ALMA Pipeline Evolution

D. Muders, A. Babai, K. Berry, S.T. Booth, P. Bos, D. Brisbin, L. Di Mascolo, H. Ezawa, H. Francke, L. Gómez, A. Guzmán, J. Hibbard, J. Hoskins, T. Hunter, I. Ihnatiev, R. Indebetouw, D. Kunneriath, A. Lipnicky, R. Loomis, C. Marocico, L. Maud, F. Montesino Pouzols, T. Nakazato, M. Neeleman, P. Sheehan, K. Sugimoto, K. Taniguchi, K. Torstensson, A. Vibhute, R. Xue, I. Yoon, A. Yoshino

Max-Planck-Institut für Radioastronomie

MAX-PLANCK-GESELLSCHAFT

Overview

The **CASA/ALMA Pipeline** is the standard data processing software for ALMA, developed by an international team over 20+ years. It automatically calibrates and images more than 95% of all ALMA datasets through a parallelized stage/task design with automatic data selection and data-driven heuristics.

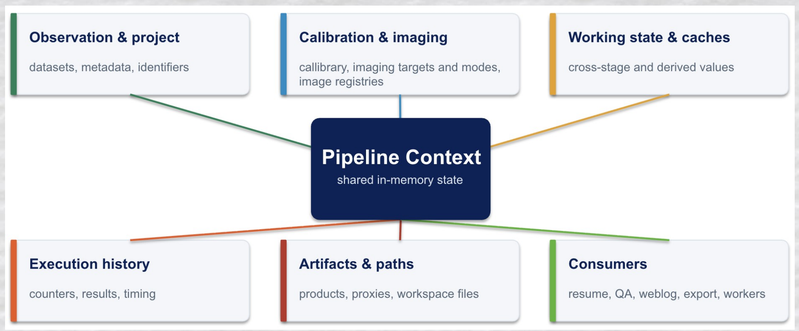
The ALMA2030 Wideband Sensitivity Upgrade (WSU) with much larger observing bandwidth and better spectral resolution will increase ALMA data volumes by orders of magnitude. Since the current pipeline does not scale beyond single-node setups, a new workflow-based processing system is foreseen. We present ongoing work to prepare the pipeline for this transition:

Faster imaging — refactoring interferometry imaging (the most time-consuming step) with faster continuum detection and more efficient parallelization

Scalable Pipeline Context — modularizing and distributing the central hub that records results and transports intermediate products

Robust QA scores — improved quality assessment for reliable "Fast Lane" delivery of pipeline products directly to PIs

Modern documentation — a “Read the Docs”-based restructuring to streamline content and ease migration to the new system



Current Pipeline Context Responsibilities

Scalable Pipeline Context

The ALMA pipeline is stateful: it records the data being processed and the results of each step in the "Pipeline Context" entity. During data import, metadata are analyzed into an in-memory model of the observations (datasets, antennas, fields, sources, spectral windows, scans). Task results are linked to the Context; calibration tables and images/cubes reside in lookup libraries, and globally useful quantities (beam sizes, sensitivities) are shared via a central hub for reuse.

The current implementation — a Python class hierarchy serialized via pickling — has grown in extent and scope, and its parallelization requires continuous pickle/unpickle cycles. This is untenable for the expected workflow setup. Our in-depth analysis identified four major refactor areas:

- **Domain/non-domain state separation** — decoupling observation and processing state from execution bookkeeping (task history, scheduling), which orchestration software such as Prefect can take over.
- **Explicit task interfaces** — declaring what each task reads and produces, with a defined way of combining concurrent changes, enabling independent scheduling, retries, and reruns.
- **Artifact records with provenance** — unique identity per product, independent of file location, including the producing task and its inputs — improving traceability and enabling partial reprocessing from an intermediate data calibration and flagging state.
- **Versioned serialization** — replacing pickles with a documented schema and migration strategy for restoration by later software versions.

These recommendations support parallel execution, task retries, partial reprocessing, and recovery — while preserving the scientific information required by existing pipeline algorithms.

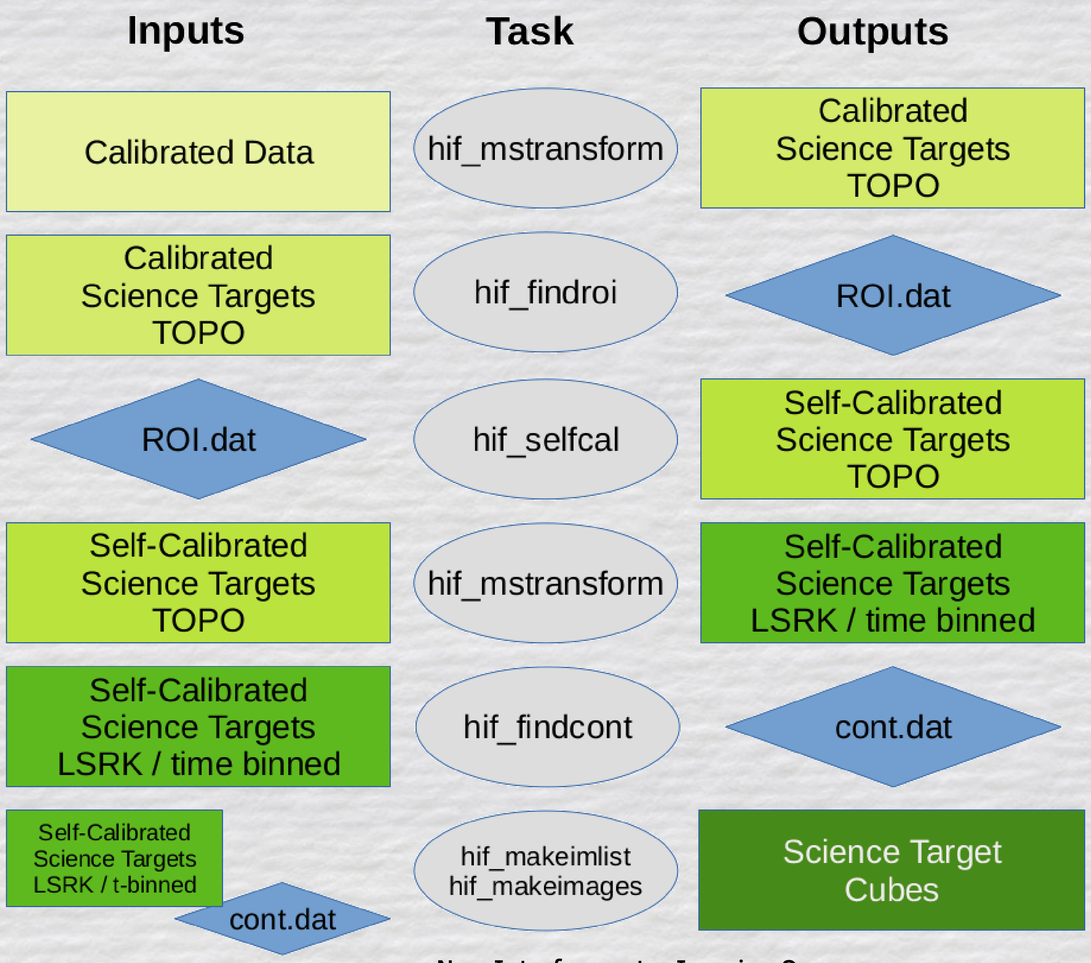
Faster Imaging

Imaging will be the dominant processing cost for future ALMA data. The Radio Astronomy Data Processing System (RADPS), under development at NRAO for ALMA and the ngVLA, is workflow-based: existing pipeline heuristics can be largely reused, but the algorithms must be reviewed to remove convoluted data flows and simplify steps. This is the aim of our ***Imaging Refactor*** effort.

The interferometry imaging pipeline consists of five building blocks — image pre-check, continuum frequency range finding, continuum subtraction, self-calibration, and cleaning — with the preparatory steps currently consuming significant time.

Key improvements:

- **Efficient pre-checks** — weight arrays computed once and reused instead of repetitive synthesized-beam calculations; finer sampling of imaging parameters to find optimum values
- **Fast continuum finding** — experimental method (2026 pipeline) derives frequency ranges from coarsely sampled data, sufficient for self-calibration; "Regions of Interest" may later include spatial information for selective processing
- **Smarter mitigation** — matrix lookups replace step-by-step walks through parameter space to avoid extremely large data products
- **Earlier self-calibration** — allows subsequent time binning to reduce the cleaning workload
- **Early frame conversion** — TOPO → LSRK/SOURCE conversion upfront avoids iterative forward/backward calculations during cleaning
- **Efficient parallelization** — splitting datasets by source and spectral window avoids file locking issues



New Interferometry Imaging Sequence

Robust QA Scores

Normalized numerical Quality Assessment (QA) scores with defined thresholds (perfect / good / medium / bad) rate the quality of each pipeline step and of the overall run. Years of iteration have made these scores trustworthy enough to automatically release pipeline products to the ALMA archive and the PI — the so-called "Fast Lane" processing.

This year's work focused on the single dish pipeline QA scores to increase the fraction of projects deliverable in Fast Lane mode:

- **New sensitivity score** — based on the ratio of observed to theoretical sensitivities for the imaging step
- **Readable messages** — QA score explanations reorganized and aggregated for easier interpretation
- **Fewer false positives** — fine-tuned detection of atmospheric line contamination in the data

Modern Documentation

In 2026, the ALMA/VLA/NRO Pipeline overhauled its documentation infrastructure, transitioning from disconnected static PDFs and Wiki/Confluence pages to a continuous "docs-as-code" architecture integrated in the version-controlled codebase and hosted on “Read the Docs”. Previously, task heuristics, WebLog descriptions, and QA scoring criteria were maintained in external LaTeX documents that lagged behind code releases. Key elements of the new system:

- **LLM-assisted migration** — agentic LLM workflows extracted and converted task-level heuristic descriptions into Google/Sphinx-style Python docstrings across all pipeline tasks, refined through human review by domain scientists and software developers
- **Reproducible builds** — Pixi-managed local and CI/CD builds produce dual formats from MyST Markdown, reStructuredText, and docstrings: a responsive HTML site (client-side search, interactive Mermaid diagrams, automated ADS citation resolution) and two LaTeX/PDF manuals (User's Guide, Task Reference Manual)
- **Public/internal variants** — BUILD_INTERNAL_DOCS=1 segregates developer notes and internal API trees
- **Version-pinned docs** — builds map to specific pipeline releases, so archive investigators can access documentation matching the version that processed their observations
- **Clean runtime help** — CASA terminal wrappers strip raw markup to plain text while preserving rich formatting for online and print manuals

Content by NI, language improvements
by AI (Chat-AI from GWDG with GLM 5.3 Flash)

ALMA ATACAMA LARGE MILLIMETER ARRAY · ESO · National Astronomical Observatory of Japan · NRAO · Max-Planck-Institut für Radioastronomie · MAX-PLANCK-GESELLSCHAFT