Th229-ScanBench: A Public Dataset and Benchmark for Detecting Time-Varying Signals in Thorium-229 Nuclear Clock Records

Abstract

Th229-ScanBench is a public dataset and benchmark built from the JILA scan-level frequency-reproducibility record for solid-state thorium-229 nuclear clocks. It addresses a practical gap: thorium-229 spectroscopy is becoming relevant for time-domain tests of fundamental physics, but public records remain too sparse for unconstrained machine-learning development. The main task uses the 55 temperature-corrected peak-b measurements and asks whether an additive sinusoidal frequency modulation is present at the observed timestamps. Each example preserves the measured cadence, crystal identities, temperatures, and frequency uncertainties; controlled injections form 6,336 labeled examples with fixed train, validation, and test splits. The benchmark targets a broader AI problem in precision science: sparse, irregular, heteroscedastic time series with trusted metadata but few observations. Among leakage-free baselines, a hierarchical sinusoid model with crystal-specific offsets and jitter is strongest, reaching test AUROC 0.774 and average precision 0.827. A generic random forest on periodogram and summary features reaches AUROC 0.712 and average precision 0.769. We also report a C10+C13-only diagnostic because the X2-only observed series remains poorly calibrated under the selected null; we treat this as excess scatter or model mismatch, not as a signal.

1 Introduction

The JILA thorium-229 frequency-reproducibility measurement released a public scan-level record of the nuclear transition in ^{229}Th:CaF_2 crystals (Ooi et al., 2026, 2025). After applying the published-subset logic, it contains 73 scan-level line measurements, of which 55 are for peak b. This record is too small for unconstrained high-capacity supervised learning, but large enough for a focused injection question: if a weak time-domain modulation were added at the same observation times, could detection methods recover it?

This benchmark is useful now because thorium-229 nuclear-clock spectroscopy is becoming a promising setting for time-domain tests of fundamental physics, while public records remain sparse. Future experiments may produce longer clock records where algorithms for weak, time-varying frequency shifts will matter, but those methods need realistic test cases before such records are available. Th229-ScanBench provides one such test case by fixing the measured cadence, crystal identities, temperature metadata, scan-level uncertainties, and instrument-dependent scatter, then creating labels through controlled injections.

The resulting task is also useful beyond thorium spectroscopy. Many physics measurements produce short, unevenly spaced time series with pointwise uncertainties, nuisance conditions, and instrument-dependent scatter. Standard ML benchmarks often assume dense grids, exchangeable examples, or large training sets. Th229-ScanBench instead keeps the actual measurement structure and asks a simple detection question that is hard enough to compare physics-informed detectors with generic feature-based classifiers.

Submitted to 2026 Conference on Physics and AI (PAI26), Datasets and Benchmarks Track.

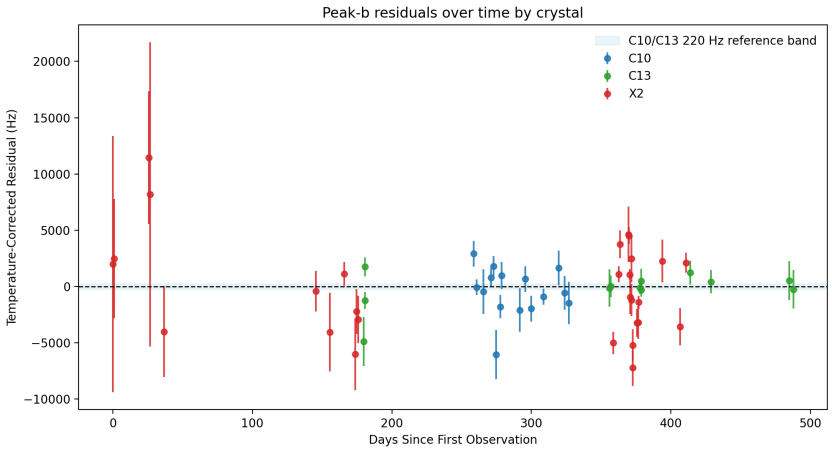

Figure 1: Peak-b temperature-corrected residuals on the observed JILA cadence. Error bars show reported frequency uncertainty; the shaded band marks the ± 220 Hz C10/C13 reproducibility scale.

We remove the measured temperature-dependent frequency trend, then create labeled examples by adding or not adding a sinusoidal frequency shift at the observed JILA cadence. This is not a claim that the observed series contains a beyond-Standard-Model signal. The physical motivation is modest: ^{229}Th spectroscopy is relevant to searches for slowly varying fields, including ultralight dark matter, but converting an injected Hz-level modulation into a coupling limit is model dependent (Fuchs et al., 2025).

2 What measurements we use

The source CSV contains 428 rows because scans appear with several correction and fit choices. The readme states that the published subset uses Lorentzian fits with the `mem int` correction, except that the earliest X2 rows use memory-corrected Lorentzians because no intensity data were available. Applying this literally gives 72 rows. The CSV and folder inventory contain 73 unique Lorentzian scan keys, matching the JILA paper's 73 line scans; the extra row is a 2024-09-25 X2 peak-c scan with `mem` but no `mem int`. We use the canonical 73-row scan-level subset and record both rules in `subset_reconciliation_scan_keys.csv`. The main peak-b count is 55 under both choices.

Peak b is the default task because the JILA paper identifies it as the lower-temperature-sensitivity line; peak c is optional/exploratory only. The 55 peak-b rows span 487.73 days and contain 15 C10, 12 C13, and 28 X2 measurements. We fit a quadratic temperature trend using canonical C10 and C13 records, subtract it, and remove a target-peak weighted mean offset. The strict quality-control subset with $\chi^2_{\mathrm{red}} \leq 10$ contains 41 rows but is not the default. The public release includes the canonical 73-row subset, the primary 55-row peak-b table, and the fixed benchmark catalog used in this paper.

3 How the benchmark is formed

One benchmark example is a length-55 residual vector on the observed peak-b timestamps. Null examples are sampled from the selected crystal-aware null model. Signal examples add

$$\delta\nu(t) = A\cos(2\pi f t + \phi)$$

to the same sampled residuals. The period grid is 7, 10, 14, 21, 30, 45, 60, 90, 120, 180, 240, and 365 days. The amplitude grid is 100, 200, 350, 500, 750, 1000, 1500, 2200, 3200, 4500, and 6500 Hz. For each period-amplitude cell, we use 24 phases and both labels, giving 6,336 examples.

The null model is fitted only on an observation-level training subset of the measured peak-b residuals. The selected model is Gaussian with crystal-specific extra variance and an X2-only Gaussian outlier component. Splits are fixed by phase index inside every period-amplitude cell: 14 phases

Table 1: Representative-period detection performance. AUROC/AP compare injected test examples at period P against all held-out null examples. A_{95} is the amplitude needed for 95% detection at validation-calibrated 5% false-positive rate; intervals are 16–84% bootstrap ranges. RF A_{95} is omitted because no RF bootstrap curve was computed.

Method	P	AUROC	AP	A_{95} (Hz)
Weighted harmonic	30 d	0.767	0.469	2950 (2025–3075)
Weighted harmonic	180 d	0.599	0.211	4419 (4392–4435)
Gen. Lomb-Scargle	30 d	0.762	0.480	2950 (2025–3075)
Gen. Lomb-Scargle	180 d	0.721	0.421	3075 (2950–3117)
Hier. sinusoid	30 d	0.825	0.651	969 (938–979)
Hier. sinusoid	180 d	0.727	0.449	2950 (2156–3075)
Random forest	30 d	0.781	0.513	–
Random forest	180 d	0.689	0.331	–

for training, 5 for validation, and 5 for test. This gives 1,848 null and 1,848 injected training examples, 660 null and 660 injected validation examples, and 660 null and 660 injected test examples. Detection thresholds are set only from validation null examples.

Dataset card. The main track contains 55 real peak-b scans and 6,336 synthetic examples: 3,168 null and 3,168 injected. Each input is a length-55 residual-frequency vector. The task is binary modulation detection; splits are fixed by phase index (14/5/5 per period-amplitude cell); metrics are AUROC, AP, and A_{95}. Curated data follow the upstream CC BY 4.0 terms, and code is MIT licensed.

4 Detection methods and evaluation

We evaluate three interpretable sinusoid baselines. Weighted harmonic regression fits a sinusoid at each searched frequency with uncertainty weights and crystal-specific offsets. Generalized Lomb-Scargle uses the same weighted search but with one global offset (Lomb, 1976; Scargle, 1982; Zechmeister and Kürster, 2009). The hierarchical sinusoid model uses a shared sinusoid with crystal-specific offsets and the selected crystal-aware uncertainty scale; all three methods score the best residual-improvement over the search grid.

We also include a small generic ML baseline to show how the resource can be used outside hand-designed periodogram tests. The random forest uses 100 trees and is trained on the training split only. Its features are weighted periodogram powers at the 12 benchmark periods plus mean, variance, minimum, maximum, range, peak-to-peak amplitude, and lag-1 autocorrelation. The validation split is used only to calibrate a 5% null threshold, matching the other baselines.

We report AUROC, AP, test false-positive rate at a validation-calibrated 5% false-positive threshold, and A_{95}, the interpolated injected amplitude at which the test detection rate reaches 95%. A_{95} intervals resample validation-null scores and the five test phases per cell; they do not include uncertainty in the measured frequencies, temperature correction, or null-model choice.

5 Results

The hierarchical sinusoid model is the strongest safe baseline on the full held-out benchmark: AUROC 0.774, AP 0.827, and test-null false-positive rate 0.042 at the validation-calibrated threshold. Generalized Lomb-Scargle reaches AUROC 0.724 and AP 0.776, and weighted harmonic regression reaches AUROC 0.700 and AP 0.748. The random forest reaches AUROC 0.712 and AP 0.769. This is below the hierarchical model but competitive with the simpler periodogram baselines, showing that the benchmark can expose differences across method families on a real sparse cadence without requiring a large neural model.

The A_{95} curves show where the sparse cadence is most favorable. For the hierarchical sinusoid model, the best period in this grid is 30 days, with $A_{95} = 969$ Hz and bootstrap interval 938–979 Hz. Longer periods are harder: at 180 days the same model has $A_{95} = 2950$ Hz with a wider interval.

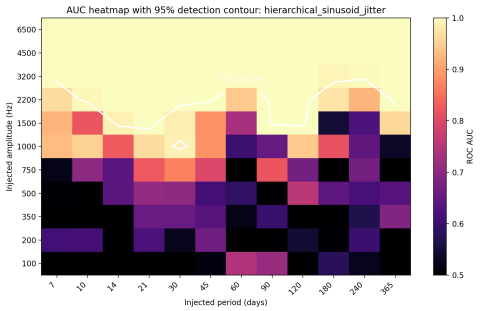

Figure 2: Cell-wise ROC AUC for the hierarchical sinusoid model across the injection grid. The white contour marks 95% detection at validation-calibrated 5% false-positive rate.

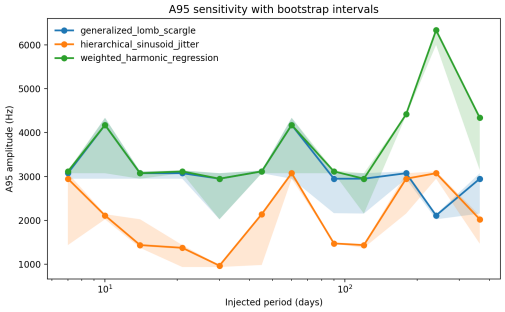

Figure 3: A_{95} sensitivity curves for safe sinusoid baselines. Lower is better; shaded bands show validation-threshold and test-phase resampling intervals.

We also score the observed JILA residual series without synthetic injection. This diagnostic is not a detection test. Under the selected null model, the all-peak-b series has empirical p-value 0.674 and the C10+C13-only subset has p-value 0.947. The X2-only subset has a low empirical p-value, 0.0039, with the strongest score near a 15.4-day period. Because X2 is already known to have larger scatter than C10 and C13, we treat this as residual scatter or model mismatch, not as a signal. Reporting the C10+C13-only diagnostic alongside the default 55-row task makes that caveat visible without removing published measurements from the benchmark.

6 Scope, reuse, and limitations

This benchmark is intentionally narrow. It uses scan-level frequency records, not photon-count or lineshape arrays. The uploaded pkl scan files are inventoried for future work but are not promoted into the v1 task. Injected amplitudes are reported in Hz; coupling conversions require extra model assumptions.

The reusable part of the resource is the separation between measured metadata and injected signal family. Future users can keep the same 55 timestamps, crystals, uncertainties, and split rules while adding non-sinusoidal injections, stress tests, or clean-subset evaluations such as C10+C13-only. New thorium records can be added by replacing the scan table and rerunning the same checks. Richer lineshape or photon-count inputs can become a separate track once the pkl files are fully validated.

The null model allows crystal-specific scatter and an X2 outlier component, but it does not fully explain the X2-only observed periodogram. We keep X2 in the default 55-row task because it is part of the published peak-b record, while making this caveat explicit. The benchmark is small: its value is not training large models from scratch, but fixing a reproducible detection problem on a scientifically meaningful cadence.

Availability: The curated dataset, benchmark generation code, baseline implementations, and evaluation scripts are available at `https://github.com/nurkyzaz/th229-scanbench` under the MIT license. The upstream JILA source data are from Ooi et al. (Ooi et al., 2025) under CC BY 4.0.

References

Tian Ooi, Jack F. Doyle, Chuankun Zhang, Jacob S. Higgins, Jun Ye, Kjeld Beeks, Tomas Sikorsky, and Thorsten Schumm. Frequency reproducibility of solid-state thorium-229 nuclear clocks. *Nature*, 650:72–78, 2026. doi:10.1038/s41586-025-09999-5.

Tian Ooi, Jack F. Doyle, Chuankun Zhang, Jacob S. Higgins, Jun Ye, Kjeld Beeks, Tomas Sikorsky, and Thorsten Schumm. Th-229 nuclear clock frequency record (JILA). Zenodo, 2025. doi:10.5281/zenodo.15751158.

Elina Fuchs, Fiona Kirk, Eric Madge, Chaitanya Paranjape, Ekkehard Peik, Gilad Perez, Wolfram Ratzinger, and Johannes Tiedau. Searching for dark matter with the thorium-229 nuclear lineshape from laser spectroscopy. *Physical Review X*, 15:021055, 2025. doi:10.1103/PhysRevX.15.021055.

Nicholas R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39:447–462, 1976. doi:10.1007/BF00648343.

Jeffrey D. Scargle. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982. doi:10.1086/160554.

Mathias Zechmeister and Manfred Kürster. The generalised Lomb-Scargle periodogram. *Astronomy and Astrophysics*, 496:577–584, 2009. doi:10.1051/0004-6361:200811296.