PDF Annotation and Data Harmonization (PANDHA)

September 2021
Version 1.2

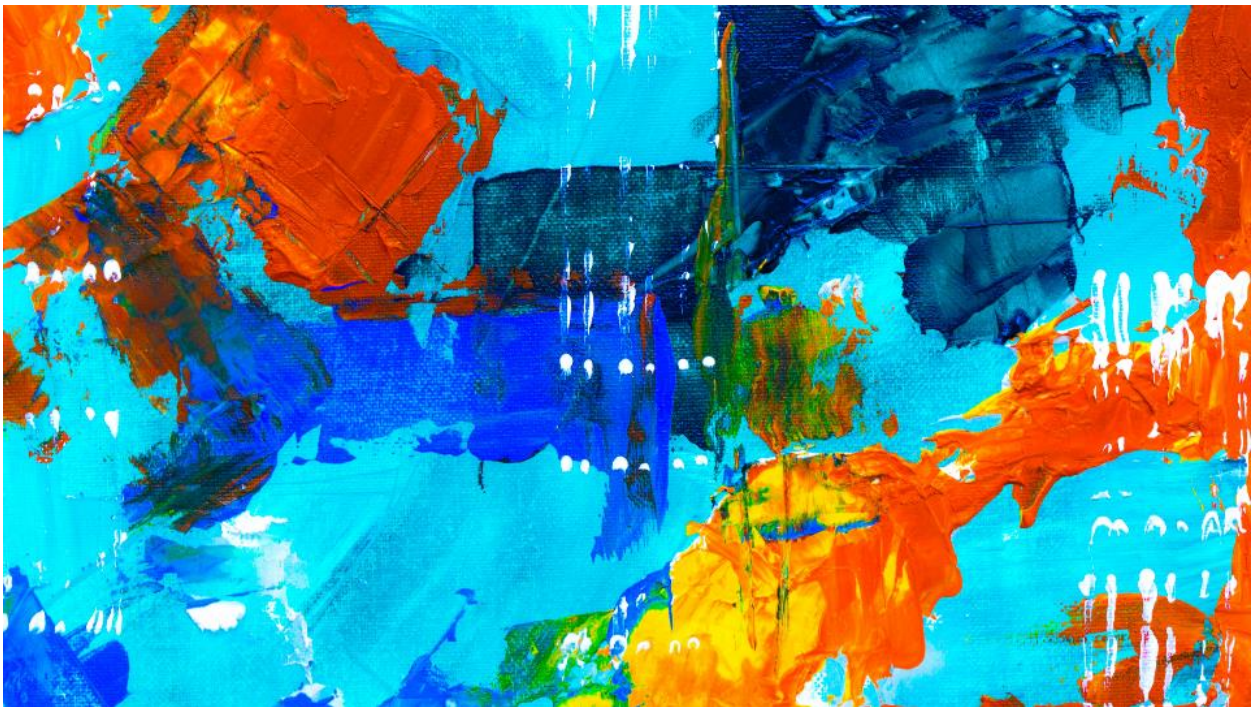


Table of Contents

Table of Contents 1
Document Revision 2
1 Installation 3
1.1 Scope and Purpose 3
1.2 Source Code Access 3
1.2 Installation MacOSX * need to work out zip issue 3
1.3 Installation Windows 4
2 Application 8
2.0 Load PDF 8
2.1 Annotate PDF 10
2.2 Save Annotations 12
2.3 Load Annotations 13

Document Revision

Date	Version Number	Document Changes
01/05/2021	1.0	Initial draft
07/20/2021	1.1	Updated to beta PANDHA V1
09/29/2021	1.2	Update to beta PANDHA V1.0.2

1 Installation

1.1 Scope and Purpose

The purpose of this project is to further the research and development of tools that NCEA can use in their creation of machine-readable datasets and machine learning research. This effort consists of the following objectives:

- Research and develop software for NCEA that provides the ability to annotate scientific publications for use in machine learning algorithms. This software should be able to accept a list of tags provided by NCEA, allow the user to apply these tags to PDF documents in a web interface, and then extract out the information needed in machine-readable formats that can be used for machine learning.

1.2 Source Code Access

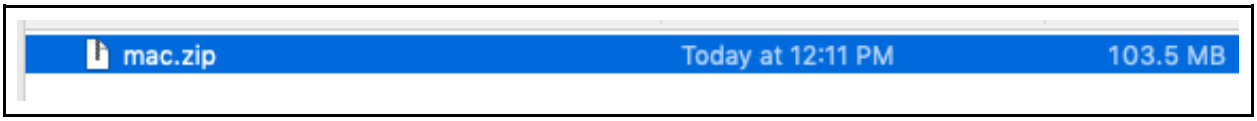
This describes the process of accessing the source code.

1. https://github.com/USEPA/pdf-data-extraction

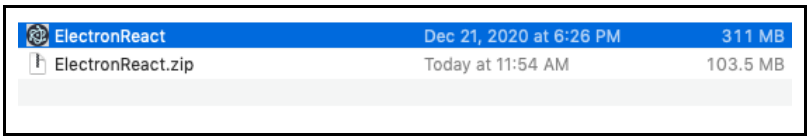
1.2 Installation MacOSX * need to work out zip issue

1. Download latest version from

2. Double click ‘mac.zip’ to unzip the file.



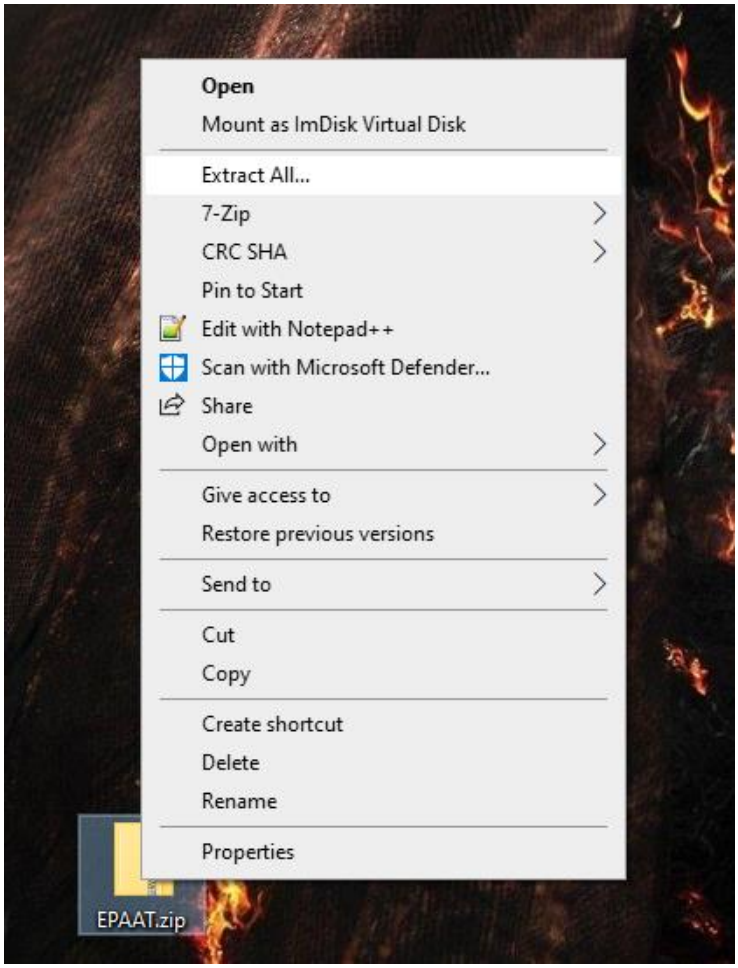
3. Double click ‘ElectronReact’ to launch the application.



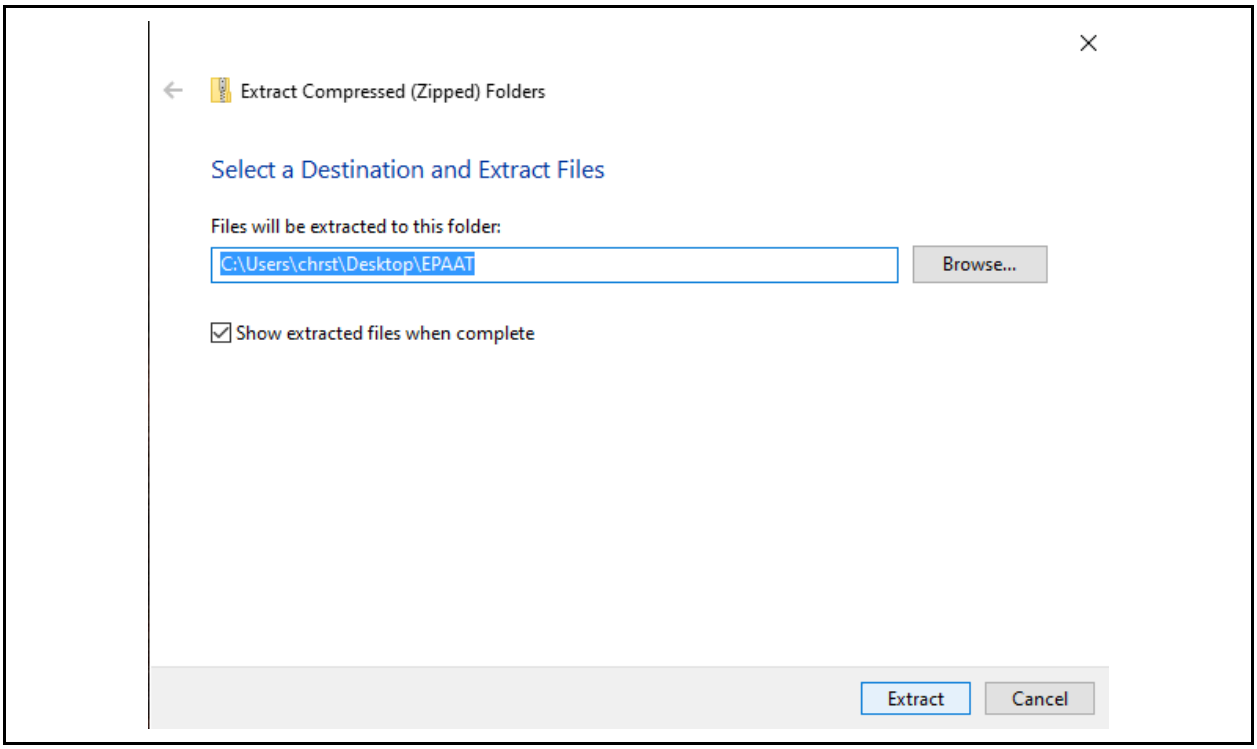
1.3 Installation Windows

1. Download latest version from (PANDHA_1.0.2)
https://usepa.sharepoint.com/sites/DOEORNLCPAD/Shared%20Documents/General/Annotation%20Tool%20Beta/Windows

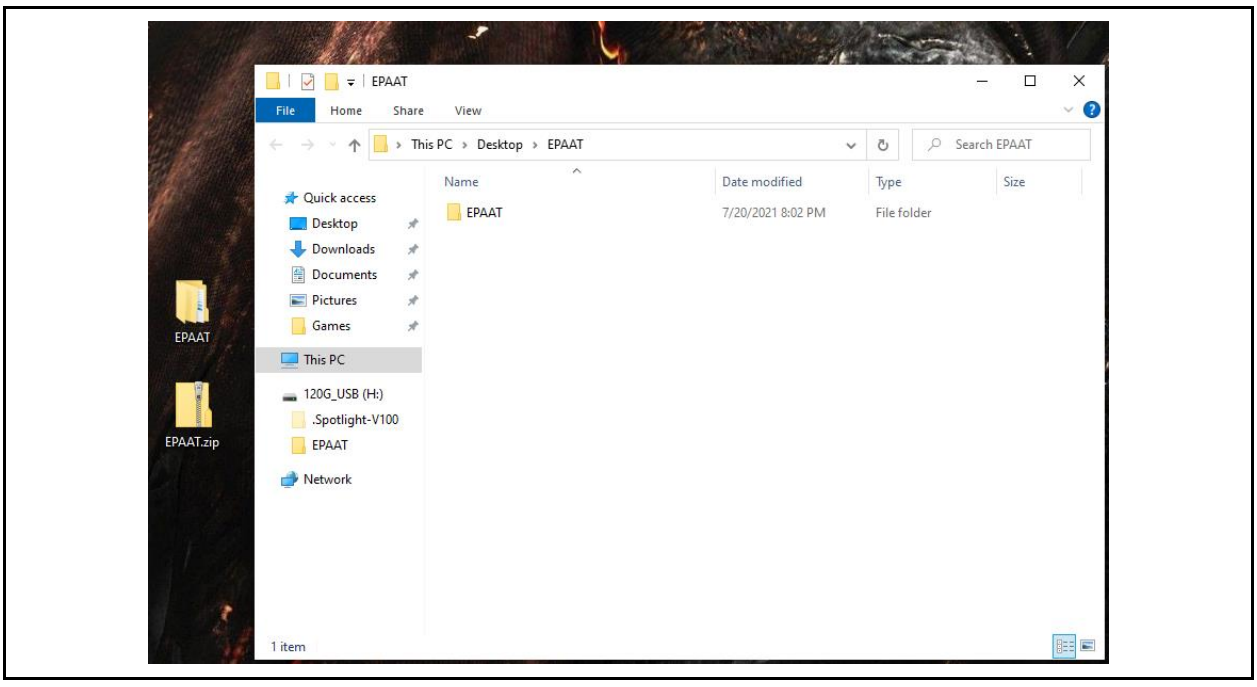
2. Right click ‘Windows.zip’ and select ‘Extract All’



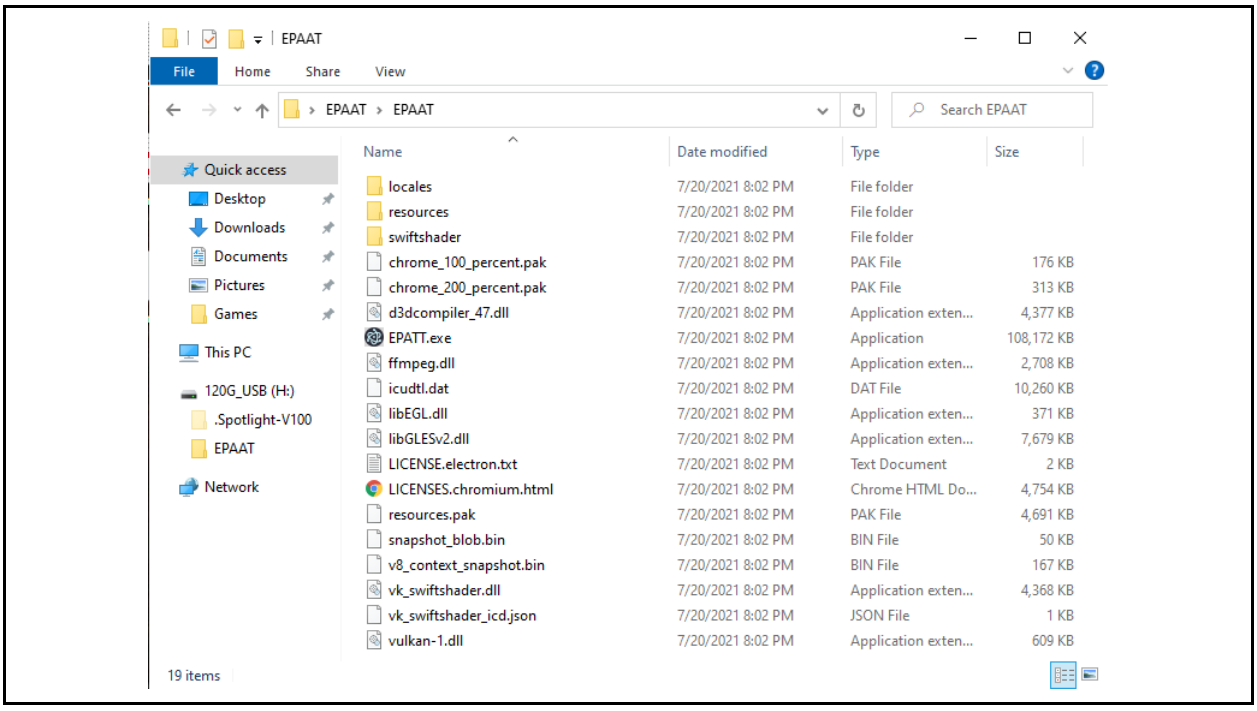
3. Select location and hit *Extract*



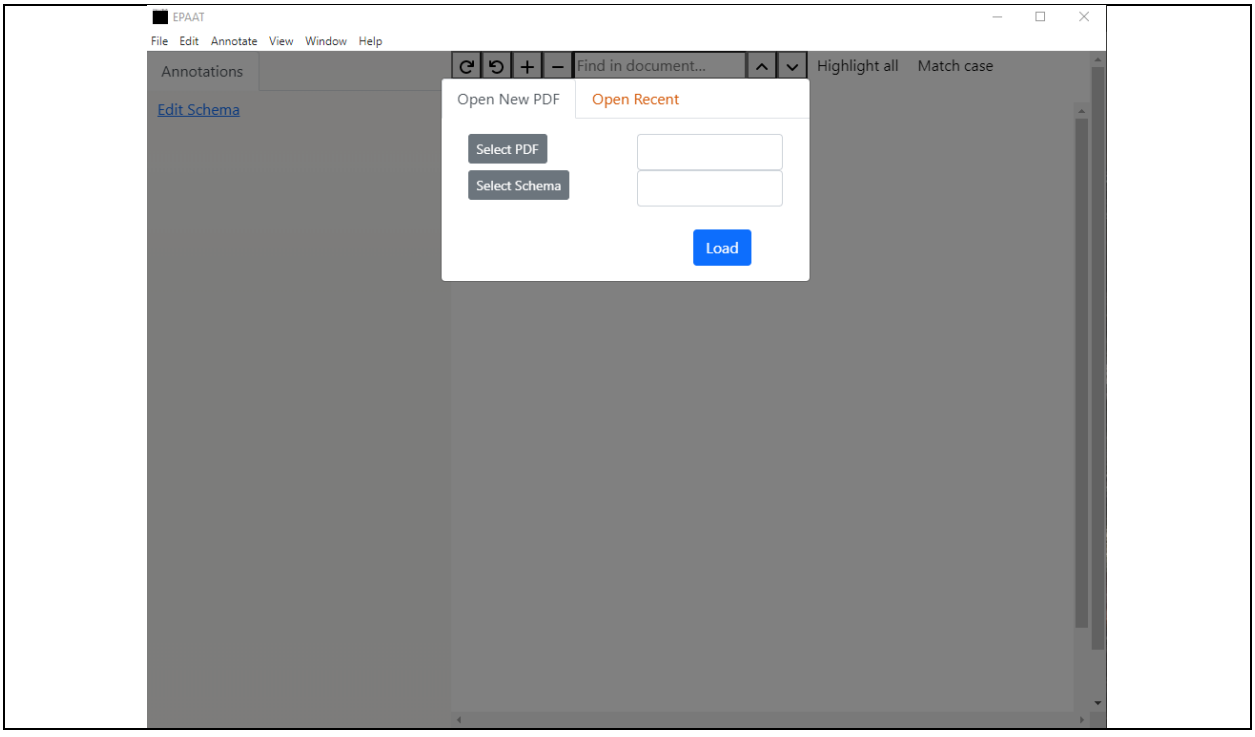
4. From the extracted location double click the *EPAAT* Folder



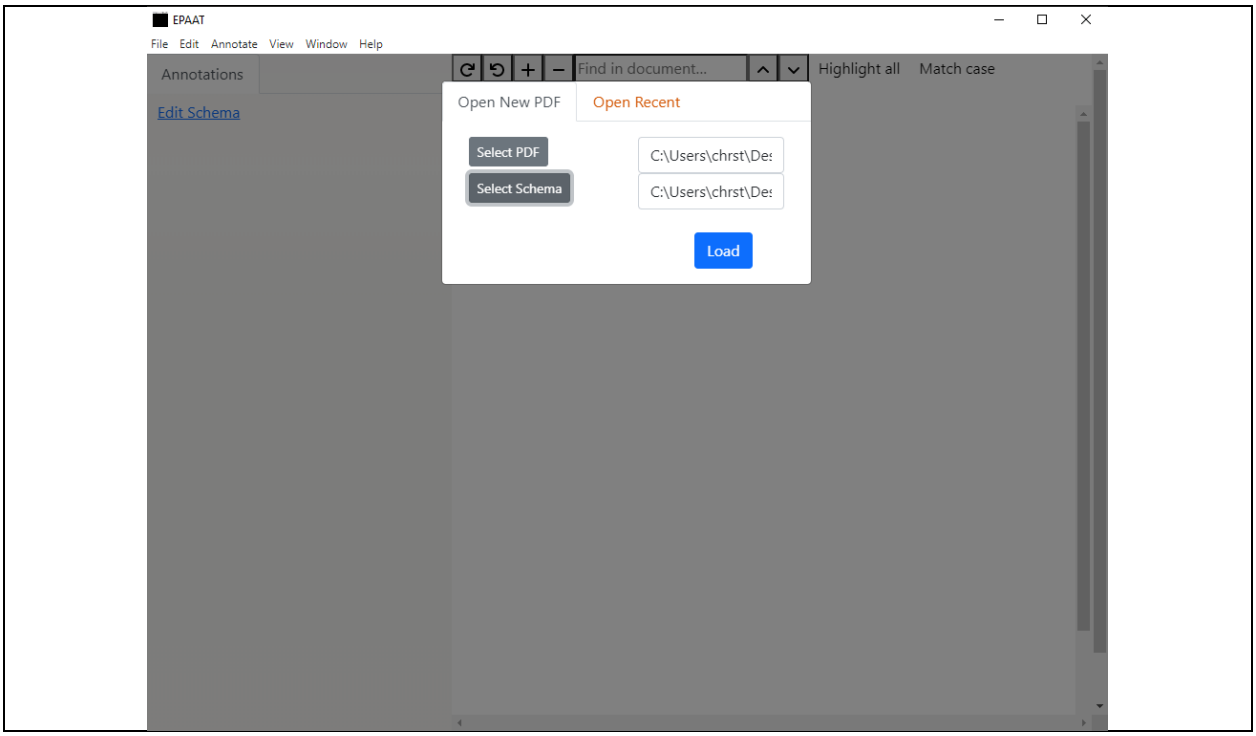
5. Double click *EPAAT*



6. Double click *EPATT.exe* to start the application



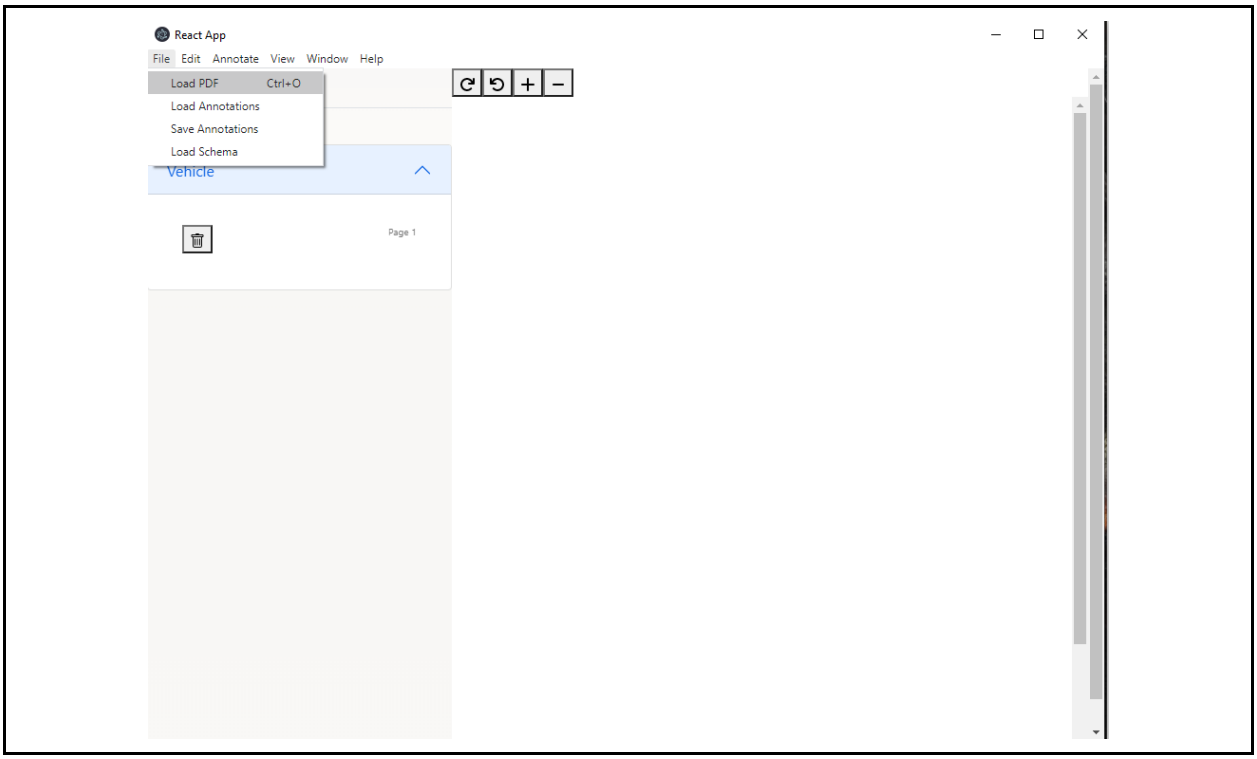
7. Select the PDF and Schema (schema.json is including in the installation folder) and click *Load*



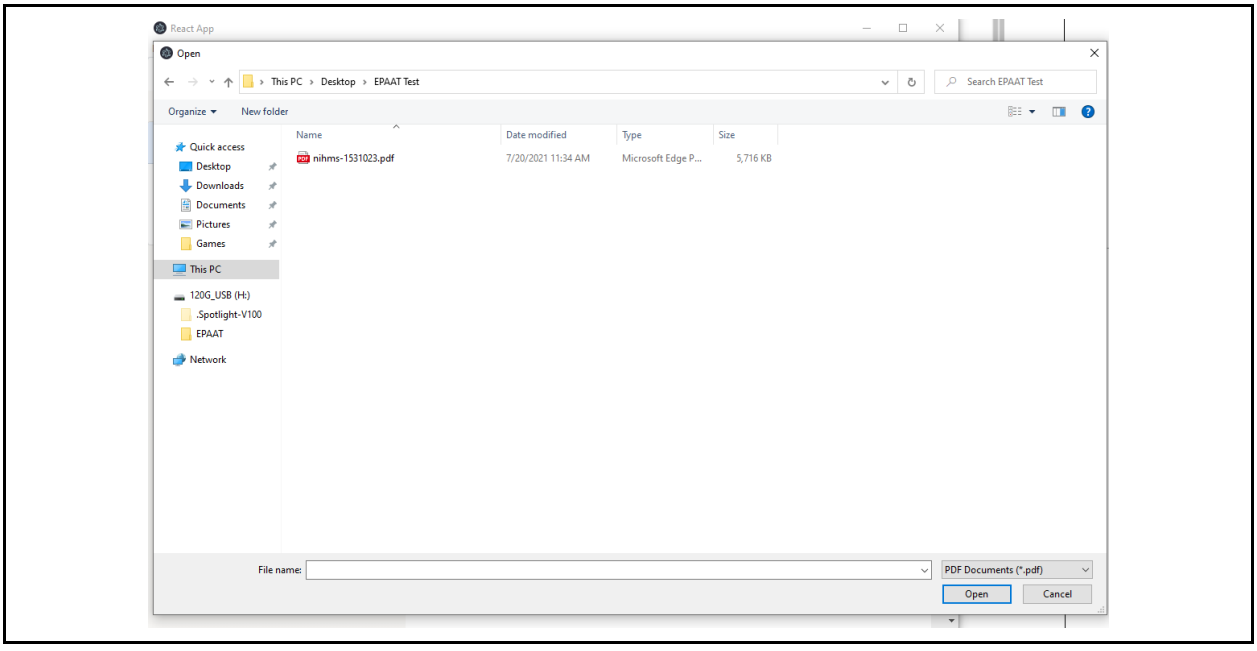
2 Application

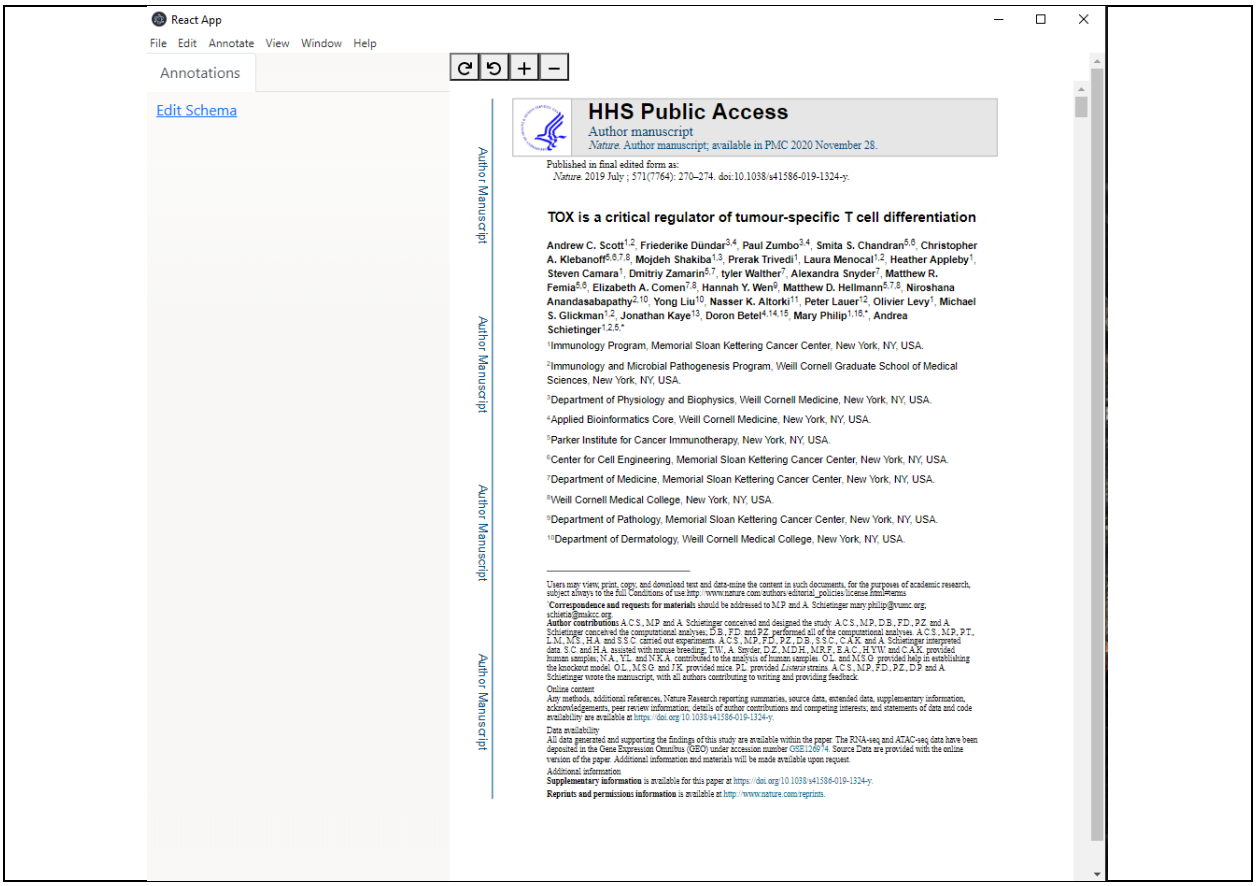
2.0 Load PDF

1. Click *File* in the menu bar and select *Load PDF*.



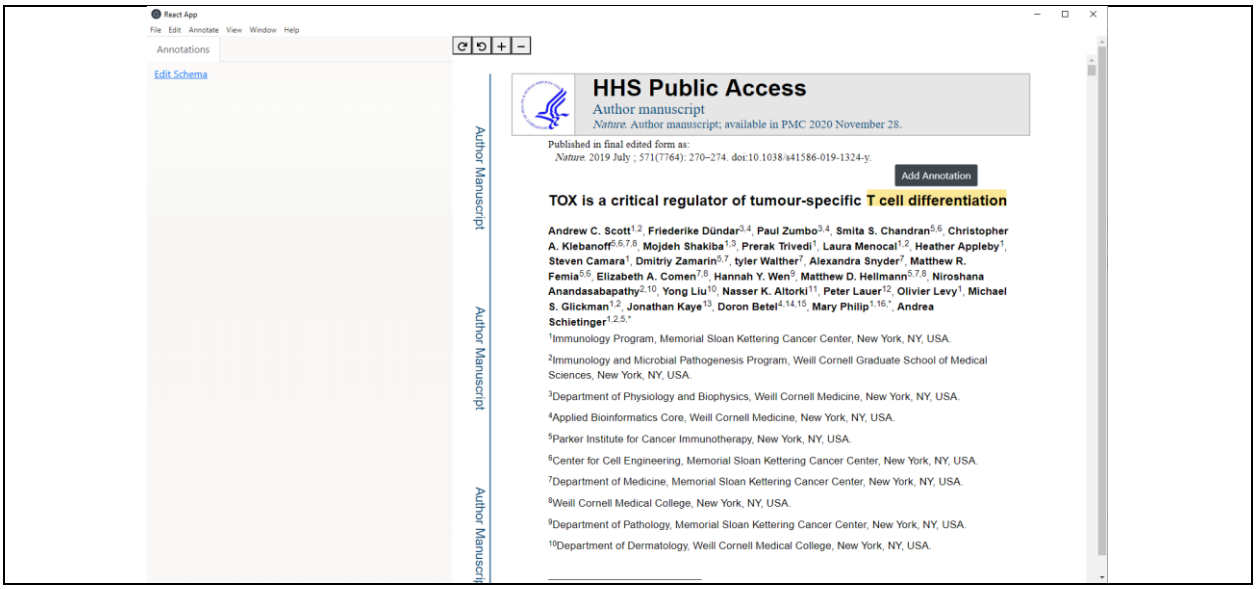
2. Select the PDF file from your computer.



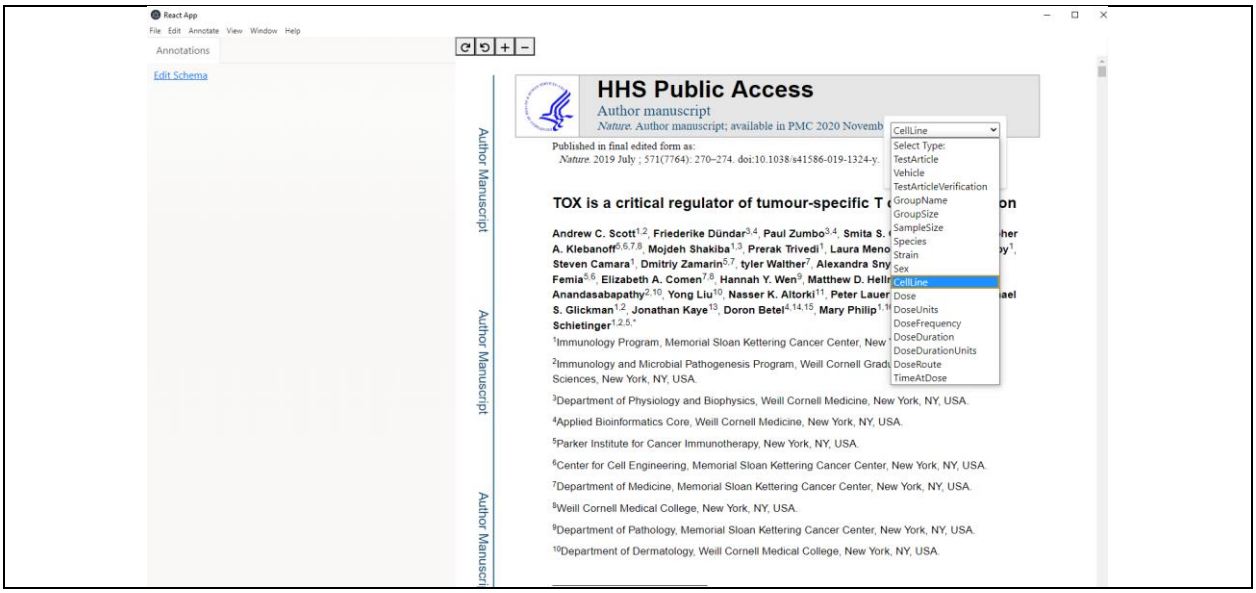


2.1 Annotate PDF

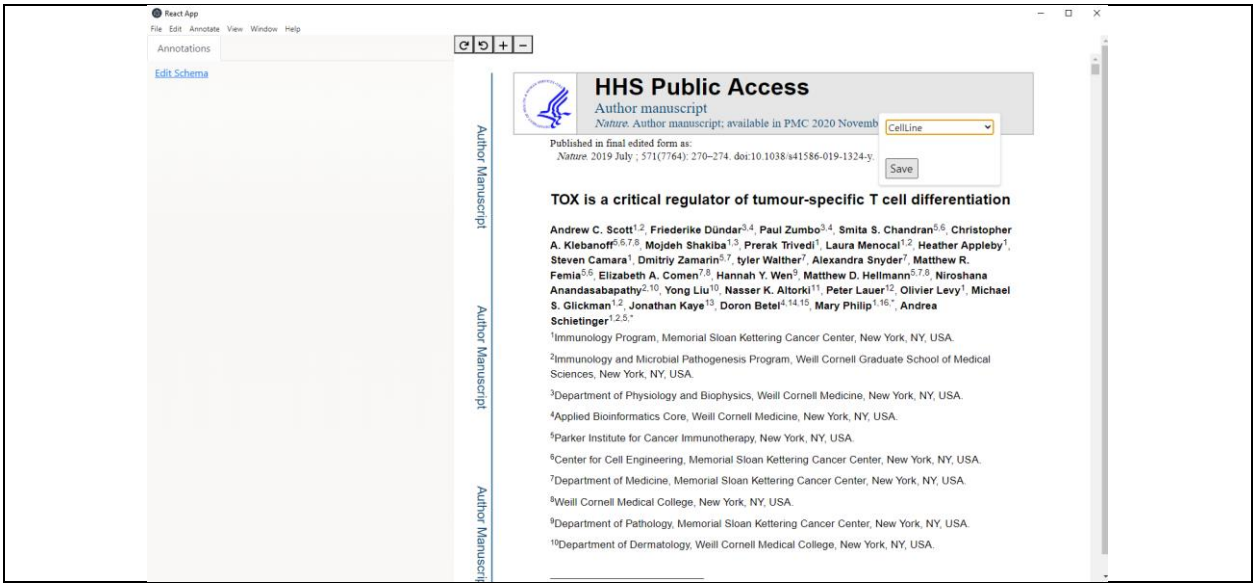
1. Highlight text you wish to annotate and select *Add Annotation*.

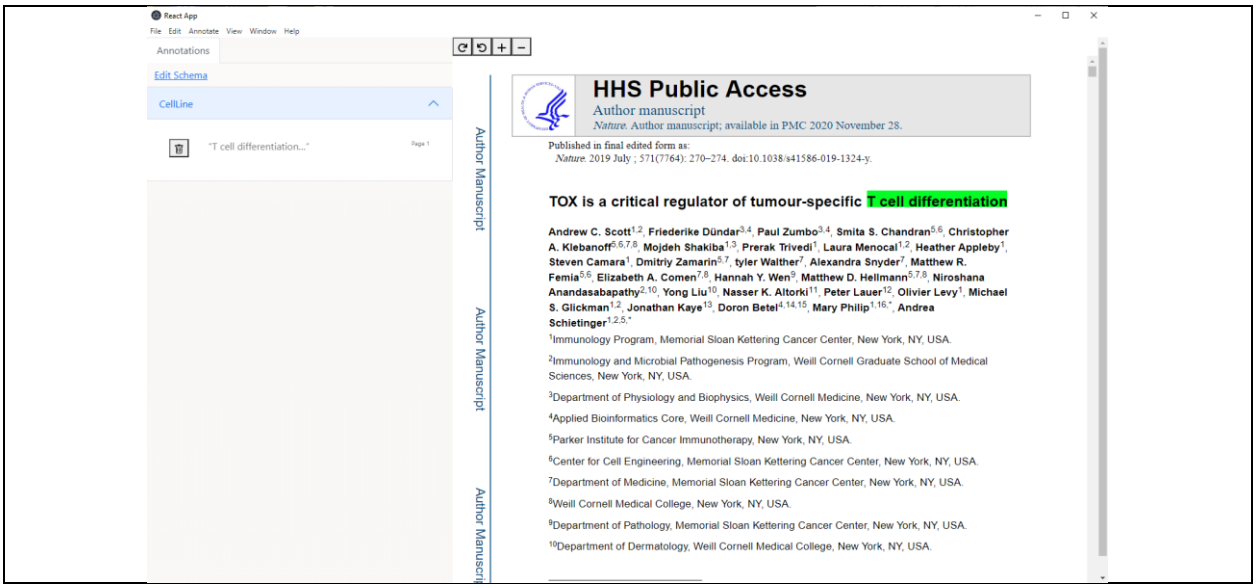


2. Select the annotation type.



3. Hit save

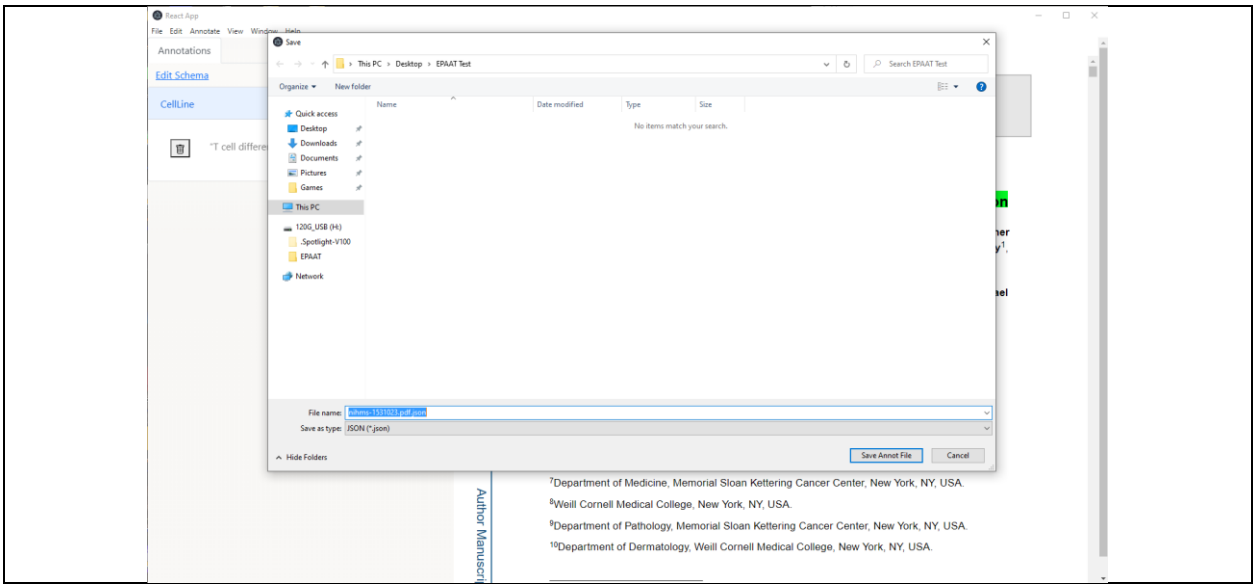




2.2 Save Annotations

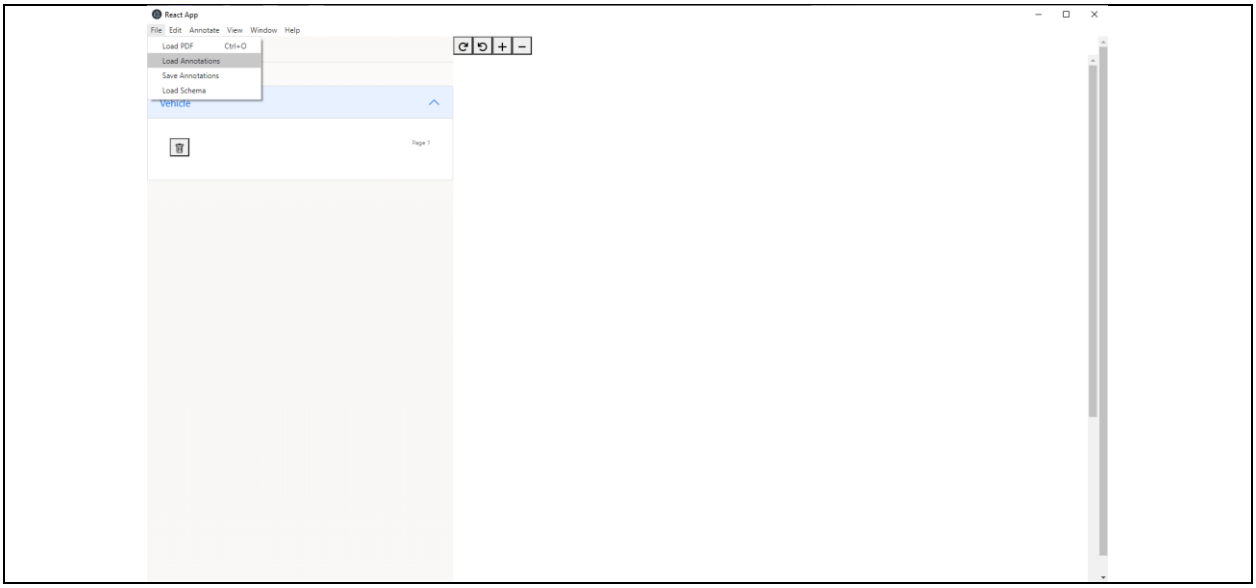
1. Click *File* in the menu bar and select *Save Annotations*.

2. Select a save location on your computer and click *Save Annot File*.

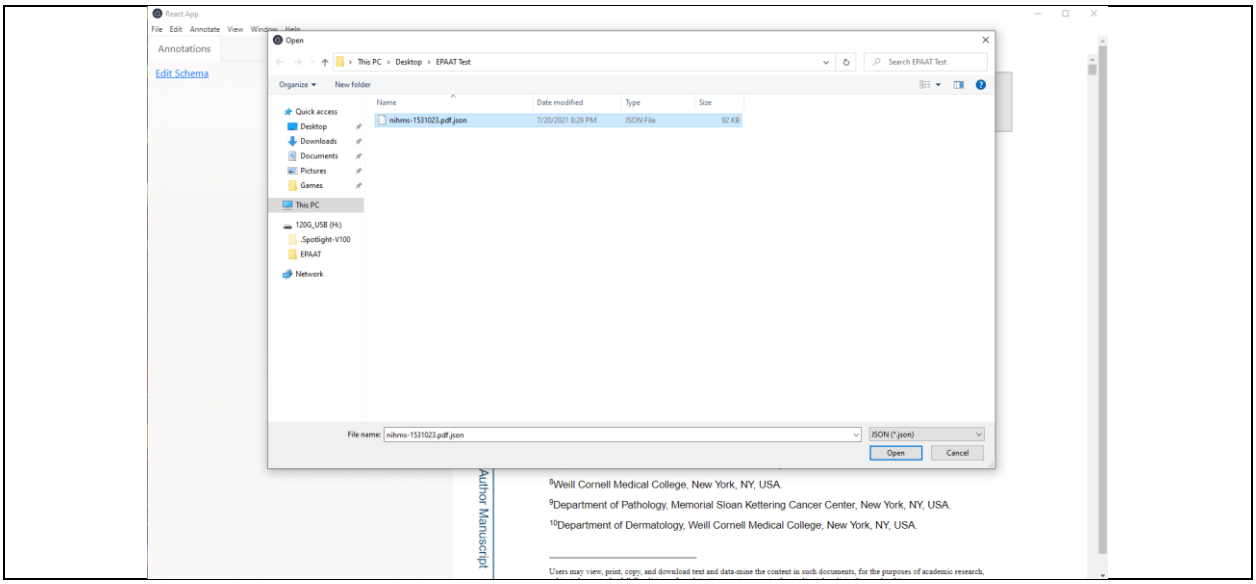


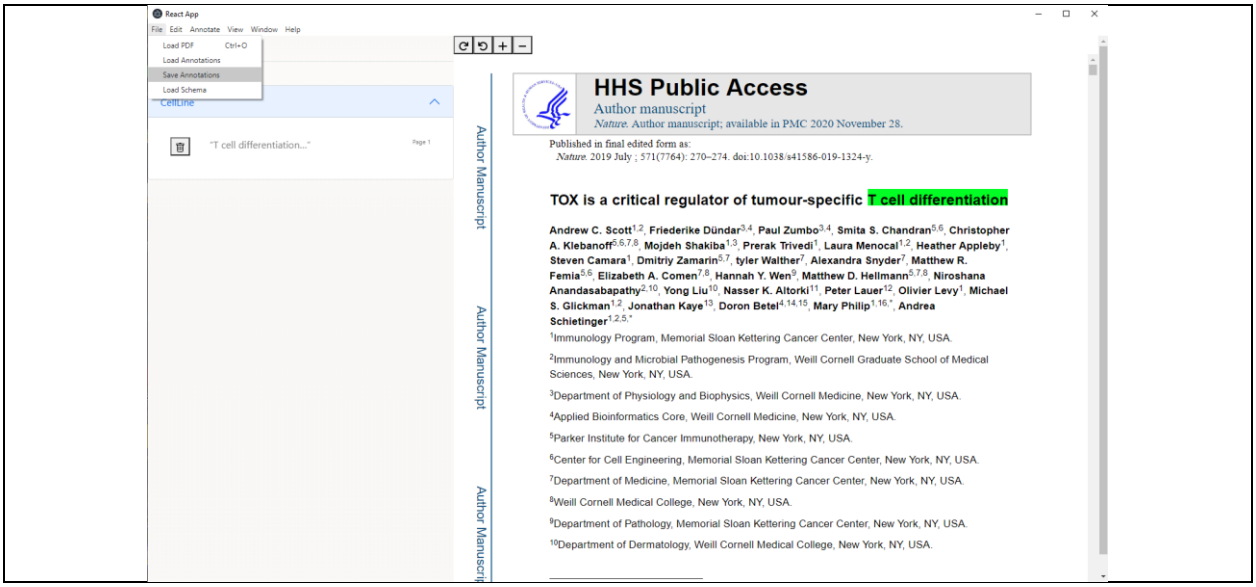
2.3 Load Annotations

1. Click *File* in the menu bar and select *Load Annotations*.



2. Select an annotation file


React App
File Edit Annotate View Window Help
Annotations
Edit Schema
Open
This PC › Desktop › EPAAT Test
Search EPAAT Test
Organize
New folder
Quick access
Desktop
Downloads
Documents
Pictures
Games
This PC
120G_USB (H:)
.Spotlight-V100
EPAAT
Network
Name
Date modified
Type
Size
nihms-1531023.pdf.json
7/20/2021 8:29 PM
JSON File
92 KB
File name: nihms-1531023.pdf.json
JSON (*.json)
Open
Cancel
Author Manuscript
[8]Weill Cornell Medical College, New York, NY, USA.
[9]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[10]Department of Dermatology, Weill Cornell Medical College, New York, NY, USA.
Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research,


React App
File Edit Annotate View Window Help
Load PDF Ctrl+O
Load Annotations
Save Annotations
Load Schema
CellLine
"T cell differentiation..."
Page 1
Author Manuscript
Author Manuscript
Author Manuscript
HHS Public Access
Author manuscript
Nature. Author manuscript; available in PMC 2020 November 28.
Published in final edited form as:
Nature. 2019 July ; 571(7764): 270–274. doi:10.1038/s41586-019-1324-y.
TOX is a critical regulator of tumour-specific T cell differentiation
Andrew C. Scott[1,2], Friederike Dündar[3,4], Paul Zumbo[3,4], Smita S. Chandran[5,6], Christopher A. Klebanoff[5,6,7,8], Mojdeh Shakiba[1,3], Prerak Trivedi[1], Laura Menocal[1,2], Heather Appleby[1], Steven Camara[1], Dmitriy Zamarin[5,7], tyler Walther[7], Alexandra Snyder[7], Matthew R. Femia[5,6], Elizabeth A. Comen[7,8], Hannah Y. Wen[9], Matthew D. Hellmann[5,7,8], Niroshana Anandasabapathy[2,10], Yong Liu[10], Nasser K. Altorki[11], Peter Lauer[12], Olivier Levy[1], Michael S. Glickman[1,2], Jonathan Kaye[13], Doron Betel[4,14,15], Mary Philip[1,16,*], Andrea Schietinger[1,2,5,*]
[1]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[2]Immunology and Microbial Pathogenesis Program, Weill Cornell Graduate School of Medical Sciences, New York, NY, USA.
[3]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA.
[4]Applied Bioinformatics Core, Weill Cornell Medicine, New York, NY, USA.
[5]Parker Institute for Cancer Immunotherapy, New York, NY, USA.
[6]Center for Cell Engineering, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[7]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[8]Weill Cornell Medical College, New York, NY, USA.
[9]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
[10]Department of Dermatology, Weill Cornell Medical College, New York, NY, USA.