



Data Enrichment Updates

ADSUG Meeting, November 2023
Felix Grezes, Tom Allen, Golnaz Shapurian





Data Enrichment: Introduction

ADS continues its efforts to enrich our data:

by using Machine Learning technologies to develop internal tools.

- astroBERT+ for planetary names detection
- astroBERT for UAT keyword assignment
- astroBERT for SciX categorization

by collaborating with external partners.

- WIESP 2023 and FOCAL
- NASA SMD foundational model efforts
- Universe TBD
- Summer PhD student internship

Data Enrichment – astroBERT Review

astroBERT is a language model trained on our astronomy text data.

- ~4B tokens from 395,499 recent astronomy papers with XML sources
- openly available: 🤗 <https://huggingface.co/adsabs/astroBERT>
 - includes tutorials
- base for downstreams tasks
 - Detecting Entities in the Astrophysics Literature:
<https://ui.adsabs.harvard.edu/WIESP/2022/SharedTasks>

Data Enrichment: ML Models into Production

Planetary Names

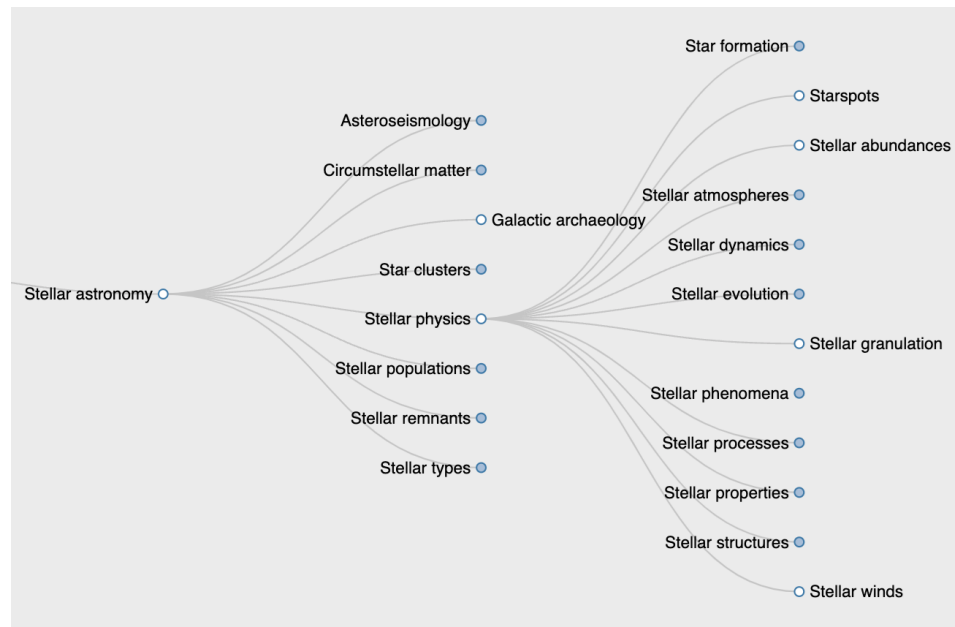
Allow searching on planetary feature names and linking to [USGS Gazetteer](#)

Unified Astronomy Thesaurus Keywords

Automate generation of keywords using UAT terms

SciX categorizer

Automate time-intensive curation processes



Section of the UAT hierarchy.

External Collaborations: 2nd WIESP 2023

- We co-organized the 2nd Workshop on Information Extraction from Scientific Publications
 - <https://ui.adsabs.harvard.edu/WIESP/2023/>
 - partnership with Dr. Tirthankar Ghosal of Oak Ridge National Lab
 - 16 participating papers

- Keynote talk by Dr. Yuan-Sen Ting from the Australian National University
 - *Can Artificial Intelligence Generate Meaningful Scientific Hypotheses?*



Dr. Tirthankar Ghosal



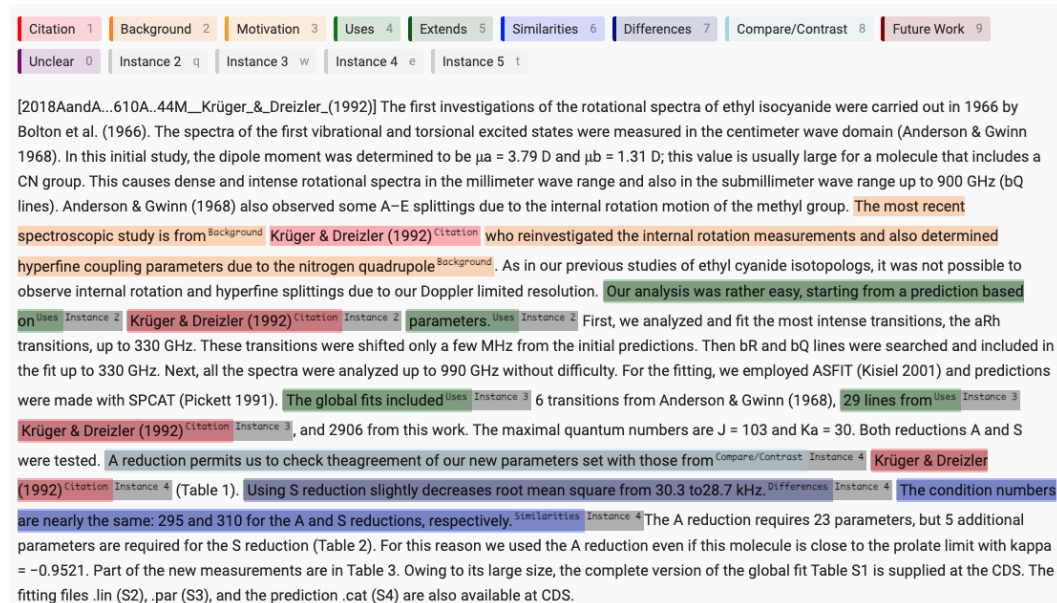
Dr. Yuan-Sen Ting

External Collaborations: FOCAL@WIESP2023

Function Of Citation in the Astrophysical Literature (FOCAL)

Addresses questions relating to the function, polarity, or impact of a given citation within a given work by analyzing the context surrounding the citation.

- Over 6000 annotated citations
- Addresses citation function
- Used for WIESP 2023 Shared Task
- [Publicly Available Dataset](#)



[2018AandA...610A..44M_Krüger_&Dreizler_(1992)] The first investigations of the rotational spectra of ethyl isocyanide were carried out in 1966 by Bolton et al. (1966). The spectra of the first vibrational and torsional excited states were measured in the centimeter wave domain (Anderson & Gwinn 1968). In this initial study, the dipole moment was determined to be $\mu_a = 3.79$ D and $\mu_b = 1.31$ D; this value is usually large for a molecule that includes a CN group. This causes dense and intense rotational spectra in the millimeter wave range and also in the submillimeter wave range up to 900 GHz (bQ lines). Anderson & Gwinn (1968) also observed some A–E splittings due to the internal rotation motion of the methyl group. The most recent spectroscopic study is from Background Krüger & Dreizler (1992) Citation who reinvestigated the internal rotation measurements and also determined hyperfine coupling parameters due to the nitrogen quadrupole Background. As in our previous studies of ethyl cyanide isotopologs, it was not possible to observe internal rotation and hyperfine splittings due to our Doppler limited resolution. Our analysis was rather easy, starting from a prediction based on Uses Instance 2 Krüger & Dreizler (1992) Citation Instance 2 parameters. Uses Instance 3 First, we analyzed and fit the most intense transitions, the aRh transitions, up to 330 GHz. These transitions were shifted only a few MHz from the initial predictions. Then bR and bQ lines were searched and included in the fit up to 330 GHz. Next, all the spectra were analyzed up to 990 GHz without difficulty. For the fitting, we employed ASFIT (Kisiel 2001) and predictions were made with SPCAT (Pickett 1991). The global fits included Uses Instance 3 6 transitions from Anderson & Gwinn (1968), 29 lines from Uses Instance 3 Krüger & Dreizler (1992) Citation Instance 3, and 2906 from this work. The maximal quantum numbers are $J = 103$ and $K_a = 30$. Both reductions A and S were tested. A reduction permits us to check the agreement of our new parameters set with those from Compare/Contrast Instance 4 Krüger & Dreizler (1992) Citation Instance 4 (Table 1). Using S reduction slightly decreases root mean square from 30.3 to 28.7 kHz, Differences Instance 4 The condition numbers are nearly the same: 295 and 310 for the A and S reductions, respectively. Similarities Instance 4 The A reduction requires 23 parameters, but 5 additional parameters are required for the S reduction (Table 2). For this reason we used the A reduction even if this molecule is close to the prolate limit with $\kappa = -0.9521$. Part of the new measurements are in Table 3. Owing to its large size, the complete version of the global fit Table S1 is supplied at the CDS. The fitting files .lin (S2), .par (S3), and the prediction .cat (S4) are also available at CDS.

Example of text annotated for citation function analysis.

NASA SMD Efforts for a Foundational LLM

- We are part of NASA's efforts to build a large foundational language model for the Science Mission Directorate's needs.
 - Led by Dr. Rahul Ramachandran at Marshall Space Flight Center
 - collaboration with IBM and Dr. Bhatta Bhattacharjee



Dr. Rahul Ramachandran



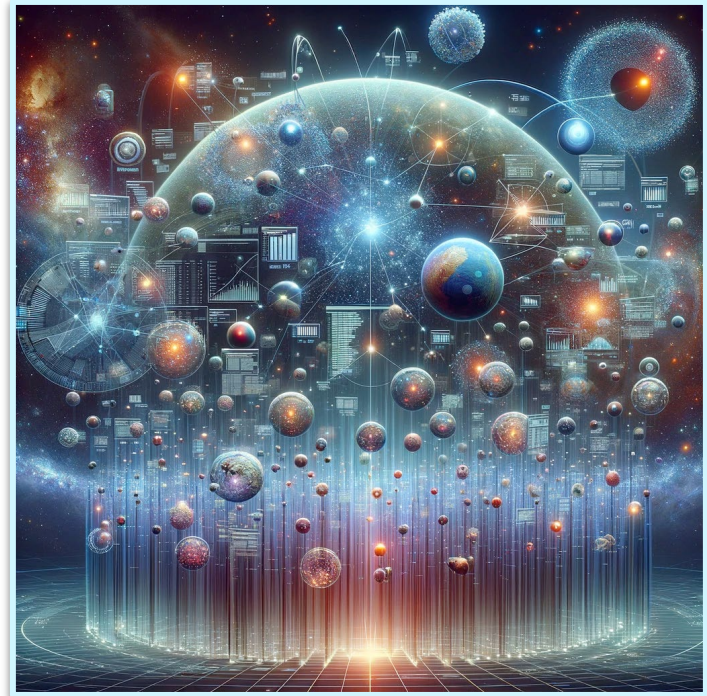
Dr. Bhatta Bhattacharjee

Open Corpus

Two datasets to be shared for language model training with collaborators.

- Over 2.8M articles that are open access
- Full Text of each article
- Abstract, References and Citations for articles, when available

- Contains question answer pairs for sample of 50 articles for model validation



Dall-E generated image inspired by the Open Corpus dataset.

Open Corpus – Question/Answer Pairs

Text: One method to constrain the hot wind properties directly is by X-ray observations. Recently, Strickland & Heckman (2009) constrained the wind parameters in the archetypal nearby starburst galaxy M82 using hard X-ray observations of its central region, finding a high thermalization efficiency (~ 1) and a mass-loading efficiency of $M^{\text{hot}}/\text{SFR} \sim 0.5$. However, superwinds in other galaxies with star formation rates (SFRs) of $1 - 1000 M_{\odot} \text{ yr}^{-1}$ at both low and high redshift are much less well studied, and a more generic approach needs to be introduced to constrain the hot wind properties and to understand their dynamical importance for rapidly star-forming galaxies. Therefore, we apply the CC85 model across a wide range of galaxies from dwarf starbursts to ultra-luminous infrared galaxies (ULIRGs). By using the observed X-ray properties of galaxies we constrain the thermalization efficiency and mass loading of hot winds.

Can be Answered

Q: What type of observation can directly constrain hot wind properties?

A: X-ray observations can constrain hot wind properties.

Cannot be Answered

Q: What is the mass of the black hole at the center of the Milky Way galaxy?

A: It is a little over 4 Million Solar masses.

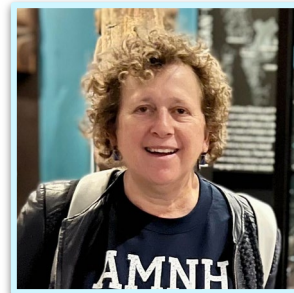
Universe TBD

International team that aims to democratize the use of LLM for astronomy.

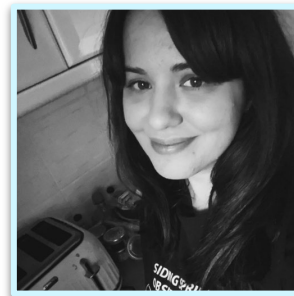
- Alberto Accomazzi collaborated on the release of AstroLLaMa
- Awarded grant for GPT4 on Azure via Dr. Alyssa Goodman of Harvard
- We hosted talks by Dr. Jo Ciucă and Dr. Yuan-Sen Ting this summer
- Sergi Blanco-Cuaresma and Kelly Lockhart collaborated on the development of SciX Brain



AstroLLaMA



Dr. Alyssa Goodman

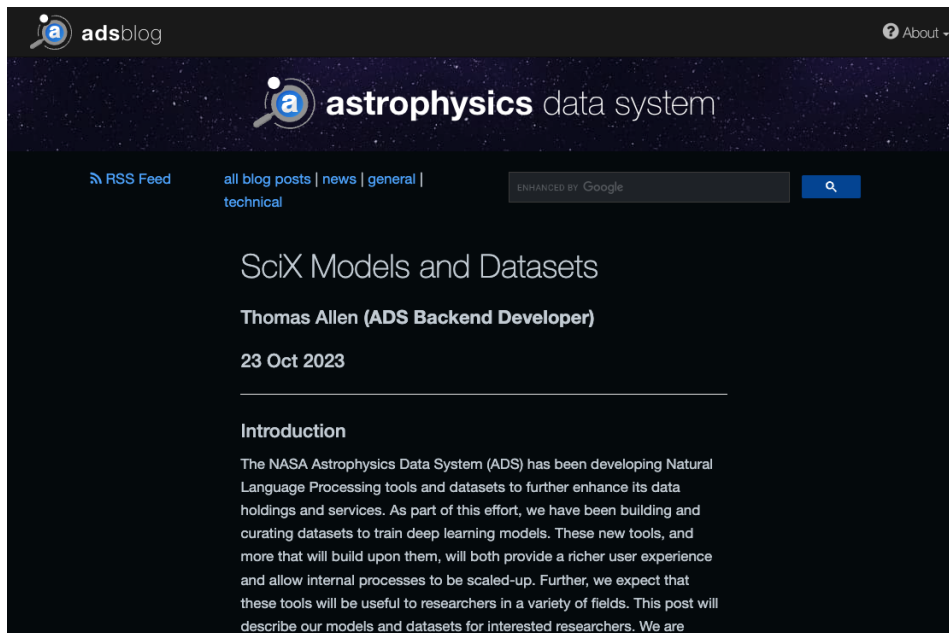


Dr. Jo Ciucă

Models & Datasets

Blog Post with links to more information about ADS models and datasets.

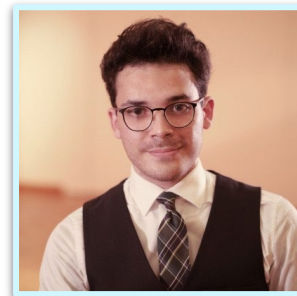
<https://ui.adsabs.harvard.edu/blog/ads-models-and-datasets>



The screenshot shows the top portion of a blog post on the ADS website. The header includes the 'adsblog' logo and an 'About' link. Below this is a dark banner with the 'astrophysics data system' logo. The navigation bar contains an RSS Feed icon, a search bar, and links for 'all blog posts', 'news', 'general', and 'technical'. The main content area features the title 'SciX Models and Datasets', the author 'Thomas Allen (ADS Backend Developer)', and the date '23 Oct 2023'. The 'Introduction' section begins with the text: 'The NASA Astrophysics Data System (ADS) has been developing Natural Language Processing tools and datasets to further enhance its data holdings and services. As part of this effort, we have been building and curating datasets to train deep learning models. These new tools, and more that will build upon them, will both provide a richer user experience and allow internal processes to be scaled-up. Further, we expect that these tools will be useful to researchers in a variety of fields. This post will describe our models and datasets for interested researchers. We are

Summer PhD Student Internship

- We hosted an internship for Atilla Kaan Alkan from the Université Paris-Saclay
 - designed and implemented tools to help with information extraction from Astronomer's Telegram
 - reference extraction
 - co-reference resolution
 - entity linking
 - relation extraction



Atilla Kaan Alkan