SF2A 2025
A. Siebert, K. Baillié, M. Béthermin, F. Cantalloube, E. Josselin, N. Lagarde, J. Malzac, J. Richard, L. Selliez, O. Venot (eds)

DATA FROM NENUFAR, AN SKA PATHFINDER: A CHALLENGE IN A CONSTRAINED LANDSCAPE.

B. Cecconi[1]

Abstract. NenuFAR (New Extension in Nançay Upgrading LOFAR) is a low-frequency radio telescope (10 to 85 MHz) located at the Observatoire de Radioastronomie de Nançay (ORN). With over 15 thematic groups, the instrument covers all fields of radio astronomy (from the solar system to the early Universe). NenuFAR can operate as an interferometer, as a phased array, or even as an extension of LOFAR. This flexibility in configuration multiplies the types of data products and associated processing, with data rates that can approach 15 TB per hour. The various receivers produce such large volumes of data that they have to be reduced before being transferred out of the ORN. Some processing is carried out on the fly, on the data stream, but in the majority of cases, scientists run their scripts a posteriori to produce the reduced data. NenuFAR's data infrastructure is made up of 3 main elements: the receivers that produce the initial data stream; the Nançay data center, which provides storage and computing resources to users for the initial data reduction; and finally, a distributed data center split between the Observatoire de Paris and the Data Centre of the Région-Centre-Val-de-Loire. While the computing resources at Nançay are conventional (independent servers with shared storage space), the distributed data center is built on a cloud architecture. This heterogeneous architecture requires coordinated management of data, processing software and computing and storage resources.

Keywords: Radio Astronomy, Data Processing, Machine Learning, Datacentre, NenuFAR, SKA

1 Introduction

The NenuFAR radio telescope (Zarka et al. 2012) is as an SKA pathfinder located in Nançay, France. It operates between 10 and 85 MHz (3–30 m wavelengths), with an open call twice per year, and a series of long term key projects covering a large span of science topics: Cosmic Dawn (Mertens et al. 2021; Munshi et al. 2024), Exoplanets and stars (Turner et al. 2023; Louis et al. 2025; Zhang et al. 2025), Pulsars, transients, fast radio bursts (FRBs) (Agar et al. 2021; Decoene et al. 2023; Wu et al. 2023; Brionne et al. 2025), Jupiter, planetary lightning, and solar observations (Lamy et al. 2023; Murphy et al. 2023, 2024), Radio recombination lines and galaxies (Bonnassieux et al. 2021; Stanislavsky et al. 2024; Cros et al. 2025), etc. NenuFAR is a genuine digital radio interferometer composed of more than 2000 full sky low frequency antennas. The electric signals captured by the antennas are combined in real time through different parallel data paths, defining several radio instruments that can observe simultaneously the same beam on the Sky: a standalone beamformer (single pixel, high temporal-spectral resolution), a radio imager (interferometric visibilities with coarse temporal resolution), a superstation mode for LOFAR (Zarka et al. 2012; van Haarlem et al. 2013) as well as a SETI receiver.

Science teams develop their pipelines and process their data on the Nançay Data Center (CDN). The NenuFAR team also provides diverse tools for imaging or beamforming data processing. A real-time detection framework (for, e.g., solar spikes, Jupiter S-bursts) is also available. Eventually, the reduced data will be pushed to the NenuFAR Data Center, which is currently under development.

In this paper, we present the data management challenges raised while developing and operating NenuFAR and its data centre.

[1] LIRA, Observatoire de Paris, Université PSL, Sorbonne Université, Université Paris Cité, CY Cergy Paris Université, CNRS, 92190 Meudon, France

2 Data and processing diversity

The NenuFAR radio telescope generates massive raw data streams, with a single radio source observed 8 hours a day producing approximately 68 GB/s of raw data. This corresponds to around 715 PB of raw data per year. To manage this data, it is decimated in real-time and sent to various receivers, such as the radio imager (18.5 GB/s) and the beamformer (1.1 GB/s). The final acquisition data rate depends on the observation configuration but will still need to be reduced through post-processing by the science teams to meet the planned data production of about 2 PB per year.

The majority of the data processing is performed using the resources of the CDN in Nançay, which provides a computing facility with a dozen servers and approximately 2 PB of mass storage. The NenuFAR data can be processed using community tools developed for other instruments, such as PRESTO (Ransom 2011), AOFlager (Offringa et al. 2012), KERN (Molenaar 2023), DP3 (Dijkema et al. 2023), killMS (Tasse 2023), DDFacet (Tasse et al. 2023), and others. The CDN technical team maintains these community software using various software configuration managers or containers, such as Singularity (Kurtzer et al. 2017) or Ansible.

In addition, the NenuFAR team provides a Python package named NenuPy (Loh et al. 2023) and a set of IDL routines called `readnuspec.pro`.

According to the NenuFAR science management plan (available from the NenuFAR instrument web page: `https://nenufar.obs-nancay.fr/en/astronomer/`), the NenuFAR data will be made public eventually (after 1 year in most cases). Since each team is developing their own pipeline based on their expertise in their science cases, the NenuFAR team needs to store this knowledge in useful metadata. This is the role of a data management plan (DMP). The NenuFAR DMP is still under development by the CDN team. The main data processing levels are: *Level 0* (Raw data); *Level 1* (Pre-processed data); and *Level 2* (Reduced data). In general, only *Level 2* data will be pushed to the NenuFAR data facility. Since each science team is defining its own data management processing strategies and science data products, ongoing interactions between the CDN team and the NenuFAR science teams are necessary to define the elements of the NenuFAR DMP and the sub-plans per team.

The DMP is prepared to comply with the maDMP recommendation (Miksa et al. 2020). An online DMP management tool, *Data Stewardship Wizard* (Pergl et al. 2019), will be used, thanks to developments conducted within the OSTrails project (`https://ostrails.eu`) at Observatoire de Paris. The goal is to use this tool to store all the metadata required to manage the data life cycle, including their provenance metadata. Virtual observatory interfaces will eventually be implemented on top of the NenuFAR data facility. However, compared to the current DMP design tasks, we believe this last part should be straightforward, given the expertise already in place at Observatoire de Paris. These interfaces will include the classical astronomy standards developed by the IVOA (International Virtual Observatory Alliance) and other standard interfaces dedicated to the heliophysics community (as recommended by the IHDEA, International Heliophysics Data Environment Alliance). Such interfaces are already implemented at CDN and Observatoire de Paris on other similar datasets.

The draft data architecture of NenuFAR is illustrated in Figure 1. This infrastructure is notable for its distributed nature. The NenuFAR instrument, its backend, receivers, and the CDN facility form the “edge” component. Conversely, the NenuFAR distributed data facility and its associated components constitute the “cloud” component. Currently, the cloud component comprises storage and compute resources in Meudon (Observatoire de Paris), storage and compute resources within the Centre-Val-de-Loire regional datacenter hosted by BRGM (Bureau des Ressources Géologiques et Minières), and cold storage at MPCDF (Max Planck Computing and Data Facility) in Germany.

3 The EXTRACT contribution

The EXTRACT (`https://extract-project.eu`) project is developing a software platform that simplifies the complexities of big data environments. The proposed framework enables the orchestration of “extreme data” processing across a “compute continuum”. Extreme data encompasses distributed or big data (highly heterogeneous, high data rate, large volume of data, etc.), while the compute continuum highlights the use of edge, cloud, or high-performance computing (HPC) resources. Figure 2 illustrates the inclusion of data management policies, data mining frameworks, security, traceability, and other relevant components within this project.

Within the TASKA use-case (Transient Astrophysics with an SKA pathfinder), the project focusses on the data processing of NenuFAR. The primary goals are to develop a real-time detection framework implemented on the beamforming receiver (edge component) and an optimised computing framework for the imaging pipelines on the distributed data facility (cloud component).

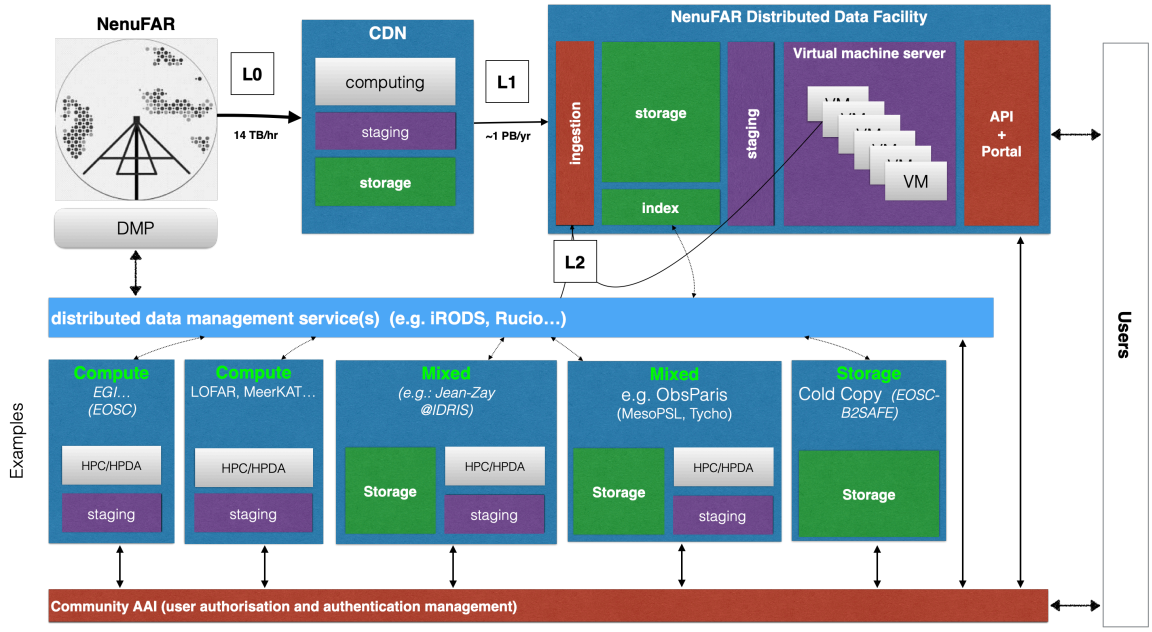

Fig. 1. NenuFAR distributed data architecture. The top row shows the NenuFAR instrument and its receivers, which feed the CDN storage with Level 0 data. The CDN facility produces Level 1 datasets, while the NenuFAR distributed data facility stores and publishes the data. The lower row shows the distributed infrastructures, which connect multiple nodes and provide additional resources.

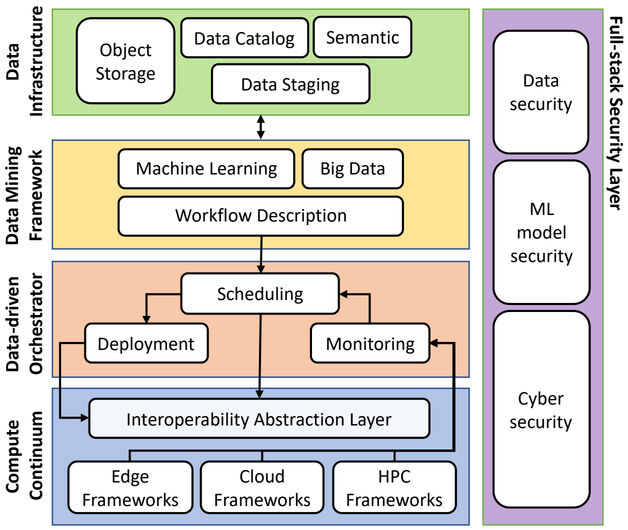

Fig. 2. EXTRACT framework conceptual stack

NenuFAR's beamformer has been equipped with real-time data mining capabilities, leveraging EXTRACT technologies (Mauduit 2025). This real-time data stream is analysed to detect "solar spikes" (Murphy et al. 2023, 2024) or Jupiter "S-bursts" (Mauduit et al. 2023). The system can trigger events, populate event databases, and make decisions about data reduction. When using artificial intelligence (AI) algorithms, model serving technology is employed to update AI coefficients. Currently, the component is implemented on a spare receiver.

NenuFAR's distributed data centre component will eventually integrate the upcoming French SKA Regional Centre infrastructure, currently being developed through the NumPEX PEPR project with HPC and exascale solutions. The EXTRACT solution developed for this purpose offers another viable option for distributed computing, particularly in cloud-based solutions.

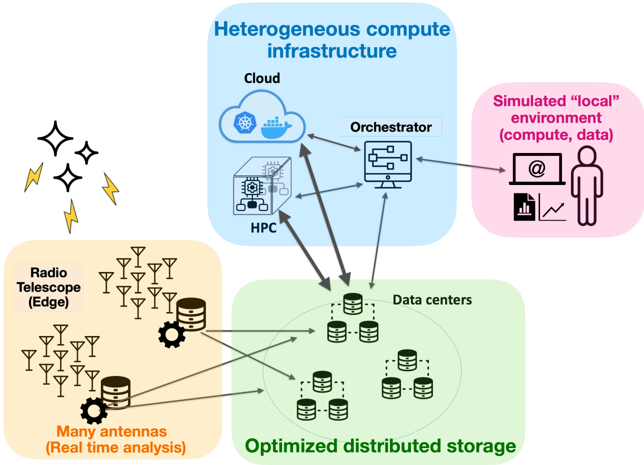

Fig. 3. NenuFAR data and compute continuum

EXTRACT's goal is to provide a dynamic and transparent distributed workflow system. This system can handle tedious tasks such as radio interferometric imaging and calibration across multiple sub-bands. Before running the workflow in batch mode, users can test and validate their results. The workflow is based on a Jupyter notebook interface, and each workflow step is embedding generic radio astronomy libraries and software, and implements the required interfaces with the EXTRACT framework. Orchestration is handled by COMPSs at workflow level (Tejedor et al. 2017), and Lithops at task level (Sampé et al. 2023), cloud data proxy by SkyStore (Liu et al. 2025), data catalogue by Nuvla.io (`https://nuvla.io/`), and data partitioning by Dataplug (Arjona et al. 2024). Local notebooks can be used for remote processing, and optimisation is based on processing time.

In addition, as reported by Nammour (2025), an improved Deep Learning radio imaging algorithm for transient sources is developed using video leveraging temporal structure, and will be implemented as a new workflow step within TASKA.

Figure 3 summarises schematically where the EXTRACT framework is complementing the NenuFAR data facilities and capabitilies.

4 Constraints

Compared to other similar telescopes, NenuFAR is relatively inexpensive due to its design, currently estimated to be around 7 Meuro . However, the cost of the data infrastructure is similar to other instruments, amounting to about 3 M€. This means that NenuFAR's data facility incurs a proportionally much higher cost (much higher than the usual 10% of the total cost), making it challenging to secure funding.

In the context of global warming, the NenuFAR data facility must be developed in an environmentally sustainable manner. Building a new datacenter is not an option, so the project team decided early on to reuse existing facilities. This approach not only reduces the carbon impact but also allows the facility to scale up when needed, taking advantage of unused shared resources.

The distributed NenuFAR data facility, spread over Meudon and Orléans, is currently designed to hold approximately 10 PB of data. The compute infrastructure includes HPC and cloud capabilities, and access is planned through JupyterLab technology. This infrastructure is supported by the Centre-Val-de-Loire region, CNRS, and Observatoire de Paris through the ORION-CVL* project. Most of the hardware has been acquired, but installation and configuration are still ongoing. Additionally, the maintenance costs are not covered by the initial funding. The contracting scheme did not allow for the inclusion of personnel costs nor sustainability costs.

Another challenge is to encourage science teams to adhere to open science, as well as good data and software management practices, which are often perceived as additional burdens. While open data is well-accepted in the astronomy community, researchers are usually unwilling to actively participate in data publication and rely

*ORION-CVL: *Un Observatoire Radio International en régiOn Centre-Val de Loire*, a CPER (Contrat Plan État Région) contract with the Centre-Val-de-Loire region.

on data centres for curation. However, for an instrument like NenuFAR, we need to start from the teams' needs. We have begun drafting Data Management Plans for each science team, but most of the work remains unfinished to date.

Concerning data security and privacy issues, NenuFAR data does not include any personal data, so we should be safe on this aspect. However, one of the major yet-to-be-solved issues will be to grant access to facilities for NenuFAR users, who could be from various countries. Security policies in data centres are often very strict, especially for researchers from certain countries (such as, e.g., China).

The NenuFAR data facility is currently a prototype due to a lack of manpower. While all the expertise is present, insufficient personnel prevents the project from progressing. Additional resources, such as cloud and HPC system administrators, and AI/workflow software engineers, are required.

In conclusion, the primary constraints on NenuFAR's implementation are the lack of manpower, financial support difficulties, security policies of hosting datacentres, and environmental factors.

5 Opportunities

Thanks to the EXTRACT project, we have developed and tested a workflow engine for a distributed infrastructure. This engine is fully applicable to the distributed NenuFAR data facility. We tested our minimal demonstration workflow on several cloud infrastructures provided by EGI (the European Grid Infrastructure) and CESNET (based on Rancher technology), EOSC (based on OKD technology), and the Observatoire de Paris (also based on OKD). This testing allowed us to consolidate our workflow engine, enabling it to run on multiple cloud clusters — a crucial aspect of a distributed data infrastructure. Currently, our workflow optimisation scheme is based on processing time, but the Lithops library was originally designed to facilitate optimisation on various metrics, such as the operational costs of cloud providers. We plan to introduce an optimisation based on carbon impact. We started to collaborate with the NumPEX PEPR team to bridge with their HPC-based optimisation schemes. The two approaches are very complementary. In a scientific workflow, some steps do not require parallel computing and can be implemented simply by splitting the input data into many small tasks. The EXTRACT framework can efficiently orchestrate tasks based on the computing requirements between cloud and HPC.

The EXTRACT project also provided us with the chance to implement a real-time processing analogue to the information framework within the beamform receiver of NenuFAR. Additionally, we began developing an innovative radio interferometry imaging algorithm for transient events, inspired by machine learning technologies and similar to video reconstruction techniques.

Managing NenuFAR's data is a pathfinder project on its own. We benefitted from the development of OSTrails. This project is structured into three pillars: planning, tracking, and assessment. For planning, we adopted a DMP management tool. We also prototyped a scientific knowledge graph database that connects datasets, publications, institutions, persons, and instruments. This database can be used to track the impact of datasets. Finally, we defined astronomy-tailored FAIR metrics to assess our research products.

Finally, the work for the French contribution to the SKA Regional Centre network (FrSRC) began a few months ago. The FrSRC facility will be a distributed facility that will also host the SKA precursors and pathfinders datasets. NenuFAR data will thus eventually be stored, processed, and published within FrSRC. The NenuFAR data facility team is ready to participate in this critical task and provide feedback based on their tests and experience.

6 Conclusion

Developing the NenuFAR data facility is a massive ongoing task, but we have overcome most challenges. Our main obstacle is a lack of manpower, despite our teams having all the necessary expertise. EXTRACT has shown us the potential of the cloud, and we should strengthen collaboration with the NumPEX team within the FrSRC project for the benefit of the entire community.

This work would not have been possible with the huge work provided by the technical and scientific NenuFAR team, the DIO team at Observatoire de Paris, the BRGM team as well as the colleagues in Orléans, at LPC2E and OSUC. This work has been supported by the OSTrails project (Open Science Plan-Track-Assess Pathways, doi:10.3030/101130187), which has received funding from the European Union's Horizon Europe programme under grant agreement number 101130187; and the EXTRACT project (A distributed data-mining software platform for EXTReme dAta Across the Compute conTinuum, doi:10.3030/101093110), which has received funding from the European Union's Horizon Europe programme under grant agreement number 101093110. We also thank the organisers of the SF2A 2025 for giving me the opportunity to report in this topic.

References

Agar, C. H., Weltevrede, P., Bondonneau, L., et al. 2021, MNRAS, 508, 1102

Arjona, A., García-López, P., & Barcelona-Pons, D. 2024, in 2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing (CCGrid), 567–576

Bonnassieux, E., Tremou, E., Girard, J. N., et al. 2021, Galaxies, 9, 105

Brionne, M., Grießmeier, J. M., Cognard, I., et al. 2025, A&A, 693, A96

Cros, L., Gusdorf, A., Salomé, P., et al. 2025, arXiv e-prints, arXiv:2506.08895

Decoene, V., Zarka, P., Ng, C., et al. 2023, in SF2A-2023: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. M. N'Diaye, A. Siebert, N. Lagarde, O. Venot, K. Bailliée, M. Béthermin, E. Lagadec, J. Malzac, & J. Richard, 137–141

Dijkema, T. J., Nijhuis, M., van Diepen, G., et al. 2023, DP3: Streaming processing pipeline for radio interferometric data, Astrophysics Source Code Library, record ascl:2305.014

Kurtzer, G. M., Sochat, V., & Bauer, M. W. 2017, PLoS ONE, 12, e0177459

Lamy, L., Duchêne, A., Mauduit, E., et al. 2023, in Planetary, Solar and Heliospheric Radio Emissions IX, ed. C. K. Louis, C. M. Jackman, G. Fischer, A. H. Sulaiman, & P. Zucca, 103097

Liu, S., Mo, X., Hershcovitch, M., et al. 2025, arXiv e-prints, arXiv:2502.20818

Loh, A., Coutouly, L., Girard, J. N., Mauduit, E., & Viou, C. 2023, AlanLoh/nenupy: 2023 release

Louis, C. K., Loh, A., Zarka, P., et al. 2025, arXiv e-prints, arXiv:2503.18733

Mauduit, E. 2025, Real-time detection of Solar and Jovian radio bursts with NenuFAR: advancing astrophysical data mining with the EXTRACT project.

Mauduit, E., Zarka, P., Lamy, L., & Hess, S. L. G. 2023, Nature Communications, 14, 5981

Mertens, F. G., Semelin, B., & Koopmans, L. V. E. 2021, in SF2A-2021: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. A. Siebert, K. Baillié, E. Lagadec, N. Lagarde, J. Malzac, J. B. Marquette, M. N'Diaye, J. Richard, & O. Venot, 211–214

Miksa, T., Walk, P., & Neish, P. 2020, RDA DMP Common Standard for Machine-actionable Data Management Plans

Molenaar, G. 2023, KERN: Radio telescope toolkit, Astrophysics Source Code Library, record ascl:2305.012

Munshi, S., Mertens, F. G., Koopmans, L. V. E., et al. 2024, A&A, 681, A62

Murphy, P. C., Aicardi, S., Cecconi, B., Briand, C., & Peccoux, T. 2024, The Open Journal of Astrophysics, 7, 51

Murphy, P. C., Cecconi, B., Briand, C., & Aicardi, S. 2023, in Planetary, Solar and Heliospheric Radio Emissions IX, ed. C. K. Louis, C. M. Jackman, G. Fischer, A. H. Sulaiman, & P. Zucca, 04055

Nammour, F. 2025, Radio imaging cloud workflow for multiband transient source reconstruction

Offringa, A. R., van de Gronde, J. J., & Roerdink, J. B. T. M. 2012, A&A, 539, A95

Pergl, R., Hooft, R., Suchánek, M., Knaisl, V., & Slifka, J. 2019, Data Science Journal

Ransom, S. 2011, PRESTO: PulsaR Exploration and Search TOolkit, Astrophysics Source Code Library, record ascl:1107.017

Sampé, J., Sánchez-Artigas, M., Vernik, G., Yehekzel, I., & García-López, P. 2023, IEEE Transactions on Cloud Computing, 11, 1026

Stanislavsky, L. A., Bubnov, I. N., Stanislavsky, A. A., et al. 2024, A&A, 683, A7

Tasse, C. 2023, killMS: Direction-dependent radio interferometric calibration package, Astrophysics Source Code Library, record ascl:2305.005

Tasse, C., Hugo, B., Mirmont, M., et al. 2023, DDFacet: Facet-based radio imaging package, Astrophysics Source Code Library, record ascl:2305.008

Tejedor, E., Becerra, Y., Alomar, G., et al. 2017, IJHPCA, 31, 66

Turner, J. D., Zarka, P., Grießmeier, J. M., et al. 2023, in Planetary, Solar and Heliospheric Radio Emissions IX, ed. C. K. Louis, C. M. Jackman, G. Fischer, A. H. Sulaiman, & P. Zucca, 04048

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2

Wu, Z., Coles, W. A., Verbiest, J. P. W., et al. 2023, MNRAS, 520, 5536

Zarka, P., Girard, J. N., Tagger, M., & Denis, L. 2012, in SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. S. Boissier, P. de Laverny, N. Nardetto, R. Samadi, D. Valls-Gabaud, & H. Wozniak, 687–694

Zhang, X., Zarka, P., Girard, J. N., et al. 2025, arXiv e-prints, arXiv:2506.07912