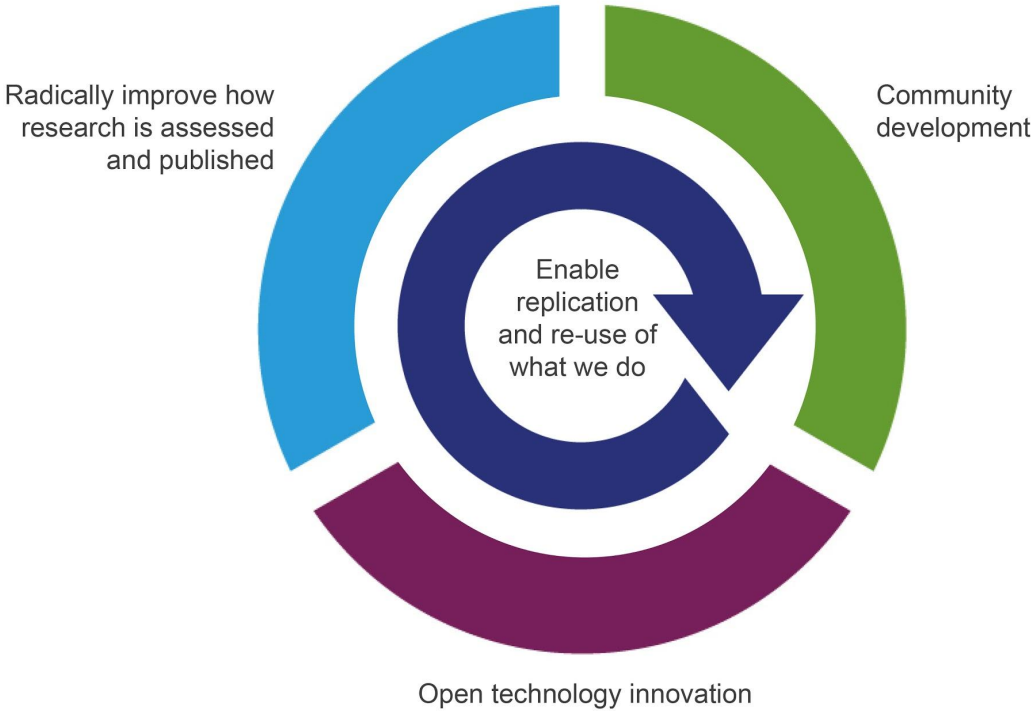

Helping scientists accelerate discovery by
operating a platform for research communication
that encourages and recognises the most
responsible behaviours in science.

eLife’s motivations

- Leverage the **power of web technology** to accelerate research and discovery
- **Support open-access** publishing
- Build a **community-owned infrastructure** for research communication

LIBERO

eLIFE


Radically improve how research is assessed and published
Community development
Enable replication and re-use of what we do
Open technology innovation

Data Science projects at eLife

Amy
Author

- Amy submits a manuscript to a journal, maybe not the first attempt
- Amy would like to reduce form filling
- **ScienceBeam** helps pre-populating fields
- **ScienceBeam** may also help to add a semantic structure early on, to make the manuscript more “accessible”

Hubert
Reviewing Editor / Editorial Staff

- Hubert has been assigned, as the reviewing editor
- Hubert now needs to assign reviewers
- **PeerScout** helps to reduce bias and extend pool to find peers at different stages (editors, reviewers etc)

Philip
Reader

- Philip wants to stay on top of his field of interest
- There are many published manuscripts
- With **PeerTax**, Philip can better incorporate the views of reviewers, when reading manuscripts
- While **Citation Sentiment** provides a better picture of why manuscripts were cited

ScienceBeam

Open-Source Data Science Projects at eLife

Add semantics to PDF articles

Foit MobilePDF

eLIFE elifesciences.org

RESEARCH ARTICLE

Enhanced ER proteostasis and temperature differentially impact the mutational tolerance of influenza hemagglutinin

Angela M Phillips[1], Michael B Doud[2,3], Luna O Gonzalez[4], Vincent L Butty[5], Yu-Shan Lin[6], Jesse D Bloom[2,3], Matthew D Shoulders[1]*

[1]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, United States; [2]Fred Hutchinson Cancer Research Center, Seattle, United States; [3]Department of Genome Sciences, University of Washington, Seattle, United States; [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, United States; [5]BioMicro Center, Massachusetts Institute of Technology, Cambridge, United States; [6]Department of Chemistry, Tufts University, Medford, United States

Abstract We systematically and quantitatively evaluate whether endoplasmic reticulum (ER) proteostasis factors impact the mutational tolerance of secretory pathway proteins. We focus on influenza hemagglutinin (HA), a viral membrane protein that folds in the host's ER via a complex pathway. By integrating chemical methods to modulate ER proteostasis with deep mutational scanning to assess mutational tolerance, we discover that upregulation of ER proteostasis factors broadly enhances HA mutational tolerance across diverse structural elements. Remarkably, this proteostasis network-enhanced mutational tolerance occurs at the same sites where mutational tolerance is most reduced by propagation at fever-like temperature. These findings have important implications for influenza evolution, because influenza immune escape is contingent on HA possessing sufficient mutational tolerance to evade antibodies while maintaining the capacity to fold and function. More broadly, this work provides the first experimental evidence that ER proteostasis mechanisms define the mutational tolerance and, therefore, the evolution of secretory pathway proteins.

DOI: https://doi.org/10.7554/eLife.38795.001

***For correspondence:** mshoulde@mit.edu

Competing interests: The authors declare that no competing interests exist.

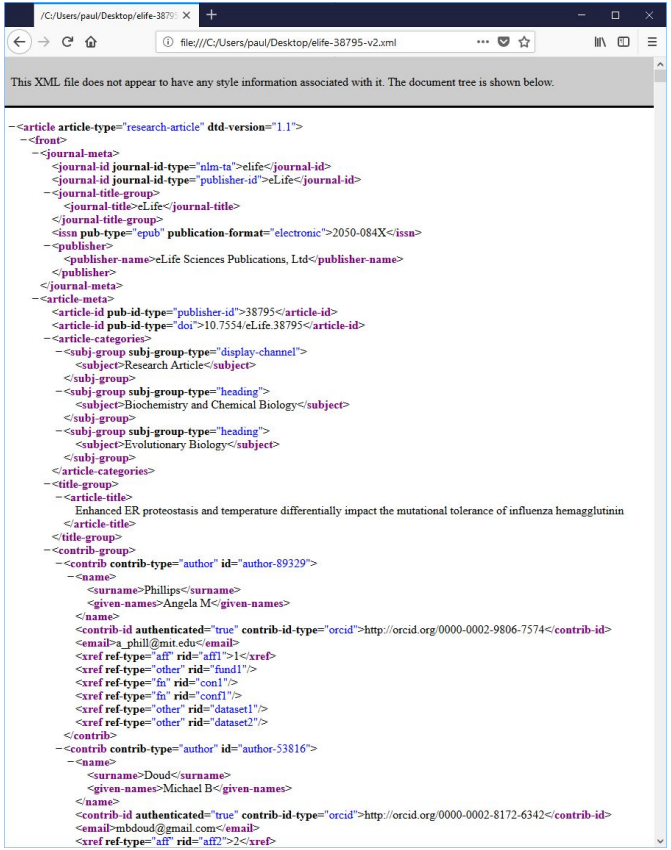
Funding: See page 20

Received: 06 June 2018
Accepted: 29 August 2018
Published: 06 September 2018

Reviewing editor: Franz-Ulrich Hartl, Max Planck Institute for Biochemistry, Germany

Introduction

Protein evolution is necessarily constrained by the inherent biophysical properties of polypeptide sequences (*Wylie and Shakhnovich, 2011*; *Wyganowski et al., 2013*; *Bloom and Glassman, 2009*; *DePristo et al., 2005*; *Tokuriki and Tawfik, 2009a*; *Gong et al., 2013*). Gene variants that encode proteins unable to properly fold are selected against (*Figure 1A*) (*Powers et al., 2009*), irrespective of any theoretically beneficial new function the encoded protein may otherwise have acquired (*Wylie and Shakhnovich, 2011*; *Wyganowski et al., 2013*; *Bloom and Glassman, 2009*; *DePristo et al., 2005*; *Tokuriki and Tawfik, 2009a*; *Gong et al., 2013*). Importantly, the consequences of amino acid substitutions for protein folding depend not just on inherent biophysical properties, but also on the environment in which the nascent chain attempts to fold. In cells, this environment is defined most importantly by the proteostasis network, which consists of a subcellular compartment-specific array of chaperones and quality control factors (*Balch et al., 2008*; *Wong et al., 2018*; *Hartl et al., 2011*; *Shoulders et al., 2013*). These proteostasis factors assist



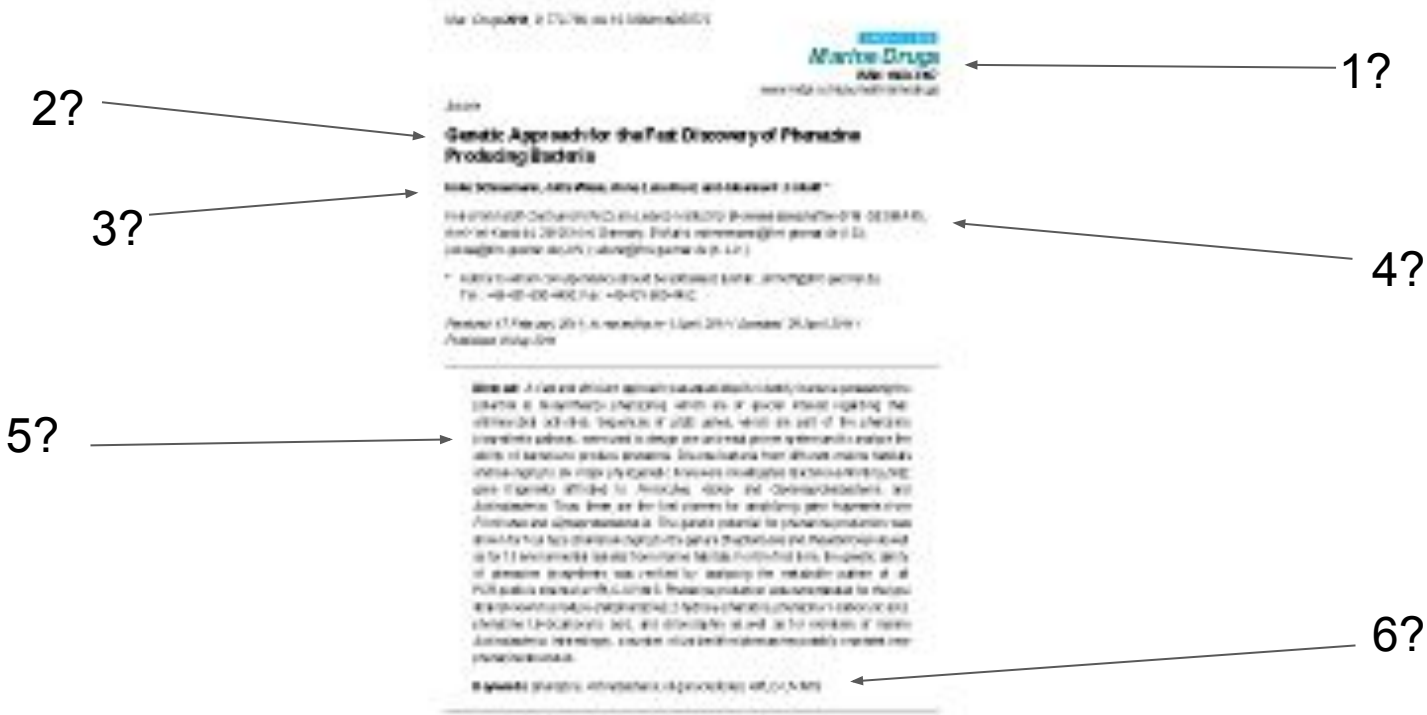
/C:/Users/paul/Desktop/elife-38795-v2.xml

file:///C:/Users/paul/Desktop/elife-38795-v2.xml

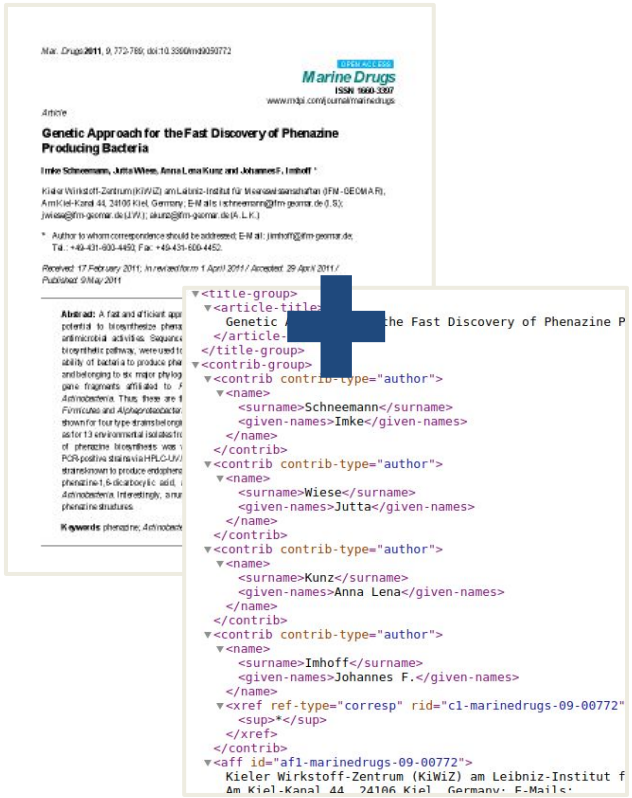
This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
−<article article-type="research-article" dtd-version="1.1">
 −<front>
  −<journal-meta>
    <journal-id journal-id-type="nlm-ta">elife</journal-id>
    <journal-id journal-id-type="publisher-id">eLife</journal-id>
   −<journal-title-group>
      <journal-title>eLife</journal-title>
    </journal-title-group>
    <issn pub-type="epub" publication-format="electronic">2050-084X</issn>
   −<publisher>
      <publisher-name>eLife Sciences Publications, Ltd</publisher-name>
    </publisher>
  </journal-meta>
 −<article-meta>
    <article-id pub-id-type="publisher-id">38795</article-id>
    <article-id pub-id-type="doi">10.7554/eLife.38795</article-id>
   −<article-categories>
     −<subj-group subj-group-type="display-channel">
        <subject>Research Article</subject>
      </subj-group>
     −<subj-group subj-group-type="heading">
        <subject>Biochemistry and Chemical Biology</subject>
      </subj-group>
     −<subj-group subj-group-type="heading">
        <subject>Evolutionary Biology</subject>
      </subj-group>
    </article-categories>
   −<title-group>
     −<article-title>
        Enhanced ER proteostasis and temperature differentially impact the mutational tolerance of influenza hemagglutinin
      </article-title>
    </title-group>
   −<contrib-group>
     −<contrib contrib-type="author" id="author-89329">
       −<name>
          <surname>Phillips</surname>
          <given-names>Angela M</given-names>
        </name>
        <contrib-id authenticated="true" contrib-id-type="orcid">http://orcid.org/0000-0002-9806-7574</contrib-id>
        <email>a_phill@mit.edu</email>
        <xref ref-type="aff" rid="aff1">1</xref>
        <xref ref-type="other" rid="fund1"/>
        <xref ref-type="fn" rid="con1"/>
        <xref ref-type="fn" rid="conf1"/>
        <xref ref-type="other" rid="dataset1"/>
        <xref ref-type="other" rid="dataset2"/>
      </contrib>
     −<contrib contrib-type="author" id="author-53816">
       −<name>
          <surname>Doud</surname>
          <given-names>Michael B</given-names>
        </name>
        <contrib-id authenticated="true" contrib-id-type="orcid">http://orcid.org/0000-0002-8172-6342</contrib-id>
        <email>mbdoud@gmail.com</email>
        <xref ref-type="aff" rid="aff2">2</xref>
```

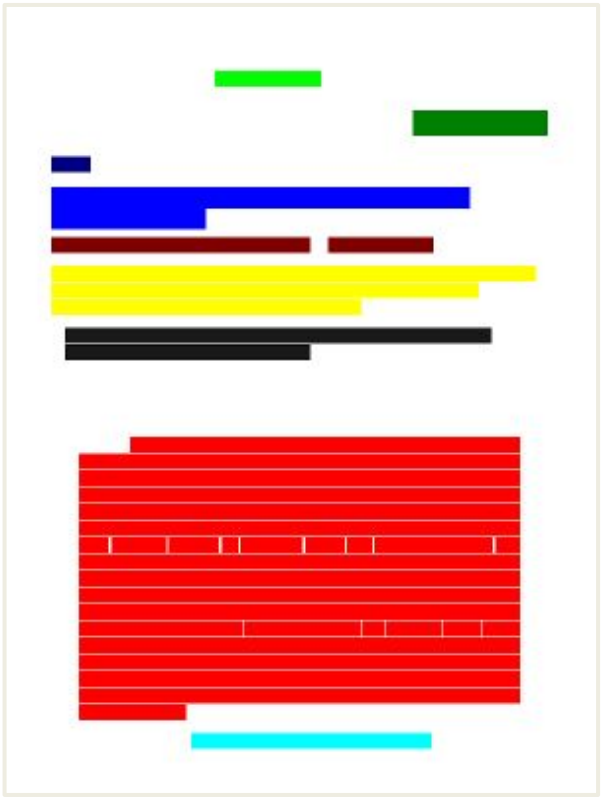
Can you guess...?



Generate training data from PDF & XML



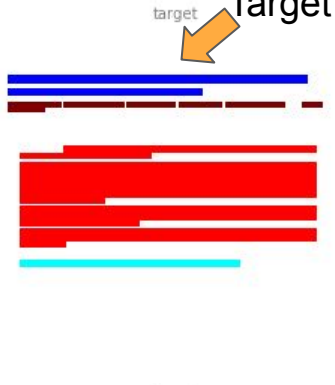
Fuzzy match



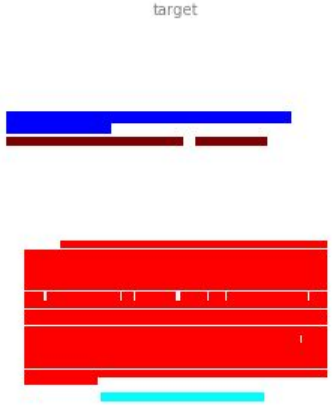
CV model qualitative results

Input (PDF page)

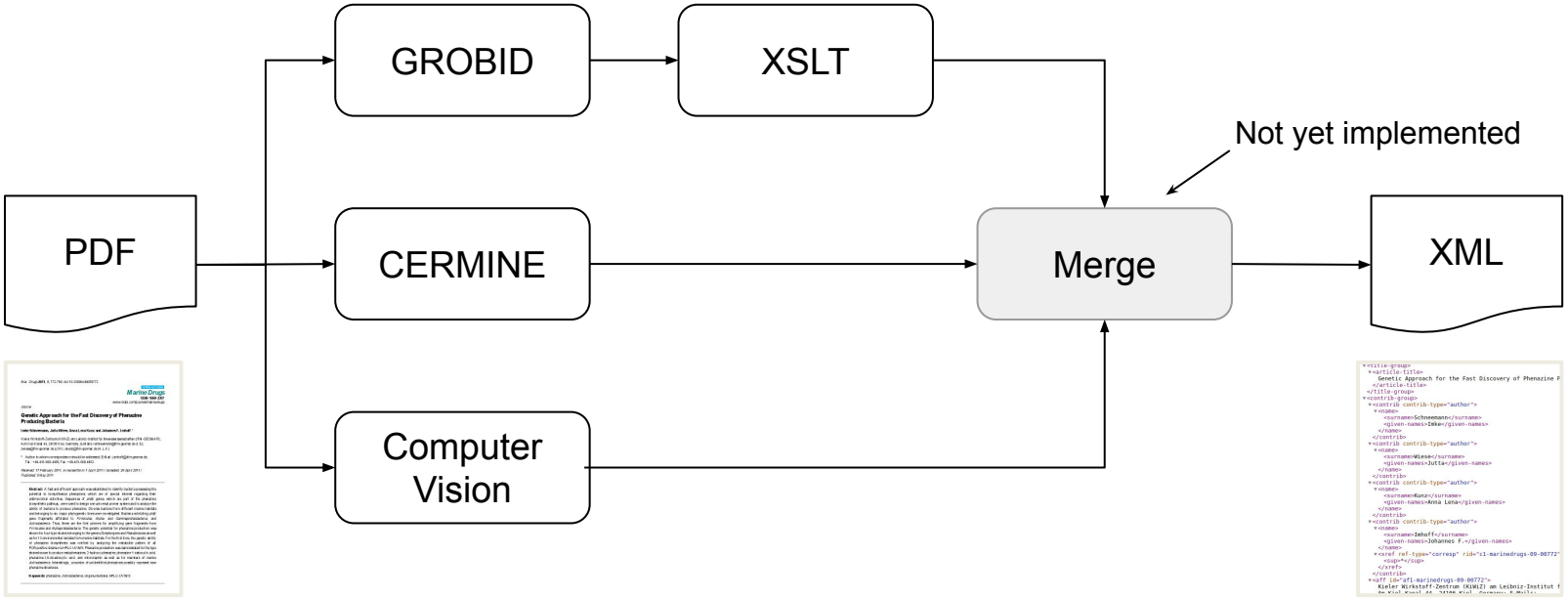
Target (expected)



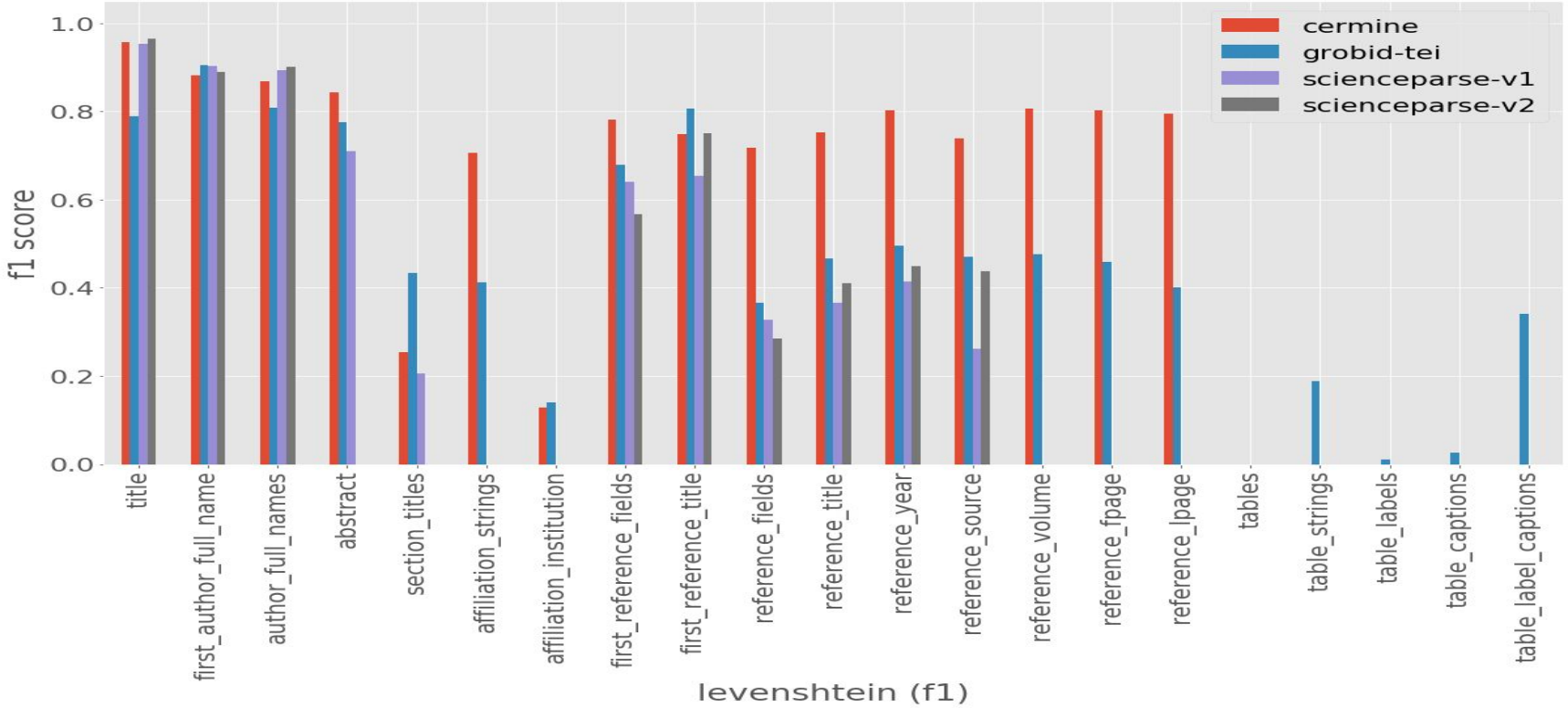
Prediction (actual)



ScienceBeam - potential pipeline

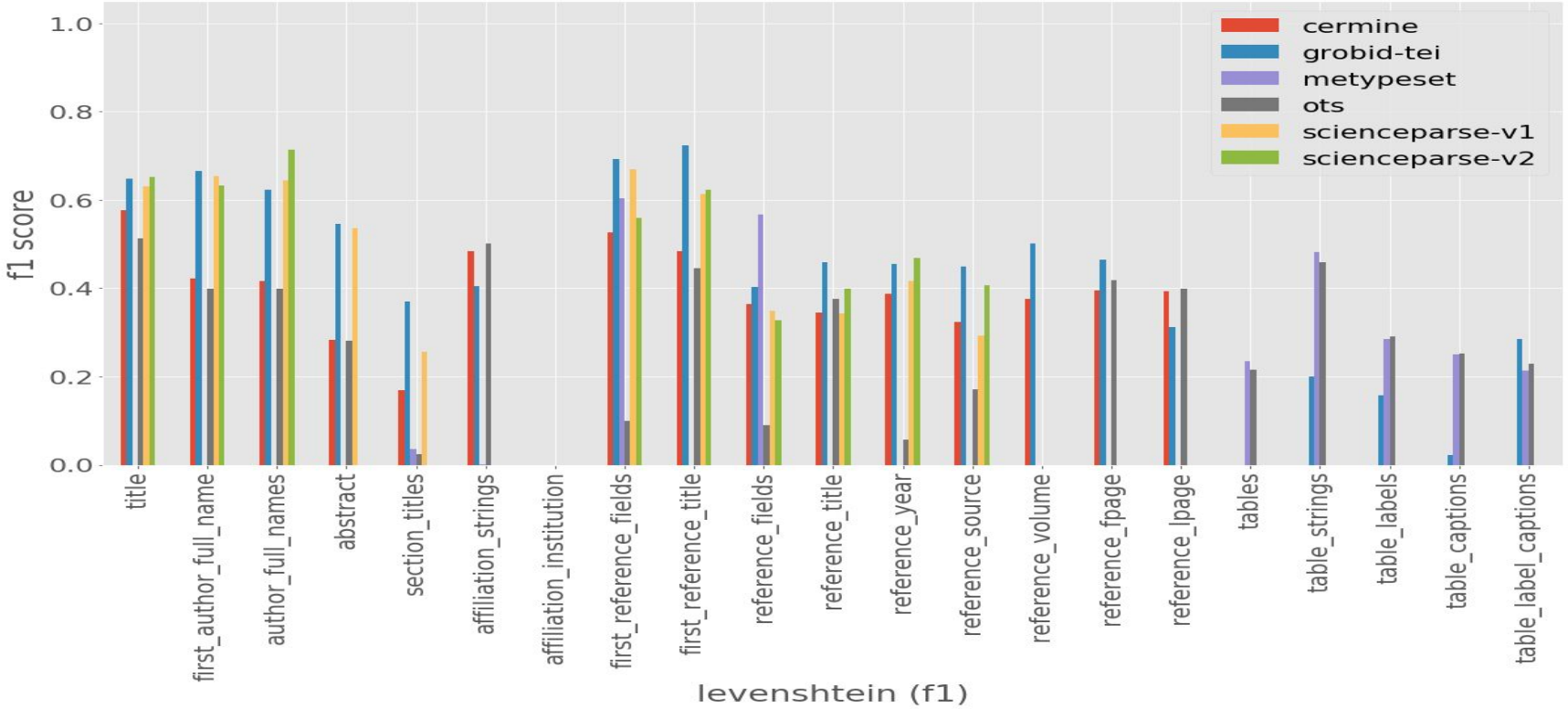


Evaluation - PMC_sample_1943

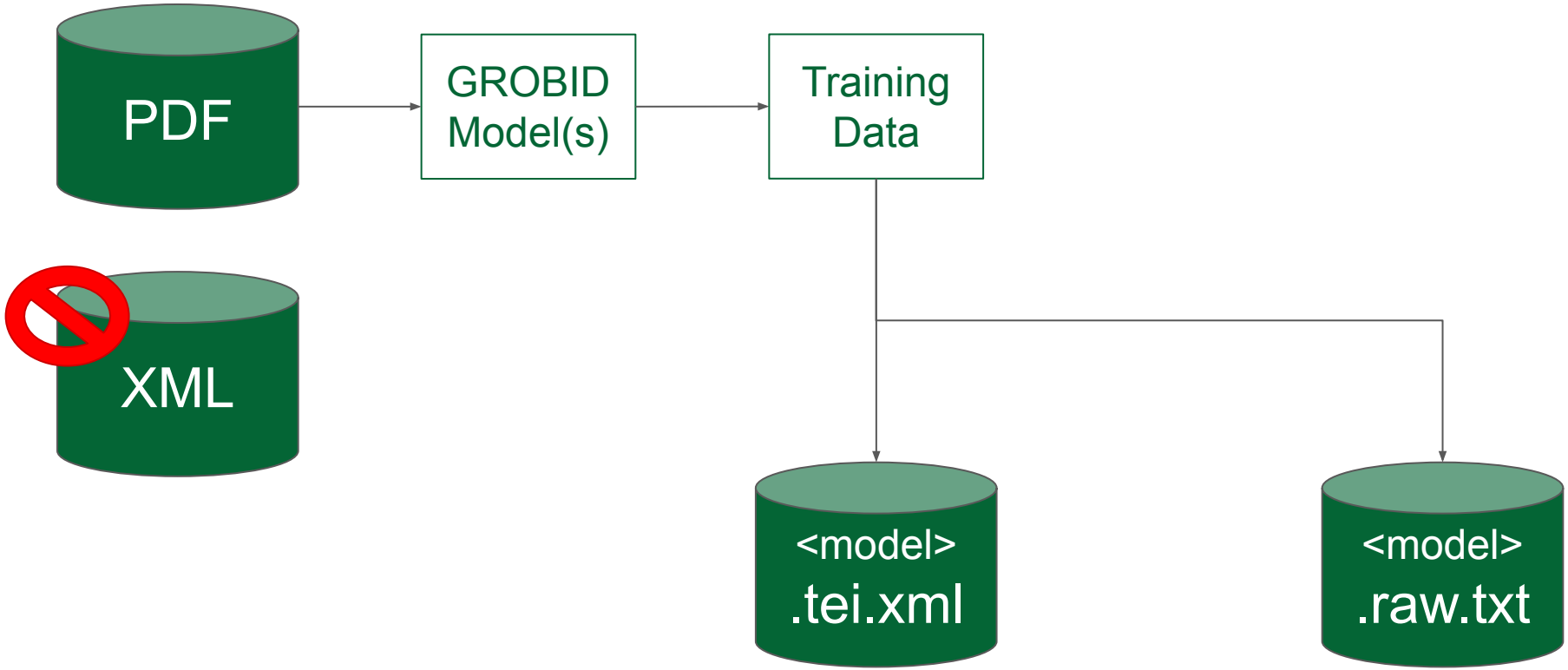


*CERMINE and ScienceParse are trained on PMC manuscripts

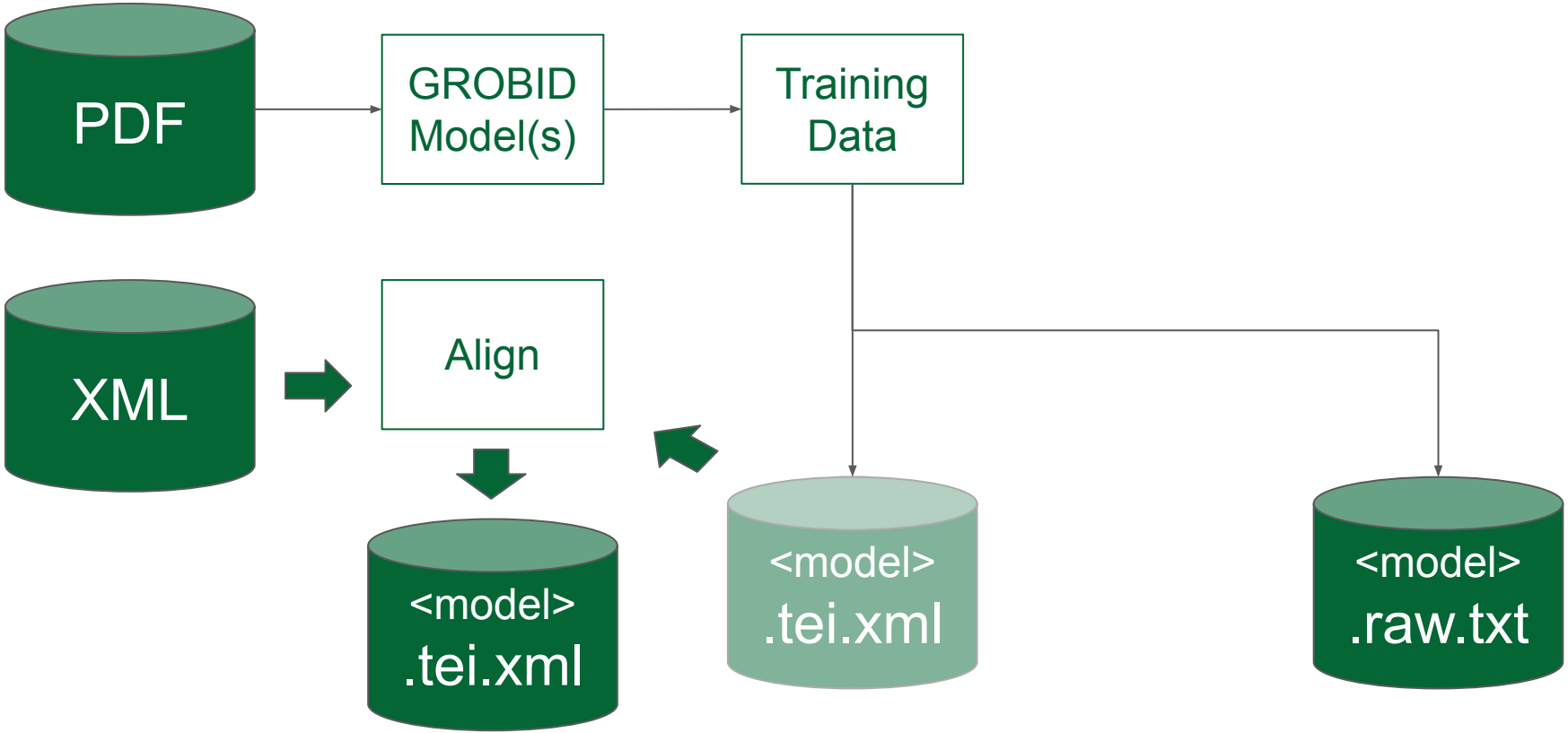
Evaluation - PKP coaction (Word)



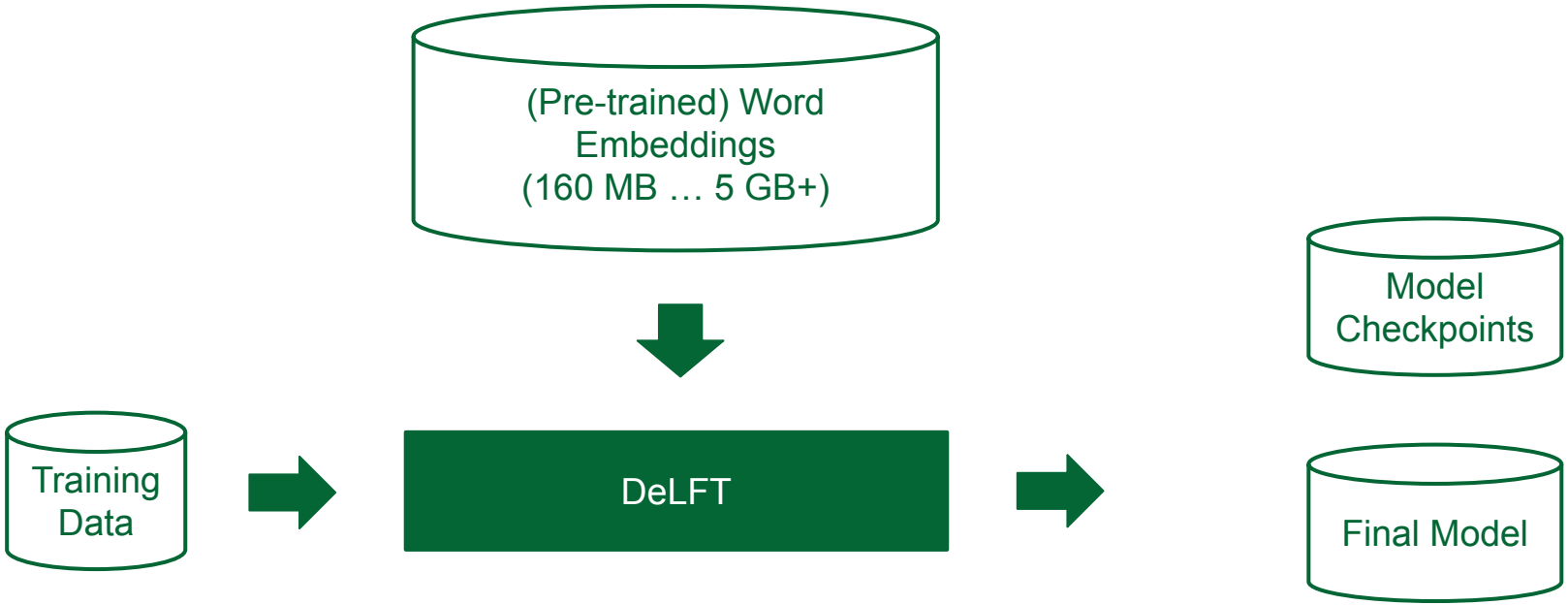
General GROBID Training Data Generation



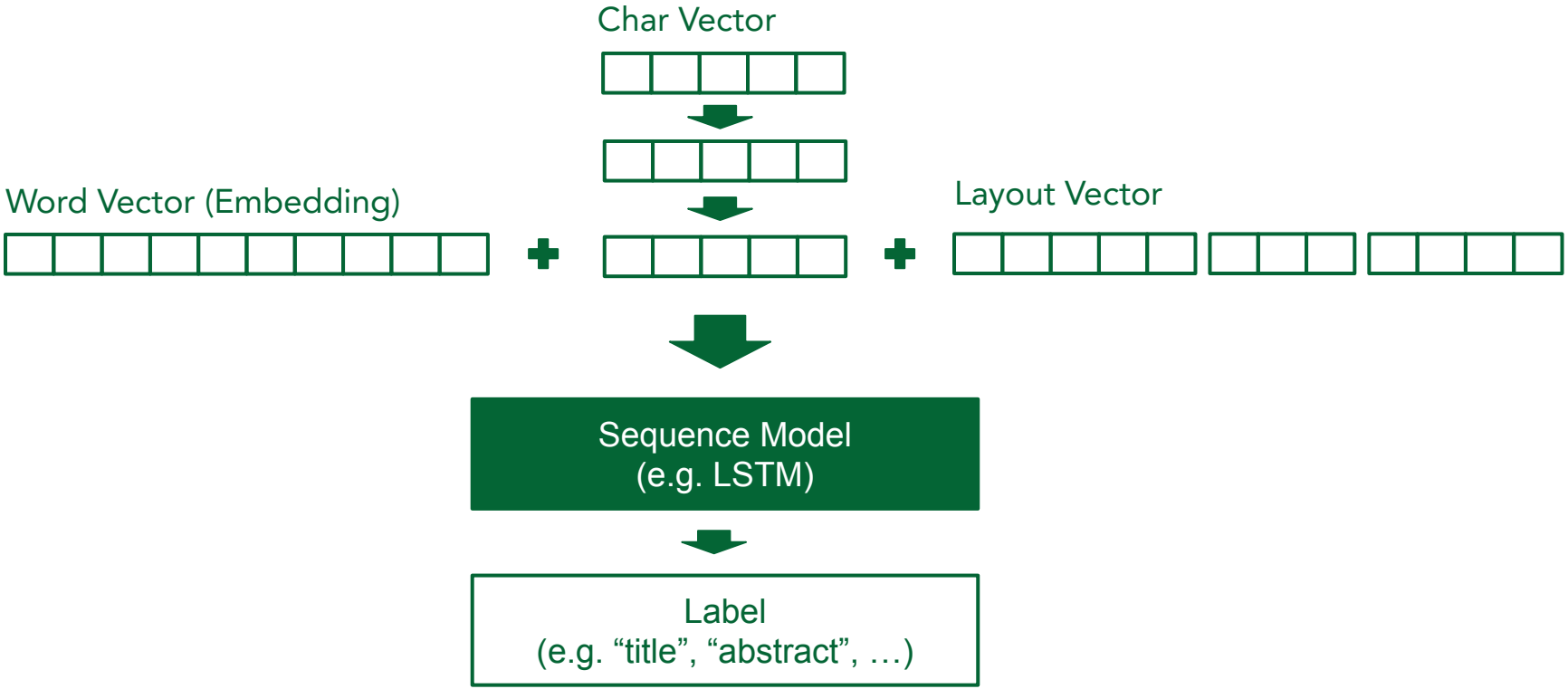
Use (JATS) XML to auto-annotate TEI training XML



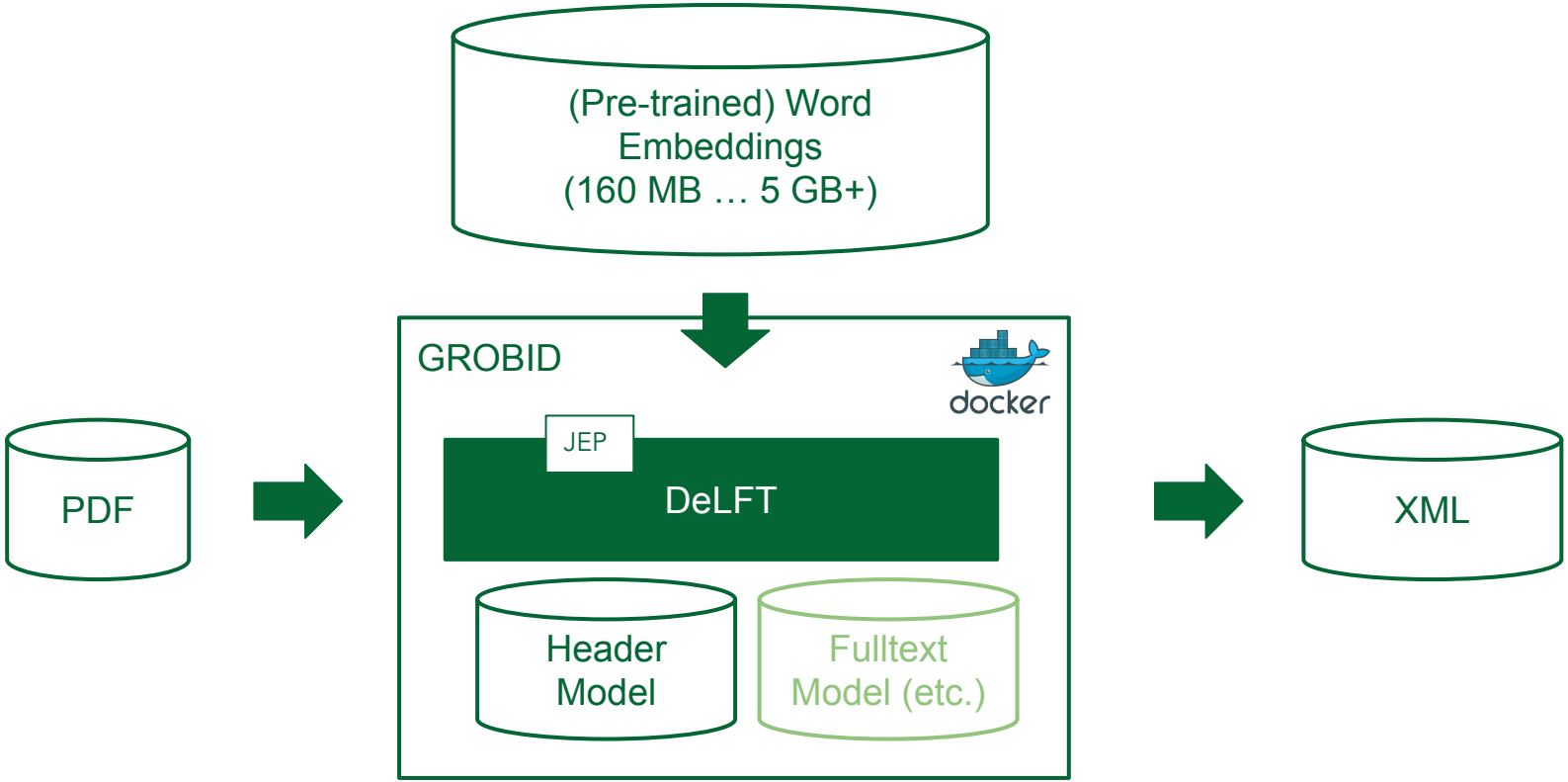
Training DeLFT Model



Adding layout features

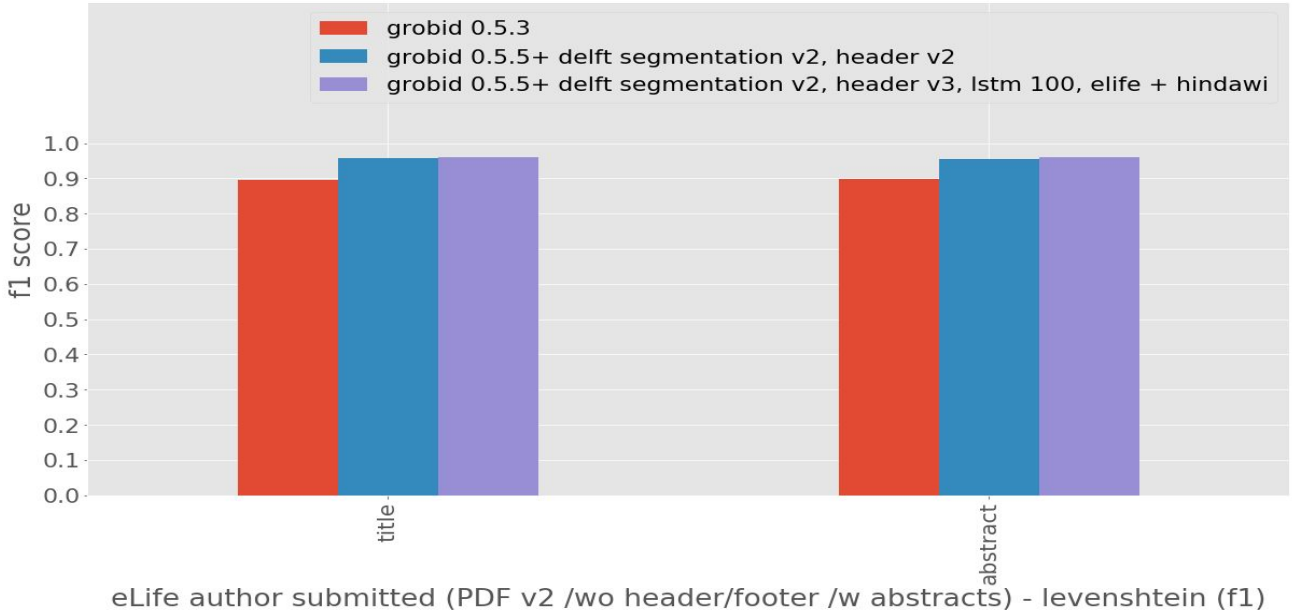


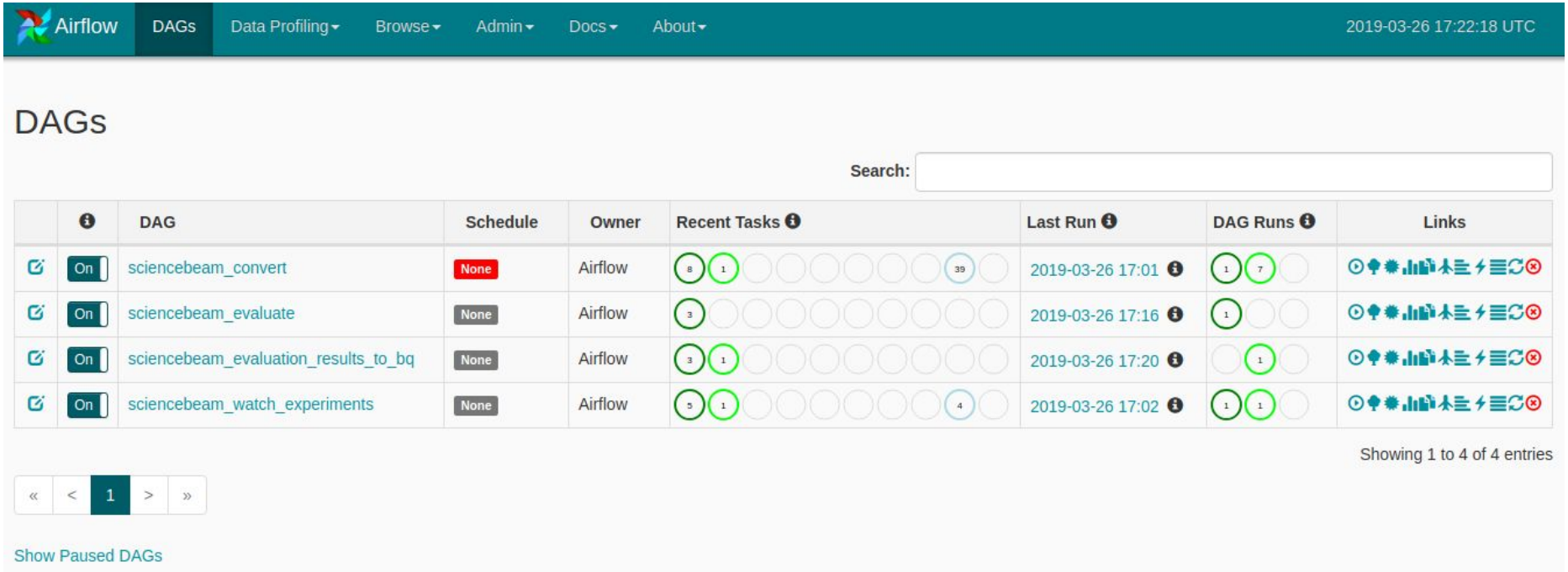
GROBID with DeLFT (in-progress)

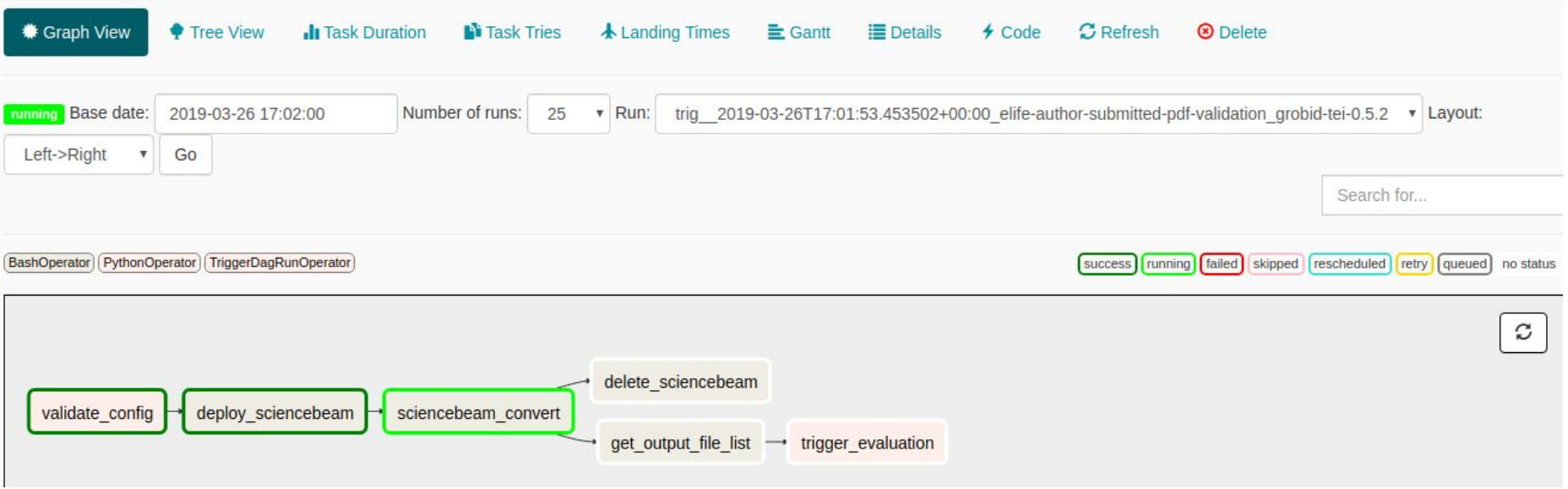


Selected models, /wo header/footer, with abstract only

	grobid 0.5.3	grobid 0.5.5+ delft segmentation v2, header v2	grobid 0.5.5+ delft segmentation v2, header v3, lstm 100, elife + hindawi
title	0.895225	0.957947	**0.960798**
abstract	0.899202	0.954903	**0.961373**




Airflow
DAGs
Data Profiling
Browse
Admin
Docs
About
2019-03-26 17:22:18 UTC
DAGs
Search:
DAG
Schedule
Owner
Recent Tasks
Last Run
DAG Runs
Links
On
sciencebeam_convert
None
Airflow
2019-03-26 17:01
On
sciencebeam_evaluate
None
Airflow
2019-03-26 17:16
On
sciencebeam_evaluation_results_to_bq
None
Airflow
2019-03-26 17:20
On
sciencebeam_watch_experiments
None
Airflow
2019-03-26 17:02
Showing 1 to 4 of 4 entries
«
<
1
>
»
Show Paused DAGs


Graph View
Tree View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Refresh
Delete
running
Base date:
2019-03-26 17:02:00
Number of runs:
25
Run:
trig__2019-03-26T17:01:53.453502+00:00_elife-author-submitted-pdf-validation_grobid-tei-0.5.2
Layout:
Left->Right
Go
Search for...
BashOperator
PythonOperator
TriggerDagRunOperator
success
running
failed
skipped
rescheduled
retry
queued
no status
validate_config
deploy_sciencebeam
sciencebeam_convert
delete_sciencebeam
get_output_file_list
trigger_evaluation


¿y como generamos XML?
Semantic Extraction Working Group
eLIFE
SciELO
PKP
PUBLIC KNOWLEDGE PROJECT
érudit
+ algunos otros
JUAN PABLO ALPERIN
N PABLO ALPERIN
@juancommander @pkp #scholcommlab


Artificial
Intelligence
CONFERENCE
O'REILLY
intel AI
érudit

Some of the technologies used..

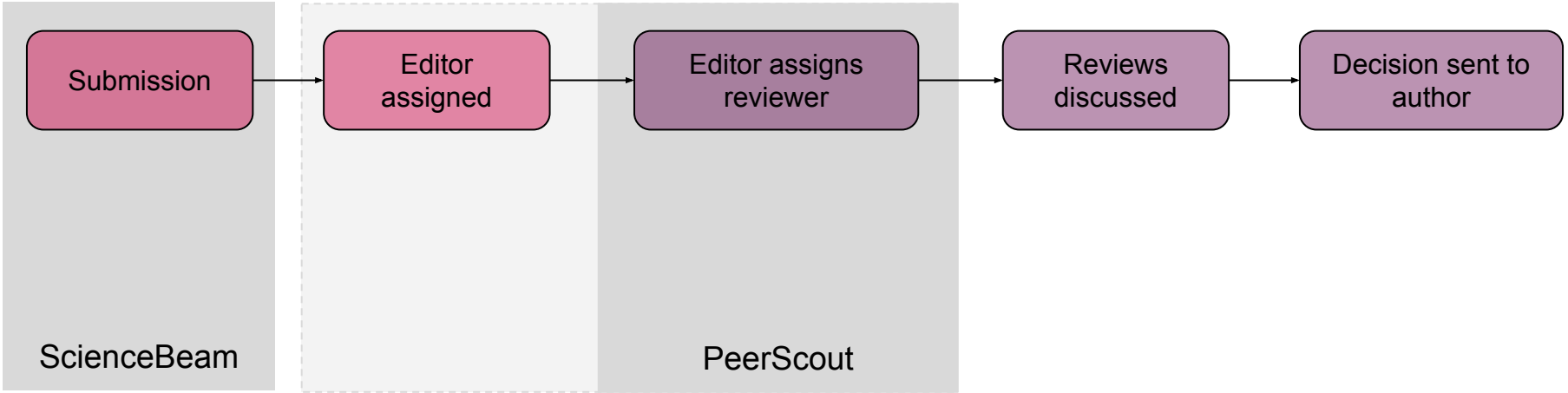
Conclusion

- Lessons learned:
  - Active community more important than higher initial score
  - Setting up training pipeline slow, compromise using Jupyter + Kubernetes for training and Airflow + Kubernetes for evaluation
  - Significantly improved performance using GROBID DL + layout features
- Next:
  - Train on larger data source
  - Improve other elements
  - Extend use-cases

PeerScout

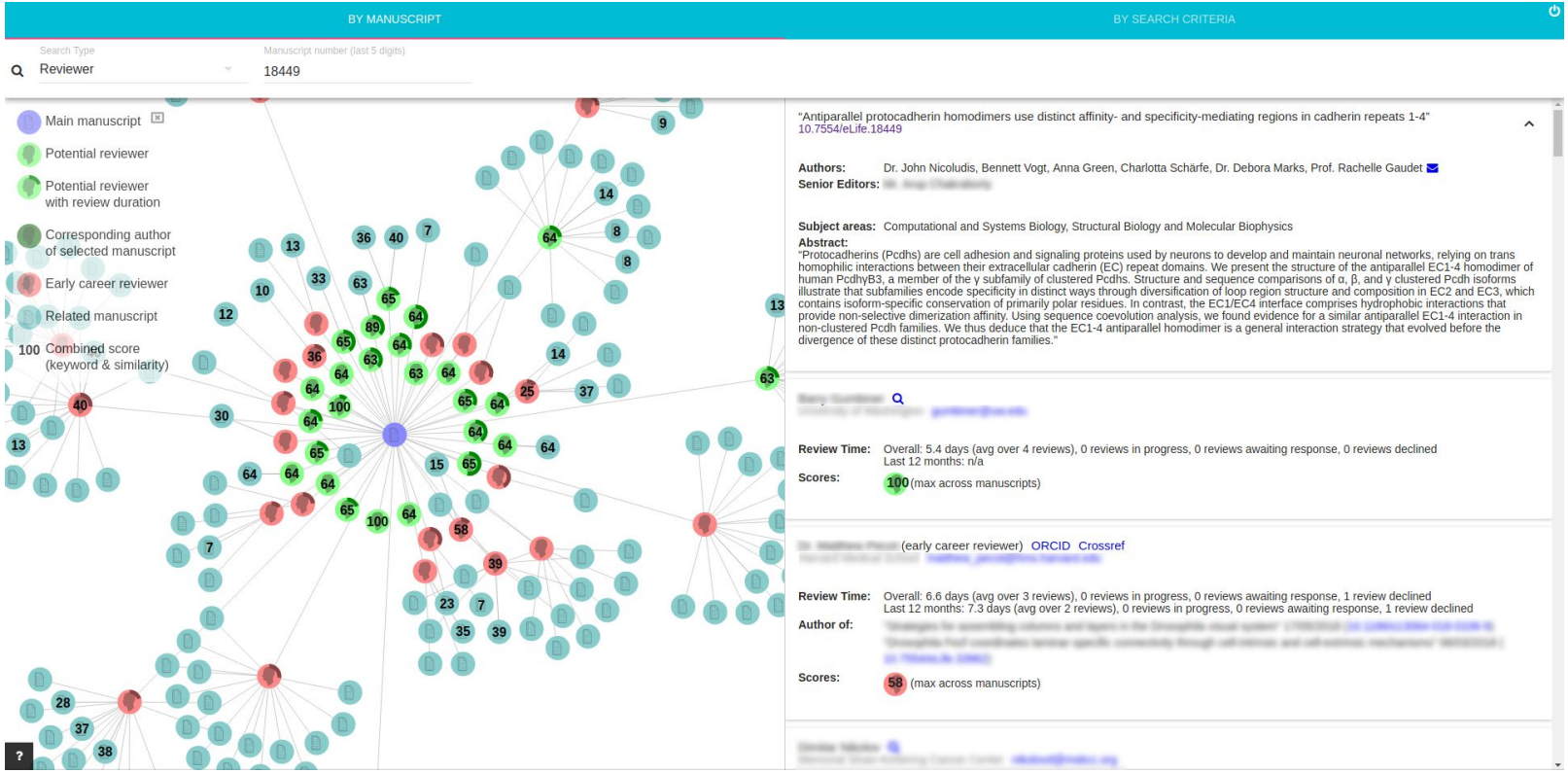
Open-Source Data Science Projects at eLife

Submission process (simplified)

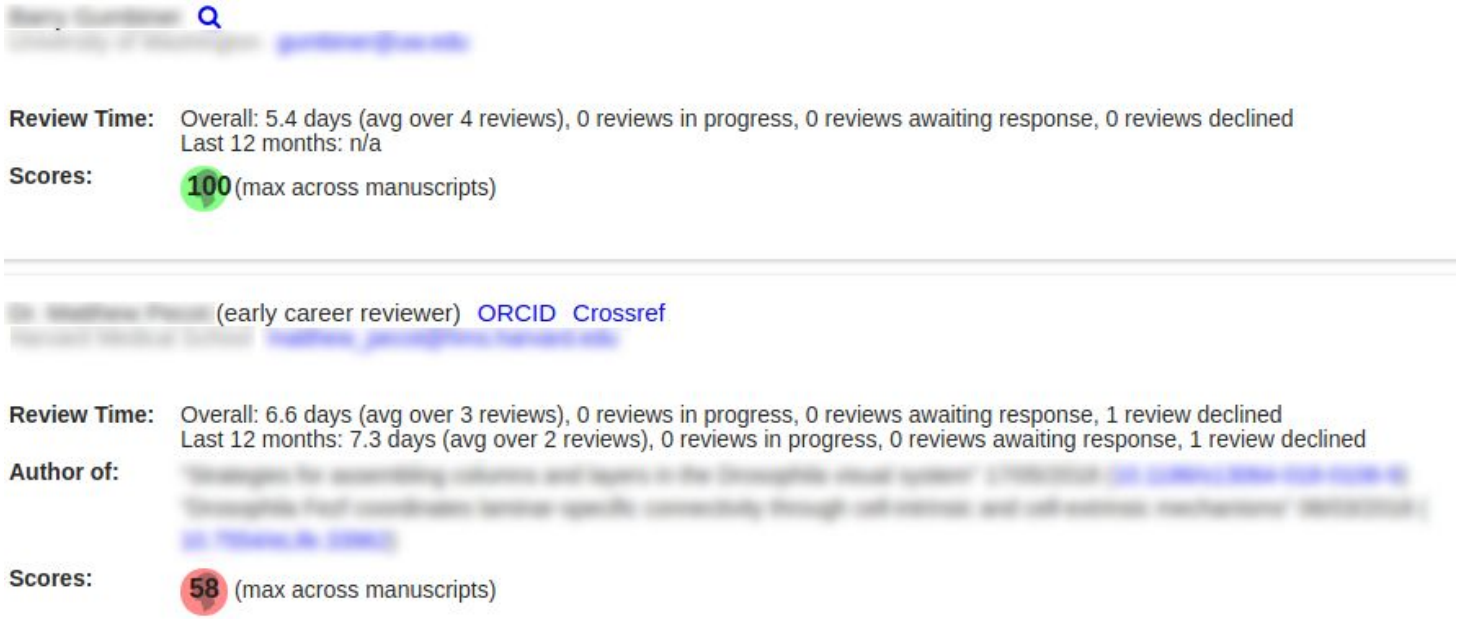


Also see Labs post: [Peer Review: New initiatives to enhance the value of eLife's process]

PeerScout - Overview



Suggested reviewers



Conclusion

- Lessons learned:
  - Work more closely with editors
  - Need for more integrated tools
- Next:
  - Integrate with our new data tools
  - More interpretable results to increase trust

PeerTax

(work by Alessio Caciagli)

Reviews can be long.. (e.g. 918 words)

In this manuscript Barry, Behet and colleagues address an important question aiming to possibly appoint an effective role in protection from clinical malaria to liver-stage immunity acquired naturally in malaria endemic areas. I believe the work presented is a significant contribution to our understanding of naturally acquired immunity to pre-erythrocytic stages of *P. falciparum* and I make a few comments and suggestions that may improve the current version of the manuscript.

The data shows that ABs developed during past malaria cases can reduce sporozoite motility and hepatocyte invasion *in vitro* suggesting that ABs acquired during natural infections can reduce new liver-stage infections. However, the contribution of a possible similar effect *in vivo* is, in my view, less clear from the data presented, and more caution may be needed to discuss the results. All the children followed throughout the study became parasite positive by qPCR, indicating that even children with strong inhibiting ABs were unable to block liver-stage infection efficiently. Furthermore, *high response to asexual stage lysate* might be confounding the analysis. I would add to the manuscript (or supplemental) figures the association of CSP ABs with asexual stage lysate ABs, and also the association of whole SPZ ABs with asexual stage lysate ABs, to give the reader an idea of how close these parameters are.

I also suggest including, if that has not been done already, the *Reported bed net use* in the multivariate analyses, as this could also be a factor increasing time to PCR positive and clinical malaria.

The authors cite the study by Tran et al.[1] where it was shown that, in Mali, time to PCR positive was independent of age, while time to clinical malaria increased with age, and where as stated *it was concluded that there no or very limited evidence for an age dependent acquisition of immunity protecting from infection.* Similarly, in the present manuscript, *in vitro* functional data of higher humoral response against pre-erythrocytic stages does not (independently of blood-stage immunity) protect from infection. So, I would rephrase the last sentence in first paragraph of the discussion to add a bit more caution in interpreting what may be causing partial protection.

At the end of section *Evidence of natural risk-modifying pre-erythrocytic immunity*the authors should, in my view, clearly state that *High gliding inhibition activity*does not independently associate in a statistically significant way, with *protection against falciparum infection*on the multivariate analyses where blood stage immunity was included; the P value is above 0.05 (0.055) and the CI includes 1, making the relative risk not statistically significant.

I believe the manuscript could be improved by presenting the quantitative analysis of the *18s* qPCR upon first parasite detection and determine if there is a negative association with the inhibitory capacity of the individuals' ABs. It would also be very interesting to question if time from first PCR positive to time of presentation of symptoms is different between poor and strong *in vitro* inhibitors. If the *in vitro* data showing gliding inhibition and reduced hepatocyte invasion are significant *in vivo*, one would expect a lower inoculum in the liver and thus a lower parasitaemia on the first PCR positive time-point. And then potentially a slower progression to clinical malaria. I believe with the data generated in this manuscript these analyses could be done, and would enrich the story.

It is not totally clear to me how individuals were selected for the flow cytometry assays. Survival, gliding inhibition, CSP, LSA1 and asexual lysate ELISAs were performed for the 51 participants, but flow cytometry data presented in fig2 D and E was obtained from 17 Burkinabes only; how were those selected and what is their time to PCR+ in the survival analysis. If they are the 8 poor and 8 strong inhibitors as defined by their gliding inhibition it should be stated in the methods (seems to be so, given supFig3, but there is one extra?).

I also suggest to pinpoint these 8 poor and 8 strong inhibitors in fig1 so that the reader would be informed of their time to PCR+ and time to malaria symptoms.

I would be more cautious when citing ref 31, I believe the study by Michael Stewart et al.[2] shows that non-motile SPZ are unable to invade, but is not clearly showing a direct association between % of human AB affecting motility and those levels correlating directly with invasion either.

Minor points:

In table 1, I would not refer to the 6 children who were PCR positive at the 3 weeks after treatment time-point as *Persisting parasites post-treatment*as I do not think that it can be excluded that the children were re-infected after clearance of PQ.

I recommend adding a brief description of the method in ref 45 in the section *In vitro sporozoite infectivity assay of a human hepatoma cell line* inmaterial and methods.

On page 3 below table 1 there is mention to *field PCR* which may be a mistake.

The data from the *in vitro* gliding inhibition by LSA IgG seems to be not shown. I think it should be clarified in the text that that is indeed the case. Likewise, if the LSA-1-specific IgG antibodies correlation with sporozoite invasion inhibition is data not shown I would clearly state it in the text.

In figS2A I would specify that is IgG in the figure x axis and use the label CSP IgG titer instead of CSP antibody titer.

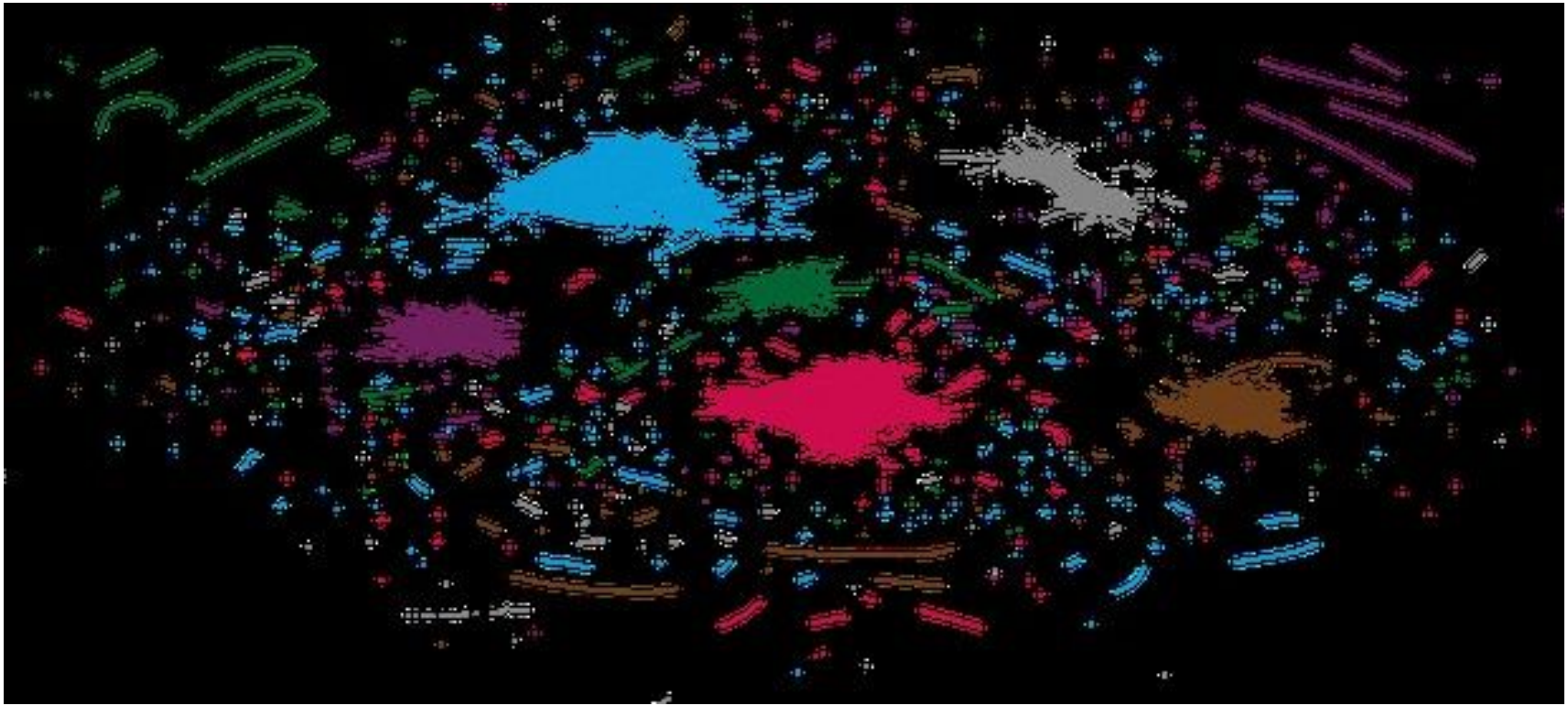
Figure S3D is called before Figure S3C, I would call figures in ascending and alphabetical order instead.

https://doi.org/10.21956/wellcomeopenres.16280.r34444

Main clusters identified

- Figures/Non-textual content
- Stats/Analysis/Techniques
- Impact/Novelty
- Text/Exposition clarity
- Previous literature
- Main discussion

HDBSCAN Output (main topics plus residuals)



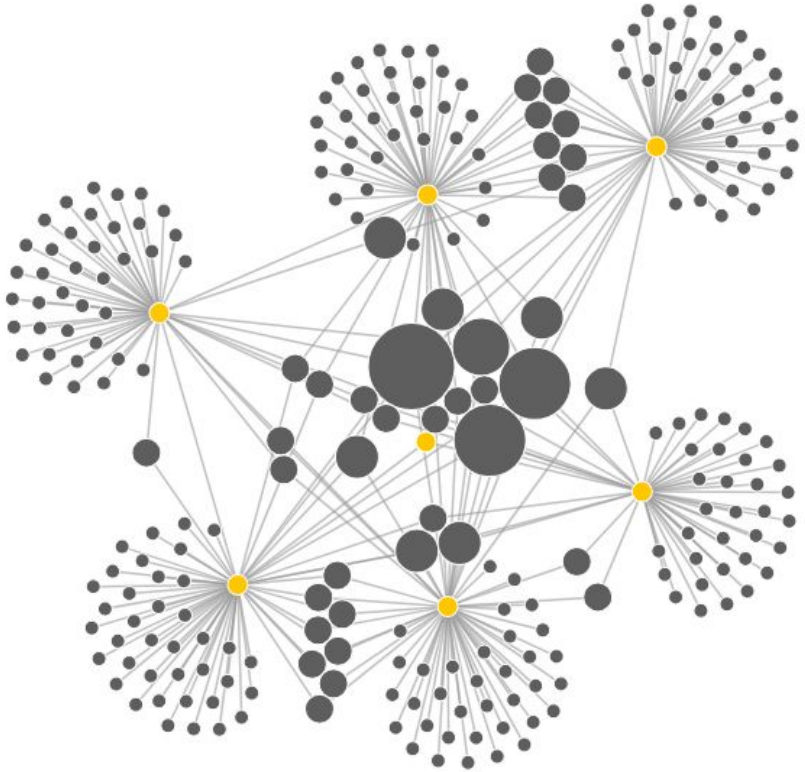
Summary

- Explored automatically structure peer-review content
- Created good first model
- Next:
  - More open training data
  - Better categories
  - More labelled data
  - Supervised model
  - Define use-cases

Citation Sentiment

(work by David Ciudad)

Example citation network visualisation



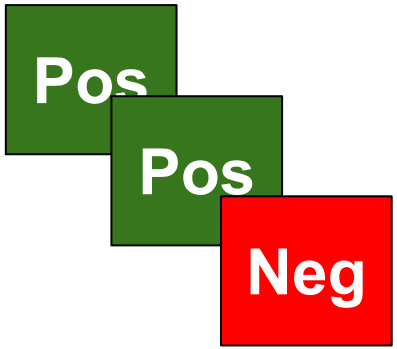
citation gecko

http://citationgecko.com/

Example citation

"Our results contrast with the high rate of XMRV detection reported by *Lombard et al.* among both CFS patients and controls, but are in agreement with recent data reported in two large studies in the UK and in the Netherlands..." (Switzer et al., 2010)

Why does the sentiment matter?



Citation count

3

“impact”

“important”

?

Negative citations count towards citation count

Summary

- Explored initial model that is better than off-the-shelf models
- Current datasets not large enough
- Next:
  - Find or create larger dataset
  - Extend to citation function
  - Define good use-cases

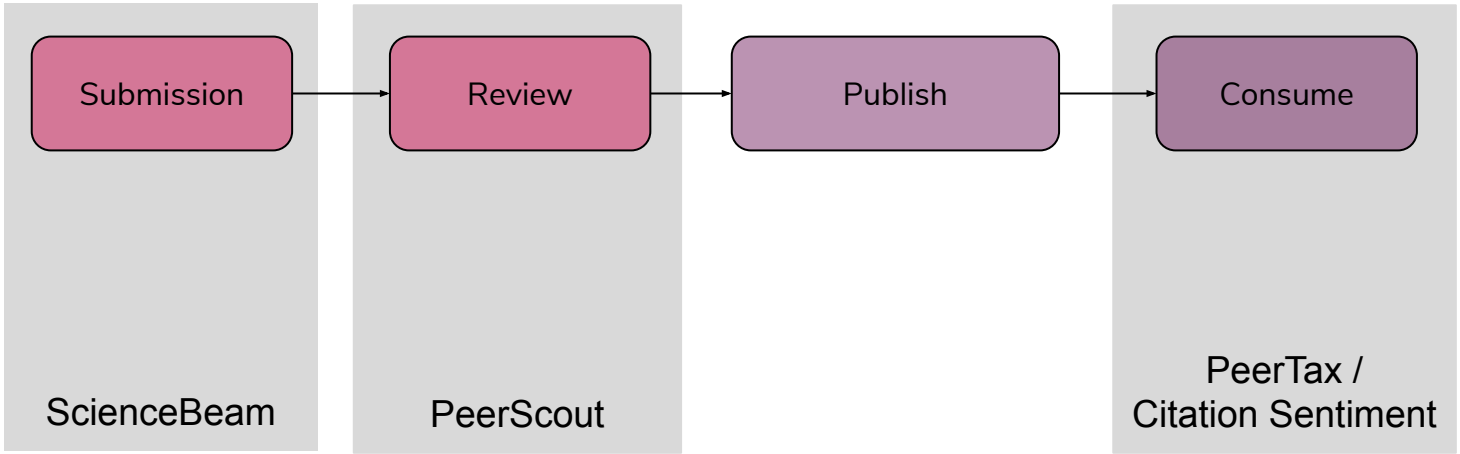
Summary

Data Science projects in context

Amy
Author

Hubert
Reviewing Editor / Editorial Staff

Philip
Reader



Also see Labs post: [Peer Review: New initiatives to enhance the value of eLife's process]

Resources

- ScienceBeam Labs blog
- PeerTax slides (and labs blog)
- Citation Sentiment slides for longer talk (as part of Workshop on Open Citations)

eLIFE

Thank you

Background: CC0, https://www.pexels.com/photo/night-television-tv-video-8158/